Online Supplement: Managing Hospital Inpatient Bed Capacity through Partitioning Care into Focused Wings

Thomas J. Best, Burhaneddin Sandıkcı, Donald D. Eisenstein
University of Chicago Booth School of Business,
thomas.j.best@chicagobooth.edu, burhan@chicagobooth.edu, don.eisenstein@chicagobooth.edu

David O. Meltzer
University of Chicago Medical Center, dmelzer@medicine.bsd.uchicago.edu

Appendix A: Proofs of Propositions 1 and 2

Let $N_j$ denote the expected number of occupied beds in wing $j$. Using Little’s law $N_j = d_j(1 - p_j)$ for wing $j$, where $d_j = \sum_{i \in T} \lambda_i' s_i'$ is the bed-demand and $p_j$ is the abandonment probability as noted earlier in the paper.

**Proposition 1** (rephrased). Consider an instance with $C = \{1, 2\}$ under the queueing system of §4.3. If $\frac{\lambda'_1}{m'} = \frac{\lambda'_2}{m'}$, $q \geq \max\{m^1, m^2\}$, and there are no focus effects, then $U^{1,2} \geq U^1 + U^2$, where $U^i$ is the expected utility from a wing for type $i \in C$ with $b^i$ beds, and $U^{1,2}$ is the expected utility from a pooled wing with $b^1 + b^2$ beds.

**Proof of Proposition 1.** Under the queueing system of §4.3, the total expected utility from the pooled wing is $U^{1,2} = \lambda^1(1 - p^{1,2})u^1 + \lambda^2(1 - p^{1,2})u^2$. Substituting for $u^i$ using equation (4), and noting that $\gamma^i = 0$ for all $i \in C$ when there are no focus effects, we find

$$U^{1,2} = (\lambda^1 v^1 + \lambda^2 v^2)(1 - p^{1,2}) = \frac{\lambda^1}{m} (\lambda^1 m^1 + \lambda^2 m^2)(1 - p^{1,2}),$$

where the last equality follows from the assumption $\frac{\lambda^1}{m^1} = \frac{\lambda^2}{m^2}$. Since $\alpha' = 0$ for all $i \in C$ when there are no focus effects, equation (4) implies that $m^i = s_i'$, and therefore $\lambda^i m^1 + \lambda^2 m^2 = d^{1,2}$, from which we conclude $U^{1,2} = \frac{\lambda^1}{m} N^{1,2}$. Similarly, $U^1 = \frac{\lambda^1}{m} N^1$ and $U^2 = \frac{\lambda^2}{m} N^2$. Since $\frac{\lambda^1}{m^1} = \frac{\lambda^2}{m^2}$ is assumed, we write $U^1 + U^2 = \frac{\lambda^1}{m}(N^1 + N^2)$. Thus, $U^{1,2} \geq U^1 + U^2$ if $N^{1,2} \geq N^1 + N^2$, which is shown to hold in the proof of Proposition 2. □

In the rest of this section, for clarity, we explicitly write out the arguments of some functions (e.g., we write $p_j$ as $p(d_j, \Lambda, q_j, b_j)$). We also indicate a one-wing formation with the subscript 0 (e.g., $q_0$ denotes the average willingness-to-wait under the one-wing formation). Before proving Proposition 2, we recall two technical results from the literature.

**Lemma 1.** [Armony et al. (2009, Proposition 2)] If $q_0 \geq \max_{i \in T} \{m^i\}$, then $p$ is convex in the number of servers.

**Lemma 2.** [Smith and Whitt (1981, pg. 53)] $\mathbb{E}\left(d_0, b_j \frac{d_0}{\sum_{i \in T} \lambda_i'} \right) \geq \mathbb{E}\left(\sum_{i \in T} \lambda_i' s_i', b_j\right)$.


Proposition 2 (restated). Consider an instance of \((P)\) under the queueing system of §4.3, and the one-wing solution (denoted ‘O’) and a multi-wing solution (denoted ‘M’). If the mean willingness-to-wait in ‘O’ is not smaller than the mean willingness-to-wait in any wing in ‘M’ or the nominal mean length-of-stay of any care type, and the focus-adjusted mean length-of-stay of each care type in ‘O’ is not smaller than that for ‘M’, then the bed occupancy of ‘O’ is higher than the weighted average bed occupancy of ‘M’.

