

Optimizing Family Medicine Residency Schedules under the Clinic First Principles

Mohammad Al Syouf* Ankit Bansal[†] Bjorn P. Berg[‡] Jean-Philippe Richard[§]

Abstract

Family medicine residency programs must balance educational and operational requirements while providing residents with consistent exposure to continuity clinics, a central principle of the **Clinic First Model**. We study the Family Medicine Residency Scheduling problem and develop a binary integer programming (BIP) framework that incorporates Clinic First principles through two criteria: Interleaved Rotation Balancing (IRB), which promotes consistent clinic involvement within resident schedules, and Level Loading (LL), which promotes a balanced distribution of total resident clinic sessions across blocks. To provide a simple and interpretable schedule structure, we integrate the proposed BIP framework with the *staggered scheduling framework*, in which resident schedules are generated by successive temporal shifts of a common **base rotation schedule**. We establish a theoretical bound on the loss in IRB resulting from the *staggered scheduling framework* and derive valid inequalities that strengthen the resulting BIP formulation. Computational experiments based on data from the University of Minnesota Family Medicine Residency program demonstrate substantial computational benefits of the valid inequalities and show that the *staggered scheduling framework* incurs only a small loss in IRB. Compared with a schedule used in current practice, the proposed approach substantially improves IRB while producing a more balanced distribution of clinic sessions. Comparisons with an existing scheduling objective further demonstrate the benefits of explicitly incorporating IRB and LL.

Keywords: Family Medicine Residency Scheduling, Clinic First Model, Staggered Scheduling, Valid Inequalities, Convex Hull

Acknowledgment: This material is based upon work supported by the National Science Foundation under Grant Numbers 2427201, 2427202. Any opinions, findings, and conclusions or recom-

*School of Systems Science and Industrial Engineering, State University of New York-Binghamton, Binghamton, NY, USA

[†]School of Systems Science and Industrial Engineering, State University of New York-Binghamton, Binghamton, NY, USA, abansal@binghamton.edu (*Corresponding author*)

[‡]Division of Health Policy and Management, School of Public Health, University of Minnesota, Minneapolis, MN, USA

[§]Department of Industrial and Systems Engineering, University of Minnesota, Minneapolis, MN, USA

mentations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

1 Introduction

Primary care is a cornerstone of the U.S. healthcare system, promoting population health, preventing chronic disease, and serving as the primary point of contact for patients seeking medical care (Manolakis and Skelton 2010, Stange et al. 2023, Starfield et al. 2005). Strong primary care systems are associated with improved health outcomes, more equitable access to care, and lower healthcare costs (Starfield 2012, Bodenheimer and Sinsky 2014, Brown et al. 2019). Despite its importance, the United States faces a persistent shortage of primary care physicians (PCPs). The American Association of Medical Colleges projects a shortage of 21,100 to 55,200 PCPs by 2032 (Dall et al. 2017), highlighting the need to strengthen the primary care workforce.

Family medicine physicians constitute the largest segment of the primary care workforce, accounting for approximately 40% of all PCPs. Consequently, improving family medicine residency training is widely recognized as a critical strategy for addressing the projected shortage of PCPs (Adam et al. 2022, Zeller et al. 2019). A major challenge, however, is that many residency programs emphasize inpatient training at the expense of outpatient **continuity clinic** experiences, leading to reduced patient continuity, irregular clinic schedules, resident dissatisfaction, and ultimately fewer graduates pursuing careers in primary care (Gupta et al. 2016, Adam et al. 2022, Zeller et al. 2019, Hofkamp et al. 2020). These scheduling challenges are driven by the design of rotation clinic *templates*. Each rotation’s template dictates the specific half-day shifts a resident spends in their family medicine clinic during each rotation.

The **Clinic First Model** has emerged as a promising educational framework for addressing this imbalance by prioritizing **continuity clinic** experiences while maintaining required inpatient and specialty training (Gupta et al. 2016). Of its six guiding principles, we focus on the following two: “(1) *design resident schedules that prioritize continuity of care and eliminate tension between inpatient and outpatient duties, and (2) create operationally excellent clinics*”. Recent implementations of the **Clinic First Model** have demonstrated improvements in resident satisfaction, continuity of care, clinic operations, and preparation for careers in primary care (Paul et al. 2021, Hofkamp et al. 2020, Hersch et al. 2023, Adam et al. 2022). Realizing these benefits in practice requires embedding Clinic First principles into the annual rotation scheduling process. However, annual rotation schedules must simultaneously satisfy stringent accreditation requirements, specialty-specific educational objectives,

service staffing obligations, and continuity clinic availability (Bard et al. 2016b, 2017, Proano and Agarwal 2018, Guo et al. 2023). Incorporating Clinic First objectives while satisfying these competing operational and educational constraints substantially increases the combinatorial complexity of generating annual schedules.

In current practice, annual rotation schedules in family medicine residency programs are developed primarily through manual processes, with implementation supported by commercial scheduling systems such as Amion and QGenda (Day et al. 2006, Proano and Agarwal 2018). Creating schedules manually demands considerable time. Consequently, it prevents program administrators from systematically evaluating alternative scheduling policies. It also offers little adaptability when unexpected changes occur. As a result, opportunities to implement the **Clinic First** principles are often missed, adversely affecting resident experiences and the operational performance of the **continuity clinics**.

We use the term *Family Medicine Residency Scheduling* (**FMRs**) to denote the problem of determining annual rotation schedules for all residents in a family medicine residency program. Each academic year is divided into equal-length time periods referred to as blocks and an annual rotation schedule specifies the rotation assigned to each resident in every block of the academic year. Typical programs span three years, with cohorts of 6–10 residents per program year (PGY). Rotation-specific templates allocate resident time between specialty training, programmatic duties, and their family medicine clinic. Because these templates govern resident availability across the planning horizon, rotation scheduling directly impacts both educational outcomes and clinic operational performance.

In this paper, we formulate a binary integer programming (BIP) model for **FMRs** that explicitly incorporates **Clinic First Model** principles into annual rotation design. Specifically, the proposed model evaluates and structures schedules around two key criteria:

- **Interleaved Rotation Balancing (IRB)**: The required number of clinic sessions varies across rotations due to specialty-specific training requirements and faculty availability. Consequently, assigning a resident to several consecutive rotations with high clinic demands causes prolonged fatigue, whereas consecutive light-clinic rotations disrupt patient continuity. To prevent extended stretches of either extreme, we introduce the Interleaved Rotation Balancing (**IRB**) metric, defined as the sum of absolute differences in the number of clinic sessions between consecutive rotations. Maximizing **IRB** encourages schedules that systematically alternate between high- and low-demand rotations, thereby distributing clinical involvement more evenly over the planning horizon while avoiding multi-month fatigue clusters. This criterion directly aligns with guiding principle (1) of the **Clinic First Model**.

- **Level Loading (LL):** The rotation schedule also determines the total number of resident clinic sessions available during each block and, consequently, the staffing levels of the **continuity clinic**. Large fluctuations in available clinic capacity across blocks may reduce patient access, create operational inefficiencies, and adversely affect residents’ clinical experiences. To manage clinic-wide operations, we introduce the *Level Loading* (LL) criterion, which evaluates the aggregate number of resident clinic sessions offered in each block. Our model enforces a lower bound on total block sessions through a set of constraints, ensuring a balanced, predictable supply of clinic capacity throughout the planning horizon. This criterion directly aligns with guiding principles (1) and (2) of the **Clinic First Model**.

To bridge the gap between complex optimization models and practical administration, we integrate our BIP model with a *staggered scheduling framework*, wherein each resident’s schedule is generated by cyclic-shifting a single base rotation sequence across the planning horizon. This structure provides every resident with identical curricular exposure in distinct temporal orders, yielding transparent and easily communicable schedules. While restricting the decision space to staggered configurations trades off full combinatorial flexibility, it preserves the intuitive simplicity typical of manual heuristics. To quantify this design trade-off, we establish a theoretical bound on the potential loss in **IRB** resulting from the staggered framework. The resulting BIP formulation is computationally challenging due to the complex interplay of resident and clinic constraints. Integrating the *staggered scheduling framework* provides an initial structural advantage: because every resident’s schedule is a cyclic permutation of a single base template, both model dimensionality and scheduling symmetry are substantially reduced. Nevertheless, solving realistic, full-scale instances remains computationally challenging. Hence, we derive a set of valid inequalities to tighten the BIP formulation and show that they provide the convex hull of a subset of the constraints in the formulation.

Computational experiments based on data from a family medicine residency program at the University of Minnesota (UMN) demonstrate that the proposed valid inequalities substantially improve the computational efficiency of the branch-and-bound algorithm for solving the BIP formulation. The experiments further show that the proposed approach produces schedules with substantially higher **IRB** and a more balanced distribution of clinic sessions than the schedule currently implemented in practice. Additionally, we demonstrate the benefits of incorporating the **IRB** and **LL** criteria by comparing the schedules generated by our proposed approach with those obtained using an alternative objective from the existing literature.

The remainder of the paper is organized as follows. Section 2 reviews related literature. Section 3

describes the **FMR**S problem, integrating the two **Clinic First** criteria and the *staggered scheduling framework*. Sections 4 and 5 introduce the proposed BIP formulation and the valid inequalities used to strengthen it, respectively. Section 6 presents the computational results, and Section 7 provides concluding remarks.

2 Literature Review

Physician scheduling has received considerable attention in the literature over the past several decades (Erhard et al. 2018). Resident rotation scheduling constitutes an important class of physician scheduling problems and involves assigning residents to rotations over a sequence of blocks while satisfying program-specific requirements and optimizing relevant objectives. Among the early contributions, Franz and Miller (1993) developed an integer programming (IP) formulation for resident rotation scheduling that focuses on optimizing resident preferences. Subsequently, Guo et al. (2014) studied a generic rotation scheduling problem incorporating constraints commonly encountered in residency programs and established that such a problem is *NP-hard*.

There is no universally applicable model for resident rotation scheduling, as program requirements and scheduling priorities vary across medical specialties and countries (Li et al. 2025, Kraul 2020, Guo et al. 2023, Hong et al. 2019). Consequently, the literature has considered a variety of objectives, including resident preferences (Akbarzadeh and Maenhout 2021a,b, Lemay et al. 2017, Smalley and Keskinocak 2016), workload management (Kraul 2020), and equity among residents (Li et al. 2025, Smalley and Keskinocak 2016). Despite these advances, relatively limited attention has been devoted to rotation scheduling in specialties where experience in the **continuity clinic** is an integral component of residency training.

Family medicine and internal medicine are two specialties in which residents’ experiences in the **continuity clinic** play an important role in their decision to continue practicing in the specialty after completing residency (Peccoraro et al. 2013, Gupta et al. 2016, Adam et al. 2022). From a rotation scheduling perspective, an important distinction between the two specialties lies in how **continuity clinic** assignments are incorporated into rotations. In family medicine, rotations are typically associated with predefined templates that specify the **continuity clinic** sessions to be completed during each week of the rotation. In contrast, internal medicine rotation scheduling generally does not rely on such rotation-specific clinic templates (Bard et al. 2017).

Within the internal medicine setting, Bard et al. (2016a, 2017) develop IP-based approaches for

annual resident rotation scheduling that explicitly account for clinic assignments. Their models seek, among other objectives, to balance the total number of clinic sessions across residents over the year and the number of patients seen by each clinic across months. Bard et al. (2016a) consider a 4 + 1 scheduling structure and obtain optimal solutions by decomposing the underlying model into two subproblems. In contrast, Bard et al. (2017) consider a more general scheduling setting that results in a substantially more complex optimization problem, which they address using a heuristic. Proano and Agarwal (2018) develop a multi-objective integer programming framework for annual rotation scheduling of internal medicine residents. They employ the Analytical Hierarchy Process (AHP) to incorporate multiple scheduling criteria and promote equity among residents. Their objectives include, among others, maximizing resident satisfaction with vacation assignments and increasing assignments to elective rotations. Wang et al. (2022) also develop an IP formulation for annual block rotation scheduling of internal medicine residents. Their formulation seeks to avoid assigning residents to heavy-workload rotations in consecutive blocks and to distribute rotations evenly across the planning horizon, with the goal of maintaining the quality of patient care. They employ a heuristic to solve the resulting model.

Within the family medicine setting, Bard et al. (2016b) develop an IP-based approach for determining annual rotation schedules for family medicine residents. Their objective is to minimize deviations from predetermined rotation templates while balancing the number of patients seen per clinic session across the planning horizon. Given the computational complexity of the resulting formulation, they develop a heuristic solution approach.

Despite these contributions, aforementioned scheduling models in internal and family medicine rarely model the specific continuity dynamics emphasized by the **Clinic First Model**, nor do they account for how schedule structure impacts longitudinal resident experience in the **continuity clinic**. Furthermore, while existing optimization approaches promote uniform training experiences across residents, they often yield irregular schedule patterns that increase administrative complexity and reduce schedule predictability for residents and clinic leadership. Our work addresses these gaps by incorporating **Clinic First** objectives (specifically balancing individual continuity and aggregate clinic loading) into a unified framework that produces curricularly uniform, structurally predictable, and operationally viable schedules.

