

Managing Appointment-based Services with Electronic Visits

Yun Cai^a, Haiqing Song^b and Shan Wang^{b,*}

^aLingnan College, Sun Yat-sen University, Guangzhou 510275, China

^bSchool of Business, Sun Yat-sen University, Guangzhou 510275, China

ARTICLE INFO

Keywords:

OR in health services

E-visits

appointment scheduling

stochastic programming

ABSTRACT

Electronic visits, or “E-visits” for short, have emerged as a promising channel for accessing healthcare and can significantly impact daily operations in healthcare facilities. However, there is a lack of research on how to efficiently manage appointments for outpatient care providers when faced with E-visits that exhibit different waiting cost patterns. Our study investigates how providers can use appointment scheduling as a “passive” control when patients have full access to the E-visit channel, to better utilise resources and reduce patient waiting. Specifically, we demonstrate that multimodularity still applies to the model with E-visits, despite their waiting costs being typically nonlinear. Furthermore, we analyse how providers can “actively” control the arrival of E-visits by scheduling their time windows. By examining the structures of the optimal joint schedule of appointments and E-visit time windows, and reformulating the problem into a two-stage program, we have designed an Accelerated Cut Generation Algorithm, which is shown to be efficient in our numerical study. To the best of our knowledge, this is the first study to explore the optimal scheduling of both appointments and E-visit time windows. By implementing proper scheduling, the negative impact of E-visits can be mitigated, their benefits to the provider can be enhanced, and overall operational efficiency can be improved.


1. Introduction

The rapid development of electronic portals has prompted many institutions and companies to offer their services in a remote and/or online format. To name a few examples, Starbucks introduced mobile ordering and payment services in 2015, which has now become the second most popular mobile payment service with 31.2 million users. According to Starbucks’ Q4 and Full Year Fiscal 2022 Results¹, mobile orders in North America accounted for 72% of the company’s total sales volume. This trend is also prevalent in the financial industry. China Banking Association (2022) revealed that banking financial institutions’ off-the-counter transactions reached 257.282 billion yuan, making an 11.46% increase year-on-year. The industry’s average electronic channel diversion rate was 90.29%. Offering services remotely and/or online provides benefits to service providers. It can boost revenue by stimulating additional demand through online channels, fully utilise capacity and resources by serving customers via electronic portals during idle time, and improve customer satisfaction by providing an additional service channel.

One service industry that has seen a significant increase in remote and online visits is healthcare. In particular, E-visits with primary care providers have become a popular option for non-emergency concerns. This online option enables patients to access care without having to physically travel to a clinic, saving both time and money. According to *American Medical Association*², the largest and only national association,

*This work was supported by the National Natural Science Foundation of China [grant numbers 72001220, 71771222].

*Corresponding author

 caiy37@mail2.sysu.edu.cn (Y. Cai); songhq@mail.sysu.edu.cn (H. Song); wangsh337@mail.sysu.edu.cn (S.

Wang)

¹Starbucks Q4 and Full Fiscal Year 2022 Earnings Conference Call

²Telehealth Survey Report 2021

85% of physicians employ telehealth in their practice. In addition, patients are encouraged by a significant proportion of physicians (more than 80%) to adopt telehealth when accessing care, given its potential benefits. Telehealth statistics³ demonstrate that about 74% of U.S. patients embrace telehealth services, as it provides a convenient option for those living in rural areas who would otherwise have to bear the cost of travelling long distances to access medical care. Third-party E-visit platforms, such as Teladoc and Amwell, have played an instrumental role in promoting the rapid development of telemedicine by offering convenient and efficient E-visit services that complement traditional in-person care. With these platforms, the patient can easily access care by submitting a request and waiting for a response from the physicians. E-visit platforms offer faster and more flexible ways to provide outpatient care, improve patient access, reduce the burden on clinics, and lower the cost of care. According to Zocchi et al. (2020), telemedicine, including E-visits, can reduce healthcare costs by an average of 13% by decreasing in-person visits. E-visits have become a critical access channel for outpatient care, and this trend is expected to continue even in the post-pandemic future.

Despite the increasing significance and prevalence of E-visits in outpatient care, there is still a lack of knowledge regarding how to manage appointments when primary providers also serve patients from the E-visit channel. Unlike patients who make an appointment in advance, those who utilise E-visits typically do not schedule an appointment beforehand. For example, on Teladoc and PlushCare, the third-party telemedicine platforms in the U.S., patients can book an online same-day visit with a primary care provider and then wait for the service. However, managing such patients should be approached differently than managing walk-in patients who do not have an appointment either. Because E-visit patients benefit from the convenience of remote access, they are often willing to wait for the services. As a result, the waiting cost for E-visit patients should differ from that of walk-in patients who are physically present in the clinic. Current practice deals with E-visits by setting up appointment schedules and reserving the last few slots to serve them, but this approach lacks scientific understanding of how to manage daily operations in the face of uncertain arrivals from different channels. Without careful planning for E-visits, daily operations will be disrupted, patients may experience extremely long wait times, providers may need to work overtime, and the quality of care may be negatively impacted. Existing literature has developed models and insights for managing the daily operations of clinics, but these models and insights do not account for E-visits. To manage daily operations in the presence of E-visits, new modelling and solution approaches are needed.

In this paper, we develop an optimisation model to handle uncertain E-visit arrivals. Specifically, we consider a single provider who caters to three types of patients – scheduled patients with advance appointments, walk-in patients who head in the clinic, and E-visit patients who request services remotely. During working hours, walk-in and E-visit patients may arrive without prior notice. To prepare for such potential arrivals, the provider decides on how many appointments to allocate to each slot. In addition to the “passive” control of reserving slots for E-visits during appointment scheduling, the provider can also use “proactive” management by scheduling the available time window for E-visits. This means that only E-visits who arrive during the time window will be accepted. Considering costs due to patient waiting, provider idling and overtime, the provider needs to make the best use of these two operational levers. By optimising the schedule of appointments and the schedule of time windows for E-visits jointly, the provider can achieve a minimum total cost.

Our paper makes several key contributions. Firstly, we identify the optimal appointment schedules for a provider serving patients from multiple channels, and address the general arrival process of E-visits, together with walk-in and no-show behaviour of in-clinic patients. We demonstrate that even when E-visit patients exhibit different and non-linear patterns of waiting costs compared to walk-ins, the mathematical properties of our optimisation model remain unchanged. Specifically, we prove that the objective function is multimodular in the number of scheduled patients in each slot. This enables us to use local search to find

³Telehealth Statistics and Telemedicine Trends 2023

the optimal schedule efficiently. To the best of our knowledge, we are the first to extend the property of multimodularity to the context of non-linear waiting costs in appointment scheduling.

Furthermore, we show that by jointly optimising the schedules of appointments and time windows for E-visits, the provider can further enhance operational efficiency. However, this extension presents a significant challenge, as the dimension of the decision variables increases substantially, and the property of multimodularity no longer holds. To solve this problem, we identify elegant structures of the optimal solution that can be used to design an accelerated cut generation algorithm for obtaining the optimal joint schedule of appointments and time windows for E-visits.

Our analysis reveals that, E-visit patients can benefit the provider by reducing idle time and improving the utilisation of healthcare resources. However, they also introduce a new source of uncertainty into the system. Appointment scheduling serves as a “passive” control that can offset some adverse effects of E-visits. Additionally, better management of E-visits requires a “proactive” control by optimising the available time windows for E-visits. By properly scheduling the time windows as well as appointments, the negative impact of E-visits can be mitigated, their benefits to the provider can be enhanced, and overall operational efficiency can be improved.

The remainder of this paper is organised as follows: Section 2 is a literature review. Section 3 develops a basic appointment-scheduling model and discusses its mathematical properties and solution approaches. Section 4 incorporates the scheduling of time windows for E-visits into the model, explores the structures of the optimal schedule, and designs an efficient algorithm to solve this problem. Section 5 conducts numerical experiments and reveals how to manage the E-visits passively and proactively. Section 6 conducts a case study using a synthetic dataset to investigate the practical importance of the proposed model and solutions. Section 7 makes conclusions and discussions. All proofs of the analytical results can be found in Online Appendix.

2. Literature Review

Our work relates to two major streams of literature: (outpatient) appointment scheduling and healthcare operation management with E-visits.

2.1. Appointment Scheduling

The provision of pre-scheduled appointments can improve healthcare operations by reducing demand variability and allowing providers to better plan their daily operations. This viewpoint is supported by comprehensive reviews, including Cayirli and Veral (2003), Gupta and Denton (2008), Pinedo et al. (2015), and Ahmadi-Javid et al. (2017). To optimise outpatient appointment scheduling, two types of decision variables are considered: the appointment time for each patient or the number of patients scheduled for each time slot. Our study focuses on the latter aspect, as indicated by previous work by Laganga and Lawrence (2012), Zacharias and Pinedo (2014, 2017), and Zacharias and Yunes (2020).

Previous research has considered various features, such as no-show behaviour and walk-ins, within this aspect. For instance, studies by Hassin and Mendel (2008), Liu et al. (2010), Luo et al. (2012), Jiang et al. (2017), Liu (2016), and Kong et al. (2020) have addressed the issue of scheduled patients cancelling or not showing up for their appointments, which can create uncertainty for providers when making schedules. Strategies such as reducing appointment intervals (e.g., Jiang et al., 2017; Kong et al., 2020), overbooking (e.g., Laganga and Lawrence, 2012; Zacharias and Pinedo, 2014) and equal-waiting scheduling system (e.g., Zhang et al., 2022) are proposed to mitigate the negative effects of no-show behaviour. Additionally, managing services for walk-ins without appointments has also attracted the attention of researchers. Çil and Lariviere (2013) find that some reservation requests may be rejected to prioritise walk-in demand. Previous literature such as Cayirli et al. (2012), Wang et al. (2020), Pan et al. (2020), and Li et al. (2021) address the problem of managing appointments in the presence of walk-ins. Liu et al. (2023a) studies how to allocate

service capacity when patients may strategically choose to walk in or schedule an appointment. However, previous studies have not explicitly considered the trend of E-visits, which can increase uncertainty for providers who work both online and offline. Our study explicitly considers three types of patients with different arrival patterns: scheduled patients with no-show behaviour, walk-in patients who head in the clinic, and online patients who request services from E-visit platforms and have a tolerance for waiting.

A key tradeoff in the literature on appointment scheduling is the balance between maximising service utilisation and minimising patients' waiting. One common approach to address this issue is to assign different cost rates to each factor and then minimise the weighted sum of the expected cost components (e.g., Kaandorp and Koole, 2007; Laganga and Lawrence, 2012; Zacharias and Pinedo, 2014; Wang et al., 2020). It should be noted that all of these studies consider linear waiting costs. However, literature reviews by Ahmadi-Javid et al. (2017) and Cayirli and Veral (2003) suggest that future research could relax the assumption of linear waiting cost setting. For instance, the waiting cost for the short term should not be the same as that for the long term. We are the first to consider the non-linear waiting cost when modelling the unit waiting cost for E-visits.

Due to the complexity of appointment scheduling problems, mathematical techniques are often used in previous studies to find the optimal solutions, including establishing properties, and/or deriving upper (lower) bounds for the problem. These analytical results are used to develop efficient algorithms, such as those demonstrated by Huh et al. (2013), Feldman et al. (2014) and Zhou et al. (2021). In the appointment-scheduling model for a service system, Kaandorp and Koole (2007) demonstrate that their scheduling model, which considers exponentially distributed service times and patient no-show behaviour, is multimodal. Koeleman and Koole (2012) extend the model to include emergency arrivals and general service times and propose a local search algorithm to find the global optimum with the help of the multimodularity property of the objective function. Our model also uses the property of multimodularity which, is further discussed by Chen and Li (2021) and Zacharias and Yunes (2020). Our study demonstrates that multimodularity holds even when the waiting cost for E-visits is not linear, and we present a more concise proof of this property compared to previous literature.

2.2. Healthcare OM with E-visit

The implementation of E-visits in healthcare operations management has received increased attention as a means of improving patient care and operational efficiency. Previous studies have shown that E-visits can enhance access to care, improve patient satisfaction, and reduce costs for both patients and providers. For instance, Hwang et al. (2022) investigate the implications of teleconsultations on healthcare disparity, while Erdogan et al. (2018) and Liu et al. (2020) demonstrate that the implementation of E-visits is associated with reduced healthcare costs and improved patient outcomes. Rajan et al. (2019) find that introducing telemedicine increases overall social welfare. Additionally, Zhong et al. (2017) study E-visits in primary care clinics and propose scheduling policies to improve the performance of patient length of visit.

Several studies have also focused on how E-visits impact physicians' operations and evaluate the performance of E-visit implementation. Zhong (2018) and Yu and Bayram (2021) address the capacity planning problem by considering the presence of E-visits while Qin et al. (2022) optimise the intra-day sequencing rule between telemedicine and in-person patients. Zhong et al. (2018) investigate E-visits' impact on primary care delivery operations and provide guides for physicians' panel size. Bavafa et al. (2018) investigate the impact of E-visit adoption on a health system depending on the capacity of physicians and the pattern of fee-for-service. Bavafa et al. (2021) capture the usage of E-visits and nonphysician providers and quantify physicians' expected earnings and patients' expected health outcomes. Delana et al. (2022) examine the impact of introducing a telemedicine centre and find that telemedicine triggers more visits in total while reducing visits to the clinic.

In conclusion, implementing E-visits in healthcare operations management has the potential to revolutionise healthcare services by enhancing patient access to care, increasing patient satisfaction, and reducing costs. However, more research is needed to fully comprehend its impact on daily operations, and identify best practices for implementation. Our study aims to fill the gap in this area by examining the impact of E-visits on healthcare providers' daily operations and providing insights to improve their implementation and advance healthcare operations management. Two studies related to our work are Savin et al. (2021), which investigates the interactions among the online healthcare platform, physicians and patients, and Liu et al. (2023b), which develops a stylised queuing-game model while incorporating patient strategic choice between the two service channels (virtual and in-person services). Our work differs significantly from these two studies, as we concentrate on scheduling appointments and E-visit time windows to minimise costs associated with patient waiting and provider utilisation.

Recent work by Shen et al. (2023) explored the optimal appointment times for both online and offline patients in a multi-server environment, aiming to minimise the linear costs of patient waiting and provider idling and overtime. Although our work shares similarities with theirs, our model setting and solution approaches are substantially different. While they focus on provider scheduling E-visits at certain times while E-visits have no-show behaviour, we concentrate on scheduling available time windows for E-visits while considering their waiting pattern. Their model balances different costs by adjusting the providers allocated to each channel, whereas ours emphasises the trade-off between scheduled patients' known arrival times and potential no-shows, and E-visits' lower waiting costs but uncertain arrival times. In terms of methodology, they formulate the problem as a mixed integer program and use Bender Decomposition to solve it. In contrast, we analyse the mathematical properties of the problem and explore the structures of the optimal solution.

3. Basic Model

This section develops a basic optimisation model that determines the number of appointments scheduled in each time slot, while taking into account the no-show and walk-in behaviour of in-clinic patients, as well as the uncertain arrival of online E-visits. At present, we assume that an E-visit can arrive at any time. However, in Section 4, we will consider the case in which only E-visits arriving within specific time windows are accepted, and further study the optimal schedule of these time windows. Throughout, we use the lowercase (uppercase) Greek letters to denote random variables (calculated values), lowercase (uppercase) Latin letters to denote variables (Constants), and bold-faced lowercase (uppercase) letters to denote vectors (matrices). All notations are summarised in Table A.1 in Online Appendix.

3.1. Model Formulation

Consider a clinic session with a single provider. In practice, the length of a clinic session is often measured by the number of appointment slots, and patients are scheduled to arrive at the beginning of each slot. Following this convention, we consider a clinic session with $T > 0$ appointment slots, where T is a predetermined number. For example, in a clinic session from 9:00am to 12:00pm, suppose the length of an appointment slot is 15 minutes, then $T = 12$. The provider needs to schedule n patients in these slots, and n is a decision variable. We assume that the provider can schedule at most N appointments per day, which means that n should be less than or equal to N .

In addition to n , the provider also needs to decide $\mathbf{x} = (x_1, x_2, \dots, x_T)$, where x_t is the number of appointments scheduled at the beginning of slot t and $\sum_{t=1}^T x_t = n$. However, scheduled patients may not show up for their pre-booked appointments. Let $\alpha(\mathbf{x}) = (\alpha_1(x_1), \alpha_2(x_2), \dots, \alpha_T(x_T))$ denote the number of show-up patients. That is, $\alpha_t(x_t)$ is the number of show-ups given x_t , which follows some distribution known by the provider. Here, we assume that $\alpha_t(x_t)$ is independent of others, meaning that the patient's no-show behaviour is homogeneous and time-independent, which is a common assumption adopted in previous literature (see e.g., Laganga and Lawrence, 2012; Zacharias and Pinedo, 2014; Feldman et al., 2014, etc.).

As discussed in Section 1, we consider that E-visits arrive at any time without an appointment. Thus, random in-clinic walk-ins and online E-visits may arrive throughout the clinic session. For tractability, we assume that they always arrive at the beginning of each appointment slot. Let β_t denote the number of in-clinic walk-ins who arrive at slot t , and γ_t denote the number of online E-visits at t . We note that β_t and γ_t are random variables, and they follow some distributions which are known by the provider. For now, we assume that β_t and γ_t are independent of others. In Section 3.3, we will relax this assumption and allow them to be correlated. We denote the vector of β_t and γ_t as $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_T)$ and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \dots, \gamma_T)$, respectively. We assume $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ as a whole are exogenously given and independent of the provider's decision.

Following Robinson and Chen (2010), Laganga and Lawrence (2012) and Zacharias and Pinedo (2014, 2017), we assume that the service time of each patient is exactly one appointment slot. It is a reasonable assumption because, in practice, the provider can adjust the conversation content and speed to control the consultation time (see, e.g., Gupta and Denton, 2008). For tractability, we focus on the impact of the uncertain arrivals of E-visits, although, similar to Wang et al. (2020), our models and solutions can be easily extended to consider exponentially distributed service time.

The main costs considered in the literature are patient's waiting, provider's idling and overtime. A common way to address this is to assign different cost rates to them and minimise the weighted sum of the expected cost components. Without loss of generality, we normalise the unit wait time cost for scheduled patients to be 1. For walk-ins, the unit wait time cost is C_W . As for E-visits, these patients benefit from the convenience and flexibility of such a channel. Thus they can accept a certain period of time to wait for services, meaning they can wait P slots at no cost. After P slots, the unit wait time cost is C_E . Figure 1 shows an example of $P = 3$. An E-visit patient arrives at the beginning of slot t , and there are 3 patients in front of him, he will wait for 4 slots if there is no patient arriving after him. As $P = 3$, only the last slot of waiting for this patient occurs cost. In summary, the cost of waiting for E-visits can be represented as $C_E \times (\text{wait time} - P)^+$, which is not linear. It is worth noting that, in the special case of $P = 0$, the waiting cost becomes linear in the wait time with a unit cost C_E .



Figure 1: The Waiting Cost of an E-visit Patient When $P = 3$

Following Wang et al. (2020), we assume that the waiting cost of scheduled patients is higher than that of walk-ins. Because online E-visits can wait anywhere, their waiting cost should be lower than that of walk-ins who must stay in the clinic (Maister, 1985; Bavafa et al., 2018). Thus, we assume that $C_E \leq C_W \leq 1$ throughout this paper. Let C_I and C_O be the unit idle time cost and overtime cost for the provider, respectively.

Given an appointment schedule \mathbf{x} , let Γ_S and Γ_W be the expected total wait times of scheduled patients and walk-ins, respectively. Let Γ_E be the expected total wait time to be counted (i.e., the wait time after P slots) for E-visits. Let Γ_I and Γ_O be the expected idle time and overtime of the provider. Then the expected total weighted cost of the provider can be expressed as follows:

$$\Gamma_S + C_W \Gamma_W + C_E \Gamma_E + C_I \Gamma_I + C_O \Gamma_O. \quad (1)$$

To calculate the value of the objective function above, we first need to specify the service priority among scheduled patients, walk-ins and online E-visits. Since $C_E \leq C_W \leq 1$, based on $c\mu$ rule, we know that it is optimal to serve scheduled patients before walk-ins and to serve walk-ins before E-visits.

To calculate Γ_S , we evaluate $\Psi_t(k)$, which represents the probability that there are k **scheduled** patients waiting at the end of slot t . Given a schedule \mathbf{x} , let $p_t(i, x_t)$ be the probability that i of these x_t scheduled patients show up at t , that is, $p_t(i, x_t) = \Pr(\alpha_t(x_t) = i)$. Let \bar{N} be a sufficiently large number such that it suffices to consider at most \bar{N} in the system. Let \bar{T} also be a sufficiently large number such that all patients can be served before \bar{T} . Then we can write $\Psi_t(k)$ for $k = 0, \dots, \bar{N}$ and $t = 1, \dots, \bar{T}$ recursively as

$$\Psi_t(k) = \sum_{i=0}^{x_t} \Psi_{t-1}(k-i+1) p_t(i, x_t) + \begin{cases} \Psi_{t-1}(0) p_t(0, x_t), & \text{if } k = 0, \\ 0, & \text{otherwise,} \end{cases} \quad (2)$$

where $\Psi_0(0) = 1$. The first part in the right-hand side (RHS) of (2) calculates the joint probability of $k-i+1$ scheduled patients waiting at the end of slot $t-1$, i scheduled patients showing up at the beginning of slot t , and one scheduled patient being served at t . The second part in the RHS of (2) adds one more term for the case of $k = 0$, which is the joint probability of no scheduled patient waiting at the end of slot $t-1$ and no scheduled patient showing up at slot t . We can now use the expected number of scheduled patients waiting at each slot to evaluate Γ_S . Given a schedule \mathbf{x} , Γ_S can be calculated as

$$\Gamma_S = \sum_{t=1}^{\bar{T}} \sum_{k=1}^{\bar{N}} k \Psi_t(k). \quad (3)$$

To calculate Γ_W , we now evaluate $\Pi_t(k)$, which represents the probability that there are k **in-clinic** patients waiting at the end of slot t (i.e., scheduled patients plus walk-ins). Let $q_t(j)$ denote the probability of j walk-ins arriving at t , that is, $q_t(j) = \Pr(\beta_t = j)$. We can write $\Pi_t(k)$ for $k = 0, \dots, \bar{N}$ and $t = 1, \dots, \bar{T}$ recursively as

$$\begin{aligned} \Pi_t(k) = & \sum_{i=0}^{x_t} \sum_{j=0}^{k-i+1} \Pi_{t-1}(k-i-j+1) p_t(i, x_t) q_t(j) \\ & + \begin{cases} \Pi_{t-1}(0) p_t(0, x_t) q_t(0), & \text{if } k = 0, \\ 0, & \text{otherwise,} \end{cases} \end{aligned} \quad (4)$$

where $\Pi_0(0) = 1$. The first part in the RHS of (4) calculates the joint probability of $k-i-j+1$ in-clinic patients waiting at the end of slot $t-1$, i scheduled patients and j walk-ins arriving at the beginning of slot t , and one in-clinic patient being served at t . The second part adds one more term for the case of $k = 0$, which is the joint probability of no in-clinic patient waiting at the end of slot $t-1$ and no scheduled or walk-in patient arriving at slot t . We are ready to use the expected number of scheduled and walk-in patients waiting at each slot to evaluate Γ_W . Given a schedule \mathbf{x} , Γ_W can be calculated as

$$\Gamma_W = \sum_{t=1}^{\bar{T}} \sum_{k=1}^{\bar{N}} k \Pi_t(k) - \Gamma_S. \quad (5)$$

For Γ_E , since there is a threshold for calculating the waiting cost, we need to keep track of how long the E-visits have waited. To do this, we define $\Phi_t^p(k)$, which represents the probability that there are k in-clinic waiting patients plus E-visits who have waited at least p slots at the end of slot t . Note that E-visits who have waited at least p slots include E-visits who have waited at least $p+1$ slots. We also note that there is no need

to consider $p > P + 1$, because the unit cost is the same for E-visits when the waiting exceeds P slots. Let $o_t(l)$ denote the probability of l E-visits arriving at t , that is, $o_t(l) = \Pr(\gamma_t = l)$. Then we can recursively write $\Phi_t^1(k)$, which is indeed the probability of k patients (including all in-clinic and online patients) waiting at the end of slot t , for $k = 0, \dots, \bar{N}$ and $t = 1, \dots, \bar{T}$, as

$$\begin{aligned} \Phi_t^1(k) = & \sum_{i=0}^{x_t} \sum_{j=0}^{k-i+1} \sum_{l=0}^{k-i-j+1} \Phi_{t-1}^1(k-i-j-l+1) p_t(i, x_t) q_t(j) o_t(l) \\ & + \begin{cases} \Phi_{t-1}^1(0) p_t(0, x_t) q_t(0) o_t(0), & \text{if } k = 0, \\ 0, & \text{otherwise,} \end{cases} \end{aligned} \quad (6)$$

where $\Phi_0^1(0) = 1$. The first part in the RHS of (6) calculates the joint probability of $k - i - j - l + 1$ patients waiting at the end of slot t , i scheduled patients, j walk-ins and l E-visits arriving at the beginning of slot t , and one patient being served at t . The second part adds one more term for the case of $k = 0$, which is the joint probability of no patient waiting at the end of slot $t - 1$ and no new patient arriving at slot t . For $p = 2, \dots, P + 1$, the calculation for $\Phi_t^p(k)$ is simpler, as the E-visits who have waited at least p slots at t are exactly the E-visits who have waited at least $p - 1$ slots at $t - 1$. Then we can recursively write $\Phi_t^p(k)$ for $k = 0, \dots, \bar{N}$ and $t = 1, \dots, \bar{T}$ as

$$\begin{aligned} \Phi_t^p(k) = & \sum_{i=0}^{x_t} \sum_{j=0}^{k-i+1} \Phi_{t-1}^{p-1}(k-i-j+1) p_t(i, x_t) q_t(j) \\ & + \begin{cases} \Phi_{t-1}^{p-1}(0) p_t(0, x_t) q_t(0), & \text{if } k = 0, \\ 0, & \text{otherwise,} \end{cases} \end{aligned} \quad (7)$$

where $\Phi_t^1(k)$ is defined in (6) and $\Phi_0^p(0) = 1$. The first part in the RHS of (7) calculates the joint probability of $k - i - j + 1$ in-clinic waiting patients plus E-visits who have waited at least $p - 1$ slots at the end of slot t , i scheduled patients and j walk-ins arriving at the beginning of slot t , and one patient being served at t . The second part adds one more term for the case of $k = 0$, which is the joint probability of no in-clinic patient or E-visit who has waited at least $p - 1$ slots at the end of slot $t - 1$ and no in-clinic patient arriving at slot t . Then we can use the expected number of E-visits who have waited at least $P + 1$ slots to evaluate Γ_E , i.e.,

$$\Gamma_E = \sum_{t=1}^{\bar{T}} \sum_{k=1}^{\bar{N}} k \Phi_t^{P+1}(k) - \Gamma_S - \Gamma_W. \quad (8)$$

We can also use the expected number of all patients waiting at the end of the regular session to evaluate Γ_I and Γ_O :

$$\Gamma_I = T + \sum_{k=1}^{\bar{N}} k \Phi_T^1(k) - \mathbb{E} \left[\sum_{t=1}^T (\alpha_t(x_t) + \beta_t + \gamma_t) \right], \quad (9)$$

$$\Gamma_O = \sum_{k=1}^{\bar{N}} k \Phi_T^1(k). \quad (10)$$

Thus, our optimisation model can be formulated as,

$$\min_{x \in \mathbb{Z}^+} \Gamma_S + C_W \Gamma_W + C_E \Gamma_E + C_I \Gamma_I + C_O \Gamma_O \quad (P1)$$

$$\text{s.t. } \sum_{t=1}^T x_t \leq N$$

$\Gamma_S, \Gamma_W, \Gamma_E, \Gamma_I$ and Γ_O are defined in (3), (5), (8), (9) and (10), respectively.