Proof of Proposition 2. The bed occupancy of the one-wing solution is \(N_0\), where \(N_0\) is the expected number of occupied beds. Consider a multi-wing solution with \(w\) wings and \(b_j\) beds in wing \(j = 1, \ldots, w\) such that \(\sum_{j=1}^{w} b_j = B\). The weighted average bed occupancy in ‘M’ is \(\sum_{j:b_j>0}^w \frac{N_j b_j}{B} = \frac{\sum_{j=1}^{w} N_j}{B}\). Thus the proof will be complete if \(N_0 > \sum_{j=1}^{w} N_j\).

\[
N_0 = d_0 \cdot [1 - p(d_0, \Lambda_0, q_0, b_0)] \geq d_0 \left[1 - \sum_{j=1}^{w} \left(\sum_{i \in T_j} \lambda_i (1 - \alpha_i^0) m_i^j \right) \cdot p(d_0, \Lambda_j, q_0, b_j) \frac{d_0}{\sum_{i \in T_j} \lambda_i (1 - \alpha_i^0) m_i^j} \right],
\]

where last inequality follows from Lemma 1. Since \(\frac{\partial p}{\partial q} \leq 0\) as shown in the proof of Theorem 1, we replace \(\Lambda_0\) in the right hand-side of inequality (1) with the smaller terms \(\Lambda_j\) and find

\[
N_0 \geq d_0 - \sum_{j=1}^{w} \left(\sum_{i \in T_j} \lambda_i (1 - \alpha_i^0) m_i^j \right) \cdot p(d_0, \Lambda_j, q_0, b_j) \frac{d_0}{\sum_{i \in T_j} \lambda_i (1 - \alpha_i^0) m_i^j}.
\]

Recall now equation (3) for the abandonment probability \(p\). Note that substituting \(\sum_{i \in T_j} \lambda_i (1 - \alpha_i^0) m_i^j\) for \(d_0\) and \(b_j\) for \(b_j \sum_{i \in T_j} \lambda_i (1 - \alpha_i^0) m_i^j\) in the abandonment probability in the right hand-side of inequality (2) only impacts the \(E\) term. Lemma 2 implies that the \(\mathcal{E}\) term decreases, and therefore \(p\) increases, after this substitution. So, we find

\[
N_0 > d_0 - \sum_{j=1}^{w} \left(\sum_{i \in T_j} \lambda_i (1 - \alpha_i^0) m_i^j \right) \cdot \left[1 - p \left(\sum_{i \in T_j} \lambda_i (1 - \alpha_i^0) m_i^j, \Lambda_j, q_0, b_j\right)\right],
\]

where equality (4) follows because \(d_0 = \sum_{i \in \mathcal{C}} \lambda_i (1 - \alpha_i^0) m_i = \sum_{j=1}^{w} \sum_{i \in T_j} \lambda_i (1 - \alpha_i^0) m_i\). Direct differentiation shows that \(\frac{\partial p}{\partial q} \leq 0\), and therefore substituting the smaller \(q_j\) terms for \(q_0\) in the right hand-side of (4), we find

\[
N_0 \geq \sum_{j=1}^{w} \left(\sum_{i \in T_j} \lambda_i (1 - \alpha_i^0) m_i^j \right) \cdot \left[1 - p \left(\sum_{i \in T_j} \lambda_i (1 - \alpha_i^0) m_i^j, \Lambda_j, q_j, b_j\right)\right].
\]

For \(j = 1, \ldots, w\), the terms \(\sum_{i \in T_j} \lambda_i (1 - \alpha_i^0) m_i^j\) and \(p \left(\sum_{i \in T_j} \lambda_i (1 - \alpha_i^0) m_i^j, \Lambda_j, q_j, b_j\right)\) can be thought of as the bed-demand and abandonment probability of wing \(j\) under the length-of-stay scaling factors of the one-wing solution. Therefore, the right hand-side of inequality (5) equals the total expected number of occupied beds \(\sum_{j=1}^{w} N_j\) under the one-wing scaling factors. Direct differentiation shows that \(\frac{\partial \mathcal{L}}{\partial b} \leq 0\). Thus, replacing \((1 - \alpha_i^0)\) with \((1 - \alpha^i)\) for each \(i \in \mathcal{C}\) yields

\[
N_0 \geq \sum_{j=1}^{w} \left(\sum_{i \in T_j} \lambda_i (1 - \alpha^i) m_i^j \right) \cdot \left[1 - p \left(\sum_{i \in T_j} \lambda_i (1 - \alpha^i) m_i^j, \Lambda_j, q_j, b_j\right)\right] = \sum_{j=1}^{w} d_j \cdot [1 - p_j] = \sum_{j=1}^{w} N_j,
\]

which completes the proof. \( \square \)

Appendix B: An upper bound on \(Z\)

