Proceedings of the 2025 Winter Simulation Conference
E. Azar, A. Djanatliev, A. Harper, C. Kogler, V. Ramamohan, A. Anagnostou, and S. J. E. Taylor, eds.

A SIMULATION-BASED EVALUATION OF STRATEGIES FOR COMMUNICATING APPOINTMENT SLOTS TO OUTPATIENTS

Aparna Venkataraman[1,2], Sisira Edirippulige[2], and Varun Ramamohan[1]

[1]Department of Mechanical Engineering, Indian Institute of Technology Delhi, New Delhi, INDIA
[2]Centre for Online Health, Faculty of Health, Medicine and Behavioural Sciences, The University of Queensland, Brisbane, AUSTRALIA

ABSTRACT

In this paper, we consider an outpatient consultation scheduling system with equal-length slots wherein a set of slots each day are reserved for walk-ins. Specifically, we consider the following questions in deciding slot start times to communicate to scheduled patients: (a) should information regarding patient arrival with respect to the slot start time communicated to them (arrival offset with respect to slot start – i.e., are they typically late or early) be considered in deciding the slot start time for communication, and (b) what impact does rounding the slot start time to the nearest 5^{th} or 10^{th} minute have on relevant outcomes? We answer these questions using a validated discrete-event simulation of an FCFS outpatient appointment system in a hospital accommodating both scheduled and walk-in patients. We also describe the development of the simulation itself, which is designed to optimize policies regarding management of walk-in patients and integration of telemedicine.

1 INTRODUCTION & BACKGROUND

Outpatient appointment scheduling has the potential to enhance operational efficiency in healthcare systems, with direct implications for patient access, service quality, and staff productivity. Over the years, many scheduling models have been developed, focusing on optimizing performance metrics such as patient waiting times, provider idle time, overtime, and clinic throughput (Cayirli and Veral 2009; Gupta and Denton 2008). These models have explored approaches such as static and dynamic slot allocation (Liu and Liu 1998), variable slot lengths (Klassen and Yoogalingam 2014), overbooking, double booking (Muthuraman and Lawley 2008), and insertion of buffers to handle uncertainty (Burdett and Kozan 2015). Recent studies have also proposed a heuristic for outpatient appointment systems based on patient classification and dynamic slot recalibration (Oleskovicz et al. 2024) and a stochastic programming framework for multi-resource allocation and care sequencing under uncertainty (Yao et al. 2024). In particular, discrete-event simulation (DES) is frequently employed in this domain due to its capacity to model the inherent randomness of outpatient systems and simultaneously evaluate policies that attempt to optimize trade-offs across multiple performance metrics.

Several comprehensive reviews have described research in outpatient scheduling. Cayirli and Veral (2009) and Gupta and Denton (2008) categorize outpatient scheduling models based on queueing theory, simulation, and optimization methods. More recently, Ahmadi-Javid et al. (2017) have classified outpatient scheduling into patient-centric and provider-centric perspectives, underscoring the increasing emphasis on patient preferences and behaviors. Another study by Berg et al. (2014) has emphasized multi-objective trade-offs, including no-show risks, appointment adherence, and demand uncertainty.

While extensive work has been done to design robust scheduling policies, relatively little emphasis has been paid to the effect of the appointment time communicated to the patient, especially in terms of how the communicated slot start time is constructed, referred to in this study as the "communicated slot." For instance, most simulation models assume punctual arrivals or use probabilistic assumptions to model arrival

variability (Kim et al. 2018) without linking it to the appointment time communicated. However, clinics routinely round appointment times to the nearest 5 or 10 minutes, either for administrative convenience or to avoid communicating slot start times such as 09:07 or 13:14 that patients may find 'strange'. Such rounding might alter patient arrival patterns in subtle ways, potentially affecting waiting times, idle time, and server utilization. Additionally, the combined effect of rounding rules and patient arrival offset to account for habitual early or late arrival behaviors observed across patient cohorts on outpatient scheduling performance has not been rigorously investigated.

To address this gap, this study uses a discrete-event simulation model to systematically assess the operational impact of different communicated slot timing strategies, constructed by modifying the originally assigned appointment slot using combinations of arrival offset and rounding to the nearest 5 or 10 minutes. This simulated "communicated slot" system aims to reflect how real-world systems often buffer and reshape appointment communication to balance punctuality, service capacity, and patient expectations.

The model in this study systematically analyses four communicated slot scenarios: (i) baseline communication of the exact appointment slot without modification, (ii) rounding of the appointment slot to the nearest 5 or 10 minutes, (iii) adjustment using the average patient arrival offset, and (iv) adjustment using the average arrival offset followed by rounding. By testing these scenarios under a split-pool scheduling policy with a fixed slot duration and limited walk-in accommodation, the study isolates the effect of communicated slot construction on patient and server performance metrics, such as patient waiting times and length of stay, server overtime, and utilization.

Additionally, the analysis is not limited to the effect of offset and rounding; the model also accounts for other operational complexities, such as walk-in demand, appointment carryovers, and telemedicine-specific disruptions. The model incorporates the day-to-day carryover of unscheduled patients to the next day's schedule with priority over new appointments. This layering of service priority, beginning with emergency patients, followed by carryovers, scheduled patients, and finally, walk-ins, mirrors real clinical workflows and increases the model's fidelity. The model also integrates stochastic disruptions in telemedicine consultations, capturing the service degradation experienced during network outages or connectivity issues (Shao et al. 2024). This issue becomes increasingly relevant in settings where in-person and remote consultations are handled by the same resources, making the system vulnerable to disruption propagation. While such disruptions are acknowledged in healthcare operations literature (Qiao et al. 2023), they have not been incorporated within discrete-event scheduling models. This study takes a step toward addressing this issue by incorporating stochastic service disruptions in telemedicine consultations in the model.

