Coordinated and Priority-based Surgical Care:
An Integrated Distributionally Robust Stochastic Optimization Approach

Appendix

Appendix A: Omitted Proofs of the Analytical Results

A.1. Proof of Proposition 1

When the probability measure $P \in \Phi(\boldsymbol{\mu}, \boldsymbol{\sigma}, \Theta)$ is defined by the polyhedral support set (15), we can explicitly rewrite the moment problem (18) under each individual scenario $s \in \mathcal{S}$ as follows:

$$\max_{P \in \Phi(\boldsymbol{\mu}, \boldsymbol{\sigma}, \Theta)} \mathbb{E}_P \Big[f_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{d}) \Big] = \max_{P} \left\{ \int_{\Theta} f_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{d}) \, dP(d) \right\} \quad (27)$$
$$\text{s.t.} \quad \text{constraints } (16a) - (16c),$$

as a linear program maximizing over all plausible distributions P in the set $\Phi(\boldsymbol{\mu}, \boldsymbol{\sigma}, \Theta)$, and with the expectation of the function $f_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{d})$ taken over this distribution as the objective function. Since all $\boldsymbol{y_s}$, $\widehat{\boldsymbol{y}}$, and $\boldsymbol{d}$ vectors are input parameters to the moment problem (18), and it contains the continuous decision variable P and linear constraints (16a)-(16c), we can take the dual of linear program (27) by associating the dual variable vectors $\delta_s \in \mathbb{R}$, $\boldsymbol{\alpha_s} \in \mathbb{R}^{|\Gamma| \times |\mathcal{K}|}$ and $\boldsymbol{\beta_s} \in \mathbb{R}^{|\Gamma| \times |\mathcal{K}|}$ with the constraints (16a)-(16c), respectively, for each individual scenario $s \in \mathcal{S}$. Thus, its dual is a *semi-infinite linear* program as follows:

$$\min_{\boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\beta}} \ \delta_s + \sum_{\gamma \in \Gamma} \sum_{k \in \mathcal{K}} \mu_{\gamma,k} \, \alpha^k_{\gamma,s} + \sum_{\gamma \in \Gamma} \sum_{k \in \mathcal{K}} (\mu^2_{\gamma} + \sigma^2_{\gamma,k}) \, \beta^k_{\gamma,s} \quad (28a)$$

$$\text{s.t.} \quad \delta_s + \sum_{\gamma \in \Gamma} \sum_{k \in \mathcal{K}} d_{\gamma,k} \, \alpha^k_{\gamma,s} + \sum_{\gamma \in \Gamma} \sum_{k \in \mathcal{K}} d^2_{\gamma,k} \, \beta^k_{\gamma,s} \ \geq \ f_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{d}), \ \forall \boldsymbol{d} \in \Theta. \quad (28b)$$

Using the strong duality theorem, we next substitute the inner maximization moment problem (18) with (28a)-(28b) in the min-max IMSDRO model (17a)-(17b), and merge the minimization objective (28a) with the minimization objective in the IMSDRO model (17a)-(17b) to obtain a reformulation of the min-max IMSDRO model (17a)-(17b) under the ambiguity set $\Phi(\boldsymbol{\mu}, \boldsymbol{\sigma}, \Theta)$ as

$$\min_{\boldsymbol{x}, \boldsymbol{y}, \widehat{\boldsymbol{y}}, \boldsymbol{q}, \boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\beta}} \ \sum_{s \in \mathcal{S}} \pi_s \left\{ \sum_{k \in \mathcal{K}} \sum_{m \in \mathcal{L}} q^k_{m,s} + \delta_s + \sum_{\gamma \in \Gamma} \sum_{k \in \mathcal{K}} \mu_{\gamma,k} \, \alpha^k_{\gamma,s} + \sum_{\gamma \in \Gamma} \sum_{k \in \mathcal{K}} (\mu^2_{\gamma,k} + \sigma^2_{\gamma,k}) \, \beta^k_{\gamma,s} \right\} \quad (29a)$$

$$\text{s.t.} \ \delta_s + \sum_{\gamma \in \Gamma} \sum_{k \in \mathcal{K}} d_{\gamma,k} \, \alpha^k_{\gamma,s} + \sum_{\gamma \in \Gamma} \sum_{k \in \mathcal{K}} d^2_{\gamma,k} \, \beta^k_{\gamma,s} \ \geq \ f_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{d}), \ \forall \boldsymbol{d} \in \Theta, \, s \in \mathcal{S} \quad (29b)$$

$$(\boldsymbol{x_s}, \, \boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \, \boldsymbol{q_s}) \in \mathcal{R}_s, \ \forall s \in \mathcal{S} \quad (29c)$$

$$\delta_s \in \mathbb{R}, \, \boldsymbol{\alpha_s}, \, \boldsymbol{\beta_s} \in \mathbb{R}^{|\Gamma| \times |\mathcal{K}|}, \ \forall s \in \mathcal{S}. \quad (29d)$$

However, the reformulation (29a)-(29c) of the IMSDRO model is still *intractable* since constraint (29b) is a semi-infinite constraint, meaning that it should be satisfied for any possible realization of d from the polyhedral support set Θ in (15). To obtain a tractable reformulation, since constraint (29b) should be satisfied for all realization of $\boldsymbol{d} \in \Theta$, it should be satisfied for the worst-case possible value of $\boldsymbol{d} \in \Theta$. Hence, we move all the terms which contain $\boldsymbol{d}$ to the right-hand side of constraint (29b) to obtain the minimization reformulation (19a)-(19d), which completes the proof.

Q.E.D.

A.2. Proof of Proposition 2

For each individual scenario $s \in \mathcal{S}$, we can derive the following simple equivalent linear program (LP) for the function $f_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{d})$ using the surgical overtime definition for surgeons as follows:

$$f_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{d}) = \min_{\boldsymbol{o}} \sum_{k\in\mathcal{K}}\sum_{n\in\mathcal{L}} o^k_{n,s} \tag{30a}$$

$$\text{s.t. } o^k_{n,s} \geq \sum_{\gamma\in\Gamma} d_{\gamma,k}\Big(\sum_{t\in\mathcal{U}}\sum_{m\in\mathcal{T}\setminus\{t_0\}} \widetilde{y}^{k,n}_{\gamma,t,m} + \sum_{t\in\mathcal{T}}\sum_{m\in\mathcal{T}\setminus\{t_0\}} y^{k,n}_{\gamma,t,m,s} + \sum_{t\in\mathcal{U}\cup\{t_0\}} \widehat{y}^{k,n}_{\gamma,t,t_0}\Big) - V^k_n, \forall k\in\mathcal{K}, n\in\mathcal{L} \tag{30b}$$

$$o^k_{n,s} \geq 0,\ \forall k\in\mathcal{K}, n\in\mathcal{L}. \tag{30c}$$

Next, we take the dual formulation of the LP (30a)-(30c) by associating dual variables $\lambda^k_{n,s} \in \mathbb{R}_+$ to the constraints (30b) as follows for each individual scenario $s \in \mathcal{S}$:

$$f_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{d}) = \max_{\lambda_s} \sum_{k\in\mathcal{K}}\sum_{n\in\mathcal{L}} \Big\{\sum_{\gamma\in\Gamma} d_{\gamma,k}\Big(\sum_{t\in\mathcal{U}}\sum_{m\in\mathcal{T}\setminus\{t_0\}} \widetilde{y}^{k,n}_{\gamma,t,m} + \sum_{t\in\mathcal{T}}\sum_{m\in\mathcal{T}\setminus\{t_0\}} y^{k,n}_{\gamma,t,m,s} + \sum_{t\in\mathcal{U}\cup\{t_0\}} \widehat{y}^{k,n}_{\gamma,t,t_0}\Big) - V^k_n\Big\}\lambda^k_{n,s} \tag{31a}$$

$$\text{s.t. } 0 \leq \lambda^k_{n,s} \leq 1,\ \forall k\in\mathcal{K}, n\in\mathcal{L}. \tag{31b}$$

