Beyond Isolated Operating Rooms: Risk-Aware Surgical Episode Scheduling in Single-Entry Networks

Pedram Peiroasfia,[a,b] Maryam Daryalal,[a,b]
[a]Department of Decision Sciences, HEC Montréal, Montréal, Québec H3T 2A7, Canada
[b]Groupe d'études et de recherche en analyse des décisions (GERAD), Montréal, Québec H3T 2A7, Canada
pedram.peiroasfia@hec.ca, maryam.daryalal@hec.ca

Vahid Roshanaei
Rotman School of Management, University of Toronto, Toronto, Ontario M5S 3E6, Canada
vahid.roshanaei@rotman.utoronto.ca

Long wait times for elective surgery are a persistent challenge in publicly funded health systems, where hospitals must coordinate limited capacity before, during, and after the operation, under considerable uncertainty. We study how a network of collaborating hospitals, such as the University Health Network in the City of Toronto, can centralize intake and jointly schedule same-day elective surgeries to improve access and efficiency under uncertainty. We formulate the problem using stochastic Mixed-Integer Programming (MIP) and Constraint Programming (CP) to capture the many linked resources used across a patient's full surgical episode, from pre-operative preparation through the operating room (OR) to post-operative recovery. Network-wide completion times of surgical episodes are controlled by incorporating coherent risk measures in the objective function, such as the risk-averse Conditional Value-at-Risk (CVaR). To solve this large-scale problem, we develop logic-based Benders decomposition methods that integrate MIP and CP in two- and three-level schemes, strengthened with problem-specific cuts, valid inequalities, and a graph-based timing reformulation. Our experiments show that the decomposition design matters more than the choice of solver. The multi-level decomposition improves scalability, with the three-level approach giving the most reliable performance. We also find that uncertainty in ORs can propagate and intensify downstream, causing congestion that models limited to OR scheduling might miss. These results show the importance of risk-aware network-level scheduling, especially in publicly funded healthcare systems where capacity, efficiency, and equitable access must be balanced.

Key words: Single-entry networks, Healthcare operations management, Operating room scheduling, Stochastic mixed-integer programming, Risk measures

1. Introduction

Healthcare systems worldwide are facing compounding pressures from reduced access, fiscal strain, and demographic shifts, particularly in publicly funded models such as those in Canada. The Canadian Institute for Health Information (CIHI) projects that Canadian healthcare spending will reach \$399 billion in 2025, equivalent to 12.7% of GDP, with public expenditure comprising about 71% of that total. It also identifies inflation, population growth, aging, and rising service utilization as principal cost drivers (CIHI, 2025a). These financial pressures coincide with rapid demographic change: Canadians aged 65 and older now make up 19.5% of the population (Statistics Canada, 2025), a group that tends to need more frequent and complex care. Access constraints have worsened alongside these trends, with CIHI reporting that median wait times for several priority cancer surgeries were longer in 2024 than in 2019, and that performance against established wait-time benchmarks has declined for selected procedures (CIHI, 2025b). Beyond these aggregate delays, patients waiting for the same procedure in the same city can face wait times that differ by many months to over a year depending on which surgeon they are referred to, disproportionately affecting patients of color, recent immigrants, and patients with limited proficiency in the dominant local language (CIHI, 2023; Duraikannan, 2022). Meeting such growing and increasingly complex surgical demand under tightening fiscal conditions requires more than classical operating room (OR) scheduling in isolation. It demands coordinated, network-wide decisions that account for capacity across the full care pathway and remain robust to the uncertainty inherent in surgical durations and patient flow.

Dual inefficiencies, namely long patient waits and underused resources, can and often do occur simultaneously in surgical care. This is largely a consequence of how perioperative care is structured, in which a patient moves through a sequence of stages, including pre-operative holding, surgery, and post-operative recovery, each relying on resources such as ORs, post-anesthesia care unit (PACU), intensive care unit, and ward beds. An OR schedule that ignores downstream requirements may be perfectly efficient in theory but create bottlenecks in practice. Naderi et al. (2021) show that OR schedules built without downstream capacity constraints are feasible for the full perioperative pathway in only 24% of instances, even under *deterministic* conditions. In reality, stage durations are inherently uncertain (Xu et al., 2026), and delays at one stage propagate and amplify downstream. We, therefore, define a *surgical episode* as the complete perioperative pathway from admission to discharge, so that these dependencies across stages are embedded in the scheduling problem itself. We use the term *episode* as a short form for surgical episode, and reserve *operation* for the stage-specific activity an episode undergoes at a single stage.

Despite extensive work on surgical suite optimization, integrated models that capture the full perioperative pathway under uncertainty within a network of collaborating hospitals remain limited. Existing research has developed along related but separate streams. Surgical planning is typically studied across three levels: strategic decisions regarding which procedure types a hospital performs and in what volumes; tactical decisions regarding how OR time is allocated across surgical services or individual surgeons over a planning horizon; and operational decisions regarding the assignment of patients to ORs and their sequencing within a given OR block. Deterministic integrations of these decisions have been solved using branch-and-check, logic-based Benders decomposition (LBBD), branch-and-price, and related exact methods (Roshanaei et al., 2020a; Maenhout et al., 2026). At the tactical level, OR scheduling is done through either block or open scheduling systems. Block systems assign dedicated OR time to surgeons or services in advance, which provides predictability at the cost of pooling efficiency, whereas open systems allow more flexible booking but complicate assignment and sequencing decisions (Denton et al., 2010; Day et al., 2012; Batun et al., 2010; Roshanaei and Naderi, 2021).

Later studies have extended these ideas to broader network settings, including single-entry referral systems, collaborative multi-hospital OR scheduling, and multi-hospital block-scheduling models (Santibáñez et al., 2007; Urbach et al., 2025; Roshanaei et al., 2017b,a, 2020b; Wang et al., 2016; Guo et al., 2021; Rana et al., 2025). Single-entry systems pool patients into a common queue and match them with the next available qualified surgeon. This improves access and equity by replacing decentralized referrals with coordinated intake (Urbach et al., 2025). In practice, moving from siloed, per-surgeon waitlists to a common queue has been reported to reduce average surgical wait times by 20-30% while narrowing the variation in waits across providers, and single-entry intake has already been adopted successfully for cardiac surgery, joint replacement, transplantation, and cancer surgery in Canada (CIHI, 2023; Duraikannan, 2022). Population-level referral data further show that referring physicians typically direct patients to only one or a few surgeons regardless of patient-specific clinical needs, undermining the common objection that single-entry models sacrifice clinically meaningful surgeon choice (Seyedi et al., 2024). Multi-hospital OR scheduling similarly shows that pooling patients, surgeons, and OR capacity across hospitals can improve utilization and reduce costs (Roshanaei et al., 2017b,a, 2020b). However, existing stochastic and simulation-based extensions either evaluate deterministic schedules under uncertainty after it is realized, or model stochastic OR-capacity risk without jointly optimizing the full multi-stage perioperative pathway across hospitals (Wang et al., 2016; Guo et al., 2021).

Incorporating uncertainty makes the problem substantially harder to solve. Batun et al. (2010) show that realistic stochastic OR scheduling instances are difficult to solve without exploiting problem structure and valid inequalities. Rath et al. (2017) and Tsang et al. (2024) develop large-scale stochastic and distributionally robust models for OR-anesthesiologist coordination. Hashemi Doulabi et al. (2020) introduce a stochastic synchronization model with an OR application solvable up to 20 surgeries. Recent reviews show that uncertainty, synchronization, and multi-resource interactions quickly create computational bottlenecks, with solution performance depending heavily on the planning horizon, source of uncertainty, and modeling approach (Xu et al., 2026). To the best of our knowledge, no existing work jointly optimizes the full multi-stage surgical episode across a network of hospitals under stage duration uncertainty while incorporating risk-averse objectives, such as conditional value at risk (CVaR), to control system-wide tail outcomes.

Addressing this gap requires both the right model and the right solution method. Both Mixed-integer Programming (MIP) and Constraint Programming (CP) are widely used in scheduling, with comparative evidence showing that their relative performance depends strongly on the structure of assignment and sequencing decisions (Naderi et al., 2023). In healthcare scheduling, exact approaches have increasingly relied on decomposition frameworks. The LBBD method, introduced by Hooker and Ottosson (2003) and further developed for planning and scheduling (Hooker, 2007), has become a prominent exact paradigm in healthcare and surgical scheduling, with applications spanning outpatient scheduling, collaborative OR scheduling, integrated OR planning and scheduling, balanced distributed OR scheduling, generalized OR planning and scheduling, and stochastic distributed OR scheduling (Riise et al., 2016; Roshanaei et al., 2017b; Roshanaei and Naderi, 2021; Roshanaei et al., 2020b; Naderi et al., 2021; Guo et al., 2021). Yet the structure of our problem, where risk-averse objectives, multi-stage perioperative coordination, and multi-hospital resource pooling interact, differs fundamentally from any existing LBBD application, and solving it at scale requires new methodological developments.

Building on this body of work, this paper makes the following contributions.

- We introduce *Risk-Aware Surgical Episode Scheduling in Single-Entry Networks (RA-SEEN)*, a network-wide planning problem in which a single-entry intake assigns each elective same-day patient uniquely to one of the collaborating hospitals and schedules their full perioperative pathway, from admission through surgery to discharge, under stage-duration uncertainty. The framework captures interactions among upstream, intraoperative, and downstream resources that OR-centric models ignore, and supports any coherent risk measure on the network-wide makespan, including the expectation, CVaR, and the essential supremum.
- We develop both MIP and CP formulations of the problem that share the same here-and-now hospital assignment and non-anticipative resource assignment and sequencing decisions, allowing the same problem structure to be solved with either technology and compared on equal footing.
- We design a two-level LBBD that separates hospital assignment from intra-hospital scheduling and strengthen it with various enhancements, including a problem-specific optimality cut that dominates the standard no-good cut, and a valid lower bound on the full network scheduling subproblem at a fraction of its cost, which produces a class of valid inequalities.
- For instances on which the network subproblem itself becomes intractable, we develop a three-level scheme that further decomposes the scheduling subproblem, and design a risk-support

variant of the L-shaped method for solving it. We further enhance this inner L-shaped method using a longest-path computation on a directed acyclic graph, which reduces each scenario evaluation to a linear-time dynamic program.

- We evaluate the framework on practice-inspired instances calibrated to the University Health Network (UHN) and report computational and managerial findings. The results show that scalability is driven primarily by decomposition depth, auxiliary bounding, and structural reformulation rather than solver choice alone. The two-level decomposition is fastest when the network subproblem is tractable, while the three-level method is the only consistently viable approach on our largest instances. We further find that deterministic planning can substantially understate makespans on adverse days, and that disruptions propagate most strongly through the surgery-recovery segment, with PACU overtaking the OR as the binding resource under conservative risk criteria. We demonstrate that stochastic scheduling delivers substantial value, which grows with system scale and risk aversion. The ability of our framework to solve instances at the scale and structure of UHN, Canada's largest network of collaborating hospitals, demonstrates its potential for supporting complex integrated healthcare systems and other multi-hospital surgical networks.

The remainder of the paper is organized as follows. Section 2 formalizes the problem, and Section 3 presents the MIP and CP formulations. Section 4 develops the two- and three-level LBBD methods. Section 5 reports the computational study and the managerial insights, and Section 6 concludes the paper. Supporting formulations, proofs, illustrative examples, and additional numerical results are collected in the e-companion.

2. Problem Statement

Single-entry, multi-hospital surgical episode scheduling centralizes access and coordinates capacity across hospitals and care stages. In this setting, patient assignment, resource selection, and sequencing decisions determine how duration uncertainty propagates through the prescribed care pathway. We use the network-wide makespan as the performance metric, defined as the completion time of the last surgical episode anywhere in the network. Minimizing the makespan penalizes congestion and delays throughout the care pathway, and encourages balanced workload allocation across hospitals. This motivates a risk-aware model that works with the *distribution* of the makespan rather than only its *nominal* value.

2.1. Operational Setting and Clinical Pathway

Our setting is a network of collaborating hospitals operating under a single-entry system. We motivate and calibrate the model using UHN in the City of Toronto, a publicly funded system comprising three sites: Toronto General Hospital, Toronto Western Hospital, and Princess Margaret Cancer Centre, which share standardized patient-education materials for surgical preparation (UHN, 2026b). Each hospital $h \in \mathcal{H}$ organizes care as an ordered sequence of clinical stages $\mathcal{S} = \{1, \ldots, |\mathcal{S}|\}$. At stage s within hospital h, care is delivered by a pool of identical parallel resources $\mathcal{R}_{hs}$, such as ORs or recovery beds, reserved for the same-day pathway modeled here. Other patient categories, such as emergency, inpatient, and ambulatory cases outside this pathway, are assumed to use separate or pre-allocated capacity. A finite set of surgical episodes $\mathcal{P}$ must be scheduled across this network. Each episode $p \in \mathcal{P}$ is assigned to exactly one hospital and then moves through that hospital's prescribed sequence of care stages, occupying one resource at each stage. We refer to this problem as *Risk-Aware Surgical Episode Scheduling in Single-Entry Networks* (RA-SEEN).

The clinical pathway we draw on is the UHN same-day perioperative pathway (UHN, 2024b), which consists of five sequential stages on the day of surgery: (*i*) arrival and registration (REG), (*ii*) admission to the preoperative holding unit (POCU), where baseline clinical assessments are performed, (*iii*) surgery in the OR, (*iv*) immediate postoperative monitoring and anesthesia recovery in the PACU, and (*v*) phase-II recovery and discharge in the day surgery unit (DSU). This pathway covers a broad range of elective procedures, including joint replacement, anterior cruciate ligament repair, thyroid surgery, breast reconstruction, cataract surgery, and deep brain stimulation internal pulse generator battery replacement (UHN, 2026b, 2024a, 2026a), and is representative of same-day surgical care across a wide variety of hospital networks. Pre-admission activities, such as phone or in-person assessments and preoperative testing, take place before the day of surgery and fall outside the scope of the operational model. Figure 1 illustrates the overall structure of the RA-SEEN problem, showing how surgical episodes are routed through the single-entry system and processed across the multi-stage perioperative pathway at each hospital in the network.

2.2. Duration Uncertainty and Risk Measures

The primary source of uncertainty in this setting is that stage durations cannot be known exactly in advance. We model this through a probability space $(\Omega, \Sigma, \mathbb{P})$, where Ω is a finite set of scenarios, each occurring with probability $\mathfrak{p}(\omega) = \mathbb{P}(\{\omega\}) > 0$ satisfying $\sum_{\omega \in \Omega} \mathfrak{p}(\omega) = 1$. For each surgical episode $p \in \mathcal{P}$, hospital $h \in \mathcal{H}$, and stage $s \in \mathcal{S}$, the random stage duration is denoted by $D_{phs} : \Omega \to \mathbb{R}_+$, and $\mathbf{D}(\omega) := \big(D_{phs}(\omega)\big)_{(p,h,s)} \in \mathbb{R}_+^{|\mathcal{P}||\mathcal{H}||\mathcal{S}|}$ collects all realized durations in scenario ω. We write $\mathbb{E}[X] = \sum_{\omega \in \Omega} \mathfrak{p}(\omega) X(\omega)$ for the expectation of any random variable X.

Hospital assignment, resource selection, and sequencing must all be decided before any scenario is realized and are therefore non-anticipative. For a fixed schedule, the network-wide makespan $C_{\max} : \Omega \to \mathbb{R}_+$ is itself a random variable, with $C_{\max}(\omega)$ equal to the maximum completion time

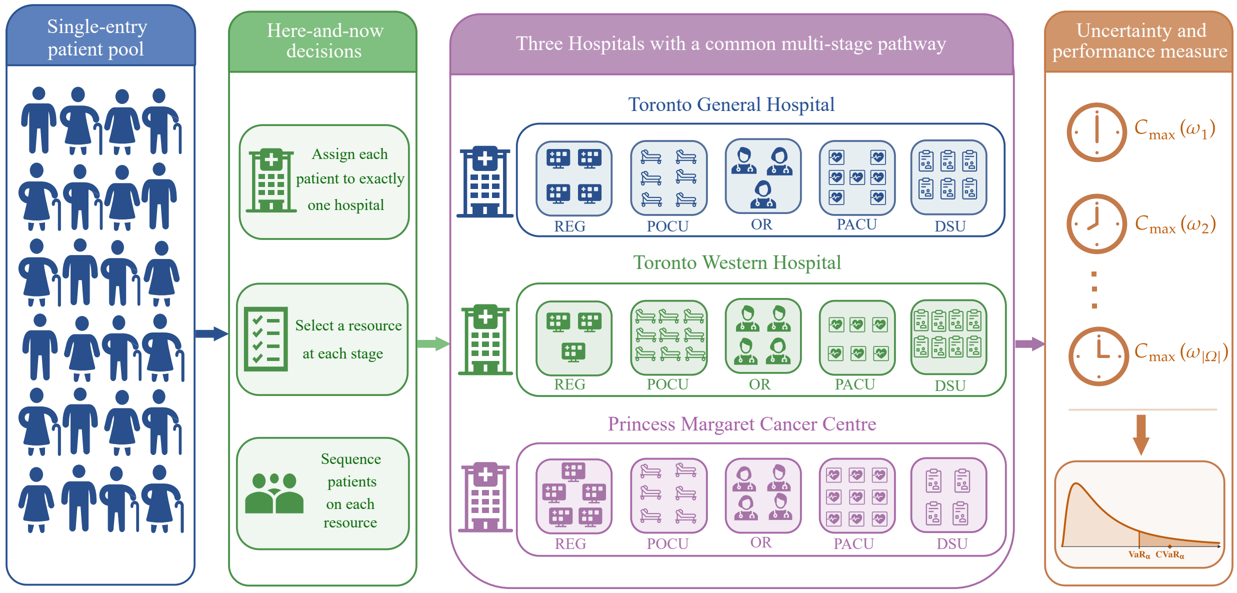

Figure 1 Illustration of the single-entry network surgical episode scheduling problem.

across all surgical episodes and hospitals in scenario ω. We assess schedule quality using a coherent risk measure $\rho\big((C_{\max}(\omega))_{\omega\in\Omega}\big)$, where larger values represent worse outcomes (Artzner et al., 1997; Shapiro et al., 2009). This class encompasses the expected value, the essential supremum, and the family of CVaR measures, which allows planners to express different attitudes toward adverse outcomes within a unified framework. In particular, the CVaR concentrates on the tail of the makespan distribution by penalizing realizations beyond the Value-at-Risk (VaR). For a loss random variable X, the CVaR at confidence level $\alpha \in (0,1)$ is defined as follows (Rockafellar and Uryasev, 2000)

$$\mathrm{CVaR}_\alpha(X) = \inf_{\eta\in\mathbb{R}} \left\{ \eta + \frac{1}{1-\alpha}\mathbb{E}\big([X-\eta]^+\big) \right\}, \qquad [a]^+ := \max\{a, 0\}. \tag{1}$$

Under the finite-scenario assumption, the CVaR admits a linear representation by introducing an auxiliary threshold η and excess variables $\tau(\omega) = \max\{0, X(\omega)-\eta\}$:

$$\mathrm{CVaR}_\alpha(X) = \min_{\eta\in\mathbb{R},\ \tau(\omega)\geq 0} \left\{ \eta + \frac{1}{1-\alpha}\sum_{\omega\in\Omega}\mathfrak{p}(\omega)\,\tau(\omega):\ \tau(\omega) \geq X(\omega)-\eta \ \ \forall \omega\in\Omega \right\}. \tag{2}$$

Since Ω is finite, the feasible region of (2) is a nonempty closed polyhedron and the objective is bounded below, which together guarantee that a minimizer exists and the infimum in (1) is attained.

2.3. Operational and Clinical Considerations

We adopt the following clinical and operational considerations. Patients, once assigned to a site, complete all stages there. Within each hospital-stage pair (h, s), clinical resources are homogeneous.

Once an operation begins on a given resource, it runs to completion without interruption, reflecting the clinical continuity required within a care stage. Each stage-specific resource serves at most one operation at a time. Inter-stage transfer times are absorbed into the stage durations $D_{phs}(\omega)$, and patients may wait between consecutive stages when downstream capacity is unavailable. We assume all surgical episodes follow an identical, prespecified sequence of care stages. Finally, all episodes are available for scheduling at time zero, meaning that the schedule determines each patient's planned admission time, communicated in advance, rather than requiring physical arrival at the start of the day. Resource homogeneity, a common care pathway across episodes, and simultaneous availability are simplifying assumptions adopted for clarity. They can be relaxed with straightforward model modifications without changing the main logic of our framework.

3. Mathematical Formulations

We now present two monolithic formulations for RA-SEEN. Both share the same here-and-now hospital assignment decisions and enforce non-anticipativity of resource selection and sequencing across scenarios, while differing in how they represent the underlying scheduling structure. The MIP formulation relies on binary assignment and precedence variables alongside scenario-dependent completion times, whereas the CP formulation reasons directly over operation intervals on the time line using optional presences and global sequencing constraints. In both cases, the scenario-wise makespan is evaluated through a finite-scenario representation of the chosen risk measure.

3.1. MIP Formulation

We introduce binary variables y_{ph} to denote assignment of surgical episode p to hospital h. Binary variables z_{phsr} indicate whether the operation of surgical episode p at stage s in hospital h uses resource r. Resource-level sequencing is captured through Manne-type precedence variables $x_{pp'hs}$ (Manne, 1960; Naderi et al., 2023): for $p > p'$, $x_{pp'hs} = 1$ if the operation of p is processed after that of p' at stage s in hospital h, with the precedence enforced only when both operations share the same resource. For each scenario ω, decision variable $c_{phs}(\omega)$ denotes the completion time of the operation of surgical episode p at (h, s), and $C_{\max}(\omega)$ the corresponding network makespan.

For a coherent risk measure $\rho(\cdot)$ admitting a finite-scenario epigraph representation, we seek a schedule that minimizes the risk-adjusted makespan $\rho(\mathbf{C}_{\max})$, where $\mathbf{C}_{\max} = (C_{\max}(\omega))_{\omega\in\Omega}$. The RA-SEEN problem can be formulated as below:

$$\min \quad \rho\big(\mathbf{C}_{\max}\big) \tag{3a}$$

$$\text{s.t.} \quad \sum_{h\in\mathcal{H}} y_{ph} = 1 \qquad \forall p \in \mathcal{P}, \tag{3b}$$

$$\sum_{r \in \mathcal{R}_{hs}} z_{phsr} = y_{ph} \qquad \forall p \in \mathcal{P}, h \in \mathcal{H}, s \in \mathcal{S}, \tag{3c}$$

$$c_{ph,1}(\omega) \geq D_{ph,1}(\omega)\, y_{ph} \qquad \forall p \in \mathcal{P}, h \in \mathcal{H}, \omega \in \Omega, \tag{3d}$$

$$c_{phs}(\omega) \geq c_{ph,s-1}(\omega) + D_{phs}(\omega)\, y_{ph} \qquad \forall p \in \mathcal{P}, h \in \mathcal{H}, s \in \mathcal{S} \setminus \{1\}, \omega \in \Omega, \tag{3e}$$

$$\begin{aligned} c_{phs}(\omega) \geq\ & c_{p'hs}(\omega) + D_{phs}(\omega) && \forall h \in \mathcal{H},\ s \in \mathcal{S},\ r \in \mathcal{R}_{hs}, \\ & - M_{pp'hs}(\omega)\big(3 - x_{pp'hs} - z_{phsr} - z_{p'hsr}\big) && p, p' \in \mathcal{P}:\ p > p',\ \omega \in \Omega, \end{aligned} \tag{3f}$$

$$\begin{aligned} c_{p'hs}(\omega) \geq\ & c_{phs}(\omega) + D_{p'hs}(\omega) && \forall h \in \mathcal{H},\ s \in \mathcal{S},\ r \in \mathcal{R}_{hs}, \\ & - M'_{pp'hs}(\omega)\big(2 + x_{pp'hs} - z_{phsr} - z_{p'hsr}\big), && p, p' \in \mathcal{P}:\ p > p',\ \omega \in \Omega, \end{aligned} \tag{3g}$$

$$C_{\max}(\omega) \geq c_{ph,|\mathcal{S}|}(\omega) \qquad \forall p \in \mathcal{P}, h \in \mathcal{H}, \omega \in \Omega, \tag{3h}$$

$$c_{phs}(\omega) \geq 0 \qquad \forall p \in \mathcal{P}, h \in \mathcal{H}, s \in \mathcal{S}, \omega \in \Omega, \tag{3i}$$

$$y_{ph} \in \{0,1\} \qquad \forall p \in \mathcal{P}, h \in \mathcal{H}, \tag{3j}$$

$$z_{phsr} \in \{0,1\} \qquad \forall p \in \mathcal{P}, h \in \mathcal{H}, s \in \mathcal{S}, r \in \mathcal{R}_{hs}, \tag{3k}$$

$$x_{pp'hs} \in \{0,1\} \qquad \forall p, p' \in \mathcal{P} : p > p', h \in \mathcal{H}, s \in \mathcal{S}, \tag{3l}$$

where $M_{pp'hs}(\omega)$ and $M'_{pp'hs}(\omega)$ are sufficiently large constants (conservative values are derived in Appendix B). Constraint (3b) assigns each surgical episode to exactly one hospital. Constraint (3c) then selects exactly one resource at each stage within that hospital, forcing all remaining resource indicators to zero. Constraints (3d)-(3e) propagate completion times along the care pathway within a hospital under each scenario ω: the first stage completes no earlier than its realized duration, and each subsequent stage completes no earlier than the sum of the previous stage's completion time and its own realized duration. Constraints (3f)-(3g) ensure that on each resource under every scenario, operations do not overlap. For any pair $p > p'$, stage s, and resource r, these two inequalities form a disjunction: if both operations are assigned to the same resource, then the precedence variable $x_{pp'hs}$ equal to one enforces that p' precedes p, whereas $x_{pp'hs} = 0$ activates the opposite ordering. If at least one of the two operations is not assigned to resource r at (h, s), the Big-M terms relax the corresponding constraint so that no ordering is imposed. Constraint (3h) defines the scenario makespan as the maximum final-stage (denoted $|S|$) completion time across all surgical episodes. Constraints (3i)-(3l) define the support of the decision variables. See Appendix A.1 for the deterministic version, where the model is stated with averaged durations instead of the scenario-dependent $D_{phs}(\omega)$.

In model (3), the assignment, sequencing, and timing constraints are the same across all risk measures. Substituting an appropriate finite-scenario epigraph for ρ yields a concrete optimization formulation; in particular, risk measures with linear epigraphs, such as expectation, CVaR, and the essential supremum, yield mixed-integer linear programs. However, the size of this model grows rapidly with the number of scenarios, surgical episodes, and resources, particularly since the precedence variables scale quadratically in episodes per stage. Furthermore, the Big-M disjunctions in

Table 1 CP primitives used in the formulation.

CP Functions	
`IntervalVar`$(\cdot)$	Creates an interval decision variable (non-optional / optional).
`PresenceOf`(I)	Returns 1 if optional interval I is present in the solution, else 0.
`Alternative`$(O,\mathcal{I})$	Selects exactly one interval in $\mathcal{I}$ to realize O and synchronizes their start and end times.
`EndBeforeStart`(A,B)	Enforces precedence: interval A ends before B starts.
`SequenceVar`$(\mathcal{I})$	Creates a sequence variable ordering intervals in $\mathcal{I}=\{I_i\}_i$.
`NoOverlap`(Q)	Enforces no overlap among intervals in sequence Q.
`SameSequence`$(Q(\omega),Q(\bar{\omega}))$	Enforces the same relative order as the reference-scenario sequence.
`EndOf`(O)	Returns end time of interval O.

(3f)-(3g) typically produce weak linear programming (LP) relaxations. These limitations motivate the CP formulation developed next.

3.2. CP Formulation

Rather than tracking completion times through pairwise Big-M disjunctions (3f)-(3g) and binary precedence variables, the CP formulation reasons directly over operation intervals on the time line and delegates resource capacity to global sequencing constraints. Throughout, we adopt CP Optimizer notation (Laborie et al., 2018), with the key primitives summarized in Table 1. At a high level, the hospital-assignment variables $\mathbf{y}$ remain here-and-now decisions, while resource selections and processing orders are represented through optional interval presences and sequence variables defined for each scenario. Non-anticipativity is enforced by synchronizing these decision variables across scenarios via a designated reference scenario $\bar{\omega}$.