Thus, this work addresses the following set of research questions relevant to outpatient system design. What is the operational impact of integrating walk-in patients alongside scheduled demand in a split-pool scheduling system? And what trade-offs emerge when systems attempt to optimize communication strategy (via communicated slot construction) while also adapting to stochastic service and demand conditions, infrastructure disruptions and appointment prioritization rules?

Existing literature addresses some of these components individually, but none has addressed the interaction of these factors within a single, unified outpatient scheduling model. For instance, studies on walk-in management (Su and Shih 2003; Morikawa and Takahashi 2017; Cayirli and Gunes 2014; Cayirli et al. 2019; Wing and Vanberkel 2022) have emphasized queueing dynamics, resource blocking and the need for flexible buffers to accommodate unscheduled arrivals. Research on telemedicine scheduling (Guo et al. 2024; Qiao et al. 2021; Bayram et al. 2019; Erdogan et al. 2018; Zhong et al. 2017) has focused on virtual care access. Likewise, many models have incorporated stochastic arrival times (Kim et al. 2018) but do not explicitly analyze the structure of the communicated appointment slot. Importantly, prior simulation studies have not systematically considered how rounding rules or arrival offsets, when applied in practice, interact with operational parameters like stochastic arrivals, walk-in demand, and service type (in-person versus telemedicine), nor have they incorporated the stochastic impact of infrastructure disruptions in virtual consultations. This gap between model fidelity and practical complexity underscores the need for simulation studies that explicitly incorporate these elements into a unified, operationally relevant framework.

To realistically capture these interactions, the simulation model developed in this study includes the following: a split-pool scheduling policy that differentiates between scheduled and walk-in patients; arrival distributions driven by offsets around assigned slots; communication of appointment times using defined rounding strategies; prioritization of emergency and carryover patients; and telemedicine-specific service disruptions.

Thus, this paper makes two main contributions. First, it introduces a framework to evaluate the effect of communicated slot construction via arrival offsets and rounding on operational metrics under realistic scheduling constraints. Second, it demonstrates how this analysis can be embedded within a broader simulation model that also accounts for walk-in arrivals, telemedicine disruptions, and dynamic service prioritization.

2 METHODOLOGY

This section elucidates the simulation structure, key model assumptions and parameterization, model validation framework, the scheduling policy, and the communicated slot construction mechanism.

2.1 Model Overview and Assumptions

The simulation was implemented in Python using the Salabim package (Van Der Ham 2018) and replicates the operational dynamics of an outpatient department in a multi-specialty tertiary care hospital in New Delhi, India. The model simulates a 42-day rolling horizon consisting of six-day work weeks (five weekdays and one Saturday), including a 12-day warm-up period. Each day consists of a single-server outpatient clinic operating for a duration of five hours, from 9:00 AM to 2:00 PM, yielding a total operational time of $H = 300$ minutes. The scheduling model divides this horizon into slots of fixed duration, $X = 14$ minutes. The number of appointment slots per day is denoted by $N = H/X$.

The system accommodates both scheduled patients (either new or follow-up) and walk-in patients, including emergencies. Consultation types are classified into six categories: Type 1 (New In-person), Type 2 (Follow-up In-person), Type 3 (New Telemedicine), Type 4 (Follow-up Telemedicine), and emergency and non-emergency walk-in consultations. Three patient classes (scheduled, walk-ins, and emergencies) arrive dynamically during the day. Scheduled patients may not arrive, governed by consultation type-specific no-show probabilities.

In the scheduling model, the fixed slot duration X was derived from the data on the consultation duration for the four types of consultation. Specifically, summary statistics (mean, maximum, and minimum) were computed across these types, yielding an average of 14.398 minutes, a maximum of 17.206 minutes, and a minimum of 11.590 minutes. The slot duration was rounded to the nearest minute and set to $X = 14$ minutes to reflect the central tendency across all consultation types. This value was selected to approximate the mean service time in a balanced manner to ensure compatibility with the different types of consultation. Here, walk-in arrivals follow a Poisson process and scheduled patients are pre-assigned slots and arrive based on the communicated slot construction mechanism detailed in Section 2.4. The model includes a rolling-horizon approach, carrying over unserved scheduled patients to subsequent days and prioritizing Emergency $>$ Carryover $>$ New Scheduled $>$ Walk-in appointments in a First-Come-First-Served (FCFS) manner.

The model encompasses stochastic variability in consultation durations, patient arrivals, and infrastructure disruptions during telemedicine appointments. Consultation durations follow lognormal or Weibull distributions depending on the consultation type. Telemedicine disruptions occur randomly, with a 45.2% chance, and uniformly distributed durations between 1-10 minutes. The key parameters of the model are included in Table 1, and the simulation model was parameterized using data from the outpatient department's historical records, in-person observation, measurement and data collection at the hospital, and discussions with hospital administrators. Slot durations, daily patient volumes, no-show rates, and walk-in proportions were derived from data collected via the hospital's management information system and verified through

discussions with clinic staff and specialists. Consultation durations for each type of appointment were recorded manually during clinic sessions on designated days using a stopwatch.