In §B.1, we transform an instance of \((P)\) into a new instance yielding an upper bound for the original instance. In §B.2, we develop a method to efficiently compute this bound and, in §B.3, we further tighten the bound. We use this upper bound to evaluate the accuracy of the heuristic presented in §§4.1-4.2, and report the resulting optimality gaps in §7.
B.1. Constructing an upper bound on Z

Let $\Theta = \{B, C, (\lambda')_{i \in C}, (m')_{i \in C}, (v')_{i \in C}\}$ denote an instance of problem $(P)$ to be evaluated under the queueing system of §4.3. Figure S.1 presents pseudocode for transforming $\Theta$ into a new instance $\hat{\Theta}$ which returns an upper bound $Z(\hat{\Theta})$ on the optimal objective value $Z(\Theta)$ of $\Theta$. The main idea of our transformation is to disaggregate the original types in $C$ into a total of $\hat{T} \geq T$ sub-types, so that every new sub-type has the same arrival rate $\lambda$ and the same mean length-of-stay $g$. The term $n' \geq 1$ in step (T-1) denotes the number of sub-types created by original type $i \in C$, and $f(\cdot)$ in step (T-6) is a mapping that keeps track of the origin of each sub-type $k$. In order to prevent unnecessarily inflating the total utility of the hospital, step (T-7) sets the nominal utility of sub-type $k$ to a fraction of the nominal utility of its original type. Finally, $\hat{C}$ in step (T-8) is the index set of all sub-types created during the transformation.

To generate a valid upper bound for $Z(\Theta)$, the effects of focus on utility and length-of-stay must also be transformed appropriately before solving $\hat{\Theta}$. Let $\Psi(\Theta)$ and $\Psi(\hat{\Theta})$ denote the feasible regions, as defined in §3, for instances $\Theta$ and $\hat{\Theta}$, respectively. We say that a solution $\hat{\psi} = (\hat{w}, \hat{b}, \hat{T}) \in \Psi(\hat{\Theta})$ is equivalent to a solution $\psi = (w, b, T) \in \Psi(\Theta)$ if (i) $\hat{w} = w$, and, for $j = 1, \ldots, n$, (ii) $\hat{b}_j = b_j$, and (iii) $\hat{T}_j = T_j$, where the ‘≡’ relationship holds if substituting $T_j$ for $C$ in the ‘MAIN STEP’ of Figure S.1 yields $\hat{C} = \hat{T}_j$. Using this definition of equivalence, Figure S.2 provides an appropriate transformation of the scaling factors that model the effects of focused care. Step (F-1) assigns the focus-adjusted utility $u'(\psi)$ to $\hat{u}'(\hat{\psi})$ and length-of-stay $s'(\psi)$ to $\hat{s}'(\hat{\psi})$ for all sub-types $k \in \hat{C}$ that have emanated from the original type $i \in C$, when $\psi$ is equivalent to $\hat{\psi}$, which occurs only if the set of sub-types $\hat{T}_j$ for each wing $j$ is entirely composed of the sub-types that are created from the original types in $T_j$. The intuition behind this step is the need to maintain the integrity of the effects of focused care. In other words, the effects of focus should be immune to giving any (dis)advantage to a wing considered in the transformed instance $\hat{\Theta}$ that is essentially equivalent to a wing in the original instance $\Theta$.

Lemma 3 provides a number of technical results regarding our transformation. These results indicate that, for two equivalent wings, one of which is considered for $\Theta$ and the other for $\hat{\Theta}$, the transformations presented in Figures S.1-S.2 keep the overall bed-demand (Lemma 3.a.–3.b.), hence the overall load (Lemma 3.c.), for each wing the same.

**INPUT.** An instance $\Theta = \{B, C, (\lambda')_{i \in C}, (m')_{i \in C}, (v')_{i \in C}\}$ of problem $(P)$.

**INITIALIZE.**

\[
\begin{align*}
T & \leftarrow 0; 
 k & \leftarrow 0 
 g & \leftarrow \text{min}_{i \in C} \{m'_i\} 
 \lambda & \leftarrow \text{GCD}_{i \in C} \left( \frac{\lambda'}{\mu'} \right), \text{where GCD stands for the greatest common divisor}
\end{align*}
\]

**MAIN STEP.** For each original type $i \in C$

\[
\begin{align*}
n'_i & \leftarrow \frac{x_w}{\lambda'} 
 \hat{T} & \leftarrow T + n'_i
\end{align*}
\]