To summarize, the primary contributions of this paper are fourfold:

1. **Integrated FMRS Formulation:** We present a BIP model for family medicine residency scheduling that operationalizes the core principles of the Clinic First Model. Specifically, the formulation

maximizes IRB while enforcing a lower bound LL.

2. *Practical Staggered Scheduling Framework* : We integrate a cyclic-shifted base scheduling framework that guarantees identical curricular exposure across residents while yielding transparent, easily communicable schedules. We formally establish a theoretical bound on the maximum potential loss in IRB incurred by restricting the solution space to staggered configurations.
3. *Polyhedral Theoretical Analysis*: We derive a family of valid inequalities to tighten the linear programming relaxation of the BIP model and prove that these inequalities define the convex hull for a key subset of constraints.
4. *Empirical Validation & Comparative Insights*: Using real-world data from a University of Minnesota family medicine residency program, we demonstrate the computational effectiveness of our valid inequalities. Furthermore, we benchmark our schedules against current administrative practice and existing literature (Bard et al. 2016a), providing actionable insights into the structural trade-offs between continuity (IRB) and clinic capacity (LL).

3 Problem Description and Staggered Scheduling Framework

Let Y denote the set of PGYs. For each PGY $y \in Y$, let R_y and S_y represent the sets of rotations and residents, respectively. Define $\gamma_y := |S_y|$ as the number of residents in PGY y . For each rotation $r \in R_y$, let $\bar{c}_r^y$ denote the number of clinic sessions associated with that rotation. Finally, let $M = 1, \dots, |M|$ denote the set of equal-length discrete blocks within an academic year. Once the rotation schedules are determined, let $\mathbf{r}(i, y, m) \in R_y$ denote the rotation assigned to resident $i \in S_y$ during block $m \in M$. Since each resident is assigned to exactly one rotation in each block and each rotation is assigned exactly once over the calendar year, it follows that $|R_y| = |M|$ for every $y \in Y$. This is without loss of generality, as multiple rotations within the same specialty during an academic year can be represented as distinct rotations in the model. Accordingly, the schedule of resident $i \in S_y$ is represented by the sequence $\omega_i^y := \{\mathbf{r}(i, y, m)\}_{m \in M}$, where ω_i^y is a permutation of the rotations in R_y .

In traditional residency schedules, outpatient clinic responsibilities are often subordinated to inpatient requirements, leading to irregular and fragmented clinic exposure. The **Clinic First Model** addresses this imbalance by placing greater emphasis on **continuity clinic** while better managing the competing demands of outpatient clinic responsibilities and inpatient training. To assess the quality of a resident schedule, we consider two performance measures that capture the primary goals of

the **Clinic First Model**. The first measure, **IRB**, quantifies how evenly clinic sessions are distributed over the course of the academic year. In our optimization model, **IRB** is the objective to be maximized. A higher value of **IRB** corresponds to a more balanced distribution of clinic sessions across blocks for each resident, preventing extended intervals with either too few or too many clinic sessions. Consequently, maximizing **IRB** promotes regular and consistent outpatient clinic participation throughout the year. For a given schedule, $\text{IRB} := \sum_{y \in Y} \sum_{i \in S_y} \sum_{m=2}^{|M|} \left| \bar{c}_{\mathbf{r}(i,y,m)}^y - \bar{c}_{\mathbf{r}(i,y,m-1)}^y \right|$.

The second performance measure, *Level Loading* (**LL**), evaluates the distribution of clinic capacity across blocks. Specifically, **LL** is defined as the minimum total number of clinic sessions scheduled in any block, *i.e.*, $\text{LL} := \min_{m \in M} \left(\sum_{y \in Y} \sum_{i \in S_y} \bar{c}_{\mathbf{r}(i,y,m)}^y \right)$. Because the number of clinic sessions in each rotation varies, the total number clinic sessions per block across all residents could be inconsistent across the year if individual resident schedules are not coordinated at the clinic level. Too few clinic sessions in a block results in poor access to care for patients, while too many clinic sessions in a block can result in congestion in the clinic and over-utilization of shared staff and resources. In our proposed model, we require **LL** to exceed a prescribed threshold $\bar{L}$, thereby guaranteeing a minimum level of clinic capacity in every block. Enforcing a lower bound, $\bar{L}$ on monthly resident clinic sessions forces the model to balance annual capacity, naturally suppressing extreme monthly spikes by redistributing sessions across the year. This requirement is enforced through the constraints $\sum_{y \in Y} \sum_{i \in S_y} \bar{c}_{\mathbf{r}(i,y,m)}^y \geq \bar{L}, \forall m \in M$.

To construct resident schedules, we adopt a *staggered scheduling framework*, in which each PGY is represented by a single **base rotation schedule** and the schedules of all residents are obtained by successive one-block right shifts of this sequence. In this study, we restrict these successive shifts to one block, leaving the exploration of larger shifts for future research. As an illustration, consider a PGY y with $|R_y| = |M| = 5$ and $\gamma_y = 4$. Let the **base rotation schedule** be $(\hat{r}_1^y, \hat{r}_2^y, \hat{r}_3^y, \hat{r}_4^y, \hat{r}_5^y)$, where $\hat{r}_j^y \in R_y \forall j \in \{1, 2, \dots, 5\}$ and $\hat{r}_i^y \neq \hat{r}_j^y$ whenever $i \neq j$. Figure 1 illustrates how the schedules of all four residents are generated from this common sequence. The **base rotation schedule** is assigned to Resident 1, while the schedule of Resident 2 is obtained by shifting Resident 1's schedule one block to the right. Thus, if Resident 1 follows $(\hat{r}_1^y, \hat{r}_2^y, \hat{r}_3^y, \hat{r}_4^y, \hat{r}_5^y)$, then Resident 2 follows $(\hat{r}_5^y, \hat{r}_1^y, \hat{r}_2^y, \hat{r}_3^y, \hat{r}_4^y)$. The schedules of the remaining residents are generated analogously by repeatedly applying a one-block right shift to the preceding resident's schedule. Building on the *staggered scheduling framework*, the next section presents a BIP formulation for determining an optimal **base rotation schedule** for each PGY, while maximizing **IRB** and satisfying the **LL** constraints.

While our formulation does not explicitly embed program-specific constraints due to their idiosyn-

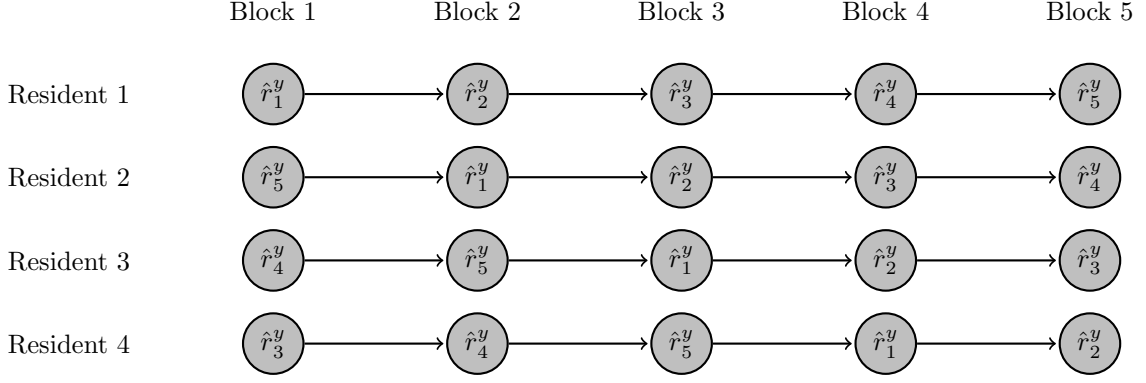


Figure 1: Example of the *staggered scheduling framework*.

cratic nature, its design allows such rules to be easily integrated. Common examples of such requirements include governing contractual coverage agreements (e.g., staffing the emergency department or OB/GYN shifts), rotation spacing guidelines, or mandatory educational sequencing. Incorporating these tailored requirements within the *staggered scheduling framework* underscores a modeling tradeoff: restrict the solution space to cyclic-shifted base schedules reduces the total space of possible schedules, but in exchange delivers substantial computational benefit in reduced dimensionality and breaking symmetry between residents. Furthermore, as the model determines an optimal base schedule without assigning specific residents to individual schedules, it maintains discretion of program leadership to include preferences, elective choices, or resident accommodations within the optimal schedule. The *staggered scheduling framework* introduced in this paper provides a foundation for developing more flexible scheduling models capable of generating a broader range of rotation schedules, as discussed in the final paragraph of Section 7.

4 BIP Formulation for Staggered Scheduling

In this section, we present a BIP formulation for FMRS under the *staggered scheduling framework* that maximizes IRB while enforcing a minimum level of LL. From a modeling perspective, the *staggered scheduling framework* has the advantage that only a single **base rotation schedule** needs to be determined for each PGY, from which the schedules of all residents in that PGY are generated. This substantially reduces the number of scheduling decisions compared with an analogous formulation of FMRS without the *staggered scheduling framework*.

Let $P := M$ denote the set of positions in a **base rotation schedule**. For each PGY $y \in Y$, let x_{rp}^y be a binary variable that equals 1 if rotation $r \in R_y$ is assigned to position $p \in P$ in the **base**

rotation schedule, and 0 otherwise. To model consecutive rotations, define $I_y := \{(r_1, r_2) \in R_y \times R_y : r_1 < r_2\}$, the set of all unordered pairs of distinct rotations in PGY y . For each $(r_1, r_2) \in I_y$ and $p \in P$, let $\eta_{r_1 r_2 p}^y$ be a binary variable that equals 1 if rotations r_1 and r_2 occupy consecutive positions in the **base rotation schedule**, with the first rotation appearing in position p , and 0 otherwise. Here, position 1 is treated as the successor of position $|P|$. Since each rotation is assigned to exactly one position in the **base rotation schedule**, we have $\eta_{r_1 r_2 p}^y = (x_{r_1 p}^y + x_{r_1 p+1}^y)(x_{r_2 p}^y + x_{r_2 p+1}^y)$.

To model the LL bound, we identify the rotations occurring in each block across the resident schedules generated from a **base rotation schedule** under the *staggered scheduling framework*. To this end, for each PGY $y \in Y$ and block $m \in M$, let $\bar{P}_m^y \subseteq P$ denote the set of positions in the **base rotation schedule** whose assigned rotations appear in block m in the resident schedules generated under the *staggered scheduling framework*. For example, consider the staggered schedule in Figure 1. During the first block, the assigned rotations are $\hat{r}_1^y$, $\hat{r}_5^y$, $\hat{r}_4^y$, and $\hat{r}_3^y$. These rotations occupy positions 1, 5, 4, and 3 in the **base rotation schedule**, respectively, yielding $\bar{P}_1^y = \{1, 5, 4, 3\}$. Thus, under the *staggered scheduling framework*, $\bar{P}_m^y = \{m - i + 1 \mid i \in \{1, 2, \dots, m\}\} \cup \{|P| - (i - m - 1) \mid i \in \{m + 1, m + 2, \dots, \gamma_y\}\}$

Let $\hat{c}_{r_1 r_2 p}^y$ denote the contribution to IRB across all residents in PGY y resulting from assigning rotations r_1 and r_2 to consecutive positions p and $p+1$, respectively, in the **base rotation schedule**. Thus, for each PGY $y \in Y$, $(r_1, r_2) \in I_y$, and $p \in P$,

$$\hat{c}_{r_1 r_2 p}^y = \begin{cases} \gamma_y |\bar{c}_{r_1}^y - \bar{c}_{r_2}^y|, & \text{if } p \leq |P| - \gamma_y, \\ (\gamma_y - 1) |\bar{c}_{r_1}^y - \bar{c}_{r_2}^y|, & \text{if } p > |P| - \gamma_y. \end{cases} \quad (1)$$

The weighting by γ_y in (1) accounts for the number of resident schedules to which each consecutive rotation pair contributes under the *staggered scheduling framework*. If the pair begins at a position $p \leq |P| - \gamma_y$ in the **base rotation schedule**, it appears consecutively in the schedules of all γ_y residents. Otherwise, it appears consecutively in only $\gamma_y - 1$ resident schedules. For ease of reference, Table 1 summarizes the notation used throughout the paper. Using these notations, we formulate the following BIP model to determine an optimal **base rotation schedule** for each PGY.

$$(\mathcal{H}) \quad \max \sum_{y \in Y} \sum_{(r_1, r_2) \in I_y} \sum_{p \in P} \hat{c}_{r_1 r_2 p}^y \eta_{r_1 r_2 p}^y \quad (2a)$$

Notation	Description
Sets and Indices	
Y	Set of PGYs
R_y	Set of rotations in PGY y
S_y	Set of residents in PGY y
M	Set of blocks
P	Set of positions in the base rotation schedule
I_y	Set of unordered pairs of distinct rotations in R_y , i.e., $\{(r_1, r_2) \in R_y \times R_y : r_1 < r_2\}$
$\bar{P}_m^y$	Set of positions in the base rotation schedule of PGY y whose rotations are assigned to residents in block m
Parameters	
γ_y	Number of residents in PGY y
$\bar{c}_r^y$	Number of clinic sessions associated with rotation r in PGY y
$\bar{L}$	Lower bound on LL
$\hat{c}_{r_1 r_2 p}^y$	Contribution to IRB from assigning rotations r_1 and r_2 consecutively starting at position p of the base rotation schedule for PGY y
Decision Variables	
x_{rp}^y	Binary variable equal to 1 if rotation r is assigned to position p in the base rotation schedule of PGY y , and 0 otherwise
$\eta_{r_1 r_2 p}^y$	Binary variable equal to 1 if rotations r_1 and r_2 are scheduled consecutively starting at position p (in either order) in the base rotation schedule of PGY y , and 0 otherwise