Additional data on the proportion of emergency walk-in patients and the occurrence of service disruptions in telemedicine consultations were sourced from national healthcare reports (Department of Emergency Medicine, JPNATC, AIIMS. 2021) and recently published peer-reviewed literature (Grover et al. 2022), respectively. Anderson-Darling goodness-of-fit tests were used to identify best-fit distributions for arrival and service processes. The data indicated non-stationary patient arrival patterns and loads on weekdays and Saturdays, reflecting higher demand for scheduled appointments on Saturdays. Consequently, the model was structured into six active clinic days per week (five weekdays and one Saturday) to reflect these operational dynamics.

Table 1: Key model parameters and notation. *Notes.* MOE: margin of error.

Parameter & Notation	**Value / Distribution**		**Source / Notes**
	Weekdays	**Saturday**	
Clinic working hours per day	09:00 – 14:00 (300 mins)		Hospital records
Slot duration (X) (min)	14		Mean of min, average, and max durations (Types 1–4)
Total slots per day (T)	$T = 300/X = 21$		Computed based on working hours and slot duration
Walk-in proportion (p)	0.49	0.3	Retrospective analysis of patient attendance records
Emergency walk-in proportion	0.07		(Department of Emergency Medicine, JPNATC, AIIMS. 2021)
Proportion of Consultation types (Type 1–4)	Type 1: 0.4, Type 2: 0.29, Type 3: 0.11, Type 4: 0.20		Retrospective analysis of patient attendance records
No-show probabilities	Type 1: 0.27 Type 2: 0.13 Type 3: 0.30 Type 4: 0.30	Type 1: 0.35 Type 2: 0 Type 3: 0.30 Type 4: 0.30	Retrospective analysis of patient attendance records
Disruption probability (telemedicine consultations)	0.452		(Grover et al. 2022)
Disruption duration (min)	$\mathscr{U}(1,10)$		Discussion with a specialist
Patient arrival offset (ϑ) (min)	$\mathscr{N}(-0.875, 19.51)$	$\mathscr{N}(-9.33, 22.23)$	Estimated from the patient arrival time
Consultation duration (min)	Type 1, 3: Lognormal$(2.689, 0.559)$ Type 2, 4: Weibull$(2.17, 13.084)$		Stopwatch-based observation
Number of replications	30		Sample size to achieve 5% MOE for output metrics
Warm-up period	2 weeks (12 days)		Based on visual inspection of convergence

2.2 Model Validation

The simulation model was validated using hospital data, comprising 131 data points collected over seven days. Three key performance indicators, average waiting time, average length of stay (LOS), and server utilization, were evaluated. Simulation results from 20 replications were compared against hospital data using Welch's t-tests, chosen for its robustness to unequal variances and sample sizes. The Shapiro-Wilk test assessed the normality of hospital data, and when non-normality was detected, the Mann-Whitney U test was applied to compare medians.

Initial comparisons revealed high equivalence for LOS and average waiting time, with 100% and 95% of simulation replications respectively showing statistical equivalence in means at the 95% confidence level. However, only 55% of replications achieved equivalence for server utilization, necessitating adjustments to better align the simulation with observed hospital performance.

Discrepancies observed between the simulated and recorded hospital waiting times were primarily attributed to random operational delays within the hospital environment that were not explicitly represented in the simulation model. These delays included interruptions such as physicians being temporarily unavailable due to administrative responsibilities, emergency inpatient interventions, or phone calls. Although these delays contributed to observed waiting times in the hospital, they are external to the modeled patient–server interactions and, therefore, not captured within the simulation framework.

To improve the alignment between simulation outputs and observed hospital data, an adjustment was introduced to the wait times recorded from the hospital for validation purposes to account for such unexplained delays. Hospital data revealed that 22% of the consultations encountered unexplained delays prior to the start of their service, with a mean of $\mu = 33.7$ minutes ($\sigma = 18.14$ minutes), with 12.3% of cases exhibiting delays greater than the mean.

Based on this distribution, the following adjustment was applied to the wait time validation data recorded from the hospital to yield an adjusted waiting time wt_{adjusted}:

$$wt_{\text{adjusted}} = \begin{cases} 0 & \text{if } wt > \mu + \sigma \\ wt(1 - 0.123) & \text{if } \mu < wt \leq \mu + \sigma \\ wt & \text{otherwise.} \end{cases}$$

This adjustment was defined under the rationale that:

- If the observed waiting time wt exceeds $\mu + \sigma$, it is assumed to be influenced by hospital-side delays not modeled in the simulation, and hence it is excluded.
- For delays wt marginally above the mean but within one standard deviation (between μ and $\mu + \sigma$), it is proportionally reduced by 12.3%, corresponding to the observed proportion of high-delay cases.
- If wt is within expected bounds (i.e., less than or equal to μ), no adjustment is made.

The above adjustment ensures that only delays arise due to the dynamics of the system that correspond to considerations incorporated in the modeled system are retained, rather than systemic inefficiencies external to the model logic, thereby enhancing the comparability between simulated and real-world data.

Server utilization adjustments addressed differences in operational duration. While the simulation assumed a five-hour working window, the hospital data for utilization was recorded over a shorter mean window of 2.5 hours. Thus, hospital utilization values were scaled accordingly as

$$\text{Scaled Utilization} = \left(\frac{\text{Utilization}}{\text{Duration (hours)}} \right) \times 5.$$

After applying these adjustments, 100% of replications showed statistical equivalence in all three metrics, confirming the simulation model's validity and ability to replicate the outpatient system under realistic constraints.

2.3 Split-Pool Scheduling Policy

In the scheduling policy employed in this study, the total T slots in each day are partitioned into a Scheduled Pool (of $N = T - W^*$ slots) and a Walk-in Pool (of $W^* = \lfloor T \times p \rfloor$ slots), as depicted in Figure 1.