For each surgical episode p, stage s, and scenario ω, the non-optional interval $O_{ps}(\omega)$ represents the operation of surgical episode p at stage s. This interval represents the episode's stage-s requirement within the overall pathway but does not itself select a hospital or resource. That selection is made by optional intervals $T_{phsr}(\omega)$, each with duration $D_{phs}(\omega)$. An `Alternative` constraint selects exactly one of these variables to realize $O_{ps}(\omega)$, and aligns their start and end times. The presence of $T_{phsr}(\omega)$ then determines the hospital and resource assigned to that stage. For each hospital-stage-resource triple (h,s,r), the sequence variable $Q_{hsr}(\omega)$ orders the optional intervals $T_{phsr}(\omega)$ on resource r in scenario ω. The CP formulation for RA-SEEN is:

$$\min \quad \rho(\mathbf{C}_{\max}) \tag{4a}$$

$$\text{s.t.} \quad (3b),\ (3j), \tag{4b}$$

$$O_{ps}(\omega) = \texttt{IntervalVar}(\text{non-optional}) \qquad \forall p\in\mathcal{P},\ s\in\mathcal{S},\ \omega\in\Omega, \tag{4c}$$

$$T_{phsr}(\omega) = \texttt{IntervalVar}(\text{optional},\ \text{size}=D_{phs}(\omega)) \quad \forall p\in\mathcal{P},\ h\in\mathcal{H},\ s\in\mathcal{S},\ r\in\mathcal{R}_{hs},\ \omega\in\Omega, \tag{4d}$$

$$\texttt{Alternative}\Big(O_{ps}(\omega),\ \{T_{phsr}(\omega)\}_{h\in\mathcal{H},r\in\mathcal{R}_{hs}}\Big) \qquad \forall p\in\mathcal{P},\ s\in\mathcal{S},\ \omega\in\Omega, \tag{4e}$$

$$\texttt{EndBeforeStart}\big(O_{ps}(\omega),\, O_{p,s+1}(\omega)\big) \qquad \forall p\in\mathcal{P},\ s=1,\ldots,|\mathcal{S}|-1,\ \omega\in\Omega, \tag{4f}$$

$$\sum_{r\in\mathcal{R}_{hs}} \texttt{PresenceOf}\big(T_{phsr}(\bar{\omega})\big) = y_{ph} \qquad \forall p\in\mathcal{P},\ h\in\mathcal{H},\ s\in\mathcal{S}, \tag{4g}$$

$$Q_{hsr}(\omega) = \texttt{SequenceVar}\big(\{T_{phsr}(\omega)\}_{p\in\mathcal{P}}\big) \qquad \forall h\in\mathcal{H},\ s\in\mathcal{S},\ r\in\mathcal{R}_{hs},\ \omega\in\Omega, \tag{4h}$$

$$\texttt{NoOverlap}\big(Q_{hsr}(\omega)\big) \qquad \forall h\in\mathcal{H},\ s\in\mathcal{S},\ r\in\mathcal{R}_{hs},\ \omega\in\Omega, \tag{4i}$$

$$\texttt{PresenceOf}\big(T_{phsr}(\omega)\big) = \texttt{PresenceOf}\big(T_{phsr}(\bar{\omega})\big) \qquad \forall p\in\mathcal{P},\ h\in\mathcal{H},\ s\in\mathcal{S},\ r\in\mathcal{R}_{hs},\ \omega\in\Omega\setminus\{\bar{\omega}\}, \tag{4j}$$

$$\texttt{SameSequence}\big(Q_{hsr}(\omega),\, Q_{hsr}(\bar{\omega})\big) \qquad \forall h\in\mathcal{H},\ s\in\mathcal{S},\ r\in\mathcal{R}_{hs},\ \omega\in\Omega\setminus\{\bar{\omega}\}, \tag{4k}$$

$$C_{\max}(\omega) \geq \max_{p\in\mathcal{P}}\big\{\texttt{EndOf}\big(O_{p,|\mathcal{S}|}(\omega)\big)\big\} \qquad \forall\omega\in\Omega. \tag{4l}$$

Constraints (4c), (4d), and (4h) define the stage-level intervals, optional resource intervals, and resource sequences. Constraint (4e) assigns each stage operation to exactly one selected resource interval and aligns their start and end times. The care pathway structure is enforced by (4f), which requires each stage to finish before the next begins. This is the CP counterpart of the MIP flow constraints (3d)-(3e). Resource capacity is then handled by constraint (4i) using the `NoOverlap` function, so each resource processes at most one operation at a time. Therefore, the CP formulation replaces all the pairwise Big-M disjunctive constraints (3f)-(3g) of the MIP with a single global constraint per resource and scenario. We establish non-anticipativity through the reference scenario $\bar{\omega}$. First, constraint (4g) links resource presences in $\bar{\omega}$ to the here-and-now hospital assignments by enforcing that if $y_{ph}=1$, then exactly one resource interval appears for (p,h,s), and none otherwise. Constraints (4j)-(4k) then propagate both this resource selection and its corresponding ordering to all other scenarios, so neither the chosen resource nor the processing sequence can vary across scenarios. Finally, (4l) defines the scenario makespan, and the objective (4a) evaluates the resulting random makespan under the selected risk measure.

Two implementation notes are worth highlighting. First, the per-scenario resource sequences in (4) are needed because resource choices and processing orders must remain non-anticipative. In deterministic settings, this is unnecessary, and the constraints can be replaced by a more compact `Pulse` representation, in which each operation contributes one unit of cumulative resource usage over its processing interval and the total usage is capped by the number of parallel resources at the corresponding stage (Naderi et al., 2023). We present this deterministic variant in Appendix A.2. Second, if all stage durations are integers, with $\rho=\mathrm{CVaR}_\alpha$, there exists an optimal schedule with integer scenario makespans and an integer optimal VaR threshold $\eta\in\mathbb{Z}_+$ (Rockafellar and Uryasev, 2002, Prop. 8). So, η can be declared as an integer CP variable without loss of optimality.

Even though the CP formulation avoids the expensive disjunctive constraints of (3), its search space still grows combinatorially with the number of episodes, hospitals, stages, and scenarios, which makes direct solves impractical for realistic networks. We therefore turn to exact decomposition methods that exploit the layered decision structure of the problem: hospital assignment, resource selection and sequencing, and scenario-wise timing evaluation. Rather than solving each layer independently, we group them into subproblems, with the grouping itself varying by decomposition scheme.

4. Solution Approaches

We develop exact decomposition algorithms for the proposed risk-aware scheduling models. We first present a two-level framework, in which the master problem (MP) assigns surgical episodes to hospitals and the subproblem (SP) jointly optimizes the resulting intra-hospital schedules. We then introduce a three-level scheme that further separates resource assignment and sequencing from scenario-wise timing evaluation.

4.1. Two-level Decomposition

In model (3), fixing $\mathbf{y}$ separates the resource-selection and sequencing constraints by hospital, so that each hospital's intra-hospital timing problem can be formulated independently. The network-wide risk measure, however, still couples these per-hospital solutions through the scenario-wise maximum completion time across hospitals. Each hospital's problem remains a stochastic scheduling SP, in which resource selection and sequencing are binary. This rules out classical Benders decomposition (Benders, 1962), which places the complicating (integer) variables in a MP and the continuous variables in a SP, alternatively solves them, and uses LP duality on the SP to generate cuts for the MP. Its stochastic counterpart, the L-shaped method, applies the same idea scenario by scenario, but still relies on a continuous SP. One could in principle restore that structure by keeping all binary resource and sequencing variables in the master and leaving only the continuous timing variables in the SP. We do not pursue this route for two reasons. First, such a lifting would add $|\mathcal{P}| \sum_{h\in\mathcal{H}} \sum_{s\in\mathcal{S}} |\mathcal{R}_{hs}| + \binom{|\mathcal{P}|}{2}|\mathcal{H}||\mathcal{S}|$ binary resource-assignment and sequencing variables to the MP. Moreover, pairwise precedence variables alone do not prevent cyclic orderings once decoupled from the timing subproblem. Enforcing a consistent sequence per resource requires adding ordering variables, which further increases the size of the MP formulation. Without further strengthening, this enlarged master would contain limited information about scenario-dependent timing, because the constraints linking these binary decisions to completion times would remain in the timing subproblems. The resulting formulation would therefore rely heavily on accumulated L-shaped cuts, whose effectiveness depends strongly on the formulation and the strength of the generated cuts (Rahmaniani et al., 2017).

Second, keeping resource selection and sequencing inside the SP preserves a self-contained scheduling problem that can be addressed with technologies tailored to that structure, including CP, MIP-CP hybrids, and other acceleration techniques (see Section 4.3). Moving these variables to the MP would eliminate this scheduling subproblem. We therefore adopt LBBD, which preserves the natural separation between hospital assignment and intra-hospital scheduling and allows each level to be addressed with methods tailored to its structure. LBBD extends the classical Benders paradigm to combinatorial SPs by replacing LP duality with logic-based inference (Hooker and Ottosson, 2003; Hooker, 2007).

At each iteration κ of the LBBD method, the MP produces a candidate hospital assignment $\mathbf{y}^{\kappa}$. The corresponding SP is then solved, and the inference returned is expressed as a linear inequality in the MP variables, which is added to the MP for the next iteration. Such an inequality is a *feasibility cut* if it eliminates assignments under which the SP is infeasible, and an *optimality cut* if it provides a valid lower bound (LB) on the second-level objective at $\mathbf{y}^{\kappa}$. The MP thus progressively builds a lower approximation of the recourse function, with exactness ensured by the accumulation of optimality cuts. In our setting, the considerations and assumptions of Section 2.3 guarantee that every feasible hospital assignment admits at least one feasible SP schedule: processing the assigned episodes one at a time on a single eligible resource at each stage of their assigned hospital satisfies all capacity and precedence constraints and yields finite completion times in every scenario. The two-level decomposition, therefore, has relatively complete recourse, and the outer LBBD only needs optimality cuts.

A MP based on optimality cuts alone can nonetheless be very weak in early iterations, when little information about the SP has been revealed. We strengthen it by adding analytical LBs that hold for any hospital assignment. Let θ denote the MP approximation of the optimal risk-adjusted makespan, and let $\mathbf{C}^{\mathrm{MP}}_{\max} = (C^{\mathrm{MP}}_{\max}(\omega))_{\omega\in\Omega}$ collect scenario-wise LBs. The MP at iteration κ is then

$$
\begin{aligned}
\min\quad & \theta && && \text{(5a)}\\
\text{s.t.}\quad & \text{(3b), (3j)}, && && \text{(5b)}\\
& \theta \geq 0, && && \text{(5c)}\\
& \theta \geq \mathcal{H}^{\iota}(\mathbf{y}) && \forall \iota = 1,\ldots,\kappa-1, && \text{(5d)}\\
& C^{\mathrm{MP}}_{\max}(\omega) \geq \sum_{h\in\mathcal{H}}\sum_{s\in\mathcal{S}} D_{phs}(\omega)\, y_{ph} && \forall p\in\mathcal{P},\ \omega\in\Omega, && \text{(5e)}\\
& C^{\mathrm{MP}}_{\max}(\omega) \geq \frac{1}{|\mathcal{R}_{hs}|}\sum_{p\in\mathcal{P}} D_{phs}(\omega)\, y_{ph} && \forall h\in\mathcal{H},\ s\in\mathcal{S},\ \omega\in\Omega, && \text{(5f)}\\
& \theta \geq \rho\big(\mathbf{C}^{\mathrm{MP}}_{\max}\big), && && \text{(5g)}
\end{aligned}
$$

where $\mathcal{H}^{\iota}(\mathbf{y})$ denotes a valid LB generated from the second-level problem at an earlier outer iteration $\iota \in 1,\ldots,\kappa-1$. Inequalities (5e)-(5f) are structural scenario LBs from parallel-machine scheduling (Pinedo, 2012). Constraint (5e) is a surgical episode-based bound, since the makespan dominates the total processing time of any episode along its hospital's care pathway. Constraint (5f) is a workload-capacity bound, because even perfectly balancing the workload across the $|\mathcal{R}_{hs}|$ parallel resources at (h,s), the makespan is at least the average per-resource workload. Finally, constraint (5g) aggregates the scenario-wise LBs through the chosen risk measure. Since each $C^{\mathrm{MP}}_{\max}(\omega)$ underestimates the true makespan in scenario ω, monotonicity of $\rho(\cdot)$ implies that $\rho\big(\mathbf{C}^{\mathrm{MP}}_{\max}\big)$ is itself a valid LB on the true risk-adjusted second-level objective.

Once an incumbent solution $\mathbf{y}^{\kappa}$ is obtained, each surgical episode p has a unique assigned hospital $h^{\kappa}(p)$ derived from $y^{\kappa}_{p,h^{\kappa}(p)}=1$. Fixing $\mathbf{y}^{\kappa}$ reduces the scheduling problem because hospital-indexed variables can be restricted to the assigned hospital, and the SP can be written with variables z_{psr}, $x_{pp's}$, and $c_{ps}(\omega)$, defined relative to $h^{\kappa}(p)$. Resource selection and sequencing range only over $\mathcal{R}_{h^{\kappa}(p)s}$ and remain scenario-independent, so non-anticipativity is preserved. Let $\mathrm{SP}^{\star}(\mathbf{y})$ denote the SP optimal objective value for a given $\mathbf{y}$, and let $\mathbf{C}^{\mathrm{SP}}_{\max}=(C^{\mathrm{SP}}_{\max}(\omega))_{\omega\in\Omega}$. Concretely, at the incumbent assignment we solve

$$\mathrm{SP}^{\star}(\mathbf{y}^{\kappa}) := \min_{\mathbf{z},\mathbf{x}} \rho\big(\mathbf{C}^{\mathrm{SP}}_{\max}\big) \quad \text{s.t. network scheduling feasibility under } \mathbf{y}^{\kappa}, \tag{6}$$

where feasibility comprises resource selection at every stage, pathway precedence, non-overlap of the operations sharing a resource, and the scenario-wise definition of $C^{\mathrm{SP}}_{\max}(\omega)$. Similar to Section 3, we can formulate both MIP and CP versions of the SP, which can be used interchangeably within the outer LBBD loop. The MIP version restricts (3) to the assigned hospitals, and the CP version restricts (4). Both formulations are stated explicitly in Appendices A.3 and A.4.

Solving SP (6) to optimality returns the value $\mathrm{SP}^{\star}(\mathbf{y}^{\kappa})$, which is used to provide a feedback to the MP. The simplest LBBD optimality cut is a no-good cut that enforces this value whenever the same hospital assignment is repeated:

$$\theta \geq \mathrm{SP}^{\star}(\mathbf{y}^{\kappa})\Big(1-\sum_{p\in\mathcal{P}}\big(1-y_{p,h^{\kappa}(p)}\big)\Big). \tag{7}$$

If the MP reproduces the same assignment, the term in parentheses equals one and the cut enforces $\theta \geq \mathrm{SP}^{\star}(\mathbf{y}^{\kappa})$ which prevents the MP from underestimating the known recourse value of $\mathbf{y}^{\kappa}$. The cut is inactive for any reassignment, since the right-hand side becomes non-positive.

Although valid, the cut (7) is quite weak, as it gives no bounding information for structurally similar assignments. To obtain a stronger cut and strengthen the MP relaxation across neighboring assignments, we adapt a deletion-penalty logic (Roshanaei et al., 2017a; Guo et al., 2021) to the multi-stage surgical setting. The idea is to charge a per-patient penalty whenever an episode is reassigned away from its incumbent hospital, where the penalty upper-bounds that patient's contribution to $\mathrm{SP}^{\star}(\mathbf{y}^{\kappa})$. Since reassigning p away from $h^{\kappa}(p)$ deletes p from that hospital's workload, its contribution cannot exceed the time it spends traversing the hospital's stages. We therefore define the *full pathway duration* of p at its incumbent hospital under scenario ω as

$$\mathrm{path}_p^{\kappa}(\omega) := \sum_{s \in \mathcal{S}} D_{p,h^{\kappa}(p),s}(\omega).$$

Since $\mathrm{SP}^{\star}(\mathbf{y}^{\kappa})$ aggregates scenario outcomes through ρ, we aggregate the scenario-wise pathway durations through the same functional to obtain the *risk-adjusted deletion penalty* $\gamma_p^{\kappa} := \rho\big((\mathrm{path}_p^{\kappa}(\omega))_{\omega \in \Omega}\big)$. The strengthened cut charges penalty γ_p^{κ} only when p is deleted from $h^{\kappa}(p)$, i.e., when the deletion indicator $(1 - y_{p,h^{\kappa}(p)})$ equals one:

$$\theta \geq \mathrm{SP}^{\star}(\mathbf{y}^{\kappa}) - \sum_{p \in \mathcal{P}} \gamma_p^{\kappa}\big(1 - y_{p,h^{\kappa}(p)}\big). \tag{8}$$

At the incumbent assignment, where all deletion indicators are zero, the cut recovers $\mathrm{SP}^{\star}(\mathbf{y}^{\kappa})$ exactly. When some surgical episodes are assigned to other hospitals, the cut relaxes $\mathrm{SP}^{\star}(\mathbf{y}^{\kappa})$ by at most the total risk-adjusted pathway duration of the reassigned episodes. Therefore, unlike the no-good cut (7), the cut (8) remains informative for partial reassignments.

Theorem 1. *Let ρ be a coherent risk measure, and let $\mathrm{SP}^{\star}(\mathbf{y}^{\kappa})$ be the optimal SP value at the master assignment $\mathbf{y}^{\kappa}$. Then:*

(i) Inequality (8) is valid for the integer epigraph of the LBBD MP (5).

(ii) Its right-hand side dominates that of the no-good cut (7) pointwise over the master assignment region.

(iii) If every SP is solved to optimality, the resulting outer LBBD algorithm terminates finitely with a globally optimal solution of RA-SEEN.

When (5) admits multiple optimal assignments $\mathbf{y}^{\kappa}$, we break ties in favor of balanced allocations of episodes across hospitals. Concentrating episodes onto few hospitals produces a larger, more heavily loaded scheduling SP for those hospitals, whereas a more even allocation keeps each hospital's SP smaller and easier to solve. Scenario-wise completion times also better reflect realistic hospital workloads. This choice, in turn, affects the strength of the information fed back to the master, since cut (8) depends on the specific assignment $\mathbf{y}^{\kappa}$ returned by the MP. Let θ^{κ} denote the optimal objective value of (5) at outer iteration κ. We can therefore re-solve (5) with its objective fixed at θ^{κ} and minimize the total absolute deviation between the number of surgical episodes assigned to each hospital and the network average $|\mathcal{P}| \,/\, |\mathcal{H}|$ (see Appendix A.5 for details). Because the primary MP objective is fixed, this step does not change the master bound passed to the LBBD loop, and only determines which optimal assignment $\mathbf{y}^{\kappa}$ is forwarded to the SP.

4.2. Per-Hospital Decomposition of the SP

The second-level scheduling SP is the main computational bottleneck of the two-level scheme. Since resource selection and sequencing are local to each hospital once $\mathbf{y}^\kappa$ is fixed, a natural idea to obtain cheaper feedback for the MP is to solve one smaller scheduling problem per hospital and combine the results. In the following, we will show that, although the resulting per-hospital decomposition is not exact in general, it yields a valid LB that can be used both to bound the SP value and to generate cuts for the MP.

Fix $\mathbf{y}^\kappa$ and let $\mathcal{P}_h(\mathbf{y}) := \{p \in \mathcal{P} : y_{ph} = 1\}$ be the episodes assigned to hospital h. For any feasible scheduling decision $(\mathbf{z}, \mathbf{x})$ and scenario ω, let $C_{\max,h}(\omega)$ denote the makespan of hospital h, and write $\mathbf{C}_{\max,h} = (C_{\max,h}(\omega))_{\omega\in\Omega}$. Since $C_{\max}(\omega) = \max_{h\in\mathcal{H}} C_{\max,h}(\omega)$ in every scenario, the original SP minimizes $\rho(\mathbf{C}_{\max})$ over scheduling decisions that are jointly feasible across all hospitals. A per-hospital decomposition AUX-SP would instead solve one local problem per hospital and aggregate the resulting risk-adjusted values through the maximum. However, the two operations do not commute,

$$\rho\Big(\big(\max_{h\in\mathcal{H}} C_{\max,h}(\omega)\big)_{\omega\in\Omega}\Big) \neq \max_{h\in\mathcal{H}} \rho(\mathbf{C}_{\max,h}),$$

because hospitals remain coupled through the scenario-wise maximum, whereas local optimization treats them independently. Appendix D.1 gives an example showing the gap between the per-hospital decomposition and the original SP. Monotonicity of ρ is nevertheless enough to recover a one-sided inequality.

Proposition 1. *Let ρ be a monotone risk measure and fix a feasible MP assignment $\mathbf{y}$. Then*

$$\mathrm{SP}^\star(\mathbf{y}) = \min_{\mathbf{z},\mathbf{x}} \rho\big(\mathbf{C}_{\max}\big) \geq \max_{h\in\mathcal{H}} \min_{\mathbf{z}_h,\mathbf{x}_h} \rho\big(\mathbf{C}_{\max,h}\big). \tag{9}$$

The minimization on the left of inequality (9) is over all scheduling decisions feasible under $\mathbf{y}$, and the inner minimization on the right is over local scheduling decisions for the surgical episodes assigned to h under $\mathbf{y}$. Concretely, for each $h \in \mathcal{H}$ we solve

$$\mathrm{AUX}_h^\star(\mathbf{y}^\kappa) := \min_{\mathbf{z}_h,\mathbf{x}_h} \rho(\mathbf{C}_{\max,h}) \quad \text{s.t. local scheduling feasibility at hospital } h \text{ for } \mathcal{P}_h(\mathbf{y}^\kappa),$$

and aggregate the hospital values into

$$\mathrm{AUX}^\star(\mathbf{y}^\kappa) := \max_{h\in\mathcal{H}} \mathrm{AUX}_h^\star(\mathbf{y}^\kappa).$$

By Proposition 1, $\mathrm{AUX}^\star(\mathbf{y}^\kappa)$ is a valid LB on the SP value at $\mathbf{y}^\kappa$. The per-hospital problems are independent and can be solved in parallel using either the MIP or the CP SP (6) formulation, restricted to a single hospital.

The per-hospital bound (9) is not necessarily strict. A sufficient condition for it to be exact for every fixed hospital assignment is that ρ commute with finite maxima:

$$\rho\left(\big(\max_{h\in\mathcal{H}} C_{\max,h}(\omega)\big)_{\omega\in\Omega}\right) = \max_{h\in\mathcal{H}} \rho(\mathbf{C}_{\max,h}).$$

The essential supremum has this property: since Ω is finite, we have ess-sup $X = \max_{\omega\in\Omega} X(\omega)$, and both sides reduce to $\max_{h\in\mathcal{H}} \max_{\omega\in\Omega} C_{\max,h}(\omega)$. Moreover, once $\mathbf{y}$ is fixed, the feasible region factorizes across hospitals: each $(\mathbf{z}_h, \mathbf{x}_h)$ is constrained only by $\mathcal{P}_h$, and $C_{\max,h}$ depends only on that block. Minimizing a maximum of terms defined on disjoint blocks over a product set can therefore be carried out blockwise. Thus, AUX-SP is an exact reformulation of SP under $\rho =$ ess-sup. Expectation and CVaR do not commute with finite maxima in general, so the same universal exactness result does not hold for these risk measures. Nevertheless, (9) may still be tight for a particular assignment or instance. For example, exactness follows if there exist locally optimal hospital schedules for which one binding hospital determines the network makespan in every scenario.

Because AUX-SP is typically cheaper to solve than the full SP, we also use it to generate cuts for the MP so that lower-bounding information can be returned without invoking the exact SP. The mechanism is analogous to the path-deletion cut (8), in that the bound $\mathrm{AUX}^\star(\mathbf{y}^\kappa)$ is attained at the binding hospital and remains valid as long as the surgical episodes assigned to that hospital are not reassigned elsewhere. Let $h^{\star,\kappa} \in \arg\max_{h\in\mathcal{H}} \mathrm{AUX}_h^\star(\mathbf{y}^\kappa)$ denote the binding hospital at iteration κ.

Proposition 2. *The inequality*

$$\theta \geq \mathrm{AUX}^\star(\mathbf{y}^\kappa)\left(1 - \sum_{p\in\mathcal{P}_{h^{\star,\kappa}}(\mathbf{y}^\kappa)} (1 - y_{p,h^{\star,\kappa}})\right) \tag{10}$$

is valid for the integer master epigraph (5)*, and dominates the auxiliary no-good inequality*

$$\theta \geq \mathrm{AUX}^\star(\mathbf{y}^\kappa)\left(1 - \sum_{p\in\mathcal{P}} (1 - y_{p,h^\kappa(p)})\right). \tag{11}$$

The per-hospital decomposition reduces the cost of obtaining feedback for the MP by replacing the full network SP with hospital-level subproblems whose values bound the SP from below and yield the valid inequality (10). On the largest instances, however, even a single hospital subproblem can remain expensive, because resource selection, sequencing, and scenario-wise timing are still solved jointly under uncertainty. The next subsection further refines this by separating the remaining binary scheduling decisions from the scenario-wise timing evaluation, which leads to a three-level decomposition.

Remark 1. To accelerate the solution times of SP and AUX-SP, we can warm-start them using a solution obtained from a deterministic variant. Given an incumbent assignment $\mathbf{y}^\kappa$, we solve, independently and in parallel for each hospital, a deterministic CP model in which all stage durations are set to their scenario averages and resources are aggregated via a cumulative-capacity `Pulse` constraint, adopted for its superior computational efficiency relative to alternative reformulations (Naderi et al., 2023); see Appendix A.2 for the explicit formulation. Because the `Pulse` representation does not distinguish between individual resources, the resulting schedule does not directly specify values for the binary variables $(\mathbf{z}, \mathbf{x})$ in the stochastic model. We assign these values by partitioning the realized intervals at each hospital-stage pair into labeled resources based on start times, and then extracting the corresponding pairwise ordering. The resulting $(\mathbf{y}^\kappa, \mathbf{z}, \mathbf{x})$ is feasible for both SP and AUX-SP and is provided to the solver as a warm start.

4.3. Three-level Decomposition via Inner L-Shaped Reformulations

The two-level scheme spends most of its computational effort inside the SP and AUX-SP, which still couple resource selection, sequencing, and scenario-wise timing under uncertainty. The structure of both subproblems, however, is amenable to a further level of decomposition. Once the binary scheduling decisions $(\mathbf{z}, \mathbf{x})$ are also fixed in addition to $\mathbf{y}^\kappa$, the only variables remaining in the SP and AUX-SPare continuous, subject to linear constraints. Indeed, the both subproblems are two-stage stochastic programs with continuous linear recourse, and can be solved using the L-shaped method. Therefore, in the three-level scheme, we retain the outer MP (5) for the hospital assignment $\mathbf{y}$, and decompose SP into an inner master problem MP-SP over the binary scheduling decisions $(\mathbf{z}, \mathbf{x})$, and scenario-wise timing subproblems SP-SP(ω) to evaluate the resulting schedule in each scenario. The same architecture serves both the full SP and, when invoked, the per-hospital subproblem AUX-SP; for brevity, we present it only for the full SP.

For a fixed outer assignment $\mathbf{y}^\kappa$, the inner L-shaped procedure proceeds in iterations $\nu = 1, 2, \ldots$. At each iteration, the inner master MP-SP is solved to optimality and returns a feasible binary scheduling decision $(\mathbf{z}^{\kappa,\nu}, \mathbf{x}^{\kappa,\nu})$, which we call the *inner incumbent* of iteration ν. For this incumbent, every scenario timing subproblem SP-SP(ω) is then evaluated, and the dual information obtained is used to generate L-shaped optimality cuts that are appended to MP-SP for iteration $\nu+1$. Concretely, SP-SP(ω) is the linear program obtained from the SP (6) after fixing all binary variables, so that only the pathway-precedence inequalities and the non-overlap inequalities activated by (z, x) remain, with $c_{ps}(\omega)$ and $C^{\mathrm{SP}}_{\max}(\omega)$ as its only decision variables. We denote its optimal value by $Q^\kappa(\mathbf{z}, \mathbf{x}; \omega)$, and write $\mathbf{Q}^{\kappa,\nu} := \big(Q^\kappa(\mathbf{z}^{\kappa,\nu}, \mathbf{x}^{\kappa,\nu}; \omega)\big)_{\omega\in\Omega}$ for the scenario vector evaluated at the inner incumbent of iteration ν.

Because SP-SP(ω) is a linear program, LP duality provides an affine LB on $Q^\kappa(\cdot;\omega)$ as a function of $(\mathbf{z},\mathbf{x})$. For each $\omega\in\Omega$, let $\lambda^{\kappa,\nu}(\omega)$ be an optimal dual solution of SP-SP(ω) evaluated at $(\mathbf{z}^{\kappa,\nu},\mathbf{x}^{\kappa,\nu})$, and let $\Theta^{\kappa,\nu}(\mathbf{z},\mathbf{x};\omega)$ be the affine function obtained by substituting $\lambda^{\kappa,\nu}(\omega)$ into the parametric dual objective. Weak duality yields the scenario-wise affine bound

$$\Theta^{\kappa,\nu}(\mathbf{z},\mathbf{x};\omega)\le Q^\kappa(\mathbf{z},\mathbf{x};\omega)\quad \forall \text{ feasible } (\mathbf{z},\mathbf{x}),\qquad \Theta^{\kappa,\nu}(\mathbf{z}^{\kappa,\nu},\mathbf{x}^{\kappa,\nu};\omega)=Q^\kappa(\mathbf{z}^{\kappa,\nu},\mathbf{x}^{\kappa,\nu};\omega),\tag{12}$$

and the equality on the right follows from strong duality. Replacing $Q^\kappa(\mathbf{z},\mathbf{x};\omega)$ in (12) by $C_{\max}^{\text{MP-SP}}(\omega)$ and keeping only the LB yields the corresponding L-shaped optimality cut,

$$C_{\max}^{\text{MP-SP}}(\omega)\ \ge\ \Theta^{\kappa,\nu}(\mathbf{z},\mathbf{x};\omega).\tag{13}$$

The cut (13) is valid because at any feasible $(\mathbf{z},\mathbf{x})$ the inner master enforces $C_{\max}^{\text{MP-SP}}(\omega)\ge Q^\kappa(\mathbf{z},\mathbf{x};\omega)$ (the inner master must dominate the true scenario makespan it predicts), and (12) then provides the affine LB on Q^κ that makes the cut linear in the decision variables of MP-SP. It is also tight at the inner incumbent by the equality in (12), so it strictly improves the inner-master relaxation around the solution that generated it.