For each sub-type in $\{1, \ldots, n'_i\}$

\[
\begin{align*}
k & \leftarrow k + 1 
 \hat{\lambda} & \leftarrow \lambda 
 \hat{\mu} & \leftarrow g 
 f(k) & \leftarrow i 
 \hat{s}' & \leftarrow s'(\hat{b}) \frac{g}{\mu'}
\end{align*}
\]

End-For-loop

\[
\begin{align*}
\hat{C} & \leftarrow \{1, \ldots, \hat{T}\}
\end{align*}
\]

**OUTPUT.** A new instance $\hat{\Theta} = \{B, \hat{\Theta}, (\hat{\lambda}')_{i \in \hat{C}}, (\hat{m}')_{i \in \hat{C}}, (\hat{v}')_{i \in \hat{C}}\}$ of problem $(P)$.

![Figure S.1 Pseudocode for transforming an instance $\Theta$ of problem $(P)$ into a new instance $\hat{\Theta}$](image-url)
However, the wing considered for \( \hat{\Theta} \) experiences a higher arrival rate than for \( \Theta \) (Lemma 3.d.), a shorter average length-of-stay (since \( g \leq m^j \) for all \( i \in C \)), and a lower utility per patient (Lemma 3.e.), although the total uncapacitated utility of the wing is unchanged (Lemma 3.f.).

**Lemma 3.** Given a pair of instances \((\Theta, \hat{\Theta})\) of the problem \((P)\) and a pair of solutions \((\psi, \hat{\psi}) \in (\Psi(\Theta), \Psi(\hat{\Theta}))\) such that \( \psi \equiv \hat{\psi} \), the following hold for wing \( j = 1, \ldots, w \):

\[
\begin{align*}
a. & \quad \sum_{i \in T_j} \lambda_i m^i = \sum_{k \in \hat{T}_j} \hat{\lambda}_k \hat{m}^k. \\
b. & \quad d_j = \hat{d}_j. \\
c. & \quad \rho_j = \hat{\rho}_j. \\
d. & \quad \Lambda_j \leq \hat{\Lambda}_j. \\
e. & \quad u^\prime \geq \hat{u}^\prime. \\
f. & \quad \sum_{i \in T_j} \lambda_i u^i = \sum_{k \in \hat{T}_j} \hat{\lambda}_k \hat{u}^k.
\end{align*}
\]

**Proof of Lemma 3.**

\[
\begin{align*}
a. & \quad \sum_{k \in \hat{T}_j} \hat{\lambda}_k \hat{m}^k = \sum_{i \in T_j} \lambda_i m^i = \sum_{k \in \hat{T}_j} \hat{\lambda}_k \hat{m}^k = \sum_{i \in T_j} \lambda_i m^i = \sum_{i \in T_j} \lambda_i n^i = \sum_{i \in T_j} \lambda_i m^i, \quad \text{where the last equality follows from step (T-1) in Figure S.1.}
\end{align*}
\]

\[b.-f. \quad \text{Similar to the proof of Lemma 3.a.} \]

**Theorem 1 (restated).** \( Z(\hat{\Theta}) \geq Z(\Theta) \).

**Proof of Theorem 1.** For any feasible solution \( \psi \in \Psi(\Theta) \), consider the construction in Figure S.3 that produces a feasible solution \( \hat{\psi} \in \Psi(\hat{\Theta}) \). We proceed to show that the objective function \( \sum_{i=1}^w U_j \) of problem \((P)\) evaluated at \( \hat{\psi} \) is not smaller than at \( \psi \). First note that the set of sub-types assigned to wing \( j \) of \( \hat{\psi} \) is simply the collection of sub-types obtained by disaggregating the original types assigned to wing \( j \) of \( \psi \), hence \( \psi \equiv \hat{\psi} \). Rewriting equation (1) under the queueing system of §4.3, we have \( U_j = (1 - \rho_j) \sum_{i \in T_j} \lambda_i u^i \) for \( j = 1, \ldots, w \). Lemma 3.f. implies that we only need to check the abandonment probability \( \rho_j \), as \( \sum \lambda_i u^i \) is the same under both solutions \( \psi \) and \( \hat{\psi} \), because \( \psi \equiv \hat{\psi} \).