For instance, with $X = 14$ minutes and walk-in proportion $p = 0.49$ (weekday), we get:

$$T = \left\lfloor \frac{300}{14} \right\rfloor = 21, \quad W^* = \lfloor 21 \cdot 0.49 \rfloor = 10, \text{ and } N = 11.$$

For $X = 14$ minutes and $p = 0.3$ (Saturday),

$$T = 21, \quad W^* = \lfloor 21 \cdot 0.3 \rfloor = 6, \text{ and } N = 15.$$

Scheduled patients (including carryover) are assigned to the scheduled pool N, while walk-in arrivals are allocated only to W^*.

Carryover patients from the previous day are allocated scheduled slots first, followed by new scheduled patients. Walk-ins are allocated to any remaining walk-in slots and emergency walk-ins can preempt the queue. Walk-in patients who do not receive service by the end of the day are treated as walkaways and are not rescheduled. The algorithm for the split-pool scheduling policy is elucidated in Algorithm 1.

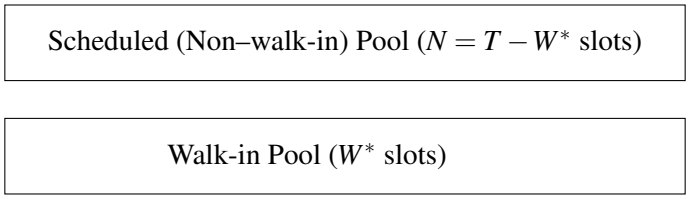

Figure 1: Slot allocation in split-pool outpatient scheduling.

2.4 Communicated Slot Time Construction

Within this scheduling framework, each scheduled patient is assigned an appointment slot starting at time t_a, and a communicated slot time t_c that is shared with the patient. Let the arrival offset ϑ be a random variable representing patient arrival deviations relative to t_a, with mean μ_ϑ. When arrival offsets are considered, the modified slot start time is given by:

$$t'_a = t_a - \mu_\vartheta.$$

Here $\vartheta \sim \mathcal{N}(\mu_\vartheta, \sigma^2)$, truncated between $[-40, 40]$ minutes. From observations at the hospital, we see that $\mu_\vartheta = -0.875$ minutes for weekdays and -8.33 minutes for Saturdays. The above approach for incorporating arrival offset implies that if habitual late arrival behavior is observed, the communicated slot starts earlier, and vice versa when habitual early arrival behavior is observed.

The communicated slot time t_c is then computed as:

$$t_c = round(t'_a,\ 5 \times \alpha), \alpha \in \{1, 2\}.$$

The above approach captures actual operational practice wherein optimal slot durations are determined to be numbers not divisible by 5 or 10, but the slot start time is rounded to the nearest 5 or 10 minutes so as to arrive at slot start times that in general patients are more comfortable with (i.e., patients may find being asked to arrive at, say, 10:28 AM rather than 10:30 or 10:25 AM, 'strange').

3 RESULTS

This section presents the simulation results across all experimental scenarios, highlighting the operational impact of varying communicated slot timing strategies. The results are structured to first describe the experimental setup, followed by a comparative analysis of key performance indicators.

3.1 Experimental Scenarios

To evaluate the impact of communicated slot timing on clinic operations, the model evaluates six scenarios grouped into four conceptual categories under the split-pool scheduling policy. These scenarios differ in whether the patient arrival offset μ_ϑ and/or rounding to the nearest $\alpha \times 5$ minutes ($\alpha \in \{1, 2\}$) is applied

Algorithm 1 Split-pool outpatient scheduling policy.

Input: Total slots T, walk-in proportion p, scheduled demand D_{sch}
Service Priority: Emergency > Carryover > Scheduled > Walk-in
$W^* \leftarrow \lfloor T \cdot p \rfloor$ ▷ Designated walk-in slots
$N \leftarrow T - W^*$ ▷ Scheduled pool slots
for each day **do**
    Initialize emergency queue, carryover list C and walkaway list R
    for $i = 1$ to N **do** ▷ Process scheduled pool
        if $C \neq \varnothing$ **then**
            Schedule the earliest carried-over scheduled request into slot i
        else if $D_{sch} \neq \varnothing$ **then**
            Schedule a new scheduled request into slot i
        else
            break ▷ No further scheduled requests
        end if
    end for
    for $i = 1$ to W^* **do** ▷ Process walk-in pool
        Schedule a walk-in request into walk-in slot i
    end for
    Merge all scheduled and walk-in requests into a master queue Q ▷ Merge arrivals and sort
    Sort Q: (i) service priority (ii) arrival time
    while (server is free) **do**
        Dequeue the next request r from Q
        Activate in priority
    end while
    for each request r in Q **do** ▷ End-of-day handling
        if r is scheduled (non-walk-in) **then**
            Add r to C
        else
            Mark r as walkaway and add to R
        end if
    end for
end for

in computing the communicated slot time t_c. The six scenarios are defined as follows:

Scenario 1: $t_c = t_a$ (exact appointment time communicated)
Scenario 2a: $t_c = \text{round}(t_a, 5)$ (rounding to nearest 5 minutes only)
Scenario 2b: $t_c = \text{round}(t_a, 10)$ (rounding to nearest 10 minutes only)
Scenario 3: $t_c = t_a - \mu_\vartheta$ (offset only)
Scenario 4a: $t_c = \text{round}(t_a - \mu_\vartheta, 5)$ (offset and rounding to nearest 5 minutes)
Scenario 4b: $t_c = \text{round}(t_a - \mu_\vartheta, 10)$ (offset and rounding to nearest 10 minutes)

Each scenario is tested using the same slot duration of 14 minutes, and a walk-in proportion of 0.49 on weekdays and 0.3 on Saturdays, ensuring the effect of communicated slot on average waiting time, and length of stay for walk-in and scheduled patients, server overtime, and server utilization is isolated for analysis.