In a standard multi-cut L-shaped implementation, one such cut (13) is appended for every scenario at every iteration. The MP-SP then aggregates the scenario makespan variables $\mathbf{C}_{\max}^{\text{MP-SP}}$ through the coherent risk measure $\rho(\cdot)$ in its objective: tightening each $C_{\max}^{\text{MP-SP}}(\omega)$ from below tightens $\rho(\mathbf{C}_{\max}^{\text{MP-SP}})$ from below as well, since ρ is monotone in its argument. The number of cuts added per iteration, however, scales linearly with $|\Omega|$, and many of those cuts contribute nothing to the risk-adjusted objective at the current incumbent because their scenarios carry no weight in the risk evaluation. To identify and discard such cuts, we exploit the dual envelope representation of coherent risk measures (Rockafellar, 2007, Thm. 4(a)),

$$\rho(\mathbf{q})=\max_{\pi\in\Pi_\rho}\sum_{\omega\in\Omega}\pi(\omega)q(\omega),\qquad \mathbf{q}\in\mathbb{R}^{|\Omega|},\tag{14}$$

where $\Pi_\rho\subset\mathbb{R}_+^{|\Omega|}$ is a nonempty compact convex set of probability vectors that depends only on the risk measure. This expresses ρ as a pointwise maximum of linear functions of $\mathbf{q}$, so that fixing any $\pi\in\Pi_\rho$ yields a linear underestimator of ρ. Composing such an underestimator with the scenario-wise bounds (12) produces, for each fixed π, a single linear LB on $\rho\big(\mathbf{Q}^\kappa(\mathbf{z},\mathbf{x};\cdot)\big)$ as a function of $(\mathbf{z},\mathbf{x})$. Among all underestimators in (14), we select at iteration ν the one that is tight at the inner incumbent, namely the envelope vector that attains the maximum when evaluated at $\mathbf{Q}^{\kappa,\nu}$,

$$\pi^{\kappa,\nu}\in\arg\max_{\pi\in\Pi_\rho}\sum_{\omega\in\Omega}\pi(\omega)\,Q^\kappa(\mathbf{z}^{\kappa,\nu},\mathbf{x}^{\kappa,\nu};\omega).\tag{15}$$

By construction, the inner product with $\pi^{\kappa,\nu}$ equals $\rho(\mathbf{Q}^{\kappa,\nu})$ at the incumbent, while at any other feasible $(\mathbf{z},\mathbf{x})$ the same inner product is a valid LB on the true risk value. The scenarios that this maximizer activates,

$$\Gamma^{\kappa,\nu} := \{\omega \in \Omega : \pi^{\kappa,\nu}(\omega) > 0\}, \tag{16}$$

form the *risk-support* set of the incumbent: they are the only scenarios whose values enter the risk evaluation with strictly positive weight at iteration ν. For example, $\Gamma^{\kappa,\nu} = \Omega$ under expectation; under ess-sup, $\Gamma^{\kappa,\nu}$ is the support of a selected optimal distribution over worst-case scenarios; and under CVaR_α, it is the optimal tail support, with VaR-binding scenarios included only when the envelope optimizer assigns them strictly positive weight. The risk-support set determines which cuts we actually append at iteration ν. For any feasible $(\mathbf{z},\mathbf{x})$,

$$\rho\big(\mathbf{Q}^{\kappa}(\mathbf{z},\mathbf{x};\cdot)\big) \ge \sum_{\omega\in\Omega} \pi^{\kappa,\nu}(\omega)\, Q^{\kappa}(\mathbf{z},\mathbf{x};\omega) \ge \sum_{\omega\in\Gamma^{\kappa,\nu}} \pi^{\kappa,\nu}(\omega)\, \Theta^{\kappa,\nu}(\mathbf{z},\mathbf{x};\omega).$$

Cuts for $\omega \notin \Gamma^{\kappa,\nu}$ are inactive at iteration ν in the sense that they do not enter this lower-bounding chain, so omitting them preserves a valid LB on the inner-master objective while reducing the cut count from $|\Omega|$ to $|\Gamma^{\kappa,\nu}|$ per iteration. Using fewer cuts makes MP-SP quicker to re-solve, at the cost of a weaker inner relaxation and potentially more iterations to close the gap (Mínguez et al., 2021).

In formulating the MP-SP, in addition to the L-shaped cuts and the scenario-wise LBs, the precedence binaries $\mathbf{x}$ decide on pairwise orderings. However, they can produce inconsistent triples such as p before p', p' before p'', and p'' before p. Such cyclic decisions do not correspond to a valid schedule. To prevent them, we introduce auxiliary ranking variables u_{psr} that assign the operation of surgical episode p at stage s a position on resource r. Together with Miller-Tucker-Zemlin (MTZ) inequalities (Miller et al., 1960), these variables tie $\mathbf{x}$ to a consistent linear order on every active stage-resource pair, which guarantees an acyclic precedence structure. For a fixed outer assignment $\mathbf{y}^{\kappa}$, the inner master at iteration ν is as follows:

$$\min \quad \theta^{\text{MP-SP}} \tag{17a}$$

$$\text{s.t.} \quad \sum_{r\in\mathcal{R}_{h^{\kappa}(p)s}} z_{psr} = 1 \qquad \forall p \in \mathcal{P},\ s \in \mathcal{S}, \tag{17b}$$

$$C_{\max}^{\text{MP-SP}}(\omega) \ge \Theta^{\kappa,\ell}(\mathbf{z},\mathbf{x};\omega) \qquad \forall \ell = 1,\dots,\nu-1,\ \omega \in \Gamma^{\kappa,\ell}, \tag{17c}$$

$$C_{\max}^{\text{MP-SP}}(\omega) \ge \sum_{\substack{p\in\mathcal{P}:\\ h^{\kappa}(p)=h}} D_{p,h,s}(\omega)\, z_{psr} \qquad \forall h \in \mathcal{H},\ s \in \mathcal{S},\ r \in \mathcal{R}_{hs},\ \omega \in \Omega, \tag{17d}$$

$$C_{\max}^{\text{MP-SP}}(\omega) \ge \sum_{s\in\mathcal{S}} \sum_{r\in\mathcal{R}_{h^{\kappa}(p)s}} D_{p,h^{\kappa}(p),s}(\omega)\, z_{psr} \qquad \forall p \in \mathcal{P},\ \omega \in \Omega, \tag{17e}$$

$$\theta^{\text{MP-SP}} \ge \rho\big(\mathbf{C}_{\max}^{\text{MP-SP}}\big), \tag{17f}$$

$$
\begin{aligned}
&u_{psr} \geq u_{p'sr} + 1 - M_u\big(3 - x_{pp's} - z_{psr} - z_{p'sr}\big) && \forall p, p' \in \mathcal{P}:\ p > p',\ h^\kappa(p) = h^\kappa(p') && \text{(17g)}\\
& && \forall s \in \mathcal{S},\ r \in \mathcal{R}_{h^\kappa(p)s}, &&\\
&u_{p'sr} \geq u_{psr} + 1 - M_u\big(2 + x_{pp's} - z_{psr} - z_{p'sr}\big) && \forall p, p' \in \mathcal{P}:\ p > p',\ h^\kappa(p) = h^\kappa(p') && \text{(17h)}\\
& && \forall s \in \mathcal{S},\ r \in \mathcal{R}_{h^\kappa(p)s}, &&\\
&z_{psr} \in \{0,1\} && \forall p \in \mathcal{P},\ s \in \mathcal{S},\ r \in \mathcal{R}_{h^\kappa(p)s}, && \text{(17i)}\\
&x_{pp's} \in \{0,1\} && \forall p, p' \in \mathcal{P}:\ p > p',\quad h^\kappa(p) = h^\kappa(p'),\ s \in \mathcal{S}, && \text{(17j)}\\
&0 \leq u_{psr} \leq |\mathcal{P}| - 1 && \forall p \in \mathcal{P},\ s \in \mathcal{S},\ r \in \mathcal{R}_{h^\kappa(p)s}. && \text{(17k)}
\end{aligned}
$$

Constraint (17b) selects exactly one resource per stage within each surgical episode's assigned hospital, and (17i)-(17j) define the support of the scheduling binaries. The inequalities (17c) are the L-shaped optimality cuts generated in prior inner iterations, restricted at each iteration ℓ to the risk-support set $\Gamma^{\kappa,\ell}$. Inequalities (17d) and (17e) are counterparts of (5f) and (5e) in the outer MP respectively, so the inner master carries useful timing information even before any L-shaped cut has been generated. Constraint (17f) aggregates the scenario-wise LBs through ρ, using the appropriate epigraph representation for the selected risk measure. The ranking inequalities (17g)-(17h) are active only when p and p' are assigned to the same resource r at stage s, such that if $x_{pp's} = 1$, then (17g) ranks p after p', and if $x_{pp's} = 0$, then (17h) ranks p' after p. Directed cycles are therefore excluded on every active stage-resource pair. With $0 \leq u_{psr} \leq |\mathcal{P}| - 1$ and $M_u = |\mathcal{P}|$, the same inequalities become relaxed whenever the two operations do not share the resource or the corresponding pairwise order is inactive.

Two observations make it straightforward to argue that the inner loop converges. First, for any feasible $(\mathbf{z}, \mathbf{x}, \mathbf{u})$ of MP-SP, the binary structure imposed by (17b) together with the ranking inequalities (17g)-(17h) induces an acyclic precedence structure, so each scenario timing LP SP-SP(ω) has a finite optimum and the inner decomposition has relatively complete recourse. The only feedback the inner master receives from the scenario level SP-SP(ω) therefore consists of the optimality cuts (17c). Second, once the cuts generated at an inner incumbent $(\mathbf{z}^{\kappa,\ell}, \mathbf{x}^{\kappa,\ell})$ are present in MP-SP, revisiting the same incumbent at a later iteration cannot improve the inner-master LB beyond $\rho(\mathbf{Q}^{\kappa,\ell})$. The upper and lower bounds maintained by the inner algorithm must therefore coincide whenever an already evaluated binary decision is selected again. Because the set of feasible binary scheduling decisions is finite, only finitely many distinct incumbents can be visited, and the procedure terminates.

Theorem 2. *Fix an outer assignment $\mathbf{y}^\kappa$. Let $\mathcal{D}^\kappa$ denote the set of binary scheduling decisions $(\mathbf{z}, \mathbf{x})$ for which there exist ranking variables $\mathbf{u}$ satisfying (17b) and (17g)-(17k). The risk-support inner L-shaped algorithm generates a non-decreasing sequence of valid LBs on $\beta^{\star,\kappa} := \min_{(\mathbf{z},\mathbf{x})\in\mathcal{D}^\kappa} \rho\big(\mathbf{Q}^\kappa(\mathbf{z}, \mathbf{x}; \cdot)\big)$ and terminates finitely at a decision $(\mathbf{z}, \mathbf{x}) \in \mathcal{D}^\kappa$ attaining $\beta^{\star,\kappa}$.*

The inner L-shaped method is illustrated in Figure 2. It iterates between MP-SP and the scenario-wise SP-SP(ω) and returns the converged incumbent and risk value to the outer LBBD loop, summarized in Figure 3. At each outer iteration, the MP proposes a hospital assignment $\mathbf{y}^\kappa$, then the AUX-SP is solved and, if needed, returns the cut (10). Since $\text{AUX}^\star(\mathbf{y}^\kappa)$ generally underestimates $\text{SP}^\star(\mathbf{y}^\kappa)$, the MP could repeatedly re-propose $\mathbf{y}^\kappa$ without ever tightening θ^κ, so we cache each assignment's AUX-SP and SP values to detect repeats and skip directly to the exact SP evaluation. Once the auxiliary loop terminates, the full SP is solved and the LBBD cut (8) is appended to the MP.

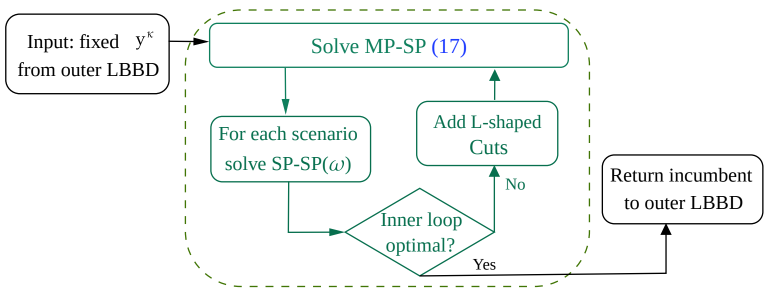

Figure 2 Inner L-shaped method used to solve an AUX-SP or SP for a fixed hospital assignment.

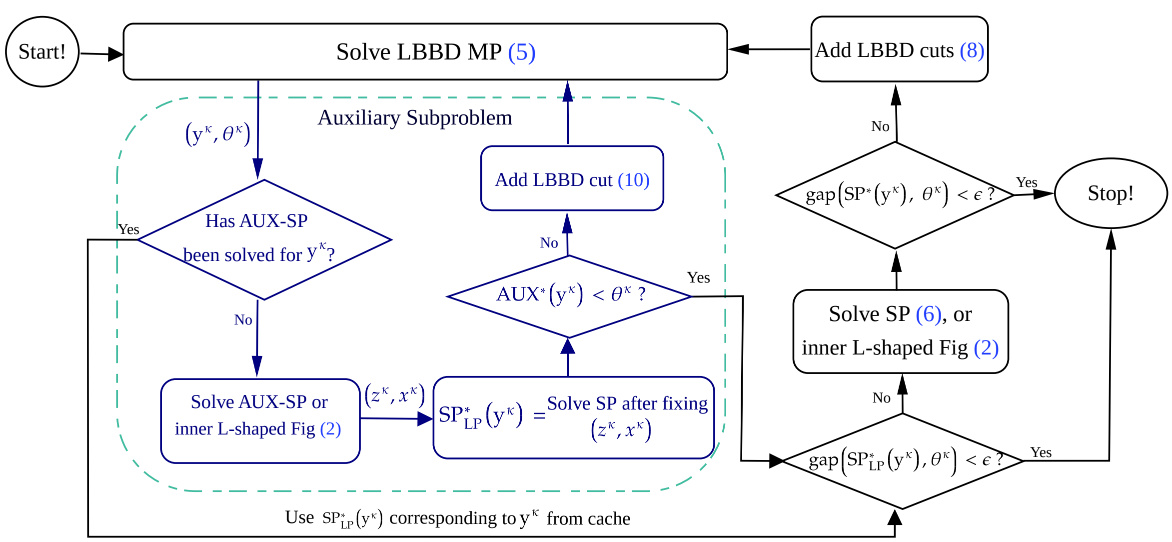

Figure 3 Outer LBBD framework with the AUX-SP.

4.4. Graph-based Evaluation of SP-SP(ω) and L-shaped Cuts

The inequalities (17g)-(17h) impose a strict order on the operations that share an active stage-resource pair, so for any fixed outer assignment $\mathbf{y}^{\kappa}$ and inner incumbent $(\mathbf{z}^{\kappa,\nu}, \mathbf{x}^{\kappa,\nu})$ the induced precedence structure is acyclic. This means that each scenario timing problem SP-SP(ω) can be evaluated as a longest-path problem on a directed acyclic graph (DAG), rather than as a LP. The same DAG also returns, at no extra cost, the dual information needed to build the L-shaped cuts (17c).

For fixed $(\mathbf{y}^{\kappa}, \mathbf{z}^{\kappa,\nu}, \mathbf{x}^{\kappa,\nu})$ and scenario ω, let

$$\mathcal{G}^{\kappa,\nu}(\omega) = \big(\mathcal{V}^{\kappa,\nu}(\omega), \mathcal{A}^{\kappa,\nu}(\omega)\big), \qquad \mathcal{V}^{\kappa,\nu}(\omega) := \{o,t\} \cup \{v_{ps} : p \in \mathcal{P},\ s \in \mathcal{S}\},$$

where o and t are a source and a sink, and each v_{ps} represents the operation of surgical episode p at stage s of its assigned hospital. The arc set $\mathcal{A}^{\kappa,\nu}(\omega)$ is built from four families. For each $p \in \mathcal{P}$, it contains the source arc $(o, v_{p,1})$ with weight $D_{p,h^{\kappa}(p),1}(\omega)$ and the sink arc $(v_{p,|\mathcal{S}|}, t)$ with zero weight. For each $p \in \mathcal{P}$ and $s = 2, \dots, |\mathcal{S}|$, it contains the pathway arc $(v_{p,s-1}, v_{ps})$ with weight $D_{p,h^{\kappa}(p),s}(\omega)$. Finally, for each pair $p > p'$ assigned to the same hospital whose operations share resource r at stage s, i.e., $z^{\kappa,\nu}_{psr} = z^{\kappa,\nu}_{p'sr} = 1$, it contains the resource-precedence arc dictated by $\mathbf{x}^{\kappa,\nu}$, such that if $x^{\kappa,\nu}_{pp's} = 1$, the arc $(v_{p's}, v_{ps})$ is added with weight $D_{p,h^{\kappa}(p),s}(\omega)$, otherwise, $(v_{ps}, v_{p's})$ is added with weight $D_{p',h^{\kappa}(p'),s}(\omega)$. Because the ranking variables remove directed cycles on every active stage-resource pair, $\mathcal{G}^{\kappa,\nu}(\omega)$ is a DAG. Appendix D.2 illustrates this construction on a small instance.

The scenario makespan equals the length of the longest o-t path in $\mathcal{G}^{\kappa,\nu}(\omega)$. Let $w^{\kappa,\nu}_{qv}(\omega)$ be the weight of arc (q, v) and

$$\delta^{\kappa,\nu}(v;\omega) := \{q \in \mathcal{V}^{\kappa,\nu}(\omega) : (q,v) \in \mathcal{A}^{\kappa,\nu}(\omega)\}$$

the set of immediate predecessors of node v. With $L^{\kappa,\nu}_{o}(\omega) = 0$, the longest-path labels satisfy

$$L^{\kappa,\nu}_{v}(\omega) = \max_{q \in \delta^{\kappa,\nu}(v;\omega)} \big\{L^{\kappa,\nu}_{q}(\omega) + w^{\kappa,\nu}_{qv}(\omega)\big\}, \qquad \forall v \in \mathcal{V}^{\kappa,\nu}(\omega) \setminus \{o\},$$

and the scenario value is obtained as $Q^{\kappa}(\mathbf{z}^{\kappa,\nu}, \mathbf{x}^{\kappa,\nu}; \omega) = L^{\kappa,\nu}_{t}(\omega)$. A single topological-order pass over $\mathcal{G}^{\kappa,\nu}(\omega)$ computes all labels in $O\big(|\mathcal{V}^{\kappa,\nu}(\omega)| + |\mathcal{A}^{\kappa,\nu}(\omega)|\big)$ time (Cormen et al., 2009, Sec. 24.2), so each scenario evaluation becomes a linear-time graph computation instead of a generic LP solve.

The longest-path computation also provides the dual information required by the L-shaped cut. The dual of the longest-path representation is a unit-flow problem on $\mathcal{G}^{\kappa,\nu}(\omega)$. Any longest o-t path therefore yields an extreme optimal dual solution, with arc multiplier equal to 1 on the selected path and 0 elsewhere. Substituting these path-incidence multipliers into the parametric dual objective constructs the affine function $\Theta^{\kappa,\nu}(\mathbf{z}, \mathbf{x}; \omega)$ used in (17c).

5. Computational Results

We evaluate the proposed frameworks on practice-inspired instances of RA-SEEN at the University Health Network in the City of Toronto. Section 5.1 describes the test instances, Section 5.2 reports implementation details, and Section 5.3 identifies a representative configuration for each decomposition level. Section 5.4 compares the best configurations with the compact MIP and CP benchmarks across problem regimes, and Section 5.5 examines the operational implications of the resulting schedules.

5.1. Dataset

To the best of our knowledge, no public benchmark set exists for RA-SEEN or for closely related multi-hospital, multi-stage surgical episode scheduling problems. We therefore develop a practice-informed instance generator based on three sources: publicly available information on the UHN surgical pathway and activity levels, published evidence on perioperative service times and resource congestion, and consultation with a domain expert. Where site-level numerical data were unavailable, we use expert-informed calibration rules designed to preserve realistic relationships among patient volumes, stage durations, resource capacities, and upstream and downstream congestion. The complete generator configuration and generated instances are made publicly available; further details are provided in Appendix E.

The network contains three hospitals corresponding to the UHN sites in our study. We consider either the five-stage UHN day-of-surgery pathway (UHN, 2024b), described in Section 2.1, or a four-stage variant that aggregates PACU and DSU into one recovery stage. Instance sizes vary over the number of surgical episodes $|\mathcal{P}| \in \{10, 15, 20, 40, 60, 80, 100, 120, 140\}$, and the number of scenarios used to represent duration uncertainty $|\Omega| \in \{100, 250, 500\}$. The upper end of the patient range is consistent with UHN's reported 31,004 acute day-surgery cases in 2022/23, corresponding to approximately 100-120 cases per weekday (UHN, 2023). The scenario sets are nested, so larger sets extend the smaller ones without changing the underlying base instance.

OR capacity scales with the number of episodes, and the capacities of REG, POCU, PACU, and DSU are calibrated relative to OR capacity. These relationships reflect the shorter and more flexible nature of upstream activities while retaining sufficiently constrained postoperative capacity for downstream congestion to emerge. This design is consistent with evidence that PACU flow can be limited by bed availability, staffing, discharge readiness, and downstream coordination (Weissman et al., 2019; Dexter et al., 2005; Heyba et al., 2024). Each episode is assigned to one of four same-day surgery classes representing short, medium, long, and complex cases. Registration is modeled using a triangular distribution, while the remaining baseline stage durations follow class- and stage-specific lognormal distributions. The duration parameters were calibrated through expert consultation and assessed, where comparable evidence was available, against reported preoperative assessment, OR, and postoperative recovery times (Taylor et al., 2018; Bhattacharyya, 2011; White et al., 2003). Lognormal distributions capture the right-skewed behavior observed in surgical durations (Strum et al., 2000; Roshanaei et al., 2017b), and a Gaussian copula introduces positive within-episode dependence across stages (Nelsen, 2006). Persistent hospital- and stage-level differences are represented through unit-mean multiplicative factors.

Finally, scenarios are generated through episode-level random variation with structured operating-day conditions, including regular days, OR-centered and recovery-centered disruptions, system-wide slowdowns, and hospital-specific disruptions. These patterns allow durations at related stages to increase jointly rather than varying independently, and are motivated by evidence on operational drivers of surgery and postoperative recovery (Kayıcs et al., 2015; Wang et al., 2020; Pavlin et al., 1998; Johnson and Calixte, 2023).

5.2. Implementation Details

All algorithms are implemented in Python 3.10.13 and executed on the Rorqual cluster operated by Calcul Québec as part of the Digital Research Alliance of Canada (Calcul Québec, 2026). We use Gurobi 12.0.0 (Gurobi Optimization, LLC, 2026) for MIP models and IBM ILOG CP Optimizer 22.1.1.0 (IBM, 2022) for CP models. Independent instances are dispatched in parallel through GNU Parallel (Tange, 2023) with 32 GB of RAM per run.

The algorithmic comparisons of Section 5.3 use a branch-and-cut (B&C) implementation of every decomposition variant with a one-hour overall time limit, evaluated on 30 instances per setting. The performance experiments of Section 5.4 use the selected configurations from Section 5.3 with a five-hour time limit on 5 instances per setting, and the outer LBBD loop is run as a cutting-plane scheme. The relative optimality-gap tolerance is set to 0.01% for small instances and 5% otherwise. On medium and large instances, each SP or AUX-SP solve is limited to 45 and 60 minutes, respectively, with the deterministic warm start of Remark 1 given a five-minute time limit per hospital. When an SP terminates early without providing the optimal solution, its LB is used to generate cuts and its best incumbent feeds the upper bound in the reported gaps. In the three-level LBBD, inner L-shaped cuts and the acyclicity constraints (17g)-(17h) are generated dynamically within the B&C framework rather than imposed upfront. Throughout, we evaluate the computational performance of the compact models and decomposition algorithms, as well as the resulting managerial insights, under $\rho \in \{\mathbb{E}, \mathrm{CVaR}_{90\%}, \mathrm{CVaR}_{95\%}, \text{ess-sup}\}$, spanning risk-neutral to worst-case planning.

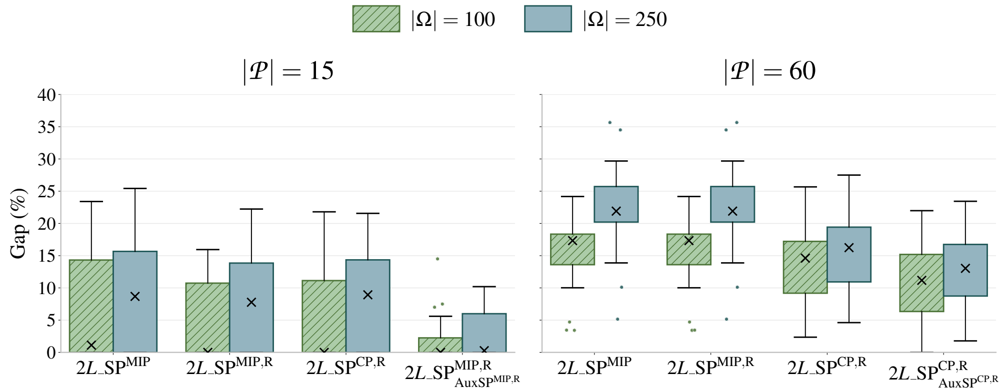

Figure 4 Relative optimality gaps of the main two-level decomposition variants for $|\mathcal{P}| = 15$ and $|\mathcal{P}| = 60$, with $|\Omega| \in \{100, 250\}$ and $\alpha = 95\%$.

5.3. Algorithm Configuration Choice

Before comparing decomposition methods with the compact benchmarks, we identify one representative implementation for each decomposition level. We evaluate the candidate variants on small and medium instances with $|\mathcal{P}| = 15$ and $|\mathcal{P}| = 60$, under $|\Omega| \in \{100, 250\}$ and $\alpha = 95\%$, using the final relative optimality gap as the comparison metric. We label each variant systematically: the prefix $2L$ or $3L$ marks the two-level or three-level scheme; the SP superscript gives the subproblem technology (MIP or CP); a superscript R marks the path-deletion cut derived from SP; a superscript L marks an inner L-shaped decomposition and Net its longest-path graph evaluation (Section 4.4); an AuxSP subscript marks that the auxiliary subproblem is solved before the SP; and, when R appears as a superscript on AuxSP itself, it marks the valid inequality derived from AUX-SP rather than from SP.

For the two-level decomposition, the most effective implementation depends on instance size, as shown in Figure 4. On small instances, the MIP-based SP remains tractable and is most effective when combined with the path-deletion cut (8) and the auxiliary cut (10), namely $2L\_\mathrm{SP}^{\mathrm{MIP,R}}_{\mathrm{AuxSP}^{\mathrm{MIP},R}}$; for $|\mathcal{P}| = 15$ this configuration closes the gap almost entirely at $|\Omega| = 100$ and keeps the mean gap below 1% at $|\Omega| = 250$. On medium and large instances the CP-based SP $2L\_\mathrm{SP}^{\mathrm{CP,R}}_{\mathrm{AuxSP}^{\mathrm{CP,R}}}$ becomes more reliable: for $|\mathcal{P}| = 60$ it reduces the mean gap to roughly 14%-16%, compared with 17%-22% for the MIP-based variants. We therefore use the MIP-based variant for small instances and the CP-based variant for medium and large instances.

For the three-level decomposition, we keep the outer LBBD enhancements selected above and evaluate the setup of the inner decomposition. Figure 5 isolates its effect: relative to the LP-based inner evaluation ($3L\_\mathrm{SP}^{\mathrm{L,R}}$), replacing the scenario-evaluation LPs with the longest-path graph evaluation of Section 4.4 ($3L\_\mathrm{SP}^{\mathrm{L,Net,R}}$) reduces both the level and the variability of the final gaps in both SP and AUX-SP, with the best performance obtained when the AUX-SP is also used ($3L\_\mathrm{SP}^{\mathrm{L,Net,R}}_{\mathrm{AUX\text{-}SP}^{\mathrm{L,Net,R}}}$). Figure 6 identifies the source of this improvement: relative to the LP-based evaluation, the dynamic-programming implementation is 10-12 times faster for $|\mathcal{P}| = 15$ and 33-42 times faster for $|\mathcal{P}| = 60$, and the speedup grows with the number of surgical episodes and scenarios. We therefore adopt the graph-based variant as the three-level implementation in the experiments that follow. Going forward, the selected decomposition methods will be denoted by 2LVL and 3LVL in the figures and tables.

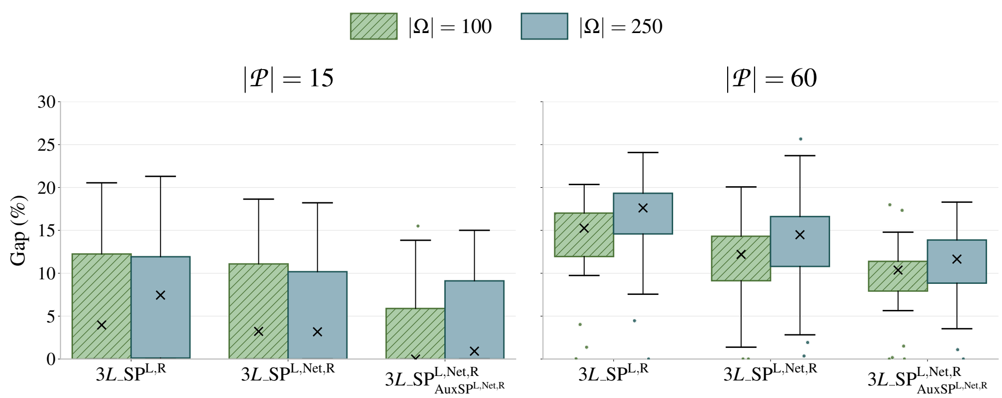

Figure 5 Relative optimality gaps of the main three-level decomposition variants for $|\mathcal{P}| = 15$ and $|\mathcal{P}| = 60$, with $|\Omega| \in \{100, 250\}$ and $\alpha = 95\%$.

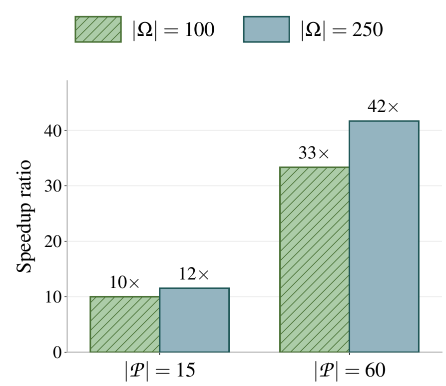

Figure 6 Speedup from evaluating the three-level scenario subproblems through the longest-path graph formulation instead of solving their LP formulations, for $|\mathcal{P}| \in \{15, 60\}$ and $|\Omega| \in \{100, 250\}$.

5.4. Computational Performance

We now compare performance of the selected decomposition methods, with the compact MIP and CP models. Detailed result tables are reported in Appendix F. The discussion below summarizes the main findings on reliability, run time, optimality gaps, and the quality of the auxiliary LB.

5.4.1. Computational Reliability and Solution Time We first assess reliability in the most challenging setting, with $|\mathcal{S}|=5$ and $|\Omega|=500$. Figure 7 reports, for each configuration, the number of instances out of five that terminate within a 5% gap (top panel) and the number that fail to return a feasible incumbent (bottom panel), with one bar per methodology (runs not included in either panel return feasible solutions with gaps larger than 5%). Compact MIP and CP perform well only on the smallest instances and deteriorate quickly as $|\mathcal{P}|$ grows. The two-level method remains effective over a wider range but begins to struggle on the larger instances. In contrast, the three-level method remains the most reliable approach at scale, and is the only viable option when the compact models and two-level method fail to find feasible solutions.