Table 1: Summary of notation.

$$\sum_{r \in R_y} x_{rp}^y = 1 \quad \forall y \in Y, p \in P \quad (2b)$$

$$\sum_{p \in P} x_{rp}^y = 1 \quad \forall y \in Y, r \in R_y \quad (2c)$$

$$\eta_{r_1 r_2 p}^y \leq x_{r_1 p}^y + x_{r_1 p+1}^y \quad \forall p \in P, (r_1, r_2) \in I_y, y \in Y \quad (2d)$$

$$\eta_{r_1 r_2 p}^y \leq x_{r_2 p}^y + x_{r_2 p+1}^y \quad \forall p \in P, (r_1, r_2) \in I_y, y \in Y \quad (2e)$$

$$\sum_{y \in Y} \sum_{p \in \bar{P}_m^y} \sum_{r \in R_y} \bar{c}_r^y x_{rp}^y \geq \bar{L} \quad \forall m \in M \quad (2f)$$

$$\eta_{r_1 r_2 p}^y \in \{0, 1\} \quad \forall p \in P, (r_1, r_2) \in I_y, y \in Y \quad (2g)$$

$$x_{rp}^y \in \{0, 1\} \quad \forall r \in R_y, p \in P, y \in Y. \quad (2h)$$

The objective function (2a) maximizes the IRB. For each PGY, Constraints (2b) and (2c) ensure that each position is assigned exactly one rotation and each rotation is assigned to exactly one position. Constraint (2d) permits $\eta_{r_1 r_2 p}^y$ to equal 1 only if rotation r_1 is assigned to either position p or $p+1$, while Constraint (2e) imposes the analogous condition for rotation r_2 . Together, these constraints linearize the bilinear inequality

$$\eta_{r_1 r_2 p}^y \leq \left(x_{r_1 p}^y + x_{r_1 p+1}^y \right) \left(x_{r_2 p}^y + x_{r_2 p+1}^y \right) \quad \forall p \in P, (r_1, r_2) \in I_y, y \in Y, \quad (3)$$

ensuring that $\eta_{r_1 r_2 p}^y$ is 1 only when rotations r_1 and r_2 occupy consecutive positions in the **base rotation schedule** starting at position p . Since the objective function coefficients in (2a) are non-negative, we can assume, without loss of generality, that this inequality is binding at optimality. Finally, Constraint (2f) enforces the LL requirement by ensuring that the total number of clinic sessions scheduled across all residents in every block are at least the prescribed lower bound $\bar{L}$.

The following proposition establishes the equivalence of the right- and left-shifting counterparts of the *staggered scheduling framework* in terms of the optimal objective value and resident rotation schedules generated at optimality. Therefore, without loss of generality, we restrict our attention to the right-shifting *staggered scheduling framework* throughout this paper. Let $\mathcal{H}^L$ denote the formulation analogous to $\mathcal{H}$ under the left-shifting counterpart of the *staggered scheduling framework*, where the objective function coefficients $\hat{c}_{r_1 r_2 p}^y$ and the sets $\bar{P}_m^y$ are redefined to account for their left-shifting. All other components of the formulation remain unchanged.

Proposition 1. *A set of resident rotation schedules can be generated under the right-shifting staggered scheduling framework if and only if they can be generated under the left-shifting staggered scheduling framework. Consequently, $\mathcal{H}$ and $\mathcal{H}^L$ have the same optimal objective value and generate the same set of resident rotation schedules at optimality.*

Proof. Fix a PGY $y \in Y$ and consider any feasible **base rotation schedule** $\bar{\omega}_1^y$ under the right-shifting *staggered scheduling framework*. Let $\bar{\omega}_i^y, i = 1, \dots, \gamma_y$, denote the resident rotation schedules generated by successively right-shifting $\bar{\omega}_1^y$. Taking $\bar{\omega}_{\gamma_y}^y$ as the **base rotation schedule** under the left-shifting *staggered scheduling framework* and successively left-shifting it generates the same set of resident rotation schedules. Hence, every set of resident rotation schedules that can be generated under the right-shifting *staggered scheduling framework* can also be generated under the left-shifting *staggered scheduling framework*. The reverse follows analogously. Given any feasible **base rotation schedule** under the left-shifting *staggered scheduling framework*, taking the last schedule generated by successive left shifts as the **base rotation schedule** under the right-shifting *staggered scheduling framework* and successively right-shifting it recovers the same set of resident rotation schedules.

As $\mathcal{H}$ and $\mathcal{H}^L$ generate the same sets of resident rotation schedules and maximize IRB over these schedules, they have the same optimal objective value and generate same resident rotation schedules at optimality. $\square$

Next, we establish a theoretical bound on the loss in IRB due to the *staggered scheduling framework*. Let $\mathcal{P}$ (presented in Appendix A) denote the BIP formulation for FMRS without the *staggered scheduling*

framework, which maximizes IRB subject to the LL constraints. Since $\mathcal{P}$ is not restricted to schedules satisfying the *staggered scheduling framework*, we refer to it as the *unrestricted formulation*.

Proposition 2. For each PGY $y \in Y$, let $\underline{\nu}^y = \min_{\substack{r_1, r_2 \in R_y \\ r_1 \neq r_2}} |\bar{c}_{r_1}^y - \bar{c}_{r_2}^y|$ and $\bar{\nu}^y = \max_{\substack{r_1, r_2 \in R_y \\ r_1 \neq r_2}} |\bar{c}_{r_1}^y - \bar{c}_{r_2}^y|$. Furthermore, let $z_{\mathcal{P}}^*$ and $z_{\mathcal{H}}^*$ denote the optimal objective values of $\mathcal{P}$ and $\mathcal{H}$, respectively. Then, $z_{\mathcal{P}}^* - z_{\mathcal{H}}^* \leq \sum_{y \in Y} ((\gamma_y |R_y| - 1) \bar{\nu}^y - (\gamma_y - 1) \underline{\nu}^y)$.

Proof. Let $z_{\mathcal{H}_{NL}}^*$ denote the optimal objective value of $\mathcal{H}$ without the LL constraints (2f) and $\tilde{\omega}_1^y$ be the rotation schedule for PGY y with the maximum single resident IRB value, $\bar{C}^y$. We have the following inequalities:

$$z_{\mathcal{P}}^* \leq \sum_{y \in Y} \gamma_y \bar{C}^y, \quad (4a)$$

$$z_{\mathcal{H}_{NL}}^* \geq \sum_{y \in Y} \gamma_y \bar{C}^y - \sum_{y \in Y} (\gamma_y - 1) (\bar{\nu}^y - \underline{\nu}^y), \quad (4b)$$

$$z_{\mathcal{H}_{NL}}^* \leq \sum_{y \in Y} \gamma_y (|R_y| - 1) \bar{\nu}^y. \quad (4c)$$

Inequality (4a) follows immediately from the definition of $\bar{C}^y$. To establish inequality (4b), we exploit the structure of the *staggered scheduling framework*. Clearly, $\tilde{\omega}_1^y$ is a feasible **base rotation schedule** for $\mathcal{H}$ without the LL constraints (2f) where the **base rotation schedule** attains IRB of $\bar{C}^y$. Each subsequent schedule is obtained by a one-block right shift of the preceding schedule and therefore differs from the **base rotation schedule** by the removal of one consecutive rotation pair and the addition of another. Consequently, the IRB value of each shifted schedule can decrease by at most $\bar{\nu}^y - \underline{\nu}^y$, implying that every such schedule has IRB of at least $\bar{C}^y - (\bar{\nu}^y - \underline{\nu}^y)$. Since there are $\gamma_y - 1$ shifted schedules in PGY y , summing the IRB contribution of the **base rotation schedule** and these lower bounds over all other schedules establishes inequality (4b).

Inequality (4c) follows from the observation that each resident schedule in PGY y contains exactly $|R_y| - 1$ consecutive rotation pairs. Since the contribution of any consecutive rotation pair to the IRB is at most $\bar{\nu}^y$, the IRB of an individual resident schedule is bounded above by $(|R_y| - 1) \bar{\nu}^y$. Summing this bound over all γ_y residents and then over all PGYs establishes inequality (4c).

From (4a) and (4b), we obtain $z_{\mathcal{P}}^* - z_{\mathcal{H}_{NL}}^* \leq \sum_{y \in Y} (\gamma_y - 1) (\bar{\nu}^y - \underline{\nu}^y)$. From (4c) and $z_{\mathcal{H}}^* \geq 0$ we obtain $z_{\mathcal{H}_{NL}}^* - z_{\mathcal{H}}^* \leq \sum_{y \in Y} \gamma_y (|R_y| - 1) \bar{\nu}^y$. Adding these two inequalities, we obtain that $z_{\mathcal{P}}^* - z_{\mathcal{H}}^* \leq \sum_{y \in Y} ((\gamma_y |R_y| - 1) \bar{\nu}^y - (\gamma_y - 1) \underline{\nu}^y)$. $\square$

Proposition 2 provides a bound on the loss in IRB due to the *staggered scheduling framework*.

This bound decreases as the number of residents, the number of rotations, and $\bar{\nu}^y$ decrease, and as $\underline{\nu}^y$ increases. Furthermore, as expected, the bound on the loss in IRB under the *staggered scheduling framework* approaches zero as $\bar{\nu}^y$ approaches zero. When $\bar{\nu}^y$ is small, the number of clinic sessions differs only slightly across rotations. Consequently, replacing one pair of consecutive rotations with another can result in only a small change in the corresponding IRB contribution, thereby limiting the loss associated with the *staggered scheduling framework*. Consistent with this result, our computational experiments in Section 6 show that imposing the *staggered scheduling framework* results in only a small loss in IRB relative to the *unrestricted formulation*.

5 Valid Inequalities for $\mathcal{H}$

Our computational experiments indicate that the linear programming (LP) relaxation of $\mathcal{H}$ is weak, leading to poor computational performance. To address this weakness, we strengthen the linearization of the bilinear inequality (3) by deriving two families of valid inequalities, VI-a and VI-b. We show that VI-b are strong in the sense that they provide a complete description of the convex hull of the discrete set defined by VI-a, and a subset of constraints (2c)–(2e), thereby yielding a stronger linearization of (3) than the one given by (2d)–(2e). Computational experiments (described in Section 6) demonstrate that incorporating VI-b substantially reduces the time required to solve $\mathcal{H}$.

Proposition 3. *The following inequalities are valid for $\mathcal{H}$:*

$$(VI-a) \quad \sum_{r \in R_y: r < k} \eta_{rkp}^y + \sum_{r \in R_y: r > k} \eta_{krp}^y \leq 1 \quad \forall p \in P, \forall y \in Y, \forall k \in R_y. \quad (5)$$

Proof. Let $(\mathbf{x}^*, \boldsymbol{\eta}^*)$ be a feasible solution to $\mathcal{H}$ such that $\sum_{r \in R_y: r < k} \eta_{rkp}^{y*} + \sum_{r \in R_y: r > k} \eta_{krp}^{y*} > 1$ for some $y \in Y$, $k \in R_y$, and $p \in P$. By (2g), there exist distinct rotations $r_1, r_2 \in R_y \setminus k$ satisfying one of the following:

- $r_1 < r_2 < k$, with $\eta_{r_1kp}^{y*} = 1$ and $\eta_{r_2kp}^{y*} = 1$;
- $r_1 < k < r_2$, with $\eta_{r_1kp}^{y*} = 1$ and $\eta_{kr_2p}^{y*} = 1$;
- $k < r_1 < r_2$, with $\eta_{kr_1p}^{y*} = 1$ and $\eta_{kr_2p}^{y*} = 1$.

In the first case where $r_1 < r_2 < k$, from (2d) and (2e) imply $x_{r_1p}^{y*} + x_{r_1p+1}^{y*} \geq 1$, $x_{r_2p}^{y*} + x_{r_2p+1}^{y*} \geq 1$, $x_{kp}^{y*} + x_{kp+1}^{y*} \geq 1$. Together with (2c) and (2h), this implies that rotations r_1 , r_2 , and k must all

be assigned to positions p and $p + 1$, contradicting (2b). The remaining two cases lead to similar contradictions. $\square$

The following example demonstrates that VI-a can eliminate some fractional solutions that are feasible to the LP relaxation of $\mathcal{H}$ where $\mathbf{x}_i^y = \{x_{ip}^y\}_{p \in P}$.