We then substitute the dual problem (31a)-(31b) into the maximization problem $\Psi_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{\alpha_s}, \boldsymbol{\beta_s})$ on the right hand side of the constraints (19b), which results in the following equivalent problem:

$$\begin{aligned}\Psi_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{\alpha_s}, \boldsymbol{\beta_s}) = \max_{d\in\Theta}\Big\{ \max_{\lambda_s\in\Lambda_s} \sum_{k\in\mathcal{K}}\sum_{n\in\mathcal{L}} \Big\{\sum_{\gamma\in\Gamma} d_{\gamma,k}\Big(\sum_{t\in\mathcal{U}}\sum_{m\in\mathcal{T}\setminus\{t_0\}} \widetilde{y}^{k,n}_{\gamma,t,m} + \sum_{t\in\mathcal{T}}\sum_{m\in\mathcal{T}\setminus\{t_0\}} y^{k,n}_{\gamma,t,m,s} \\ + \sum_{t\in\mathcal{U}\cup\{t_0\}} \widehat{y}^{k,n}_{\gamma,t,t_0}\Big) - V^k_n\Big\}\cdot\lambda^k_{n,s} - \sum_{\gamma\in\Gamma}\sum_{k\in\mathcal{K}} d_{\gamma,k}\,\alpha^k_{\gamma,s} - \sum_{\gamma\in\Gamma}\sum_{k\in\mathcal{K}} d^2_{\gamma,k}\,\beta^k_{\gamma,s}\Big\}.\end{aligned} \tag{32}$$

Swapping the order of maximizations in (32) does not affect the optimal solution because the polyhedron-shaped support set Θ of $\boldsymbol{d}$ is a *compact and bounded set*. Thus, as we have a separable structure for the polyhedron-shaped support set Θ of $\boldsymbol{d}$, we can maximize the objective function (32) first over $\lambda_s \in \Lambda_s$, and then over $\boldsymbol{d} \in \Theta$, separately as follows:

$$\begin{aligned}\Psi_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{\alpha_s}, \boldsymbol{\beta_s}) = \max_{\lambda_s\in\Lambda_s} \Big\{ \max_{d\in\Theta} \Big\{ \sum_{\gamma\in\Gamma}\sum_{k\in\mathcal{K}} \Big(\sum_{n\in\mathcal{L}} \Big\{ \sum_{t\in\mathcal{U}}\sum_{m\in\mathcal{T}\setminus\{t_0\}} \widetilde{y}^{k,n}_{\gamma,t,m} + \sum_{t\in\mathcal{T}}\sum_{m\in\mathcal{T}\setminus\{t_0\}} y^{k,n}_{\gamma,t,m,s} \\ + \sum_{t\in\mathcal{U}\cup\{t_0\}} \widehat{y}^{k,n}_{\gamma,t,t_0}\Big\}\cdot\lambda^k_{n,s} d_{\gamma,k} - \alpha^k_{\gamma,s} d_{\gamma,k} - \beta^k_{\gamma,s} d^2_{\gamma,k}\Big)\Big\} - \sum_{k\in\mathcal{K}}\sum_{n\in\mathcal{L}} V^k_n \lambda^k_{n,s}\Big\}.\end{aligned} \tag{33}$$

Next, given the fact that the polyhedron-shaped support set Θ of $\boldsymbol{d}$ is defined by independent lower and upper bounds in each dimension of $\gamma \in \Gamma$ and $k \in \mathcal{K}$ pair (see (13)), the inner maximization problem over $\boldsymbol{d} \in \Theta$ in the problem (33) is a separable optimization problem by the patient class $\gamma \in \Gamma$ and surgeon $k \in \mathcal{K}$ indices. We can separate this inner max problem in the problem (33) into $|\Gamma| \times |\mathcal{K}|$ maximization problems, each of them is over the interval $d^{LB}_{\gamma,k} \leq d_{\gamma,k} \leq d^{UB}_{\gamma,k}$, and make a summation over the indices $\gamma \in \Gamma$ and $k \in \mathcal{K}$, which results in the optimization problem (20) and completes the proof. **Q.E.D.**

A.3. Proof of Theorem 1

For each pair of patient class $\gamma \in \Gamma$ and surgeon $k \in \mathcal{K}$, we define binary variables $\eta^k_{\gamma,i} \in \{0,1\}$ corresponding to each segment point $\tilde{d}_{\gamma,k}(i)$, $i = 0, \cdots, H$ such that $\eta^k_{\gamma,i} = 1$ if the i^{th} segment point $\tilde{d}_{\gamma,k}(i)$ in the set $\Upsilon_{\gamma,k} = \{\tilde{d}_{\gamma,k}(i)\}_{i=0}^{H}$ yields the maximum value for the problem (22) and $\eta^k_{\gamma,i} = 0$ otherwise. Consequently, the approximation problem (22) for the given $\boldsymbol{y}, \boldsymbol{\alpha}$, and $\boldsymbol{\beta}$ decisions is equivalent to the following problem for each pair of patient class $\gamma \in \Gamma$ and surgeon $k \in \mathcal{K}$ under scenario $s \in \mathcal{S}$:

$$\max_{\boldsymbol{\eta}} \Bigg\{ \Bigg(\sum_{i=0}^{H} \Bigg(\sum_{n\in\mathcal{L}} \Big\{ \sum_{t\in\mathcal{U}} \sum_{m\in\mathcal{T}\setminus\{t_0\}} \widetilde{y}^{k,n}_{\gamma,t,m} + \sum_{t\in\mathcal{T}} \sum_{m\in\mathcal{T}\setminus\{t_0\}} y^{k,n}_{\gamma,t,m,s} + \sum_{t\in\mathcal{U}\cup\{t_0\}} \widehat{y}^{k,n}_{\gamma,t,t_0} \Big\} \lambda^k_{n,s} \Bigg) \tilde{d}_{\gamma,k}(i) - \alpha^k_{\gamma,s}\, \tilde{d}_{\gamma,k}(i) - \beta^k_{\gamma,s}\, \tilde{d}_{\gamma,k}(i)^2 \Bigg) \eta^k_{\gamma,i} \Bigg\} \tag{34a}$$

$$\text{s.t.} \sum_{i=0}^{H} \eta^k_{\gamma,i} = 1, \tag{34b}$$

$$\eta^k_{\gamma,i} \in \{0,1\}, \ \forall i = 0, \cdots, H. \tag{34c}$$

To ensure that exactly one of the segment points $\tilde{d}_{\gamma,k}(i)$, $i \in \{0, \cdots, H\}$ is selected for each pair (γ, k) to maximize the objective function of problem (22), we require the constraints (34b). Note that each set of segment points $\Upsilon_{\gamma,k} = \{\tilde{d}_{\gamma,k}(i)\}_{i=0}^{H}$ for each pair of class γ and surgeon k is a Specially Ordered Set of Type 1 (SOS1), containing binary variables that sum to one (see constraints (34b)). When the segment point $\tilde{d}_{\gamma,k}(i')$ maximizes (22), i.e., $\eta^k_{\gamma,i'} = 1$, the objective function of (34a) gets the value of $\sum_{i=0}^{H} \Big(\sum_{n\in\mathcal{L}} \Big\{ \sum_{t\in\mathcal{U}} \sum_{m\in\mathcal{T}\setminus\{t_0\}} \widetilde{y}^{k,n}_{\gamma,t,m} + \sum_{t\in\mathcal{T}} \sum_{m\in\mathcal{T}\setminus\{t_0\}} y^{k,n}_{\gamma,t,m,s} + \sum_{t\in\mathcal{U}\cup\{t_0\}} \widehat{y}^{k,n}_{\gamma,t,t_0} \Big\} \lambda^k_{n,s}\, \tilde{d}_{\gamma,k}(i) - \alpha^k_{\gamma,s}\, \tilde{d}_{\gamma,k}(i) - \beta^k_{\gamma,s}\, \tilde{d}_{\gamma,k}(i)^2 \Big)$ and other $\eta^k_{\gamma,i}$ variables are zero for $i \in \{0, \cdots, H\}$, $i \neq i'$.