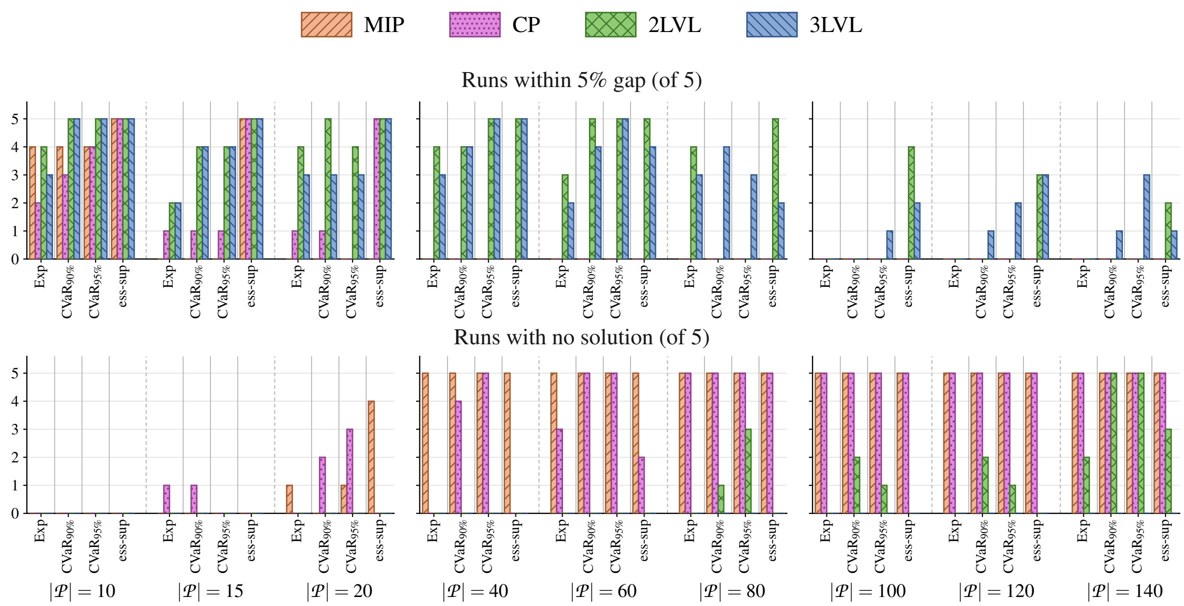

Figure 7 Number of 5%-optimal runs (top) and runs with no feasible solution (bottom), out of five, for $|\mathcal{S}|=5$ and $|\Omega|=500$, by surgical episode size and risk measure. Each bar reports the count for one methodology (MIP, CP, two-level, three-level).

Decomposition methods are not only essential for solving larger instances, but they are also generally faster than the compact models even when they can return feasible solutions. Figure 8 reports median solution times under $\text{CVaR}_{95\%}$ on small instances. The compact models often hit the time limit as $|\mathcal{P}|$, $|\mathcal{S}|$, or $|\Omega|$ grows, while the two-level method usually terminates within seconds or minutes. The three-level method incurs additional overhead from its inner decomposition and is therefore sometimes slower than two-level on small instances, but it consistently outperforms the compact models.

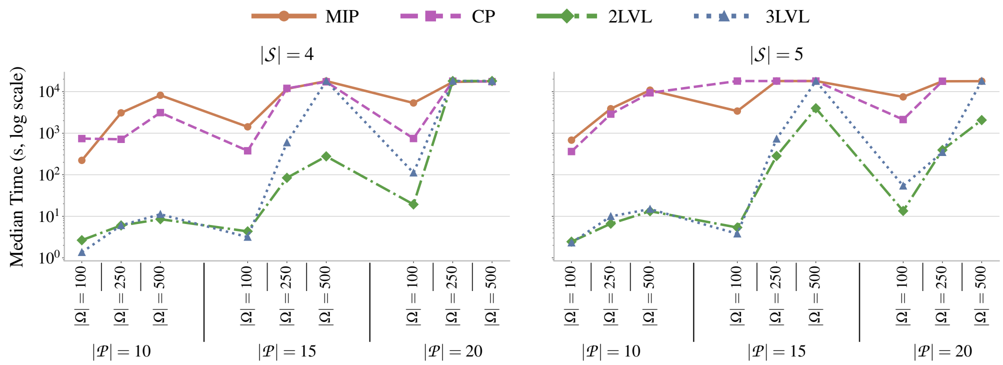

Figure 8 Median solution time (seconds, log scale) for the $\text{CVaR}_{95\%}$ model on small instances. Algorithms are terminated either when optimality is proven or when the five-hour time limit is reached.

5.4.2. Optimality Gaps We measure solution quality through the relative optimality gap. Since RA-SEEN is a minimization problem, the best incumbent objective, obtained from the SP, is an upper bound on the optimal value, and the best relaxation bound, obtained from the MP, is a LB. The optimal value therefore lies between these bounds. The gap is defined as

$$\text{optimality gap} := \frac{\text{best incumbent objective} - \text{best relaxation bound}}{\text{best incumbent objective}} \times 100\%.$$

Figures 9 and 10 report average final optimality gaps under two levels of problem difficulty, controlled by the number of stages and scenarios. In both figures, gray cells indicate configurations without a feasible incumbent, and black outlines identify the smallest average gap for each surgical episode size within each risk-measure panel. For moderate difficulty ($|\mathcal{S}| = 4$, $|\Omega| = 100$), compact MIP solves only the smaller instances, while compact CP remains functional over a wider range but with increasing gaps. Decomposition is more reliable beyond small instances, with the two-level method typically returning the smallest gaps, since the full scheduling SP is tractable at this difficulty level. The three-level method is still competitive, though its additional inner decomposition does not consistently improve performance. For higher difficulty ($|\mathcal{S}| = 5$, $|\Omega| = 500$), compact MIP and CP are unavailable for most medium and large instances, and the two-level method, while still effective when the SP is solvable, sees its gaps widen or become unavailable on the largest risk-averse cases. The three-level method performs well throughout, reaching average gaps at $|\mathcal{P}| = 140$ of 13.43%, 9.36%, 6.54%, and 6.95% under expectation, $\text{CVaR}_{90\%}$, $\text{CVaR}_{95\%}$, and ess-sup, respectively. Overall, the two-level method is preferable when the SP is manageable, while the three-level method becomes the only option once the computational cost of solving the SP becomes too high.

5.4.3. Quality of the Per-Hospital Auxiliary Lower Bound We examine the quality of AUX-SP as a surrogate for the full SP. For each incumbent assignment, we take the schedule

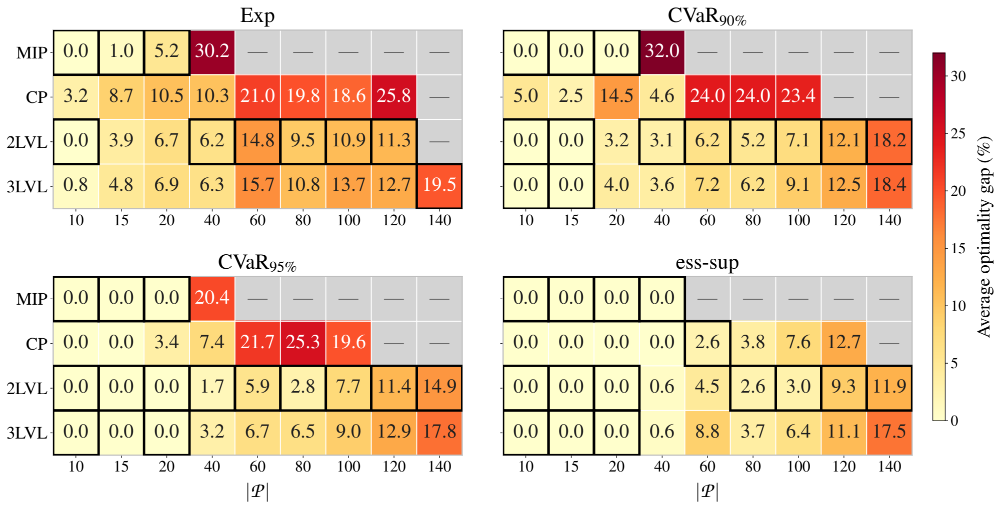

Figure 9 Average optimality gaps (%) for $|\mathcal{S}| = 4$ and $|\Omega| = 100$, by method, risk measure, and number of surgical episodes.

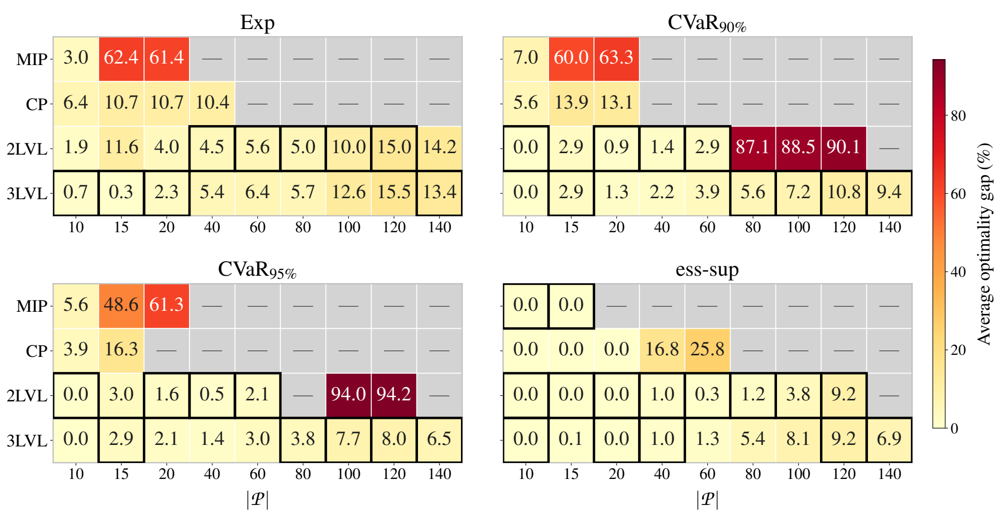

Figure 10 Average optimality gaps (%) for $|\mathcal{S}| = 5$ and $|\Omega| = 500$, by method, risk measure, and number of surgical episodes.

returned by AUX-SP and report the relative difference between the SP and AUX-SP objective values at that same schedule, normalized by the SP value. As shown in Figure 11, the average difference remains below 0.2% under expectation, $\text{CVaR}_{90\%}$, and $\text{CVaR}_{95\%}$, with no monotone pattern across risk measures. The difference is effectively zero under ess-sup, which is consistent with the tightness condition discussed in Section 4.2. Thus AUX-SP provides a much cheaper but empirically tight signal for guiding the MP.

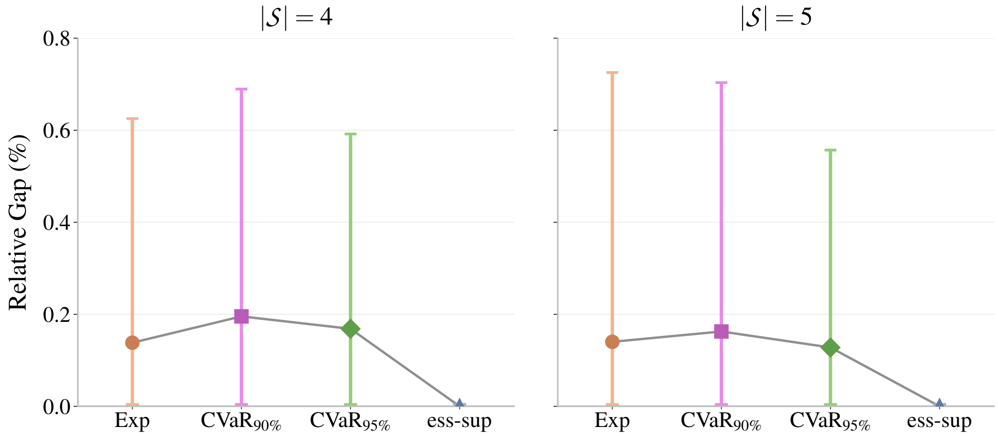

Figure 11 Relative difference between the SP and AUX-SP objective values at the schedule returned by AUX-SP, normalized by the SP value. The left and right panels correspond to $|\mathcal{S}| = 4$ and $|\mathcal{S}| = 5$, respectively. Markers show averages, error bars show variability across instances.

5.5. Managerial Insights

Because the proposed methods scale to realistic instance sizes, the managerial analysis below is based on optimal or low-gap solutions. We use the optimized schedules to examine three operational questions: how much average-duration planning underestimates the impact of adverse outcomes, where local disruptions propagate most strongly along the surgical episode pathway, and how much value is gained by optimizing schedules under uncertainty.

5.5.1. Gap Between Average-Case and Adverse-Case Performance A natural question is how much longer the schedule runs on adverse days than on an average day. For a fixed schedule and risk measure ρ, we report the relative gap

$$\big(\rho(\mathbf{C}_{\max})/\mathbb{E}[\mathbf{C}_{\max}] - 1\big) \times 100,$$

which gives the percentage by which the risk-adjusted makespan exceeds the expected makespan. A value of 25%, for example, indicates that the risk-adjusted makespan on adverse days is 25% longer than on an average day. Figure 12 reports this gap for $|\mathcal{S}| = 5$, with three trends emerging. First, as expected the gap grows with risk aversion: CVaR$_{95\%}$ is consistently above CVaR$_{90\%}$, and ess-sup is largest in every group. Second, the gap grows with the number of scenarios, most visibly for ess-sup, because larger scenario sets are more likely to contain extreme realizations. Third, the largest *relative* gap occurs on medium-sized instances rather than the larger ones. On large instances, the expected makespan increases faster than the risk-adjusted makespan, which reduces the relative gap even though the absolute difference remains large. Across all settings, the gap ranges from about 18% to 37%. Under 500 scenarios, ess-sup reaches 34%, 37%, and 31% on small, medium, and large instances, respectively. Consequently, planning around average durations alone understates how long the system runs on adverse days by roughly one-fifth to one-third, with the largest understatement under worst-case criteria.

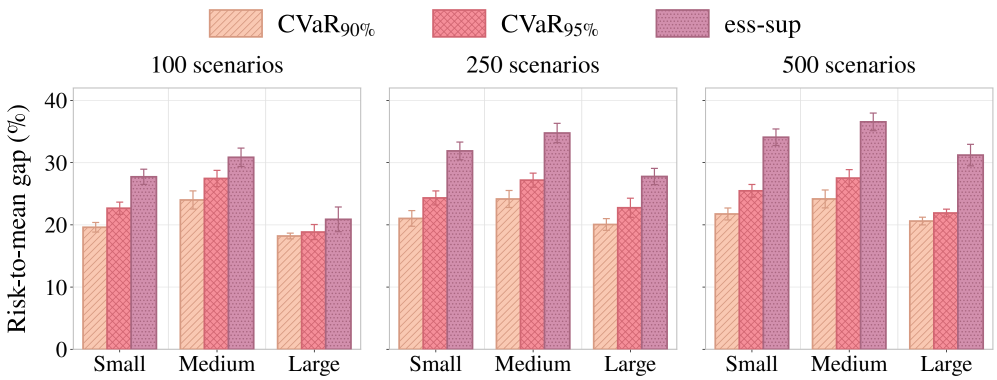

Figure 12 Risk-to-mean gap (%) for $|\mathcal{S}| = 5$, grouped by instance size and scenario count. Bars correspond to $\mathrm{CVaR}_{90\%}$, $\mathrm{CVaR}_{95\%}$, and ess-sup; error bars show variability across instances.

5.5.2. Disruption Propagation Across the Surgical Pathway We next ask which stages drive system-wide delays when their durations are perturbed. Let σ denote a fixed schedule and $\phi > 0$ a proportional disruption level. The disruption propagation index (DPI) of stage s is

$$\mathrm{DPI}_s(\sigma) := \frac{\rho(\mathbf{C}_{\max}^{\sigma,s,\phi}) - \rho(\mathbf{C}_{\max}^{\sigma})}{\phi\, \rho(\mathbf{C}_{\max}^{\sigma})} \times 100,$$

where $\mathbf{C}_{\max}^{\sigma,s,\phi}$ is the makespan vector after increasing every duration at stage s by the factor $1+\phi$. The value DPI_s is the elasticity of the risk-adjusted makespan with respect to stage-s durations at disruption level ϕ: a value of 50% means that increasing every stage-s duration by $\phi = 10\%$ lengthens the risk-adjusted makespan by 5%. Figure 13 depicts the average DPI by stage under four disruption levels. We can see four patterns. First, REG and POCU have DPI values below 3% and 10%, respectively, across all disruption levels and risk measures. Therefore, delays at intake and pre-operative check-in are absorbed before patients reach the operating room, and slack in earlier stages prevents these delays from reaching the bottleneck resources. Second, OR, PACU, and DSU are the dominant propagators, with DPI values of roughly 34-42%, 31-44%, and 25-32%, respectively. Each extra minute in surgery or recovery shifts every subsequent operation on the same resource, so disruptions to these stages translate almost directly into a longer day. Third, the ranking of OR and PACU depends on the risk measure. Under expectation, OR is the most disruptive stage, consistent with ORs being the scarcest resource on a typical day. Under risk-averse $\mathrm{CVaR}_{90\%}$, $\mathrm{CVaR}_{95\%}$, and ess-sup, PACU is more disruptive than OR in every panel and

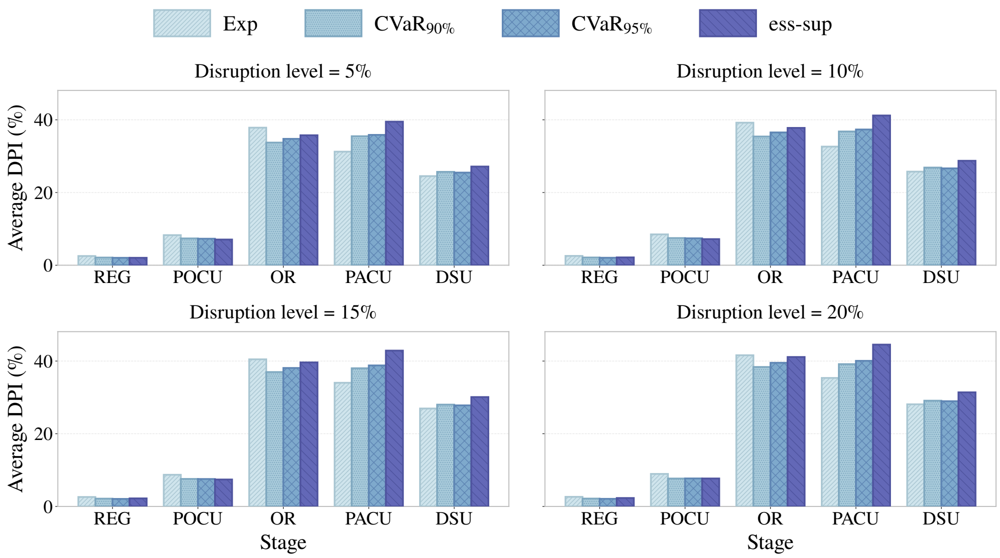

Figure 13 Average DPI (%) by stage for $|\mathcal{S}| = 5$, under four disruption levels and four risk measures.

reaches 44% at the 20% disruption level, meaning that on adverse days the recovery unit, not the operating room, governs how late the system runs. Fourth, the DPI of each stage barely changes as the disruption level grows from 5% to 20%. This means that the stages that propagate delay are determined by the schedule itself, not by how large the delay happens to be, so the same stages need to be protected whether disruptions are small or large. Taken together, the results show that protecting OR capacity alone is insufficient, and resilience requires coordinated planning of OR, PACU, and DSU, with the recovery and discharge stages becoming the binding constraint on the adverse days that planners most want to guard against.

5.5.3. Value of Stochastic Scheduling We close by measuring the value of the stochastic solution (VSS), defined as the relative improvement of the stochastic schedule over the deterministic schedule built from average durations, with both schedules evaluated on the same stochastic scenario set. A larger VSS means that explicitly modeling uncertainty provides a schedule that performs noticeably better than one built on mean values. Figure 14 presents the average VSS of instances with $|\Omega| = 100$. It shows that the VSS is smallest under expectation and largest under ess-sup for every $|\mathcal{P}|$, meaning that it increases with risk aversion. A schedule built from average durations places no buffer where adverse scenarios happen, so the gap between deterministic and stochastic planning widens as the planner cares more about longer durations. The VSS also grows consistently with system size up to $|\mathcal{P}| = 100$, reaching about 20% under expectation and nearly 35% under ess-sup. Larger instances have more sequencing and assignment choices, and the stochastic model exploits this flexibility to shift risk away from the bottleneck resources, while the deterministic model cannot. On the largest instances, the VSS slightly decreases but stays above

17%. We attribute this drop to the optimality gap on the largest instances, which are harder to solve to proven optimality within the time limit. The stochastic solutions reported are therefore conservative estimates of the true VSS. Even so, the value of uncertainty-aware planning is substantial across all instance sizes, which suggests that the approach is valuable in practice even when only approximate solutions are available.

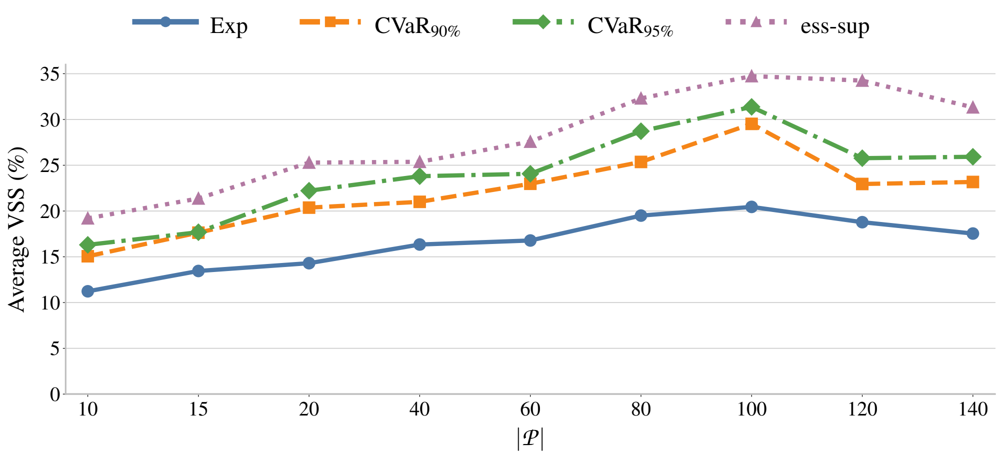

Figure 14 Average VSS by risk measure for instances with $|\mathcal{S}| = 5$ and $|\Omega| = 100$. VSS is computed as the relative improvement of the stochastic solution over the deterministic benchmark, with both evaluated under the stochastic scenario set.

6. Conclusion

We studied risk-aware surgical episode scheduling across a network of collaborating hospitals, motivated by the practice at the University Health Network in the City of Toronto. We modeled the full same-day perioperative pathway under uncertain stage durations, and developed MIP and CP formulations for a broad class of coherent risk measures, and proposed two- and three-level exact decomposition methods that separate hospital assignment from resource assignment, sequencing, and scenario-wise timing. The algorithms are strengthened by valid inequalities, logic-based optimality cuts, auxiliary hospital-level valid inequalities, and a directed-acyclic-graph reformulation of scenario timing as a longest-path problem. Compact formulations remain useful only on small instances, and the two-level method is effective when the scheduling subproblem is tractable, whereas the three-level method is the most robust at scale, attaining an average optimality gap of 6.54% on the largest instances with 140 surgical episodes and 500 scenarios under $\text{CVaR}_{95\%}$.

The computational study also offers three managerial insights. First, planning around average durations can understate makespans on adverse days, by one-fifth to one-third in our instances, with the largest gap under worst-case criteria. Second, delays do not stay local, as they propagate along the pathway through the OR, PACU, and DSU. Protecting OR capacity alone is therefore insufficient. Third, the value of the stochastic solution grows with both system scale and risk aversion, reaching about 35% under worst-case planning. Together, these findings argue for end-to-end, risk-aware planning of the full perioperative pathway, particularly in publicly funded systems where capacity is shared and access is the primary policy concern.

The framework can be extended in several directions. A multistage stochastic formulation would allow cancellations, rebooking, and backlog to be carried across operating days, and balance same-day efficiency against longer-horizon access targets. Stage capacities, currently inputs of the problem, could be optimized jointly with the schedule to study the trade-off between adding OR time and adding recovery beds. Finally, our model treats uncertainty as exogenous, whereas scheduling decisions can themselves shape stage durations, for example, through crowding effects in the PACU or fatigue-induced slowdowns when surgeries are sequenced back-to-back. Capturing this feedback would require an endogenous-uncertainty formulation and new solution methodologies tailored to it.

Acknowledgment

Maryam Daryalal is supported by the Natural Sciences and Engineering Research Council of Canada [Discovery Grant RGPIN-2024-05908] and by the Canada Research Chairs program [CRC-2024-00178]. This research was enabled in part by support provided by Calcul Québec (calculquebec.ca) and the Digital Research Alliance of Canada (alliancecan.ca), using the Rorqual supercomputer.

References

Ph. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Thinking coherently. *RISK*, 10:68–71, 1997.

S. Batun, B. T. Denton, T. R. Huschka, and A. J. Schaefer. Operating room pooling and parallel surgery processing under uncertainty. *INFORMS Journal on Computing*, 23(2):220–237, 2010.

J. F. Benders. Partitioning procedures for solving mixed-variables programming problems. *Numerische Mathematik*, 4(1):238–252, 1962.

N. Bhattacharyya. Benchmarks for the durations of ambulatory surgical procedures in otolaryngology. *Annals of Otology, Rhinology & Laryngology*, 120(11):727–731, 2011. doi: 10.1177/000348941112001106.

Calcul Québec. Rorqual: Cluster and services description (Rorqual User Portal), 2026. Available at https://metrix.rorqual.calculquebec.ca/ (last accessed date: February 23, 2026).

CIHI. Surgical wait times in Canada — Dr. David Urbach. Canadian Health Information Podcast, March 2023. Available at https://www.cihi.ca/en/podcast/surgical-wait-times-in-canada-dr-david-urbach (last accessed date: July 28, 2026).

CIHI. National health expenditure trends, November 2025a. Available at https://www.cihi.ca/en/national-health-expenditure-trends (last accessed date: April 17, 2026).

CIHI. Wait times for priority procedures in Canada, June 2025b. Available at https://www.cihi.ca/en/wait-times-for-priority-procedures-in-canada-2025 (last accessed date: May 17, 2026).

T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to algorithms*. MIT Press, Cambridge, MA, 3rd edition, 2009.

R. Day, R. Garfinkel, and S. Thompson. Integrated block sharing: A win–win strategy for hospitals and surgeons. *Manufacturing & Service Operations Management*, 14(4):567–583, 2012.

B. T. Denton, A. J. Miller, H. J. Balasubramanian, and T. R. Huschka. Optimal allocation of surgery blocks to operating rooms under uncertainty. *Operations Research*, 58(4-part-1):802–816, 2010.

F. Dexter, R. H. Epstein, E. Marcon, and R. de Matta. Strategies to reduce delays in admission into a postanesthesia care unit from operating rooms. *Journal of PeriAnesthesia Nursing*, 20(2):92–102, April 2005.

Anita Duraikannan. Single file: An equitable approach to surgical wait times. Women's College Hospital Annual Report 2022, 2022. URL https://annualreport2022.womenscollegehospital.ca/innovating-toward-equity/.

C. Guo, M. Bodur, D. M. Aleman, and D. R. Urbach. Logic-based Benders decomposition and binary decision diagram based approaches for stochastic distributed operating room scheduling. *INFORMS Journal on Computing*, 33(4):1551–1569, 2021.

Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2026. Available at https://www.gurobi.com.

H. Hashemi Doulabi, G. Pesant, and L.-M. Rousseau. Vehicle routing problems with synchronized visits and stochastic travel and service times: Applications in healthcare. *Transportation Science*, 54(4):1053–1072, 2020.

M. Heyba, F. Qasem, A. S. Ibrahim, T. Habib, H. Akl, and S. M. Al-Matouq. Improving postanesthesia care unit (PACU) delays: A quality improvement project. *Journal of PeriAnesthesia Nursing*, 39(5):716–721, October 2024.

J. Hooker. Planning and scheduling by logic-based Benders decomposition. *Operations Research*, 55(3):588–602, 2007.

J. Hooker and G. Ottosson. Logic-based Benders decomposition. *Mathematical Programming*, 96(1):33–60, 2003.

IBM. *IBM ILOG CP Optimizer User's Manual*. IBM, 2022. Version 22.1.1.0. Available at https://www.ibm.com/docs/en/icos/22.1.1?topic=manual-welcome-cp-optimizer-users.

G. Johnson and T. Calixte. Leaning on postanesthesia care unit hold times. *Journal of PeriAnesthesia Nursing*, 38(4):e4, 2023.

E. Kayıcs, T. T. Khaniyev, J. Suermondt, and K. Sylvester. A robust estimation model for surgery durations with temporal, operational, and surgery team effects. *Health Care Management Science*, 18(3):222–233, September 2015.

P. Laborie, J. Rogerie, P. Shaw, and P. Vilím. IBM ILOG CP optimizer for scheduling. 20+ years of scheduling with constraints at IBM/ILOG. *Constraints*, 23(2):210–250, Apr 2018.

B. Maenhout, P. Šůcha, V. Valdmanova, O. Tkadlec, and J. T. Rozlivkova. A bilevel approach to integrated surgeon scheduling and surgery planning solved via branch-and-price. *Omega*, 138:103424, 2026. ISSN 0305-0483.

A. S. Manne. On the job-shop scheduling problem. *Operations Research*, 8(2):219–223, 1960.

Clair E. Miller, Albert W. Tucker, and Richard A. Zemlin. Integer programming formulation of traveling salesman problems. *Journal of the ACM (JACM)*, 7(4):326–329, 1960. doi: 10.1145/321043.321046.

R. Mínguez, W. van Ackooij, and R. García-Bertrand. Constraint generation for risk averse two-stage stochastic programs. *European Journal of Operational Research*, 288(1):194–206, 2021. ISSN 0377-2217.

B. Naderi, V. Roshanaei, M. A. Begen, D. M. Aleman, and D. R. Urbach. Increased surgical capacity without additional resources: Generalized operating room planning and scheduling. *Production and Operations Management*, 30(8):2608–2635, 2021.

B. Naderi, R. Ruiz, and V. Roshanaei. Mixed-integer programming vs. constraint programming for shop scheduling problems: New results and outlook. *INFORMS Journal on Computing*, 35(4):817–843, 2023.