The findings from the experiments are highlighted in Table 3 and in Figures 2 and 3. These results serve as the foundation for the subsequent discussion in Section 4, where the scenario-specific comparisons and underlying trade-offs are examined. Table 2 shows how the communicated slot times vary across the six scenarios, highlighting the effects of offset and rounding for different appointment slots on a weekday. Table 3 summarizes the mean and standard deviation (SD) of the key performance indicators, while Figures 2 and 3 provide comparative visualizations across the various rounding methods and scenarios.

3.2 Impact on Scheduled and Walk-in Patients

From Table 3 and Figure 2, Scenario 1, where the exact appointment time is communicated without offset or rounding, consistently achieves the lowest average wait time (6.73 min) and length of stay (LOS) (19.29 min) for scheduled patients. In contrast, Scenario 3, which uses offset-only construction, results in prolonged delays for scheduled patients (Wait time: 19.27 min, LOS: 32.83 min). This trend in Scenario 3 reinforces that introducing offset alone without a compensatory rounding mechanism may exacerbate patient congestion due to misalignment between slot timing and actual arrival behavior.

For walk-in patients, the results present a more nuanced picture. Under 5-minute rounding, Scenario 4 achieves the lowest average wait time (58.01 min) and LOS (68.14 min), while Scenario 3 again exhibits the worst performance (Wait: 60.14 min, LOS: 70.67 min). Interestingly, Scenario 1 performs slightly better than Scenario 2 on walk-in metrics while also maintaining superior scheduled patient performance. This suggests that adding 5-minute rounding without offset correction in Scenario 2 may not improve efficiency and slightly degrade performance across patient groups.

When switching to 10-minute rounding, performance degradation is observed across all patient metrics. As illustrated in Figure 3, Scenario 4 with 10-minute rounding sees wait times for non-walk-ins increase to 20.46 min and LOS to 32.53 min, over three times the baseline performance in Scenario 1. Walk-in metrics also deteriorate under this scenario, reaching the highest observed LOS (77.96 min). These findings highlight that aggressive rounding of 10 minutes, even with offset correction, may distort arrival patterns and resource allocation enough to impair performance.

In both Scenario 2 and Scenario 4, switching from 5-minute to 10-minute rounding increases patient delays across all categories. The comparative plots in Figure 3 reinforce this. For instance, the wait time for scheduled patients increases from 7.04 to 9.29 min in Scenario 2 and from 12.36 to 20.46 min in Scenario 4, while the server overtime remains relatively stable. These results emphasize the trade-off between improved coordination through communicated slot timing and the variability it introduces, mainly when offset and rounding mechanisms interact.

3.3 Impact on Server Metrics

From a provider perspective, server overtime remains relatively stable across all scenarios, fluctuating between 21.05 and 21.83 minutes. However, server utilization presents more variability. As shown in Table 3 and Figure 2, Scenario 4 with 10-minute rounding, despite its poor patient outcomes, records the highest utilization (1.027), while Scenario 2 with 10-minute rounding records the lowest (0.912). This inverse relationship suggests that higher utilization does not necessarily correlate with improved patient outcomes, particularly when increased workload stems from inefficient slot structuring or misaligned arrival time.

Interestingly, Scenario 4 exhibits increased server utilization when shifting from 5-minute to 10-minute rounding, despite a reduction in server overtime. This could suggest a redistribution of service rather than an actual improvement in operational efficiency. The rounding appears to consolidate appointment times and likely elevates the load on the server, increasing utilization while key performance metrics such as patient wait times and length of stay continue to deteriorate.

Table 2: Example communicated slot construction on weekdays ($X = 14$, $\mu_\vartheta = -0.875$). *Notes.* X: slot duration, μ_ϑ: mean arrival offset, a: appointment slot, c: communicated slot, SY: Scenario Y.

Slot (a)	**S1:** $c = a$ $\bar{X} = 14$	**S2a:** **round**$_5(a)$ $\bar{X} = 14.29$	**S2b:** **round**$_{10}(a)$ $\bar{X} = 15$	**S3:** $a + \mu_\vartheta$ $\bar{X} = 14$	**S4a:** **round**$_5(a + \mu_\vartheta)$ $\bar{X} = 13.75$	**S4b:** **round**$_{10}(a + \mu_\vartheta)$ $\bar{X} = 15$
09:00	09:00	09:00	09:00	08:59	09:00	09:00
09:14	09:14	09:15	09:10	09:13	09:15	09:10
09:28	09:28	09:30	09:30	09:27	09:25	09:30
09:42	09:42	09:40	09:40	09:41	09:40	09:40

Table 3: Summary of performance metrics across experimental scenarios (Mean (SD)). *Notes.* LOS: Length of stay, WI: walk-in, SY: Scenario Y.