Furthermore, we see that there is a bi-linear expression $\lambda^k_{n,s}\eta^k_{\gamma,i}$ in the objective function (34a) for the given $\boldsymbol{y}, \widehat{\boldsymbol{y}}, \boldsymbol{\alpha}$, and $\boldsymbol{\beta}$ decisions. However, we can reformulate these bi-linear terms by adding some *McCormick type inequalities* because $\eta^k_{\gamma,i}$ is a binary variable and we have lower and upper bounds, i.e., $0 \leq \lambda^k_{n,s} \leq 1$, for the variable $\lambda^k_{n,s}$ based on the polyhedron set Λ_s defined in Proposition 2. To do so, we define auxiliary variables $\tau^k_{n,s,\gamma,i}$ such that $\tau^k_{n,s,\gamma,i} = \lambda^k_{n,s}\, \eta^k_{\gamma,i}$ for all $i \in \{0, \cdots, H\}$, and $\gamma \in \Gamma$. To remove theses bi-linear terms in objective function (34a), we need to add the McCormick type constraints (23c)-(23e), and $\tau^k_{n,s,\gamma,i} \geq 0$, which guarantee that $\tau^k_{n,s,\gamma,i} = \lambda^k_{n,s}\, \eta^k_{\gamma,i}$ for all $i \in \{0, \cdots, H\}$, $k \in \mathcal{K}$ and $\gamma \in \Gamma$. More precisely, when the binary variable $\eta^k_{\gamma,i} = 1$, constraints (23c)-(23d) make sure that $0 \leq \lambda^k_{n,s} \leq 1$, and when $\eta^k_{\gamma,i} = 0$, constraints (23e) and $\tau^k_{n,s,\gamma,i} \geq 0$ guarantee that $\lambda^k_{n,s} = 0$. Thus, there exists an equivalence relation between the bi-linear term $\tau^k_{n,s,\gamma,i} = \lambda^k_{n,s}\, \eta^k_{\gamma,i}$ and constraints (23c)-(23d), and $\tau^k_{n,s,\gamma,i} \geq 0$.

Next, if we do the change of variables $\tau^k_{n,s,\gamma,i} = \lambda^k_{n,s}\, \eta^k_{\gamma,i}$ in the objective function (34a), and replace it in the objective function (20), we obtain the objective function $\chi_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{\alpha_s}, \boldsymbol{\beta_s}; \boldsymbol{\tau_s}, \boldsymbol{\eta_s}, \boldsymbol{\lambda_s})$ is defined for each scenario $s \in \mathcal{S}$ defined by (24). Therefore, using the approximation of (21) with the problem (34a)-(34c) and the McCormick type constraints (23c)-(23e), and $\tau^k_{n,s,\gamma,i} \geq 0$, we are able to approximate the optimal objective function value of the maximization problem $\Psi_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{\alpha_s}, \boldsymbol{\beta_s})$ in (20) by the MILP (23a)-(23f).

Q.E.D.

A.4. Proof of Theorem 2

We need to prove two things, which include that (i) the scenario cuts derived by the scenario cut-generating problem $\widetilde{\Psi}_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{\alpha_s}, \boldsymbol{\beta_s})$ or (23a)-(23f) for solving the IMSDRO-APRX model are *valid*, and (ii) *finitely many* scenario cuts suffice to reach a feasible solution that satisfies constraints (25b).

(i) For any value of $(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{\alpha_s}, \boldsymbol{\beta_s})$ satisfying constraints (26c) and (26d), the optimization problem $\Psi_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{\alpha_s}, \boldsymbol{\beta_s})$ can be approximated by $\widetilde{\Psi}_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{\alpha_s}, \boldsymbol{\beta_s})$ according to the result of Theorem 1, and so the scenario cuts (26b) are valid.

(ii) The scenario cut-generating problem $\widetilde{\Psi}_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{\alpha_s}, \boldsymbol{\beta_s})$ is not dependent on the values of $\boldsymbol{y_s}$, $\widehat{\boldsymbol{y}}$, $\boldsymbol{\alpha_s}$, $\boldsymbol{\beta_s}$. The number of binary variables of the problem $\widetilde{\Psi}_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{\alpha_s}, \boldsymbol{\beta_s})$ is limited to $|H| \times |\Gamma| \times |\mathcal{K}|$, and for any value for binary variables that satisfy $\sum_{i=0}^{H} \eta_{\gamma,i}^{k} = 1$, the feasible region of this optimization problem is a polyhedron with finite extreme points. Therefore, the maximum number of scenario cuts corresponding to the extreme points of the feasible region of $\widetilde{\Psi}_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{\alpha_s}, \boldsymbol{\beta_s})$ for any value of binary variables are limited.

Therefore, the constraint generation Algorithm 1 terminates after finite number of iterations.

Q.E.D.

Appendix B: Scenario Tree and Ambiguity Set Construction Approach

One often starts from a full description of stochastic random variable (the number of patient referrals in the CAS problem). Solving a stochastic model with such a full description of a stochastic random variable is however next to impossible. We therefore need a *scenario construction algorithm* to translate the full representation of the stochastic variable into a set of discrete realizations (i.e., scenarios) of that stochastic variable. To adequately represent the appointment request stochastic process, we need to generate a sufficient number of scenarios; that is, the set of scenarios needs to cover the most plausible realizations of the stochastic process. This often requires a very large number of scenarios, typically generated by a scenario construction algorithm. However, the computational burden of solving such a stochastic model with a large number of scenarios is extremely high and for practical purposes often impossible. To avoid such intractability, a *scenario reduction algorithm* is then deployed so as to reduce the cardinality of the set of scenarios. In general, the goal of a scenario construction algorithm is to minimize the error caused by the approximation of the true stochastic process with a scenario tree.

In this Appendix, we first explain our approach along with an example for how we generate a scenario tree for the number of patient arrivals in B.1, B.2 and B.3. In B.4, we pinpoint how a scenario generation and reduction can be evaluated. Finally, in B.5, we explain our method for generating an ambiguity set for the surgery duration.

B.1. Scenario Reduction Heuristics

We consider a T-dimensional stochastic process $\boldsymbol{\xi} = \{\xi_t\}_{t=1}^{T}$ with a distribution function F and a finite support for the number of patient appointment requests over T days. This finite support is presented by S discrete scenarios through $supp(\boldsymbol{\xi}) = \{\xi^{(1)}, \xi^{(2)}, \cdots, \xi^{(S)}\}$ where $\xi^{(s)} = \{\xi_t^{(s)}\}_{t=1}^{T}$ for $s \in \mathcal{S} = \{1, \cdots, S\}$. The corresponding scenario probability is denoted by π_s and $\sum_{s=1}^{S} \pi_s = 1$. Assume P is the distribution function of another T-dimensional stochastic process $\tilde{\boldsymbol{\xi}} = \{\tilde{\xi}_t\}_{t=1}^{T}$. Let $supp(\tilde{\boldsymbol{\xi}}) = \{\tilde{\xi}^{(1)}, \tilde{\xi}^{(2)}, \cdots, \tilde{\xi}^{(S')}\}$ where

$\tilde{\xi}^{(s')} = \{\tilde{\xi}_t^{(s')}\}_{t=1}^T$, and S' be the number of discrete scenarios with corresponding scenario probabilities $\tilde{\pi}_{s'}$ for $s' \in \mathcal{S}' = \{1, \cdots, S'\}$ and $\sum_{s'=1}^{S'} \tilde{\pi}_{s'} = 1$.

The *Kantorovich distance* $D_T(F, P)$ between the above-mentioned stochastic processes F and P is the optimal solution of the following linear transportation problem:

$$D_T(F,P) = \min_{\rho} \Bigg\{ \sum_{s=1}^{S} \sum_{s'=1}^{S'} \rho_{s,s'} \cdot d_{|\mathcal{T}|}(\xi^{(s)}, \tilde{\xi}^{(s')}), \\ s.t. \ \sum_{s=1}^{S} \rho_{s,s'} = \pi_{s'}, \ \sum_{s'=1}^{S'} \rho_{s,s'} = \tilde{\pi}_s, \ \rho_{s,s'} \geq 0, \ \forall s \in \mathcal{S}, \forall s' \in \mathcal{S}' \Bigg\}, \tag{35}$$

where $d_t(\xi^{(s)}, \tilde{\xi}^{(s')}) = \sum_{v=1}^{t} || \ \xi_v^{(s)} - \tilde{\xi}_v^{(s')} ||$, $t \in \mathcal{T} = \{1, \cdots, T\}$, and $||.||$ is a norm function over $\mathbb{R}^T$. Thus, $d_{|\mathcal{T}|}(\xi^{(s)}, \tilde{\xi}^{(s')})$ is total distance between scenarios s and s'.