R. B. Nelsen. *An introduction to copulas*. Springer Series in Statistics. Springer, New York, NY, 2nd edition, 2006.

D. J. Pavlin, S. E. Rapp, N. L. Polissar, J. A. Malmgren, M. Koerschgen, and H. Keyes. Factors affecting discharge time in adult outpatients. *Anesthesia & Analgesia*, 87(4):816–826, 1998. doi: 10.1097/00000539-199810000-00014.

M. L. Pinedo. *Scheduling: Theory, algorithms, and systems*. Springer, New York, 4th edition, 2012.

R. Rahmaniani, T. G. Crainic, M. Gendreau, and W. Rei. The Benders decomposition algorithm: A literature review. *European Journal of Operational Research*, 259(3):801–817, 2017. ISSN 0377-2217.

V. Rana, C. A. Yano, and A. M. Newman. Block scheduling in a multi-hospital setting. *Operations Research, Data Analytics and Logistics*, 45:200477, 2025. ISSN 3050-7847.

S. Rath, K. Rajaram, and A. Mahajan. Integrated anesthesiologist and room scheduling for surgeries: Methodology and application. *Operations Research*, 65(6):1460–1478, 2017.

A. Riise, C. Mannino, and L. Lamorgese. Recursive logic-based Benders' decomposition for multi-mode outpatient scheduling. *European Journal of Operational Research*, 255(3):719–728, 2016.

R. T. Rockafellar. Coherent approaches to risk in optimization under uncertainty. In *INFORMS TutORials in Operations Research 2007*, pages 38–61. INFORMS, 2007.

R. T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2:21–42, 2000.

R. Tyrrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471, 2002. ISSN 0378-4266. doi: 10.1016/S0378-4266(02)00271-6.

V. Roshanaei and B. Naderi. Solving integrated operating room planning and scheduling: Logic-based Benders decomposition versus branch-price-and-cut. *European Journal of Operational Research*, 293(1): 65–78, 2021.

V. Roshanaei, C. Luong, D. Aleman, and D. Urbach. Propagating logic-based Benders' decomposition approaches for distributed operating room scheduling. *European Journal of Operational Research*, 257(2): 439–455, 2017a.

V. Roshanaei, C. Luong, D. M. Aleman, and D. R. Urbach. Collaborative operating room planning and scheduling. *INFORMS Journal on Computing*, 29(3):558–580, 2017b.

V. Roshanaei, K. E. C. Booth, D. M. Aleman, D. R. Urbach, and J. C. Beck. Branch-and-check methods for multi-level operating room planning and scheduling. *International Journal of Production Economics*, 220:107433, 2020a.

V. Roshanaei, C. Luong, D. M. Aleman, and D. R. Urbach. Reformulation, linearization, and decomposition techniques for balanced distributed operating room scheduling. *Omega*, 93:102043, 2020b.

P. Santibáñez, M. Begen, and D. Atkins. Surgical block scheduling in a system of hospitals: An application to resource and wait list management in a British Columbia health authority. *Health Care Management Science*, 10(3):269–282, September 2007.

Pardis Seyedi, Dionne Aleman, Nancy Baxter, Chaim Bell, Merve Bodur, Andrew Calzavara, Robert Campbell, Michael Carter, Pieter de Jager, Scott Emerson, Anna Gagliardi, Jonathan Irish, Danielle Martin, Samantha Lee, Marcy Saxe-Braithwaite, Julie Takata, Suting Yang, Claudia Zanchetta, and David Urbach. Referral patterns for common surgical procedures in Ontario: A cross-sectional population-level study. *Canadian Journal of Surgery*, 67(6):E397–E405, 2024. doi: 10.1503/cjs.002324.

A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on stochastic programming: Modeling and theory*. MPS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics, 2009.

Statistics Canada. Canada's population estimates: Age and gender, July 1, 2025, September 2025. Available at https://www150.statcan.gc.ca/n1/daily-quotidien/250924/dq250924a-eng.htm (last accessed date: May 17, 2026).

D. P. Strum, J. H. May, and L. G. Vargas. Modeling the uncertainty of surgical procedure times: Comparison of log-normal and normal models. *Anesthesiology*, 92(4):1160–1167, April 2000.

O. Tange. GNU Parallel 20230522 ('Charles'), May 2023.

S. K. Taylor, J. C. Andrzejowski, M. D. Wiles, S. Bland, G. L. Jones, and S. C. Radley. A prospective observational study of the impact of an electronic questionnaire (ePAQ-PO) on the duration of nurse-led pre-operative assessment and patient satisfaction. *PLoS One*, 13(10):e0205439, 2018.

M. Y. Tsang, K. S. Shehadeh, F. E. Curtis, B. R. Hochman, and T. E. Brentjens. Stochastic optimization approaches for an operating room and anesthesiologist scheduling problem. *Operations Research*, 73(3): 1430–1458, 2024.

UHN. Report to our community 2022–2023. Organizational annual report, 2023. Available at https://www.uhn.ca/corporate/AboutUHN/Documents/report-to-our-community-2022-23.pdf (last accessed date: July 23, 2026).

UHN. Cataract surgery. Patient information webpage, 2024a. Available at https://www.uhn.ca/PatientsFamilies/Health_Information/Health_Topics/Pages/brochure.aspx?itemid=339 (last accessed date: July 30, 2026).

UHN. Day surgery at Toronto General Hospital: For patients preparing for day surgery. Patient education brochure (PDF), 2024b. Form D-5080, revised 02/2024. Available at https://www.uhn.ca/PatientsFamilies/Health_Information/Health_Topics/Documents/Day_Surgery_at_Toronto_General_Hospital.pdf (last accessed date: February 23, 2026).

UHN. Deep brain stimulation surgery at Krembil Brain Institute. Patient information webpage, 2026a. Available at https://www.uhn.ca/Krembil/Treatments_Procedures/Deep_Brain_Stimulation_Surgery (last accessed date: February 23, 2026).

UHN. Surgery guides, 2026b. Available at https://www.uhn.ca/Surgery/My_Surgery/Surgery_Guides (last accessed date: May 15, 2026).

D. R. Urbach, K. M. Devon, and B. N. Pakes. Single-entry models as an ethical approach to decreasing surgical wait times. *CMAJ*, 197(42):E1458–E1459, December 2025.

J. Wang, J. Cabrera, K. L. Tsui, H. Guo, M. Bakker, and J. B. Kostis. Clinical and nonclinical effects on operative duration: Evidence from a database on thoracic surgery. *Journal of Healthcare Engineering*, 2020:3582796, February 2020.

S. Wang, V. Roshanaei, D. Aleman, and D. Urbach. A discrete event simulation evaluation of distributed operating room scheduling. *IIE Transactions on Healthcare Systems Engineering*, 6(4):236–245, 2016.

C. Weissman, J. Scemama, and Y. G. Weiss. The ratio of PACU length-of-stay to surgical duration: Practical observations. *Acta Anaesthesiologica Scandinavica*, 63(9):1143–1151, 2019.

P. F. White, S. Rawal, J. Nguyen, and A. Watkins. PACU fast-tracking: An alternative to “bypassing” the PACU for facilitating the recovery process after ambulatory surgery. *Journal of PeriAnesthesia Nursing*, 18(4):247–253, 2003.

X. Xu, Y. Yin, D. Wang, T. C. E. Cheng, and X. Sima. Distributionally robust multi-period operating room scheduling with multiple surgical disciplines under uncertain surgery durations. *Omega*, 138:103420, 2026. ISSN 0305-0483.

e-Companion

This e-companion provides supplementary formulations, technical derivations, proofs, computational details, and additional numerical material deferred from the main paper titled "Beyond Isolated Operating Rooms: Risk-Aware Surgical Episode Scheduling in Single-Entry Networks".

Appendix A: Supporting Formulations and Warm-start Decoding

This appendix collects the formulations deferred from the main text. Appendices A.1 and A.2 present the deterministic MIP and CP formulations underlying the models in the main text. Appendix A.2 additionally describes how a pulse-based CP schedule is decoded into MIP-compatible binary decisions for warm-starting SP and AUX-SP. Appendices A.3 and A.4 state the MIP and CP versions of the second-level scheduling SP of the two-level decomposition, and Appendix A.5 gives the workload-balancing model used to break ties among master-optimal assignments.

A.1. Deterministic MIP Formulation

The deterministic surgical episode scheduling problem in single-entry networks can be formulated as the following MIP, with notation summarized in Table A.1.

$$\min \quad C_{\max} \tag{A.1a}$$

$$\text{s.t.} \quad \sum_{h\in\mathcal{H}} y_{ph} = 1 \qquad \forall p\in\mathcal{P}, \tag{A.1b}$$

$$\sum_{r\in\mathcal{R}_{hs}} z_{phsr} = y_{ph} \qquad \forall p\in\mathcal{P},\ h\in\mathcal{H},\ s\in\mathcal{S}, \tag{A.1c}$$

$$c_{ph,1} \geq D_{ph,1} y_{ph} \qquad \forall p\in\mathcal{P},\ h\in\mathcal{H}, \tag{A.1d}$$

$$c_{phs} \geq c_{ph,s-1} + D_{phs} y_{ph} \qquad \forall p\in\mathcal{P},\ h\in\mathcal{H},\ s\in\mathcal{S}\setminus\{1\}, \tag{A.1e}$$

$$c_{phs} \geq c_{p'hs} + D_{phs} - M\left(3 - x_{pp'hs} - z_{phsr} - z_{p'hsr}\right) \qquad \forall h\in\mathcal{H},\ s\in\mathcal{S},\ r\in\mathcal{R}_{hs},\ p,p'\in\mathcal{P}:\ p>p', \tag{A.1f}$$

$$c_{p'hs} \geq c_{phs} + D_{p'hs} - M\left(2 + x_{pp'hs} - z_{phsr} - z_{p'hsr}\right) \qquad \forall h\in\mathcal{H},\ s\in\mathcal{S},\ r\in\mathcal{R}_{hs},\ p,p'\in\mathcal{P}:\ p>p', \tag{A.1g}$$

$$C_{\max} \geq c_{ph,|\mathcal{S}|} \qquad \forall p\in\mathcal{P},\ h\in\mathcal{H}, \tag{A.1h}$$

$$c_{phs} \geq 0 \qquad \forall p\in\mathcal{P},\ h\in\mathcal{H},\ s\in\mathcal{S}, \tag{A.1i}$$

$$y_{ph} \in \{0,1\} \qquad \forall p\in\mathcal{P},\ h\in\mathcal{H}, \tag{A.1j}$$

$$z_{phsr} \in \{0,1\} \qquad \forall p\in\mathcal{P},\ h\in\mathcal{H},\ s\in\mathcal{S},\ r\in\mathcal{R}_{hs}, \tag{A.1k}$$

$$x_{pp'hs} \in \{0,1\} \qquad \forall p,p'\in\mathcal{P}:\ p>p',\ h\in\mathcal{H},\ s\in\mathcal{S}. \tag{A.1l}$$

Table A.1 Notation for the deterministic MIP formulation.

Sets	
$\mathcal{P}$	Set of surgical episodes, indexed by p.
$\mathcal{H}$	Set of hospitals, indexed by h.
$\mathcal{S}$	Ordered set of care stages, indexed by s.
$\mathcal{R}_{hs}$	Resources within stage s in hospital h, indexed by r.
Parameters	
D_{phs}	Deterministic duration of surgical episode p at (h,s).
M	Sufficiently large constants.
Decision Variables	
$y_{ph} \in \{0,1\}$	1 if surgical episode p is assigned to hospital h, 0 otherwise.
$z_{phsr} \in \{0,1\}$	1 if surgical episode p uses resource r at stage s in hospital h, 0 otherwise.
$x_{pp'hs} \in \{0,1\}$	1 if surgical episode p is processed after surgical episode p' at stage s in hospital h.
$c_{phs} \geq 0$	Completion time of surgical episode p at stage s in hospital h.
$C_{\max} \geq 0$	Deterministic makespan.

A.2. Deterministic CP Formulation

The CP counterpart of MIP (A.1) uses a cumulative-capacity (`Pulse`) representation in which resources do not need to be modeled individually. Instead, the `Pulse` function aggregates identical resources at each (h,s) pair. Time quantities are assumed to be integer-valued after scaling if needed. With notation summarized in Table A.2, the CP formulation is given below.

Table A.2 Notation for the deterministic CP formulation.

CP Decision Variables	
O_{ps}	Non-optional interval for surgical episode p at stage s.
T_{phs}	Optional hospital-stage interval for surgical episode p at (h,s) in the pulse model.
CP Functions	
$\texttt{IntervalVar}(\cdot)$	Creates an interval decision variable, either optional or non-optional.
$\texttt{PresenceOf}(I)$	Returns 1 if optional interval I is present, and 0 otherwise.
$\texttt{Alternative}(O,\mathcal{I})$	Selects exactly one interval in $\mathcal{I}$ to realize O and synchronizes their start and end times.
$\texttt{EndBeforeStart}(A,B)$	Enforces that interval A ends before interval B starts.
$\texttt{Pulse}(I,1)$	Adds one unit of cumulative resource usage over interval I.
$\texttt{EndOf}(I)$	Returns the end time of interval I.

$$\min \quad C_{\max} \tag{A.2a}$$

$$\text{s.t.} \quad \sum_{h\in\mathcal{H}} y_{ph} = 1 \qquad \forall p\in\mathcal{P}, \tag{A.2b}$$

$$O_{ps} = \texttt{IntervalVar}(\text{non-optional}) \qquad \forall p\in\mathcal{P},\ s\in\mathcal{S}, \tag{A.2c}$$

$$T_{phs} = \texttt{IntervalVar}(\text{optional},\ \text{size} = D_{phs}) \qquad \forall p\in\mathcal{P},\ h\in\mathcal{H},\ s\in\mathcal{S}, \tag{A.2d}$$

$$\texttt{PresenceOf}(T_{phs}) = y_{ph} \qquad \forall p\in\mathcal{P},\ h\in\mathcal{H},\ s\in\mathcal{S}, \tag{A.2e}$$

$$\texttt{Alternative}(O_{ps},\ \{T_{phs}\}_{h\in\mathcal{H}}) \qquad \forall p\in\mathcal{P},\ s\in\mathcal{S}, \tag{A.2f}$$

$$\texttt{EndBeforeStart}(O_{ps},\ O_{p,s+1}) \qquad \forall p\in\mathcal{P},\ s=1,\ldots,|\mathcal{S}|-1, \tag{A.2g}$$

$$\sum_{p\in\mathcal{P}} \texttt{Pulse}(T_{phs}, 1) \le |\mathcal{R}_{hs}| \qquad \forall h \in \mathcal{H},\ s \in \mathcal{S}, \tag{A.2h}$$

$$C_{\max} \ge \max_{p\in\mathcal{P}} \big\{ \texttt{EndOf}(O_{p,|\mathcal{S}|}) \big\}, \tag{A.2i}$$

$$y_{ph} \in \{0,1\} \qquad \forall p \in \mathcal{P},\ h \in \mathcal{H}. \tag{A.2j}$$

A pulse-based CP schedule can be decoded into the binary variables $(\mathbf{y}, \mathbf{z}, \mathbf{x})$ required by the MIP representation. Because the pulse formulation enforces only aggregate stage capacity and does not distinguish between individual resources, the resource-selection and sequencing binaries cannot be read off directly. We can however reconstruct them from the realized intervals: hospital assignments are read from the CP solution, resource assignments are recovered by greedily labeling the realized intervals at each hospital-stage pair in order of start time, and the sequencing binaries are then set from the resulting order on each decoded resource. Feasibility for the MIP is evident from this construction, as the resource-capacity limit enforced by the `Pulse` constraint guarantees that at most $|\mathcal{R}_{hs}|$ resource labels are ever open at each hospital-stage pair, and the greedy labeling assigns each surgical episode to exactly one open label, so constraint (A.2h)'s aggregate limit is preserved as a per-resource assignment. The decoding is not unique, since resources within a hospital-stage pair are interchangeable in the pulse formulation and any relabeling of r yields an equally valid $(\mathbf{z}, \mathbf{x})$.

A.3. MIP Formulation of the SP

This appendix provides the MIP form of the SP (6) at a fixed hospital assignment $\mathbf{y}^\kappa$. Since the hospital is fixed, the index h is dropped from the scheduling variables, which are written relative to $h^\kappa(p)$. The MIP SP is the restriction of (3) to the assigned hospitals:

$$\mathrm{SP}^\star(\mathbf{y}^\kappa) = \min \quad \rho\big(\mathbf{C}^{\mathrm{SP}}_{\max}\big) \tag{A.3a}$$

$$\text{s.t.} \quad \sum_{r\in\mathcal{R}_{h^\kappa(p)s}} z_{psr} = 1 \qquad \forall p \in \mathcal{P},\ s \in \mathcal{S}, \tag{A.3b}$$

$$c_{p1}(\omega) \ge D_{p,h^\kappa(p),1}(\omega) \qquad \forall p \in \mathcal{P},\ \omega \in \Omega, \tag{A.3c}$$

$$c_{ps}(\omega) \ge c_{p,s-1}(\omega) + D_{p,h^\kappa(p),s}(\omega) \qquad \forall p \in \mathcal{P},\ s \in \mathcal{S}\setminus\{1\},\ \omega \in \Omega, \tag{A.3d}$$

$$\begin{aligned} c_{ps}(\omega) \ge\ & c_{p's}(\omega) + D_{p,h^\kappa(p),s}(\omega) && \forall p, p' \in \mathcal{P}:\ p > p',\ s \in \mathcal{S} \text{ with} \\ & - M_{pp',h^\kappa(p),s}(\omega)\big(3 - x_{pp's} - z_{psr} - z_{p'sr}\big) && h^\kappa(p) = h^\kappa(p'),\ r \in \mathcal{R}_{h^\kappa(p)s}, \\ & && \omega \in \Omega \end{aligned} \tag{A.3e}$$

$$\begin{aligned} c_{p's}(\omega) \ge\ & c_{ps}(\omega) + D_{p',h^\kappa(p'),s}(\omega) && \forall p, p' \in \mathcal{P}:\ p > p',\ s \in \mathcal{S} \text{ with} \\ & - M'_{pp',h^\kappa(p),s}(\omega)\big(2 + x_{pp's} - z_{psr} - z_{p'sr}\big) && h^\kappa(p) = h^\kappa(p'),\ r \in \mathcal{R}_{h^\kappa(p)s}, \\ & && \omega \in \Omega \end{aligned} \tag{A.3f}$$

$$C^{\mathrm{SP}}_{\max}(\omega) \ge c_{p,|\mathcal{S}|}(\omega) \qquad \forall p \in \mathcal{P},\ \omega \in \Omega, \tag{A.3g}$$

$$c_{ps}(\omega) \ge 0 \qquad \forall p \in \mathcal{P},\ s \in \mathcal{S},\ \omega \in \Omega, \tag{A.3h}$$

$$z_{psr} \in \{0,1\} \qquad \forall p \in \mathcal{P},\ s \in \mathcal{S},\ r \in \mathcal{R}_{h^\kappa(p)s}, \tag{A.3i}$$

$$x_{pp's} \in \{0,1\} \qquad \forall p,p' \in \mathcal{P}:\ p > p',\ h^\kappa(p) = h^\kappa(p'),\ s \in \mathcal{S}. \tag{A.3j}$$

Constraint (A.3b) selects exactly one resource per stage within each surgical episode's assigned hospital. Constraints (A.3c)-(A.3d) enforce the pathway precedence along the care stages of the assigned hospital, and (A.3e)-(A.3f) are the disjunctive non-overlap constraints for every pair of surgical episodes that may share a resource at a stage of the same hospital. Their Big-M coefficients are derived in Appendix B, where fixing $\mathbf{y}^\kappa$ allows the assignment-independent hospital horizon to be replaced by the smaller assignment-specific one. Constraint (A.3g) defines the scenario-wise network makespan, and (A.3h)-(A.3j) define the support of the decision variables. Resource selection and sequencing range only over $\mathcal{R}_{h^\kappa(p)s}$, and z_{psr} and $x_{pp's}$ remain scenario-independent so non-anticipativity is preserved.

A.4. CP Formulation of the SP

This appendix gives the CP form of the SP (6), which restricts (4) to the assigned hospitals and describes the same feasible set and optimal value $\mathrm{SP}^\star(\mathbf{y}^\kappa)$ as (A.3). Optional resource intervals are defined only for $r \in \mathcal{R}_{h^\kappa(p)s}$, with duration $D_{p,h^\kappa(p),s}(\omega)$, and non-anticipativity is propagated through a reference scenario $\bar{\omega}$:

$$\min \quad \rho\big(\mathbf{C}^{\mathrm{SP}}_{\max}\big) \tag{A.4a}$$

$$\text{s.t.} \quad O_{ps}(\omega) = \texttt{IntervalVar}(\text{non-optional}) \qquad \forall p \in \mathcal{P},\ s \in \mathcal{S},\ \omega \in \Omega \tag{A.4b}$$

$$T_{psr}(\omega) = \texttt{IntervalVar}\Big(\text{optional},\ \text{size} = D_{p,h^\kappa(p),s}(\omega)\Big) \quad \forall p \in \mathcal{P},\ s \in \mathcal{S},\ r \in \mathcal{R}_{h^\kappa(p)s},\ \omega \in \Omega \tag{A.4c}$$

$$\texttt{Alternative}\Big(O_{ps}(\omega),\ \{T_{psr}(\omega)\}_{r \in \mathcal{R}_{h^\kappa(p)s}}\Big) \qquad \forall p \in \mathcal{P},\ s \in \mathcal{S},\ \omega \in \Omega \tag{A.4d}$$

$$\texttt{EndBeforeStart}\big(O_{ps}(\omega),\, O_{p,s+1}(\omega)\big) \qquad \forall p \in \mathcal{P},\ s = 1,\dots,|\mathcal{S}|-1,\ \omega \in \Omega \tag{A.4e}$$

$$Q_{hsr}(\omega) = \texttt{SequenceVar}\big(\{\, T_{psr}(\omega) : p \in \mathcal{P},\ h^\kappa(p) = h \,\}\big) \quad \forall h \in \mathcal{H},\ s \in \mathcal{S},\ r \in \mathcal{R}_{hs},\ \omega \in \Omega \tag{A.4f}$$

$$\texttt{NoOverlap}\big(Q_{hsr}(\omega)\big) \qquad \forall h \in \mathcal{H},\ s \in \mathcal{S},\ r \in \mathcal{R}_{hs},\ \omega \in \Omega \tag{A.4g}$$

$$\texttt{PresenceOf}\big(T_{psr}(\omega)\big) = \texttt{PresenceOf}\big(T_{psr}(\bar{\omega})\big) \qquad \forall p \in \mathcal{P},\ s \in \mathcal{S},\ r \in \mathcal{R}_{h^\kappa(p)s},\ \omega \in \Omega \setminus \{\bar{\omega}\} \tag{A.4h}$$

$$\texttt{SameSequence}\big(Q_{hsr}(\omega),\, Q_{hsr}(\bar{\omega})\big) \qquad \forall h \in \mathcal{H},\ s \in \mathcal{S},\ r \in \mathcal{R}_{hs},\ \omega \in \Omega \setminus \{\bar{\omega}\} \tag{A.4i}$$

$$C^{\mathrm{SP}}_{\max}(\omega) \geq \max_{p \in \mathcal{P}} \big\{ \texttt{EndOf}\big(O_{p,|\mathcal{S}|}(\omega)\big) \big\} \qquad \forall \omega \in \Omega. \tag{A.4j}$$

The `Alternative` constraint selects exactly one resource in $\mathcal{R}_{h^\kappa(p)s}$ per stage, and `PresenceOf` and `SameSequence` link scenario-specific resource selections and sequences to those in $\bar{\omega}$. Stage durations and completion times remain scenario-dependent, while assignment and sequencing decisions are common across Ω.

A.5. Workload-Balancing Tie-Breaking Model

When (5) admits multiple optimal assignments at outer iteration κ, the tie-breaking step of Section 4.1 re-solves the MP with its objective fixed at θ^κ and minimizes the total workload imbalance across hospitals. For each hospital $h \in \mathcal{H}$, define $N_h = \sum_{p\in\mathcal{P}} y_{ph}$ as the number of surgical episodes assigned to hospital, and let $\bar{N} = \frac{|\mathcal{P}|}{|\mathcal{H}|}$ be the average number of surgical episodes per hospital. We introduce imbalance variables $b_h \geq 0$ and solve

$$\min \quad \sum_{h\in\mathcal{H}} b_h \tag{A.5a}$$

$$\text{s.t.} \quad (5b) - (5g)$$

$$\theta = \theta^\kappa \tag{A.5b}$$

$$N_h = \sum_{p\in\mathcal{P}} y_{ph} \qquad \forall h \in \mathcal{H},$$

$$b_h \geq N_h - \bar{N} \qquad \forall h \in \mathcal{H}, \tag{A.5c}$$

$$b_h \geq \bar{N} - N_h \qquad \forall h \in \mathcal{H}, \tag{A.5d}$$

$$b_h \geq 0 \qquad \forall h \in \mathcal{H}. \tag{A.5e}$$

Constraints (A.5c)-(A.5d) make b_h capture the absolute deviation between the number of episodes assigned to hospital h and the network average, i.e., $b_h \geq |N_h - \bar{N}|$ for all $h \in \mathcal{H}$.

Appendix B: Big-M Values for the Disjunctive Constraints

This appendix derives sufficient Big-M coefficients for the disjunctive non-overlap constraints (3f)-(3g). Fix $h \in \mathcal{H}$, $s \in \mathcal{S}$, $r \in \mathcal{R}_{hs}$, $p, p' \in \mathcal{P}$ with $p > p'$, and $\omega \in \Omega$. The two disjunctive non-overlap constraints are

$$c_{phs}(\omega) \geq c_{p'hs}(\omega) + D_{phs}(\omega) - M_{pp'hs}(\omega)\big(3 - x_{pp'hs} - z_{phsr} - z_{p'hsr}\big), \tag{B.6}$$

$$c_{p'hs}(\omega) \geq c_{phs}(\omega) + D_{p'hs}(\omega) - M'_{pp'hs}(\omega)\big(2 + x_{pp'hs} - z_{phsr} - z_{p'hsr}\big). \tag{B.7}$$

The activation of these inequalities depends jointly on the precedence variable and the resource-assignment variables. If both episodes use resource r at (h,s), that is, $z_{phsr} = z_{p'hsr} = 1$, then the Big-M multipliers in (B.6) and (B.7) reduce to $1 - x_{pp'hs}$ and $x_{pp'hs}$, respectively. Hence, $x_{pp'hs} = 1$ activates (B.6) and enforces that p' precedes p, whereas $x_{pp'hs} = 0$ activates (B.7) and enforces that p precedes p'. If the episodes do not both use resource r, both multipliers are at least one, regardless of the value of $x_{pp'hs}$, and both inequalities must be deactivated by their Big-M terms.

The original formulation does not explicitly upper-bound completion times, so a feasible timing vector may contain arbitrary delays. We therefore derive the coefficients using an earliest feasible timing realization. For each $p \in \mathcal{P}$, $h \in \mathcal{H}$, $s \in \mathcal{S}$, and $\omega \in \Omega$, define the cumulative pathway duration $L_{phs}(\omega) := \sum_{t=1}^{s} D_{pht}(\omega)$, the assignment-dependent hospital horizon $H_{hs}(\mathbf{y},\omega) := \sum_{k\in\mathcal{P}} y_{kh} L_{khs}(\omega)$, and its assignment-independent counterpart $\overline{H}_{hs}(\omega) := \sum_{k\in\mathcal{P}} L_{khs}(\omega) = \sum_{k\in\mathcal{P}} \sum_{t=1}^{s} D_{kht}(\omega)$.

Lemma EC.1. *Fix a scenario $\omega \in \Omega$ and hospital-assignment, resource-assignment, and sequencing decisions satisfying the corresponding discrete constraints. Suppose that the active precedence graph resulting from these decisions is acyclic, where the graph contains the pathway arcs and the selected resource-order arcs. Then the graph admits an earliest completion-time vector satisfying*

$$\widehat{c}_{phs}(\omega) \geq L_{phs}(\omega) y_{ph}, \tag{B.8}$$

$$\widehat{c}_{phs}(\omega) \leq H_{hs}(\mathbf{y},\omega), \tag{B.9}$$

$$0 \leq \widehat{c}_{phs}(\omega) \leq \overline{H}_{hs}(\omega) y_{ph}, \tag{B.10}$$

for every $p \in \mathcal{P}$, $h \in \mathcal{H}$, and $s \in \mathcal{S}$.

Proof. If $y_{ph} = 1$, recursively applying the pathway-precedence relations gives

$$\widehat{c}_{phs}(\omega) \geq \sum_{t=1}^{s} D_{pht}(\omega) = L_{phs}(\omega).$$

If $y_{ph} = 0$, the bound reduces to nonnegativity. Because the active precedence graph is acyclic, the earliest completion time of each assigned operation equals the length of a longest directed path ending at its node. A path ending at an operation in hospital h and stage s can contain only operations assigned to h and belonging to stages $1,\ldots,s$: resource-order arcs connect operations within the same stage, while pathway arcs move from stage t to stage $t+1$. Moreover, no operation can appear more than once on an acyclic path. Consequently,

$$\widehat{c}_{phs}(\omega) \leq \sum_{k\in\mathcal{P}} y_{kh} \sum_{t=1}^{s} D_{kht}(\omega) = H_{hs}(\mathbf{y},\omega)$$

for every assigned operation. Finally, if $y_{ph} = 0$, episode p has no operation at hospital h, and we set $\widehat{c}_{phs}(\omega) = 0$. Combining these bounds with $H_{hs}(\mathbf{y},\omega) \leq \overline{H}_{hs}(\omega)$ establishes (B.10). $\square$

We next derive coefficients that make every inactive disjunctive inequality valid for the earliest completion-time vector constructed in Lemma EC.1. Consider first (B.6). Its multiplier is $3 - x_{pp'hs} - z_{phsr} - z_{p'hsr}$. It is zero exactly when $x_{pp'hs} = z_{phsr} = z_{p'hsr} = 1$, in which case the inequality enforces that p' precedes p on resource r. In every other case, the multiplier is at least one. Using (B.8) and (B.9), we obtain

$$\begin{aligned}
&\widehat{c}_{p'hs}(\omega) + D_{phs}(\omega) - \widehat{c}_{phs}(\omega) \\
&\quad \leq H_{hs}(\mathbf{y},\omega) + D_{phs}(\omega) - L_{phs}(\omega) y_{ph}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{k\in\mathcal{P}\setminus\{p\}} y_{kh} L_{khs}(\omega) + D_{phs}(\omega) \\
&\leq \sum_{k\in\mathcal{P}\setminus\{p\}} L_{khs}(\omega) + D_{phs}(\omega). \qquad \text{(B.11)}
\end{aligned}
$$

It is therefore sufficient to set

$$
M_{pp'hs}(\omega) = \sum_{k\in\mathcal{P}\setminus\{p\}} L_{khs}(\omega) + D_{phs}(\omega) = \sum_{k\in\mathcal{P}\setminus\{p\}} \sum_{t=1}^{s} D_{kht}(\omega) + D_{phs}(\omega). \qquad \text{(B.12)}
$$

Whenever (B.6) is inactive, its multiplier is at least one, and (B.11) guarantees that the inequality is satisfied by the earliest completion-time vector.