Scenario	**Wait Time (Non-WI) (min)**	**Wait Time (WI) (min)**	**LOS (Non-WI) (min)**	**LOS (WI) (min)**	**Server Overtime (min)**	**Utilization**
S1	6.728 (0.062)	58.358 (9.756)	19.287 (0.031)	67.722 (10.521)	21.741 (4.331)	0.951 (0.002)
S2a	7.039 (0.076)	58.704 (9.650)	22.689 (0.056)	69.450 (10.500)	21.424 (4.254)	0.943 (0.002)
S2b	9.287 (0.062)	59.689 (9.808)	23.168 (0.025)	69.587 (10.567)	21.050 (4.188)	0.912 (0.002)
S3	19.273 (0.035)	60.139 (10.055)	32.825 (0.038)	70.665 (10.922)	21.833 (4.332)	0.990 (0.002)
S4a	12.358 (0.053)	58.011 (9.735)	24.435 (0.051)	68.136 (10.549)	21.530 (4.284)	0.997 (0.002)
S4b	20.456 (0.036)	66.872 (10.465)	32.527 (0.036)	77.961 (11.308)	21.192 (4.245)	1.027 (0.002)

4 DISCUSSION

4.1 Effect of Arrival Offset

It might be expected that incorporating a negative arrival offset in Scenarios 3, 4a, and 4b would lead scheduled patients to arrive earlier and therefore reduce their waiting time. However, the results in Table 3 indicate the opposite: applying this offset leads to higher waiting times and increased LOS for scheduled patients, compared to Scenario 1. This is likely because the baseline utilization itself is very high - around 95% - and incorporating offset decreases the likelihood of the server being idle between consultations, thereby increasing utilization. When utilization increases from an already high baseline, we observe the substantial increases in wait time that are seen in Table 3.

4.2 Effect of Combining Offset and Rounding

In Scenarios 4a and 4b, the offset is combined with rounding of communicated slots to the nearest 5 or 10 minutes. Table 2 shows an effective slot duration of approximately 13.75 minutes in Scenario 4a and 15 minutes in Scenario 4b. This rounding causes more patients to arrive at similar times, further increasing the queue length. The coarser rounding in Scenario 4b effectively overlaps more arrivals, which further increases the server utilization and leads to longer waiting times. Therefore, although the offset may help align patient arrival with system working hours, its combination with rounding causes longer queues and reduces the expected benefit.

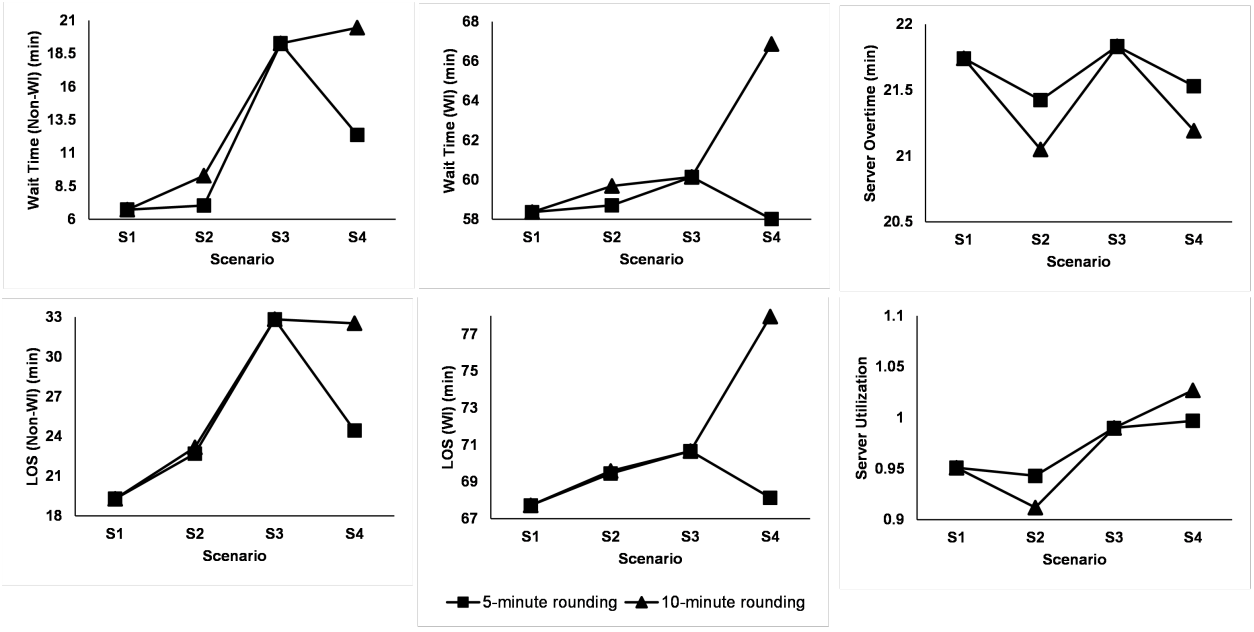

Figure 2: Comparison of average performance metrics across scenarios using 5- and 10-minute communicated slot rounding. *Notes.* LOS: Length of stay, WI: walk-in.