3.2. The Multimodularity

Our main result is given in Proposition 1, which suggests that Problem (P1) has the property of multimodularity. This property enables us to use local search to solve this combinatorial optimisation problem. Multimodularity has been shown previously in Kaandorp and Koole (2007) and Wang et al. (2020), and we are the first to extend the result to the context incorporating E-visits whose waiting cost is not linear. Additionally, while Kaandorp and Koole (2007) and Wang et al. (2020) prove multimodularity by considering all sample paths, we utilise a more concise way to prove it, which is more elegant. The concept of multimodularity, which was first introduced by Hajek (1985), has been a helpful tool in the study of queuing systems and other operations management problems (see, e.g., Murota, 2005; Chen and Li, 2021, et al.). One of the most significant features of multimodularity is that it ensures the local optimum is the global optimum and the problem can be solved by polynomial-time algorithms.

Definition 1 (Hajek, 1985). Define vectors $\mathbf{u}_0, \mathbf{u}_1, \dots, \mathbf{u}_T$ in \mathbb{Z}_+^T by

$$\begin{Bmatrix} \mathbf{u}_0 \\ \mathbf{u}_1 \\ \mathbf{u}_2 \\ \dots \\ \mathbf{u}_{T-1} \\ \mathbf{u}_T \end{Bmatrix} = \begin{Bmatrix} (-1, 0, 0, \dots, 0, 0) \\ (1, -1, 0, \dots, 0, 0) \\ (0, 1, -1, \dots, 0, 0) \\ \dots \\ (0, 0, 0, \dots, 1, -1) \\ (0, 0, 0, \dots, 0, 1) \end{Bmatrix}, \quad (11)$$

and let \mathcal{U} be the set $\{\mathbf{u}_0, \mathbf{u}_1, \dots, \mathbf{u}_T\}$. We say that a function f on \mathbb{Z}_+^T is multimodular if for all \mathbf{x} in \mathbb{Z}_+^T ,

$$f(\mathbf{x} + \mathbf{u}_i) - f(\mathbf{x}) \geq f(\mathbf{x} + \mathbf{u}_j + \mathbf{u}_i) - f(\mathbf{x} + \mathbf{u}_j) \quad (12)$$

when $\mathbf{u}_i, \mathbf{u}_j \in \mathcal{U}$, $\mathbf{x} + \mathbf{u}_i, \mathbf{x} + \mathbf{u}_j, \mathbf{x} + \mathbf{u}_i + \mathbf{u}_j \in \mathbb{Z}_+^T$ and $\mathbf{u}_i \neq \mathbf{u}_j$.

The next Proposition suggests that our optimisation problem (P1) has such a property.

Proposition 1. $\Gamma_S + C_W \Gamma_W + C_E \Gamma_E + C_I \Gamma_I + C_O \Gamma_O$ is multimodular in $\mathbf{x} \in \mathbb{Z}_+^T$, where $\Gamma_S, \Gamma_W, \Gamma_E, \Gamma_I$ and Γ_O are defined in (3), (5), (8), (9) and (10), respectively.

Previous literature (e.g., Kaandorp and Koole, 2007; Wang et al., 2020) has shown that, in appointment scheduling, the objective function is multimodular with some specific features (such as no-show behaviour and walk-ins). It should be noted that all of these studies consider linear waiting costs (see Section 2 for a detailed discussion). We complement and advance this literature by using a more concise way to show that this elegant property also holds with nonlinear waiting costs for E-visits.

According to Definition 1, a multimodular function has the property of an increasing difference, which makes minimising such functions tractable. Murota (2005) has shown that for a multimodular function, a local minimum on its domain is also a global minimum. Therefore, we can design a local search algorithm to find the optimal solution. To do that, we need to define the neighbourhood of a schedule.

Definition 2 (Neighbourhood of schedule \mathbf{x}). We say $\tilde{\mathbf{x}}$ is a feasible neighbour of \mathbf{x} if $\mathbf{x} \in \mathbb{Z}_+^T$ and $\tilde{\mathbf{x}} = \mathbf{x} + \sum_{\mathbf{u} \in \mathcal{U}'} \mathbf{u}$ such that $\mathbf{x} + \sum_{\mathbf{u} \in \mathcal{U}'} \mathbf{u} \geq 0$ for some $\mathcal{U}' \subsetneq \mathcal{U}$, where \mathcal{U} is defined in Definition 1.

Since \mathcal{U} is a non-empty strict subset of \mathcal{U} , $\sum_{u \in \mathcal{U}} u$ represents the possible combinations of a single slot shift from the current schedule \mathbf{x} . Wang et al. (2020) have shown that, with this well-chosen neighbourhood, the local search algorithm would terminate at a global optimum for a problem with multimodularity.

Corollary 1. *If \mathbf{x} is a feasible schedule for (P1) and the objective function value of it is not greater than that of any feasible neighbour schedule (as defined in Definition 2), then \mathbf{x} is a globally optimal schedule for (P1).*

However, even though we can use local search to find the optimal schedule, the computation is still time-consuming. Given an initial feasible schedule \mathbf{x} , we need to use recursive functions (2-7) to calculate the distributions of Ψ , Π , and Φ and then derive the expected total cost of schedule \mathbf{x} . Additionally, a feasible solution has at most $2^{T+1} - 1$ neighbours. Thus, using local search still suffers from inefficiency. In the next section, we will reformulate this problem as a two-stage program that can be solved by more efficient algorithms.

3.3. Two-stage Programming Reformulation

Following Wang et al. (2020), we reformulate the problem as a two-stage program, in which the first stage problem is a binary linear program and the second stage is a linear program. This reformulation is based on sample path representation. We use Ω to denote the set of all possible sample scenarios for α , β and γ . Let $\omega \in \Omega$ be an arbitrary sample scenario, and let $\alpha(\omega)$, $\beta(\omega)$, and $\gamma(\omega)$ be the vectors of show-up patients, walk-ins and E-visits under scenario ω , respectively. When solving the problem practically, we adopt the commonly-used Sample Average Approximation (SAA) approach to randomly generate a sufficient number of samples to represent Ω and then to minimise the average cost of these samples. With some abuse of notation, we utilise \mathbb{E}_ω to denote computing the average over these samples.

To record the state of the system at the end of each slot, we define a series of auxiliary variables. Let y_t^S denote the number of scheduled patients waiting at the end of slot t , and let y_t^I denote the number of in-clinic patients (i.e., scheduled patients plus walk-ins) waiting at the end of slot t . Then we have $y_t^S(\omega)$ and $y_t^I(\omega)$ under scenario ω with the following Lindley equations (Asmussen, 2003),

$$y_t^S(\omega) = \begin{cases} (y_{t-1}^S(\omega) + \alpha_t(x_t|\omega) - 1)^+, & \forall 1 \leq t \leq T, \\ (y_{t-1}^S(\omega) - 1)^+, & \forall T < t \leq \bar{T}, \end{cases} \quad (13)$$

and

$$y_t^I(\omega) = \begin{cases} (y_{t-1}^I(\omega) + \alpha_t(x_t|\omega) + \beta_t(\omega) - 1)^+, & \forall 1 \leq t \leq T, \\ (y_{t-1}^I(\omega) - 1)^+, & \forall T < t \leq \bar{T}, \end{cases} \quad (14)$$

where $y_0^S(\omega) = 0$ and $y_0^I(\omega) = 0$. The first case of (13) and (14) pertains to the clinic session when new patients arrive and one patient would be served, if any. The second case of (13) and (14) is for overtime when new patients are not allowed.

We next define y_t^p as the number of in-clinic waiting patients and E-visits who have waited at least p slots at the end of slot t (i.e., all in-clinic patients plus E-visits who arrived p slots ago), for $p = 1, 2, \dots, P + 1$. Note that y_t^1 denotes all waiting patients at the end of t ; and y_t^{P+1} denotes all waiting patients who incur waiting costs at the end of t . Then we have

$$y_t^1(\omega) = \begin{cases} (y_{t-1}^1(\omega) + \alpha_t(x_t|\omega) + \beta_t(\omega) + \gamma_t(\omega) - 1)^+, & \forall 1 \leq t \leq T, \\ (y_{t-1}^1(\omega) - 1)^+, & \forall T < t \leq \bar{T}, \end{cases} \quad (15)$$

where $y_0^1(\omega) = 0$ for $p = 1$; and for $p = 2, \dots, P + 1$

$$y_t^p(\omega) = \begin{cases} (y_{t-1}^{p-1}(\omega) + \alpha_t(x_t|\omega) + \beta_t(\omega) - 1)^+, & \forall 1 \leq t \leq T, \\ (y_{t-1}^{p-1}(\omega) - 1)^+, & \forall T < t \leq \bar{T}, \end{cases} \quad (16)$$

where where $y_0^p(\omega) = 0$.

Based on y_t^S , y_t^I and y_t^p , we can now derive the cost components as follows. The expected wait time of scheduled patients is equal to the sum of the expected number of scheduled patients waiting at the end of each t , i.e., the sum of y_t^S :

$$\Gamma_S = \mathbb{E}_\omega \left[\sum_{t=1}^{\bar{T}} y_t^S(\omega) \right]. \quad (17)$$

Similarly, the expected wait time of walk-ins equals the sum of the expected number of walk-ins waiting at the end of each t , i.e., the sum of $y_t^I - y_t^S$:

$$\Gamma_W = \mathbb{E}_\omega \left[\sum_{t=1}^{\bar{T}} \left(y_t^I(\omega) - y_t^S(\omega) \right) \right]. \quad (18)$$

The expected wait time to be counted for E-visits is the sum of the expected number of E-visits who have waited at least $P + 1$ slots at the end of each t , i.e., the sum of $y_t^{P+1} - y_t^I$:

$$\Gamma_E = \mathbb{E}_\omega \left[\sum_{t=1}^{\bar{T}} \left(y_t^{P+1}(\omega) - y_t^I(\omega) \right) \right]. \quad (19)$$

The expected over time equals the expected number of patients waiting at the end of a regular clinic session, i.e., y_T^1 :

$$\Gamma_O = \mathbb{E}_\omega \left[y_T^1(\omega) \right]. \quad (20)$$

The expected idle time is the expected off time of the session, i.e., regular session length plus overtime, minus the expected working time, i.e., the expected number of served patients:

$$\Gamma_I = T + \mathbb{E}_\omega \left[\left(y_T^1(\omega) - \sum_{t=1}^T \left(\alpha_t(x_t|\omega) + \beta_t(\omega) + \gamma_t(\omega) \right) \right) \right]. \quad (21)$$

The mean of $\sum_{t=1}^T (\beta_t(\omega) + \gamma_t(\omega))$ is always a constant given by the distributions of walk-ins and E-visits. Thus, we can simplify the weighted sum of the above cost components as the following objective function,

$$\mathbb{E}_\omega \left[\sum_{t=1}^{\bar{T}} \left((1 - C_W) y_t^S(\omega) + (C_W - C_E) y_t^I(\omega) + C_E y_t^{P+1}(\omega) \right) + (C_I + C_O) y_T^1(\omega) - C_I \sum_{t=1}^T \alpha_t(x_t|\omega) \right]. \quad (22)$$

Now our problem is to minimise (22) subject to the Lindley equations (13-16). The challenge we face in this optimisation problem arises from the non-linearity, which is present in both the operator $(x)^+$ for auxiliary variables and the function $\alpha_t(x_t|\omega)$ itself within each constraint. To make this optimisation problem tractable, we first employ a common trick for the auxiliary variables, which is relaxing $y = (x)^+$ to two

essential constraints: $y \geq x$ and $y \geq 0$. It is important to note that this relaxation is exact for our problem because the coefficients of these auxiliary variables within the objective function (as defined in equation (22)) are invariably positive. As for $\alpha_t(x_t|\omega)$, we adopt the reformulation approach used in Wang et al. (2020). Instead of directly optimising the variable x_t , we introduce a new set of decision variables $z_{i,t}$ for $i = 1, 2, \dots, N$, where N is the maximum number of patients to be scheduled, to replace it. Specifically, we use binary variable $z_{i,t}$ to denote whether patient i is scheduled to time slot t . Then x_t can be expressed as the sum of these binary variables, i.e., $x_t = \sum_{i=1}^N z_{i,t}$ by definition. Note that for any \mathbf{x} we can obtain a corresponding \mathbf{z} and for any \mathbf{z} we can obtain a corresponding \mathbf{x} , thus, the local search on \mathbf{x} can be modified to be applied on \mathbf{z} . Additionally, we use binary variable $\delta_i(\omega)$ to denote whether patient i shows up or not under scenario ω . With these reformulations in place, we can express $\alpha_t(x_t|\omega)$ as a linear combination of $\delta_i(\omega)$ and $z_{i,t}$, i.e., $\alpha(x_t|\omega) = \sum_{i=1}^N \delta_i(\omega)z_{i,t}$. The reformulations for $\alpha_t(x_t|\omega)$, together with the relaxations for auxiliary variables, linearise our optimisation problem, allowing us to leverage standard optimisation techniques and commercial solvers for finding optimal solutions.

Since the decision variable x_t is changed to $z_{i,t}$, we have new constraints $\mathbf{z} \in \mathbb{Z}^+$ and $\sum_{t=1}^T z_{i,t} \leq 1, \forall i$, requiring that one patient can be scheduled to at most one slot. By replacing $\alpha(x_t|\omega)$ with $\sum_{i=1}^N \delta_i(\omega)z_{i,t}$ and relaxing the $(x)^+$ operator in the Lindley equations in (13-16), we can reformulate the original problem (P1) as a two-stage programming problem (P2),

$$\begin{aligned} \min_{\mathbf{z} \in \mathbb{Z}^+} \quad & \mathbb{E}_\omega \left[\Upsilon(\mathbf{z}, \omega) - C_I \sum_{t=1}^T \sum_{i=1}^N \delta_i(\omega)z_{i,t} \right] \\ \text{s.t.} \quad & \sum_{t=1}^T z_{i,t} \leq 1, \quad \forall i \end{aligned} \quad (\text{P2})$$

where $\Upsilon(\mathbf{z}, \omega) =$

$$\min_{y \geq 0} \sum_{t=1}^{\bar{T}} \left((1 - C_W)y_t^S + (C_W - C_E)y_t^I + C_E y_t^{P+1} \right) + (C_I + C_O)y_T^1 \quad (\text{Sub.LP})$$

$$\text{s.t.} \quad y_t^S \geq y_{t-1}^S + \sum_{i=1}^N \delta_i(\omega)z_{i,t} - 1, \quad \forall t \quad (23)$$

$$y_t^I \geq y_{t-1}^I + \sum_{i=1}^N \delta_i(\omega)z_{i,t} + \beta_t(\omega) - 1, \quad \forall t \quad (24)$$

$$y_t^1 \geq y_{t-1}^1 + \sum_{i=1}^N \delta_i(\omega)z_{i,t} + \beta_t(\omega) + \gamma_t(\omega) - 1, \quad \forall t \quad (25)$$

$$y_t^p \geq y_{t-1}^{p-1} + \sum_{i=1}^N \delta_i(\omega)z_{i,t} + \beta_t(\omega) - 1, \quad \forall t \text{ and } p \geq 2 \quad (26)$$

Proposition 2. (P2) is equivalent to (P1).

Remark 1 (For details in (Sub.LP)).

1. (23-26) are relaxed from (13-16).
2. $\sum_{i=1}^N \delta_i(\omega)z_{i,t}$, $\beta_t(\omega)$ and $\gamma_t(\omega)$ can be regarded as 0 for $t > T$.

3. There should be integer constraints for the decision variables y_t^S , y_t^I and y_t^p , which represent the number of waiting patients. One can easily check that the coefficient matrix of decision variables is totally unimodular. That is, in each constraint, the coefficient of the decision variable is either 1 or -1, and there are at most two decision variables. Thus the integer constraints can be ignored and the second stage problem becomes a linear program.

Remark 2 (For solving (P2)). In (P2), the second stage problem (Sub.LP) solves a linear program to obtain $Y(\mathbf{z}, \omega)$ for the given \mathbf{z} and ω ; the first stage problem is to find an optimal \mathbf{z} that minimises $\mathbb{E}_\omega[Y(\mathbf{z}, \omega) - C_I \sum_{t=1}^T \sum_{i=1}^N \delta_i(\omega) z_{i,t}]$. Such a two-stage programming problem can be solved efficiently.

1. Both stages are minimisation problems. By introducing auxiliary variables $\mathbf{y}(\omega)$ for each scenario, the second stage problem can be integrated into the first stage, resulting in the entire problem becoming an MBILP (mixed binary integer linear program). Then, widely used commercial solvers like Gurobi and Cplex can be employed to find the optimal schedule.
2. The objective function is shown to be multimodular in \mathbf{x} (see Proposition 1). For any \mathbf{z} , we have \mathbf{x} by $x_t = \sum_{i=1}^N z_{i,t}$; and for any \mathbf{x} , it is easy to generate a corresponding \mathbf{z} . Thus we can use local search in the first stage to search for the optimal schedule.
3. The second stage problem is an LP. Following Wang et al. (2020), we are able to design an efficient algorithm which is called Cut Generation Algorithm (CGA), that can cut the sub-optimal solutions and bound the objective value faster by deriving the dual problem for (Sub.LP). More details can be found in Appendix B.1.

Remark 3 (General arrival patterns of walk-ins and E-visits). In Section 3.1, we assume that the arrival patterns of walk-ins and E-visits are independent and uncorrelated when deriving the recursive functions. However, this assumption can be relaxed by using the sample path representation. By considering the distribution of all walk-ins and E-visits as a single entity, all results, including multimodularity, still hold as long as the entire distribution is given exogenously.

4. Model Incorporating E-visit Time Windows

In this section, we will consider the case where only E-visits arriving within the available time windows are accepted and analyse how to optimally schedule the time windows. In addition to deciding the appointment schedule \mathbf{x} (or equivalently \mathbf{z}), we also decide the time window schedule $\mathbf{s} = (s_1, s_2, \dots, s_T)$ where s_t is a binary variable. If $s_t = 1$, slot t is available for E-visits; otherwise, slot t is not available for E-visits. With the time window schedule \mathbf{s} , the number of E-visits arriving in slot t is $\gamma_t s_t$ ⁴.

We will first formulate the model with recursive functions. Since $\alpha(\mathbf{x})$ which represents the vector of show-up patients, and β , which represents the vector of walk-ins, are the same as those in Section 3, the calculations for $\Psi_t(k)$ (the probability of k scheduled patients waiting at the end of slot t), $\Pi_t(k)$ (the probability of k in-clinic patients waiting at the end of slot t), and $\Phi_t^p(k)$ (the probability of k in-clinic waiting patients plus E-visits who have waited at least p slots at the end of t) for $p \geq 2$ are exactly the same as (2), (4) and (7).

The calculation for $\Phi_t^1(k)$, the probability of k patients waiting at the end of slot t , is influenced by the E-visit time window schedule \mathbf{s} . Let $o_t(l, s_t)$ denote the probability of l E-visits arriving at slot t , that is,

$$o_t(l, s_t = 1) = \Pr(\gamma_t = l);$$

⁴We assume that the arrival patterns of E-visits are exogenously given throughout the paper, so the distribution of γ_t will not be influenced by the decision variable \mathbf{s} . For cases where E-visits may be strategically shifted to other channels due to healthcare provider decision changes, we leave it for future investigation.

$$o_t(l, s_t = 0) = \begin{cases} 1, & \text{if } l = 0, \\ 0, & \text{otherwise.} \end{cases} \quad (27)$$

Following (6), with a slight abuse of notation, we can write $\Phi_t^1(k)$ for $k = 0, \dots, \bar{N}$ and $t = 1, \dots, \bar{T}$ recursively as

$$\begin{aligned} \Phi_t^1(k) = & \sum_{i=0}^{x_t} \sum_{j=0}^{k-i+1} \sum_{l=0}^{k-i-j+1} \Phi_{t-1}^1(k-i-j-l+1) p_t(i, x_t) q_t(j) o_t(l, s_t) \\ & + \begin{cases} \Phi_{t-1}^1(0) p_t(0, x_t) q_t(0) o_t(0, s_t), & \text{if } k = 0, \\ 0, & \text{otherwise,} \end{cases} \end{aligned} \quad (28)$$

where $\Phi_0^1(0) = 1$.

For the cost components, we note that the calculations for $\Gamma_S, \Gamma_W, \Gamma_E$ and Γ_O are unchanged and can directly follow (3), (5), (8) and (10), with the redefined Φ_t^1 in (28). However Γ_I will be different because only E-visits that arrive within the time windows will be served. Then we have

$$\Gamma_I = T + \sum_{k=1}^{\bar{N}} k \Phi_T^1(k) - \mathbb{E} \left[\sum_{t=1}^T (\alpha_t(x_t) + \beta_t + \gamma_t s_t) \right], \quad (29)$$

where Φ_t^1 is defined in (28).

Now, the model incorporating scheduling E-visit time windows can be formulated as follows,

$$\begin{aligned} \min_{\mathbf{x}, \mathbf{s} \in \mathbb{Z}^+} \quad & \Gamma_S + C_W \Gamma_W + C_E \Gamma_E + C_I \Gamma_I + C_O \Gamma_O \\ \text{s.t.} \quad & \sum_{t=1}^T x_t \leq N \\ & s_t \leq 1, \quad \forall t \end{aligned} \quad (P3)$$

$\Gamma_S, \Gamma_W, \Gamma_E, \Gamma_I$ and Γ_O are defined in (3), (5), (8), (29) and (10), respectively.

Incorporating a schedule of E-visit time windows can benefit both patients and the provider in several ways. Firstly, it can reduce patient wait times since access to E-visits is limited to specific time windows. Additionally, this can help to shorten providers' overtime. One potential challenge with appointment scheduling is the fact that some patients may not show up, even though their arrival time is known in advance. In contrast, the arrival times of E-visits are uncertain, but the cost of waiting for them is generally lower. By coordinating the schedules of both types of visits, the provider can better balance the waiting costs of scheduled patients and E-visits. Overall, the decision to schedule E-visit time windows can help to further improve patient satisfaction while also optimising resource utilisation for the provider.

4.1. Structure of the Optimal Joint Schedule of Appointments and E-visit Time Windows

After incorporating scheduling E-visit time windows into the model, the objective function of (P3) does not inherit the elegant property of multimodularity in (\mathbf{x}, \mathbf{s}) as a whole. Although we can still enumerate \mathbf{s} and solve \mathbf{x} for the given \mathbf{s} using the solution approaches developed in Section 3, it is not efficient as we need to enumerate 2^T times in the search for the optimal \mathbf{s} . To speed up the search, we explore some structures of it.