If we assume that P is a *reduced* distribution function of F, its discrete support then includes scenarios $\tilde{\xi}^{(s')} = \{\tilde{\xi}_t^{(s')}\}_{t=1}^T$, $s' \in \{1, \cdots, S\} \backslash del(\mathcal{S})$ where $del(\mathcal{S})$ is the set of deleted scenarios from the original scenario set $\mathcal{S}$. For a pre-specified set $del(\mathcal{S}) \subset \mathcal{S}$, the Kantorovich distance between stochastic processes F and P can be calculated by the following expression:

$$D_T(F,P) = \sum_{s \in \ del(\mathcal{S})} \pi_s \cdot \min_{s' \notin \ del(\mathcal{S})} \Big\{ d_{|\mathcal{T}|}(\xi^{(s)}, \tilde{\xi}^{(s')}) \Big\}. \tag{36}$$

Moreover, the scenario probabilities $\tilde{\pi}_{s'}$, $s' \notin \ del(\mathcal{S})$ for the reduced set of scenarios $\{\tilde{\xi}^{(s')}\}_{s' \notin \ del(\mathcal{S})}$ are given by $\tilde{\pi}_{s'} = \pi_{s'} + \sum_{s \in del_{s'}(\mathcal{S})} \pi_s$, where $del_{s'}(\mathcal{S}) = \{s \in del(\mathcal{S}) : s' = s'(s)\}$, and also $s'(s) \in \arg\min_{s' \notin del(\mathcal{S})} d_{|\mathcal{T}|}(\xi^{(s)}, \tilde{\xi}^{(s')})$ for each scenario $s \in del(\mathcal{S})$ is a selection from the the index set of nearest scenarios to the scenario $\xi^{(s)}$ for all $s \in del(\mathcal{S})$. The optimal set $del(\mathcal{S})$ of deleted scenarios with cardinality $\kappa = |del(\mathcal{S})|$ is obtained by solving the following *scenario reduction problem*:

$$\min \Big\{ \sum_{s \in \ del(\mathcal{S})} \pi_s \cdot \min_{s' \notin \ del(\mathcal{S})} d_{|\mathcal{T}|}(\xi^{(s)}, \tilde{\xi}^{(s')}) \ \ s.t. \ del(\mathcal{S}) \subset \mathcal{S} = \{1, \cdots, S\}, \kappa = S - L \Big\}, \tag{37}$$

where $L = S - \kappa$ is the number of remaining scenarios after reduction.

Dupačová et al. (2003) proved the NP-hardness of the scenario reduction problem (37) by showing its equivalence to the set covering problem. However, this problem can be solved efficiently for two special cases of $\kappa = 1$ (i.e., deleting one scenario), and $\kappa = S - 1$ (i.e., keeping one scenario). They proposed two heuristics called backward scenario reduction and forward scenario selection algorithms for solving the reduction problem efficiently. In the *backward reduction*, optimal deletion of one scenario is recursively repeated until deleting $\kappa = S - L$ scenarios while in the *forward selection*, optimal selection of one scenario is recursively done until achieving L scenarios.

B.2. Scenario Tree Construction Approach

In §6.1, Latin Hypercube Sampling (LHS) method (Helton and Davis 2003) is used to construct a set of discrete scenarios for the corresponding multivariate stochastic parameters (number of patient requests) as a scenario fan. We then convert this scenario fan into a scenario tree, and use forward selection method to obtain an appropriate number of scenarios (Dupačová et al. 2003).

Let F be the probability distribution for a scenario fan of multivariate stochastic parameters, then each scenario $s \in \mathcal{S} = \{1, \cdots, S\}$ is presented by $\xi^{(s)} = \{\xi_0^{(s)}, \xi_1^{(s)}, \cdots, \xi_T^{(s)}\}$ with probability π_s. Since all scenarios are the same at the first node, i.e., $\xi_0^{(1)} = \xi_0^{(2)} = \cdots = \xi_0^{(S)}$ in the scenario fan, the total number of nodes is $S \times T + 1$, where $T = |\mathcal{T}|$, in the scenario fan. The goal of scenario tree construction is to generate a scenario tree with probability distribution $F_{\boldsymbol{\zeta}}$ based on the scenario fan in which the number of scenarios is reduced, and also the Kantorovich distance between F and $F_{\boldsymbol{\zeta}}$ is less than a pre-specified value $\boldsymbol{\zeta}$ (i.e., $D_T(F, F_{\boldsymbol{\zeta}}) \leq \boldsymbol{\zeta}$).

To this aim, the forward scenario reduction is used at each period $t \in \{1, \cdots, T\}$, and successive clustering of scenarios is then exploited to convert a scenario fan into a scenario tree. To construct a scenario tree with $D_T(F, F_{\boldsymbol{\zeta}}) \leq \boldsymbol{\zeta}$, at each period t, $\boldsymbol{\zeta}_t$ is considered for implementing forward scenario reduction under the criterion $\sum_{t=1}^{T} \boldsymbol{\zeta}_t \leq \boldsymbol{\zeta}$. This means that at each period t, maximal reduction strategy is applied such that $\sum_{s \in\, del(\mathcal{S})} \pi_s \cdot \min_{s' \notin\, del(\mathcal{S})} d_t(\xi^{(s)}, \tilde{\xi}^{(s')}) \leq \boldsymbol{\zeta}_t$, where the distance between two scenarios $\xi^{(s)}$ and $\tilde{\xi}^{(s')}$ is calculated by $d_t(\xi^{(s)}, \tilde{\xi}^{(s')}) = \sum_{v=1}^{t} || \xi_v^{(s)} - \tilde{\xi}_v^{(s')} ||$ at each period t, and $\tilde{\boldsymbol{\xi}}$ is the *reduced* version of $\boldsymbol{\xi}$ after implementing the reduction strategy.

Furthermore, we use $\boldsymbol{\zeta} = \boldsymbol{\zeta}_{rel} \cdot \boldsymbol{\zeta}_{max}$ where $0 < \boldsymbol{\zeta}_{rel} < 1$ is a constant parameter, which presents a scale for the amount of reduction in the scenario fan, and $\boldsymbol{\zeta}_{max}$ is the optimal distance between probability distribution of scenario fan and one of its scenarios with probability one. To generate the scenario tree, at each period t, $\boldsymbol{\zeta}_t$ is then computed by the following relation:

$$\boldsymbol{\zeta}_t = \frac{\boldsymbol{\zeta}}{T+1} \Big(\frac{1}{2} + \delta \left(1 - \frac{t}{T+1} \right) \Big), \ \forall t, \tag{38}$$

where $\delta \in [0, 1]$ is a constant parameter, which is set to one in our implementation.

B.3. Illustration of Generating a Scenario Tree for the Number of Patient Referrals

In this section, we explain the details along with an example on how we generate a scenario tree for the number of appointment referrals that will be used in the IMSDRO-APRX model as well as the MS-MIP model. For each patient class, we first fit a Poisson probability distribution over the number of appointment referrals from that patient class.

The LHS method is then used to generate a set of discrete scenarios for the corresponding multivariate stochastic parameters as a scenario fan. It is essential to efficiently reduce the number of scenarios in order to avoid computationally intractable stochastic programs. We then deploy a forward scenario tree construction heuristic to convert the scenario fan into a scenario tree, and so reduce the number of generated scenarios (see §B.2 for details). The strategy is to modify the fan of scenarios via bundling scenarios, which produces scenario trees with fewer scenarios than initial scenario fans.