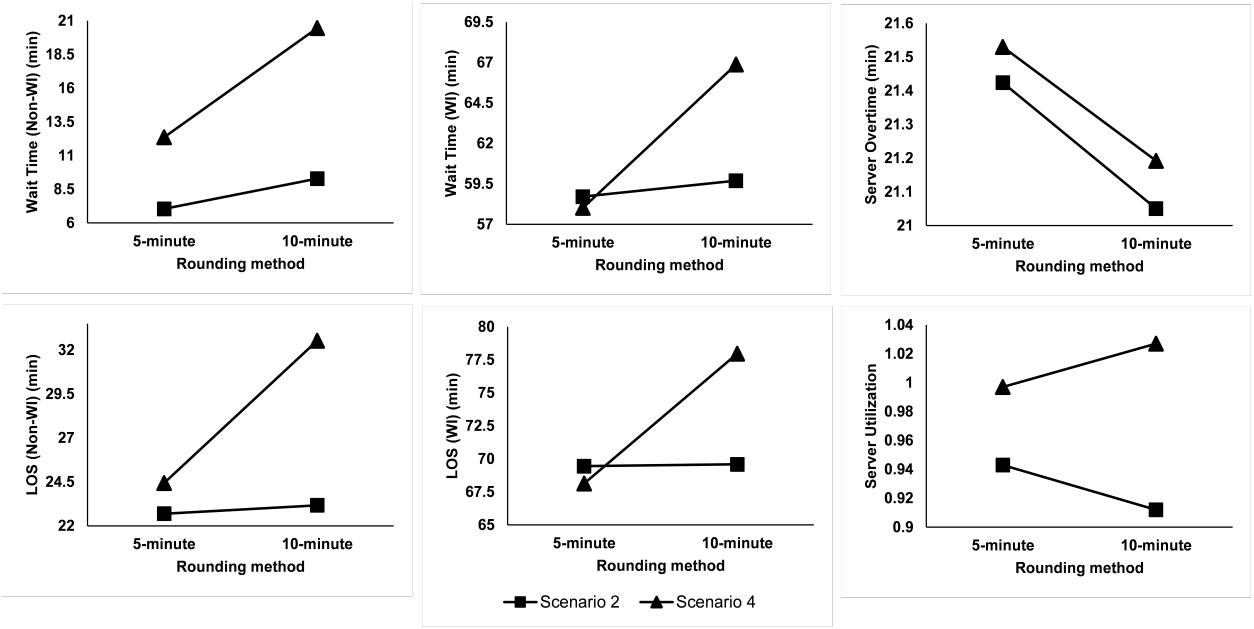

Figure 3: Comparison of influence of 5- and 10-minute communicated slot rounding on average performance metrics. *Notes.* LOS: Length of stay, WI: walk-in.

4.3 Effect of Rounding Alone

In Scenarios 2a and 2b, where only rounding is applied (without incorporating offset), the performance worsens compared to Scenario 1. At the beginning of the day, rounding can create gaps in arrival times that lead to server idle time. Later in the day, rounding causes many patients to arrive around the same time, resulting in queues. These

variations in load, between idle time and congestion, lead to lower utilization and higher waiting times. This is more visible in Scenario 2b (10-minute rounding), where the larger rounding granularity increases the spacing between arrivals and decreases server utilization. Although Scenario 2a (5-minute rounding) performs slightly better, it still underperforms compared to the baseline scenario.

Note that at such high server utilizations, the decrease in utilization (at the expense of substantially higher patient wait times) with coarser rounding may also provide providers with much-needed respite, in turn implying that this trade-off may need to be optimized by attaching (artificial, if required) costs to changes in each metric.

4.4 Interaction Between Offset and Rounding

Overall, Scenario 1 shows the most balanced performance across all metrics, where communicated slots are equal to the scheduled appointment times without offset or rounding. Scenario 4a, which applies both offset and 5-minute rounding, may serve as a practical compromise to balance scheduled and walk-in demands in situations where rounding is operationally required. However, the performance drop observed in Scenario 4b demonstrates that the system is highly sensitive to the rounding granularity. This is clearly illustrated in Figure 2 and Figure 3, where even modest changes in communicated slot timing leads to measurable differences in both patient and server metrics. These results indicate that the interaction between offset, rounding, and slot granularity must be carefully considered during scheduling system design.

5 LIMITATIONS AND FUTURE WORK

While this study provides valuable insights into communicated slot timing, it is limited by its use of a fixed slot duration (14 minutes) and a split-pool scheduling policy. These assumptions, while representative, may restrict the generalizability of findings to broader clinical settings. Future studies should evaluate these communicated slot constructions under varying slot durations, walk-in proportions, and alternative scheduling policies to fully capture their operational implications.

Moreover, the observed performance shifts across rounding methods suggest that Scenarios 2 and 4 may depend highly on slot granularity. If these scenarios appear suboptimal in certain configurations, it may not be due to the communicated slot logic *per se*, but rather due to their interplay with slot duration and arrival offsets. Further experiments varying the underlying slot duration would be essential to validate or refute this dependency.

Furthermore, individual-level modeling of patient behavior with respect to communicated slot start times, including demographic and cultural influences on punctuality, could further strengthen the realism and applicability of the simulation framework. For example, it would be interesting to consider whether, when provided with a slot start time that is not a multiple of 5, patients perform rounding themselves to a multiple of 5 and arrive with respect to this self-modified slot start time.

REFERENCES

Ahmadi-Javid, A., Z. Jalali, and K. J. Klassen. 2017. "Outpatient Appointment Systems in Healthcare: A Review of Optimization Studies". *European Journal of Operational Research* 258(1):3–34.

Bayram, A., S. Deo, S. Iravani, and K. Smilowitz. 2019. "Managing Virtual Appointments in Chronic Care". *IISE Transactions on Healthcare Systems Engineering* 10(1):1–17.

Berg, B. P., B. T. Denton, S. Ayca Erdogan, T. Rohleder, and T. Huschka. 2014. "Optimal Booking and Scheduling in Outpatient Procedure Centers". *Computers & Operations Research* 50:24–37.

Burdett, R., and E. Kozan. 2015. "Techniques to Effectively Buffer Schedules in the Face of Uncertainties". *Computers & Industrial Engineering* 87:16–29.