Let $Y(\mathbf{x}, \mathbf{s})$ denote the objective function value given the joint schedule (\mathbf{x}, \mathbf{s}) . Let \mathbf{e}_t denote the t -th T -dimensional unit vector, i.e., $\mathbf{e}_t = \{0, \dots, 1, \dots, 0\}$ where the t -th coordinate is 1. Then $\mathbf{x} + j \times \mathbf{e}_t$ means scheduling j more appointments at slot t . $\mathbf{s} \vee \mathbf{e}_t$ means opening slot t for E-visits, and $(\mathbf{s} - \mathbf{e}_t)^+$ means closing slot t for E-visits.

Proposition 3. *In the optimal joint schedule (\mathbf{x}, \mathbf{s}) , we must have*

$$\Upsilon(\mathbf{x} + j \times \mathbf{e}_{t-}, \mathbf{s} \vee \mathbf{e}_t) - \Upsilon(\mathbf{x}, \mathbf{s} \vee \mathbf{e}_t) \geq \Upsilon(\mathbf{x} + j \times \mathbf{e}_{t-}, (\mathbf{s} - \mathbf{e}_t)^+) - \Upsilon(\mathbf{x}, (\mathbf{s} - \mathbf{e}_t)^+),$$

for $\forall t \in [1, T]$, $\forall j \geq 0$ and $\forall t^- \in [1, t]$.

Proposition 3 states that the cost gap resulting from scheduling more patients at slot t or some slot before t when slot t is open for E-visits is larger than when slot t is closed for E-visits. This indicates that if scheduling more appointments at slot t or some slot before t is worse when slot t is closed for E-visits, then it must also be worse when slot t is open for E-visits. Likewise, if scheduling more appointments at slot t or some slot before t is better when slot t is open for E-visits, then it must also be better when slot t is closed for E-visits.

To visually illustrate Proposition 3, Figure 2 shows an example. The bins represent the number of appointments scheduled at each slot, and the circle (cross) symbol means that the slot is open (closed) for E-visits. The first schedule (a) has one more patient at slot 3 and opens slot 5 for E-visits; the second schedule (b) opens slot 5 for E-visits; the third schedule (c) has one more patient at slot 3 and closes slot 5 for E-visits; the last schedule (d) closes slot 5 for E-visits. According to Proposition 3, we will know the cost difference between schedule (a) and (b) is higher than that between schedule (c) and (d). And if schedule (c) is worse than schedule (d) then schedule (a) must be worse than schedule (b); if schedule (a) is better than schedule (b) then schedule (c) must be better than schedule (b).

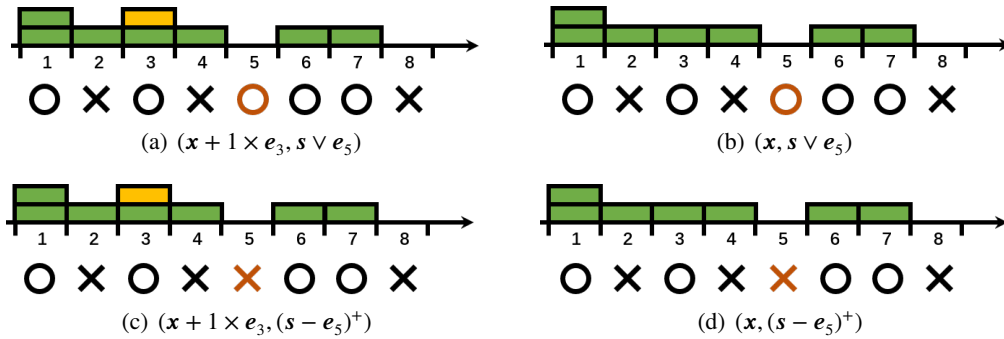


Figure 2: Visual illustration for Proposition 3

We note that the structure in Proposition 3 can be further extended to a general form.

Corollary 2. *Let \mathcal{T}^- and \mathcal{T}^+ denote two subsets of $[1, T]$, such that $\mathcal{T}^- \subseteq [1, t]$ and $\mathcal{T}^+ \subseteq [t, T]$ for some $t \in [1, T]$. Let j_t denote a non-negative integer. In the optimal joint schedule (\mathbf{x}, \mathbf{s}) , we must have*

$$f\left(\mathbf{x} + \sum_{t \in \mathcal{T}^-} (j_t \times \mathbf{e}_t), \mathbf{s} \vee \sum_{t \in \mathcal{T}^+} \mathbf{e}_t\right) - f\left(\mathbf{x}, \mathbf{s} \vee \sum_{t \in \mathcal{T}^+} \mathbf{e}_t\right) \geq f\left(\mathbf{x} + \sum_{t \in \mathcal{T}^-} (j_t \times \mathbf{e}_t), (\mathbf{s} - \sum_{t \in \mathcal{T}^+} \mathbf{e}_t)^+\right) - f\left(\mathbf{x}, (\mathbf{s} - \sum_{t \in \mathcal{T}^+} \mathbf{e}_t)^+\right),$$

for $\forall j_t, \forall \mathcal{T}^-$ and $\forall \mathcal{T}^+$.

Corollary 2 states that the cost gap resulting from scheduling more patients at or before slot t when some slots at or after t are open for E-visits is larger than when these slots are closed for E-visits. This indicates that if scheduling more appointments at or before slot t is worse when some slots at or after t are closed for E-visits, then doing so must be worse when these slots are open for E-visits. Likewise, if scheduling more

appointments at or before slot t is better when some slots at or after t are open for E-visits, then doing so must be better when these slots are closed for E-visits.

Figure 3 shows an example. Schedule (a) and schedule (c) have two more appointments at slot 2 and one more appointment at slot 3, compared to schedule (c) and (d); schedule (a) and schedule (b) open slots 4, 5 and 7 for E-visits while schedule (c) and schedule (d) close these slots. According to Corollary 2, we will know the cost difference between schedule (a) and (b) is higher than that between schedule (c) and (d). And if schedule (c) is worse than schedule (d) then schedule (a) must be worse than schedule (b); if schedule (a) is better than schedule (b) then schedule (c) must be better than schedule (d).

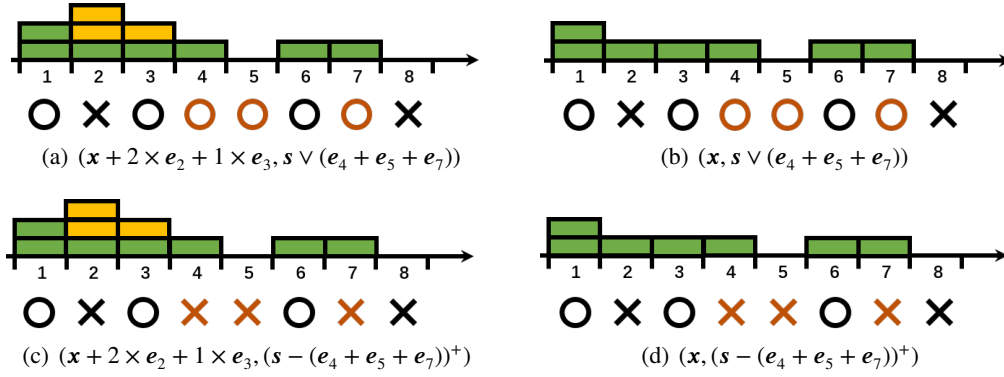


Figure 3: Visual illustration for Corollary 2

With Proposition 3 and Corollary 2, one can further speed up the search for the optimal joint schedule of appointments and E-visit time windows. However, this approach is still not efficient enough when using recursive functions. By leveraging these structures together with the property of multimodularity in \mathbf{x} for a given \mathbf{s} , we are able to design an efficient algorithm to obtain the optimal joint schedule.

4.2. Solution Approaches

Following Section 3.3, we can reformulate (P3) into a two-stage programming model with sample average approximation by replacing \mathbf{x} with \mathbf{z} . The reformulated model shares a similar structure to (P2), with the main distinction being the use of $\gamma_t(\omega)s_t$ instead of $\gamma_t(\omega)$ in (P4). It is important to note that $\mathbb{E}[\gamma_t(\omega)s_t]$ is dependent on the decision variables and is no longer a constant. Therefore, it should be included in the objective function.

Corollary 3. *The following two-stage programming model (P4) is equivalent to (P3).*

$$\begin{aligned} \min_{\mathbf{z}, \mathbf{s} \in \mathbb{Z}^+} \quad & \mathbb{E}_\omega \left[\Upsilon(\mathbf{z}, \mathbf{s}, \omega) - C_I \sum_{t=1}^T \left(\sum_{i=1}^N \delta_i(\omega) z_{i,t} + \gamma_t(\omega) s_t \right) \right] \\ \text{s.t.} \quad & \sum_{t=1}^T z_{i,t} \leq 1, \quad \forall i \\ & s_t \leq 1, \quad \forall t \end{aligned} \quad (\text{P4})$$

where $\Upsilon(\mathbf{z}, \mathbf{s}, \omega) =$

$$\min_{y \geq 0} \sum_{t=1}^{\bar{T}} \left((1 - C_W) y_t^S + (C_W - C_E) y_t^I + C_E y_t^{P+1} \right) + (C_I + C_O) y_T^1 \quad (\text{Sub.LP.2})$$

$$s.t. \ y_t^S \geq y_{t-1}^S + \sum_{i=1}^N \delta_i(\omega) z_{i,t} - 1, \quad \forall t \quad (30)$$

$$y_t^I \geq y_{t-1}^I + \sum_{i=1}^N \delta_i(\omega) z_{i,t} + \beta_t(\omega) - 1, \quad \forall t \quad (31)$$

$$y_t^1 \geq y_{t-1}^1 + \sum_{i=1}^N \delta_i(\omega) z_{i,t} + \beta_t(\omega) + \gamma_t(\omega) s_t - 1, \quad \forall t \quad (32)$$

$$y_t^p \geq y_{t-1}^{p-1} + \sum_{i=1}^N \delta_i(\omega) z_{i,t} + \beta_t(\omega) - 1, \quad \forall t \text{ and } p \geq 2 \quad (33)$$

Similar to the basic model, both stages in (P4) are minimisation problems. The second stage problem can be embedded into the first stage by adding $\mathbf{y}(\omega)$ for each scenario, transforming the entire problem into a mixed binary integer linear program (MBILP). This allows the use of commercial solvers like Gurobi and Cplex to find the optimal joint schedule.

Moreover, in (P4), only the first stage problem involves binary integer decision variables, while the second stage problem does not. Therefore, we can leverage the dual of the second stage problem (Sub.LP.2) to design an efficient algorithm that gradually generates valid cuts to eliminate non-optimal schedules. By utilising the property of multimodularity on \mathbf{x} , we can employ a local search procedure to find the optimal appointment schedule within a given E-visit time window schedule. Additionally, the structural properties of the optimal joint schedule presented in Proposition 3 and Corollary 2 can be used to efficiently eliminate non-optimal schedules. These elegant properties of the optimal solution expedite the search for the optimal joint schedule. Incorporating the structural properties into the cut generation procedure, we have designed an efficient algorithm called the Accelerated Cut Generation Algorithm (ACGA) (Algorithm 2 in Online Appendix) to solve (P4). Appendix B.2 provides a detailed explanation of the analysis and design of this algorithm, as well as its pseudo codes.

In our numerical study, we have found that the proposed Accelerated Cut Generation Algorithm (ACGA) is significantly more powerful and efficient compared to directly solving the mixed binary integer linear program (MBILP).

5. Numerical Studies

Our numerical study aims to explore several aspects related to the scheduling of in-clinic appointments and E-visit time windows. Firstly, we compare the computational performance of different solution approaches, including pure local search, MILP solvers, and proposed algorithms. Next, we investigate the impact of factors such as E-visits arrival rate, arrival pattern and E-visit waiting cost on optimal appointment scheduling. Thirdly, we analyse the improvement achieved by adopting the “proactive” control, which involves scheduling appointments and E-visit time windows together, compared to the “passive” control, which only involves scheduling appointments and opening all E-visit time windows. Then, we examine how the patterns of optimal joint schedules change with various parameter settings. Lastly, we assess the impact of the E-visit channel on the entire system by assuming that E-visits will go to other channels in the absence of an E-visit channel, and walk-ins will become E-visits with an E-visit channel. Our goal is to provide outpatient care management with more relevant insights on how to enhance healthcare resource utilisation, reduce patient wait time, and improve satisfaction for both patients and providers.

In our numerical studies, we utilise a set of model parameters to account for various practice scenarios. To ensure clarity, we set the length of daily clinic sessions at $T = 10$ and the maximum number of patients to schedule at $N = 15$. In practice, the no-show probability of scheduled patients ranges from 3% to 60%

(Hoppen, 2018), walk-ins account for 10% ~ 60% (Wang et al., 2020) of the total patients, and E-visits range from 23.2% to 71.6% and continues to increase (OECD, 2023). Following the parameter settings in previous literature (see, e.g., Robinson and Chen, 2010; Laganga and Lawrence, 2012; Zacharias and Pinedo, 2014; Wang et al., 2020, et al.), we set the no-show probability of scheduled patients to be 10% or 50%. We normalise C_S (unit waiting cost for scheduled patients) to be 1, set C_W (unit waiting cost for walk-ins) to be 0.95, set C_I (unit idle time cost) to be 4, and set C_O/C_I (the cost ratio between idle time and overtime costs) to be 0.5, 1.0 or 2.0. We set C_E (unit waiting cost for online patients) to be 0.45 or 0.9 which is no more than C_W . This is reasonable since online patients have more flexible time and could do other things while waiting for the provider to respond to their requests (see., Maister 1985). We set the waiting patience P of E-visits to be 0, 2 or 4, during which period there is no waiting cost to be incurred. The average arrival rate of walk-ins per slot is set to be $\lambda_W = 0.1$, and the number of walk-ins in each time slot follows a Poisson distribution. To investigate the impact of arrival patterns of E-visits, we examine three patterns of E-visit arrivals: (1) “Dome” shaped, in which the arrival rate of E-visits initially rises, then decreases over time; (2) “Flat” shaped, in which the arrival rate remains stable over time; and (3) “Bowl” shaped, in which the expected number of the arrival rate declines first and subsequently increases over time. The average arrival rate of E-visits in each slot is set to be 0.2, 0.5 or 0.8. Table 1 summarises the parameter settings.

5.1. Computational Performance of Different Solution Approaches

The objective function of the basic optimisation model established in Section 3 owns the property of multimodularity, thus, the local optimum found by local search is the global optimum. We use three solution approaches to solve our problem instances: Local Search (LS), Mixed-Integer Linear Program (MILP) and Constraint Generation Algorithm (CGA). For solving the extended model which considers scheduling E-visit time windows formulated in Section 4, where the multimodularity property no longer holds, we apply the Local Search method by enumerating all feasible E-visit time window schedules s , MILP approach and Accelerated Cut Generation Algorithm (ACGA). All computations are conducted with a 64-bit computer with an AMD Core 2700X CPU at 3.60GHz and a RAM of 16 GB and programmed with R 4.2.0 and Gurobi 9.5.1.

Table A.2 in Online Appendix illustrates the computational performance of various solution approaches for $\lambda_E = 0.5$, while the remaining parameters adhere to Table 1. Overall, the proposed “CGA” and “ACGA” algorithms are the most efficient methods for models with and without optimising E-visit time windows, respectively. Solving the MILP directly via a Gurobi is slower than our algorithms because it discards the structure of the optimisation model. Although the local search procedure is similar to our algorithms, it is inefficient due to the time-consuming evaluation of the cost for a given schedule through recursive functions. Furthermore, for the model with E-visit time window scheduling, the ACGA algorithm utilises the lower bound generated by the current optimal joint schedule and the properties in Corollary 2 to eliminate non-optimal joint schedules and start with a high-quality solution. In summary, for the model without E-visit time window scheduling, the average computation times (in seconds) for local search, MILP, and CGA are 515.83, 717.93, and 257.06, respectively. As for the model with E-visit time window scheduling, the average computation time (in seconds) for ACGA is 7326.01, while local search and MILP solver are unable

Table 1
Parameter Settings

Parameter	Values	Parameter	Values
E-visit waiting patience (P)	2, 4	Overtime cost over idle cost (C_O/C_I)	0.5, 1.5, 2.0
E-visit waiting cost (C_E)	0.45, 0.9	Appointment no-show rate (p_s)	0.1, 0.5
Average E-visit arrival rate (λ_E)	0.2, 0.5, 0.8	E-visit Pattern	Dome, Flat, Bowl

Notes. Other basic parameters are set to be $T = 10$, $N = 15$, $C_W = 0.95$, $C_I = 4$, $\lambda_W = 0.1$.

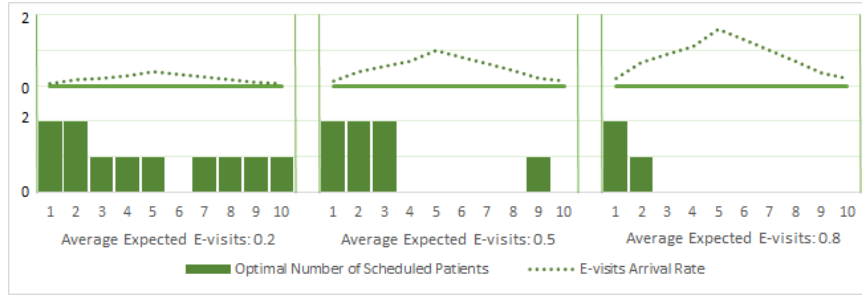
to produce the optimal schedule within 5 hours. For $P = 2$, the average cost gap is 28.16% and 9.45% for local search and MILP with a 5-hour computation time limit, respectively. When $P = 4$, the average cost gap between is as high as 70.71% and 28.43% for local search and MILP, respectively.

5.2. The Impact of E-visits on Appointment Scheduling

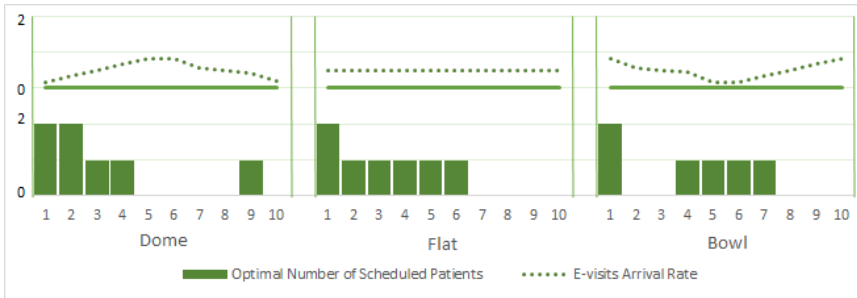
In this section, we investigate the impact of various factors related to E-visits on optimal appointment scheduling. Specifically, we examine how different E-visit arrival rate λ_E , E-visit arrival pattern, unit E-visit waiting cost C_E , and E-visit waiting tolerance P influence appointment scheduling when adopting “passive” control, i.e., only appointments can be scheduled.

Firstly, we observe that the optimal schedule tends to reserve time slots for potential E-visit arrivals when their arrival rate is high. Figure 4(a) shows the optimal schedules with different average arrival rates of E-visits ($\lambda_E \in \{0.2, 0.5, 0.8\}$). It clearly demonstrates that the higher the arrival rates of E-visits, the more “holes” are reserved. Besides, the arrival patterns of E-visits also influence the optimal schedules. Figure 4(b) shows this for a fixed average arrival rate of E-visits ($\lambda_E = 0.5$). We note that the optimal schedule reserves “holes” in the late slots and the slots with high E-visit rates for potential E-visits. And we also observe that the appointments are overbooked in the early slots to deal with no-shows. Such a phenomenon can be explained by the “Front-loading Pattern” of overbooking, i.e., overbooking is more likely to occur in the early slots, which is observed in previous literature on appointment scheduling (e.g., Hassin and Mendel, 2008; Zacharias and Pinedo, 2014, 2017; Zacharias and Yunes, 2020; Kong et al., 2020). “Front-loading Pattern” of overbooking exists in our setting even when the E-visit arrival rates in the early slots are high. The reason is that E-visits have patience to wait some slots at no cost.

We also examine the impact of cost structure (see Figure 5 below and Table A.3 in Online Appendix) on the optimal appointment schedule. We can observe that when the waiting cost of E-visits increases from



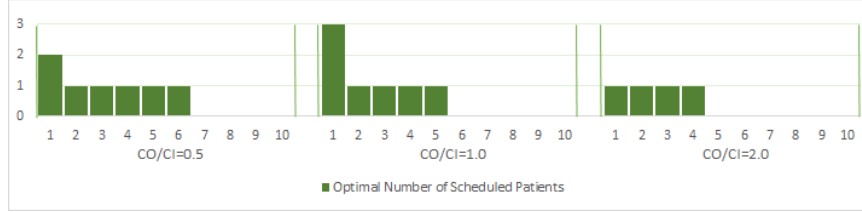
(a) Optimal appointment schedules with different average arrival rates of E-visits



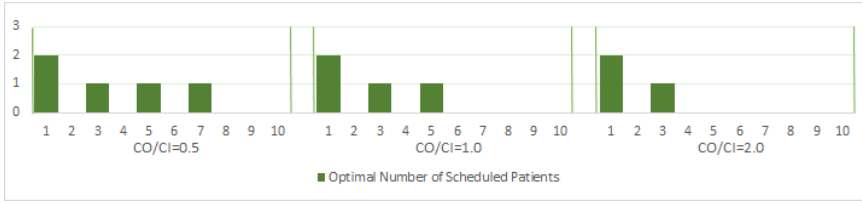
(b) Optimal appointment schedules with different arrival patterns of E-visits (average $\lambda_E = 0.5$)

Figure 4: Impact of E-visit arrivals on the optimal appointment schedule ($C_O/C_I = 0.5$, $C_E = 0.45$, $p_s = 0.5$, $P = 2$, $T = 10$)

0.45 (Figure 5(a)) to 0.9 (Figure 5(b)), fewer appointments are scheduled as E-visits become more important. When C_I is fixed and the overtime cost is relatively higher compared to the idle cost, more “holes” are reserved for E-visits to avoid overtime. Similarly, when the idle cost is relatively higher, fewer “holes” are reserved to save idle time. The impact of E-visit waiting tolerance P on the optimal appointment schedule can be found in Figure 6 below and Table A.4 in Online Appendix. When E-visits can tolerate longer waiting, the optimal appointment schedule reserves fewer “holes” and schedules more appointments to best utilise the working hours of the provider.



(a) $C_E = 0.45$



(b) $C_E = 0.9$

Figure 5: Optimal appointment schedules under different cost structures ($p_s = 0.5, P = 2, T = 10, \lambda_E = 0.5$, “Flat” shaped pattern)

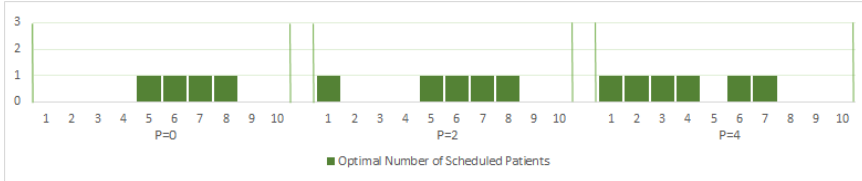


Figure 6: Optimal schedules under different patient waiting tolerance ($C_O/C_I = 0.5, C_E = 0.9, p_s = 0.5, T = 10, \lambda_E = 0.5$, “Bow” shaped pattern)

The optimal appointment schedule is significantly affected by various parameters related to E-visits. First of all, to achieve the best operational performance, it is recommended to schedule appointments during slots with low E-visit arrival rates, while reserving some “holes” for anticipated E-visits during high-arrival-rate slots. Furthermore, the ratio of unit overtime cost and unit idle time cost also plays a crucial role in determining the optimal appointment schedule. If the unit overtime cost is relatively higher, scheduling fewer appointments is more cost-effective; On the other hand, scheduling more appointments is more beneficial if the unit idle cost is higher. Last but not least, when the waiting tolerance of E-visits increases, it is essential to schedule more appointments. Overall, these observations highlight the importance of considering the arrival rate, arrival pattern, cost structure and waiting tolerance of E-visits when designing the optimal appointment schedule.

5.3. The Benefits of Scheduling E-visit Time Windows

This section compares the performance of “proactive” control, i.e., the extensive model that incorporates E-visit time window scheduling (Section 4), with that of only “passive” control, i.e., the basic model (Section 3). The purpose of this comparison is to demonstrate the improvements that can be achieved by incorporating the scheduling of E-visit time windows. Additionally, we investigate how the model parameters, such as the arrival pattern and rate of E-visits, cost structure, waiting tolerance of E-visits, etc., influence the performance differences between the two models in terms of the expected total cost (Γ^*), the total number of appointments scheduled (n^*), and the number of E-visit time windows opened (s^*). Table A.5 in Online Appendix shows the detailed results.

We first note that jointly scheduling E-visit time windows and appointments can result in an overall cost reduction of approximately 9%. Also, the “proactive” control tends to close some time slots for E-visits and schedule more appointments. It is important to note that the arrival patterns of E-visits have a significant impact on the effectiveness of “proactive” control. A “Bowl” shaped arrival pattern has the highest improvement, resulting in an overall cost reduction of 12.44%. On the other hand, a “Flat” or “Dome” shaped pattern can only reduce the cost by 6.51% or 8.09%, respectively. The reason for this is that the provider prefers E-visits to arrive during the early time slots, as they can increase the utilization of idle time without incurring excessive waiting or overtime costs. With a “Bowl” shaped pattern, such situation can be easily achieved by closing late slots for E-visits.