Example 1. Consider an instance of $\mathcal{H}$ with $|Y| = 1$, $R_1 = P = 1, 2, 3, 4$, and $\bar{L} = 0$. The solution defined by $\mathbf{x}_1^1 = \{0.5, 0, 0.5, 0\}$, $\mathbf{x}_2^1 = \{0.5, 0.5, 0, 0\}$, $\mathbf{x}_3^1 = \{0, 0, 0.5, 0.5\}$, $\mathbf{x}_4^1 = \{0, 0.5, 0, 0.5\}$, $\eta_{122}^1 = \eta_{232}^1 = \eta_{242}^1 = 0.5$, with all other η^1 variables equal to zero, is feasible for the LP relaxation of $\mathcal{H}$ but violates VI-a for $p = 2$, $k = 2$, and $y = 1$.

The inequalities VI-a enforce that each rotation $k \in R_y$ can be paired with at most one other rotation $r \neq k$ to form a consecutive pair in the **base rotation schedule**, with the earlier rotation of the pair assigned to position p . In the following proposition, we strengthen VI-a.

Proposition 4. The following inequalities are valid for $\mathcal{H}$:

$$(VI-b) \quad \sum_{r \in R_y: r > k} \eta_{krp}^y + \sum_{r \in R_y: r < k} \eta_{rkp}^y \leq x_{kp}^y + x_{kp+1}^y \quad \forall p \in P, \forall y \in Y, \forall k \in R_y. \quad (6)$$

Proof. Let $(\mathbf{x}^*, \boldsymbol{\eta}^*)$ be a feasible solution to $\mathcal{H}$ such that $\sum_{r \in R_y: r > k} \eta_{krp}^{y*} + \sum_{r \in R_y: r < k} \eta_{rkp}^{y*} > x_{kp}^{y*} + x_{kp+1}^{y*}$ for some $y \in Y$, $k \in R_y$, and $p \in P$. By VI-a and (2c), it follows that $\sum_{r \in R_y: r > k} \eta_{krp}^{y*} + \sum_{r \in R_y: r < k} \eta_{rkp}^{y*} = 1$, and $x_{kp}^{y*} = x_{kp+1}^{y*} = 0$. By (2d), (2e), and (2g), this implies $\eta_{krp}^{y*} = 0$, $\forall r \in R_y : r > k$, and $\eta_{rkp}^{y*} = 0$, $\forall r \in R_y : r < k$, contradicting $\sum_{r \in R_y: r > k} \eta_{krp}^{y*} + \sum_{r \in R_y: r < k} \eta_{rkp}^{y*} = 1$. $\square$

VI-b ensure that each rotation $k \in R_y$ can be paired with at most one other rotation $r \neq k$ to form a consecutive pair in the **base rotation schedule**, with the earlier rotation of the pair assigned to position p only if k is assigned to position p or $p + 1$. VI-b dominates VI-a, since both families share the same left-hand side, while the right-hand side of VI-b satisfies $x_{kp}^y + x_{kp+1}^y \leq 1$ due to (2c).

Next, we establish that inequalities VI-b are strong by proving that they describe the convex hull of the discrete set defined by VI-a and a subset of the constraints in (2c)-(2e). Specifically, Let $\mathcal{Z}_k^y$ denote the set defined by the following constraints:

$$\sum_{p \in P} x_{kp}^y = 1 \quad (7a)$$

$$\eta_{krp}^y \leq x_{kp}^y + x_{kp+1}^y \quad \forall p \in P, \forall r \in R_y : r > k \quad (7b)$$

$$\eta_{rkp}^y \leq x_{kp}^y + x_{kp+1}^y \quad \forall p \in P, \forall r \in R_y : r < k \quad (7c)$$

$$\sum_{r \in R_y: r < k} \eta_{rkp}^y + \sum_{r \in R_y: r > k} \eta_{krp}^y \leq 1 \quad \forall p \in P \quad (7d)$$

$$\eta_{r_1 r_2 p}^y \in \{0, 1\} \quad \forall p \in P, \forall (r_1, r_2) \in I_y : k \in \{r_1, r_2\} \quad (7e)$$

$$x_{kp}^y \in \{0, 1\} \quad \forall p \in P. \quad (7f)$$

Let $\mathcal{L}_k^y$ denote the linear relaxation of $\mathcal{Z}_k^y$. For each $y \in Y$, let $\bar{\mathcal{Z}}^y$ denote the binary set defined by constraints (2c)–(2e), VI-a, (2g), and (2h), and let $\bar{\mathcal{L}}^y$ denote its linear relaxation. Then, $\bar{\mathcal{Z}}^y = \bigcap_{k \in R_y} \mathcal{Z}_k^y$ and $\bar{\mathcal{L}}^y = \bigcap_{k \in R_y} \mathcal{L}_k^y$. Consequently, $\text{conv}(\bar{\mathcal{Z}}^y) \subseteq \bigcap_{k \in R_y} \text{conv}(\mathcal{Z}_k^y) \subseteq \bigcap_{k \in R_y} \mathcal{L}_k^y = \bar{\mathcal{L}}^y$. Thus, characterizing $\text{conv}(\mathcal{Z}_k^y)$ may yield a stronger linear relaxation of $\bar{\mathcal{Z}}^y$, and hence a stronger linearization of the bilinear inequality (3) than that provided by constraints (2d)–(2e). The following proposition establishes that replacing constraints (7b)–(7d) in (7) with VI-b and relaxing the integrality restrictions provides a complete description of $\text{conv}(\mathcal{Z}_k^y)$.

Proposition 5. *The following set of constraints*

$$\sum_{p \in P} x_{kp}^y = 1 \quad (8a)$$

$$\sum_{r \in R_y: r < k} \eta_{rkp}^y + \sum_{r \in R_y: r > k} \eta_{krp}^y \leq x_{kp}^y + x_{kp+1}^y \quad \forall p \in P \quad (8b)$$

$$\eta_{r_1 r_2 p}^y \geq 0 \quad \forall p \in P, \forall (r_1, r_2) \in I_y : k \in \{r_1, r_2\} \quad (8c)$$

$$x_{kp}^y \geq 0 \quad \forall p \in P \quad (8d)$$

completely describe $\text{conv}(\mathcal{Z}_k^y)$.

Proof. Let $\mathcal{Q}_k^y$ denote the polyhedron defined by (8), and let $\bar{\mathcal{Q}}_k^y$ be the set of binary points satisfying (8). We show that (i) $\mathcal{Z}_k^y = \bar{\mathcal{Q}}_k^y$ and (ii) *Vertices of $\mathcal{Q}_k^y$ are binary*. Together, these imply $\mathcal{Q}_k^y = \text{conv}(\bar{\mathcal{Q}}_k^y) = \text{conv}(\mathcal{Z}_k^y)$.

(i) $\mathcal{Z}_k^y = \bar{\mathcal{Q}}_k^y$: The inclusion $\mathcal{Z}_k^y \subseteq \bar{\mathcal{Q}}_k^y$ follows because (8b) is valid for $\mathcal{Z}_k^y$, as shown in Proposition 4.

Conversely, $\bar{\mathcal{Q}}_k^y \subseteq \mathcal{Z}_k^y$ because every feasible solution in $\bar{\mathcal{Q}}_k^y$ satisfies the following inequalities derived from (8b):

$$\eta_{krp}^y \leq \sum_{r \in R_y: r < k} \eta_{rkp}^y + \sum_{r \in R_y: r > k} \eta_{krp}^y \leq x_{kp}^y + x_{kp+1}^y \quad \forall p \in P, \forall r \in R_y : r > k \quad (9a)$$

$$\eta_{rkp}^y \leq \sum_{r \in R_y: r < k} \eta_{rkp}^y + \sum_{r \in R_y: r > k} \eta_{krp}^y \leq x_{kp}^y + x_{kp+1}^y \quad \forall p \in P, \forall r \in R_y : k < r \quad (9b)$$

$$\sum_{r \in R_y: r < k} \eta_{rkp}^y + \sum_{r \in R_y: r > k} \eta_{krp}^y \leq x_{kp}^y + x_{kp+1}^y \leq 1 \quad \forall p \in P. \quad (9c)$$

The first inequalities in (9a) and (9b) follow from the non-negativity of η_k^y , whereas the second inequality in (9c) follows from (8a) and the non-negativity of $\mathbf{x}_k^y$.

(ii) *Vertices of $\mathcal{Q}_k^y$ are binary:* Consider the following linear program that maximizes an arbitrary linear objective over $\mathcal{Q}_k^y$:

$$(\mathcal{W}) \quad \max \sum_{p \in P} \tilde{f}_{kp}^y x_{kp}^y + \sum_{r \in R_y: r > k} \sum_{p \in P} \hat{f}_{krp}^y \eta_{krp}^y + \sum_{r \in R_y: r < k} \sum_{p \in P} \hat{f}_{rkp}^y \eta_{rkp}^y \quad (10a)$$

$$[\alpha] \quad \sum_{p \in P} x_{kp}^y = 1 \quad (10b)$$

$$[\beta_p] \quad \sum_{r \in R_y: r > k} \eta_{rkp}^y + \sum_{r \in R_y: r < k} \eta_{krp}^y \leq x_{kp}^y + x_{kp+1}^y \quad \forall p \in P \quad (10c)$$

$$\eta_{r_1 r_2 p}^y \geq 0 \quad \forall p \in P, \forall (r_1, r_2) \in I_y : k \in \{r_1, r_2\} \quad (10d)$$

$$x_{kp}^y \geq 0 \quad \forall p \in P. \quad (10e)$$

The corresponding dual variables are shown beside the constraints. The dual of $\mathcal{W}$ is given by:

$$(\mathcal{W}_D) \quad \min \quad \alpha \quad (11a)$$

$$\alpha - \beta_p - \beta_{p-1} \geq \tilde{f}_{kp}^y \quad \forall p \in P \quad (11b)$$

$$\beta_p \geq \hat{f}_{krp}^y \quad \forall r \in R_y : r > k, \forall p \in P \quad (11c)$$

$$\beta_p \geq \hat{f}_{rkp}^y \quad \forall r \in R_y : r < k, \forall p \in P \quad (11d)$$

$$\beta_p \geq 0 \quad \forall p \in P, \quad (11e)$$

where $\beta_0 = \beta_{|P|}$.

Setting $\bar{\beta}_p = \max \left\{ 0, \max_{r \in R_y: r > k} \hat{f}_{krp}^y, \max_{r \in R_y: r < k} \hat{f}_{rkp}^y \right\} \quad \forall p \in P$ and $\bar{\alpha} = \max_{p \in P} \left\{ \tilde{f}_{kp}^y + \bar{\beta}_p + \bar{\beta}_{p-1} \right\}$ yields a feasible solution for $\mathcal{W}_D$. By the weak duality theorem, the optimal value of $\mathcal{W}$ is at most $\bar{\alpha}$. Let $\bar{p} = \operatorname{argmax}_{p \in P} \left\{ \tilde{f}_{kp}^y + \bar{\beta}_p + \bar{\beta}_{p-1} \right\}$, and without the loss of generality, assume that $(k, \bar{r}_{\bar{p}}) \in \operatorname{argmax}_{(r_1, r_2) \in I_y: k \in \{r_1, r_2\}} \hat{f}_{r_1 r_2 \bar{p}}^y$, $(k, \bar{r}_{\bar{p}-1}) \in \operatorname{argmax}_{(r_1, r_2) \in I_y: k \in \{r_1, r_2\}} \hat{f}_{r_1 r_2 \bar{p}-1}^y$. Let $(\tilde{\mathbf{x}}_k^y, \tilde{\boldsymbol{\eta}}_k^y)$, be such that

- $\tilde{x}_{k\bar{p}}^y = 1, x_{kp}^y = 0 \quad \forall p \in P, p \neq \bar{p},$
- $\tilde{\eta}_{r_1 r_2 p}^y = 0 \quad \forall p \in P \setminus \{\bar{p}, \bar{p}-1\}, (r_1, r_2) \in I_y : k \in \{r_1, r_2\},$

- $\tilde{\eta}_{k\bar{r}\bar{p}}^y = 1$ if $\hat{f}_{k\bar{r}\bar{p}}^y > 0$ and 0, otherwise; $\tilde{\eta}_{r_1, r_2 \bar{p}}^y = 0 \ \forall (r_1, r_2) \in \{I_y | k \in \{r_1, r_2\}, \bar{r}_{\bar{p}} \notin \{r_1, r_2\}\}$,
- $\tilde{\eta}_{k\bar{r}\bar{p}-1\bar{p}-1}^y = 1$ if $\hat{f}_{k\bar{r}\bar{p}-1\bar{p}-1}^y > 0$ and 0, otherwise; $\tilde{\eta}_{r_1, r_2 \bar{p}-1}^y = 0 \ \forall (r_1, r_2) \in \{I_y | k \in \{r_1, r_2\}, \bar{r}_{\bar{p}-1} \notin \{r_1, r_2\}\}$.

The solution $(\tilde{\mathbf{x}}_k^y, \tilde{\boldsymbol{\eta}}_k^y)$ is feasible for $\mathcal{W}$ and attains the objective function value $\bar{\alpha}$. Hence, it is optimal for $\mathcal{W}$. Moreover, since $(\tilde{\mathbf{x}}_k^y, \tilde{\boldsymbol{\eta}}_k^y)$ is integral and the objective function coefficients of $\mathcal{W}$ are arbitrary, all vertices of $\mathcal{Q}_k^y$ are binary.