The multiplier in (B.7) is $2 + x_{pp'hs} - z_{phsr} - z_{p'hsr}$. It is zero exactly when $x_{pp'hs} = 0, z_{phsr} = z_{p'hsr} = 1$, in which case the inequality enforces that p precedes p' on resource r. In every other case, the multiplier is at least one. Interchanging the roles of p and p' in the preceding derivation gives

$$
M'_{pp'hs}(\omega) = \sum_{k\in\mathcal{P}\setminus\{p'\}} \sum_{t=1}^{s} D_{kht}(\omega) + D_{p'hs}(\omega). \qquad \text{(B.13)}
$$

The pathway constraints and active resource-order constraints are satisfied by construction of the earliest completion-time vector, while (B.12)-(B.13) ensure that all inactive disjunctive constraints are also satisfied. Thus, every acyclic discrete schedule admits a feasible timing realization satisfying (B.8)-(B.10). Moreover, this earliest timing realization cannot have a larger scenario makespan than any other timing realization of the same discrete schedule. By monotonicity of ρ, restricting attention to these bounded completion times is optimality-preserving.

Accordingly, the bounds

$$
c_{phs}(\omega) \geq L_{phs}(\omega)\, y_{ph}, \qquad c_{phs}(\omega) \leq H_{hs}(\mathbf{y}, \omega), \qquad 0 \leq c_{phs}(\omega) \leq \overline{H}_{hs}(\omega)\, y_{ph}
$$

may be added to the monolithic formulation without changing its optimal value. With these bounds included, the coefficients (B.12)-(B.13) make the inactive disjunctive inequalities redundant over the resulting optimality-equivalent formulation. The coefficients are sufficient but are not claimed to be minimal, because the separate upper and lower bounds used in their derivation are not necessarily simultaneously attained. Although they are written with the full pair index to match (B.6)-(B.7), $M_{pp'hs}(\omega)$ does not depend on p', and $M'_{pp'hs}(\omega)$ does not depend on p. They may therefore be stored as $M_{phs}(\omega)$ and $M'_{p'hs}(\omega)$, respectively.

It is worth noting that, once the hospital assignment $\mathbf{y}^{\kappa}$ is fixed, the coefficients can be tightened by excluding episodes assigned to other hospitals. Let

$$
\mathcal{P}_h^{\kappa} := \{k \in \mathcal{P} : y_{kh}^{\kappa} = 1\}, \quad H_{hs}^{\kappa}(\omega) := \sum_{k\in\mathcal{P}_h^{\kappa}} \sum_{t=1}^{s} D_{kht}(\omega).
$$

For $p, p' \in \mathcal{P}_h^\kappa$, sufficient coefficients for the fixed-assignment scheduling subproblem are

$$M_{pp'hs}^{\kappa}(\omega) = H_{hs}^{\kappa}(\omega) - \sum_{t=1}^{s-1} D_{pht}(\omega),$$
$$M_{pp'hs}^{\prime,\kappa}(\omega) = H_{hs}^{\kappa}(\omega) - \sum_{t=1}^{s-1} D_{p'ht}(\omega).$$

The pathway-duration subtraction is valid in both the monolithic and fixed-assignment formulations. Fixing the assignment tightens the coefficients by replacing the assignment-independent hospital horizon $\overline{H}_{hs}(\omega)$ with the smaller assignment-specific horizon $H_{hs}^{\kappa}(\omega)$.

Appendix C: Proofs of Theoretical Results

This appendix presents the proofs of the theoretical results stated in the main text.

THEOREM 1. *Let ρ be a coherent risk measure, and let $\mathrm{SP}^\star(\mathbf{y}^\kappa)$ be the optimal SP value at the master assignment $\mathbf{y}^\kappa$. Then:*

(i) Inequality (8) is valid for the integer epigraph of the LBBD MP (5).

(ii) Its right-hand side dominates that of the no-good cut (7) pointwise over the master assignment region.

(iii) If every SP is solved to optimality, the resulting outer LBBD algorithm terminates finitely with a globally optimal solution of RA-SEEN.

Proof. For any feasible binary assignment $\mathbf{y}$, define $\bar{\mathcal{P}}^\kappa(\mathbf{y}) := \{p \in \mathcal{P} : y_{p,h^\kappa(p)} = 0\}$, the set of patients removed from their incumbent hospitals at iteration κ. Then

$$\sum_{p\in\mathcal{P}} \gamma_p^\kappa \big(1 - y_{p,h^\kappa(p)}\big) = \sum_{p\in\bar{\mathcal{P}}^\kappa(\mathbf{y})} \gamma_p^\kappa.$$

Validity. Fix a feasible assignment $\mathbf{y}$, and let $\widehat{\mathbf{C}}_{\max}^{\mathrm{SP}}$ be the makespan vector of an optimal schedule under $\mathbf{y}$. Thus, $\mathrm{SP}^\star(\mathbf{y}) = \rho\big(\widehat{\mathbf{C}}_{\max}^{\mathrm{SP}}\big)$. Starting from this schedule under $\mathbf{y}$, we construct a feasible candidate schedule under the incumbent assignment $\mathbf{y}^\kappa$. First, delete the episodes in $\bar{\mathcal{P}}^\kappa(\mathbf{y})$ from the schedule under $\mathbf{y}$, while retaining the resource choices, sequences, and timings of all other episodes. These retained episodes are assigned to the same hospitals under $\mathbf{y}$ and $\mathbf{y}^\kappa$, and deleting operations preserves feasibility. Hence, in scenario ω, all retained operations finish no later than $\widehat{C}_{\max}^{\mathrm{SP}}(\omega)$. Next, reinsert the episodes in $\bar{\mathcal{P}}^\kappa(\mathbf{y})$ at their incumbent hospitals. Fix an arbitrary episode order and a feasible resource at each stage, using the same choices in every scenario. Starting after the retained schedule has finished, process the reinserted episodes globally, one at a time, through all their stages. Because all retained operations have finished and at most one reinserted episode is processed at a time, this creates no resource conflicts and preserves non-anticipativity. The resulting schedule for $\mathbf{y}^\kappa$ has scenario makespan at most $\widehat{C}^{\mathrm{SP}}_{\max}(\omega) + \sum_{p\in\bar{\mathcal{P}}^\kappa(\mathbf{y})} \mathrm{path}^\kappa_p(\omega)$. Let $\mathbf{path}^\kappa_p := \big(\mathrm{path}^\kappa_p(\omega)\big)_{\omega\in\Omega}$. By optimality of $\mathrm{SP}^\star(\mathbf{y}^\kappa)$, monotonicity of ρ, and subadditivity of coherent risk measures,

$$\begin{aligned}\mathrm{SP}^\star(\mathbf{y}^\kappa) &\leq \rho\left(\widehat{\mathbf{C}}^{\mathrm{SP}}_{\max} + \sum_{p\in\bar{\mathcal{P}}^\kappa(\mathbf{y})} \mathbf{path}^\kappa_p\right) \xRightarrow{\text{subadditivity}} \\ &\leq \rho\big(\widehat{\mathbf{C}}^{\mathrm{SP}}_{\max}\big) + \sum_{p\in\bar{\mathcal{P}}^\kappa(\mathbf{y})} \rho\big(\mathbf{path}^\kappa_p\big) \\ &= \mathrm{SP}^\star(\mathbf{y}) + \sum_{p\in\bar{\mathcal{P}}^\kappa(\mathbf{y})} \gamma^\kappa_p.\end{aligned}$$

Consequently,

$$\mathrm{SP}^\star(\mathbf{y}) \geq \mathrm{SP}^\star(\mathbf{y}^\kappa) - \sum_{p\in\mathcal{P}} \gamma^\kappa_p\big(1 - y_{p,h^\kappa(p)}\big).$$

Thus every point in the integer epigraph $\theta \geq \mathrm{SP}^\star(\mathbf{y})$ satisfies (8).

Dominance. Let $\mathbf{C}^{\mathrm{SP},\kappa}_{\max}$ denote the scenario-wise makespan vector of an optimal SP schedule at $\mathbf{y}^\kappa$. For every $p \in \mathcal{P}$, the incumbent makespan vector dominates the pathway-duration vector componentwise. Monotonicity of ρ therefore gives

$$\mathbf{path}^\kappa_p \leq \mathbf{C}^{\mathrm{SP},\kappa}_{\max} \quad\Longrightarrow\quad \gamma^\kappa_p = \rho\big(\mathbf{path}^\kappa_p\big) \leq \rho\big(\mathbf{C}^{\mathrm{SP},\kappa}_{\max}\big) = \mathrm{SP}^\star(\mathbf{y}^\kappa).$$

Since $1 - y_{p,h^\kappa(p)} \geq 0$, it follows that

$$\sum_{p\in\mathcal{P}} \gamma^\kappa_p\big(1 - y_{p,h^\kappa(p)}\big) \leq \mathrm{SP}^\star(\mathbf{y}^\kappa)\sum_{p\in\mathcal{P}}\big(1 - y_{p,h^\kappa(p)}\big).$$

Hence,

$$\mathrm{SP}^\star(\mathbf{y}^\kappa) - \sum_{p\in\mathcal{P}} \gamma^\kappa_p\big(1 - y_{p,h^\kappa(p)}\big) \geq \mathrm{SP}^\star(\mathbf{y}^\kappa)\left(1 - \sum_{p\in\mathcal{P}}\big(1 - y_{p,h^\kappa(p)}\big)\right).$$

The right-hand side of the path-deletion cut therefore dominates that of the no-good cut pointwise. This comparison also holds for fractional $\mathbf{y} \in [0,1]^{|\mathcal{P}|\times|\mathcal{H}|}$.

Finite convergence. At the assignment that generates the cut, all deletion terms are zero, so

$$\theta \geq \mathrm{SP}^\star(\mathbf{y}^\kappa).$$

Thus, after $\mathbf{y}^\kappa$ has been evaluated, the master cannot return to it with a lower estimate than its exact recourse value. Suppose an evaluated assignment is selected again. The master lower bound is then at least its exact SP value, while the incumbent upper bound is no greater than that same value because the assignment has already been evaluated. Since the master lower bound cannot exceed the global optimum and the incumbent upper bound cannot be below it, the bounds coincide and the algorithm terminates. Therefore, every nonterminating iteration evaluates a previously unseen hospital assignment. Because the set of feasible binary hospital assignments is finite, the algorithm terminates after finitely many exact SP evaluations. Equal lower and upper bounds establish global optimality.

□

PROPOSITION 1. *Let ρ be a monotone risk measure and fix a feasible MP assignment $\mathbf{y}$. Then*

$$\mathrm{SP}^{\star}(\mathbf{y}) = \min_{\mathbf{z},\mathbf{x}} \rho\big(\mathbf{C}_{\max}\big) \;\geq\; \max_{h\in\mathcal{H}} \min_{\mathbf{z}_h,\mathbf{x}_h} \rho\big(\mathbf{C}_{\max,h}\big). \tag{9}$$

Proof. Let $\mathbf{T}_h(\mathbf{z}_h,\mathbf{x}_h)$ denote the scenario-wise makespan vector produced at hospital h by feasible local decisions $(\mathbf{z}_h,\mathbf{x}_h)$ under $\mathbf{y}$, and define

$$\mathrm{AUX}_h^{\star}(\mathbf{y}) := \min_{\mathbf{z}_h,\mathbf{x}_h} \rho\big(\mathbf{T}_h(\mathbf{z}_h,\mathbf{x}_h)\big).$$

Thus, $\mathbf{T}_h(\mathbf{z}_h,\mathbf{x}_h)$ is the vector denoted by $\mathbf{C}_{\max,h}$ in (9) with a given $(\mathbf{z}_h,\mathbf{x}_h)$. Now fix any feasible network schedule $(\mathbf{z},\mathbf{x})$, and denote its network makespan vector by $\widehat{\mathbf{C}}_{\max}$. Its restriction $(\mathbf{z}_h,\mathbf{x}_h)$ to hospital h is a feasible local schedule. Moreover, for every $h\in\mathcal{H}$ and every scenario ω,

$$\widehat{C}_{\max}(\omega) = \max_{h'\in\mathcal{H}} T_{h'}(\mathbf{z}_{h'},\mathbf{x}_{h'};\omega) \geq T_h(\mathbf{z}_h,\mathbf{x}_h;\omega).$$

Thus, $\widehat{\mathbf{C}}_{\max} \geq \mathbf{T}_h(\mathbf{z}_h,\mathbf{x}_h)$ componentwise. By monotonicity of ρ,

$$\rho\big(\widehat{\mathbf{C}}_{\max}\big) \geq \rho\big(\mathbf{T}_h(\mathbf{z}_h,\mathbf{x}_h)\big) \geq \mathrm{AUX}_h^{\star}(\mathbf{y}),$$

where the second inequality follows from the definition of $\mathrm{AUX}_h^{\star}(\mathbf{y})$ as the minimum over all feasible local schedules at hospital h. Since this holds for every $h\in\mathcal{H}$,

$$\rho\big(\widehat{\mathbf{C}}_{\max}\big) \geq \max_{h\in\mathcal{H}} \mathrm{AUX}_h^{\star}(\mathbf{y}) = \max_{h\in\mathcal{H}} \min_{\mathbf{z}_h,\mathbf{x}_h} \rho\big(\mathbf{C}_{\max,h}\big).$$

Because $(\mathbf{z},\mathbf{x})$ was an arbitrary feasible network schedule, minimizing the left-hand side over all such schedules proves (9). □

PROPOSITION 2. *The inequality*

$$\theta \geq \mathrm{AUX}^{\star}(\mathbf{y}^{\kappa}) \left(1 - \sum_{p\in\mathcal{P}_{h^{\star},\kappa}(\mathbf{y}^{\kappa})} (1 - y_{p,h^{\star},\kappa})\right) \tag{10}$$

is valid for the integer master epigraph (5), and dominates the auxiliary no-good inequality

$$\theta \geq \mathrm{AUX}^{\star}(\mathbf{y}^{\kappa}) \left(1 - \sum_{p\in\mathcal{P}} (1 - y_{p,h^{\kappa}(p)})\right). \tag{11}$$

Proof. Define

$$\Delta_{\mathrm{AUX}}^{\kappa}(\mathbf{y}) := \sum_{p\in\mathcal{P}_{h^{\star},\kappa}(\mathbf{y}^{\kappa})} (1 - y_{p,h^{\star},\kappa}).$$

Validity. Fix a feasible binary assignment $\mathbf{y}$. First suppose that $\Delta^{\kappa}_{\mathrm{AUX}}(\mathbf{y})=0$. Then every episode in $\mathcal{P}^{\kappa}_{h^{\star,\kappa}}$ remains assigned to $h^{\star,\kappa}$ under $\mathbf{y}$, possibly together with additional episodes. Next, consider any feasible network schedule σ under $\mathbf{y}$, and let $\mathbf{C}^{\sigma}_{\max}$ and $\mathbf{C}^{\sigma}_{\max,h^{\star,\kappa}}$ denote its network and binding-hospital makespan vectors, respectively. Delete from the binding-hospital schedule all episodes not in $\mathcal{P}^{\kappa}_{h^{\star,\kappa}}$, and denote the resulting makespan vector by $\widetilde{\mathbf{C}}^{\sigma}_{\max,h^{\star,\kappa}}$. The resulting schedule is feasible for the local problem defining $\mathrm{AUX}^{\star}_{h^{\star,\kappa}}(\mathbf{y}^{\kappa})$, and

$$\widetilde{\mathbf{C}}^{\sigma}_{\max,h^{\star,\kappa}} \leq \mathbf{C}^{\sigma}_{\max,h^{\star,\kappa}} \leq \mathbf{C}^{\sigma}_{\max}$$

componentwise. Therefore, by local optimality and monotonicity of ρ,

$$\mathrm{AUX}^{\star}(\mathbf{y}^{\kappa}) = \mathrm{AUX}^{\star}_{h^{\star,\kappa}}(\mathbf{y}^{\kappa}) \leq \rho\big(\widetilde{\mathbf{C}}^{\sigma}_{\max,h^{\star,\kappa}}\big) \leq \rho\big(\mathbf{C}^{\sigma}_{\max}\big).$$

Because this inequality holds for every feasible network schedule σ under $\mathbf{y}$, minimizing over such schedules gives

$$\mathrm{SP}^{\star}(\mathbf{y}) \geq \mathrm{AUX}^{\star}(\mathbf{y}^{\kappa}).$$

Hence, for any integer master-epigraph point associated with this assignment,

$$\theta \geq \mathrm{SP}^{\star}(\mathbf{y}) \geq \mathrm{AUX}^{\star}(\mathbf{y}^{\kappa}).$$

This is (10) with $\Delta^{\kappa}_{\mathrm{AUX}}(\mathbf{y})=0$.

If $\Delta^{\kappa}_{\mathrm{AUX}}(\mathbf{y})\geq 1$, then, since $\mathrm{AUX}^{\star}(\mathbf{y}^{\kappa})\geq 0$,

$$\mathrm{AUX}^{\star}(\mathbf{y}^{\kappa})\big(1-\Delta^{\kappa}_{\mathrm{AUX}}(\mathbf{y})\big) \leq 0,$$

so (10) follows from $\theta \geq 0$. Thus, the inequality is valid for every feasible integer assignment.

Dominance. Define the total number of episodes removed from their incumbent hospitals by

$$\Delta^{\kappa}_{\mathrm{NG}}(\mathbf{y}) := \sum_{p\in\mathcal{P}}\big(1-y_{p,h^{\kappa}(p)}\big).$$

For every $p\in\mathcal{P}_{h^{\star,\kappa}}(\mathbf{y}^{\kappa})$, we have $h^{\kappa}(p)=h^{\star,\kappa}$. Hence, $\Delta^{\kappa}_{\mathrm{AUX}}(\mathbf{y})$ contains a subset of the nonnegative terms in $\Delta^{\kappa}_{\mathrm{NG}}(\mathbf{y})$, and therefore

$$\Delta^{\kappa}_{\mathrm{AUX}}(\mathbf{y}) \leq \Delta^{\kappa}_{\mathrm{NG}}(\mathbf{y}).$$

Since $\mathrm{AUX}^{\star}(\mathbf{y}^{\kappa})\geq 0$, it follows that

$$\mathrm{AUX}^{\star}(\mathbf{y}^{\kappa})\big(1-\Delta^{\kappa}_{\mathrm{AUX}}(\mathbf{y})\big) \geq \mathrm{AUX}^{\star}(\mathbf{y}^{\kappa})\big(1-\Delta^{\kappa}_{\mathrm{NG}}(\mathbf{y})\big).$$

Thus, the right-hand side of (10) dominates that of (11) pointwise. This comparison also holds for fractional assignments $\mathbf{y}\in[0,1]^{|\mathcal{P}|\times|\mathcal{H}|}$. $\square$

THEOREM 2. *Fix an outer assignment* $\mathbf{y}^\kappa$. *Let* $\mathcal{D}^\kappa$ *denote the set of binary scheduling decisions* $(\mathbf{z},\mathbf{x})$ *for which there exist ranking variables* $\mathbf{u}$ *satisfying* (17b) *and* (17g)-(17k). *The risk-support inner L-shaped algorithm generates a non-decreasing sequence of valid LBs on* $\beta^{\star,\kappa} := \min_{(\mathbf{z},\mathbf{x})\in\mathcal{D}^\kappa} \rho\big(\mathbf{Q}^\kappa(\mathbf{z},\mathbf{x};\cdot)\big)$ *and terminates finitely at a decision* $(\mathbf{z},\mathbf{x})\in\mathcal{D}^\kappa$ *attaining* $\beta^{\star,\kappa}$.

Proof. The set $\mathcal{D}^\kappa$ is finite because $\mathbf{z}$ and $\mathbf{x}$ are binary with finite index sets. The ranking constraints ensure that each $(\mathbf{z},\mathbf{x})\in\mathcal{D}^\kappa$ results in an acyclic precedence structure, so every scenario timing problem has a finite optimum. At iteration ν, let $\mathcal{L}^{\kappa,\nu}(\mathbf{z},\mathbf{x};\omega)$ denote the largest scenario-ω LB imposed on $C_{\max}^{\text{MP-SP}}(\omega)$ by (17d)-(17e) and the L-shaped cuts generated in iterations $\ell<\nu$ for which $\omega\in\Gamma^{\kappa,\ell}$. The structural bounds are valid, and weak duality gives

$$\Theta^{\kappa,\ell}(\mathbf{z},\mathbf{x};\omega) \le Q^\kappa(\mathbf{z},\mathbf{x};\omega).$$

Therefore,

$$\mathcal{L}^{\kappa,\nu}(\mathbf{z},\mathbf{x};\omega) \le Q^\kappa(\mathbf{z},\mathbf{x};\omega) \qquad \forall\,(\mathbf{z},\mathbf{x})\in\mathcal{D}^\kappa,\ \omega\in\Omega.$$

Let $\underline{\beta}^\nu$ be the optimal value of MP-SP at iteration ν. At fixed $(\mathbf{z},\mathbf{x})$, (17a) and (17f) together with monotonicity of ρ place the optimum at $C_{\max}^{\text{MP-SP}}(\omega)=\mathcal{L}^{\kappa,\nu}(\mathbf{z},\mathbf{x};\omega)$ for every ω, so

$$\underline{\beta}^\nu = \min_{(\mathbf{z},\mathbf{x})\in\mathcal{D}^\kappa} \rho\big(\mathcal{L}^{\kappa,\nu}(\mathbf{z},\mathbf{x};\cdot)\big) \le \beta^{\star,\kappa},$$

where the inequality uses monotonicity of ρ once more. Thus, $\underline{\beta}^\nu$ is a valid LB. Since each iteration only adds cuts,

$$\mathcal{L}^{\kappa,\nu+1}(\mathbf{z},\mathbf{x};\omega) \ge \mathcal{L}^{\kappa,\nu}(\mathbf{z},\mathbf{x};\omega) \quad\Longrightarrow\quad \underline{\beta}^{\nu+1} \ge \underline{\beta}^\nu$$

Because Ω is finite and ρ is a real-valued coherent risk measure on $\mathbb{R}^{|\Omega|}$, it admits the representation (14) over a nonempty compact convex risk envelope Π_ρ. Hence, the maximum in (14) is attained. To prove finite termination, suppose that $(\mathbf{z}^{\kappa,\ell},\mathbf{x}^{\kappa,\ell})$, first evaluated at iteration ℓ, is selected again at iteration $\nu>\ell$. Let $\pi^{\kappa,\ell}\in\Pi_\rho$ be an optimizer of (15) at iteration ℓ, with support $\Gamma^{\kappa,\ell}$. Then

$$\rho\big(\mathbf{Q}^\kappa(\mathbf{z}^{\kappa,\ell},\mathbf{x}^{\kappa,\ell};\cdot)\big) = \sum_{\omega\in\Gamma^{\kappa,\ell}} \pi^{\kappa,\ell}(\omega) Q^\kappa(\mathbf{z}^{\kappa,\ell},\mathbf{x}^{\kappa,\ell};\omega).$$

Because the cuts generated at iteration ℓ are tight at this decision,

$$\mathcal{L}^{\kappa,\nu}(\mathbf{z}^{\kappa,\ell},\mathbf{x}^{\kappa,\ell};\omega) \ge Q^\kappa(\mathbf{z}^{\kappa,\ell},\mathbf{x}^{\kappa,\ell};\omega) \qquad \forall\,\omega\in\Gamma^{\kappa,\ell}.$$

Since this decision is optimal for the inner master at iteration ν, the risk-envelope representation gives

$$\begin{aligned}
\underline{\beta}^\nu &= \rho\big(\mathcal{L}^{\kappa,\nu}(\mathbf{z}^{\kappa,\ell},\mathbf{x}^{\kappa,\ell};\cdot)\big) \\
&\ge \sum_{\omega\in\Gamma^{\kappa,\ell}} \pi^{\kappa,\ell}(\omega)\mathcal{L}^{\kappa,\nu}(\mathbf{z}^{\kappa,\ell},\mathbf{x}^{\kappa,\ell};\omega) \\
&\ge \rho\big(\mathbf{Q}^\kappa(\mathbf{z}^{\kappa,\ell},\mathbf{x}^{\kappa,\ell};\cdot)\big).
\end{aligned}$$

Let $\overline{\beta}^{\nu}$ be the smallest value $\rho\big(\mathbf{Q}^{\kappa}(\mathbf{z},\mathbf{x};\cdot)\big)$ among the incumbents evaluated in iterations $\ell \le \nu$. Being the risk value of a feasible decision, it satisfies $\beta^{\star,\kappa} \le \overline{\beta}^{\nu}$. Since the present decision has already been evaluated,

$$\overline{\beta}^{\nu} \le \rho\big(\mathbf{Q}^{\kappa}(\mathbf{z}^{\kappa,\ell},\mathbf{x}^{\kappa,\ell};\cdot)\big).$$

Combining these inequalities with $\underline{\beta}^{\nu} \le \beta^{\star,\kappa} \le \overline{\beta}^{\nu}$ shows that $\underline{\beta}^{\nu} = \rho\big(\mathbf{Q}^{\kappa}(\mathbf{z}^{\kappa,\ell},\mathbf{x}^{\kappa,\ell};\cdot)\big) = \overline{\beta}^{\nu} = \beta^{\star,\kappa}$. Thus, revisiting an evaluated decision closes the inner optimality gap, and that decision attains $\beta^{\star,\kappa}$. Every nonterminating iteration must therefore evaluate a previously unseen element of the finite set $\mathcal{D}^{\kappa}$. Hence, the algorithm terminates finitely at a decision attaining $\beta^{\star,\kappa}$. □

Appendix D: Illustrative Examples

D.1. AUX-SP and SP Relation

Figure D.1 illustrates that the per-hospital decomposition can provide a strict LB on the original SP. The example contains two hospitals, two stages, four surgical episodes, and two equiprobable scenarios, with $\rho = \mathbb{E}$. Each hospital has one resource at each stage, and the hospital assignment is fixed as

$$\mathcal{P}_{h_1} = \{p_1, p_2\}, \qquad \mathcal{P}_{h_2} = \{p_3, p_4\}, \qquad \mathfrak{p}(\omega_1) = \mathfrak{p}(\omega_2) = 0.5.$$

Table D.3 reports the durations, where each entry gives the pair of stage durations (s_1, s_2).

Table D.3 Durations used in the AUX-SP illustration.

Hospital	Patient	ω_1	ω_2
h_1	p_1	$(1,3)$	$(2,1)$
h_1	p_2	$(3,1)$	$(1,7)$
h_2	p_3	$(1,4)$	$(4,1)$
h_2	p_4	$(3,2)$	$(1,4)$

In AUX-SP, each hospital is optimized independently. In this instance, the optimal local schedules happen to use the same patient order at both stages: $h_1:\ p_1 \prec p_2, h_2:\ p_4 \prec p_3$, where $p_i \prec p_j$ means that p_i is processed before p_j at both stages. The corresponding hospital makespan vectors are $\mathbf{C}_{\max,h_1} = (5,10), \mathbf{C}_{\max,h_2} = (9,6)$, so both hospitals have an expected makespan of 7.5. Therefore, $\mathrm{AUX}^{\star}(\mathbf{y}) = \max\{7.5, 7.5\} = 7.5$. The original SP instead optimizes the scenario-wise network makespan. Its optimal orders are $h_1 : p_2 \prec p_1, h_2 : p_3 \prec p_4$, which give hospital makespan vectors $\mathbf{C}_{\max,h_1} = \mathbf{C}_{\max,h_2} = (7,9)$. The network makespan vector is consequently $(7,9)$, and $\mathrm{SP}^{\star}(\mathbf{y}) = \mathbb{E}[\mathbf{C}_{\max}] = 8$. Thus,

$$\mathrm{AUX}^{\star}(\mathbf{y}) = 7.5 < 8 = \mathrm{SP}^{\star}(\mathbf{y}),$$

consistent with Proposition 1.

The schedules attaining the AUX-SP value are therefore not necessarily network-optimal. If the two locally optimal schedules are combined, the network makespan vector is $\big(\max\{5,9\},\max\{10,6\}\big)=(9,10)$, with expected makespan 9.5, compared with 8 for the network-optimal schedule. The difference arises because AUX-SP applies the risk measure separately at each hospital and then takes the maximum, whereas SP first forms the scenario-wise network maximum and then applies the risk measure.