The cost structure also influences the performance improvement of scheduling E-visit time windows. When the ratio of C_O/C_I increases from 0.5 to 2.0, the cost reduction increases from 3.82% to 15.75%. When overtime is more costly, providers have more incentives to schedule appointments and close E-visit time windows to avoid excessive demands. As a result, providers can benefit more from scheduling E-visit time windows in the environment with higher over time costs. This observation indicates that our model and solution approaches provide a greater opportunity for providers to better manage their demand, thereby reducing the overtime as well as the associated costs.

One may imagine that the cost reduction by scheduling E-visit time windows is higher when C_E is larger. This is true only with a “Dome” shaped arrival pattern. When the E-visit arrivals follow a “Flat” or “Bowl” shaped pattern, the cost reduction is higher when C_E is smaller (see Table 2). Our study found that a rise in the unit E-visit waiting cost C_E results in a reduction of the number of E-visit time windows to be opened, at the same time leading to more patients to be scheduled in advance expecting that the E-visit arrival pattern follows the “Dome” shape. The insight behind this counter-intuitive observation is that, when scheduling the time windows is not allowed, a “Dome” shaped pattern results in longer wait time for E-visits as there is only one peak of their arrivals in the middle slots, while a “Flat” and “Bowl” shaped pattern results in more over time because more E-visits may come in the late slots. Therefore, “proactive” control is more effective and can bring more cost reduction when C_E is larger for “Dome” shaped pattern, and when C_E is smaller (C_O is relatively larger) for “Flat” or “Bowl” shaped pattern.

Finally, it is worth noting that the cost reduction achieved by “proactively” controlling E-visits is higher when their arrival rate (λ_E) is higher and their waiting tolerance (P) is shorter. This is because only with “passive” control, where appointments are scheduled in response to E-visit arrivals, the uncertainty of their arrival still interrupts the daily operations of the provider. When there are many E-visits arriving but they cannot tolerate long waiting, the side-effects of E-visits become even more significant. However, with “proactive” control, the proper joint scheduling of E-visit time windows and appointments can greatly mitigate such side effects. Therefore, the improvement is more significant when the side effects are more pronounced.

Table 2The Side Effects of C_E and E-visit Arrival Patterns

Parameters Arrival Pattern	Avg. Improvement								
	Dome			Flat			Bowl		
C_E	$\Delta\Gamma^*$	Δn^*	Δs^*	$\Delta\Gamma^*$	Δn^*	Δs^*	$\Delta\Gamma^*$	Δn^*	Δs^*
0.45	-7.20%	2.44	-2.61	-7.03%	2.55	-2.83	-13.60%	2.72	-2.50
0.9	-8.99%	2.28	-2.56	-5.98%	2.67	-2.78	-11.28%	2.94	-2.67

Notes. (a) The average improvement derives from 18 cases (3 different E-visit patterns by 2 no-show probabilities by 3 C_O/C_I rates). (b) $\Delta\Gamma^*$ represents the improvement by proactive control of E-visit time window, where $\Delta\Gamma^* = \frac{\Gamma(\text{Model2}) - \Gamma(\text{Model1})}{\Gamma(\text{Model1})}$, Model1 represents the basic model while Model2 represents the extensive model; Δn^* represents an increment of scheduled patients by proactive control of E-visit time window; Δs^* represents an increment of E-visit time windows by proactive control of E-visit time window, where the negative number means the E-visit time window has been closed under these slots.

5.4. The Patterns of Optimal Joint Schedules

This section examines how different E-visit arrival patterns, cost structures, E-visit arrival rates and E-visit waiting tolerances affect the joint optimal schedules in the model incorporating scheduling E-visit time windows. To better illustrate the results, a numerical example is shown in Figure 7. The sub-figures represent different cost ratios between idle time and overtime, i.e., $C_O/C_I = \{0.5, 2.0\}$, respectively; and the plots in each sub-figure represent different arrival patterns of E-visits (“Dome”, “Flat” and “Bowl”). In each plot, the top dotted line represents the arrival rates through time; the circle/cross symbols represent the optimal schedule of E-visit time windows, where the circle symbol means opening the slot for E-visits, and the cross one means closing the slot. The bottom bars represent the number of appointments scheduled in each slot.

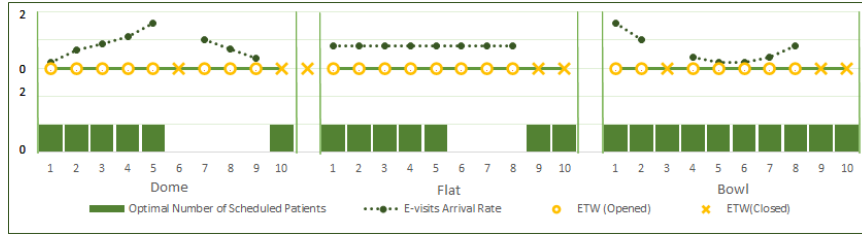
In general, under the optimal schedule of E-visit time windows, early slots and slots with a low E-visit arrival rate are opened for E-visits, while late slots and slots with a high arrival rate are closed for E-visits. This helps the provider to better manage the wait time of E-visits and in-clinic patients during peak hours. However, this structure is significantly impacted by different E-visit arrival patterns. From Figure 7, we can see that with a “Dome”-shaped E-visit pattern, some middle slots with a high arrival rate are also opened for E-visits. The reason is that, with a “Dome”-shaped pattern, if most middle slots are closed, the low arrival rates of E-visits cannot fully utilise the capacity, resulting in idle time.

The optimal appointment schedule is also highly influenced by the cost structure. As shown in Figure 7, when overtime becomes more costly, the optimal solution tends to schedule more appointments and open fewer E-visit time windows for better control of the overtime. More appointments are scheduled with the “Bowl” shaped E-visit arrival pattern compared with the “Flat” or “Dome” shaped pattern. Moreover, we observe that reserving “holes” for E-visits and overbooking still coexist even when using “proactive” control to manage the appointments and E-visits together (see Table A.6 in Online Appendix). However, overbooking is less likely to happen, and more appointments are scheduled compared to the “passive” control. This means that adopting “proactive” control can partially, if not fully, offset the uncertainties from no-shows and E-visits.

5.5. The Impact of E-visit Channel

This section investigates the impact of the E-visit channel on the entire service system. Without this channel, E-visit patients have the option to schedule an appointment, walk in for their care, or seek services from other clinics. However, with this channel, there is a possibility that walk-ins may switch to this channel. In this analysis, we consider four scenarios.

Managing Appointment-based Services with E-visits



(a) Optimal Joint Schedules Under Different E-visit Arrival Patterns ($C_O/C_I = 0.5$)



(b) Optimal Joint Schedules Under Different E-visit Arrival Patterns ($C_O/C_I = 2.0$)

Figure 7: Optimal Joint Schedules Under Different Cost Structures and Arrival Patterns ($C_E = 0.9, p_s = 0.5, P = 2, T = 10$)

The first scenario, called “no E-visits” (NE), occurs when the E-visit channel is closed and these patients seek alternative care options. The second scenario, “E-visits becoming scheduled patients” (EtoS), occurs when E-visit patients choose to schedule face-to-face appointments in the absence of the E-visit channel. Similarly, the third scenario, “E-visits choosing to walk in” (EtoW), considers the possibility of E-visit patients opting for walk-in visits. The final scenario, “walk-ins switching to E-visits” (WtoE), reflects the situation where walk-in patients start using the E-visit channel after its adoption.

To evaluate the impact of the E-visit channel on the system performance, we compare each scenario to the original system where all three types of patients exist along with the E-visit channel. Specifically, in the NE scenario, the parameter λ_E (the average E-visit arrival rate) decreases to 0. In the EtoS scenario, λ_E changes to 0, while N (the maximum number of patients to schedule) increases to $N + \lambda_E \times T$. In the EtoW scenario, λ_E decreases to 0, and λ_W (the walk-in arrival rate) becomes $\lambda_W + \lambda_E$. In the WtoE scenario, λ_W changes to 0, while λ_E increases to $\lambda_W + \lambda_E$. For each combination of model parameters (as shown in Table 1), we determine the optimal joint schedule for the original setting and the four scenarios mentioned above. We then compare the system performance in terms of the number of scheduled patients, patient wait time, provider idle time, overtime, and total cost.

The first fact is that, under current parameter settings, NE scenario and EtoS scenario have the same performance. The reason is that, without the E-visit channel, the optimal number of scheduled patients n^* is smaller than the maximum number of patients to schedule which is $N = 15$ in our setting. Thus, regardless of whether the E-visit patients are lost or go to the pool of scheduled patients, the optimal schedules are the same. One can imagine that, when N is set to be smaller than the optimal number of scheduled patients, the EtoS scenario would outperform the NE scenario as it has more potential patients to schedule. In general, in the NE or EtoS scenario, i.e., when E-visits opt out or choose to schedule an appointment, more appointments are scheduled to compensate for the absence of E-visits. We observe that even with more patients scheduled, the wait time is shorter, the overtime is shorter, and the idle time is longer. The reason behind this is that scheduled patients are more under control than E-visits. When E-visit arrival rate is low, the decreased wait time and overtime do not outweigh the idle time brought by the absence of E-visits, thus the total cost is increased. When E-visit demand is large, the reduced wait time and overtime are significant enough,

Table 3

The Change of the System Performance under Different Scenarios

λ_E	Scenario	Measurement Metrics				
		Δn^*	Δ Wait Time	Δ Overtime	Δ Idle Time	Δ Total Cost
0.2	NE	2.67	-22.25%	-47.72%	18.66%	9.61%
	EtoS	2.67	-22.25%	-47.72%	18.66%	9.61%
	EtoW	0.00	82.99%	19.91%	1.86%	21.20%
	WtoE	0.33	-11.42%	37.88%	-1.84%	-1.84%
0.5	NE	7.50	-35.23%	-75.87%	33.84%	7.42%
	EtoS	7.50	-35.23%	-75.87%	33.84%	7.42%
	EtoW	-0.83	157.73%	4.57%	19.23%	50.37%
	WtoE	0.33	6.56%	17.92%	-2.28%	1.22%
0.8	NE	9.33	-79.43%	-92.68%	97.39%	-25.84%
	EtoS	9.33	-79.43%	-92.68%	97.39%	-25.84%
	EtoW	-0.33	117.14%	6.80%	19.14%	78.07%
	WtoE	-0.17	-1.34%	8.42%	15.40%	6.76%

Notes. a) Five metrics are used to evaluate the system's performance: the optimal number of scheduled patients (n^*), the total wait time of all patients, the provider's overtime and idle time, and the total cost. b) The performance gap for each scenario per parameter setting is calculated as $\Delta M = \frac{M(\text{scenario}) - M(\text{original})}{M(\text{original})}$ where $M(\text{scenario})$ and $M(\text{original})$ represent the performance for the specific scenario and the original system, respectively. c) Each ΔM in the table is averaged from 6 cases of parameter settings ($C_O/C_I \in \{0.5, 1, 2\}$, $C_E \in \{0.45, 0.9\}$) with the following fixed parameters: "Dome"-shaped E-visit Pattern, $T = 10$, $N = 15$, $P = 2$, $p_s = 0.5$ and $C_I = 4$.

resulting in a lower total cost. The numerical observations suggest that, if the provider can persuade E-visits to schedule appointments, or she has a big pool of patients to schedule, closing E-visit channel is beneficial to the whole system when the E-visit demand is large. When the E-visit demand is small, E-visit channel is still attractive to the provider because of the lower waiting cost of E-visits.

In the EtoW scenario, i.e., when E-visits switch to the walk-in channel, fewer patients are scheduled. However, all of the wait time, overtime, idle time, and total cost increase. The reason is that E-visits can be partially scheduled by setting the optimal E-visit time windows, while walk-ins must be accepted and there is no notification for their arrivals. This observation suggests that closing the E-visit channel is undesirable for the provider when these patients will switch to the walk-in channel. Such a disadvantage becomes more significant when the E-visit demand is large.

In contrast to the EtoW scenario, where E-visits switch to the walk-in channel without the E-visit channel, in the WtoE scenario, all walk-ins are drawn to the E-visit channel. The WtoE scenario is expected to yield better performance. This holds true when the E-visit demand is low. However, as the demand for E-visits increases, the scheduling limitation for E-visit time windows becomes apparent. In such cases, accommodating walk-ins as a separate stream of arrivals can help to smooth the demand. Our observations indicate that when E-visit stream is small, encouraging walk-ins to the E-visit channel is beneficial for the provider. Nevertheless, it is important to retain some walk-ins when the E-visit demand is already substantial.

The presence of the E-visit channel has dual effects on the entire system. Generally, when the E-visit arrival rate is low, this channel can assist the provider in better managing daily demand and enhancing operational efficiency. Furthermore, the provider can benefit if walk-ins can be attracted to the E-visit channel in this case. However, when the E-visit demand increases, this channel may negatively impact the system as scheduling E-visit time windows still has limitations and cannot fully optimize the arrival of E-visits. Nonetheless, even with high E-visit demand, closing the E-visit channel is still not recommended when E-visits may choose the walk-in channel.

6. Case Study with Synthetic Data

In this section, we conduct a case study to explore the potential practical improvement by implementing the proposed optimal solutions. To achieve this, we utilise various data sources to create a synthetic dataset. The first data source pertains to E-visits, encompassing one-year patient visit records from a leading E-visit platform in China known as HaoDF⁵. We derive E-visit arrival patterns through statistical analysis of their arrival times recorded in the data. The second data source concerns walk-ins. We utilise the walk-in patterns obtained by Wang et al. (2020) based on their data set. The final data source focuses on practical schedules. By consulting with hospital practitioners, we obtain the commonly used scheduling rule, which evenly distributes appointments across available slots. Regarding E-visit time windows, we note that HaoDF allows E-visits at any time. Based on these observed scheduling policies, we are able to construct a schedule resembling those used in practice. We believe this synthetic data can partially, if not fully, capture the current practical environment.

HaoDF is a prominent E-visit platform with over 25,000 patient visits annually. The dataset spans from January 1st to December 30th, 2020, and encompasses 111,585 valid records of patient visits. Each record represents one patient electronic visit, containing the patient ID, the provider ID, and the arrival time. Within this dataset, 57,514 providers are included. We focus on three representative providers who are active throughout the year: a Dermatologist, a Paediatrician, and a Gynaecologist. We perform Poisson regression analysis on the arrivals of E-visit patients. The respective p -values for the goodness-of-fit test are 0.11, 0.19 and 0.36 for the Dermatologist, the Paediatrician and the Gynaecologist, respectively. It suggests that we cannot reject the hypothesis that the arrivals follow a Poisson distribution. Figure 8 depicts the half-hour E-visit arrival rate for each provider. We observe a decreasing pattern in the arrival rates for the Dermatologist, indicating high rates in the early stages and then followed by a decline. Regarding the Gynaecologist and Paediatrician, the arrival rate pattern follows a relatively “flat” shape.

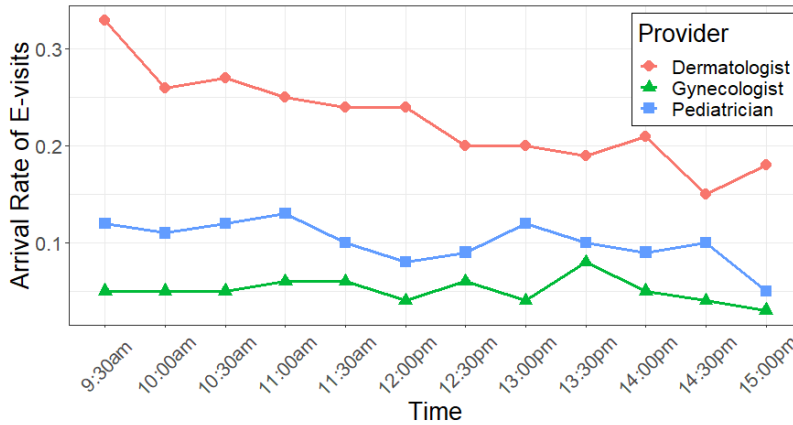


Figure 8: Expected Arrival of E-visits for Each Provider

The walk-in patterns replicate the observed Poisson process in Wang et al. (2020), with arrival rates of (0.45, 0.47, 0.48, 0.50, 0.50, 0.52, 0.52, 0.52, 0.57, 0.59, 0.54, 0.49) throughout the session. All other parameters align with the case study in Wang et al. (2020), including the session length of $T = 12$ and a no-show rate of 0.16. The cost parameters in Wang et al. (2020) are set as $(C_I, C_O) \in \{(5, 10), (5, 20), (10, 5), (10, 15)\}$ and $C_W \in \{0.5, 0.9\}$. Given that the cost of E-visit waiting is relatively higher than that of walk-in waiting, we set $C_E = 0.8$ for $C_W = 0.9$ and $C_E = 0.4$ for $C_W = 0.5$. The waiting patience of E-visits P is set to be

⁵<https://www.haodf.com>

2. As previously mentioned, the commonly used scheduling policies evenly distribute appointments across available slots and open all E-visit time windows. Therefore, the practical appointment schedule and E-visit time window schedule constructed are $\mathbf{x}^P = (1, 1, \dots, 1)$ and $\mathbf{s}^P = (1, 1, \dots, 1)$ respectively.

To evaluate the performance improvement achieved through the adoption of optimal joint schedules generated by our algorithm, we compare the cost components (total patient wait time, provider idle time, provider overtime) derived from the optimal schedule and practical schedule for each parameter setting. Additionally, we compute the total cost incurred by these two schedules. We observe that the optimal schedule results in slightly higher idle time but significantly lower wait time, and can achieve a remarkable 263% cost reduction over the practical schedule on average. Detailed numerical results are presented in Table A.7 in Online Appendix. While the practical schedule is informed by scheduling policies used in practice, it may not precisely represent the schedules employed in hospitals, potentially leading to an oversight in evaluating performance improvement. Nonetheless, it still demonstrates the significant potential of the proposed models and solution approaches in practical applications.

7. Conclusion

In this paper, we tackle the challenge of managing E-visits alongside scheduled patients and walk-ins in an outpatient care facility. Specifically, we develop a model to determine the optimal joint schedule of appointments and E-visit time windows, considering the uncertain arrivals of E-visits and walk-ins and the no-show behaviour of scheduled patients. We begin by introducing a basic model that employs “passive” control by reserving slots for walk-ins and E-visits. We then present an extended model that utilises “proactive” control by scheduling appropriate E-visit time windows. We show that given the schedule of E-visit time windows, the appointment scheduling problem is of multimodularity, enabling us to use local search to find the optimal appointment schedule. To further reduce the computational complexity, we formulate our model as a two-stage stochastic problem and utilise the Constraint Generation Method to solve it. Additionally, we introduce an Accelerated Constraint Generation method for the extended model, which takes advantage of the optimal solution’s structural properties. Our approaches allow us to handle complicated settings, including general arrival patterns of walk-ins and E-visits, non-linear unit waiting costs for E-visits, and time-dependent no-show behaviours. Overall, our models provide a framework for outpatient care facilities to effectively manage patient satisfaction while ensuring efficient resource utilisation.

Our analysis also provides valuable managerial insights for healthcare practitioners. Firstly, the presence of E-visits can create uncertainty in the service system, leading to increased costs for daily operations. Without considering the presence of E-visits when scheduling appointments, there can be excessive costs associated with the waiting of E-visits and the idling of providers. By figuring out the arrival patterns of E-visits, we may deliberately reserve some time slots for potential demand from the E-visit platform. This decision not only reduces the waiting of E-visits but also improves the operational efficiency of the clinic by better utilising time and resources.

Secondly, our research reveals that the optimal scheduling strategy for providers depends on various factors, such as E-visit arrival patterns, the cost structure and E-visit waiting tolerance. In general, when there is a high probability of E-visits with low waiting tolerance and high waiting costs, it is essential for the provider to consider the proper scheduling of the E-visit time windows. Moreover, the cost structure can influence the optimal schedule and the associated performance significantly. Therefore, it is crucial for the practitioners to determine the appropriate cost parameters when adopting our model and solution approaches.

Thirdly, when the E-visit arrival rate is low, the E-visit channel can help the provider in better managing daily demand and enhancing operational efficiency. However, when the E-visit demand increases, this channel may negatively impact the system because of the limitations of scheduling E-visit time windows.

Nevertheless, even with high E-visit demand, opening an E-visit channel is still necessary to avoid E-visits switching to the walk-in channel.

Our study highlights the importance of incorporating scheduling E-visit time windows into daily operations management. By proactively controlling for the uncertainties of E-visit arrivals and in-clinic no-shows, overall costs can be significantly reduced, healthcare resources can be better utilised, and patient satisfaction can be improved. The case study with synthetic data also demonstrates the practical importance and relevance of the proposed model and solution approaches. In conclusion, our research contributes to a better understanding of appointment scheduling in the presence of E-visits, offering useful methodologies and insights for healthcare practitioners, policymakers, and researchers.

Our research has some limitations that should be noted. First, we assume that the service time for each type of patient is the same, but it is possible that the service time for E-visits and in-clinic patients may differ. Secondly, we assume that the arrivals of E-visits are exogenous, meaning that they are not influenced by the provider's decisions on appointment schedules and E-visit time windows. However, E-visits may strategically switch to other channels when provider's decisions change. We leave these to future research.

References

- Ahmadi-Javid, A., Jalali, Z., Klassen, K.J., 2017. Outpatient appointment systems in healthcare: A review of optimization studies. *European Journal of Operational Research* 258, 3–34. doi:10.1016/j.ejor.2016.06.064.
- Altman, E., Gaujal, B., Hordijk, A., 2000. Multimodularity, Convexity, and Optimization Properties. *Mathematics of Operations Research* 25, 324–347. doi:10.1287/moor.25.2.324.12230.
- Asmussen, S., 2003. *Applied probability and queues*. volume 2. Springer. doi:10.1007/b97236.
- Bavafa, H., Hitt, L.M., Terwiesch, C., 2018. The impact of e-visits on visit frequencies and patient health: Evidence from primary care. *Management Science* 64, 5461–5480. doi:10.1287/mnsc.2017.2900.
- Bavafa, H., Savin, S., Terwiesch, C., 2021. Customizing Primary Care Delivery Using E-Visits. *Production and Operations Management* 30, 4306–4327. doi:10.1111/poms.13528.
- Cayirli, T., Veral, E., 2003. Outpatient scheduling in health care: A review of literature. *Production and Operations Management* 12, 519–549. doi:10.1111/j.1937-5956.2003.tb00218.x.
- Cayirli, T., Yang, K.K., Quek, S.A., 2012. A universal appointment rule in the presence of no-shows and walk-ins. *Production and Operations Management* 21, 682–697. doi:10.1111/j.1937-5956.2011.01297.x.
- Chen, X., Li, M., 2021. Discrete Convex Analysis and Its Applications in Operations: A Survey. *Production and Operations Management* 30, 1904–1926. doi:10.1111/poms.13234.
- China Banking Association, 2022. Report on China Banking Industry Services 2021. China Financial Publishing House, Beijing, China.
- Çil, E.B., Lariviere, M.A., 2013. Saving seats for strategic customers. *Operations Research* 61, 1321–1332. doi:10.1287/opre.2013.1218.
- Delana, K., Deo, S., Ramdas, K., Subburaman, G.B.B., Ravilla, T., 2022. Multichannel delivery in healthcare: The impact of telemedicine centers in southern india. *Management Science* 0, null. doi:10.1287/mnsc.2022.4488.
- Erdogan, S.A., Krupski, T.L., Lobo, J.M., 2018. Optimization of telemedicine appointments in rural areas. *Service Science* 10, 261–276. doi:10.1287/serv.2018.0222.
- Feldman, J., Liu, N., Topaloglu, H., Ziya, S., 2014. Appointment scheduling under patient preference and no-show behavior. *Operations Research* 62, 794–811. doi:10.1287/opre.2014.1286.
- Gupta, D., Denton, B., 2008. Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions* 40, 800–819. doi:10.1080/07408170802165880.
- Hajek, B., 1985. Extremal Splittings of Point Processes. *Mathematics of Operations Research* 10, 543–556. doi:10.1287/moor.10.4.543.
- Hassin, R., Mendel, S., 2008. Scheduling arrivals to queues: A single-server model with no-shows. *Management Science* 54, 565–572. doi:10.1287/mnsc.1070.0802.
- Hoppen, J., 2018. Medical appointment no shows. <https://www.kaggle.com/datasets/joniarroba/noshowappointments>. Accessed: 2023-03-26.
- Huh, W.T., Liu, N., Truong, V.A., 2013. Multiresource allocation scheduling in dynamic environments. *Manufacturing and Service Operations Management* 15, 280–291. doi:10.1287/msom.1120.0415.
- Hwang, E.H., Guo, X., Tan, Y., Dang, Y., 2022. Delivering Healthcare Through Teleconsultations: Implications for Offline Healthcare Disparity. *Information Systems Research* 33, 515–539. doi:10.1287/isre.2021.1055.
- Jiang, R., Shen, S., Zhang, Y., 2017. Integer programming approaches for appointment scheduling with random no-shows and service durations. *Operations Research* 65, 1638–1656. doi:10.1287/opre.2017.1656.
- Kaandorp, G.C., Koole, G., 2007. Optimal outpatient appointment scheduling. *Health Care Management Science* 10, 217–229. doi:10.1007/s10729-007-9015-x.
- Koeleman, P.M., Koole, G.M., 2012. Optimal outpatient appointment scheduling with emergency arrivals and general service times. *IIE Transactions on Healthcare Systems Engineering* 2, 14–30. doi:10.1080/19488300.2012.665154.