□

Based on Proposition 5, we strengthen $\mathcal{H}$ by replacing constraints (2d) and (2e), which linearize (3), with VI-b. The resulting formulation is stated below:

$$(\bar{\mathcal{H}}) \quad \max \sum_{y \in Y} \sum_{(r_1, r_2) \in I_y} \sum_{p \in P} \tilde{c}_{r_1 r_2 p}^y \eta_{r_1 r_2 p}^y \text{ s.t. } (2b), (2c), \text{VI-b}, (2f) - (2h). \quad (12)$$

6 Computational Experiments

The computational experiments are based on data from the UMN Family Medicine program. Each rotation spans multiple in a block, during which residents are assigned a fixed number of half-day family medicine clinic sessions per week based on the rotation's template. To construct benchmark instances, we estimate the templated weekly number of family medicine clinic sessions (see Table 7 in Appendix D) and parameterize a Poisson distribution to generate randomized clinic session counts across rotations. The Poisson distribution provides a practical choice for modeling the nonnegative integer number of clinic sessions, as it is characterized by a single parameter and avoids estimating additional distributional parameters that cannot be reliably inferred from the available rotation template data. The total number of clinic sessions assigned to a rotation is then obtained by multiplying the realized weekly count by the number of weeks in a block. We consider two discretizations of the academic year, consisting of $|M| \in \{12, 24\}$ blocks. Consequently, each PGY is partitioned into either 12 four-week rotations or 24 two-week rotations, assuming 48 working weeks in a year. We consider the same number of residents in each PGY, with $\gamma_y \in \{4, 6, 8\}$ for $|M|=12$ and $\gamma_y \in \{9, 12, 15\}$ for $|M|=24$, and therefore omit the y index from γ throughout this section. We set $\bar{L} = \lfloor \phi \tilde{L} \rfloor$, where $\phi \in \{0.90, 0.95, 0.98\}$ is a scaling factor and $\tilde{L}$ is an upper bound on LL . The computation of $\tilde{L}$ is described in Appendix C. We conduct a full-factorial experimental design over $|M|$, γ , and ϕ . For each

Case	Runtime (s)		Opt-Gap (%)		LB-I	UB-I
$\gamma \mid \phi$	$\mathcal{H}$	$\bar{\mathcal{H}}$	$\mathcal{H}$	$\bar{\mathcal{H}}$		
9 0.90	3600 (3600)	2 (5)	1313.14 (1344.32)	0 (0)	0.54 (0.15)	92.25 (91.89)
9 0.95	3600 (3600)	2 (4)	1258.43 (1354.30)	0 (0)	0.65 (0.31)	92.20 (91.86)
9 0.98	3600 (3600)	8 (10)	1028.58 (1393.45)	0 (0)	1.89 (0.76)	92.22 (91.92)
12 0.90	3600 (3600)	2 (5)	833.73 (1072.72)	0 (0)	0.28 (0.00)	92.41 (92.16)
12 0.95	3600 (3600)	2 (3)	1049.90 (1083.11)	0 (0)	0.65 (0.15)	92.32 (92.10)
12 0.98	3600 (3600)	3 (6)	1056.01 (1094.19)	0 (0)	0.80 (0.12)	92.37 (91.95)
15 0.90	3600 (3600)	2 (4)	1047.04 (1382.49)	0 (0)	0.41 (0.00)	92.35 (92.11)
15 0.95	3600 (3600)	3 (5)	1024.25 (1375.22)	0 (0)	0.41 (0.00)	92.35 (92.09)
15 0.98	3600 (3600)	4 (7)	1295.04 (1401.72)	0 (0)	0.68 (0.22)	92.40 (91.90)

Table 2: Comparison of $\mathcal{H}$ with $\bar{\mathcal{H}}$ for $|M|=24$.

parameter combination, we generate five random instances, resulting in a total of 90 test instances. All computational experiments are performed in Python using Gurobi 11.0 on a workstation equipped with an Intel Xeon Gold 6132 processor and 64 GB of RAM.

In Section 6.1, we evaluate the computational performance of $\bar{\mathcal{H}}$. In Section 6.2, we analyze the structural properties of the schedules produced by $\bar{\mathcal{H}}$. In Section 6.3, we compare the schedules produced by $\bar{\mathcal{H}}$ with those implemented in current practice. Finally, in Section 6.4, we compare the schedules produced by $\bar{\mathcal{H}}$ with those obtained from its analogous formulation using the objective function proposed by Bard et al. (2016b).

6.1 Computational Performance Analysis

In this section, we compare the computational performance of $\mathcal{H}$ and $\bar{\mathcal{H}}$. We also evaluate the computational benefits of warm-starting $\mathcal{P}$ with the optimal solution of $\bar{\mathcal{H}}$. Finally, we quantify the loss in IRB due to the *staggered scheduling framework*. All BIP formulations are solved using Gurobi’s branch-and-bound algorithm in default settings with a time limit of 3,600 seconds.

For each of the nine combinations of γ and ϕ , Tables 2 and 3 compare the computational performance of $\mathcal{H}$ and $\bar{\mathcal{H}}$ for $|M|=24$ and $|M|=12$, respectively. Columns 2 and 3 report the average (maximum) solution times, in seconds, whereas Columns 4 and 5 report the average (maximum) optimality gaps (%) for $\mathcal{H}$ and $\bar{\mathcal{H}}$, respectively at the termination of the branch-and-bound algorithm. Columns 6 and 7 report the average (minimum) percentage improvements in the best lower bound (LB-I) and best upper bound (UB-I) obtained by $\bar{\mathcal{H}}$ relative to $\mathcal{H}$, computed as $\text{LB-I} = \frac{LB^{\bar{\mathcal{H}}} - LB^{\mathcal{H}}}{LB^{\mathcal{H}}} \times 100$, $\text{UB-I} = \frac{UB^{\mathcal{H}} - UB^{\bar{\mathcal{H}}}}{UB^{\mathcal{H}}} \times 100$. Here, $LB^{\mathcal{H}}$ and $LB^{\bar{\mathcal{H}}}$ denote the best lower bounds obtained by $\mathcal{H}$ and $\bar{\mathcal{H}}$,

Case	Runtime (s)		Opt-Gap (%)		LB-I	UB-I
$\gamma \mid \phi$	$\mathcal{H}$	$\bar{\mathcal{H}}$	$\mathcal{H}$	$\bar{\mathcal{H}}$		
4 0.90	3600 (3600)	0.11 (0.17)	234.38 (276.47)	0 (0)	0.16 (0.00)	69.60 (63.52)
4 0.95	3600 (3600)	0.34 (0.46)	260.53 (278.46)	0 (0)	0.34 (0.00)	72.13 (70.88)
4 0.98	3600 (3600)	0.69 (1.42)	243.76 (277.25)	0 (0)	1.29 (0.20)	70.40 (67.05)
6 0.90	3600 (3600)	0.12 (0.27)	233.17 (281.00)	0 (0)	0.03 (0.00)	69.22 (61.48)
6 0.95	3600 (3600)	0.33 (0.49)	253.43 (297.01)	0 (0)	0.23 (0.00)	71.41 (66.89)
6 0.98	3600 (3600)	0.51 (0.70)	252.90 (294.13)	0 (0)	0.20 (0.00)	71.51 (70.06)
8 0.90	3600 (3600)	0.30 (1.23)	205.73 (302.68)	0 (0)	0.06 (0.00)	66.56 (62.96)
8 0.95	3600 (3600)	0.29 (0.68)	250.50 (336.91)	0 (0)	0.34 (0.10)	70.85 (67.45)
8 0.98	3600 (3600)	0.40 (0.53)	260.16 (332.47)	0 (0)	0.33 (0.20)	71.88 (69.89)

Table 3: Comparison of $\mathcal{H}$ and $\bar{\mathcal{H}}$ for $|M|=12$.

respectively, whereas $UB^{\mathcal{H}}$ and $UB^{\bar{\mathcal{H}}}$ denote the corresponding best upper bounds, all evaluated at the termination of the branch-and-bound algorithm.

As shown in Tables 2 and 3, replacing constraints (2d) and (2e) in $\mathcal{H}$ with VI-b yields a significant improvement in computational performance. For every test instance and for both values of $|M|$, $\mathcal{H}$ reaches the 3600-second time limit with large optimality gaps, whereas $\bar{\mathcal{H}}$ solves every instance to optimality within 10 seconds. The poor performance of $\mathcal{H}$ is primarily due to the weakness of its LP relaxation. Although the solver is able to identify high-quality feasible solutions, it is unable to sufficiently tighten the upper bound within the allotted time, resulting in large optimality gaps. By contrast, incorporating VI-b significantly strengthens the formulation. This observation is supported by Columns 6 and 7 of Tables 2 and 3. Specifically, the improvement in the best lower bound is consistently small, indicating that $\mathcal{H}$ already produces near-optimal feasible solutions. In contrast, the improvement in the best upper bound exceeds 66% on average for all test instances, demonstrating that the primary benefit of VI-b is a substantially tighter relaxation, which enables the solver to close the optimality gap rapidly.

For each of the nine combinations of γ and ϕ , Tables 4 and 5 compare the computational performance of the *unrestricted formulation*, $\mathcal{P}$, with its warm-started counterpart, denoted by $\mathcal{P}^{\bar{\mathcal{H}}}$, in which $\mathcal{P}$ is initialized with the optimal solution of $\bar{\mathcal{H}}$. The two tables report results for $|M|=24$ and $|M|=12$, respectively. Columns 2 and 3 report the percentage of instances for which the commercial solver finds a feasible solution within the 3600-second time limit (%-Inst-F) for $\mathcal{P}$ and $\mathcal{P}^{\bar{\mathcal{H}}}$, respectively. Columns 4 and 5 similarly report the percentage of instances solved to 1% optimality. Columns 6 and 7 report the average (maximum) solution time for $\mathcal{P}$ and $\mathcal{P}^{\bar{\mathcal{H}}}$, respectively. Finally,

Case $\gamma \mid \phi$	%Inst-F		%Inst-S		Runtime (seconds)		δ (%)
	$\mathcal{P}$	$\mathcal{P}^{\bar{\mathcal{H}}}$	$\mathcal{P}$	$\mathcal{P}^{\bar{\mathcal{H}}}$	$\mathcal{P}$	$\mathcal{P}^{\bar{\mathcal{H}}}$	
9 0.90	100	100	100	100	2659 (3232)	17 (17)	0.69 (0.99)
9 0.95	100	100	40	100	3516 (3602)	17 (17)	0.85 (0.99)
9 0.98	0	100	0	100	3601 (3601)	78 (320)	0.89 (1.17)
12 0.90	20	100	0	100	3603 (3610)	24 (25)	0.81 (0.99)
12 0.95	0	100	0	100	3600 (3601)	24 (25)	0.91 (1.00)
12 0.98	0	100	0	100	3600 (3601)	24 (24)	0.92 (0.99)
15 0.90	0	100	0	100	3600 (3600)	28 (29)	0.99 (1.00)
15 0.95	0	100	0	100	3600 (3600)	29 (29)	0.99 (1.00)
15 0.98	0	100	0	100	3600 (3600)	29 (29)	0.99 (1.00)

Table 4: Comparison of $\mathcal{P}$ with $\mathcal{P}^{\bar{\mathcal{H}}}$ for $|M|=24$.

Case $\gamma \mid \phi$	%Inst-F		%Inst-S		Runtime (seconds)		δ (%)
	$\mathcal{P}$	$\mathcal{P}^{\bar{\mathcal{H}}}$	$\mathcal{P}$	$\mathcal{P}^{\bar{\mathcal{H}}}$	$\mathcal{P}$	$\mathcal{P}^{\bar{\mathcal{H}}}$	
4 0.90	100	100	100	100	4 (7)	1 (2)	0.36 (1.17)
4 0.95	100	100	100	100	6 (11)	1 (2)	0.63 (1.57)
4 0.98	100	100	100	100	22 (53)	4 (15)	0.69 (1.59)
6 0.90	100	100	100	100	7 (8)	3 (6)	0.60 (1.58)
6 0.95	100	100	100	100	19 (28)	5 (13)	1.03 (2.39)
6 0.98	100	100	100	100	25 (45)	14 (37)	1.35 (2.67)
8 0.90	100	100	100	100	31 (46)	6 (8)	0.61 (0.90)
8 0.95	100	100	100	100	47 (83)	14 (40)	0.68 (1.10)
8 0.98	100	100	100	100	83 (115)	20 (48)	0.96 (1.61)

Table 5: Comparison of $\mathcal{P}$ with $\mathcal{P}^{\bar{\mathcal{H}}}$ for $|M|=12$.

Column 8 reports the average (maximum) value of $\delta = \frac{UB^{\mathcal{P}^{\bar{\mathcal{H}}}} - LB^{\bar{\mathcal{H}}}}{LB^{\bar{\mathcal{H}}}} \times 100$. Here, $UB^{\mathcal{P}^{\bar{\mathcal{H}}}}$ denotes the best upper bound obtained by $\mathcal{P}^{\bar{\mathcal{H}}}$, whereas $LB^{\bar{\mathcal{H}}}$ denotes the optimal objective value of $\bar{\mathcal{H}}$. Thus, δ measures the percentage gap between the optimal objective value under the *staggered scheduling framework* and the best upper bound obtained for $\mathcal{P}^{\bar{\mathcal{H}}}$. A small value of δ therefore indicates that the *staggered scheduling framework* achieves a **IRB** value close to the best achievable value under the *unrestricted formulation*, $\mathcal{P}$.