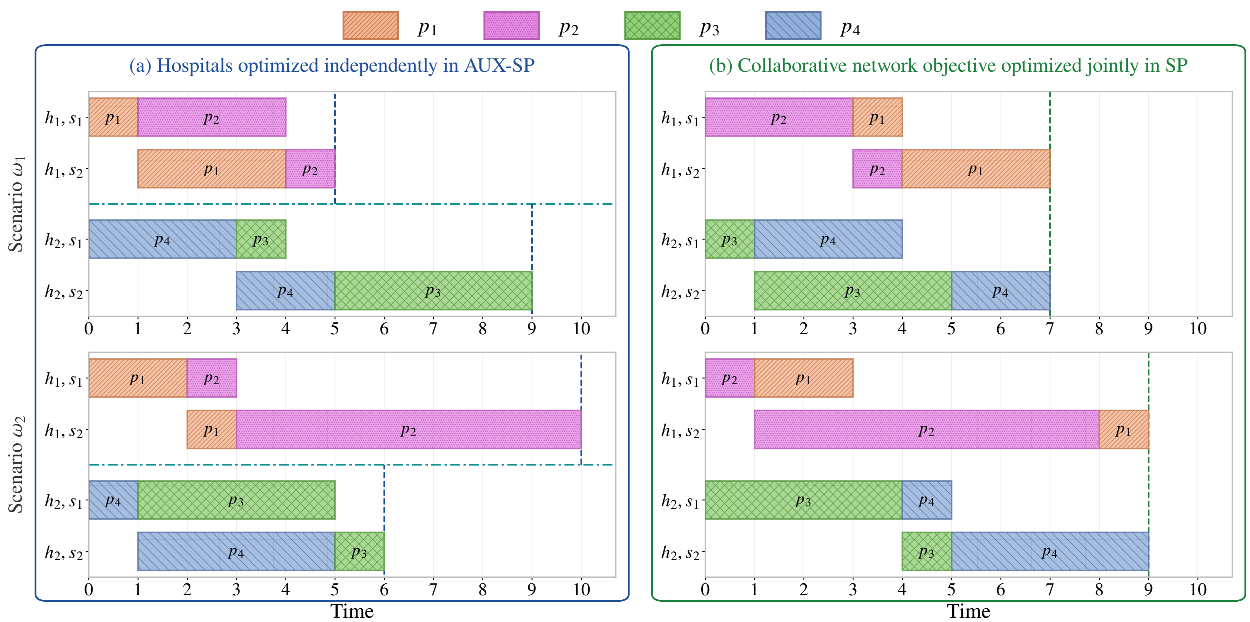

Figure D.1 Illustrative comparison between the per-hospital AUX-SP and the original integrated SP. Panel (a) optimizes hospital schedules independently, whereas the panel (b) optimizes the network-wide scenario makespan jointly.

D.2. Illustrative Construction of the Scenario DAG

Figure D.2 illustrates the scenario-wise graph used to evaluate the timing problem SP-SP(ω) for a fixed incumbent $(\mathbf{y}^{\kappa},\mathbf{z}^{\kappa,\nu},\mathbf{x}^{\kappa,\nu})$, as discussed in Section 4.4. For brevity, we suppress the superscripts (κ,ν). The example contains three surgical episodes and two stages. Each node $v_{p,s}$ represents the operation of surgical episode p at stage s in its assigned hospital $h(p)$. The source node o initializes every surgical episode at stage 1, and the sink node t collects the completion of all surgical episodes.

Black arcs represent source arcs and within-episode pathway arcs, with weights equal to the corresponding scenario-dependent durations $D_{p,h(p),s}(\omega)$. Gray arcs connect final-stage operations to the sink and carry zero weight. Dashed red arcs encode the resource-precedence decisions induced by $\mathbf{x}$. In the example, surgical episodes 1 and 2 share resource 1 at stage 1, and the incumbent sequence places episode 1 before episode 2, so the arc $(v_{1,1},v_{2,1})$ is added with weight $D_{2,h(2),1}(\omega)$.

Similarly, episodes 2 and 3 share resource 2 at stage 2 with episode 2 preceding episode 3, yielding the arc $(v_{2,2}, v_{3,2})$ with weight $D_{3,h(3),2}(\omega)$. The scenario makespan $Q(\mathbf{z}, \mathbf{x}; \omega)$ is then the length of the longest o-t path in this DAG.

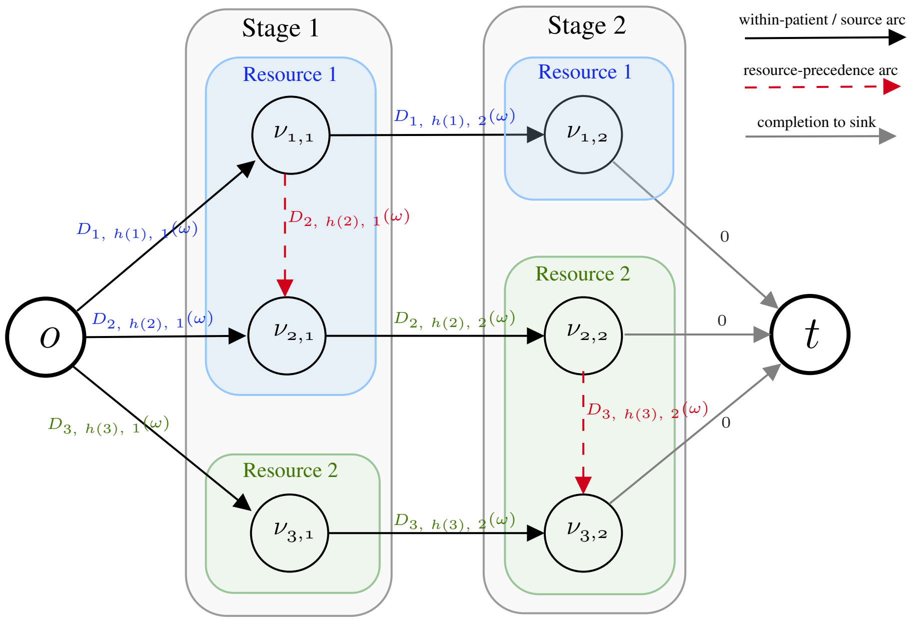

Figure D.2 Illustrative DAG $G^{\kappa,\nu}(\omega)$ for a fixed incumbent $(\mathbf{y}, \mathbf{z}, \mathbf{x})$, showing source/pathway arcs, resource-precedence arcs induced by $\mathbf{x}$, and sink arcs.

Appendix E: Data Generation

To the best of our knowledge, no public benchmark set is available for RA-SEEN or for closely related multi-hospital, multi-stage surgical episode scheduling problems. We therefore developed a practice-informed instance generator using three sources: publicly available information on the University Health Network surgical pathway and activity levels described in Section 2.1, empirical evidence from the perioperative operations literature, and consultation with a domain expert. When site-level numerical data were not publicly available, we used calibration rules designed to preserve realistic relationships among patient volumes, stage durations, resource capacities, and upstream and downstream congestion. The generated instances, parameter files, and random seeds are made publicly available to support replication and future comparisons.

Network structure and instance sizes. We consider $|\mathcal{H}| = 3$ hospitals, representing Toronto General Hospital, Toronto Western Hospital, and Princess Margaret Cancer Centre. We generate either the five-stage UHN pathway described in Section 2.1, $(\mathrm{REG}, \mathrm{POCU}, \mathrm{OR}, \mathrm{PACU}, \mathrm{DSU})$, or a four-stage variant, $(\mathrm{REG}, \mathrm{POCU}, \mathrm{OR}, \mathrm{REC})$, in which PACU and DSU are aggregated into one postoperative recovery stage. The five-stage setting represents the clinical pathway more closely, whereas the four-stage setting allows us to examine the effect of model size and pathway granularity. The number of surgical episodes and scenarios is selected from

$$|\mathcal{P}| \in \{10, 15, 20, 40, 60, 80, 100, 120, 140\}, \qquad |\Omega| \in \{100, 250, 500\}.$$

The upper end of the patient range is consistent with the scale of UHN, which reported 31,004 acute day-surgery cases in 2022/23 (UHN, 2023). For each base instance, the scenario sets are nested: the 250-scenario set contains the first 100 scenarios, and the 500-scenario set contains the first 250. This construction isolates the computational effect of increasing $|\Omega|$ without changing the underlying instance.

Stage capacities. For each hospital h, let $m_h^{\mathrm{OR}} := |\mathcal{R}_{h,\mathrm{OR}}|$ denote its OR capacity. The support of m_h^{OR} scales with the number of surgical episodes as follows:

$\lvert\mathcal{P}\rvert$	10	15	20	40	60	80	100	120	140
Support of m_h^{OR}	$\{2,3,4\}$	$\{3,4,5\}$	$\{4,5,6\}$	$\{5,6,7\}$	$\{6,7,8\}$	$\{7,8,9\}$	$\{8,9,10\}$	$\{9,10,11\}$	$\{10,11,12\}$

For each support, the low, middle, and high values are selected with probabilities 0.15, 0.50, and 0.35, respectively. The support ranges were selected to scale OR capacity with network size, using UHN activity levels, expert input, and the reported Canadian same-day throughput benchmark of approximately 2.7 patients per OR per day as reference points (Head et al., 2011). The asymmetric probabilities place slightly greater weight on the higher-capacity realization, increasing the expected OR capacity by 0.2 relative to the middle support value.

The capacities of the remaining stages are calibrated relative to OR capacity, as it is the central stage around which same-day surgical flow is organized:

$$|\mathcal{R}_{h,\mathrm{REG}}| = \max\left\{2, \left\lceil 0.4 m_h^{\mathrm{OR}} \right\rceil\right\}, \qquad |\mathcal{R}_{h,\mathrm{POCU}}| = \max\left\{2, \left\lceil 0.7 m_h^{\mathrm{OR}} \right\rceil\right\},$$
$$|\mathcal{R}_{h,\mathrm{PACU}}| = \max\left\{3, \left\lceil m_h^{\mathrm{OR}} \right\rceil\right\}, \qquad |\mathcal{R}_{h,\mathrm{DSU}}| = \max\left\{2, \left\lceil 0.8 m_h^{\mathrm{OR}} \right\rceil\right\}.$$

In the four-stage setting, we use $|\mathcal{R}_{h,\mathrm{REC}}| = \max\{3, \lceil m_h^{\mathrm{OR}} \rceil\}$. These ratios do not represent reported physical capacities at individual UHN sites. Rather, they are expert-informed calibration rules that reflect the shorter and more flexible nature of registration and preoperative activities while keeping postoperative capacity sufficiently close to OR capacity for downstream congestion to arise. The scaling factors are designed such that it is possible to face upstream and downstream congestion (Shehadeh and Padman, 2022), with consideration that capacities should not fall below a specific stage-based LB. The calibration is also consistent with evidence that OR and PACU activity are tightly linked, and that PACU delays may arise from bed availability, staffing, and discharge bottlenecks (Weissman et al., 2019; Dexter et al., 2005; Heyba et al., 2024). The resulting instances can therefore exhibit OR, PACU, and discharge bottlenecks rather than being dominated by a single stage by construction.

Surgical episode classes and baseline durations. Each episode p is assigned independently to one of four same-day surgery classes representing short, medium, long, and complex cases. The respective probabilities are 0.25, 0.35, 0.25, and 0.15, producing a case mix concentrated on medium-duration episodes while retaining short and complex cases. Such a classification scheme is analogous to those used in the literature for benchmarking perioperative scheduling under heterogeneous runtime phases (Leeftink and Hans, 2018). It is also consistent with complexity overlays used in the health care database system in Canada to classify ambulatory episodes by resource usage and clinical complexity (CIHI, 2012).

Baseline stage-duration parameters B_{ps} were calibrated through consultation with a domain expert and, where comparable evidence was available, assessed against published ambulatory perioperative service times. The values represent class-specific mean service requirements under regular operating conditions. Registration is modeled as a short administrative stage using a triangular distribution with minimum, mode, and maximum durations of (5), (8), and (15) minutes, respectively. These parameters were selected in consultation with the domain expert to represent routine registration service under regular operating conditions. For the remaining stages, baseline durations are generated from class- and stage-specific lognormal distributions. Empirical studies support the use of right-skewed distributions, particularly the lognormal distribution, for surgical durations (Strum et al., 2000; Roshanaei et al., 2017). We also use lognormal distributions for the other perioperative stages to represent their nonnegative, right-skewed durations consistently. The class-specific mean durations increase with episode complexity and range from (17)-(32) minutes in POCU, (37)-(85) minutes in the OR, (29)-(73) minutes in PACU, and (24)-(57) minutes in DSU. The POCU range is broadly consistent with the standard nurse-led preoperative assessment times reported by Taylor et al. (2018). Although measured in a preoperative assessment setting rather than specifically in a day-of-surgery POCU, these values provide a relevant benchmark for the scale of the modeled activities. The OR range overlaps with procedure-specific total OR times reported by Bhattacharyya (2011), while the PACU and DSU ranges are consistent with published evidence on Phase-I and Phase-II recovery in ambulatory surgery (White et al., 2003). These parameters define the baseline duration distributions; unusually long service realizations and adverse operating conditions are represented through the right tails of the lognormal distributions and scenario-level variation.

Stage-specific coefficients of variation range from 0.30 to 0.45, with the largest value assigned to PACU. This calibration reflects the sensitivity of postoperative recovery to patient condition, pain, drowsiness, nausea and vomiting, and other clinical factors (Pavlin et al., 1998). To capture within-episode dependence, the non-registration stage durations are generated jointly using a Gaussian copula (Nelsen, 2006). All cross-stage correlations are positive, with stronger dependence between the OR and PACU and between PACU and DSU. Thus, an episode with a longer-than-usual operation is also more likely to require longer postoperative recovery.

Hospital and stage heterogeneity. To represent persistent differences across hospitals and care stages, we independently sample a hospital-level factor β_h and a hospital-stage factor γ_{hs}:

$$\beta_h \sim \text{Lognormal}\left(-\frac{0.06^2}{2}, 0.06^2\right), \qquad \gamma_{hs} \sim \text{Lognormal}\left(-\frac{0.05^2}{2}, 0.05^2\right).$$

Both factors have expectation one, with coefficients of variation of approximately 6% and 5%, respectively. These parameters introduce moderate persistent heterogeneity while preserving the baseline duration level on average. The factor β_h captures differences affecting the full pathway at hospital h, whereas γ_{hs} captures additional stage-specific differences within that hospital. The multiplicative structure is consistent with empirical evidence that surgical durations vary systematically with operational, provider, and anesthesia-related factors, and that such variation can be proportional to baseline duration (Kayıcs et al., 2015).

The factors are sampled once for each base instance and held fixed across episodes and scenarios. The resulting nominal duration is then

$$D_{phs}^{\text{nom}} = \max\left\{\ell_s, \text{round}\left(\beta_h \gamma_{hs} B_{ps}\right)\right\},$$

where ℓ_s prevents unrealistically short realizations after the multiplicative adjustments. In the five-stage setting, $(\ell_{\text{REG}}, \ell_{\text{POCU}}, \ell_{\text{OR}}, \ell_{\text{PACU}}, \ell_{\text{DSU}}) = (3, 10, 20, 20, 20)$ minutes.

Scenario generation. Scenarios are generated from the nominal durations by adding two sources of variation: operating-day conditions and episode-level noise. Each scenario is assigned a day type defined below, following a categorical distribution that gives positive probability to all types.

- *Regular days* capture normal operating conditions. Durations may still vary, but there is no major slowdown affecting a particular part of the pathway.
- *Slower OR days* capture situations where delays are mainly concentrated around the operating room and nearby perioperative activities. (Kayıcs et al., 2015; Wang et al., 2020).
- *Slower recovery days* capture situations where delays mainly occur in PACU and DSU. These scenarios reflect the role of recovery capacity in OR flow and patient throughput, for example through PACU hold times, discharge orders, transfer decisions, and recovery transitions (Johnson and Calixte, 2023).
- *System-wide slowdown days* capture days where all stages are moderately slower, for example because of staffing pressure, coordination issues, or general execution delays. This reflects the fact that hospital care is delivered through connected activities, resources, and professionals, so bottlenecks may arise across the process rather than at one stage only (Agostinelli et al., 2020).

- *Hospital-specific disruption days* capture days where one hospital is slower across its stages. These scenarios test whether the network can benefit from shifting work toward other hospitals when one site is temporarily less efficient. This is aligned with prior studies showing that different hospital resources or operative blocks can have different temporal performance profiles (Agostinelli et al., 2020).

These patterns create correlated delays across related stages rather than perturbing all durations independently. Each day type defines a vector of stage multipliers. Regular days use unit multipliers; OR-centered disruptions apply the largest increase to the OR and smaller increases to adjacent stages; recovery-centered disruptions primarily increase PACU and DSU durations; and system-wide disruptions moderately increase durations throughout the pathway. For hospital-specific disruptions, one hospital is selected and receives stage-dependent increases, while the other hospitals remain under regular conditions. Let $F_{phs}(\omega)$ denote the resulting structured multiplier for episode p, hospital h, and stage s in scenario ω. The day-type probabilities and multiplier values are expert-informed calibration choices intended to generate a diverse mix of regular and disrupted operating conditions. The realized duration is

$$D_{phs}(\omega) = \max\left\{\ell_s, \operatorname{round}\left(D_{phs}^{\text{nom}} F_{phs}(\omega)\varepsilon_{phs}(\omega)\right)\right\}.$$

Thus, each scenario combines structured operating-day effects with episode-level random noise, in addition to the persistent hospital and stage heterogeneity embedded in the nominal durations.

Four-stage aggregation. For the four-stage setting, we first generate durations for the five-stage pathway and then define the combined postoperative recovery duration as

$$D_{ph,\text{REC}}(\omega) = \max\left\{D_{ph,\text{PACU}}(\omega), D_{ph,\text{DSU}}(\omega)\right\}.$$

This approximation represents the recovery stage by the dominant of the two postoperative phases. It is used to reduce pathway granularity for the computational comparison.

Appendix F: Supplementary Computational Tables

This appendix reports the detailed computational tables for the compact MIP, compact CP, two-level decomposition, and three-level decomposition methods, evaluated on five instances per setting. Column "# Fail" reports the number of instances for which the method failed to return a feasible incumbent, and column "# Opt" reports the number of instances whose final optimality gap is at most 5%. When at least three of the five instances fail to return a feasible incumbent for a given setting, the mean optimality gap and median solution time are omitted, as fewer than three remaining instances are not representative. The label "TL" indicates that the time limit was reached, and the label "NC" implies instances could not be compiled.

Table F.4: MIP results summary.

				# Fail			# Opt			Mean Optimality Gap (%)			Median Time (s)		
Size	Patients ($\lvert\mathcal{P}\rvert$)	Stages ($\lvert S\rvert$)	ρ	$\lvert\Omega\rvert$=100	$\lvert\Omega\rvert$=250	$\lvert\Omega\rvert$=500	$\lvert\Omega\rvert$=100	$\lvert\Omega\rvert$=250	$\lvert\Omega\rvert$=500	$\lvert\Omega\rvert$=100	$\lvert\Omega\rvert$=250	$\lvert\Omega\rvert$=500	$\lvert\Omega\rvert$=100	$\lvert\Omega\rvert$=250	$\lvert\Omega\rvert$=500
Small	10	4	Exp	0	0	0	5	5	5	0	0.31	0.27	386	5490	9211
			$\text{CVaR}_{90\%}$	0	0	0	5	5	5	0	0	0	413	3006	8664
			$\text{CVaR}_{95\%}$	0	0	0	5	5	5	0	0	0.03	222	3089	8185
			ess-sup	0	0	0	5	5	5	0	0	0	19	129	531
		5	Exp	0	0	0	5	5	4	0.16	0.86	3.03	2512	TL	TL
			$\text{CVaR}_{90\%}$	0	0	0	5	5	4	0	0	7.05	793	5195	12059
			$\text{CVaR}_{95\%}$	0	0	0	5	5	4	0	0	5.58	681	3881	10874
			ess-sup	0	0	0	5	5	5	0	0	0	46	161	1741
	15	4	Exp	0	0	0	5	1	0	0.97	5.91	35.67	TL	TL	TL
			$\text{CVaR}_{90\%}$	0	0	0	5	2	3	0	9.10	10.84	2325	TL	TL
			$\text{CVaR}_{95\%}$	0	0	0	5	5	2	0	0	17.65	1424	11699	TL
			ess-sup	0	0	1	5	5	4	0	0	0	160	860	1614
		5	Exp	0	1	0	2	1	0	5.82	10.11	62.35	TL	TL	TL
			$\text{CVaR}_{90\%}$	0	1	0	5	1	0	0.97	25.11	60.01	7796	TL	TL
			$\text{CVaR}_{95\%}$	0	1	0	5	0	0	0	22.06	48.63	3409	TL	TL
			ess-sup	0	1	0	5	4	5	0	0	0	411	1040	15655
	20	4	Exp	0	0	1	2	0	0	5.20	30.51	61.42	TL	TL	TL
			$\text{CVaR}_{90\%}$	0	2	0	5	1	0	0	24.23	52.85	13026	TL	TL
			$\text{CVaR}_{95\%}$	0	1	0	5	1	0	0	19.91	45.61	5342	TL	TL
			ess-sup	0	0	0	5	5	5	0	0	0	1239	2864	12130
		5	Exp	0	0	1	3	0	0	4.60	52.16	61.39	TL	TL	TL
			$\text{CVaR}_{90\%}$	0	0	0	3	0	0	3.90	60.46	63.25	10550	TL	TL
			$\text{CVaR}_{95\%}$	0	0	1	4	0	0	1.86	49.50	61.28	7478	TL	TL
			ess-sup	0	0	4	5	5	0	0	0	-	2473	2240	-
Medium	40	4	Exp	0	NC	NC	0	0	0	30.23	-	-	TL	-	-
			$\text{CVaR}_{90\%}$	0	NC	NC	0	0	0	31.99	-	-	TL	-	-
			$\text{CVaR}_{95\%}$	0	NC	NC	0	0	0	20.36	-	-	TL	-	-

Continued on next page

Table F.4: MIP results summary (continued).

Size	Patients ($\lvert\mathcal{P}\rvert$)	Stages ($\lvert S\rvert$)	ρ	# Fail $\lvert\Omega\rvert$=100	# Fail $\lvert\Omega\rvert$=250	# Fail $\lvert\Omega\rvert$=500	# Opt $\lvert\Omega\rvert$=100	# Opt $\lvert\Omega\rvert$=250	# Opt $\lvert\Omega\rvert$=500	Mean Optimality Gap (%) $\lvert\Omega\rvert$=100	Mean Optimality Gap (%) $\lvert\Omega\rvert$=250	Mean Optimality Gap (%) $\lvert\Omega\rvert$=500	Median Time (s) $\lvert\Omega\rvert$=100	Median Time (s) $\lvert\Omega\rvert$=250	Median Time (s) $\lvert\Omega\rvert$=500
			ess-sup	1	NC	NC	4	0	0	0	-	-	5775	-	-
		5	Exp	0	NC	NC	0	0	0	40.66	-	-	TL	-	-
			$\text{CVaR}_{90\%}$	0	NC	NC	0	0	0	55.74	-	-	TL	-	-
			$\text{CVaR}_{95\%}$	0	NC	NC	1	0	0	36.59	-	-	TL	-	-
			ess-sup	1	NC	NC	2	0	0	5.43	-	-	10250	-	-
	60	4	Exp	NC	NC	NC	0	0	0	-	-	-	-	-	-
			$\text{CVaR}_{90\%}$	NC	NC	NC	0	0	0	-	-	-	-	-	-
			$\text{CVaR}_{95\%}$	NC	NC	NC	0	0	0	-	-	-	-	-	-
			ess-sup	NC	NC	NC	0	0	0	-	-	-	-	-	-
		5	Exp	NC	NC	NC	0	0	0	-	-	-	-	-	-
			$\text{CVaR}_{90\%}$	NC	NC	NC	0	0	0	-	-	-	-	-	-
			$\text{CVaR}_{95\%}$	NC	NC	NC	0	0	0	-	-	-	-	-	-
			ess-sup	NC	NC	NC	0	0	0	-	-	-	-	-	-
	80	4	Exp	NC	NC	NC	0	0	0	-	-	-	-	-	-
			$\text{CVaR}_{90\%}$	NC	NC	NC	0	0	0	-	-	-	-	-	-
			$\text{CVaR}_{95\%}$	NC	NC	NC	0	0	0	-	-	-	-	-	-
			ess-sup	NC	NC	NC	0	0	0	-	-	-	-	-	-
		5	Exp	NC	NC	NC	0	0	0	-	-	-	-	-	-
			$\text{CVaR}_{90\%}$	NC	NC	NC	0	0	0	-	-	-	-	-	-
			$\text{CVaR}_{95\%}$	NC	NC	NC	0	0	0	-	-	-	-	-	-
			ess-sup	NC	NC	NC	0	0	0	-	-	-	-	-	-
Large	100	4	Exp	NC	NC	NC	0	0	0	-	-	-	-	-	-
			$\text{CVaR}_{90\%}$	NC	NC	NC	0	0	0	-	-	-	-	-	-
			$\text{CVaR}_{95\%}$	NC	NC	NC	0	0	0	-	-	-	-	-	-
			ess-sup	NC	NC	NC	0	0	0	-	-	-	-	-	-
		5	Exp	NC	NC	NC	0	0	0	-	-	-	-	-	-

Continued on next page

Table F.4: MIP results summary (continued).

Size	Patients ($\lvert\mathcal{P}\rvert$)	Stages ($\lvert S\rvert$)	ρ	# Fail $\lvert\Omega\rvert$=100	# Fail $\lvert\Omega\rvert$=250	# Fail $\lvert\Omega\rvert$=500	# Opt $\lvert\Omega\rvert$=100	# Opt $\lvert\Omega\rvert$=250	# Opt $\lvert\Omega\rvert$=500	Mean Optimality Gap (%) $\lvert\Omega\rvert$=100	Mean Optimality Gap (%) $\lvert\Omega\rvert$=250	Mean Optimality Gap (%) $\lvert\Omega\rvert$=500	Median Time (s) $\lvert\Omega\rvert$=100	Median Time (s) $\lvert\Omega\rvert$=250	Median Time (s) $\lvert\Omega\rvert$=500
			$\mathrm{CVaR}_{90\%}$	NC	NC	NC	0	0	0	-	-	-	-	-	-
			$\mathrm{CVaR}_{95\%}$	NC	NC	NC	0	0	0	-	-	-	-	-	-
			ess-sup	NC	NC	NC	0	0	0	-	-	-	-	-	-
	120	4	Exp	NC	NC	NC	0	0	0	-	-	-	-	-	-
			$\mathrm{CVaR}_{90\%}$	NC	NC	NC	0	0	0	-	-	-	-	-	-
			$\mathrm{CVaR}_{95\%}$	NC	NC	NC	0	0	0	-	-	-	-	-	-
			ess-sup	NC	NC	NC	0	0	0	-	-	-	-	-	-
		5	Exp	NC	NC	NC	0	0	0	-	-	-	-	-	-
			$\mathrm{CVaR}_{90\%}$	NC	NC	NC	0	0	0	-	-	-	-	-	-
			$\mathrm{CVaR}_{95\%}$	NC	NC	NC	0	0	0	-	-	-	-	-	-
			ess-sup	NC	NC	NC	0	0	0	-	-	-	-	-	-
	140	4	Exp	NC	NC	NC	0	0	0	-	-	-	-	-	-
			$\mathrm{CVaR}_{90\%}$	NC	NC	NC	0	0	0	-	-	-	-	-	-
			$\mathrm{CVaR}_{95\%}$	NC	NC	NC	0	0	0	-	-	-	-	-	-
			ess-sup	NC	NC	NC	0	0	0	-	-	-	-	-	-
		5	Exp	NC	NC	NC	0	0	0	-	-	-	-	-	-
			$\mathrm{CVaR}_{90\%}$	NC	NC	NC	0	0	0	-	-	-	-	-	-
			$\mathrm{CVaR}_{95\%}$	NC	NC	NC	0	0	0	-	-	-	-	-	-
			ess-sup	NC	NC	NC	0	0	0	-	-	-	-	-	-

Table F.5: CP results summary.

				# Fail			# Opt			Mean Optimality Gap (%)			Median Time (s)		
Size	Patients ($\|\mathcal{P}\|$)	Stages ($\|\mathcal{S}\|$)	ρ	$\|\Omega\|$=100	$\|\Omega\|$=250	$\|\Omega\|$=500	$\|\Omega\|$=100	$\|\Omega\|$=250	$\|\Omega\|$=500	$\|\Omega\|$=100	$\|\Omega\|$=250	$\|\Omega\|$=500	$\|\Omega\|$=100	$\|\Omega\|$=250	$\|\Omega\|$=500
Small	10	4	Exp	0	0	0	3	4	3	3.20	2.36	4.34	3357	321	TL
			$\text{CVaR}_{90\%}$	0	0	0	3	3	3	5.03	5.21	5.47	482	880	3325
			$\text{CVaR}_{95\%}$	0	0	0	5	3	4	0.01	5.88	4.20	742	713	3129
			ess-sup	0	0	0	5	5	5	0.01	0.01	0.01	11	72	279
		5	Exp	0	0	0	2	2	2	6.28	6.24	6.43	TL	TL	TL
			$\text{CVaR}_{90\%}$	0	0	0	3	2	3	7.19	9.48	5.63	1675	TL	14282
			$\text{CVaR}_{95\%}$	0	0	0	4	4	4	3.42	3.53	3.95	362	2879	9374
			ess-sup	0	0	0	5	5	5	0.01	0.01	0.01	26	92	304
	15	4	Exp	0	0	0	1	1	1	8.74	9.20	9.47	TL	TL	TL
			$\text{CVaR}_{90\%}$	0	0	0	4	1	2	2.50	12.03	14.80	751	TL	TL
			$\text{CVaR}_{95\%}$	0	0	0	5	2	3	0.01	11.75	9.49	378	12008	TL
			ess-sup	0	0	0	5	4	4	0.01	1.25	1.99	27	111	751
		5	Exp	0	0	1	1	1	1	10.55	11.05	10.71	TL	TL	TL
			$\text{CVaR}_{90\%}$	0	0	1	1	2	1	14.32	12.95	13.88	TL	TL	TL
			$\text{CVaR}_{95\%}$	0	0	0	1	1	1	15.56	17.06	16.29	TL	TL	TL
			ess-sup	0	0	0	5	4	5	0.01	1.76	0.01	63	311	1239
	20	4	Exp	0	0	0	1	0	0	10.53	12.83	13.59	TL	TL	TL
			$\text{CVaR}_{90\%}$	0	0	1	1	1	0	14.52	15.14	19.92	TL	TL	TL
			$\text{CVaR}_{95\%}$	0	0	1	4	2	0	3.39	11.64	22.33	746	TL	TL
			ess-sup	0	0	0	5	5	3	0.01	0.01	3.76	73	357	9929
		5	Exp	0	0	0	1	1	1	9.63	10.54	10.75	TL	TL	TL
			$\text{CVaR}_{90\%}$	0	0	2	2	1	1	9.71	15.24	13.13	TL	TL	TL
			$\text{CVaR}_{95\%}$	0	0	3	3	2	0	6.01	10.37	-	2131	TL	TL
			ess-sup	0	0	0	5	5	5	0.01	0.01	0.01	41	189	2899
Medium	40	4	Exp	0	0	0	1	0	0	10.34	11.58	13.02	TL	TL	TL
			$\text{CVaR}_{90\%}$	1	0	2	3	0	0	4.62	8.35	5.00	3544	8429	15927
			$\text{CVaR}_{95\%}$	0	0	4	3	0	0	7.44	9.58	-	2201	9085	-

Continued on next page

Table F.5: CP results summary (continued).