- Kong, Q., Li, S., Liu, N., Teo, C.P., Yan, Z., 2020. Appointment Scheduling Under Time-Dependent Patient Appointment Scheduling Under Time-Dependent Patient. *Management Science* 66, 3480–3500. doi:10.1287/mnsc.2019.3366. This.
- Laganga, L.R., Lawrence, S.R., 2012. Appointment overbooking in health care clinics to improve patient service and clinic performance. *Production and Operations Management* 21, 874–888. doi:10.1111/j.1937-5956.2011.01308.x.
- Li, N., Li, X., Zhang, C., Kong, N., 2021. Integrated optimization of appointment allocation and access prioritization in patient-centred outpatient scheduling. *Computers and Industrial Engineering* 154, 107125. doi:10.1016/j.cie.2021.107125.
- Liu, N., 2016. Optimal Choice for Appointment Scheduling Window under Patient No-Show Behavior. *Production and Operations Management* 25, 128–142. doi:10.1111/poms.12401.
- Liu, N., van Jaarsveld, W., Wang, S., Xiao, G., 2023a. Managing outpatient service with strategic walk-ins. *Management Science* doi:10.1287/mnsc.2023.4676.
- Liu, N., Wang, S., Zychlinski, N., 2023b. RL or URL: Managing Outpatient (Tele)visits with Strategic Behavior. *SSRN Preprint* doi:10.2139/ssrn.4383199.
- Liu, N., Ziya, S., Kulkarni, V.G., 2010. Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. *Manufacturing and Service Operations Management* 12, 347–364. doi:10.1287/msom.1090.0272.
- Liu, Y., Lafayette, W., Gilbert, S., Lai, G., 2020. Pricing, Quality and Competition at On-Demand Healthcare Service Platforms. *SSRN Preprint* doi:10.2139/ssrn.3253855.
- Luo, J., Kulkarni, V.G., Ziya, S., 2012. Appointment Scheduling Under Patient No-Shows and Service Interruptions. *Manufacturing & Service Operations Management* 14, 670–684. doi:10.1287/msom.1120.0394.
- Maister, D.H., 1985. The Psychology of Waiting Lines, in: *Harvard Business School*, pp. 71–78. URL: http://www.columbia.edu/~ww2040/4615S13/Psychology_of_Waiting_Lines.pdf.
- Murota, K., 2005. Note on multimodularity and L-convexity. *Mathematics of Operations Research* 30, 658–661. doi:10.1287/moor.1040.0142.
- OECD, 2023. The COVID-19 Pandemic and the Future of Telemedicine. *OECD Publishing*. doi:10.1787/ac8b0a27-en.
- Pan, X., Geng, N., Xie, X., Wen, J., 2020. Managing appointments with waiting time targets and random walk-ins. *Omega* 95, 102062. doi:10.1016/j.omega.2019.04.005.
- Pinedo, M., Zacharias, C., Zhu, N., 2015. Scheduling in the service industries: An overview. *Journal of Systems Science and Systems Engineering* 24, 1–48. doi:10.1007/s11518-015-5266-0.
- Qin, J., Chan, C.W., Dong, J., 2022. Waiting Online versus In-Person in Outpatient Clinics : An Empirical Study on Visit Incompletion. working paper , *Columbia University*. URL: http://www.columbia.edu/~cc3179/TeleIncomplete_2023.pdf.
- Rajan, B., Tezcan, T., Seidmann, A., 2019. Service systems with heterogeneous customers: Investigating the effect of telemedicine on chronic care. *Management Science* 65, 1236–1267. doi:10.1287/mnsc.2017.2979.
- Robinson, L.W., Chen, R.R., 2010. A comparison of traditional and open-access policies for appointment scheduling. *Manufacturing & Service Operations Management* 12, 330–346. doi:10.1287/msom.1090.0270.
- Savin, S., Xu, Y., Zhu, L., 2021. Delivering Multi-Specialty Care via Online Telemedicine Platforms. *SSRN Preprint* doi:10.2139/ssrn.3479544.
- Shaked, M., Shanthikumar, J.G., 2007. *Stochastic Orders*. Springer New York, NY. doi:10.1007/978-0-387-34675-5.
- Shen, X., Li, N., Xie, X., 2023. Multiserver Time Window Allowance Appointment Scheduling Problem for Coordinating Online and Offline Patients. *SSRN Preprint* doi:10.2139/ssrn.4352121.
- Wang, S., Liu, N., Wan, G., 2020. Managing appointment-based services in the presence of walk-in customers. *Management Science* 66, 667–686. doi:10.1287/mnsc.2018.3239.
- Yu, X., Bayram, A., 2021. Managing capacity for virtual and office appointments in chronic care. *Health Care Management Science* 24, 742–767. doi:10.1007/s10729-021-09546-4.
- Zacharias, C., Pinedo, M., 2014. Appointment scheduling with no-shows and overbooking. *Production and Operations Management* 23, 788–801. doi:10.1111/poms.12065.
- Zacharias, C., Pinedo, M., 2017. Managing customer arrivals in service systems with multiple identical servers. *Manufacturing and Service Operations Management* 19, 639–656. doi:10.1287/msom.2017.0629.
- Zacharias, C., Yunes, T., 2020. Multimodularity in the stochastic appointment scheduling problem with discrete arrival epochs. *Management Science* 66, 744–763. doi:10.1287/mnsc.2018.3242.
- Zhang, R., Han, X., Wang, R., Zhang, J., Zhang, Y., 2022. Please Don't make me wait! influence of customers' waiting preference and no-show behavior on appointment systems. *Production and Operations Management* doi:10.1111/poms.13928.
- Zhong, X., 2018. A queueing approach for appointment capacity planning in primary care clinics with electronic visits. *IIE Transactions* 50, 970–988. doi:10.1080/24725854.2018.1486053.
- Zhong, X., Hoonakker, P., Bain, P.A., Musa, A.J., Li, J., 2018. The impact of e-visits on patient access to primary care. *Health Care Management Science* 21, 475–491. doi:10.1007/s10729-017-9404-8.
- Zhong, X., Li, J., Bain, P.A., Musa, A.J., 2017. Electronic Visits in Primary Care: Modeling, Analysis, and Scheduling Policies. *IEEE Transactions on Automation Science and Engineering* 14, 1451–1466. doi:10.1109/TASE.2016.2555854.
- Zhou, S., Ding, Y., Huh, W.T., Wan, G., 2021. Constant Job-Allowance Policies for Appointment Scheduling: Performance Bounds and Numerical Analysis. *Production and Operations Management* 30, 2211–2231. doi:10.1111/poms.13362.
- Zocchi, M., Uscher-Pines, L., Ober, A.J., Kapinos, K.A., 2020. Costs of maintaining a high-volume telemedicine program in community health centers. *RAND*. doi:10.7249/RR100-3.

Online Appendix

Appendix A Supplementary Tables

Table A.1

List of Notations.

Constants	
T	Total number of appointment slots (i.e. the clinic session length)
\bar{T}	Maximum number of time slots ($\bar{T} > T$)
N	Maximum number of patients can be scheduled
C_S	Waiting cost per unit time for the scheduled patient, normalised to be 1
C_W	Waiting cost per unit time for the walk-in
C_E	Waiting cost per unit time for the E-visit
C_I	Idling cost per unit time for the provider
C_O	Overtime cost per unit time for the provider
P	Maximum patience time for the E-visit
(Auxiliary) Decision variables	
x_t	Number of patients to be scheduled in slot t
n	Total number of scheduled patients, $n = \sum_{t=1}^T x_t$
\mathbf{x}	Vector of x_t
$z_{i,t}$	If patient i is scheduled at slot t , then $z_{i,t} = 1$, otherwise $z_{i,t} = 0$
\mathbf{z}	Vector of $z_{i,t}$
y_t^S	Number of scheduled patients waiting at the end of the slot t
\mathbf{y}^S	Vector of $y_{s,t}$
y_t^I	Number of in-clinic patients waiting at the end of slot t
\mathbf{y}^I	Vector of y_t^I
y_t^p	Number of in-clinic patients and E-visits who have waited for more than p slots at the end of slot t , where $p = 1, \dots, P + 1$ and P is a predefined number
\mathbf{y}^p	Vector of y_t^p
Values to be calculated	
Γ_S	Expected total wait time of scheduled patients
Γ_W	Expected total wait time of walk-ins
Γ_E	Expected total wait time (more than P slots) of online patients
Γ_I	Expected idle time of the provider
Γ_O	Expected over time of the provider
$\Psi_t(k)$	Probability of k scheduled patients waiting for services at the end of slot t
$\Pi_t(k)$	Probability of k scheduled patients and walk-ins waiting for services at the end of slot t
$\Phi_t^p(k)$	Probability of k scheduled patients, walk-ins who are waiting and online patients who have waited for more than p slots at the end of slot t , where $p = 1, \dots, P + 1$ and P is a predefined number
$Y(\mathbf{x}, \mathbf{z}, \omega)$	Total cost under schedule \mathbf{x} and \mathbf{z} , when scenario ω occurs
Random variables	
β	Vector of β_t
$\alpha_t(x_t)$	Number of show-up patients in slot t under a scheduled \mathbf{x} , and its pmf is $q_t(j, x_t) = \Pr(\alpha_t(x_t) = j)$
$\alpha(\mathbf{x})$	Vector of $\alpha_t(x_t)$
$\Omega(\mathbf{x})$	Set of all possible scenarios for α , β and γ
ω	$\omega_x \in \Omega_x(\mathbf{x})$, an arbitrary scenarios in the set $\Omega_x(\mathbf{x})$
$\gamma_{i,t}(\omega)$	Indicator of whether patient i shows up at slot t under scenario ω_z
$\gamma(\omega)$	Vector of $\gamma_{i,t}(\omega)$

Table A.2

Computational performance of different solution approaches

Basic Model								
Parameters			Local Search		MILP		CGA	
P	C_O/C_I	C_E	MEAN	CV	MEAN	CV	MEAN	CV
2	0.50	0.45	537.77	0.99	831.96	0.57	147.43	0.63
2	0.50	0.90	432.62	1.04	258.79	0.24	167.55	0.65
2	1.00	0.45	451.66	1.02	384.04	0.38	126.49	0.47
2	1.00	0.90	383.70	1.04	273.22	0.03	123.97	0.62
2	2.00	0.45	376.48	1.04	282.84	0.06	90.92	0.56
2	2.00	0.90	341.92	1.07	249.69	0.04	116.78	0.81
4	0.50	0.45	780.64	0.83	2989.15	0.78	475.00	0.70
4	0.50	0.90	642.68	0.85	792.07	0.25	523.16	0.74
4	1.00	0.45	647.89	0.85	946.03	0.34	394.15	0.58
4	1.00	0.90	559.63	0.86	531.77	0.13	339.70	0.55
4	2.00	0.45	536.31	0.89	582.57	0.17	247.97	0.63
4	2.00	0.90	498.65	0.91	492.99	0.06	331.59	0.90

Model Incorporating E-visit Time Windows								
Parameters			Local Search		MILP		ACGA	
P	C_O/C_I	C_E	MEAN	GAP(%)	MEAN	GAP(%)	MEAN	CV
2	0.50	0.45	> 18000	24.58	> 18000	5.48	2819.12	0.35
2	0.50	0.90	> 18000	35.71	> 18000	10.31	2985.54	0.36
2	1.00	0.45	> 18000	24.72	> 18000	6.66	3064.05	0.39
2	1.00	0.90	> 18000	21.38	> 18000	12.42	3621.59	0.45
2	2.00	0.45	> 18000	36.86	> 18000	14.58	3183.57	0.49
2	2.00	0.90	> 18000	25.73	> 18000	7.23	2679.61	0.46
4	0.50	0.45	> 18000	57.39	> 18000	21.28	11794.39	0.35
4	0.50	0.90	> 18000	75.35	> 18000	26.55	12801.00	0.30
4	1.00	0.45	> 18000	86.52	> 18000	35.06	11135.30	0.23
4	1.00	0.90	> 18000	59.25	> 18000	26.58	10067.34	0.28
4	2.00	0.45	> 18000	82.43	> 18000	39.73	11749.75	0.27
4	2.00	0.90	> 18000	63.30	> 18000	21.40	12010.88	0.36

Notes. (a) We fix $\lambda_E = 0.5$ while other parameters follow Table 1. (b) "MEAN" represents the mean computation time (sec) which is averaged over 6 cases (3 different E-visit patterns by 2 no-show probabilities). "CV" represents the computation time's coefficient of variation. The computation time includes the time of random sample generation and model solving. (c) For the cases where we cannot obtain the optimal solution within 5 hours by Local Search or MILP, CV is not available. Thus we report the average performance "GAP" which is calculated by $\frac{\text{cost of 5-hour solution} - \text{optimal cost output by ACGA}}{\text{optimal cost output by ACGA}}$.

Table A.3

Optimal schedules under different cost structures

$C_E = 0.45$					$C_E = 0.9$				
Pattern	C_O/C_I	Schedule	n^*	Holes	Pattern	C_O/C_I	Schedule	n^*	Holes
Dome	0.5	2 1 2 0 0 0 0 0 0 1	6	6	Dome	0.5	2 1 2 0 0 0 0 0 0 0	5	7
Dome	1.0	2 1 2 0 0 0 0 0 0 0	5	7	Dome	1.0	3 0 1 0 0 0 0 0 0 0	4	8
Dome	2.0	3 0 1 0 0 0 0 0 0 0	4	8	Dome	2.0	3 0 1 0 0 0 0 0 0 0	4	8
Flat	0.5	2 1 1 1 1 1 0 0 0 0	7	4	Flat	0.5	2 0 1 0 1 0 1 0 0 0	5	6
Flat	1.0	3 1 1 1 1 0 0 0 0 0	7	5	Flat	1.0	2 0 1 0 1 0 0 0 0 0	4	7
Flat	2.0	1 1 1 1 0 0 0 0 0 0	4	6	Flat	2.0	2 0 1 0 0 0 0 0 0 0	3	8
Bowl	0.5	2 0 0 1 1 1 1 0 0 0	6	5	Bowl	0.5	1 0 0 0 1 1 1 1 0 0	5	5
Bowl	1.0	1 1 0 1 1 0 2 0 0 0	6	5	Bowl	1.0	1 0 1 0 0 1 2 0 0 0	5	6
Bowl	2.0	1 0 1 0 1 1 1 0 0 0	5	5	Bowl	2.0	1 0 1 0 1 1 1 0 0 0	5	5

Notes. (a) The basic parameters are $T = 10$, $N = 15$, $\lambda_E = 0.5$ and $P = 2$ while other parameters vary. (b) n^* represents the number of scheduled patients, while holes represent the slots deliberately reserved for uncertain arrivals.

Table A.4

Optimal schedules under different patient waiting tolerance

$C_E = 0.45$					$C_E = 0.9$				
Pattern	P	Schedule	n^*	Holes	Pattern	P	Schedule	n^*	Holes
Dome	0	2 2 1 0 0 0 0 0 0 1	6	6	Dome	0	2 1 0 0 0 0 0 0 0 1	4	7
Dome	2	2 1 2 0 0 0 0 0 0 1	6	6	Dome	2	2 1 2 0 0 0 0 0 0 0	5	7
Dome	4	2 1 2 1 0 0 0 1 0 0	7	5	Dome	4	2 1 2 0 0 0 0 1 0 0	6	6
Flat	0	2 0 1 1 1 1 0 0 0 0	6	5	Flat	0	1 1 0 1 1 0 0 0 0 0	4	6
Flat	2	2 1 1 1 1 1 0 0 0 0	7	4	Flat	2	2 0 1 0 1 0 1 0 0 0	5	6
Flat	4	2 1 1 1 1 1 1 0 0 0	8	3	Flat	4	2 1 1 1 1 1 0 0 0 0	7	4
Bowl	0	1 0 1 0 1 1 1 0 0 0	5	5	Bowl	0	0 0 0 1 1 1 1 0 0 0	4	6
Bowl	2	2 0 0 1 1 1 1 0 0 0	6	5	Bowl	2	1 0 0 0 1 1 1 1 0 0	5	5
Bowl	4	1 1 0 2 0 1 1 1 0 0	7	4	Bowl	4	1 1 1 1 0 1 1 0 0 0	6	4

Notes. (a) The basic parameters are $T = 10$, $N = 15$, $\lambda_E = 0.5$ and $C_O/C_I = 0.5$ while other parameters vary. (b) n^* represents the number of scheduled patients, while holes represent the slots deliberately reserved for uncertain arrivals.

Table A.5
Sensitive Analysis

Parameters				Improvement								
Arrival Pattern				Dome			Flat			Bowl		
C_O/C_I	C_E	P	λ_E	$\Delta\Gamma^*$	Δn^*	Δs^*	$\Delta\Gamma^*$	Δn^*	Δs^*	$\Delta\Gamma^*$	Δn^*	Δs^*
0.50	0.45	2	0.20	0.00%	0	0	-0.02%	0	-1	-0.44%	0	-1
0.50	0.45	2	0.50	-4.29%	1	-2	-0.32%	1	-2	-3.01%	2	-2
0.50	0.45	2	0.80	-8.54%	4	-3	-4.02%	3	-3	-12.35%	4	-3
0.50	0.45	4	0.20	0.00%	0	0	-0.08%	0	-1	-6.91%	1	-1
0.50	0.45	4	0.50	-0.34%	0	-1	-0.57%	1	-2	-3.47%	1	-1
0.50	0.45	4	0.80	-9.68%	2	-2	-5.49%	2	-2	-6.98%	4	-3
0.50	0.90	2	0.20	0.00%	0	0	-0.03%	0	-1	-0.54%	0	-1
0.50	0.90	2	0.50	-4.12%	1	-2	-4.07%	2	-2	-0.13%	3	-2
0.50	0.90	2	0.80	-14.69%	4	-3	-1.74%	5	-4	-10.81%	6	-4
0.50	0.90	4	0.20	0.00%	0	0	-0.07%	0	-1	-0.27%	0	-1
0.50	0.90	4	0.50	-0.37%	2	-2	-2.75%	1	-2	-1.00%	2	-1
0.50	0.90	4	0.80	-15.30%	4	-3	-9.14%	2	-2	-7.78%	3	-2
1.00	0.45	2	0.20	-0.42%	0	-1	-1.84%	0	-1	-3.93%	1	-1
1.00	0.45	2	0.50	-0.91%	2	-2	-2.46%	3	-3	-9.92%	2	-2
1.00	0.45	2	0.80	-15.51%	5	-4	-14.60%	6	-5	-26.15%	6	-5
1.00	0.45	4	0.20	-0.42%	0	-1	-1.83%	0	-1	-3.93%	0	-1
1.00	0.45	4	0.50	-2.76%	2	-3	-1.48%	3	-3	-10.35%	2	-2
1.00	0.45	4	0.80	-13.93%	5	-4	-12.25%	4	-4	-22.26%	5	-4
1.00	0.90	2	0.20	-0.42%	0	-1	-1.80%	0	-1	-1.59%	0	-1
1.00	0.90	2	0.50	-2.26%	2	-3	-0.87%	2	-2	-8.68%	3	-3
1.00	0.90	2	0.80	-21.26%	3	-3	-11.62%	6	-5	-22.38%	7	-5
1.00	0.90	4	0.20	-0.42%	0	-1	-1.79%	0	-1	-3.80%	0	-1
1.00	0.90	4	0.50	-5.72%	1	-2	-1.71%	2	-2	-6.87%	2	-2
1.00	0.90	4	0.80	-22.44%	3	-3	-2.79%	5	-4	-19.04%	4	-3
2.00	0.45	2	0.20	-2.61%	1	-2	-3.84%	1	-2	-5.88%	1	-2
2.00	0.45	2	0.50	-7.56%	5	-5	-9.47%	4	-4	-21.54%	3	-3
2.00	0.45	2	0.80	-28.28%	7	-6	-28.81%	7	-6	-41.32%	7	-5
2.00	0.45	4	0.20	-2.99%	1	-2	-4.13%	1	-2	-6.06%	1	-2
2.00	0.45	4	0.50	-5.30%	2	-3	-8.99%	4	-4	-21.54%	3	-3
2.00	0.45	4	0.80	-26.01%	7	-6	-26.42%	6	-5	-38.69%	6	-4
2.00	0.90	2	0.20	-2.31%	1	-3	-3.21%	1	-2	-5.25%	1	-2
2.00	0.90	2	0.50	-5.52%	4	-4	-9.25%	5	-5	-19.89%	4	-4
2.00	0.90	2	0.80	-34.17%	6	-6	-26.72%	7	-6	-36.54%	8	-6
2.00	0.90	4	0.20	-2.29%	1	-2	-3.74%	1	-2	-5.71%	1	-2
2.00	0.90	4	0.50	-2.19%	3	-3	-4.75%	3	-3	-17.69%	3	-3
2.00	0.90	4	0.80	-28.27%	6	-5	-21.55%	6	-5	-35.04%	6	-5
Average				-8.09%	2.36	-2.58	-6.51%	2.61	-2.81	-12.44%	2.83	-2.58

Notes. (a) $\Delta\Gamma^*$ represents the improvement by proactive control of E-visit time window, where $\Delta\Gamma^* = \frac{\Gamma(\text{Model2}) - \Gamma(\text{Model1})}{\Gamma(\text{Model1})}$, Model1 represents the basic model while Model2 represents the extensive model; Δn^* represents an increment of scheduled patients by proactive control of E-visit time window; Δs^* represents an increment of E-visit time windows by proactive control of E-visit time window, where the negative number means the E-visit time window has been closed under these slots.

Table A.6

The Optimal Joint Schedule Patterns

C_O/C_I	C_E	Schedule	ETW
0.5	0.45	2 1 1 1 1 1 1 0 1 1	1 1 1 1 1 0 1 1 1 0
1	0.45	2 1 1 1 1 1 1 0 0 1	1 1 1 1 1 0 1 1 1 0
2	0.45	2 1 1 1 1 0 1 1 0 1	1 1 1 1 1 0 1 0 0 0
0.5	0.9	2 1 1 1 1 1 1 0 1 1	1 1 1 1 1 0 1 1 1 0
1	0.9	2 1 1 1 1 1 1 0 0 1	1 1 1 1 1 0 1 1 1 0
2	0.9	2 1 1 1 1 1 1 0 1 1	1 1 1 0 1 0 1 0 0 0

Notes. (a) ETW means the optimal schedule for E-visit time windows. (b) All results were obtained when E-visits followed the "Dome" Pattern, $\lambda_E = 0.5$, $P = 4$, $p_s = 0.5$, $T = 10$, $N = 15$.

Table A.7

Performance Improvement by Adopting the Optimal Schedules over the Practical Schedules

(C_W, C_E)	(0.5, 0.4)				(0.9, 0.8)			
(C_I, C_O)	(5, 10)	(5, 20)	(10, 5)	(10, 15)	(5, 10)	(5, 20)	(10, 5)	(10, 15)
The Dermatologist								
Δ Wait Time	-74.91	-74.91	-66.77	-70.89	-74.91	-78.09	-71.2	-74.91
Δ Overtime	-5.99	-5.99	-5.41	-5.77	-5.99	-6.11	-5.75	-5.99
Δ Idle Time	2.34	2.34	1.24	1.72	2.34	3.06	1.74	2.34
% Total Cost	-326.99%	-417.25%	-162.05%	-252.69%	-381.99%	-472.27%	-207.45%	-286.85%
The Paediatrician								
Δ Wait Time	-52.46	-55.22	-43.75	-51.58	-55.22	-57.52	-47.99	-52.46
Δ Overtime	-4.68	-4.9	-4.05	-4.73	-4.9	-5.01	-4.43	-4.68
Δ Idle Time	2.04	2.66	0.99	1.99	2.66	3.39	1.45	2.04
% Total Cost	-258.62%	-379.23%	-115.62%	-202.02%	-311.41%	-415.34%	-146.32%	-223.53
The Gynaecologist								
Δ Wait Time	-46.68	-48.85	-39.39	-42.93	-46.75	-49.04	-39.9	-42.93
Δ Overtime	-4.36	-4.52	-3.83	-4.18	-4.34	-4.51	-3.81	-4.18
Δ Idle Time	2.36	3.04	1.21	1.7	2.38	3.05	1.23	1.7
% Total Cost	-238.61%	-353.96%	-97.75%	-184.38%	-270.35%	-381.21%	-125.84%	-201.73%

Notes. a) For the specific metrics, we employ absolute improvement (Δ) to evaluate the gap by adopting the optimal schedule and the practical schedule, which is calculated by $M(\text{Optimal}) - M(\text{Heuristic})$ where $M(\cdot)$ is the measurement metric. b) For the total cost, we employ the relative improvement (%), which is calculated as $\frac{\text{cost by practical policy} - \text{cost of the optimal schedule}}{\text{cost of the optimal schedule}}$.