As shown in Tables 4 and 5, the computational difficulty of $\mathcal{P}$ increases substantially with instance size. For $|M|=12$, the commercial solver solves all instances to within 1% optimality in less than 115 seconds. In contrast, for $|M|=24$, even instances with smaller values of γ require more than 2,600 seconds to reach the same optimality tolerance. For larger values of γ , the solver frequently

ϕ	IRB retained (%)
0.90	100.0
0.95	99.3
0.98	98.7

Table 6: Average percentage of IRB retained relative to $\phi = 0.90$ for $\gamma = 6$.

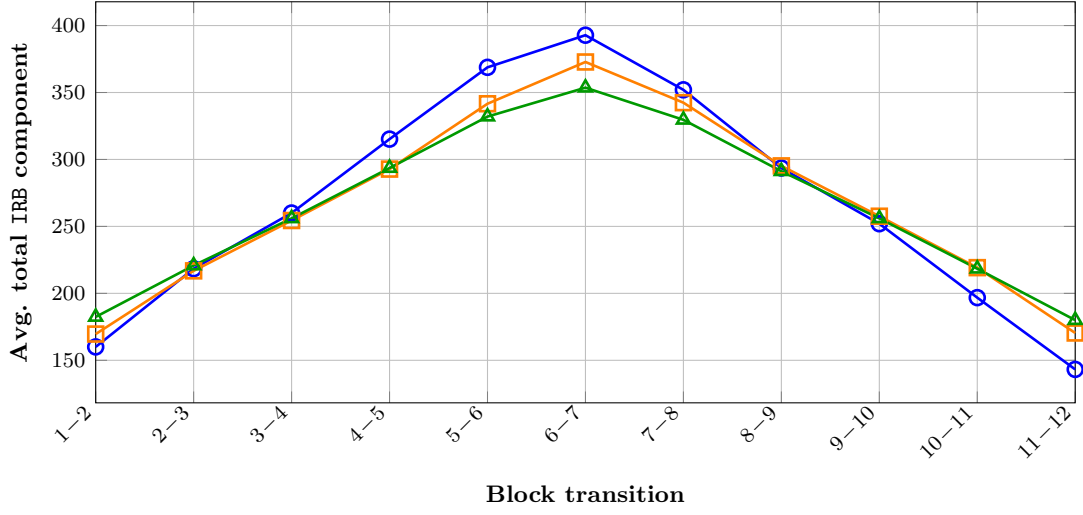
fails to even identify a feasible solution within the 3,600-second time limit. Specifically, among the 45 instances with $|M| = 24$, the commercial solver finds feasible solutions for only 11 instances using $\mathcal{P}$, all but one of which have $\gamma = 9$. Furthermore, for fixed instance size, the computational difficulty of $\mathcal{P}$ increases with ϕ . For example, when $|M| = 24$ and $\gamma = 9$, the percentage of instances solved to within 1% optimality within 3,600 seconds decreases from 100% to 0% as ϕ increases from 0.90 to 0.98.

Columns 5 and 7 of Tables 4 and 5 show that warm-starting $\mathcal{P}$ with an optimal solution of $\bar{\mathcal{H}}$ substantially improves its computational performance. With the warm start, the solver reaches the 1% optimality tolerance in less than 80 seconds on average, demonstrating the computational benefit of using $\bar{\mathcal{H}}$ to provide a high-quality solution. Moreover, the average values of δ reported in the last column of these tables are approximately 1% or lower, indicating that the solutions obtained under the *staggered scheduling framework* achieve IRB values close to the best upper bounds obtained for the *unrestricted formulation*. Thus, in addition to accelerating the solution of $\mathcal{P}$, $\bar{\mathcal{H}}$ provides high-quality feasible solutions for FMRS.

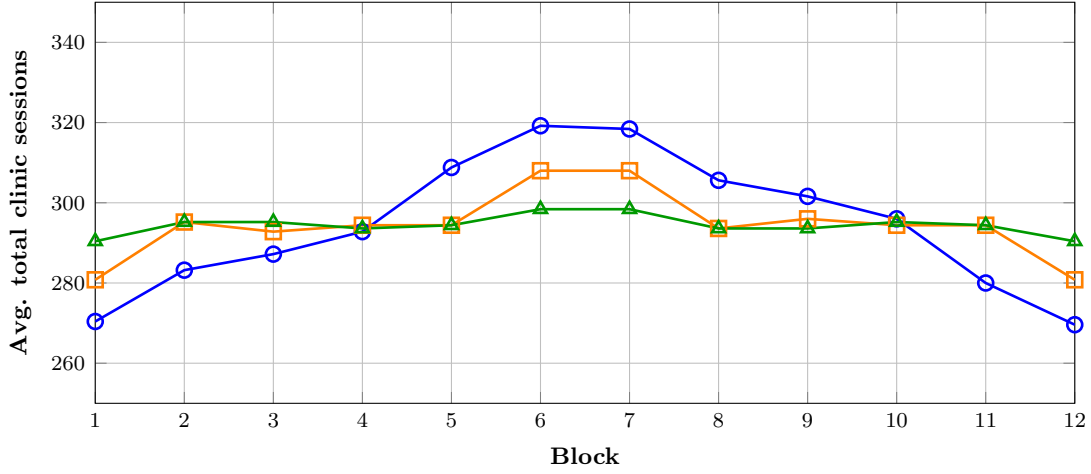
6.2 Solution Analysis

In this section, we examine the structural properties of the schedules produced by $\bar{\mathcal{H}}$. We conduct this analysis using the same five random instances considered in Section 6.1 for each value of ϕ , with $|M| = 12$ and $\gamma = 6$. Figure 2a gives the average total IRB component over five random instances for each pair of consecutive blocks in M while Figure 2b reports the average total number of clinic sessions in each block for the schedules produced by $\bar{\mathcal{H}}$. Here, the IRB component for two consecutive blocks in a rotation schedule is defined as the absolute difference in the number of clinic sessions associated with the rotations assigned to those blocks. The total IRB component for two consecutive blocks is obtained by summing these individual components across all residents. Figure 3 reports the average IRB, computed over five random instances, for the first six and the last six positions of the optimal **base rotation schedule** under each value of ϕ while Table 6 reports the corresponding average percentage of IRB retained relative to the solution obtained with $\phi = 0.90$ as ϕ increases.

Figure 3 shows that the optimal **base rotation schedule** consistently assigns rotations with



(a) Average total IRB components across block transitions.



(b) Average number of clinic sessions across blocks.

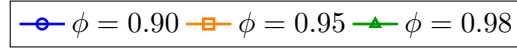


Figure 2: Properties schedules obtained from $\bar{\mathcal{H}}$ under different values of ϕ

larger IRB components to earlier positions. This behavior is a direct consequence of the objective coefficients of $\bar{\mathcal{H}}$ in (1). Under the *staggered scheduling framework*, IRB components associated with earlier positions contribute to a larger number of resident schedules than those associated with later positions and therefore receive greater weight in the objective function. Furthermore, the same figure shows that increasing ϕ decreases the average IRB associated with the first six positions of the **base rotation schedule** while increasing that of the last six positions. This redistribution explains the modest reduction in IRB reported in Table 6. Under the *staggered scheduling framework*, IRB components associated with earlier positions of the **base rotation schedule** primarily affect the middle portions of the schedules of the remaining residents, whereas those associated with later positions primarily affect the beginning and end of their schedules. Consequently, Figure 2a shows that the

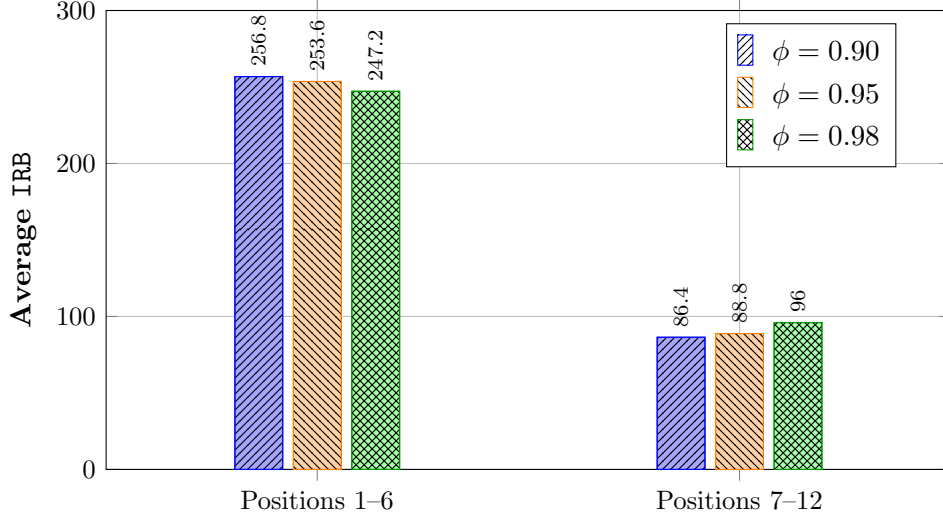


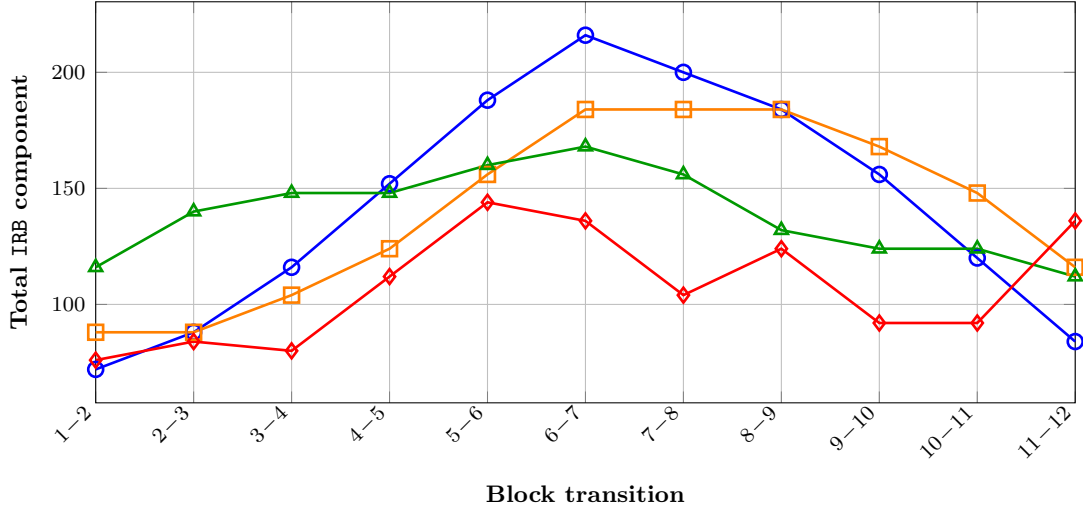
Figure 3: Average IRB over the first six and last six positions in the **base rotation schedule** for $|M|=12$, $\gamma=6$, and $\phi \in \{0.90, 0.95, 0.98\}$.

total IRB components initially increase across blocks before subsequently decreasing. As ϕ increases, however, the total IRB components decrease in the middle blocks while increasing near the beginning and end of the planning horizon. Thus, the reduction in IRB in one portion of the planning horizon is largely offset by an increase elsewhere, resulting in only a modest reduction in the overall IRB. This balance is achieved by assigning rotations with larger number of clinic sessions to later positions in the **base rotation schedule** to satisfy the increasingly restrictive lower bound on LL, resulting in a more uniform distribution of clinic sessions across blocks, as shown in Figure 2b.

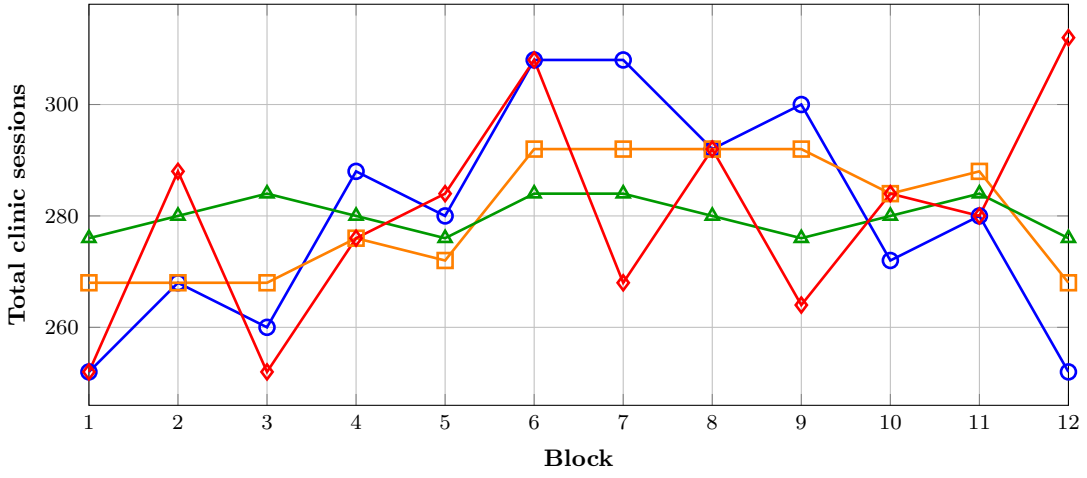
6.3 Comparison with Current Practice

In current practice, residency rotation scheduling remains a highly manual and hierarchical process that prioritizes basic operational feasibility ensuring compliance with educational requirements and clinic capacity constraints. While IRB and LL are widely recognized as desirable, current manual workflows lack the capability to directly optimize scheduling over these criteria. Figure 4 compares the schedules produced by $\bar{\mathcal{H}}$ for different values of ϕ with a schedule implemented in current practice. As in the previous section, we set $|M|=12$ and $\gamma=6$ for each value of ϕ . The weekly clinic-session count for each rotation is set equal to the value specified in the rotation template used to construct the current practice schedule. Figure 4a reports the total IRB component for each pair of consecutive blocks, while Figure 4b reports the total number of clinic sessions in each block for the schedules produced by $\bar{\mathcal{H}}$ and the current practice schedule.