Size	Patients ($\lvert\mathcal{P}\rvert$)	Stages ($\lvert\mathcal{S}\rvert$)	ρ	# Fail			# Opt			Mean Optimality Gap (%)			Median Time (s)		
				$\lvert\Omega\rvert$=100	$\lvert\Omega\rvert$=250	$\lvert\Omega\rvert$=500	$\lvert\Omega\rvert$=100	$\lvert\Omega\rvert$=250	$\lvert\Omega\rvert$=500	$\lvert\Omega\rvert$=100	$\lvert\Omega\rvert$=250	$\lvert\Omega\rvert$=500	$\lvert\Omega\rvert$=100	$\lvert\Omega\rvert$=250	$\lvert\Omega\rvert$=500
			ess-sup	0	0	0	5	0	0	0.01	5.00	17.06	155	1416	12033
		5	Exp	0	0	0	1	0	0	8.80	12.89	10.43	TL	TL	TL
			$\text{CVaR}_{90\%}$	0	1	4	1	0	0	14.14	5.00	-	TL	3854	-
			$\text{CVaR}_{95\%}$	0	0	NC	1	0	0	15.84	6.62	-	TL	4887	-
			ess-sup	0	0	0	5	0	0	0.01	5.00	16.85	172	830	12683
	60	4	Exp	0	0	0	0	0	0	21.01	25.73	26.27	TL	TL	TL
			$\text{CVaR}_{90\%}$	0	3	NC	0	0	0	24.03	-	-	TL	-	-
			$\text{CVaR}_{95\%}$	0	4	NC	0	0	0	21.71	-	-	TL	-	-
			ess-sup	0	0	0	4	1	0	2.60	14.22	9.02	2773	16171	2227
		5	Exp	0	0	3	0	0	0	13.78	18.21	-	TL	TL	-
			$\text{CVaR}_{90\%}$	0	3	NC	0	0	0	20.69	-	-	TL	-	-
			$\text{CVaR}_{95\%}$	0	1	NC	1	0	0	16.60	57.09	-	TL	TL	-
			ess-sup	0	0	2	5	0	0	0.00	5.00	25.79	569	1732	5533
	80	4	Exp	0	NC	NC	0	0	0	19.80	-	-	TL	-	-
			$\text{CVaR}_{90\%}$	2	NC	NC	0	0	0	24.00	-	-	TL	-	-
			$\text{CVaR}_{95\%}$	0	NC	NC	0	0	0	25.33	-	-	TL	-	-
			ess-sup	0	NC	NC	3	0	0	3.84	-	-	10681	-	-
		5	Exp	0	NC	NC	0	0	0	12.94	-	-	TL	-	-
			$\text{CVaR}_{90\%}$	2	NC	NC	0	0	0	18.56	-	-	TL	-	-
			$\text{CVaR}_{95\%}$	0	NC	NC	2	0	0	11.50	-	-	TL	-	-
			ess-sup	0	NC	NC	5	0	0	0.00	-	-	2138	-	-
Large	100	4	Exp	1	0	NC	0	0	0	18.56	23.85	-	TL	TL	-
			$\text{CVaR}_{90\%}$	2	NC	NC	0	0	0	23.44	-	-	TL	-	-
			$\text{CVaR}_{95\%}$	1	NC	NC	1	0	0	19.64	-	-	TL	-	-
			ess-sup	0	0	NC	4	3	0	7.58	10.92	-	4389	TL	-
		5	Exp	0	NC	NC	0	0	0	23.02	-	-	TL	-	-

Continued on next page

Table F.5: CP results summary (continued).

Size	Patients ($\|\mathcal{P}\|$)	Stages ($\|\mathcal{S}\|$)	ρ	# Fail			# Opt			Mean Optimality Gap (%)			Median Time (s)		
				$\|\Omega\|$=100	$\|\Omega\|$=250	$\|\Omega\|$=500	$\|\Omega\|$=100	$\|\Omega\|$=250	$\|\Omega\|$=500	$\|\Omega\|$=100	$\|\Omega\|$=250	$\|\Omega\|$=500	$\|\Omega\|$=100	$\|\Omega\|$=250	$\|\Omega\|$=500
			$\text{CVaR}_{90\%}$	2	NC	NC	0	0	0	28.69	-	-	TL	-	-
			$\text{CVaR}_{95\%}$	2	NC	NC	0	0	0	27.60	-	-	TL	-	-
			ess-sup	0	NC	NC	2	0	0	5.04	-	-	TL	-	-
	120	4	Exp	0	NC	NC	0	0	0	25.79	-	-	TL	-	-
			$\text{CVaR}_{90\%}$	4	NC	NC	0	0	0	-	-	-	-	-	-
			$\text{CVaR}_{95\%}$	3	NC	NC	0	0	0	-	-	-	-	-	-
			ess-sup	0	NC	NC	2	0	0	12.75	-	-	TL	-	-
		5	Exp	0	NC	NC	0	0	0	26.28	-	-	TL	-	-
			$\text{CVaR}_{90\%}$	NC	NC	NC	0	0	0	-	-	-	-	-	-
			$\text{CVaR}_{95\%}$	4	NC	NC	0	0	0	-	-	-	-	-	-
			ess-sup	0	NC	NC	1	0	0	20.88	-	-	TL	-	-
	140	4	Exp	NC	NC	NC	0	0	0	-	-	-	-	-	-
			$\text{CVaR}_{90\%}$	NC	NC	NC	0	0	0	-	-	-	-	-	-
			$\text{CVaR}_{95\%}$	NC	NC	NC	0	0	0	-	-	-	-	-	-
			ess-sup	NC	NC	NC	0	0	0	-	-	-	-	-	-
		5	Exp	NC	NC	NC	0	0	0	-	-	-	-	-	-
			$\text{CVaR}_{90\%}$	NC	NC	NC	0	0	0	-	-	-	-	-	-
			$\text{CVaR}_{95\%}$	NC	NC	NC	0	0	0	-	-	-	-	-	-
			ess-sup	NC	NC	NC	0	0	0	-	-	-	-	-	-

Table F.6: Two-level decomposition results summary.

				# Fail			# Opt			Mean Optimality Gap (%)			Median Time (s)		
Size	Patients ($\lvert\mathcal{P}\rvert$)	Stages ($\lvert\mathcal{S}\rvert$)	ρ	$\lvert\Omega\rvert=100$	$\lvert\Omega\rvert=250$	$\lvert\Omega\rvert=500$	$\lvert\Omega\rvert=100$	$\lvert\Omega\rvert=250$	$\lvert\Omega\rvert=500$	$\lvert\Omega\rvert=100$	$\lvert\Omega\rvert=250$	$\lvert\Omega\rvert=500$	$\lvert\Omega\rvert=100$	$\lvert\Omega\rvert=250$	$\lvert\Omega\rvert=500$
Small	10	4	Exp	0	0	0	5	4	4	0	2.63	2.6	34	3849	4392
			$\text{CVaR}_{90\%}$	0	0	0	5	5	5	0	0	0	5	40	57
			$\text{CVaR}_{95\%}$	0	0	0	5	5	5	0	0	0	3	7	10
			ess-sup	0	0	0	5	5	5	0	0	0	2	4	6
		5	Exp	0	0	0	4	4	4	1.18	1.28	1.9	3808	7481	7899
			$\text{CVaR}_{90\%}$	0	0	0	5	5	5	0	0	0	3	40	26
			$\text{CVaR}_{95\%}$	0	0	0	5	5	5	0	0	0	11	6	116
			ess-sup	0	0	0	5	5	5	0	0	0	2	8	8
	15	4	Exp	0	0	0	3	3	1	3.89	4.34	6.38	14518	8480	14419
			$\text{CVaR}_{90\%}$	0	0	0	5	5	5	0	0	0	281	4574	651
			$\text{CVaR}_{95\%}$	0	0	0	5	5	5	0	0.01	0	16	3684	3827
			ess-sup	0	0	0	5	5	5	0	0	0	5	10	18
		5	Exp	0	0	0	2	2	2	10.57	11.54	11.6	14420	14476	14428
			$\text{CVaR}_{90\%}$	0	0	0	4	4	4	1	2.34	2.9	3664	7228	14478
			$\text{CVaR}_{95\%}$	0	0	0	5	4	4	0.81	2.31	2.99	3619	7268	8066
			ess-sup	0	0	0	5	5	5	0	0	0	6	14	343
	20	4	Exp	0	0	0	3	2	2	6.72	7.17	8.92	TL	TL	TL
			$\text{CVaR}_{90\%}$	0	0	0	4	4	4	3.23	2.32	1.23	14406	14414	TL
			$\text{CVaR}_{95\%}$	0	0	0	5	4	5	0	2.19	0.17	58	14415	14428
			ess-sup	0	0	0	5	5	5	0	0	0	7	59	258
		5	Exp	0	0	0	4	4	4	3.08	4.14	4.02	TL	15099	TL
			$\text{CVaR}_{90\%}$	0	0	0	5	4	5	0.09	1.19	0.91	3614	7956	7300
			$\text{CVaR}_{95\%}$	0	0	0	5	5	4	0	0.44	1.65	120	7301	7918
			ess-sup	0	0	0	5	5	5	0	0	0	22	42	89
Medium	40	4	Exp	0	0	0	4	4	4	6.21	6.08	6.99	5772	5772	5795
			$\text{CVaR}_{90\%}$	0	0	0	4	5	5	3.13	0.91	0.98	4774	2206	4506
			$\text{CVaR}_{95\%}$	0	0	0	4	5	4	1.66	0.87	1.29	4256	2191	5793

Continued on next page

Table F.6: Two-level decomposition results summary (continued).

Size	Patients ($\|\mathcal{P}\|$)	Stages ($\|\mathcal{S}\|$)	ρ	# Fail $\|\Omega\|$=100	# Fail $\|\Omega\|$=250	# Fail $\|\Omega\|$=500	# Opt $\|\Omega\|$=100	# Opt $\|\Omega\|$=250	# Opt $\|\Omega\|$=500	Mean Optimality Gap (%) $\|\Omega\|$=100	Mean Optimality Gap (%) $\|\Omega\|$=250	Mean Optimality Gap (%) $\|\Omega\|$=500	Median Time (s) $\|\Omega\|$=100	Median Time (s) $\|\Omega\|$=250	Median Time (s) $\|\Omega\|$=500
			ess-sup	0	0	0	5	5	5	0.63	0.8	0.96	553	967	1137
		5	Exp	0	0	0	4	4	4	3.54	4.79	4.52	5233	5657	5801
			$\mathrm{CVaR}_{90\%}$	0	0	0	5	5	4	0.57	1.31	1.37	1394	1889	5482
			$\mathrm{CVaR}_{95\%}$	0	0	0	5	5	5	1.02	0.31	0.47	3503	1329	2449
			ess-sup	0	0	0	5	5	5	1.6	0.81	0.98	569	515	895
	60	4	Exp	0	0	0	0	0	0	14.83	14.61	12.85	TL	TL	TL
			$\mathrm{CVaR}_{90\%}$	0	0	0	4	2	4	6.23	8.77	6.56	7097	12058	8852
			$\mathrm{CVaR}_{95\%}$	0	0	0	4	3	3	5.86	8.6	20.75	6015	11422	9105
			ess-sup	0	0	0	4	3	4	4.55	7.55	4.99	4997	7940	5096
		5	Exp	0	0	0	3	3	3	5.5	6.29	5.56	12286	11461	12212
			$\mathrm{CVaR}_{90\%}$	0	0	0	5	4	5	3.2	4.68	2.92	4139	8145	5213
			$\mathrm{CVaR}_{95\%}$	0	0	0	5	5	5	3.1	1.39	2.1	1163	2768	3432
			ess-sup	0	0	0	5	5	5	1.29	2.36	0.32	154	1608	1945
	80	4	Exp	0	0	0	2	1	1	9.5	10.53	10.72	12705	15215	15322
			$\mathrm{CVaR}_{90\%}$	0	0	4	2	2	1	5.23	20.97	-	12075	13254	-
			$\mathrm{CVaR}_{95\%}$	0	0	0	4	2	2	2.81	4.63	57.92	7220	12216	13942
			ess-sup	0	0	0	5	5	3	2.59	1.63	4.65	2438	1900	9125
		5	Exp	0	0	0	4	4	4	4.9	4.85	5.03	6089	7202	7973
			$\mathrm{CVaR}_{90\%}$	0	0	2	4	3	0	1.24	10.39	87.08	5956	9051	TL
			$\mathrm{CVaR}_{95\%}$	0	0	3	5	3	0	1.89	20.64	-	1729	8987	-
			ess-sup	0	0	0	5	4	5	2.32	2.63	1.22	722	5028	3226
Large	100	4	Exp	0	0	1	0	0	0	10.92	9.72	6.85	TL	TL	TL
			$\mathrm{CVaR}_{90\%}$	0	0	3	3	2	0	7.09	8.01	-	12560	14851	-
			$\mathrm{CVaR}_{95\%}$	0	0	0	2	4	0	7.68	4.78	77.44	12558	9882	TL
			ess-sup	0	0	0	4	4	3	2.98	4.64	6.38	5372	6760	8985
		5	Exp	0	0	0	0	0	0	10.12	10.05	9.99	TL	TL	TL

Continued on next page

Table F.6: Two-level decomposition results summary (continued).

				# Fail			# Opt			Mean Optimality Gap (%)			Median Time (s)		
Size	Patients ($\|\mathcal{P}\|$)	Stages ($\|\mathcal{S}\|$)	ρ	$\|\Omega\|$=100	$\|\Omega\|$=250	$\|\Omega\|$=500	$\|\Omega\|$=100	$\|\Omega\|$=250	$\|\Omega\|$=500	$\|\Omega\|$=100	$\|\Omega\|$=250	$\|\Omega\|$=500	$\|\Omega\|$=100	$\|\Omega\|$=250	$\|\Omega\|$=500
			$\text{CVaR}_{90\%}$	0	0	2	1	1	0	6.57	6.19	88.47	15443	15424	TL
			$\text{CVaR}_{95\%}$	0	0	1	3	3	0	6.61	3.8	94.02	11063	12541	TL
			ess-sup	0	0	0	5	5	4	3.26	2.22	3.84	2546	3821	10019
	120	4	Exp	1	1	1	0	0	1	11.33	12.41	11.93	TL	TL	14753
			$\text{CVaR}_{90\%}$	0	0	1	1	2	0	12.06	23.2	89.32	15384	12427	TL
			$\text{CVaR}_{95\%}$	0	0	0	1	2	0	11.45	31.04	94.12	15442	12484	TL
			ess-sup	0	0	0	2	2	2	9.27	6.37	9.32	12561	12600	12662
		5	Exp	0	0	0	0	0	0	15.67	14.63	14.98	TL	TL	TL
			$\text{CVaR}_{90\%}$	0	0	2	2	1	0	7.68	53.79	90.15	12571	16809	TL
			$\text{CVaR}_{95\%}$	0	0	1	3	1	0	8.07	75.05	94.17	9690	16798	TL
			ess-sup	0	0	0	3	3	3	4.06	7.11	9.19	9114	9735	9071
	140	4	Exp	3	5	5	0	0	0	-	-	-	-	-	-
			$\text{CVaR}_{90\%}$	0	0	0	0	0	0	18.16	45.77	90.72	TL	TL	TL
			$\text{CVaR}_{95\%}$	0	0	1	0	0	0	14.92	62.65	95.13	TL	TL	TL
			ess-sup	0	0	0	0	1	0	11.89	17.82	18.84	TL	15495	TL
		5	Exp	0	0	2	0	1	0	12.3	11.63	14.15	TL	TL	TL
			$\text{CVaR}_{90\%}$	0	3	NC	1	0	0	8.93	-	-	15460	-	-
			$\text{CVaR}_{95\%}$	0	3	NC	4	0	0	5.5	-	-	9640	-	-
			ess-sup	0	0	3	3	3	2	5.34	5.32	-	8857	9716	-

Table F.7: Three-level decomposition results summary.

Size	Patients ($\lvert\mathcal{P}\rvert$)	Stages ($\lvert\mathcal{S}\rvert$)	ρ	# Fail			# Opt			Mean Optimality Gap (%)			Median Time (s)		
				$\lvert\Omega\rvert$=100	$\lvert\Omega\rvert$=250	$\lvert\Omega\rvert$=500	$\lvert\Omega\rvert$=100	$\lvert\Omega\rvert$=250	$\lvert\Omega\rvert$=500	$\lvert\Omega\rvert$=100	$\lvert\Omega\rvert$=250	$\lvert\Omega\rvert$=500	$\lvert\Omega\rvert$=100	$\lvert\Omega\rvert$=250	$\lvert\Omega\rvert$=500
Small	10	4	Exp	0	0	0	5	4	4	0.79	2.72	3.06	5167	3819	7363
			$\text{CVaR}_{90\%}$	0	0	0	5	5	5	0	0	0.01	4	111	259
			$\text{CVaR}_{95\%}$	0	0	0	5	5	5	0	0.01	0	3	6	10
			ess-sup	0	0	0	5	5	5	0	0	0	1	2	3
		5	Exp	0	0	0	4	4	3	2.14	0.57	0.71	10929	10986	14405
			$\text{CVaR}_{90\%}$	0	0	0	5	5	5	0.01	0.01	0.02	4	648	20
			$\text{CVaR}_{95\%}$	0	0	0	5	5	5	0	0	0.01	17	9	19
			ess-sup	0	0	0	5	5	5	0	0	0	3	15	4
	15	4	Exp	0	0	0	1	2	1	4.79	0.05	3.63	TL	TL	TL
			$\text{CVaR}_{90\%}$	0	0	0	4	5	5	0	0.02	0.03	7819	10806	4874
			$\text{CVaR}_{95\%}$	0	0	0	5	4	2	0.01	0.01	0.01	1585	7246	TL
			ess-sup	0	0	0	5	5	5	0	0	0	9	48	164
		5	Exp	0	0	0	2	2	2	0.9	0.08	0.28	TL	TL	TL
			$\text{CVaR}_{90\%}$	0	0	0	4	4	4	3.05	2.79	2.88	10979	7236	14341
			$\text{CVaR}_{95\%}$	0	0	0	4	4	4	2.26	3.02	2.88	6076	7361	11407
			ess-sup	0	0	0	5	5	5	0	0	0.11	55	354	3610
	20	4	Exp	0	0	0	3	1	1	6.91	7.94	9.56	TL	TL	TL
			$\text{CVaR}_{90\%}$	0	0	0	3	4	4	4.03	2.61	1.34	14405	14405	TL
			$\text{CVaR}_{95\%}$	0	0	0	5	3	5	0	2.75	0.2	7225	14408	14410
			ess-sup	0	0	0	5	5	5	0	0	0	2	620	4035
		5	Exp	0	0	0	4	2	3	1.25	2.45	2.33	14817	TL	TL
			$\text{CVaR}_{90\%}$	0	0	0	5	4	3	0.09	1.21	1.28	3624	14414	7289
			$\text{CVaR}_{95\%}$	0	0	0	5	5	3	0.04	0.53	2.08	3630	7300	11116
			ess-sup	0	0	0	5	5	5	0	0	0	3634	150	1065
Medium	40	4	Exp	0	0	0	4	4	4	6.35	6.37	7.13	5051	5045	5054
			$\text{CVaR}_{90\%}$	0	0	0	4	5	4	3.63	1.5	2.58	4338	1452	4706

Continued on next page

Table F.7: Three-level decomposition results summary (continued).

Size	Patients ($\lvert\mathcal{P}\rvert$)	Stages ($\lvert\mathcal{S}\rvert$)	ρ	# Fail			# Opt			Mean Optimality Gap (%)			Median Time (s)		
				$\lvert\Omega\rvert$=100	$\lvert\Omega\rvert$=250	$\lvert\Omega\rvert$=500	$\lvert\Omega\rvert$=100	$\lvert\Omega\rvert$=250	$\lvert\Omega\rvert$=500	$\lvert\Omega\rvert$=100	$\lvert\Omega\rvert$=250	$\lvert\Omega\rvert$=500	$\lvert\Omega\rvert$=100	$\lvert\Omega\rvert$=250	$\lvert\Omega\rvert$=500
			$\mathrm{CVaR}_{95\%}$	0	0	0	4	5	4	3.16	1.18	2.51	4690	1450	5057
			ess-sup	0	0	0	5	5	4	0.63	1.3	1.8	1083	1086	4091
		5	Exp	0	0	0	3	3	3	4.35	5.43	5.44	7927	8285	8293
			$\mathrm{CVaR}_{90\%}$	0	0	0	5	5	4	1.13	1.89	2.2	1452	1097	5061
			$\mathrm{CVaR}_{95\%}$	0	0	0	4	5	5	2.63	1.94	1.38	4729	1090	1462
			ess-sup	0	0	0	5	5	5	1.6	0.81	0.98	1084	727	734
	60	4	Exp	0	0	0	0	0	0	15.75	15.23	16.33	TL	TL	TL
			$\mathrm{CVaR}_{90\%}$	0	0	0	3	2	2	7.23	10.19	8.04	8501	TL	TL
			$\mathrm{CVaR}_{95\%}$	0	0	0	3	1	3	6.72	9.99	7.21	8489	TL	8524
			ess-sup	0	0	0	2	2	4	8.83	9.58	5.77	TL	TL	5012
		5	Exp	0	0	0	0	1	2	6.51	6.65	6.43	TL	TL	TL
			$\mathrm{CVaR}_{90\%}$	0	0	0	3	5	4	4.4	2.75	3.89	7564	1141	4873
			$\mathrm{CVaR}_{95\%}$	0	0	0	4	5	5	3.78	3.45	3	4018	1706	1310
			ess-sup	0	0	0	5	4	4	1.29	3.55	1.28	1610	4561	4944
	80	4	Exp	0	0	0	1	1	1	10.78	11.46	10.73	TL	TL	TL
			$\mathrm{CVaR}_{90\%}$	0	0	0	2	1	2	6.23	6.39	6.14	TL	TL	TL
			$\mathrm{CVaR}_{95\%}$	0	0	0	2	1	5	6.51	5.64	3.72	TL	TL	2149
			ess-sup	0	0	0	3	5	3	3.72	1.63	4.82	8532	1931	8652
		5	Exp	0	0	0	3	2	3	5.33	5.86	5.7	8403	TL	8599
			$\mathrm{CVaR}_{90\%}$	0	0	0	4	3	4	3.37	7.59	5.59	5286	8476	5264
			$\mathrm{CVaR}_{95\%}$	0	0	0	3	3	3	3.64	5.17	3.83	8044	8411	8491
			ess-sup	0	0	0	5	4	2	2.32	2.95	5.35	1565	4930	TL
Large	100	4	Exp	0	0	0	0	0	0	13.68	10.42	7.88	TL	TL	TL
			$\mathrm{CVaR}_{90\%}$	0	0	0	1	2	0	9.12	7.34	12.98	TL	TL	TL
			$\mathrm{CVaR}_{95\%}$	0	0	0	1	4	0	9	5.04	10.2	TL	6916	TL
			ess-sup	0	0	0	3	4	3	6.38	4.64	6.67	9164	6851	9636

Continued on next page

Table F.7: Three-level decomposition results summary (continued).

Size	Patients ($\|\mathcal{P}\|$)	Stages ($\|\mathcal{S}\|$)	ρ	# Fail			# Opt			Mean Optimality Gap (%)			Median Time (s)		
				$\|\Omega\|$=100	$\|\Omega\|$=250	$\|\Omega\|$=500	$\|\Omega\|$=100	$\|\Omega\|$=250	$\|\Omega\|$=500	$\|\Omega\|$=100	$\|\Omega\|$=250	$\|\Omega\|$=500	$\|\Omega\|$=100	$\|\Omega\|$=250	$\|\Omega\|$=500
		5	Exp	0	0	0	0	0	0	10.97	10.73	12.58	TL	TL	TL
			$\text{CVaR}_{90\%}$	0	0	0	0	1	0	8.53	6.45	7.18	TL	TL	TL
			$\text{CVaR}_{95\%}$	0	0	0	0	1	1	11.67	6.42	7.66	TL	TL	TL
			ess-sup	0	0	0	2	0	2	6.49	8.27	8.05	TL	TL	TL
	120	4	Exp	0	0	0	0	0	1	12.7	12.67	12.34	TL	TL	TL
			$\text{CVaR}_{90\%}$	0	0	0	1	2	1	12.5	9.62	9.92	TL	TL	TL
			$\text{CVaR}_{95\%}$	0	0	0	1	1	1	12.9	9.65	8.87	TL	TL	TL
			ess-sup	0	0	0	1	1	1	11.08	9.6	10.38	TL	TL	TL
		5	Exp	0	0	0	0	0	0	18.02	15.95	15.55	TL	TL	TL
			$\text{CVaR}_{90\%}$	0	0	0	1	1	1	9.01	9.39	10.84	TL	TL	TL
			$\text{CVaR}_{95\%}$	0	0	0	2	2	2	8.94	8.7	8.02	TL	TL	TL
			ess-sup	0	0	0	1	2	3	8.17	8.17	9.19	TL	TL	8991
	140	4	Exp	0	0	0	0	0	0	19.47	20.45	24.45	TL	TL	TL
			$\text{CVaR}_{90\%}$	0	0	0	0	0	0	18.42	18.52	19.18	TL	TL	TL
			$\text{CVaR}_{95\%}$	0	0	0	0	0	0	17.83	18.32	15.68	TL	TL	TL
			ess-sup	0	0	0	0	1	0	17.54	18.69	19.6	TL	TL	TL
		5	Exp	0	0	0	0	0	0	12.76	12.88	13.43	TL	TL	TL
			$\text{CVaR}_{90\%}$	0	0	0	1	2	1	9.56	7.25	9.36	TL	TL	TL
			$\text{CVaR}_{95\%}$	0	0	0	3	1	3	6.9	9.14	6.54	9555	TL	9591
			ess-sup	0	0	0	2	2	1	8.29	6.42	6.95	TL	TL	TL

References

S. Agostinelli, F. Covino, G. D'Agnese, C. De Crea, F. Leotta, and A. Marrella. Supporting governance in healthcare through process mining: A case study. *IEEE Access*, 8:186012–186025, 2020.

N. Bhattacharyya. Benchmarks for the durations of ambulatory surgical procedures in otolaryngology. *Annals of Otology, Rhinology & Laryngology*, 120(11):727–731, 2011. doi: 10.1177/000348941112001106.

CIHI. Data quality documentation, national ambulatory care reporting system - multi-year information. Technical report, Canadian Institute for Health Information, Ottawa, ON, 2012. Available at https://www.cihi.ca/sites/default/files/nacrs_multi-year_info_en_0.pdf (last accessed date: July 3, 2026).

F. Dexter, R. H. Epstein, E. Marcon, and R. de Matta. Strategies to reduce delays in admission into a postanesthesia care unit from operating rooms. *Journal of PeriAnesthesia Nursing*, 20(2):92–102, April 2005.

S. J. Head, R. Seib, J. A. Osborn, and S. K. W. Schwarz. A "swing room" model based on regional anesthesia reduces turnover time and increases case throughput. *Canadian Journal of Anesthesia*, 58(8):725–732, August 2011.

M. Heyba, F. Qasem, A. S. Ibrahim, T. Habib, H. Akl, and S. M. Al-Matouq. Improving postanesthesia care unit (PACU) delays: A quality improvement project. *Journal of PeriAnesthesia Nursing*, 39(5):716–721, October 2024.

G. Johnson and T. Calixte. Leaning on postanesthesia care unit hold times. *Journal of PeriAnesthesia Nursing*, 38(4):e4, 2023.

E. Kayıcs, T. T. Khaniyev, J. Suermondt, and K. Sylvester. A robust estimation model for surgery durations with temporal, operational, and surgery team effects. *Health Care Management Science*, 18(3):222–233, September 2015.

G. Leeftink and E. W. Hans. Case mix classification and a benchmark set for surgery scheduling. *Journal of Scheduling*, 21(1):17–33, February 2018.

R. B. Nelsen. *An introduction to copulas*. Springer Series in Statistics. Springer, New York, NY, 2nd edition, 2006.

D. J. Pavlin, S. E. Rapp, N. L. Polissar, J. A. Malmgren, M. Koerschgen, and H. Keyes. Factors affecting discharge time in adult outpatients. *Anesthesia & Analgesia*, 87(4):816–826, 1998. doi: 10.1097/00000539-199810000-00014.

V. Roshanaei, C. Luong, D. M. Aleman, and D. R. Urbach. Collaborative operating room planning and scheduling. *INFORMS Journal on Computing*, 29(3):558–580, 2017.

K. S. Shehadeh and R. Padman. Stochastic optimization approaches for elective surgery scheduling with downstream capacity constraints: Models, challenges, and opportunities. *Computers & Operations Research*, 137:105523, 2022.

D. P. Strum, J. H. May, and L. G. Vargas. Modeling the uncertainty of surgical procedure times: Comparison of log-normal and normal models. *Anesthesiology*, 92(4):1160–1167, April 2000.

S. K. Taylor, J. C. Andrzejowski, M. D. Wiles, S. Bland, G. L. Jones, and S. C. Radley. A prospective observational study of the impact of an electronic questionnaire (ePAQ-PO) on the duration of nurse-led pre-operative assessment and patient satisfaction. *PLoS One*, 13(10):e0205439, 2018.

UHN. Report to our community 2022–2023. Organizational annual report, 2023. Available at https://www.uhn.ca/corporate/AboutUHN/Documents/report-to-our-community-2022-23.pdf (last accessed date: July 23, 2026).

J. Wang, J. Cabrera, K. L. Tsui, H. Guo, M. Bakker, and J. B. Kostis. Clinical and nonclinical effects on operative duration: Evidence from a database on thoracic surgery. *Journal of Healthcare Engineering*, 2020:3582796, February 2020.

C. Weissman, J. Scemama, and Y. G. Weiss. The ratio of PACU length-of-stay to surgical duration: Practical observations. *Acta Anaesthesiologica Scandinavica*, 63(9):1143–1151, 2019.

P. F. White, S. Rawal, J. Nguyen, and A. Watkins. PACU fast-tracking: An alternative to “bypassing” the PACU for facilitating the recovery process after ambulatory surgery. *Journal of PeriAnesthesia Nursing*, 18(4):247–253, 2003.