Pick an arbitrary wing \( j \). By construction, the set of care types in the wing under \( \psi \) and \( \hat{\psi} \) are equivalent (i.e., \( T_j = \hat{T}_j \)), and the willingness-to-wait parameter \( q_j \) and the number of beds \( b_j \) allocated to the wing remain the same. Furthermore, Lemma 3.b. implies that the bed-demand \( d_j \) to the wing is the same under both \( \psi \) and \( \hat{\psi} \). Therefore, \( E_j \) in equation (3) is the same for \( \psi \) and \( \hat{\psi} \). On the other hand, Lemma 3.d. implies that the overall arrival rate \( \Lambda_j \) increases from \( \psi \) to \( \hat{\psi} \). Differentiating \( \rho_j \) with respect to \( \Lambda_j \), we find \( \frac{\partial \rho_j}{\partial \Lambda_j} = \frac{\partial \rho_j}{\partial \Lambda_j} \cdot \frac{\partial \Lambda_j}{\partial \Lambda_j} \). Lemma 5 below shows that \( \frac{\partial \rho_j}{\partial \Lambda_j} < 0 \) and direct differentiation shows that \( \frac{\partial \Lambda_j}{\partial \Lambda_j} \geq 0 \). As a result, \( \frac{\partial \rho_j}{\partial \Lambda_j} \leq 0 \). Combining this result with the result of Lemma 3.d., we find that \( p_j(\hat{\psi}) \leq p_j(\psi) \). Therefore, the wing utility \( U_j \) evaluated at \( \hat{\psi} \) is not smaller than that evaluated at \( \psi \).
 Furthermore, since condition (C3), we denote the length-of-stay scaling factor for wing $j$ as $\gamma_j$. As a result of condition (C3), we denote the utility scaling factor for wing $w$ as $\alpha_w$. We denote $T_j^{\ast}$ as the resulting solution (after trivial resequencing of the wings, or the care types within a wing, if necessary), then the result holds. Other-\[\hat{\theta} \leftarrow (\hat{\psi}) \in \Psi(\hat{\Theta}).\]wise, we show that $T_j^{\ast}$ can be transformed into a new partition $T'$ such that $T'$ is formed by making cuts in the sequence $\hat{\theta}$ and the resulting solution $(w^*, b^*, T')$ has no loss of utility.

**Proposition 1.** Consider an instance of (P) under the queueing system of \( S \). If conditions (C1-C3) hold, then there exists an optimal solution $(w^*, b^*, T^*)$ such that $T^* = \{T_1^*, \ldots, T_w^*\}$ is formed by making cuts in the sequence $\hat{S}$.

**Proof of Proposition 1.** Let $\psi = (w^*, b^*, T^*)$ be an optimal solution to (P). If $T^*$ is formed by cuts into the sequence $\hat{S}$ (after trivial resequencing of the wings, or the care types within a wing, if necessary), then the result holds. Otherwise, we show that $T^*$ can be transformed into a new partition $T'$ such that $T'$ is formed by making cuts in $\hat{S}$ and the resulting solution $(w^*, b^*, T')$ has no loss of utility.

Condition (C2) implies that all of the care types in a wing have the same utility scaling factor, hence, for notational clarity, we denote the utility scaling factor for wing $j$ as $\gamma_j(\psi') := \gamma_j(\psi')$ for $i \in T_j^{\ast}$ and $j = 1, \ldots, w^*$. Similarly, by condition (C3), we denote the length-of-stay scaling factor for wing $j$ as $\alpha_j(\psi') := \alpha_j(\psi')$ for $i \in T_j^{\ast}$ and $j = 1, \ldots, w^*$. Furthermore, since $\lambda^* = \lambda$ for all $i$ by condition (C1), equation (1) for wing $j$ under solution $\psi$ simplifies to

\[\sum_{i \in T_j^{\ast}} p_i(1 + \gamma_j(\psi')) \lambda^* \sum_{i \in T_j^{\ast}} v_i.\]
Without loss of generality, we re-index the wings of $\psi$ so that $(1 - p_i) \cdot (1 + \gamma_i(\psi)) \geq (1 - p_j) \cdot (1 + \gamma_j(\psi)) \geq \cdots \geq (1 - p_w) \cdot (1 + \gamma_w(\psi))$. Since $\mathcal{T}^*$ is not cut from $S$, there must exist at least two care types $i_1 \in \mathcal{T}^*_{j_1}$ and $i_2 \in \mathcal{T}^*_{j_2}$ where $\psi^{j_1} < \psi^{j_2}$ for wing indices $j_1 < j_2$. Consider swapping these care types, and let $\tilde{\psi}$ denote the resulting wing formation. Conditions (C2) and (C3), respectively, imply that such a swap does not change the utility scaling factors, $\gamma_{j_1}$ or $\gamma_{j_2}$, and the length-of-stay scaling factors, $\alpha_{j_1}$ or $\alpha_{j_2}$. Combining this result with condition (C1) implies that the abandonment probabilities $p_{i_1}$ and $p_{i_2}$ are also not affected by a swap. Therefore, only the term $\sum_{i \in \mathcal{T}^*_{j_1}} \psi^i$ from $U_{j_1}$ and $\sum_{i \in \mathcal{T}^*_{j_2}} \psi^i$ from $U_{j_2}$ are affected by the swap. The net change in $U_{j_1} + U_{j_2}$ is