As shown in Figure 4a, for each value of ϕ , the schedules produced by $\bar{\mathcal{H}}$ achieve substantially higher



(a) Total IRB components across block transitions.



(b) Total clinic sessions across blocks.

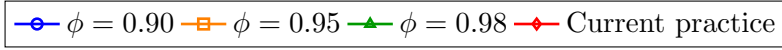


Figure 4: Comparison of schedules obtained from $\bar{\mathcal{H}}$ under different values of ϕ with the schedule implemented in current practice.

total IRB components than the current practice schedule for most consecutive time-block transitions, resulting in a significant overall improvement in IRB. Figure 4b shows that, as expected, increasing ϕ results in a more uniform distribution of clinic sessions across blocks. Compared with the current practice schedule, the schedules produced by $\bar{\mathcal{H}}$ exhibit substantially less temporal variability in clinic capacity, thereby providing a more balanced distribution of clinic sessions over time. Furthermore, as demonstrated in the previous section, the improved balance in clinic capacity achieved at higher values of ϕ comes with little loss in IRB.

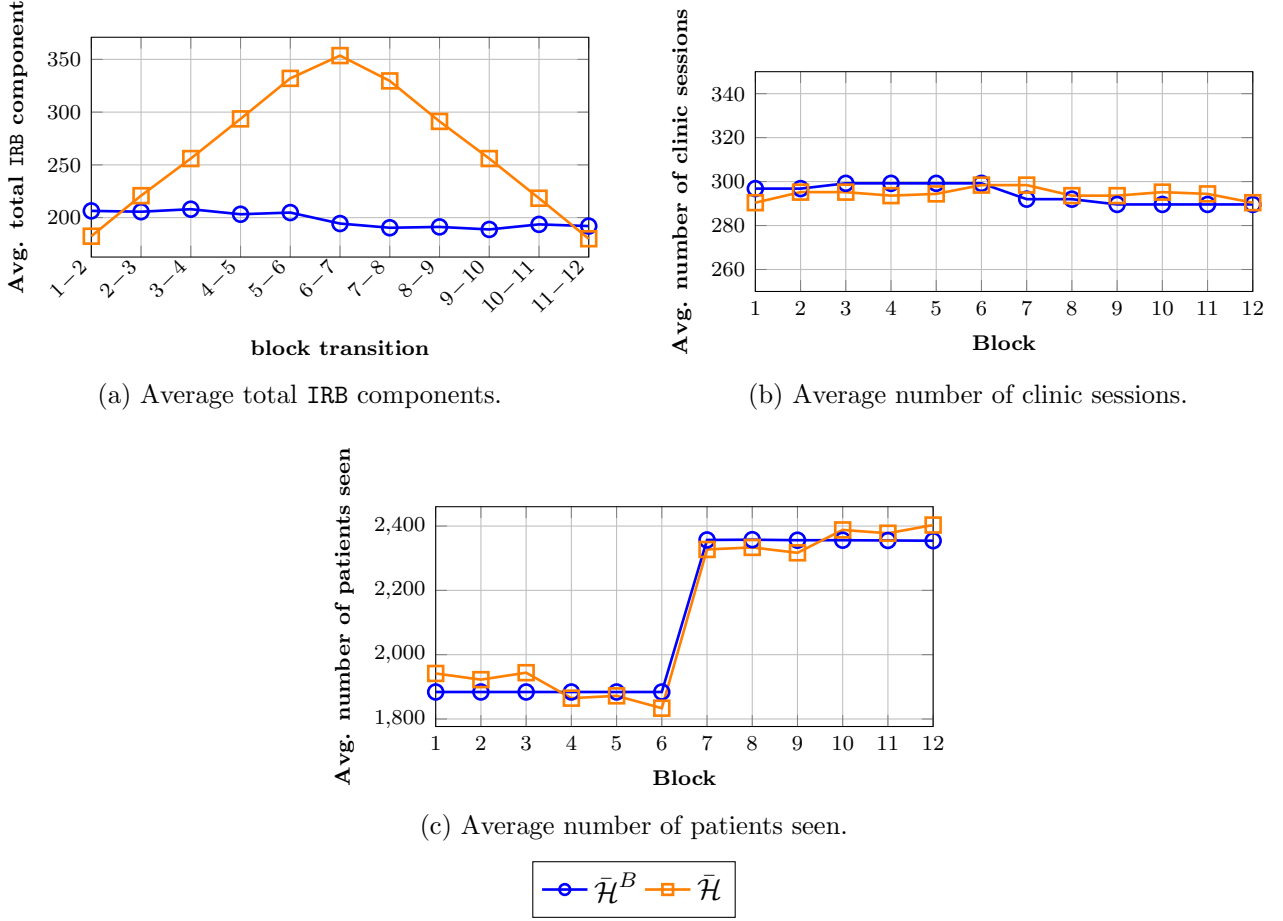


Figure 5: Comparison of schedules obtained from $\bar{\mathcal{H}}^B$ and $\bar{\mathcal{H}}$ for $\phi = 0.98$, $|M| = 12$, and $\gamma = 6$.

6.4 Comparison with the objective of Bard et al. (2016b)

In this section, we compare schedules generated under the *staggered scheduling framework* using two different objectives. The first maximizes IRB subject to a lower bound on LL, while the second uses the objective considered in Bard et al. (2016b), which minimizes the maximum deviation in the number of patients seen across half-years. Although Bard et al. (2016b) do not consider staggered scheduling, we incorporate their objective into our *staggered scheduling framework* to facilitate a direct comparison. Specifically, Appendix C presents a variant of $\bar{\mathcal{H}}$, denoted by $\bar{\mathcal{H}}^B$, in which the IRB objective is replaced by the objective of Bard et al. (2016b) and the lower-bound constraint on LL is removed. The objective function in Bard et al. (2016b) also includes a term that penalizes modifications to predefined rotation templates. Because our setting does not permit such modifications, this term is omitted from $\bar{\mathcal{H}}^B$. For the comparison, we use the data from Bard et al. (2016b) on the number of patients that can be seen by residents in each PGY, reported in Table 8, Appendix D.

Figure 5 compares the schedules obtained from $\bar{\mathcal{H}}$ and $\bar{\mathcal{H}}^B$ for $\phi = 0.98$, $|M| = 12$, and $\gamma = 6$.

Specifically, Figure 5a reports the average total **IRB** component, computed across five random instances, for each pair of consecutive blocks. Figures 5b and 5c report the corresponding average numbers of clinic sessions and patients seen, respectively, in each block.

As shown in Figures 5b and 5c, the schedules produced by $\bar{\mathcal{H}}$ and $\bar{\mathcal{H}}^B$ result in similar distributions of clinic sessions and patients seen across the planning horizon. This finding suggests that, at the relatively high value of $\phi = 0.98$, the LL constraint in $\bar{\mathcal{H}}$ is sufficient to produce a distribution of patients seen comparable to that obtained by directly minimizing deviations in the number of patients seen. This occurs because balancing clinic sessions across blocks also promotes a balanced distribution of patients seen. Importantly, these results suggest that there is considerable room to improve **IRB** beyond that achieved by the objective of Bard et al. (2016b) without substantially affecting the resulting distribution of patients seen. Indeed, Figure 5a shows that $\bar{\mathcal{H}}$ yields substantially higher **IRB** components for most consecutive time-block transitions while producing a distribution of patients seen similar to that obtained by $\bar{\mathcal{H}}^B$.

7 Conclusion

In this paper, we developed a BIP framework for **FMRS** that explicitly incorporates the guiding principles of the **Clinic First Model** through its objective function and constraints. Specifically, the proposed formulation maximizes **IRB** to ensure a consistent interleaving of family medicine clinic sessions throughout each resident’s schedule, thereby mitigating both prolonged clinical gaps and heavy session clustering. Additionally, enforcing LL stabilizes year-long operational capacity by spreading total resident clinic sessions across blocks. Integrating these criteria within **FMRS** provides benefits for both resident education experience and outpatient clinic operations, in line with **Clinic First** goals.

We integrated the proposed BIP framework with the *staggered scheduling framework*, under which resident schedules are generated by successively right-shifting a single **base rotation schedule** across the planning horizon. This structure ensures identical curricular exposure across each resident program year cohort while producing transparent, simple to manage rotation schedules. We also established a theoretical bound on the loss in **IRB** resulting from the *staggered scheduling framework*. To improve the computational performance of the resulting BIP formulation, we derived valid inequalities and showed that they describe the convex hull of a set defined by a subset of the formulation’s constraints.

Using computational experiments based on data from the UMN Family Medicine Residency pro-

gram, we demonstrated that the proposed valid inequalities substantially improve the computational performance of the BIP formulation. Our results also showed that the *staggered scheduling framework* results in only a small loss in IRB. Moreover, using its optimal solution to warm-start the *unrestricted formulation* substantially improves computational performance. Compared with a schedule implemented in current practice, the schedules generated by our approach achieve substantially higher IRB while providing a more balanced distribution of clinic sessions across blocks. Finally, our analysis of the structural properties of the resulting schedules provides further insight into how the proposed approach improves IRB while balancing the number of clinic sessions across blocks. A comparison with the objective function considered in Bard et al. (2016b) further demonstrates the benefits of explicitly incorporating IRB and LL into FMRS.

Several promising directions for future research emerge from this work. First, the proposed modeling framework could be extended by treating the number of blocks by which each resident’s schedule is shifted relative to the **base rotation schedule** as a decision variable, rather than imposing a fixed one-block shift between consecutive residents. This extension would increase the flexibility of the *staggered scheduling framework* by expanding the set of schedules it can generate. Second, the *staggered scheduling framework* could be generalized to allow independent staggered scheduling over multiple consecutive time intervals. For example, the planning horizon M could be partitioned into two intervals, $\{1, 2, \dots, \lfloor M/2 \rfloor\}$ and $\{\lfloor M/2 \rfloor + 1, \dots, M\}$, with the *staggered scheduling framework* applied independently within each interval. This would provide additional scheduling flexibility by allowing different staggered patterns across the planning horizon. Another direction is to incorporate uncertainty in the number of clinic sessions associated with each rotation, enabling the development of schedules that remain effective under variations in clinic availability. Finally, the proposed framework could be extended to jointly optimize resident team formation and annual rotation schedules within FMRS. Such an integrated approach would explicitly account for interactions among residents’ clinic schedules when determining team composition, potentially improving patient continuity of care beyond what can be achieved by optimizing rotation schedules alone.

References

- Adam P, Hersch D, Peek CJ (2022) Implementing Clinic First guiding actions across 4 family medicine residency clinics. *Academic Medicine* 97(2):233–238.
- Akbarzadeh B, Maenhout B (2021a) A decomposition-based heuristic procedure for the medical student scheduling problem. *European Journal of Operational Research* 288(1):63–79.