The process of scenario tree construction for the number of appointment referrals in the CAS problem is illustrated by Figure 11 step by step. The number of final scenarios depends on a constant parameter $\boldsymbol{\zeta}_{rel}$ between zero and one, that represents a scale for the reduction amount compared with the scenario fan. To construct a scenario tree for our case study, an initial scenario fan with 100 scenarios (the most left tree in Figure 11) is generated for the stochastic parameters over an arrival horizon of $T = |\mathcal{T}| = 5$ periods (note that in our case study we have five business days as the arrival horizon $\mathcal{T}$), and the scenario tree construction approach is then implemented with $\boldsymbol{\zeta}_{rel} = 0.7$. Finally, a scenario tree with 14 scenarios (the

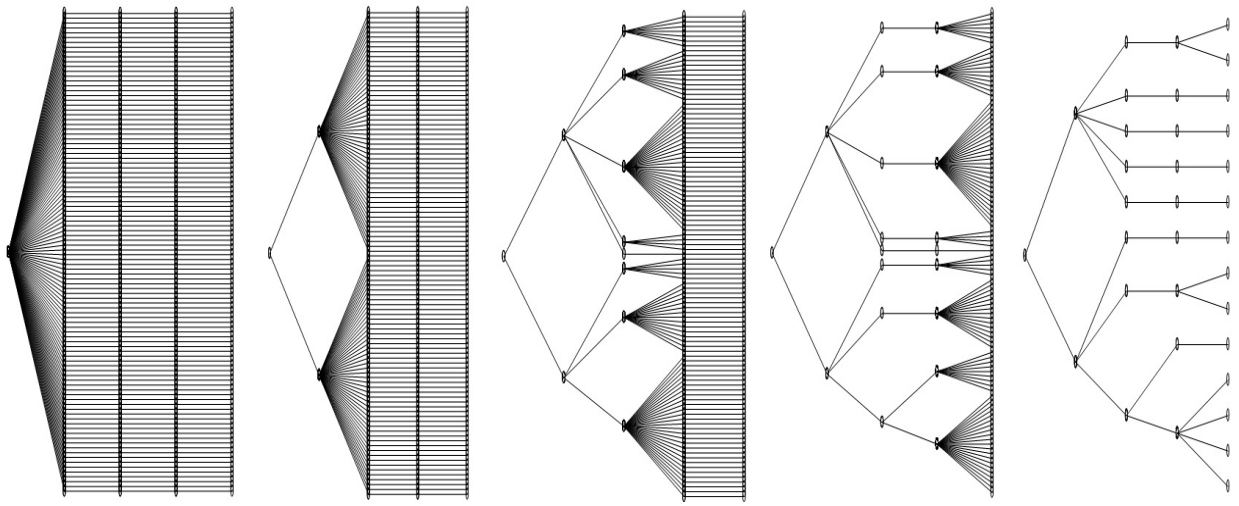


Figure 11 Illustration of scenario tree construction procedure for the number of appointment referrals in our case study over an arrival horizon of $T = |\mathcal{T}| = 5$ business days. We start with a scenario fan of 100 scenarios (the most left tree), and then turn it into a scenario tree of 14 scenarios (the most right tree).

most right tree in Figure 11) is obtained. It should be mentioned that by increasing the reduction scale ζ_{rel}, the number of obtained scenarios decreases, so the information loss increases. However, as the number of scenario decrease, we have a better computational tractability for solving the multi-stage stochastic program. Therefore, there is a trade-off between the number of scenarios and computational tractability. In §6.4, we evaluate the efficiency of this scenario construction algorithm in generating a scenario tree for the number of appointment referrals for our case study and two other random instances of our CAS problem. In particular, we find that $\zeta_{rel} = 0.7$ is a good value for the reduction scale. Refer to §6.4 for details.

B.4. Evaluation of a Scenario Construction Algorithm.

There are two main criteria in the literature of stochastic programming (Kaut and Wallace 2003) by which the efficiency of a scenario construction algorithm can be evaluated to ensure that there is not too much loss of information while constructing an *adequate* scenario tree. They include (i) *in-sample stability* and (ii) *out-of-sample stability*.

Due to the random nature of most scenario construction algorithms (such as the LHS that we used), different scenario trees will be obtained with the same input if we apply a scenario construction algorithm multiple times. Then, the *in-sample stability* guarantees that if several scenario trees are constructed with the same input, the optimal objective function values of their corresponding stochastic optimization models with these scenario trees are the same approximately. In other words, if the objective function value does not change too much, we can claim the in-sample stability. We have done this analysis in the subsection "In-sample Stability Analysis" (see Table 5) in §6.4.

Further, the *out-of-sample stability* guarantees that the objective function value obtained from implementing the scheduling policy by using our data-driven rolling-horizon procedure should be close to the optimal objective function of the stochastic model. Indeed, the out-of-sample stability ensures that the true objective value obtained from any simulation procedure (e.g., the data-driven rolling horizon algorithm in our paper)

is close to the optimal objective value of the stochastic program. This analysis is performed in the subsection "Evaluation of Objective Function Values (Overtime)" in §6.2 (see Table 3).

B.5. Ambiguity Set Generation for Surgery Durations.

We follow the procedures in the appointment scheduling literature (Denton and Gupta 2003, and Jiang et al. 2017) to generate ambiguity sets for the generation of surgery durations in the sample paths for the RHP (Algorithm 2), so that we can simulate the reality. In order to consider the distributional robustness for the surgery duration, we assume that the surgery duration can follow three classes of probability distributions: truncated normal, gamma, and log-normal, each of which can be specified by their means and standard deviations. In each CAS problem instance for the IMSDRO-APRX model, we sample realizations $(d^1_{\gamma,k}, d^2_{\gamma,k}, \cdots, d^M_{\gamma,k})$ for each class $\gamma \in \Gamma$ and surgeon $k \in \mathcal{K}$ pair. In each of M realizations for patient class γ and surgeon $k \in \mathcal{K}$ pair, we first select randomly a distribution among normal, gamma, and log-normal, and then obtain a random surgery duration from that distribution with the known mean and standard deviation.

Appendix C: Alternative Optimization Model to Balance Overtime and Access Delay

In both the problem statement in §2 and the IMSDRO formulation in §3, we have assumed that (i) all patients must obtain one clinic appointment date and (if needed) one surgery appointment date within their wait time target windows, and (ii) overtime are deployed as needed to accommodate the clinical and surgical capacities. In this appendix, we relax these assumptions by trying to balance the trade-off between patients' access delays and surgeon overtimes.

To this aim, we propose an alternative optimization model for the CAS problem in which we assume no clinical and surgical overtimes for the surgeons and they only have regular clinical and surgical capacities; however, there is a penalty for the case when we cannot meet the clinical and surgical wait time targets for patients. In particular, we incur a clinic penalty of $u^{k,m}_{p,\gamma,t,s} \geq 0$ for each class $\gamma \in \Gamma$ patient $p \in \mathcal{D}^s_{\gamma,t}$ whose request is received on day $t \in \mathcal{T}$ under scenario $s \in \mathcal{S}$ and has a clinic appointment visit on $m > t + WTS_\gamma - CSG_\gamma$ (recall that the safe range for clinic appointment visit is $m \in \big[t + WTC_\gamma, t + WTS_\gamma - CSG\gamma\big]$). Moreover, we incur a surgery penalty of $v^{k,n}_{\gamma,t,m,s} \geq 0$ for class $\gamma \in \Gamma$ patients whose requests are received on day $t \in \mathcal{U} \cup \mathcal{T}$ under $s \in \mathcal{S}$, and have clinic visit on $m \in \mathcal{T} \backslash \{t_0\}$, but surgery visit on $n > t + WTS_\gamma$ with surgeon $k \in \mathcal{K}$. We have a similar surgery penalty $e^{k,n}_{\gamma,t,t_0} \geq 0$ for class $\gamma \in \Gamma$ patients whose requests are received on day $t \in \mathcal{U} \cup \{t_0\}$, and have clinic visit on current day t_0, but surgery visit on $n > t + WTS_\gamma$ with surgeon $k \in \mathcal{K}$ (recall that the safe range for surgery appointment visit is $n \in \big[t + WTC_\gamma, t + WTS_\gamma - CSG\gamma\big]$). We summarize the new notations in Table 10. The other notations are as before (see Table 1).