$$A \cdot \left( \psi^{j_2} - \psi^{j_1} \right) \cdot \left[ (1 - p_{i_1})(1 + \gamma_{j_1}(\psi)) - (1 - p_{i_2})(1 + \gamma_{j_2}(\psi)) \right] \geq 0,$$

where the inequality holds since $A \geq 0$, $(1 - p_{i_1}) \cdot (1 + \gamma_{j_1}(\psi)) \geq (1 - p_{i_2}) \cdot (1 + \gamma_{j_2}(\psi))$, and $\psi^{j_2} > \psi^{j_1}$. We repeat such swaps until the $|\mathcal{T}^*_{j_1}|$ largest nominal utility care types are assigned to wing 1, the next $|\mathcal{T}^*_2|$ largest to wing 2, and so forth—forming a new care partition $\mathcal{T}'$ which cuts $\tilde{S}$. The resulting solution $(w', b', \mathcal{T}')$ has no loss of utility, because of inequality (8). □

One can readily confirm that the instance $\hat{\Theta}$ satisfies condition (C1). However, the scaling factors (5) and (6) may violate conditions (C2) and (C3) under instance $\hat{\Theta}$. So we construct a modified problem ($\hat{P}$) that augments the set of decision variables of the problem ($P$) in such a way that conditions (C2) and (C3) are satisfied. Let $t = \{t_1, \ldots, t_w\}$, where $t_j$ ($j = 1, \ldots, w$) is a new decision variable. Denote the vector of decision variables for problem ($\hat{P}$) by $(w, b, \mathcal{T}, t)$. Furthermore, for any $\Theta$, we create $\hat{\Theta}$ as in Figure S.1. However, we do not use Figure S.2 to adjust the scaling factors. Rather, we determine the scaling factors for $\hat{\Theta}$ inside problem ($\hat{P}$) using a set of constraints similar to equations (5) and (6). The modified problem is

$$(\hat{P}) \quad \hat{Z}(\hat{\Theta}) = \max_{(w, b, \mathcal{T}, t)} \left\{ \sum_{j=1}^{w} U_j : \{w, b, \mathcal{T}\} \in \Psi(\hat{\Theta}), \ t \in T \right\},$$

where $\Psi$ is as given in §3, and

$$T = \left\{ t \in \mathbb{Z}^w : t_j \geq 0 \text{ for } j = 1, \ldots, w, \sum_{j=1}^{w} t_j = T, \ \alpha_j = \frac{\Delta \left( 1 - \frac{t_j}{T} \right)}{1 + e^{\beta (\rho_j - c)}} \text{ and } \gamma_j = \eta \left( 1 - \frac{t_j}{T} \right) \text{ for } j = 1, \ldots, w \right\},$$

where $T$ is the number of care types in the original instance $\Theta$.

Note that for any fixed $t$, ($\hat{P}$) reduces to ($P$), therefore one can apply Proposition 1. We are now ready to prove Theorems 2 and 3.

**Theorem 2 (restated).** An instance of problem ($\hat{P}$) can be solved to optimality using the DP of §4.2 by making cuts in the utility-sorted sequence $\tilde{S}$.

**Proof of Theorem 2.** As in the proof of Proposition 1, let $(w', b', \mathcal{T}', t')$ be any claimed optimal solution to ($\hat{P}$) such that $\mathcal{T}' = \{\mathcal{T}_1', \mathcal{T}_2', \ldots, \mathcal{T}_w'\}$ is not cut from $\tilde{S}$. It is clear that swapping sub-types between wings does not change the values of $b_j$ and $t_j$ for any wing $j = 1, \ldots, w$. Therefore, the scaling factors $\alpha_j$ and $\gamma_j$ enforced by the constraints of $\mathcal{T}$ satisfy conditions (C2-C3). Furthermore, condition (C1) is satisfied by the data of $\hat{\Theta}$. As a result, we can swap types as in the proof of Proposition 1 to show the existence of an optimal partition $\mathcal{T}^*$ that is formed by cuts into $\tilde{S}$. □