Appendix B Supplementary Technical Results

B.1 Solution Approaches for (P2)

For the sake of concision and to simplify the notation, we first introduce the matrix form of (P2).

Let $\mathbf{z} = (z_{1,1}, \dots, z_{1,N}, z_{2,1}, \dots, z_{2,N}, \dots, z_{T,1}, \dots, z_{T,N})$, $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \dots, \gamma_T)$ be a $T \times N$ -dimensional vector and $\boldsymbol{\delta}(\omega) = (\delta_1(\omega), \dots, \delta_N(\omega))'$ be a N -dimensional vector. Let $\mathbf{c} = (1 - C_W, \dots, 1 - C_W, C_W - C_E, \dots, C_W - C_E, C_E, \dots, C_E, 0 \dots 0, C_I + C_O, 0 \dots 0)'$ be a $(P + 3) \times \bar{T}$ -dimensional vector where the first \bar{T} elements are $1 - C_W$, the second \bar{T} elements are $C_W - C_E$, the third \bar{T} elements are C_E , and the last $P \times \bar{T}$ elements are 0 except for the $((P + 2) \times \bar{T} + T)$ th element is $C_I + C_O$. Let $\mathbf{y} = (y_1^S, \dots, y_{\bar{T}}^S, y_1^I, \dots, y_{\bar{T}}^I, y_1^{P+1}, \dots, y_{\bar{T}}^{P+1}, y_1^P, \dots, y_{\bar{T}}^P, \dots, y_T^2, \dots, y_T^2, y_T^1, \dots, y_T^1)'$, which is a $(P + 3) \times \bar{T}$ -dimensional decision vector and each of the elements is a non-negative integer number. Let \mathbf{E} be a block matrix made up with T identity matrices that $\mathbf{E} = [\mathbf{I} \mathbf{I} \dots \mathbf{I}]$, in which \mathbf{I} is the N -dimensional identity matrix. Let \mathbf{e} be a $T \times N$ -dimensional unit vector. Let \mathbf{U} be a $(P + 3) \times \bar{T}$ -dimensional square matrix such that

$$\mathbf{U} = \begin{bmatrix} \mathbf{U}^0 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{U}^0 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{U}^1 & \mathbf{U}^2 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{U}^1 & \mathbf{U}^2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{U}^1 & \mathbf{U}^2 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{U}^0 \end{bmatrix}, \quad (\text{B.1})$$

where

$$\mathbf{U}^0 = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -1 & 1 \end{bmatrix}, \quad (\text{B.2})$$

\mathbf{U}^1 is a \bar{T} dimensional identity matrix, and

$$\mathbf{U}^2 = \mathbf{U}^0 - \mathbf{U}^1 = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ -1 & 0 & 0 & \dots & 0 & 0 \\ 0 & -1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -1 & 0 \end{bmatrix}. \quad (\text{B.3})$$

Let $\mathbf{D}(\omega)$ be a $(P + 3) \times \bar{T}$ by $T \times N$ matrix, where for all $0 \leq t \leq T, 0 \leq i \leq N, 0 \leq p \leq P + 2$, element $\mathbf{D}_{t+p\bar{T},(t-1) \times N+i}(\omega)$ is equal to $\delta_{i,t}$ and the rest elements are 0. Let $\mathbf{r}(\omega) = c(0, \dots, 0, \gamma_1(\omega), \dots, \gamma_T(\omega), 0, \dots, 0)$ be a $(P + 3) \times \bar{T}$ vector, the first $(P + 2) \times \bar{T}$ and the last $\bar{T} - T$ elements are 0, the rest T elements are $\gamma_t(\omega)$ for $0 \leq t \leq T$. Let $\mathbf{b}(\omega)$ be a $(P + 3) \times \bar{T}$ vector, where for $0 \leq t \leq T, 0 \leq p \leq P + 3$, element $\mathbf{b}_{t+p\bar{T}}(\omega)$ is $\beta_t(\omega) - 1$ and the rest elements are -1 .

Now, the matrix form of (P2) can be expressed as follows.

$$\min_{\mathbf{z} \in \mathbb{Z}^+} \quad \mathbb{E}_\omega[\mathbf{Y}(\mathbf{z}, \omega) - C_I \boldsymbol{\delta}(\omega)' \mathbf{z}] \quad (\text{M.P2})$$

$$\text{s.t.} \quad \mathbf{E} \mathbf{z} \leq \mathbf{e} \quad (\text{B.4})$$

where $\Upsilon(\mathbf{z}, \omega) =$

$$\min_{\mathbf{y} \geq \mathbf{0}} \quad \mathbf{c}'\mathbf{y} \quad (\text{M.Sub.LP})$$

$$\text{s.t.} \quad \mathbf{U}\mathbf{y} \geq \mathbf{D}(\omega)\mathbf{z} + \mathbf{r}(\omega) + \mathbf{b}(\omega) \quad (\text{B.5})$$

Noting that the first stage problem is 0 – 1 integer programming and the second stage problem is a linear program.

Let us define the feasible region for the schedule \mathbf{z} as \mathcal{Z} , i.e., $\mathcal{Z} = \{\mathbf{z} | \mathbf{z} \in \mathbb{Z}^+, \mathbf{E}\mathbf{z} \leq \mathbf{e}\}$. Then we write the dual of the second-stage problem (M.Sub.LP) for any given schedule $\mathbf{z} \in \mathcal{Z}$ and scenario $\omega \in \Omega$ as follows

$$\max_{\mathbf{v} \geq \mathbf{0}} \quad \mathbf{v}'[\mathbf{D}(\omega)\mathbf{z} + \mathbf{r}(\omega) + \mathbf{b}(\omega)] \quad (\text{Sub.Dual})$$

$$\text{s.t.} \quad \mathbf{U}'\mathbf{v} \leq \mathbf{c}, \quad (\text{B.6})$$

where \mathbf{v} is the vector of dual variables.

Let $\mathbf{v}(\mathbf{z}, \omega)$ denote the optimal solution to (Sub.Dual) for any given schedule \mathbf{z} and scenario ω . By strong duality we will have, for any $\mathbf{z} \in \mathcal{Z}$,

$$\mathbf{v}(\mathbf{z}, \omega)'[\mathbf{D}(\omega)\mathbf{z} + \mathbf{r}(\omega) + \mathbf{b}(\omega)] = \Upsilon(\mathbf{z}, \omega). \quad (\text{B.7})$$

Since (Sub.Dual) is a maximisation problem, then for any \mathbf{z} and $\mathbf{z}^0 \in \mathcal{Z}$, we will have

$$\mathbf{v}(\mathbf{z}^0, \omega)'[\mathbf{D}(\omega)\mathbf{z} + \mathbf{r}(\omega) + \mathbf{b}(\omega)] \leq \mathbf{v}(\mathbf{z}, \omega)'[\mathbf{D}(\omega)\mathbf{z} + \mathbf{r}(\omega) + \mathbf{b}(\omega)] = \Upsilon(\mathbf{z}, \omega). \quad (\text{B.8})$$

Taking the expectation over ω of the inequality (B.8), we will have, for any \mathbf{z} and $\mathbf{z}^0 \in \mathcal{Z}$,

$$\mathbb{E}_\omega[\mathbf{v}(\mathbf{z}^0, \omega)' \mathbf{D}(\omega)]\mathbf{z} + \mathbb{E}_\omega[\mathbf{v}(\mathbf{z}^0, \omega)' \mathbf{r}(\omega)] + \mathbb{E}_\omega[\mathbf{v}(\mathbf{z}^0, \omega)' \mathbf{b}(\omega)] \leq \mathbb{E}_\omega[\Upsilon(\mathbf{z}, \omega)]. \quad (\text{B.9})$$

Let us further define the following notations for any $\mathbf{z} \in \mathcal{Z}$

$$\mathbf{a}(\mathbf{z}) = \mathbb{E}_\omega[\mathbf{v}(\mathbf{z}, \omega)' \mathbf{D}(\omega)], \quad (\text{B.10})$$

$$\mathbf{g}(\mathbf{z}) = \mathbb{E}_\omega[\mathbf{v}(\mathbf{z}, \omega)' \mathbf{r}(\omega)], \quad (\text{B.11})$$

$$\mathbf{m}(\mathbf{z}) = \mathbb{E}_\omega[\mathbf{v}(\mathbf{z}, \omega)' \mathbf{b}(\omega)], \quad (\text{B.12})$$

$$\Upsilon(\mathbf{z}) = \mathbb{E}_\omega[\Upsilon(\mathbf{z}, \omega)]. \quad (\text{B.13})$$

We will have, for any \mathbf{z} and $\mathbf{z}^0 \in \mathcal{Z}$,

$$\mathbf{a}(\mathbf{z}^0)\mathbf{z} + \mathbf{g}(\mathbf{z}^0) + \mathbf{m}(\mathbf{z}^0) \leq \Upsilon(\mathbf{z}). \quad (\text{B.14})$$

Let us define $\mathbf{h} = C_I \mathbb{E}_\omega[\delta(\omega)']$, and add it into (B.14), we will have, for any \mathbf{z} and $\mathbf{z}^0 \in \mathcal{Z}$,

$$\mathbf{a}(\mathbf{z}^0)\mathbf{z} + \mathbf{g}(\mathbf{z}^0) + \mathbf{m}(\mathbf{z}^0) - \mathbf{h}\mathbf{z} \leq \Upsilon(\mathbf{z}) - \mathbf{h}\mathbf{z} = \Gamma(\mathbf{z}), \quad (\text{B.15})$$

where $\Gamma(\mathbf{z})$ is the total expected cost of schedule \mathbf{z} , and the equality holds when $\mathbf{z} = \mathbf{z}^0$.

Next we present an equivalent formulation for (M.P2), which is

$$\min_{\mathbf{z} \in \mathcal{Z}, u} \quad u \quad (\text{P2.Dual})$$

$$\text{s.t.} \quad [\mathbf{a}(\mathbf{z}^0) - \mathbf{h}]\mathbf{z} + \mathbf{g}(\mathbf{z}^0) + \mathbf{m}(\mathbf{z}^0) \leq u, \quad \forall \mathbf{z}^0 \in \mathcal{Z} \quad (\text{B.16})$$

To see the equivalence, let \mathbf{z}^* be the optimal solution to (M.P2), thus $u^* = \Upsilon(\mathbf{z}^*) - \mathbf{h}\mathbf{z}^*$. By (B.15), (\mathbf{z}^*, u^*) must be a feasible solution to (Sub.Dual). Let $(\mathbf{z}^\#, u^\#)$ be the optimal solution to (P2.Dual), there must be $[\mathbf{a}(\mathbf{z}^\#) - \mathbf{h}]\mathbf{z}^\# + g(\mathbf{z}^\#) + m(\mathbf{z}^\#) = u^\# \leq u^* = \Upsilon(\mathbf{z}^*) - \mathbf{h}\mathbf{z}^*$. Since u^* is the minimum value of $\Upsilon(\mathbf{z}) - \mathbf{h}\mathbf{z}$ for any $\mathbf{z} \in \mathcal{Z}$, then we must have $u^* = u^\#$, i.e., solving (P2.Dual) is equivalent to solving (M.P2).

For solving (P2.Dual), we design an algorithm which is similar to the one in Wang et al. (2020) as follows.

Algorithm 1: Cut Generation Algorithm (CGA)

```

1 Initialise a schedule ( $\mathbf{x}^0$ ) and obtain a corresponding  $\mathbf{z}^0$ ;
2 Let  $\mathbf{A} \leftarrow \mathbf{a}(\mathbf{z}^0) - \mathbf{h}$ ,  $\mathbf{g}_m \leftarrow g(\mathbf{z}^0) + m(\mathbf{z}^0)$ ,  $\mathbf{e} \leftarrow \mathbf{1}$ ;
3 Let  $u \leftarrow \Upsilon(\mathbf{z}^0) - \mathbf{h}\mathbf{z}^0$ ;
4 while  $flag = 1$  do
5      $flag \leftarrow 0$ ;
6     for all neighbours of  $\mathbf{x}^0$  do
7         Let  $\mathbf{x}$  denote the current neighbour and obtain its corresponding  $\mathbf{z}$ ;
8         if  $\mathbf{A}\mathbf{z} + \mathbf{g}_m < \mathbf{u}\mathbf{e}$  then
9              $\mathbf{A} \leftarrow (\mathbf{A}; \mathbf{a}(\mathbf{z}) - \mathbf{h})$ ,  $\mathbf{g}_m \leftarrow (\mathbf{g}_m; g(\mathbf{z}) + m(\mathbf{z}))$ ,  $\mathbf{e} \leftarrow (\mathbf{e}; 1)$ ;
10            if  $\Upsilon(\mathbf{z}) - \mathbf{h}\mathbf{z} < u$  then
11                 $u \leftarrow \Upsilon(\mathbf{z}) - \mathbf{h}\mathbf{z}$ ;
12                 $\mathbf{z}^0 \leftarrow \mathbf{z}$  and  $\mathbf{x}_t^0 \leftarrow \sum_{i=1}^N z_{i,t}$ ;
13                 $flag \leftarrow 1$ ;
14                break;
15            end
16        end
17    end
18 end
19 return  $\mathbf{x}^0$ 
    
```

This algorithm presents a local search procedure which leverages the property of multimodularity on \mathbf{x} of the problem and generates cuts according constraint (B.16) to speed up the search the optimal schedule. For each \mathbf{x} being checked, a corresponding \mathbf{z} is derived. If it satisfy the existing cuts, a new cut associated with this solution is created. If its cost is lower than the current upper bound, the bound and current schedule are updated accordingly. More details on this algorithm can be found in Wang et al. (2020).

B.2 Solution Approaches for (P4)

Similar to Section B.1, we can use a matrix form to represent the two-stage programming problem (P4).

$$\min_{\mathbf{z}, \mathbf{s} \in \mathbb{Z}^+} \mathbb{E}_\omega \left[\Upsilon(\mathbf{z}, \mathbf{s}, \omega) - \mathbf{C}_I \boldsymbol{\delta}(\omega)' \mathbf{z} - \mathbf{C}_I \boldsymbol{\gamma}(\omega)' \mathbf{s} \right] \quad (\text{M.P4})$$

$$\text{s.t. } \mathbf{E}\mathbf{z} \leq \mathbf{e} \quad (\text{B.17})$$

$$\mathbf{I}\mathbf{s} \leq \mathbf{e} \quad (\text{B.18})$$

where $\Upsilon(\mathbf{z}, \mathbf{s}, \omega) =$

$$\min_{\mathbf{y} \geq 0} \mathbf{c}'\mathbf{y} \quad (\text{M.Sub.LP.2})$$

$$\text{s.t. } \mathbf{U}\mathbf{y} \geq \mathbf{D}(\omega)\mathbf{z} + \mathbf{R}(\omega)\mathbf{s} + \mathbf{b}(\omega) \quad (\text{B.19})$$

The notations used in (M.P4) are very similar to those used in (M.P2). The differences result from the additional decisions \mathbf{s} . Let $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \dots, \gamma_T)'$, $\mathbf{s} = (s_1, s_2, \dots, s_T)'$. Thus in the objective function of the

master problem, we have $\gamma(\omega)'s$ to represent the expected number of served E-visits, i.e., $\sum_{t=1}^T \sum_{i=1}^N \gamma_t(\omega)s_i$; and the constraint $\mathbf{I}s \leq \mathbf{e}$ represents $s_t \leq 1$ for any t where \mathbf{I} is an identity matrix. And in the sub-problem, the only change is that, in (M.P2), we use vector $\mathbf{r}(\omega)$ to denote $\gamma_t(\omega)$ in each constraint, while in (M.P4), we use $\mathbf{R}(\omega)s$ to represent $\gamma_t(\omega)s_t$ in each constraint, where $\mathbf{R}(\omega)$ is a $(P+3) \times \bar{T}$ vector with the first $(P+2) \times \bar{T}$ and the last $\bar{T} - T$ elements are 0, the rest T elements are $\gamma_t(\omega)$ for $0 \leq t \leq T$.

Denote the set $\{\mathbf{z} | \mathbf{E}\mathbf{z} \leq \mathbf{e}, \mathbf{z} \in \mathbb{Z}^+\}$ as \mathcal{Z} and $\{\mathbf{s} | \mathbf{I}\mathbf{s} \leq \mathbf{e}, \mathbf{s} \in \mathbb{Z}^+\}$ as \mathcal{S} . Then for any $\mathbf{z} \in \mathcal{Z}$, $\mathbf{s} \in \mathcal{S}$ and $\omega \in \Omega$, we can take the dual for the sub-problem (M.Sub.LP.2). We have,

$$\max_{\mathbf{v} \geq 0} \mathbf{v}'[\mathbf{D}(\omega)\mathbf{z} + \mathbf{R}(\omega)\mathbf{s} + \mathbf{b}(\omega)] \quad (\text{Sub.Dual.2})$$

$$\text{s.t. } \mathbf{U}'\mathbf{v} \leq \mathbf{c}, \quad (\text{B.20})$$

where \mathbf{v} denotes the vector of dual variables (with slightly abusing the notations).

Denote the optimal solution of (Sub.Dual.2) for any feasible schedule (\mathbf{z}, \mathbf{s}) and scenario ω by $\mathbf{v}(\mathbf{z}, \mathbf{s}, \omega)$. By strong duality we will have, for any $\mathbf{z} \in \mathcal{Z}$ and $\mathbf{s} \in \mathcal{S}$,

$$\mathbf{v}(\mathbf{z}, \mathbf{s}, \omega)'[\mathbf{D}(\omega)\mathbf{z} + \mathbf{R}(\omega)\mathbf{s} + \mathbf{b}(\omega)] = \Upsilon(\mathbf{z}, \mathbf{s}, \omega). \quad (\text{B.21})$$

Since (Sub.Dual.2) is a maximisation problem, then for any \mathbf{z} and $\mathbf{z}^0 \in \mathcal{Z}$, \mathbf{s} and $\mathbf{s}^0 \in \mathcal{S}$, we will have

$$\mathbf{v}(\mathbf{z}^0, \mathbf{s}^0, \omega)'[\mathbf{D}(\omega)\mathbf{z} + \mathbf{R}(\omega)\mathbf{s} + \mathbf{b}(\omega)] \leq \mathbf{v}(\mathbf{z}, \mathbf{s}, \omega)'[\mathbf{D}(\omega)\mathbf{z} + \mathbf{r}(\omega)\mathbf{s} + \mathbf{b}(\omega)] = \Upsilon(\mathbf{z}, \omega). \quad (\text{B.22})$$

Taking the expectation over ω of the inequality (B.22), we will have, for any \mathbf{z} and $\mathbf{z}^0 \in \mathcal{Z}$, \mathbf{s} and $\mathbf{s}^0 \in \mathcal{S}$,

$$\mathbb{E}_\omega[\mathbf{v}(\mathbf{z}^0, \mathbf{s}^0, \omega)' \mathbf{D}(\omega)]\mathbf{z} + \mathbb{E}_\omega[\mathbf{v}(\mathbf{z}^0, \mathbf{s}^0, \omega)' \mathbf{R}(\omega)]\mathbf{s} + \mathbb{E}_\omega[\mathbf{v}(\mathbf{z}^0, \mathbf{s}^0, \omega)' \mathbf{b}(\omega)] \leq \mathbb{E}_\omega[\Upsilon(\mathbf{z}, \mathbf{s}, \omega)]. \quad (\text{B.23})$$

Let us further define the following notations for any $\mathbf{z} \in \mathcal{Z}$ and $\mathbf{s} \in \mathcal{S}$,

$$\mathbf{a}(\mathbf{z}, \mathbf{s}) = \mathbb{E}_\omega[\mathbf{v}(\mathbf{z}, \mathbf{s}, \omega)' \mathbf{D}(\omega)], \quad (\text{B.24})$$

$$\mathbf{g}(\mathbf{z}, \mathbf{s}) = \mathbb{E}_\omega[\mathbf{v}(\mathbf{z}, \mathbf{s}, \omega)' \mathbf{R}(\omega)], \quad (\text{B.25})$$

$$\mathbf{m}(\mathbf{z}, \mathbf{s}) = \mathbb{E}_\omega[\mathbf{v}(\mathbf{z}, \mathbf{s}, \omega)' \mathbf{b}(\omega)], \quad (\text{B.26})$$

$$\Upsilon(\mathbf{z}, \mathbf{s}) = \mathbb{E}_\omega[\Upsilon(\mathbf{z}, \mathbf{s}, \omega)]. \quad (\text{B.27})$$

We will have, for any \mathbf{z} and $\mathbf{z}^0 \in \mathcal{Z}$, \mathbf{s} and $\mathbf{s}^0 \in \mathcal{S}$,

$$\mathbf{a}(\mathbf{z}^0, \mathbf{s}^0)\mathbf{z} + \mathbf{g}(\mathbf{z}^0, \mathbf{s}^0)\mathbf{s} + \mathbf{m}(\mathbf{z}^0, \mathbf{s}^0) \leq \Upsilon(\mathbf{z}, \mathbf{s}), \quad (\text{B.28})$$

Let us define $\mathbf{h}_z = C_I \mathbb{E}_\omega[\delta(\omega)']$ and $\mathbf{h}_s = C_I \mathbb{E}_\omega[\gamma(\omega)']$, and add it into (B.28), we will have, for any \mathbf{z} and $\mathbf{z}^0 \in \mathcal{Z}$, \mathbf{s} and $\mathbf{s}^0 \in \mathcal{S}$,

$$\mathbf{a}(\mathbf{z}^0, \mathbf{s}^0)\mathbf{z} + \mathbf{g}(\mathbf{z}^0, \mathbf{s}^0)\mathbf{s} + \mathbf{m}(\mathbf{z}^0, \mathbf{s}^0) - \mathbf{h}_z\mathbf{z} - \mathbf{h}_s\mathbf{s} \leq \Upsilon(\mathbf{z}, \mathbf{s}) - \mathbf{h}_z\mathbf{z} - \mathbf{h}_s\mathbf{s} = \Gamma(\mathbf{z}, \mathbf{s}), \quad (\text{B.29})$$

of which the RHS is the overall cost with a joint schedule (\mathbf{z}, \mathbf{s}) .

Then we can present an equivalent formulation for (M.P4), which is

$$\min_{\mathbf{z} \in \mathcal{Z}, \mathbf{s} \in \mathcal{S}, u} u \quad (\text{P4.Dual})$$

$$\text{s.t. } [\mathbf{a}(\mathbf{z}^0, \mathbf{s}^0) - \mathbf{h}_z]\mathbf{z} + [\mathbf{g}(\mathbf{z}^0, \mathbf{s}^0) - \mathbf{h}_s]\mathbf{s} + \mathbf{m}(\mathbf{z}^0, \mathbf{s}^0) \leq u, \quad \forall \mathbf{z}^0 \in \mathcal{Z}, \mathbf{s}^0 \in \mathcal{S} \quad (\text{B.30})$$

To see the equivalence, let (z^*, s^*) be the optimal solution to (M.P4), and let $u^* = Y(z^*, s^*) - h_z z^* - h_s s^*$. It is obvious that (z^*, s^*, u^*) is a feasible solution to (P4.Dual). Let $(z^\#, s^\#, u^\#)$ be the optimal solution to (P4.Dual), then we must have $u^* \geq u^\#$. Suppose that $u^* > u^\#$, then we have $[a(z^\#, s^\#) - h_z]z^\# + [g(z^\#, s^\#) - h_s]s^\# + m(z^\#, s^\#) < u^*$, which contradicts that (z^*, s^*) is the optimal solution to (M.P4). Thus, $u^* = u^\#$ and solving (P4.Dual) is equivalent to solving (M.P4).