Multi-stage stochastic model. With these three new penalty decisions variables, the MS-MIP model (1)-(14) is turned into the following optimization model:

$$\min \sum_{s\in\mathcal{S}} \pi_s \sum_{k\in\mathcal{K}} \sum_{\gamma\in\Gamma} \Big(\sum_{m\in\mathcal{L}} \sum_{t\in\mathcal{T}} \sum_{p\in\mathcal{D}^s_{\gamma,t}} u^{k,m}_{p,\gamma,t,s} + \sum_{m\in\mathcal{L}} \sum_{t\in\mathcal{U}\cup\mathcal{T}} \sum_{n\in\mathcal{L}} v^{k,n}_{\gamma,t,m,s} + \sum_{n\in\mathcal{L}} \sum_{t\in\mathcal{U}\cup\{t_0\}} e^{k,n}_{\gamma,t,t_0} \Big) \tag{39a}$$

$$\text{s.t. } x^{k,m}_{p,\gamma,t,s}(m - t - WTS_\gamma + CSG_\gamma) \leq u^{k,m}_{p,\gamma,t,s},\ \forall \gamma \in \Gamma, t \in \mathcal{T}, s \in \mathcal{S}, p \in \mathcal{D}^s_{\gamma,t}, k \in \mathcal{K}, m \in \mathcal{L}, \tag{39b}$$

$$y^{k,n}_{\gamma,t,m,s}(n - t - WTS_\gamma) \leq v^{k,n}_{\gamma,t,m,s},\ \forall \gamma \in \Gamma, t \in \mathcal{T} \cup \mathcal{U}, s \in \mathcal{S}, k \in \mathcal{K}, m \in \mathcal{T} \backslash \{t_0\}, \tag{39c}$$

Stage decision variables	
$u^{k,m}_{p,\gamma,t,s}$	: Clinical penalty if a class $\gamma\in\Gamma$ patient p whose request is received on day $t\in\mathcal{T}$ under scenario $s\in\mathcal{S}$, has clinic visit on day $m>t+WTS_\gamma-CSG_\gamma$ with surgeon $k\in\mathcal{K}$.
$v^{k,n}_{\gamma,t,m,s}$	: Surgical penalty if class $\gamma\in\Gamma$ patients whose requests are received on day $t\in\mathcal{U}\cup\mathcal{T}$ under $s\in\mathcal{S}$ and have clinic visit on $m\in\mathcal{T}\backslash\{t_0\}$, have surgery visit on day $n>t+WTS_\gamma$ with surgeon $k\in\mathcal{K}$.
$e^{k,n}_{\gamma,t,t_0}$	: Surgical penalty if class $\gamma\in\Gamma$ patients whose requests are received on day $t\in\mathcal{U}\cup\{t_0\}$, and have clinic visit on day t_0, have surgery visit on $n>t+WTS_\gamma$ with surgeon $k\in\mathcal{K}$.

Table 6 The description of new notations used by the MS-MIP model (39a)-(39o) of the CAS problem.

$$\widehat{y}^{k,n}_{\gamma,t,t_0}(n-t-WTS_\gamma)\le e^{k,n}_{\gamma,t,t_0},\ \forall\gamma\in\Gamma,t\in\mathcal{U}\cup\{t_0\},k\in\mathcal{K},n\in\mathcal{L}, \tag{39d}$$

$$x^{k,m}_{p,\gamma,t,s}=0,\ \forall\gamma\in\Gamma,t\in\mathcal{T},s\in\mathcal{S},p\in\mathcal{D}^s_{\gamma,t},k\in\mathcal{K},m\in\big[t_0,t+WTC_\gamma-1\big], \tag{39e}$$

$$\sum_{m=t+WTC_\gamma}^{t_e}\sum_{k\in\mathcal{K}}x^{k,m}_{p,\gamma,t,s}=1,\ \forall\gamma\in\Gamma,t\in\mathcal{T},s\in\mathcal{S},p\in\mathcal{D}^s_{\gamma,t}, \tag{39f}$$

$$r_\gamma\Big(\sum_{p\in\widetilde{\mathcal{D}}_{\gamma,t}}\widetilde{x}^{k,m}_{p,\gamma,t}\Big)\ \le\ \sum_{n=m+CSG_\gamma}^{t_e}y^{k,n}_{\gamma,t,m,s},\ \forall\gamma\in\Gamma,t\in\mathcal{U},m\in\mathcal{T}\backslash\{t_0\},k\in\mathcal{K},s\in\mathcal{S}, \tag{39g}$$

$$r_\gamma\Big(\sum_{p\in\mathcal{D}^s_{\gamma,t}}x^{k,m}_{p,\gamma,t,s}\Big)\ \le\ \sum_{n=m+CSG_\gamma}^{t_e}y^{k,n}_{\gamma,t,m,s},\ \forall\gamma\in\Gamma,t\in\mathcal{T},m\in\mathcal{T}\backslash\{t_0\},k\in\mathcal{K},s\in\mathcal{S}, \tag{39h}$$

$$\widetilde{z}^{k}_{\gamma,t}\ \le\ \sum_{n=t_0+CSG_\gamma}^{t_e}\widehat{y}^{k,n}_{\gamma,t,t_0},\ \forall\gamma\in\Gamma,t\in\mathcal{U}\cup\{t_0\},k\in\mathcal{K}, \tag{39i}$$

$$\sum_{\gamma\in\Gamma}c_\gamma\Big(\sum_{t\in\mathcal{U}}\sum_{p\in\widetilde{\mathcal{D}}_{\gamma,t}}\widetilde{x}^{k,m}_{p,\gamma,t}+\sum_{t\in\mathcal{T}}\sum_{p\in\mathcal{D}^s_{\gamma,t}}x^{k,m}_{p,\gamma,t,s}\Big)\le U^k_m,\ \forall m\in\mathcal{L},k\in\mathcal{K},s\in\mathcal{S}, \tag{39j}$$

$$\sum_{\gamma\in\Gamma}d_{\gamma,k}\Big(\sum_{t\in\mathcal{U}}\sum_{m\in\mathcal{U}}\widetilde{y}^{k,n}_{\gamma,t,m}+\sum_{t\in\mathcal{T}}\sum_{m\in\mathcal{T}\backslash\{t_0\}}y^{k,n}_{\gamma,t,m,s}+\sum_{t\in\mathcal{U}\cup\{t_0\}}\widehat{y}^{k,n}_{\gamma,t,t_0}\Big)\le V^k_n,\forall n\in\mathcal{L},k\in\mathcal{K},s\in\mathcal{S}, \tag{39k}$$

$$u^{k,m}_{p,\gamma,t,s}\ge 0,\ \forall\gamma\in\Gamma,t\in\mathcal{T},s\in\mathcal{S},p\in\mathcal{D}^s_{\gamma,t},k\in\mathcal{K},m\in\mathcal{L}, \tag{39l}$$

$$v^{k,n}_{\gamma,t,m,s}\ge 0,\ \forall\gamma\in\Gamma,t\in\mathcal{T}\cup\mathcal{U},s\in\mathcal{S},k\in\mathcal{K},m\in\mathcal{T}\backslash\{t_0\}, \tag{39m}$$

$$e^{k,n}_{\gamma,t,t_0}\ge 0,\ \forall\gamma\in\Gamma,t\in\mathcal{U}\cup\{t_0\},k\in\mathcal{K},n\in\mathcal{L}, \tag{39n}$$

$$(9)-(10),(11)-(14). \tag{39o}$$

The objective function (39a) is to minimize the expected penalties due to not meeting clinical and surgical wait time targets for patients. Constraints (39b)-(39d) along with constraints (39l)-(39n) are the related constraints for making the penalty decisions $u^{k,m}_{p,\gamma,t,s}$, $v^{k,n}_{\gamma,t,m,s}$ and $e^{k,n}_{\gamma,t,t_0}$. Constraints (39e)-(39f) and (39g)-(39i) determine the clinic and surgery appointment visits, respectively. Constraints (39j)-(39k) restricts the regular clinical and surgical capacities of surgeons on each day, respectively. Similar to the MS-MIP model (1)-(14), we assume that the surgery durations are deterministic in the above MS-MIP model (39a)-(39o).