**Theorem 3 (restated).** $\hat{Z}(\hat{\Theta}) \geq Z(\Theta)$. 
Proof of Theorem 3. For any solution \( \psi \in \Psi(\Theta) \), we construct \( \hat{\psi} \in \Psi(\hat{\Theta}) \) as in Figure S.3. And, for \( j = 1, \ldots, w \), we set \( t_j \) equal to the number of original types in wing \( j \) under the \( \psi \) solution, which is feasible to \( \sum_{j=1}^{w} t_j = T \) in \( \Upsilon \). This implies that the scaling factors (5) and (6) evaluated under \( \psi \) are, respectively, equal to the scaling factor equations \( \alpha_j \) and \( \gamma_j \) in \( \Upsilon \) evaluated under \( (\hat{\psi}, t) \). Therefore, \( \hat{\psi} \) is feasible to \( \Psi(\hat{\Theta}) \) and \( t \) is feasible to \( \Upsilon \). Showing that \( \sum_{j=1}^{w} U_j \) evaluated under \( (\hat{\psi}, t) \) is at least as much as that under \( \psi \) is similar to the proof of Theorem 1. \( \square \)

B.3. Tightening the upper bound \( Z(\hat{\Theta}) \)

To tighten \( Z(\hat{\Theta}) \), we further restrict the values \( t_j \) for \( j = 1, \ldots, w \) by adding the following constraints to \( \Upsilon \):

\[
|T_j| \in \left\{ \sum_{i \in I} n_i : I \subseteq C \right\} \quad \text{for } j = 1, \ldots, w, \tag{9}
\]

\[
t_j \in \left\{ |I| : I \subseteq C \text{ and } \sum_{i \in I} n_i = |T_j| \right\} \quad \text{for } j = 1, \ldots, w. \tag{10}
\]

Recall that \( n_i \) is set in Figure S.1 (step T-1) as the number of sub-types that originated from type \( i \). Adding these constraints shrinks \( \Upsilon \), hence we obtain a tighter objective value \( Z(\hat{\Theta}) \). To check that this value is still a valid upper bound for \( Z(\Theta) \), consider any feasible solution \( \psi \in \Psi(\Theta) \). Clearly, \( \psi \) has a feasible completion \( t \) that satisfies (10). Therefore, we can construct a solution \( (\hat{\psi}, t) \) as in the proof of Theorem 3 that is feasible to \( \Psi, \Upsilon, (9), \) and (10). Showing that \( \sum_{j=1}^{w} U_j \) evaluated under \( (\hat{\psi}, t) \) is no worse than under \( \psi \) proceeds similarly to the proof of Theorem 1.

After adding these constraints, the scaling factors \( \alpha_j \) and \( \gamma_j \) enforced by the constraints of \( \Upsilon \) do not violate conditions (C2-C3). Hence, as in Theorem 2, we can find the optimal wing formation by making cuts into the sequence \( \hat{S} \).

Appendix C: Computing the abandonment probability

This section describes how to compute the abandonment probability given in equation (3). For notational convenience, we drop the wing index \( j \) throughout this section. First, estimate \( X \) recursively as illustrated in Figure S.4. The \( R_n \) term in step (X-2) is the \( n \)th ratio inside the infinite summation needed to compute \( X \) in equation (3). Note that the recursion stops in a finite number of steps, because \( R_n < 1 \) for some finite \( n \) and, beyond this step, \( R_n \) is strictly decreasing in \( n \), therefore \( R_n \leq \epsilon \in (0, 1) \) for some finite \( n \). Furthermore, it can be shown that \( X_n \) monotonically converges to \( X \) from below; and it may have an initial convex piece, but for large enough \( n \), it becomes concave.

**INPUT.** Problem parameters \( \Lambda, q, d, b \), and tolerance parameter \( \epsilon \in (0, 1) \).

**INITIALIZE.**

\[
n \leftarrow 0
\]

\[
\phi \leftarrow \frac{d}{2}, \quad A \leftarrow \Lambda q
\]

\[
X_0 \leftarrow 0
\]

\[
R_0 \leftarrow 1
\]

**MAIN STEP.** While \( R_n > \epsilon \)

\[
n \leftarrow n + 1
\]

\[
X_n \leftarrow X_{n-1} + R_{n-1}
\]

\[
R_n \leftarrow \frac{A}{\epsilon^{n+1}}
\]  \hspace{1cm} (X-1)

\[
R_n \leftarrow \frac{A}{\epsilon^{n+1}} \tag{X-2}
\]

End-While-loop

**OUTPUT.** \( \hat{X} \leftarrow X_n \).

Figure S.4 Pseudocode for estimating \( X \).
Lemma 4 shows that our estimate $\hat{X}$ recovers at least $100(1-\epsilon)$% of the actual $X$, so one can get arbitrarily close to $X$ by a sufficiently small tolerance $\epsilon$.