Algorithm 2: Accelerated Cut Generation Algorithm (ACGA)

```

1  Initialise a schedule  $(x^*, s^*)$  and generate a corresponding  $z^*$  for  $x^*$ ;
2  Let  $A \leftarrow a(z^*, s^*) - h_z$ ,  $G \leftarrow g(z^*, s^*) - h_s$ ,  $m \leftarrow m(z^*, s^*)$ ,  $e \leftarrow 1$ ;
3  Let  $\bar{u} \leftarrow (a(z^*, s^*) - h_z)z^* + (g(z^*, s^*) - h_s)s^* + m(z^*, s^*)$ ;
4  for all feasible  $s$  do
5      Initialise an appointment schedule  $x^0$  such  $(x^0, s)$  is not dominated by  $(x^*, s^*)$  according to
        Corollary 2;
6      Generate a corresponding  $z^0$  for  $x^0$ ;
7      Let  $A \leftarrow (A; a(z^0, s) - h_z)$ ,  $G \leftarrow (G; g(z^0, s) - h_s)$ ,  $m \leftarrow (m; m(z^0, s))$ ,  $e \leftarrow (e; 1)$ , flag  $\leftarrow 1$ ;
8      Let  $u \leftarrow (a(z^0, s) - h_z)z^0 + (g(z^0, s) - h_s)s + m(z^0, s)$ ;
9      if  $u < \bar{u}$  then
10          $\bar{u} \leftarrow u$ ,  $(x^*, z^*, s^*) \leftarrow (x^0, z^0, s)$ ;
11     end
12     while flag = 1 do
13         flag  $\leftarrow 0$ ;
14         for all neighbours of  $x^0$  do
15             Let  $x$  denote the current neighbour and generate a corresponding  $z$  for it;
16             if  $(x, s)$  is not dominated by  $(x^0, s)$  according to Corollary 2 then
17                 if  $Az + Gs + m < ue$  then
18                      $A \leftarrow (A; a(z, s) - h_z)$ ,  $G \leftarrow (G; g(z, s) - h_s)$ ,  $m \leftarrow (m; m(z, s))$ ,  $e \leftarrow (e; 1)$ ;
19                     if  $(a(z, s) - h_z)z + (g(z, s) - h_s)s + m(z, s) < u$  then
20                          $u \leftarrow (a(z, s) - h_z)z + (g(z, s) - h_s)s + m(z, s)$ ;
21                          $z^0 \leftarrow z$  and  $x_t^0 \leftarrow \sum_{i=1}^N z_{i,t}$ ;
22                         flag  $\leftarrow 1$ ;
23                         if  $u < \bar{u}$  then
24                              $\bar{u} \leftarrow u$ ,  $(x^*, z^*, s^*) \leftarrow (x^0, z^0, s)$ ;
25                         end
26                         break;
27                     end
28                 end
29             end
30         end
31     end
32 end
33 return  $(x^*, s^*)$ 
    
```

Solving problem (P4.Dual) can be much more efficient than solving (M.P4). Let (z^0, s^0) be the current solution checked. Let \bar{u} be the smallest objective value we have obtained so far (which serves as an upper bound). Then, we can generate a cut by (B.30). With these cuts, we can eliminate a sub-optimal solution without calculating its objective value. Recall that, due to the multimodularity of x , the optimal appointment schedule can be found by local search given the schedule of E-visit time windows. Such local search

procedure can be also applied on searching for \mathbf{z} for a given \mathbf{s} . Additionally, the structures of the optimal joint schedule explored in Proposition 3 and Corollary 2 can also be utilised to eliminate non-optimal schedules more efficiently. Motivated by these elegant properties of the optimal solution and the mathematical nature of (P4.Dual), we design an Accelerated Cut Generation Algorithm (ACGA) shown in Algorithm 2.

In Algorithm 2, we first initialise a joint schedule $(\mathbf{x}^*, \mathbf{z}^*, \mathbf{s}^*)$, generate a cut associated with it and set its cost as \bar{u} (steps 1-3). Then, we check all possible E-visit time window schedules (loop 4-32). For each E-visit time window schedule \mathbf{s} that is examined, we initialize an appointment schedule \mathbf{x}^0 which is not dominated according to 2, add the associated cut, and set its cost as u (steps 5-8). If u is lower than \bar{u} , we update \bar{u} and $(\mathbf{x}^*, \mathbf{z}^*, \mathbf{s}^*)$ (step 10). Then we use local search to find the best appointment schedule \mathbf{x} for the current \mathbf{s} (loop 12-31). For each neighbour that is examined, we first check whether it is dominated (step 16). If not, we check whether it satisfies the current cuts (step 17). If so, we add the associated cut (step 18) and check whether its cost is lower u ; if so, we update u and $(\mathbf{x}^0, \mathbf{z}^0)$ (steps 19-22). If u is further lower than \bar{u} , we update \bar{u} and $(\mathbf{x}^*, \mathbf{z}^*, \mathbf{s}^*)$ (step 24). Once all feasible \mathbf{s} have been examined, the algorithm terminates, and returns the optimal joint schedule $(\mathbf{x}^*, \mathbf{s}^*)$ (step 33).

B.3 Proofs of Analytical Results

B.3.1 Definitions and Auxiliary Results

In order to prove Proposition 3, we will need a number of definitions, as well as some ancillary results.

Definition 3. For two random variables a and b , if $\Pr\{a < b\} = 1$, we say that $a \leq b$, $a - b \leq 0$ or $b - a \geq 0$.

Lemma 1 (Shaked and Shanthikumar (2007)). If two random variables $a \leq b$, then $\mathbb{E}[h(a)] \leq \mathbb{E}[h(b)]$ for all increasing functions $h(\cdot)$ for which the expectations exist.

Lemma 2. If a and b are non-negative integer random variables such that $a - b \geq 0$, for any non-negative integer random variable β , $0 \leq (a - 1)^+ - (b - 1)^+ \leq (a + \beta - 1)^+ - (b + \beta - 1)^+ \leq a - b$ holds.

Proof 1. Proof of Lemma 2. First, let us prove $0 \leq (a - 1)^+ - (b - 1)^+$. (In the rest of the appendix, if we say “when a random variable is greater than, less than, or equal to a number”, it means “when the realisation of a random variable is greater than, less than, or equal to a number”.) When $a - 1 \leq 0$ and $b - 1 \leq 0$, then $(a - 1)^+ = (b - 1)^+ = 0$. When $a - 1 \leq 0$ and $b - 1 > 0$, $\Pr\{0 \leq (a - 1)^+ - (b - 1)^+\} = \Pr\{a \geq 1\} = 1$. When $a - 1 > 0$ and $b - 1 \leq 0$, $\Pr\{0 \leq (a - 1)^+ - (b - 1)^+\} = \Pr\{b \leq a\} = 1$. So $0 \leq (a - 1)^+ - (b - 1)^+$.

Second, let us prove $(a - 1)^+ - (b - 1)^+ \leq (a + \beta - 1)^+ - (b + \beta - 1)^+$. When $\beta = 0$, $(a - 1)^+ - (b - 1)^+ = (a + \beta - 1)^+ - (b + \beta - 1)^+$. When $\beta \geq 1$, $(a + \beta - 1)^+ - (b + \beta - 1)^+ = a - b$; when $a - 1 \leq 0$ and $b - 1 \leq 0$, then $(a - 1)^+ - (b - 1)^+ = 0 \leq a - b$; when $a - 1 > 0$ and $b - 1 \leq 0$, $(a - 1)^+ - (b - 1)^+ = a - 1 \leq a - b$; when $a - 1 > 0$ and $b - 1 > 0$, $(a - 1)^+ - (b - 1)^+ = a - b$, thus $(a - 1)^+ - (b - 1)^+ \leq a - b$. So $(a - 1)^+ - (b - 1)^+ \leq (a + \beta - 1)^+ - (b + \beta - 1)^+$.

Third, let us prove $(a + \beta - 1)^+ - (b + \beta - 1)^+ \leq a - b$. When $\beta = 0$, $(a + \beta - 1)^+ - (b + \beta - 1)^+$ becomes $(a - 1)^+ - (b - 1)^+$ which has been proved to be no greater than $a - b$. When $\beta \geq 1$, $(a + \beta - 1)^+ - (b + \beta - 1)^+ = a - b$. So, $(a + \beta - 1)^+ - (b + \beta - 1)^+ \leq a - b$. \square

Lemma 3. If a is a non-negative integer random variable, for any non-negative integer random variable β , $(a + \beta - 1)^+ \leq (a - 1)^+ + \beta \leq a + \beta$ holds.

Proof 2. Proof of Lemma 3. When $\beta = 0$, $(a + \beta - 1)^+ = (a - 1)^+ + \beta$. When $\beta \geq 1$, $(a + \beta - 1)^+ = a - 1 + \beta$, since $\Pr\{a - 1 \leq (a - 1)^+\} = 1$, then $(a + \beta - 1)^+ \leq (a - 1)^+ + \beta$. When $a \geq 1$, $(a - 1)^+ = a - 1$, then $a - (a - 1)^+ = 1$. When $a \leq 1$, $(a - 1)^+ = 0$, then $a - (a - 1)^+ = a$ with $0 \leq a \leq 1$. Then $0 \leq a - (a - 1)^+ \leq 1$. So, $(a - 1)^+ + \beta \leq a + \beta$. \square

Lemma 4. *If a and b are non-negative integer random variables such that $0 \leq a - b \leq 1$, for any non-negative integer random variable β and γ , $0 \leq (a + \beta - 1)^+ - (b + \beta - 1)^+ \leq (a + \beta + \gamma - 1)^+ - (b + \beta + \gamma - 1)^+ \leq 1$ holds.*

Proof 3. *Proof of Lemma 4. By Lemma 2, we have $0 \leq (a + \beta - 1)^+ - (b + \beta - 1)^+$ and $0 \leq (a + \beta + \gamma - 1)^+ - (b + \beta + \gamma - 1)^+$. When $\gamma = 0$, $0 \leq (a + \beta - 1)^+ - (b + \beta - 1)^+ = (a + \beta + \gamma - 1)^+ - (b + \beta + \gamma - 1)^+ \leq a - b$. When $\gamma \geq 1$, $(a + \beta + \gamma - 1)^+ - (b + \beta + \gamma - 1)^+ = a + \beta + \gamma - 1 - (b + \beta + \gamma - 1) = a - b$. Since $0 \leq a - b \leq 1$, so $0 \leq (a + \beta - 1)^+ - (b + \beta - 1)^+ \leq (a + \beta + \gamma - 1)^+ - (b + \beta + \gamma - 1)^+ \leq 1$ holds. \square*

B.3.2 Proof of Proposition 1

To prove the Proposition 1, we first consider the all-show-up case ($\alpha_t(x_t) = x_t$) and introduce Lemmas 5-7 below to help us prove the Proposition 1. With slight abuse of notations, we replicate the notations from the show-up case in the discussion.

Consider a feasible schedule $\mathbf{x} \in \mathbb{Z}_+^T$, random walk-ins $\beta = \{\beta_1, \dots, \beta_T\}$ and random E-visits $\gamma = \{\gamma_1, \dots, \gamma_T\}$. Let y_t^S denote the number of scheduled patients waiting at the end of slot t , and let y_t^I denote the number of in-clinic patients (i.e., scheduled patients plus walk-ins) waiting at the end of slot t . Denote y_t^p as the number of in-clinic waiting patients and E-visits who have waited at least p slots at the end of slot t (i.e., all in-clinic patients plus E-visits who arrived p slots ago), for $p = 1, \dots, P + 1$. According to the Lindley equations (13-16), for $t = 1, \dots, T$, we have

$$y_t^S = (y_{t-1}^S + x_t - 1)^+, \quad (\text{B.31})$$

$$y_t^I = (y_{t-1}^I + x_t + \beta_t - 1)^+, \quad (\text{B.32})$$

$$y_t^1 = (y_{t-1}^1 + x_t + \beta_t + \gamma_t - 1)^+, \quad (\text{B.33})$$

$$y_t^p = (y_{t-1}^{p-1} + x_t + \beta_t - 1)^+ \quad \forall p \geq 2. \quad (\text{B.34})$$

Lemma 5. *Let subscript i represent the schedule $\mathbf{x} + \mathbf{v}_i$, subscript j represent the schedule $\mathbf{x} + \mathbf{v}_j$, subscript ij represent the schedule $\mathbf{x} + \mathbf{v}_i + \mathbf{v}_j$ and no subscript represent schedule \mathbf{x} . If y_t^1 satisfies the following inequality for any t ,*

$$y_{t,i}^1 + y_{t,j}^1 \geq y_t^1 + y_{t,ij}^1 \quad (\text{B.35})$$

then (12) holds for all feasible schedules.

Proof 4. *Proof of Lemma 5. The wait time of each component in all-show-up case can be listed as: $\Gamma_S = \sum_{t=1}^T y_t^S$, $\Gamma_W = \sum_{t=1}^T \mathbb{E}([y_t^I] - y_t^S)$, $\Gamma_E = \sum_{t=1}^T \mathbb{E}[y_t^{P+1} - y_t^I]$, $\Gamma_O = \mathbb{E}[y_T^1]$ and $\Gamma_I = T + \mathbb{E}[y_T^1 - \sum_{t=1}^T (x_t + \beta_t + \gamma_t)]$. It is easy to check that the objective function $f(\mathbf{x})$ is a linear combination of y_t^S , $\mathbb{E}[y_t^I]$, $\mathbb{E}[y_t^{P+1}]$, $\mathbb{E}[y_T^1]$ and $-x_t$ with non-negative weights (since $C_E \leq C_W \leq 1$). If for any t , we have $y_{t,i}^S + y_{t,j}^S \geq y_t^S + y_{t,ij}^S$, $\mathbb{E}[y_{t,i}^I] + \mathbb{E}[y_{t,j}^I] \geq \mathbb{E}[y_t^I] + \mathbb{E}[y_{t,ij}^I]$, $\mathbb{E}[y_{t,i}^{P+1}] + \mathbb{E}[y_{t,j}^{P+1}] \geq \mathbb{E}[y_t^{P+1}] + \mathbb{E}[y_{t,ij}^{P+1}]$, $\mathbb{E}[y_{t,i}^1] + \mathbb{E}[y_{t,j}^1] \geq \mathbb{E}[y_t^1] + \mathbb{E}[y_{t,ij}^1]$ and $-x_{t,i} - x_{t,j} \geq -x_t - x_{t,ij}$, then (12) holds for all feasible schedules. By Lemma 1, it is equivalent to show*

$$y_{t,i}^S + y_{t,j}^S \geq y_t^S + y_{t,ij}^S, \quad (\text{B.36})$$

$$y_{t,i}^I + y_{t,j}^I \geq y_t^I + y_{t,ij}^I, \quad (\text{B.37})$$

$$y_{t,i}^{P+1} + y_{t,j}^{P+1} \geq y_t^{P+1} + y_{t,ij}^{P+1}, \quad (\text{B.38})$$

$$y_{t,i}^1 + y_{t,j}^1 \geq y_t^1 + y_{t,ij}^1, \quad (\text{B.39})$$

$$-x_{t,i} - x_{t,j} \geq -x_t - x_{t,ij}, \quad (\text{B.40})$$

Since we have $-x_{t,i} - x_{t,j} = -x_t - x_{t,ij}$ by definition, thus, we only need to prove (B.36 - B.38).

When (B.39) ($y_{t,i}^1 + y_{t,j}^1 \geq y_t^1 + y_{t,ij}^1$) holds for any t , we can prove that (B.36-B.38) hold for any t . To be specific, recall Lindley equations (B.31-B.34). By setting $\beta_t = 0$ and $\gamma_t = 0$, we will have same formulations for y_t^S and y_t^1 . And by setting $\gamma_t = 0$ we will have same formulations for $y_{t,i}^1$, $y_{t,j}^1$ and $y_{t,ij}^1$. Thus, if $y_{t,i}^1 + y_{t,j}^1 \geq y_t^1 + y_{t,ij}^1$ holds for any t , then (B.36-B.38) hold for any t . Then, if $y_{t,i}^1 + y_{t,j}^1 \geq y_t^1 + y_{t,ij}^1$ holds for any t , (12) holds for all feasible schedules. \square

In the following content, we are going to prove that $y_{t,i}^1 + y_{t,j}^1 \geq y_t^1 + y_{t,ij}^1$ holds for every possible combination of i and j with $1 \leq i, j \leq T$.

Lemma 6 (Hajek, 1985). Consider the “splitting sequence” — a deterministic, 0-1 sequence $\mathbf{z} = (z_1, z_2, \dots, z_k, \dots)$. Define χ_k as a sequence of independent random variables. Then $n_{k+1} = (n_k + z_k - \chi_k)^+$ is multimodular on \mathbf{z} .

For any \mathbf{x} , we can obtain a corresponding \mathbf{z} by mapping x_t to $z_{i,t}$, i.e., $\mathbf{z} = (\dots, z_{1,t}, z_{2,t}, \dots, z_{N,t}, \dots)$ and $x_t = \sum_{i=1}^N z_{i,t}$. Let $\mathbf{n} = (\dots, n_{1,t}, n_{2,t}, \dots, n_{N,t}, \dots)$. Let $\chi_{i,t} = 0$ for $i \neq N$ and $\chi_{i,t} = (1 - \beta_t - \gamma_t)$ for $i = N$. Define $n_{i+1,t} = (n_{i,t} + z_{i,t} - \chi_{i,t})^+$ for $i \neq N$ and $n_{1,t+1} = (n_{N,t} + z_{N,t} - \chi_{N,t})^+$ for $i = N$. By Lemma 6, $n_{i,t}$ is multimodular. According to the Lindley equation of y_t^1 : $y_{t+1}^1 = [y_t^1 + x_t - (1 - \beta_t - \gamma_t)]^+$, we know $y_t^1 = n_{N,t}$. Thus y_t^1 also has the property of multimodularity.

Lemma 7 (Altman et al., 2000). g is multimodular if and only if

$$g(\mathbf{x} + \mathbf{v}_i) - g(\mathbf{x}) \geq g(\mathbf{x} + \mathbf{v}_j + \mathbf{v}_i) - g(\mathbf{x} + \mathbf{v}_j) \quad (\text{B.41})$$

for all $\mathbf{v}_i, \mathbf{v}_j \in \mathcal{V}$, $\mathbf{v}_i \neq \mathbf{v}_j$.

Since y_t^1 has the property of multimodularity on (x_1, \dots, x_t) , then by lemma 7, $y_{t,i}^1 + y_{t,j}^1 \geq y_t^1 + y_{t,ij}^1$ holds for any t . Now we can prove that (12) holds in the all-show-up case by Lemma 5.

Next, we prove that (12) holds with no-show behaviour. Note that x_t has been replaced by a random variable $\alpha_t(x_t)$ in (B.31 - B.34). We are going to prove that for any t , $\alpha_t(x_{t,i}) + \alpha_t(x_{t,j}) \geq \alpha_t(x_t) + \alpha_t(x_{t,ij})$ and inequalities (B.36 - B.39) hold. For the some slots t ($1 \leq t \leq T$) where \mathbf{v}_i and \mathbf{v}_j do not change the number of scheduled patients such that $x_t = x_t^i = x_t^j = x_t^{ij}$, inequalities (B.36 - B.39) hold as illustrated before. What we need to focus on is when the number of scheduled patients changes by \mathbf{v}_i and \mathbf{v}_j . We denote i (j) as the patient who changes his appointment slots in schedule $\mathbf{x} + \mathbf{v}_i$ ($\mathbf{x} + \mathbf{v}_j$) compared with schedule \mathbf{x} . Now we examine whether $\alpha_t(x_{t,i}) + \alpha_t(x_{t,j}) \geq \alpha_t(x_t) + \alpha_t(x_{t,ij})$ holds case by case.

(1) If both patient i and patient j do not show up, then the four schedules \mathbf{x} , $\mathbf{x} + \mathbf{v}_i$, $\mathbf{x} + \mathbf{v}_j$ and $\mathbf{x} + \mathbf{v}_i + \mathbf{v}_j$ are identical and the inequality holds.

(2) If patient i shows up and patient j does not show up, then $\mathbf{x} = \mathbf{x} + \mathbf{v}_j$, $\mathbf{x} + \mathbf{v}_i = \mathbf{x} + \mathbf{v}_i + \mathbf{v}_j$, then the inequality holds.

(3) If patient i and patient j both show up, the problem is reduced to the all-show-up case, then the inequality holds.

By considering these three cases of that results in $\alpha_t(x_{t,i}) + \alpha_t(x_{t,j}) \geq \alpha_t(x_t) + \alpha_t(x_{t,ij})$, we can easily prove the inequalities (B.36 - B.39) hold. Thus, we can say the objective function $f(\mathbf{x})$ with no-show behaviour is multimodular for all \mathbf{x} in \mathbb{Z}_+^T by Definition 1. Summarising the above analysis, we complete the proof of Proposition 1.

B.3.3 Proof of Proposition 2

To prove the equivalence between (P2) and (P1), we first establish the equivalence between the original formulation and the reformulated formulation. Next, we show that the reformulated version of equations (13-16), when applied with the positive part function, can be transformed into the set of inequalities represented by (23-26).

Denote \mathbf{a} , \mathbf{b} , and \mathbf{c} as the vectors where a_t , b_t and c_t as the number of show-up patients, in-clinic walk-ins and E-visits for $t = 1, \dots, T$. Our objective is to establish that, for all combinations of the values of a_t , b_t , and c_t , we have

$$\Pr\{\boldsymbol{\alpha}(\mathbf{x}, \omega) = \mathbf{a}, \boldsymbol{\beta}(\omega) = \mathbf{b}, \boldsymbol{\gamma}(\omega) = \mathbf{c}\} = \Pr\left\{\sum_{i=1}^N \delta_i(\omega) \mathbf{z}_i = \mathbf{a}, \boldsymbol{\beta}(\omega) = \mathbf{b}, \boldsymbol{\gamma}(\omega) = \mathbf{c}\right\} \quad (\text{B.42})$$

As events occurring in different slots are independent of others, E-visits, walk-ins and scheduled patients are independent, we have

$$\Pr\{\boldsymbol{\alpha}(\mathbf{x}, \omega) = \mathbf{a}, \boldsymbol{\beta}(\omega) = \mathbf{b}, \boldsymbol{\gamma}(\omega) = \mathbf{c}\} = \prod_{t=1}^T \Pr\{\alpha_t(x_t, \omega) = a_t\} \cdot \Pr\{\beta_t(\omega) = b_t\} \cdot \Pr\{\gamma_t(\omega) = c_t\} \quad (\text{B.43})$$

Due to Constraints of \mathbf{z} in (Sub.LP), we know that $\sum_{i=1}^N \delta_i(\omega) \mathbf{z}_i$, $\boldsymbol{\beta}(\omega)$ and $\boldsymbol{\gamma}(\omega)$ have no overlapping terms, and thus are independent. It follows that

$$\Pr\{\boldsymbol{\alpha}(\mathbf{x}, \omega) = \mathbf{a}, \boldsymbol{\beta}(\omega) = \mathbf{b}, \boldsymbol{\gamma}(\omega) = \mathbf{c}\} = \prod_{t=1}^T \Pr\left\{\sum_{i=1}^N \delta_i(\omega) z_{t,i} = a_t\right\} \cdot \Pr\{\beta_t(\omega) = b_t\} \cdot \Pr\{\gamma_t(\omega) = c_t\} \quad (\text{B.44})$$

For any t , with the Constraint $x_t = \sum_{i=1}^N z_{t,i}$ and $\alpha(x_t|\omega) = \sum_{i=1}^N \delta_i(\omega) z_{t,i}$, we have

$$\Pr\left\{\sum_{i=1}^N \delta_i(\omega) z_{t,i} = a_t\right\} = \Pr\{\alpha(x_t|\omega) = a_t\}. \quad (\text{B.45})$$

Now we need to prove the second part. By definition, y_t^S , y_t^I and y_t^p are the number of scheduled patients, in-clinic patients and in-clinic patients together with E-visits who have waited for p slots at the end of slot t . Look at the inequality (26), the difference between the RHS of (26) and the LHS of (26) is the number of patients to be served at t . Since this number cannot be greater than 1, then it must be 0 or 1 with an integer Constraint on y_t^p . Note that the objective function is increasing in y_t^{p+1} and y_t^1 , so in the second stage problem (Sub.LP), whether to serve the patient at each time period or not influence the total weighted waiting patients. It is easy to see that the optimal solution must be serving a patient if there is any. y_t^S and y_t^I have a similar situation that we can derive the same conclusion. In a nutshell, it is optimal to serve a patient which implies (P2) is equivalent to (P1).

B.3.4 Proof of Proposition 3

Let \mathbf{e}_t denote the t -th T -dimensional unit vector, i.e., $\mathbf{e}_t = \{0, \dots, 1, \dots, 0\}$ where the t -th coordinate is 1. Then $\mathbf{x} + j \times \mathbf{e}_t$ means scheduling j more appointments at slot t . $\mathbf{s} \vee \mathbf{e}_t$ means opening slot t for E-visits, and $(\mathbf{s} - \mathbf{e}_t)^+$ means closing slot t for E-visits.

Given the joint schedule (\mathbf{x}, \mathbf{s}) where $\mathbf{x} \in \mathbb{Z}_+^T$ and $\mathbf{s} \in \{0, 1\}^T$, $f(\mathbf{x}, \mathbf{s})$ is denoted as the objective function value, by $f(\mathbf{x}, \mathbf{s}) := \Gamma_S + C_W \Gamma_W + C_E \Gamma_E + C_I \Gamma_I + C_O \Gamma_O$. To proof the inequality in Proposition

3, similar to the proof sketch in Appendix B.3.2, we first prove that for the objective function for all \mathbf{x} , \mathbf{s} , $\mathbf{x} + j \times \mathbf{e}_{t^-}$, $\mathbf{s} \vee \mathbf{e}_t$ and $(\mathbf{s} - \mathbf{e}_t)^+$ in \mathbb{Z}_+^T in the all-show-up case where $\alpha_t(x_t) = x_t$ satisfy the following inequality

$$f(\mathbf{x} + \mathbf{e}_{t^-}, \mathbf{s} \vee \mathbf{e}_t) - f(\mathbf{x}, \mathbf{s} \vee \mathbf{e}_t) \geq f(\mathbf{x} + \mathbf{e}_{t^-}, (\mathbf{s} - \mathbf{e}_t)^+) - f(\mathbf{x}, (\mathbf{s} - \mathbf{e}_t)^+), \quad (\text{B.46})$$

for $\forall t \in [1, T]$ and $\forall t^- \in [1, t]$.