Integrated multi-stage stochastic and distributionally robust model. To model the uncertainty in surgery durations presented by the moment-based ambiguity set $\Phi(\boldsymbol{\mu},\boldsymbol{\sigma},\Theta)$ in (16a)-(16c), we deploy the IMSDRO approach described in §3.2, to MS-MIP model (39a)-(39o). The resulting IMSDRO model is formulated as follows:

$$\bar{Z}^{IMSDRO}\ =\ \min_{\boldsymbol{x},\boldsymbol{y},\widehat{\boldsymbol{y}},\boldsymbol{u},\boldsymbol{v},\boldsymbol{e}}\ \Big\{w_1\Big(\sum_{s\in\mathcal{S}}\pi_s\sum_{k\in\mathcal{K}}\sum_{\gamma\in\Gamma}\Big(\sum_{m\in\mathcal{L}}\sum_{t\in\mathcal{T}}\sum_{p\in\mathcal{D}^s_{\gamma,t}}u^{k,m}_{p,\gamma,t,s}+\sum_{m\in\mathcal{L}}\sum_{t\in\mathcal{U}\cup\mathcal{T}}\sum_{n\in\mathcal{L}}v^{k,n}_{\gamma,t,m,s} \tag{40a}$$

$$+\sum_{n\in\mathcal{L}}\sum_{t\in\mathcal{U}\cup\{t_0\}} e^{k,n}_{\gamma,t,t_0}\Big)\Big)+w_2\Big(\sum_{s\in\mathcal{S}}\pi_s \max_{P\in\Phi(\boldsymbol{\mu},\boldsymbol{\sigma},\Theta)}\mathbb{E}_P\Big[f_s(\boldsymbol{y_s},\widehat{\boldsymbol{y}},\boldsymbol{d})\Big]\Big)\Big\},$$

$$\text{s.t. } (\boldsymbol{x_s},\ \boldsymbol{y_s},\widehat{\boldsymbol{y}},\ \boldsymbol{u_s},\ \boldsymbol{v_s},\ \boldsymbol{e})\ \in\ \mathcal{O}_s,\ \forall s\in\mathcal{S} \tag{40b}$$

where $\mathcal{O}_s$ is the feasible region defined by the constraints (39a)-(39j) and (39l)-(39o). The objective function (40a) is obtained by removing constraints (39k) from the MS-MIP model (39a)-(39o) and adding its worst-case expected value over the set of plausible surgery duration distributions $P\in\Phi(\boldsymbol{\mu},\boldsymbol{\sigma},\Theta)$ into the objective function, which results in the min-max IMSDRO model (40a)-(40b).

An important feature of IMSDRO model (40a)-(40b) compared with IMSDRO model (17a)-(17b) is that two parts of the objective functions (40a) weighted by w_1 and w_2 are in fact two *conflicting objectives*. The first part weighted by w_1 is to minimize the expected penalties incurred due to not meeting the clinical and surgical wait time targets for patients, and the second one weighted by w_2 is to minimize the maximum penalties incurred due to not satisfying the regular surgical capacities of surgeons. Indeed, it is the case that either we can meet the clinical and surgical wait time targets for patients, or we can make regular capacities for surgeons.

Following the steps of IMSDRO approach described in Propositions 1 and 2 and Theorem 1 the min-max IMSDRO model (40a)-(40b) can be approximated by the following optimization model:

$$\tilde{\tilde{Z}}^{IMSDRO}=\min_{\boldsymbol{x},\boldsymbol{y},\widehat{\boldsymbol{y}},\boldsymbol{u},\boldsymbol{v},\boldsymbol{e},\boldsymbol{\delta},\boldsymbol{\alpha},\boldsymbol{\beta}}\Big\{w_1\Big(\sum_{s\in\mathcal{S}}\pi_s\sum_{k\in\mathcal{K}}\sum_{\gamma\in\Gamma}\Big(\sum_{m\in\mathcal{L}}\sum_{t\in\mathcal{T}}\sum_{p\in\mathcal{D}^s_{\gamma,t}} u^{k,m}_{p,\gamma,t,s}+\sum_{m\in\mathcal{L}}\sum_{t\in\mathcal{U}\cup\mathcal{T}}\sum_{n\in\mathcal{L}} v^{k,n}_{\gamma,t,m,s}$$
$$+\sum_{n\in\mathcal{L}}\sum_{t\in\mathcal{U}\cup\{t_0\}} e^{k,n}_{\gamma,t,t_0}\Big)\Big)+w_2\Big(\sum_{s\in\mathcal{S}}\pi_s\sum_{\gamma\in\Gamma}\sum_{k\in\mathcal{K}}\big(\mu_{\gamma,k}\,\alpha^k_{\gamma,s}+(\mu^2_{\gamma,k}+\sigma^2_{\gamma,k})\,\beta^k_{\gamma,s}\big)+\sum_{s\in\mathcal{S}}\pi_s\delta_s\Big)\Big\}, \tag{41a}$$

$$\text{s.t. } \delta_s\ \geq\ \widetilde{\Psi}_s(\boldsymbol{y_s},\widehat{\boldsymbol{y}},\boldsymbol{\alpha_s},\boldsymbol{\beta_s}),\ \forall s\in\mathcal{S} \tag{41b}$$

$$(\boldsymbol{x_s},\ \boldsymbol{y_s},\widehat{\boldsymbol{y}},\ \boldsymbol{u_s},\ \boldsymbol{v_s},\ \boldsymbol{e})\ \in\ \mathcal{O}_s,\ \forall s\in\mathcal{S} \tag{41c}$$

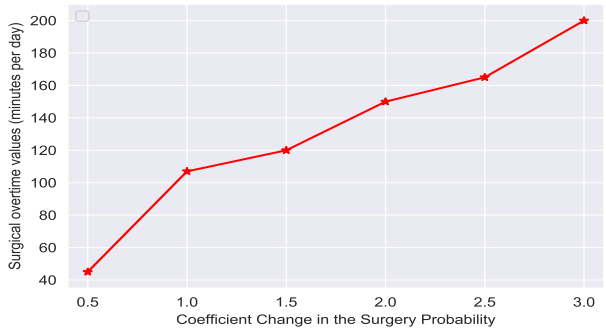
$$\delta_s\ \in\ \mathbb{R},\ \boldsymbol{\alpha_s},\ \boldsymbol{\beta_s}\ \in\ \mathbb{R}^{|\Gamma|\times|\mathcal{K}|},\ \forall s\in\mathcal{S}. \tag{41d}$$

The above IMSDRO-APRX model can be solved by Algorithm 1 described in §4.

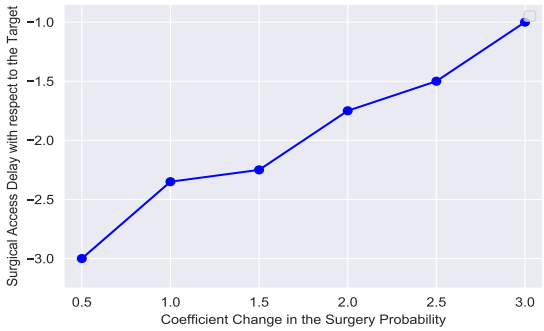
Appendix D: Additional Analyses

In this appendix, we provide additional analyses. They include the sensitivity analysis on the probability of needing a surgical procedure, and the number of intervals for segment points to approximate the support of surgery duration distribution. Furthermore, we compare the single versus multi-cut versions of our proposed constraint generation algorithm.

Sensitivity analysis on the surgery probability. In our approach, we model the need for a surgical procedure through a Bernoulli random variable. This assumption is made by others as well (see e.g., Wang et al. 2015 and Kazemian et al. 2017). We evaluate this modeling choice by a sensitivity analysis on the surgery probability with respect to both surgical overtimes and surgical access time delays. Figure 12 illustrates how the surgical overtime and average surgical access time delay change as the surgery probability alters. We observe that as the surgery probability increases, we require more surgical overtime while the surgical access delay increases.



(a) Surgical overtime versus the surgery prob.



(b) Surgical access delay versus the surgery prob.

Figure 12 The sensitivity analysis around the surgery probability with respect to (a) surgical overtime, and (b) surgical access measure (negative values indicate earliness, i.e., the model grants access to surgery within the maximum wait time target for each patient.)