- Akbarzadeh B, Maenhout B (2021b) An exact branch-and-price approach for the medical student scheduling problem. *Computers & Operations Research* 129:105209.
- Bard JF, Shu Z, Morrice DJ, Leykum LK (2016a) Annual block scheduling for internal medicine residents with 4+ 1 templates. *Journal of the Operational Research Society* 67(7):911–927.
- Bard JF, Shu Z, Morrice DJ, Leykum LK (2017) Constructing block schedules for internal medicine residents. *IIE Transactions on Healthcare Systems Engineering* 7(1):1–14.
- Bard JF, Shu Z, Morrice DJ, Leykum LK, Poursani R (2016b) Annual block scheduling for family medicine residency programs with continuity clinic considerations. *IIE Transactions* 48(9):797–811.
- Bodenheimer T, Sinsky C (2014) From triple to quadruple aim: care of the patient requires care of the provider. *The Annals of Family Medicine* 12(6):573–576.
- Brown SR, Bodenheimer T, Kong M (2019) High-performing primary care residency clinics: a collaboration. *Annals of Family Medicine* 17(5):470.
- Dall T, Chakrabarti R, Iacobucci W, Hansari A, West T (2017) The complexities of physician supply and demand: projections from 2015 to 2030. *Association of American Medical Colleges* .
- Day TE, Napoli JT, Kuo PC (2006) Scheduling the resident 80-hour work week: an operations research algorithm. *Current Surgery* 63(2):136–141.
- Erhard M, Schoenfelder J, Fügner A, Brunner JO (2018) State of the art in physician scheduling. *European Journal of Operational Research* 265(1):1–18.
- Franz LS, Miller JL (1993) Scheduling medical residents to rotations: solving the large-scale multiperiod staff assignment problem. *Operations Research* 41(2):269–279.
- Guo J, Morrison DR, Jacobson SH, Jokela JA (2014) Complexity results for the basic residency scheduling problem. *Journal of Scheduling* 17:211–223.
- Guo J, Pozehl W, Cohn A (2023) A two-stage partial fixing approach for solving the residency block scheduling problem. *Health Care Management Science* 26(2):363–393.
- Gupta R, Barnes K, Bodenheimer T (2016) Clinic first: 6 actions to transform ambulatory residency training. *Journal of Graduate Medical Education* 8(4):500–503.
- Hersch D, Klemenhausen K, Martin C, Berg B, Adam P (2023) Impact of set-day clinic on physician continuity in a family medicine residency clinic. *Family Medicine* 55(9):612–615.
- Hofkamp H, Verdick A, Cheng A, Ramalingam N, Rdesinski R, Garvin R (2020) Effects of clinic first educational interventions on resident wellness and engagement. *Family Medicine* 52(6):422–426.
- Hong YC, Cohn A, Gorga S, O’Brien E, Pozehl W, Zank J (2019) Using optimization techniques and multidisciplinary collaboration to solve a challenging real-world residency scheduling problem. *INFORMS Journal on Applied Analytics* 49(3):201–212.

- Kraul S (2020) Annual scheduling for anesthesiology medicine residents in task-related programs with a focus on continuity of care. *Flexible Services and Manufacturing Journal* 32(1):181–212.
- Lemay B, Cohn A, Epelman M, Gorga S (2017) New methods for resolving conflicting requests with examples from medical residency scheduling. *Production and Operations Management* 26(9):1778–1793.
- Li S, Shehadeh KS, Curtis FE, Hochman BR (2025) Equity-promoting integer programming approaches for medical resident rotation scheduling. *Health Care Management Science* 28(4):890–929.
- Manolakis PG, Skelton JB (2010) Pharmacists’ contributions to primary care in the United States collaborating to address unmet patient care needs: the emerging role for pharmacists to address the shortage of primary care providers. *American Journal of Pharmaceutical Education* 74(10).
- Paul KJ, Hidaka BH, Ford P, Morris C (2021) Measuring greater patient-provider continuity in a clinic-first family medicine residency curriculum. *The Permanente Journal* 25.
- Peccoraro LA, Tackett S, Ward L, Federman A, Helenius I, Christmas C, Thomas DC (2013) Resident satisfaction with continuity clinic and career choice in general internal medicine. *Journal of General Internal Medicine* 28(8):1020–1027, URL <http://dx.doi.org/10.1007/s11606-012-2280-5>.
- Proano RA, Agarwal A (2018) Scheduling internal medicine resident rotations to ensure fairness and facilitate continuity of care. *Health Care Management Science* 21:461–474.
- Smalley HK, Keskinocak P (2016) Automated medical resident rotation and shift scheduling to ensure quality resident education and patient care. *Health Care Management Science* 19(1):66–88.
- Stange KC, Miller WL, Etz RS (2023) The role of primary care in improving population health. *The Milbank Quarterly* 101(S1):795–840.
- Starfield B (2012) Primary care: an increasingly important contributor to effectiveness, equity, and efficiency of health services. SESPAS report 2012. *Gaceta Sanitaria* 26:20–26.
- Starfield B, Shi L, Macinko J (2005) Contribution of primary care to health systems and health. *The Milbank Quarterly* 83(3):457–502.
- Wang F, Huang YL, Ju F, Grega JC (2022) Resident rotation scheduling for categorical internal medicine residency program. *IEEE Robotics and Automation Letters* 7(2):4142–4148.
- Zeller T, Ewing J, Asif I (2019) Prevalence of clinic first curricula: a survey of AFMRD members. *Family Medicine* 51(4):338–343.

Appendices

Appendix A: *Unrestricted Formulation* for FMRS

Let $\kappa_{ir_1r_2m}^y$ be a binary variable that equals 1 if resident $i \in S_y$ in PGY $y \in Y$ is assigned to rotation $r_1 \in R_y$ in block $m - 1$ and to rotation $r_2 \in R_y$ in block m , and 0 otherwise. Define $c_{r_1r_2}^y = |\bar{c}_{r_1}^y - \bar{c}_{r_2}^y|$ as the contribution to IRB obtained by scheduling rotations $r_1, r_2 \in R_y$ consecutively. Also define $[a] = \{1, 2, 3, \dots, a\}$, $\bar{M} = M \setminus \{1\}$, and $\vec{M} = \bar{M} \setminus \{|M|\}$. With this notation, the *unrestricted formulation* for FMRS is:

$$(\mathcal{P}) \quad \max \sum_{y \in Y} \sum_{i \in S_y} \sum_{r_1 \in R_y} \sum_{r_2 \in R_y \setminus \{r_1\}} \sum_{m \in \bar{M}} c_{r_1r_2}^y \kappa_{ir_1r_2m}^y \quad (13a)$$

$$\sum_{i \in S_y} \sum_{r_2 \in R_y \setminus \{r_1\}} \kappa_{ir_1r_22}^y \leq 1, \quad \forall r_1 \in R_y, \forall y \in Y \quad (13b)$$

$$\sum_{i \in S_y} \sum_{r_1 \in R_y \setminus \{r_2\}} \kappa_{ir_1r_2m}^y \leq 1, \quad \forall r_2 \in R_y, \forall m \in \bar{M}, \forall y \in Y \quad (13c)$$

$$\sum_{r_1 \in R_y \setminus \{r_2\}} \kappa_{ir_1r_22}^y + \sum_{r_1 \in R_y \setminus \{r_2\}} \sum_{m \in \bar{M}} \kappa_{ir_1r_2m}^y = 1, \quad \forall r_2 \in R_y, \forall i \in S_y, \forall y \in Y \quad (13d)$$

$$\sum_{r_1 \in R_y \setminus \{r_2\}} \kappa_{ir_1r_2m}^y = \sum_{r_3 \in R_y \setminus \{r_2\}} \kappa_{ir_3r_3(m+1)}^y, \quad \forall m \in \vec{M}, \forall r_2 \in R_y, \forall i \in S_y, \forall y \in Y \quad (13e)$$

$$\sum_{y \in Y} \sum_{r_2 \in R_y} \sum_{r_1 \in R_y \setminus \{r_2\}} \sum_{i \in S_y} \bar{c}_{r_1}^y \kappa_{ir_1r_22}^y \geq \bar{L}, \quad (13f)$$

$$\sum_{y \in Y} \sum_{r_2 \in R_y} \sum_{r_1 \in R_y \setminus \{r_2\}} \sum_{i \in S_y} \bar{c}_{r_2}^y \kappa_{ir_1r_2m}^y \geq \bar{L}, \quad \forall m \in \bar{M} \quad (13g)$$

$$\sum_{r_2 \in R_y \setminus \{r_1\}} \kappa_{kr_1r_22}^y = 0, \quad \forall r_1 \in [k-1], \forall k \in S_y, \forall y \in Y \quad (13h)$$

$$\sum_{r_2 \in R_y \setminus \{r_1\}} \kappa_{kr_1r_22}^y \leq \sum_{r_3=1}^{r_1-1} \sum_{r_2 \in R_y \setminus \{r_3\}} \kappa_{lr_3r_22}^y, \quad \forall l \in [k-1], \forall r_1 \in R_y \setminus [k-1], \quad (13i)$$

$$\kappa_{ir_1r_2m}^y \in \{0, 1\}, \quad \forall i \in S_y, \forall r_1 \neq r_2 \in R_y, \quad (13j)$$

$$\forall m \in M, \forall y \in Y$$

Constraints (13b) and (13c) ensure that each rotation in R_y is assigned to at most one resident in each block, whereas Constraints (13d) ensure that each resident is assigned to every rotation in R_y exactly once over the planning horizon. Constraints (13e) enforce consistency in rotation assignments across consecutive blocks, while Constraints (13f) and (13g) impose the lower bound on LL. Constraints (13h) and (13i) are symmetry-breaking constraints. Specifically, Constraints (13h) prevent resident k from being assigned to any rotation $r_1 < k$ in the first block. Constraints (13i) ensure that if resident k is assigned to rotation r_1 in the first block, then every resident $l < k$ is assigned to a rotation $r_3 < r_1$ in that block. The objective function (13a) maximizes IRB.

Appendix B: Upper Bound on the Maximum Level Loading

The following proposition gives an upper bound on LL.

Proposition 6. $LL \leq \tilde{LL} = \left\lfloor \frac{\sum_{y \in Y} \gamma_y \left(\sum_{r \in R_y} \bar{c}_r^y \right)}{|M|} \right\rfloor$

Proof. Let $TC = \sum_{y \in Y} \gamma_y \left(\sum_{r \in R_y} \bar{c}_r^y \right)$ denote the total number of clinic sessions across the planning horizon. Indeed, for each PGY $y \in Y$, $\gamma_y \left(\sum_{r \in R_y} \bar{c}_r^y \right)$ is the total number of clinic sessions assigned to all residents in that PGY. Suppose, by contradiction, that $LL > \left\lfloor \frac{TC}{|M|} \right\rfloor$. Since LL is integer-valued, it follows that there exists some $\zeta > 0$ such that $LL = \frac{TC}{|M|} + \zeta$. By definition, LL is the minimum total number of clinic sessions assigned in any block. Therefore, every block must contain at least LL clinic sessions, implying $TC \geq |M|LL = |M| \left(\frac{TC}{|M|} + \zeta \right) = TC + |M|\zeta$. This implies $|M|\zeta \leq 0$, which is a contradiction. $\square$

Appendix C: *Staggered Scheduling* formulation with the objective from Bard et al. (2016b)

Let $L = \{1, 2\}$ denote the set of half-years, and let $\mathcal{M}_l$ denote the set of blocks in half-year $l \in L$. For each PGY y and half-year l , let ρ_l^y denote the number of patients seen per clinic session, and let μ_l denote the average number of patients seen per block in half-year l . Finally, let Δ denote the maximum absolute deviation in the number of patients seen in any block from the corresponding half-year average μ_l . Using these notations, the following formulation determines a **base rotation schedule** under the *staggered scheduling framework* that minimizes Δ .

$$(\bar{\mathcal{H}}^B) \quad \min \quad \Delta \tag{14a}$$

$$(2b), (2c), (2h) \tag{14b}$$

$$\mu_l = \frac{1}{|M|} \sum_{m \in \mathcal{M}_l} \sum_{y \in Y} \sum_{p \in \bar{P}_m^y} \sum_{r \in R_y} \rho_l^y \bar{c}_r^y x_{rp}^y \quad \forall l \in L \tag{14c}$$

$$\Delta \geq \sum_{y \in Y} \sum_{p \in \bar{P}_m^y} \sum_{r \in R_y} \rho_l^y \bar{c}_r^y x_{rp}^y - \mu_l \quad \forall l \in L, m \in \mathcal{M}_l \tag{14d}$$

$$\Delta \geq \mu_l - \sum_{y \in Y} \sum_{p \in \bar{P}_m^y} \sum_{r \in R_y} \rho_l^y \bar{c}_r^y x_{rp}^y \quad \forall l \in L, m \in \mathcal{M}_l. \tag{14e}$$

$\bar{\mathcal{H}}^B$ includes a subset of the constraints from $\bar{\mathcal{H}}$, as specified in (14b). Constraints (14c) determine

the average number of patients seen per block for each half-year. Finally, Constraints (14d)–(14e), together with the objective function (14a), minimize the maximum absolute deviation in the number of patients seen in each block from the corresponding half-year average, μ_l .

Appendix D: Supplementary Tables

Rotation	PGY-1	PGY-2	PGY-3	Avg. Weekly Clinic Sessions
Inpt1	2	0	0	1
ICU	1	0	0	1
Emed	1	0	0	10
Clinic1	1	0	0	8
OB	2	0	0	3
Card1	1	0	0	2
NICU	1	0	0	1
Ped	1	0	0	1
Surg	1	0	0	3
Neur1	1	0	0	3
AO	0	1	0	4
Bmed	0	1	0	4
ELEC2	0	2	0	4
Derm	0	1	0	4
CHlth	0	1	0	4
Ger	0	1	0	4
Emed2	0	1	0	4
Ortho	0	1	0	6
PedSv	0	1	0	3
Inpt2	0	2	0	3
Clinic3	0	0	1	5
Card3	0	0	1	4
ELEC3	0	0	3	4
Inpt3	0	0	2	4
CoUro	0	0	1	5
Gyn	0	0	1	6
Nephro	0	0	1	4
PedER	0	0	1	4
PMed	0	0	1	5

Table 7: Family medicine rotations by PGYs and average weekly clinic sessions in each rotation.

PGY	Blocks 1–6	Blocks 7–12
PGY-1	3	5
PGY-2	6	8
PGY-3	9	10

Table 8: Number of patients seen per clinic session by PGY and block (Bard et al. 2016b).