**Lemma 4.** $\hat{X} \geq (1-\epsilon)X$.

**Proof of Lemma 4.** Let $n^*$ denote the iterate in Figure S.4 upon termination of the procedure, so $R_{n^*} \leq \epsilon$. The error in our estimate is given by

$$X - \hat{X} = \sum_{n=n^*}^{\infty} R_n = R_{n^*} \left[ 1 + \sum_{n=n^*+1}^{\infty} \prod_{k=n^*+1}^{n} \left( \frac{A}{\phi + k} \right) \right] \leq R_{n^*} \left[ 1 + \sum_{n=n^*+1}^{\infty} \prod_{k=1}^{n} \left( \frac{A}{\phi + k} \right) \right] = R_{n^*} X \leq \epsilon X,$$

where the second equality follows from step (X-2) in Figure S.4, the first inequality results from adding the nonnegative terms for $n = 1, 2, \ldots, n^*$, and last inequality follows from $R_{n^*} \leq \epsilon$. □

Lemma 5 presents a technical result, which is used in proving the main result of this section in Theorem 1.

**Lemma 5.** $\frac{\partial p}{\partial \phi} < 0$.

**Proof of Lemma 5.** Differentiating equation (3) with respect to $X$, we obtain

$$\frac{\partial p}{\partial X} = \frac{\frac{\phi}{b}(E - 1) - E}{(E + \frac{\phi}{b})X(E + \frac{\phi}{b}X)},$$

which is negative if and only if the numerator is negative. For clarity, in the rest of this proof, we write $E$ as $E(b)$ to be explicit about its dependence on $b$ and note that $E(b+1) = 1 + \frac{\phi}{b}\frac{d}{db}$ for $b = 1, 2, \ldots$. Since $E(1) = 1$, the numerator of equation (11) is negative for $b = 1$. Assume that $\frac{\phi}{b}[E(b) - 1] - E(b) < 0$ for $b = 2, 3, \ldots, b^*$. For $b = b^* + 1$,

$$\frac{d}{db^*} [E(b^*+1) - 1] = \frac{b^*}{b^*+1}E(b^*) - \frac{E(b^*)}{b^*+1} - 1 < \frac{b^*}{b^*+1}E(b^*) - [E(b^*) - 1] - 1 = \frac{-E(b^*)}{b^*+1} \leq 0,$$

where the first equation follows from substituting $1 + \frac{\phi(b^*+1)}{db^*}$ for $E(b^*+1)$, and the strict inequality follows from the induction hypothesis that implies $\frac{\phi(b^*)}{db^*} > E(b^*) - 1$. □

Let $\hat{p}$ denote the estimated abandonment probability obtained from substituting $\hat{X}$ for $X$ in equation (3). Theorem 1 establishes that $\hat{p}$ overestimates the actual abandonment probability $p$ by at most $\epsilon$.

**Theorem 1.** $\hat{p} - p \in [0, \epsilon]$.

**Proof of Theorem 1.** The nonnegativity of $\hat{p} - p$ follows from Lemma 5 along with the fact that $\hat{X} \leq X$.

To show that $\hat{p} - p \leq \epsilon$, first note that Lemmas 4 and 5 together imply

$$\hat{p} = \frac{1 + (\phi - 1)\hat{X}}{E + \phi \hat{X}} \leq \frac{1 + (\phi - 1)(1-\epsilon)X}{E + \phi(1-\epsilon)X},$$

from which we find

$$\hat{p} - p \leq \frac{1 + (\phi - 1)(1-\epsilon)X}{E + \phi(1-\epsilon)X} - \frac{1 + (\phi - 1)X}{E + \phi X} = \left( \frac{E - \phi(E - 1)}{E + \phi(1-\epsilon)X} \right) \cdot \epsilon \left( \frac{X}{E + \phi X} \right).$$

Since $E \geq 1$, $X \geq 1$, $\phi \geq 0$, and $\epsilon \in (0, 1)$, we find

$$\frac{E - \phi(E - 1)}{E + \phi(1-\epsilon)X} \leq 1.$$

If $\phi \geq 1$, then $\frac{X}{E + \phi X} \leq 1$. If, on the other hand, $\phi < 1$, then $p \geq 0$ implies $X \leq \frac{1}{1-\phi}$, which leads to $\frac{X}{E + \phi X} \leq 1$. So, we find

$$\frac{X}{E + \phi X} \leq 1.$$

Substituting inequalities (14)-(15) into (13), we conclude that $\hat{p} - p \leq \epsilon$. □
References