Cayirli, T., P. Dursun, and E. D. Gunes. 2019. "An Integrated Analysis of Capacity Allocation and Patient Scheduling in Presence of Seasonal Walk-Ins". *Flexible Services and Manufacturing Journal* 31(2):524–561.

Cayirli, T., and E. D. Gunes. 2014. "Outpatient Appointment Scheduling in Presence of Seasonal Walk-Ins". *Journal of the Operational Research Society* 65(4):512–531.

Cayirli, T., and E. Veral. 2009. "Outpatient Scheduling in Health Care: A Review of Literature". *Production and Operations Management* 12(4):519–549.

Department of Emergency Medicine, JPNATC, AIIMS. 2021. "Emergency and Injury Care at Secondary and Tertiary Level Centres in India". Report, NITI Aayog. https://www.niti.gov.in/sites/default/files/2021-12/AIIMS_STUDY_1.pdf, accessed 15.01.2024.

Erdogan, S. A., T. L. Krupski, and J. M. Lobo. 2018. "Optimization of Telemedicine Appointments in Rural Areas". *Service Science* 10(3):261–276.

Grover, S., C. Naskar, S. Sahoo, and A. Mehra. 2022. "Clinician's Experience of Telepsychiatry Consultations". *Asian Journal of Psychiatry* 75:103207.

Guo, H., Y. Xie, B. Jiang, and J. Tang. 2024. "When Outpatient Appointment Meets Online Consultation: A Joint Scheduling Optimization Framework". *Omega* 127:103101.

Gupta, D., and B. Denton. 2008. "Appointment Scheduling in Health Care: Challenges and Opportunities". *IIE Transactions* 40(9):800–819.

Kim, S.-H., W. Whitt, and W. C. Cha. 2018. "A Data-Driven Model of an Appointment-Generated Arrival Process at an Outpatient Clinic". *INFORMS Journal on Computing* 30(1):181–199.

Klassen, K. J., and R. Yoogalingam. 2014. "Strategies for Appointment Policy Design with Patient Unpunctuality". *Decision Sciences* 45(5):881–911.

Liu, L., and X. Liu. 1998. "Dynamic and Static Job Allocation for Multi-Server Systems". *IIE Transactions* 30(9):845–854.

Morikawa, K., and K. Takahashi. 2017. "Scheduling Appointments for Walk-Ins". *International Journal of Production Economics* 190:60–66.

Muthuraman, K., and M. Lawley. 2008. "A Stochastic Overbooking Model for Outpatient Clinical Scheduling with No-Shows". *IIE Transactions* 40(9):820–837.

Oleskovicz, M., M. C. Pedroso, and J. L. Biazzi. 2024. "Outpatient Appointment Systems: A New Heuristic with Patient Classification". *Operations Research for Health Care* 43:100443.

Qiao, Y., L. Ran, J. Li, and Y. Zhai. 2021. "Design and Comparison of Scheduling Strategy for Teleconsultation". *Technology and Health Care* 29(5):939–953.

Qiao, Y., Y. Zhai, R. Ma, M. Ji, and W. Lu. 2023. "Optimizing Teleconsultation Scheduling to Make Healthcare Greener". *Journal of Cleaner Production* 422:138569.

Shao, C. C., M. H. Katta, B. P. Smith, B. A. Jones, L. T. Gleason, A. Abbas, *et al*. 2024. "Reducing No-Show Visits and Disparities in Access: The Impact of Telemedicine". *Journal of Telemedicine and Telecare* 31(7):1041–1049.

Su, S., and C.-L. Shih. 2003. "Managing A Mixed-Registration-Type Appointment System in Outpatient Clinics". *International Journal of Medical Informatics* 70(1):31–40.

Van Der Ham, R. 2018. "Salabim: Discrete Event Simulation and Animation in Python". *Journal of Open Source Software* 3(27):767.

Wing, J., and P. Vanberkel. 2022. "Simulation Optimisation for Mixing Scheduled and Walk-In Patients". *Health Systems* 11(4):276–287.

Yao, X., K. S. Shehadeh, and R. Padman. 2024. "Multi-Resource Allocation and Care Sequence Assignment in Patient Management: A Stochastic Programming Approach". *Health Care Management Science* 27(3):352–369.

Zhong, X., J. Li, P. A. Bain, and A. J. Musa. 2017. "Electronic Visits in Primary Care: Modeling, Analysis, and Scheduling Policies". *IEEE Transactions on Automation Science and Engineering* 14(3):1451–1466.

AUTHOR BIOGRAPHIES

APARNA VENKATARAMAN is a UQ-IITD joint Ph.D. student and a Prime Minister's Research Fellow affiliated with the Centre for Health Services Research at The University of Queensland (UQ) and Industrial Engineering in the Department of Mechanical Engineering at the Indian Institute of Technology Delhi (IITD). Her research interests include simulation and simulation optimization. Her e-mail address is aprna.v@uqidar.iitd.ac.in.

SISIRA EDIRIPPULIGE is an Associate Professor affiliated with the Centre for Health Services Research in the Faculty of Health, Medicine and Behavioural Sciences at The University of Queensland. His research interests include the development, promotion and integration of e-health in the health care sector. His e-mail address is s.edirippulige@uq.edu.au.

VARUN RAMAMOHAN is an Associate Professor in the Department of Mechanical Engineering at the Indian Institute of Technology Delhi, located in New Delhi, India. His research interests include probabilistic modeling, simulation and simulation optimization, with applications in healthcare operations, health economics and outcomes research, and public transportation. His e-mail address is varunr@mech.iitd.ac.in and his website is https://web.iitd.ac.in/~varunr.