Let subscript $\{i, k\}$ represent the schedule $(\mathbf{x} + \mathbf{e}_{t^-})$ that adding one extra patient to slot t^- under schedule \mathbf{x} and E-visit time window is set to be available at t ; let subscript $\{0, k\}$ represent the schedule \mathbf{x} and E-visit time window is set to be available at k , subscript $\{i, \bar{k}\}$ represent adding one extra patient to slot i under schedule \mathbf{x} and E-visit time window is set to be unavailable at k , subscript $\{0, \bar{k}\}$ represent the schedule \mathbf{x} and E-visit time window is set to be unavailable at t . Let $\{I_x, I_s\}$ denote the pair of index, that is,

$$\{I_x, I_s\} = \begin{cases} \{i, k\} : & (\mathbf{x} + \mathbf{e}_{t^-}, \mathbf{s} \vee \mathbf{e}_t). \\ \{i, \bar{k}\} : & (\mathbf{x} + \mathbf{e}_{t^-}, (\mathbf{s} - \mathbf{e}_t)^+). \\ \{i, 0\} : & (\mathbf{x} + \mathbf{e}_{t^-}, \mathbf{s}). \\ \{0, k\} : & (\mathbf{x}, \mathbf{s} \vee \mathbf{e}_t). \\ \{0, \bar{k}\} : & (\mathbf{x}, (\mathbf{s} - \mathbf{e}_t)^+). \\ \{0, 0\} : & (\mathbf{x}, \mathbf{s}). \end{cases} \quad (\text{B.47})$$

Lemma 8. Let y_t^S , y_t^I , and y_t^P follow the Lindley equations (B.31), (B.32) and (B.34), with y_t^1 be redefined in (B.48) for $\forall t \in [1, T]$.

$$y_t^1 = (y_{t-1}^1 + x_t + \beta_t + \gamma_t s_t - 1)^+. \quad (\text{B.48})$$

with $y_0^1 = 0$. If y_t^1 satisfies the following inequality for any t ,

$$y_{t,\{i,k\}}^1 - y_{t,\{0,k\}}^1 \geq y_{t,\{i,\bar{k}\}}^1 - y_{t,\{0,\bar{k}\}}^1 \quad (\text{B.49})$$

then (B.46) holds for all feasible schedules.

Proof 5. Proof of Lemma 8. We can derive all the cost components of the model with a E-visit time window schedule: $\Gamma_S = \sum_{t=1}^{\bar{T}} y_t^S$, $\Gamma_W = \sum_{t=1}^{\bar{T}} (\mathbb{E}[y_t^I] - y_t^S)$, $\Gamma_E = \sum_{t=1}^{\bar{T}} \mathbb{E}[y_t^{P+1} - y_t^I]$, $\Gamma_O = \mathbb{E}[y_T^1]$ and $\Gamma_I = T + \mathbb{E}[y_T^1 - \sum_{t=1}^T (x_t + \beta_t + \gamma_t s_t)]$. Recall that the arrival pattern of E-visits is exogenously given and the distribution of γ_t is independent with \mathbf{s} , then we can show that the objective function $f(\mathbf{x}, \mathbf{s})$ of (P3) is a linear combination of y_t^S , $\mathbb{E}[y_t^I]$, $\mathbb{E}[y_t^{P+1}]$, $\mathbb{E}[y_T^1]$, $-x_t$ and $-s_t$ with non-negative weights (recall that $C_E \leq C_W \leq 1$). Following Appendix B.3.2 and Lemma 1, if for any t , we have

$$y_{t,\{i,k\}}^S - y_{t,\{0,k\}}^S \geq y_{t,\{i,\bar{k}\}}^S - y_{t,\{0,\bar{k}\}}^S, \quad (\text{B.50})$$

$$y_{t,i}^I - y_{t,j}^I \geq y_{t,\{i,\bar{k}\}}^I - y_{t,\{0,\bar{k}\}}^I, \quad (\text{B.51})$$

$$y_{t,\{i,k\}}^{P+1} - y_{t,\{0,k\}}^{P+1} \geq y_{t,\{i,\bar{k}\}}^{P+1} - y_{t,\{0,\bar{k}\}}^{P+1}, \quad (\text{B.52})$$

$$y_{t,\{i,k\}}^1 - y_{t,\{0,k\}}^1 \geq y_{t,\{i,\bar{k}\}}^1 - y_{t,\{0,\bar{k}\}}^1, \quad (\text{B.53})$$

$$-x_{t,\{i,k\}} + x_{t,\{0,k\}} \geq -x_{t,\{i,\bar{k}\}} + x_{t,\{0,\bar{k}\}}, \quad (\text{B.54})$$

then (B.46) holds for all feasible schedules.

Note that decision x_t is not influenced by the decision s_k , thus $-x_{t,\{i,k\}} + x_{t,\{0,k\}} = -x_{t,\{i,\bar{k}\}} + x_{t,\{0,\bar{k}\}}$, i.e., (B.54) holds.

If (B.53) $(y_{t,\{i,k\}}^1 - y_{t,\{0,k\}}^1 \geq y_{t,\{i,\bar{k}\}}^1 - y_{t,\{0,\bar{k}\}}^1)$ holds for any t , then we can prove (B.50) by setting $\beta_t = 0$ and $\gamma_t = 0$. Similarly, we can prove (B.51) and (B.52) hold by setting $\gamma_t = 0$.

Thus we have when $y_{t,\{i,k\}}^1 - y_{t,\{0,k\}}^1 \geq y_{t,\{i,\bar{k}\}}^1 - y_{t,\{0,\bar{k}\}}^1$ for any t , then (B.46) holds for all feasible schedules. \square

In the following content, we will show (B.53) $(y_{t,\{i,k\}}^1 - y_{t,\{0,k\}}^1 \geq y_{t,\{i,\bar{k}\}}^1 - y_{t,\{0,\bar{k}\}}^1)$ holds for any t . For ease of description, we replace the notation $y_{t,\{I_x, I_s\}}^1$ by $y_t^{I_x, I_s}$ and replace $y_t^{0,0}$ by y_t . Thus, we are going to show that the following inequality $y_t^{i,k} - y_t^{0,k} \geq y_t^{i,\bar{k}} - y_t^{0,\bar{k}}$ holds for any t with $1 \leq t \leq T$ and every possible combination of i and k with $1 \leq i \leq k \leq T$.

Case A: $t < i \leq k$. In this case, $y_t = y_t^{i,k} = y_t^{0,k} = y_t^{i,\bar{k}} = y_t^{0,\bar{k}}$, we have $y_t = (y_{t-1} + x_t + \beta_t + \gamma_t s_t - 1)^+$.

Case B: $t = i = k$. In this case, we have $y_t^{i,t} - y_t^{i,\bar{T}} = \gamma_t$ and $y_t^{0,t} - y_t^{0,\bar{T}} \leq \gamma_t$ by Lemma 3, thus $y_t^{i,t} - y_t^{0,t} \geq y_t^{i,\bar{T}} - y_t^{0,\bar{T}}$. Besides, we can show that $y_t^{i,t} \geq \{y_t^{0,t}, y_t^{i,\bar{T}}\} \geq y_t^{0,\bar{T}}$. We have $y_t^{i,t} = y_{t-1} + x_t + \beta_t + \gamma_t$, $y_t^{0,t} = (y_{t-1} + x_t + \beta_t + \gamma_t - 1)^+$, $y_t^{i,\bar{T}} = y_{t-1} + x_t + \beta_t$, $y_t^{0,\bar{T}} = (y_{t-1} + x_t + \beta_t - 1)^+$.

Case C: $t = i < k$. In this case, $y_t^{i,k} = y_t^{i,\bar{k}}$, $y_t^{0,k} = y_t^{0,\bar{k}}$; by Lemma 3, we have $y_t^{i,k} - y_t^{0,k} = y_t^{i,\bar{k}} - y_t^{0,\bar{k}}$. Specifically, $y_t^{i,k} = y_{t-1} + x_t + \beta_t + \gamma_t s_t$, $y_t^{0,k} = (y_{t-1} + x_t + \beta_t + \gamma_t s_t - 1)^+$, $y_t^{i,\bar{k}} = y_{t-1} + x_t + \beta_t + \gamma_t s_t$ and $y_t^{0,\bar{k}} = (y_{t-1} + x_t + \beta_t + \gamma_t s_t - 1)^+$.

Case D: $t = i + 1 = k$. In this case, since $y_{t-1}^{t-1,0} = y_{t-2} + x_{t-1} + \beta_{t-1} + \gamma_{t-1} s_{t-1}$ and $y_{t-1}^{0,t} = (y_{t-2} + x_{t-1} + \beta_{t-1} + \gamma_{t-1} s_{t-1} - 1)^+$, then $0 \leq y_{t-1}^{t-1,0} - y_{t-1}^{0,t} \leq 1$. By Lemma 4, we have $y_t^{i,t} - y_t^{0,t} \geq y_t^{i,\bar{T}} - y_t^{0,\bar{T}}$. Specifically, $y_t^{i,t} = (y_{t-1}^{t-1,0} + x_t + \beta_t + \gamma_t - 1)^+$, $y_t^{0,t} = (y_{t-1} + x_t + \beta_t + \gamma_t - 1)^+$, $y_t^{i,\bar{T}} = (y_{t-1}^{t-1,0} + x_t + \beta_t - 1)^+$, and $y_t^{0,\bar{T}} = (y_{t-1} + x_t + \beta_t - 1)^+$.

Case E: $t = i + 1 < k$. In this case, we have $y_t^{i,k} - y_t^{i,\bar{k}} = y_t^{0,k} - y_t^{0,\bar{k}}$. Specifically, $y_t^{i,k} = (y_{t-1}^{t-1,0} + x_t + \beta_t + \gamma_t s_t - 1)^+$, $y_t^{0,k} = (y_{t-1} + x_t + \beta_t + \gamma_t s_t - 1)^+$, $y_t^{i,\bar{k}} = (y_{t-1}^{t-1,0} + x_t + \beta_t + \gamma_t s_t - 1)^+$ and $y_t^{0,\bar{k}} = (y_{t-1} + x_t + \beta_t + \gamma_t s_t - 1)^+$.

Case F: $i + 1 < t = k$. In this case, we have $y_t^{i,t} - y_t^{0,t} = y_t^{i,\bar{T}} - y_t^{0,\bar{T}}$. Specifically, $y_t^{i,t} = (y_{t-1}^{i,t} + x_t + \beta_t + \gamma_t - 1)^+$, $y_t^{0,t} = (y_{t-1} + x_t + \beta_t + \gamma_t - 1)^+$, $y_t^{i,\bar{T}} = (y_{t-1}^{i,\bar{T}} + x_t + \beta_t - 1)^+$ and $y_t^{0,\bar{T}} = (y_{t-1} + x_t + \beta_t - 1)^+$.

Case G: $i + 1 < t < k$. In this case, we have $y_t^{i,k} - y_t^{i,\bar{k}} = y_t^{0,k} - y_t^{0,\bar{k}} = 0$. Specifically, $y_t^{i,k} = (y_{t-1}^{i,k} + x_t + \beta_t + \gamma_t s_t - 1)^+$, $y_t^{0,k} = (y_{t-1} + x_t + \beta_t + \gamma_t s_t - 1)^+$, $y_t^{i,\bar{k}} = (y_{t-1}^{i,\bar{k}} + x_t + \beta_t + \gamma_t s_t - 1)^+$, $y_t^{0,\bar{k}} = (y_{t-1} + x_t + \beta_t + \gamma_t s_t - 1)^+$.

Case H: $i < k < t$. In this case, we have $y_t^{i,k} - y_t^{i,\bar{k}} = y_t^{0,k} - y_t^{0,\bar{k}} = 0$. Specifically, $y_t^{i,k} = (y_{t-1}^{i,k} + x_t + \beta_t + \gamma_t s_t - 1)^+$, $y_t^{0,k} = (y_{t-1}^{0,k} + x_t + \beta_t + \gamma_t s_t - 1)^+$, $y_t^{i,\bar{k}} = (y_{t-1}^{i,\bar{k}} + x_t + \beta_t + \gamma_t s_t - 1)^+$, $y_t^{0,\bar{k}} = (y_{t-1}^{0,\bar{k}} + x_t + \beta_t + \gamma_t s_t - 1)^+$.

Case I: $i = k < t = i + 1$. In this case, following Case B, we have $y_t^{t-1,t-1} - y_{t-1}^{t-1,\bar{t}-1} \geq y_t^{0,t-1} - y_{t-1}^{0,\bar{t}-1}$, $y_t^{t-1,t-1} \geq y_{t-1}^{0,t-1}$ and $y_{t-1}^{t-1,\bar{t}-1} \geq y_{t-1}^{0,\bar{t}-1}$. By Lemma 5, we have $y_t^{t-1,t-1} - y_{t-1}^{t-1,\bar{t}-1} \geq y_t^{0,t-1} - y_{t-1}^{0,\bar{t}-1}$. Specifically, $y_t^{t-1,t-1} = (y_{t-1}^{t-1,t-1} + x_t + \beta_t + \gamma_t s_t - 1)^+$, $y_t^{0,t-1} = (y_{t-1}^{0,t-1} + x_t + \beta_t + \gamma_t s_t - 1)^+$, $y_{t-1}^{t-1,\bar{t}-1} = (y_{t-1}^{t-1,\bar{t}-1} + x_t + \beta_t + \gamma_t s_t - 1)^+$, $y_{t-1}^{0,\bar{t}-1} = (y_{t-1}^{0,\bar{t}-1} + x_t + \beta_t + \gamma_t s_t - 1)^+$.

Case J: $i = k < i + 1 < t$. In this case, we have $y_t^{i,k} - y_t^{i,\bar{k}} = y_t^{0,k} - y_t^{0,\bar{k}} = 0$. $y_t^{i,k} = (y_{t-1}^{i,k} + x_t + \beta_t + \gamma_t s_t - 1)^+$, $y_t^{0,k} = (y_{t-1}^{0,k} + x_t + \beta_t + \gamma_t s_t - 1)^+$, $y_t^{i,\bar{k}} = (y_{t-1}^{i,\bar{k}} + x_t + \beta_t + \gamma_t s_t - 1)^+$, $y_t^{0,\bar{k}} = (y_{t-1}^{0,\bar{k}} + x_t + \beta_t + \gamma_t s_t - 1)^+$.

Till now, we examine all the cases and show that $y_{t,\{i,k\}}^1 - y_{t,\{0,k\}}^1 \geq y_{t,\{i,\bar{k}\}}^1 - y_{t,\{0,\bar{k}\}}^1$ holds for any t . Thus, by Lemma 8, (B.46) holds for any schedule.

Then for $\forall t \in [1, T]$, $\forall j \geq 0$ and $\forall t^- \in [1, t]$, by (B.46), we have the following equation

$$\begin{aligned} f(\mathbf{x} + j \times \mathbf{e}_{t^-}, s \vee \mathbf{e}_t) - f(\mathbf{x} + (j-1) \times \mathbf{e}_{t^-}, s \vee \mathbf{e}_t) \\ \geq f(\mathbf{x} + j \times \mathbf{e}_{t^-}, (s - \mathbf{e}_t)^+) - f(\mathbf{x} + (j-1) \times \mathbf{e}_{t^-}, (s - \mathbf{e}_t)^+), \end{aligned} \quad (\text{B.55})$$

by regarding $\mathbf{x} + (j-1) \times \mathbf{e}_{t^-}$ as \mathbf{x}' . Thus we have

$$\begin{aligned} f(\mathbf{x} + j \times \mathbf{e}_{t^-}, s \vee \mathbf{e}_t) - f(\mathbf{x} + (j-1) \times \mathbf{e}_{t^-}, s \vee \mathbf{e}_t) \\ \geq f(\mathbf{x} + j \times \mathbf{e}_{t^-}, (s - \mathbf{e}_t)^+) - f(\mathbf{x} + (j-1) \times \mathbf{e}_{t^-}, (s - \mathbf{e}_t)^+), \\ f(\mathbf{x} + (j-1) \times \mathbf{e}_{t^-}, s \vee \mathbf{e}_t) - f(\mathbf{x} + (j-2) \times \mathbf{e}_{t^-}, s \vee \mathbf{e}_t) \\ \geq f(\mathbf{x} + (j-1) \times \mathbf{e}_{t^-}, (s - \mathbf{e}_t)^+) - f(\mathbf{x} + (j-2) \times \mathbf{e}_{t^-}, (s - \mathbf{e}_t)^+), \\ \dots \\ f(\mathbf{x} + 2 \times \mathbf{e}_{t^-}, s \vee \mathbf{e}_t) - f(\mathbf{x} + 1 \times \mathbf{e}_{t^-}, s \vee \mathbf{e}_t) \\ \geq f(\mathbf{x} + 2 \times \mathbf{e}_{t^-}, (s - \mathbf{e}_t)^+) - f(\mathbf{x} + 1 \times \mathbf{e}_{t^-}, (s - \mathbf{e}_t)^+), \\ f(\mathbf{x} + 1 \times \mathbf{e}_{t^-}, s \vee \mathbf{e}_t) - f(\mathbf{x}, s \vee \mathbf{e}_t) \\ \geq f(\mathbf{x} + 1 \times \mathbf{e}_{t^-}, (s - \mathbf{e}_t)^+) - f(\mathbf{x}, (s - \mathbf{e}_t)^+). \end{aligned}$$

Then we have the following inequality by adding up the LHS and the RHS of the inequalities above:

$$f(\mathbf{x} + j \times \mathbf{e}_{t^-}, s \vee \mathbf{e}_t) - f(\mathbf{x}, s \vee \mathbf{e}_t) \geq f(\mathbf{x} + j \times \mathbf{e}_{t^-}, (s - \mathbf{e}_t)^+) - f(\mathbf{x}, (s - \mathbf{e}_t)^+), \quad (\text{B.56})$$

To show the inequality holds when the objective function is incorporated with no-show behaviour, where the number of scheduled patients in t becomes a random variable $\alpha_t(x_t)$. If all patients added at i do not show up, then the LHS equals RHS; if more than one patient shows up at i , then the inequality holds. This completes the proof.

B.3.5 Proof of Corollary 2

Next, we prove the connection between the scheduled slots and the opening of E-visit time windows in the joint schedule optimization problem. Given that the inequality (B.56) holds true when both indices, i and j are equal to slot t , we can extend the validity of the summation.

Let j_t denote a non-negative integer. In the optimal joint schedule (\mathbf{x}, s) , we first prove for any t' , the following inequality:

$$f\left(\mathbf{x} + \sum_{t \in \mathcal{T}^-} (j_t \times \mathbf{e}_t), s \vee \mathbf{e}_{t'}\right) - f\left(\mathbf{x}, s \vee \mathbf{e}_{t'}\right) \geq f\left(\mathbf{x} + \sum_{t \in \mathcal{T}^-} (j_t \times \mathbf{e}_t), (s - \mathbf{e}_{t'})^+\right) - f\left(\mathbf{x}, (s - \mathbf{e}_{t'})^+\right), \quad (\text{B.57})$$

where $\mathcal{T}^- \subseteq [1, t']$.

With slight abuse of notations, let us renumber the indexes in \mathcal{T}^- as $t = 1, 2, 3, \dots, k'$.

By (B.56) and regarding $\mathbf{x} + \sum_{t=1}^{k'-1} (j_t \times \mathbf{e}_t)$ as \mathbf{x}' , we have

$$\begin{aligned} f\left(\mathbf{x} + \sum_{t=1}^{k'} (j_t \times \mathbf{e}_t), s \vee \mathbf{e}_{t'}\right) - f\left(\mathbf{x} + \sum_{t=1}^{k'-1} (j_t \times \mathbf{e}_t), s \vee \mathbf{e}_{t'}\right) \\ \geq f\left(\mathbf{x} + \sum_{t=1}^{k'} (j_t \times \mathbf{e}_t), (s - \mathbf{e}_{t'})^+\right) - f\left(\mathbf{x} + \sum_{t=1}^{k'-1} (j_t \times \mathbf{e}_t), (s - \mathbf{e}_{t'})^+\right), \end{aligned} \quad (\text{B.58})$$

for any $k' \leq t'$. Then we have

$$\begin{aligned}
 & f\left(\mathbf{x} + \sum_{t=1}^k (j_t \times e_t), s \vee e_{t'}\right) - f\left(\mathbf{x} + \sum_{t=1}^{k-1} (j_t \times e_t), s \vee e_{t'}\right) \\
 & \geq f\left(\mathbf{x} + \sum_{t=1}^k (j_t \times e_t), (s - e_{t'})^+\right) - f\left(\mathbf{x} + \sum_{t=1}^{k-1} (j_t \times e_t), (s - e_{t'})^+\right), \\
 & f\left(\mathbf{x} + \sum_{t=1}^{k-1} (j_t \times e_t), s \vee e_{t'}\right) - f\left(\mathbf{x} + \sum_{t=1}^{k-2} (j_t \times e_t), s \vee e_{t'}\right) \\
 & \geq f\left(\mathbf{x} + \sum_{t=1}^{k-1} (j_t \times e_t), (s - e_{t'})^+\right) - f\left(\mathbf{x} + \sum_{t=1}^{k-2} (j_t \times e_t), (s - e_{t'})^+\right), \\
 & \dots \\
 & f\left(\mathbf{x} + \sum_{t=1}^2 (j_t \times e_t), s \vee e_{t'}\right) - f\left(\mathbf{x} + j_1 \times e_1, s \vee e_{t'}\right) \\
 & \geq f\left(\mathbf{x} + \sum_{t=1}^2 (j_t \times e_t), (s - e_{t'})^+\right) - f\left(\mathbf{x} + j_1 \times e_1, (s - e_{t'})^+\right), \\
 & f\left(\mathbf{x} + j_1 \times e_1, s \vee e_{t'}\right) - f\left(\mathbf{x}, s \vee e_{t'}\right) \\
 & \geq f\left(\mathbf{x} + j_1 \times e_1, (s - e_{t'})^+\right) - f\left(\mathbf{x}, (s - e_{t'})^+\right).
 \end{aligned}$$

Now we have (B.57) by adding up the LHS and the RHS of the inequalities above.

Next, we prove

$$\begin{aligned}
 & f\left(\mathbf{x} + \sum_{t \in \mathcal{T}^-} (j_t \times e_t), s \vee \sum_{t \in \mathcal{T}^+} e_t\right) - f\left(\mathbf{x} + \sum_{t \in \mathcal{T}^-} (j_t \times e_t), (s - \sum_{t \in \mathcal{T}^+} e_t)^+\right) \geq \\
 & f\left(\mathbf{x}, s \vee \sum_{t \in \mathcal{T}^+} e_t\right) - f\left(\mathbf{x}, (s - \sum_{t \in \mathcal{T}^+} e_t)^+\right), \tag{B.59}
 \end{aligned}$$

for $\mathcal{T}^- \subseteq [1, t']$ and $\mathcal{T}^+ \subseteq [t', T]$.

With slight abuse of notations, let us renumber the indexes in \mathcal{T}^+ as $t = 1, 2, 3, \dots, k$.

By (B.57) and regarding $s \vee e_2$ as s' , we have

$$\begin{aligned}
 & f\left(\mathbf{x} + \sum_{t \in \mathcal{T}^-} (j_t \times e_t), (s \vee e_2) \vee e_1\right) - f\left(\mathbf{x} + \sum_{t \in \mathcal{T}^-} (j_t \times e_t), ((s \vee e_2) - e_1)^+\right) \geq \\
 & f\left(\mathbf{x}, (s \vee e_2) \vee e_1\right) - f\left(\mathbf{x}, ((s \vee e_2) - e_1)^+\right),
 \end{aligned}$$

Regarding $(s - e_1)^+$ as s' , we have

$$\begin{aligned}
 & f\left(\mathbf{x} + \sum_{t \in \mathcal{T}^-} (j_t \times e_t), (s - e_1)^+ \vee e_2\right) - f\left(\mathbf{x} + \sum_{t \in \mathcal{T}^-} (j_t \times e_t), ((s - e_1)^+ - e_2)^+\right) \geq \\
 & f\left(\mathbf{x}, (s - e_1)^+ \vee e_2\right) - f\left(\mathbf{x}, ((s - e_1)^+ - e_2)^+\right),
 \end{aligned}$$

It is easy to check that $(s - e_1)^+ \vee e_2 = ((s \vee e_2) - e_1)^+$, $(s \vee e_2) \vee e_1 = s \vee \sum_{t=1}^2 e_t$ and $((s - e_1)^+ - e_2)^+ = (s - \sum_{t=1}^2 e_t)^+$. Thus, by adding up the LHS and RHS of the inequalities above, we have

$$\begin{aligned} f\left(x + \sum_{t \in \mathcal{T}^-} (j_t \times e_t), s \vee \sum_{t=1}^2 e_t\right) - f\left(x + \sum_{t \in \mathcal{T}^-} (j_t \times e_t), (s - \sum_{t=1}^2 e_t)^+\right) \geq \\ f\left(x, s \vee \sum_{t=1}^2 e_t\right) - f\left(x, (s - \sum_{t=1}^2 e_t)^+\right). \end{aligned}$$

Replicating the above procedure by regrading $\sum_{t=1}^2 e_t$ as e'_1 and regarding e_3 as e'_2 , we will have

$$\begin{aligned} f\left(x + \sum_{t \in \mathcal{T}^-} (j_t \times e_t), s \vee \sum_{t=1}^3 e_t\right) - f\left(x + \sum_{t \in \mathcal{T}^-} (j_t \times e_t), (s - \sum_{t=1}^3 e_t)^+\right) \geq \\ f\left(x, s \vee \sum_{t=1}^3 e_t\right) - f\left(x, (s - \sum_{t=1}^3 e_t)^+\right). \end{aligned}$$

By induction, we will have (B.59) hold.

This completes the proof.

B.3.6 Proof of Corollary 3

It directly follows Proposition 2.