Analysis of support discretization for the surgery duration. In §3, we reformulated the objective function (21) into a piece-wise linear function with H equal intervals through discretizing the support set $[d^{LB}_{\gamma,k}, d^{UB}_{\gamma,k}]$ of the surgery distribution $d_{\gamma,k}$ for each pair of class $\gamma \in \Gamma$ and surgeon $k \in \mathcal{K}$ into $H+1$ segment points $\Upsilon_{\gamma,k} = \{\tilde{d}_{\gamma,k}(i)\}_{i=0}^{H}$. Here, we provide sensitivity analysis results on varying the number of segment points and investigate the trade-off between the solution quality and the computational time of solving the IMSDRO-APRX model. Intuitively, when the number of segment points H increases, it results in achieving a more precise approximation for the objective function, but with a longer computational time.

Test instance	# of segment points	Objective fun.	# of Iterations	CPU time
The case study	5	4,625	7	1,014
	10	4,451	9	1,345
	20	4,375	13	2,116
Test instance A	5	6,757	6	1,546
	10	6,421	10	2,005
	20	6,235	15	2,643
Test instance B	5	2,389	5	1,189
	10	2,265	8	2,115
	20	2,218	16	2,954

Table 7 The sensitivity analysis on the number of segment points and how it impacts the objective function value, computational time, and the number of iterations for the case study and the test instances A and B.

Table 7 reports the objective function value, computational time (in seconds), and the number of iterations required for solving the IMSDRO-APRX model for the case study and test instances A and B with various number of segment points for the surgery duration. As we increase the number of segment points for the surgery duration (so the support of the surgery duration is approximated more accurately), we obtain a more precise approximation of the objective function with a larger number of iterations. It also requires a longer computational time; however, note that it grows slower than linearly. As we can see in Table 7, by doubling the number of segment points (i.e., increasing it from 5 to 10 and from 10 to 20), the objective

function value alters by less than 5% in the case study. Results are similar for test instances A and B. Table 7 demonstrates that our choice of using 10 segment points as the default in the case study is appropriate.

Comparison of single versus multi-scenario cuts. In §4, we developed a constraint generation algorithm with *multi-scenario cuts* for solving the IMSDRO-APRX model, in which the cut-generating problem (23a)-(23f) obtains at most one scenario cut per scenario and passes it back to the RMP (26a)-(26d). Similar to the L-shaped decomposition methods, our algorithm can have two versions: (i) a multi-cut version in which multiple cuts are added to the RMP (at most one cut per scenario); and (ii) a single cut version in which one aggregated cut is added to the RMP.

In this part, we introduce the single cut version of the proposed constraint generation algorithm, which is similar to Algorithm 1, except that it passes back one aggregated cut in the form of (42b) to the following RMP (42a)-(42d) at each iteration:

$$\bar{Z}^{RMP} = \min_{\boldsymbol{x},\boldsymbol{y},\widehat{\boldsymbol{y}},\boldsymbol{q},\boldsymbol{\delta},\boldsymbol{\alpha},\boldsymbol{\beta}} \sum_{s\in\mathcal{S}} \pi_s \left\{ \sum_{k\in\mathcal{K}} \Big(\sum_{m\in\mathcal{L}} q^k_{m,s} + \sum_{\gamma\in\Gamma} \Big(\mu_{\gamma,k}\, \alpha^k_{\gamma,s} + (\mu^2_{\gamma,k} + \sigma^2_{\gamma,k})\, \beta^k_{\gamma,s} \Big) \Big) \right\} + \Pi \tag{42a}$$

$$\text{s.t. } \Pi \geq \sum_{s\in\mathcal{S}} \pi_s\, \chi_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{\alpha_s}, \boldsymbol{\beta_s}; \boldsymbol{\tau_s^{(r)}}, \boldsymbol{\eta_s^{(r)}}, \boldsymbol{\lambda_s^{(r)}}),\ \forall r = 1, \cdots, R-1 \tag{42b}$$

$$(\boldsymbol{x_s},\, \boldsymbol{y_s}, \widehat{\boldsymbol{y}},\, \boldsymbol{q_s}) \in \mathcal{R}_s,\ \forall s \in \mathcal{S} \tag{42c}$$

$$\delta_s \in \mathbb{R},\, \boldsymbol{\alpha_s},\, \boldsymbol{\beta_s} \in \mathbb{R}^{|\Gamma|\times|\mathcal{K}|},\ \forall s \in \mathcal{S}. \tag{42d}$$

We next compare the empirical performance of single versus multi-scenario cut versions of the proposed constraint generation algorithm. To this aim, we generate 10 random instances of the CAS problem with different number of surgeons, number of patient classes, and number of scenarios. Each instance is specified by a tuple ($|\mathcal{K}|$,$|\Gamma|$,$|\mathcal{T}|$,$|\mathcal{S}|$) denoting the number of surgeons, patient classes, days in the arrival arrival, and scenarios for patient appointment requests, respectively. We then solve the IMSDRO-APRX model for each of these instances by using both multi-cut and single-cut versions of the constraint generation algorithm (Algorithm 1). Table 8 indicates the number of iterations ("# of Iterations") for them until the algorithm converges to the optimal policy as well as the CPU time in seconds for each instance. Note that the instance numbers 5, 8 and 9 refer to the test instance A, the case study and the test instance B, respectively.

Instance Number	($\|\mathcal{K}\|$, $\|\Gamma\|$, $\|\mathcal{T}\|$, $\|\mathcal{S}\|$)	multi cuts		single cut	
		# of Iterations	CPU time	# of Iterations	CPU time
1	(4, 12, 4, 8)	5	654	9	1,177
2	(4, 12, 5, 10)	7	1,021	12	1,897
3	(4, 12, 5, 15)	8	1,254	18	2,257
4	(4, 24, 5, 15)	6	1,578	13	2,957
5	(4, 24, 5, 20)	10	2,005	20	3,609
6	(8, 12, 5, 8)	4	755	13	1,177
7	(8, 24, 5, 10)	6	1,176	13	2,215
8	(8, 24, 5, 14)	9	1,345	15	2,421
9	(10, 24, 5, 10)	8	2,115	16	3,484
10	(10, 24, 5, 15)	7	2,321	13	3,752

Table 8 The comparison of the multi-cut and single-cut versions of the constraint generation Algorithm 1 for different instances of the CAS problem in terms of the number of iterations and CPU time in seconds.

Table 8 demonstrates that we are able to solve a range of suitable instances for the IMSDRO-APRX model within a reasonable number of iterations. We also observe that since the multi-cut version offers more information about the feasible region, we require fewer number of iterations compared to the single-cut version. The average number of iterations for the multi-cut version was 6, while it was 14 for the single-cut version. Moreover, the single-cut version takes more CPU time compared to the multi-cut version.

References

Denton, Brian, Diwakar Gupta. 2003. A sequential bounding approach for optimal appointment scheduling. *IIE transactions* **35**(11) 1003–1016.

Dupačová, Jitka, Nicole Gröwe-Kuska, Werner Römisch. 2003. Scenario reduction in stochastic programming. *Mathematical programming* **95**(3) 493–511.

Helton, Jon C, Freddie Joe Davis. 2003. Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. *Reliability Engineering & System Safety* **81**(1) 23–69.

Jiang, Ruiwei, Siqian Shen, Yiling Zhang. 2017. Integer programming approaches for appointment scheduling with random no-shows and service durations. *Operations Research* **65**(6) 1638–1656.

Kaut, Michal, Stein W Wallace. 2003. Evaluation of scenario-generation methods for stochastic programming.

Kazemian, Pooyan, Mustafa Y Sir, Mark P Van Oyen, Jenna K Lovely, David W Larson, Kalyan S Pasupathy. 2017. Coordinating clinic and surgery appointments to meet access service levels for elective surgery. *Journal of biomedical informatics* **66** 105–115.

Wang, Bing, Xingbao Han, Xianxia Zhang, Shaohua Zhang. 2015. Predictive-reactive scheduling for single surgical suite subject to random emergency surgery. *Journal of Combinatorial Optimization* **30**(4) 949–966.