European Journal of Operational Research 258 (2017) 3–34

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Invited Review

Outpatient appointment systems in healthcare: A review of optimization studies

Amir Ahmadi-Javid[a,*], Zahra Jalali[b], Kenneth J Klassen[c]

[a] Department of Industrial Engineering, Amirkabir University of Technology, Tehran, Iran
[b] Department of Industrial Engineering, Sharif University of Technology, Tehran, Iran
[c] Goodman School of Business, Brock University, St. Catharines, Canada

ARTICLE INFO

Article history:
Received 26 May 2015
Accepted 28 June 2016
Available online 5 July 2016

Keywords:
Healthcare management
Appointment scheduling/sequencing
Service operations management
Operations research
Stochastic models

ABSTRACT

Outpatient appointment scheduling problems have recently gained increasing attention. This paper provides a comprehensive review of recent analytical and numerical optimization studies that present decision-support tools for designing and planning outpatient appointment systems (OASs). A structure for organizing the recent literature according to various criteria is provided, including a framework that classifies decisions at the strategic, tactical, and operational levels. The OAS literature is evaluated from four perspectives: problem settings, environmental factors, modeling approaches, and solution methods. In addition, research gaps and areas of opportunity for future research are discussed.

1. Introduction

Developing efficient healthcare systems has become more important in the last few decades for two major reasons: 1) the rapid increase in healthcare expenditures in more developed countries, and 2) the simultaneous growth of demand for healthcare services and patients' expectations of service quality (Hulshof, Kortbeek, Boucherie, Hans, & Bakker, 2012). As a result, governments and healthcare decision makers are constantly seeking to develop more efficient healthcare systems. A popular field for such development is operations research (OR), which provides numerous methodologies and solution techniques to simultaneously reduce costs and improve access to healthcare services. Prior reviews of OR in healthcare include Brailsford and Vissers (2011), Hulshof, Kortbeek et al. (2012), and Rais and Viana (2011).

In recent years, outpatient clinics have become more central in healthcare systems due to an emphasis on preventive medical practices, shorter hospital stays, and service provision on an outpatient basis (Cayirli & Veral, 2003). Appointment systems (ASs) are important components for efficient care delivery in outpatient clinics.

Outpatient appointment system (OAS) problems are an attractive research area, having been studied for more than half a century (since the seminal paper by Bailey (1952)). A review of the literature and of open research questions related to OAS problems can be found in Cayirli and Veral (2003) and Gupta and Denton (2008). The main goal of Cayirli and Veral (2003) is to review prior formulations and modeling considerations for outpatient appointment scheduling, whereas Gupta and Denton (2008) focus on describing the most common types of healthcare appointment systems, paying particular attention to the factors complicating OAS planning, and listing research opportunities. The current paper adds to these by updating the literature review and introducing a broader framework which is organized according to whether the OAS decisions – made in designing and planning OASs – are strategic, tactical, or operational. We also discuss the most common modeling approaches, solution methods, and environmental factors (patient unpunctuality, physician lateness, interruption, patient no-show and cancellation, patient preference, random service time, patient heterogeneity, and type of appointment required by patients) in the OAS literature.

The focus of this review is on recent papers which provide optimization-based decision tools for OAS decision makers. This review covers a wide range of analytical and numerical optimization studies, including studies using simulation optimization. Papers that do not provide a decision support tool (e.g., simulations and case studies) are not included. While there would be benefits to including all of the recent OAS research, the recent work is so extensive that it would be impossible to review it all in one paper. By focusing on optimization studies, we are able to provide a comprehensive review of these papers. However, it will be useful to have a separate literature review that includes simulations and/or case studies, for comparison purposes with the results of this

* Corresponding author.
E-mail address: ahmadi_javid@aut.ac.ir (A. Ahmadi-Javid).

http://dx.doi.org/10.1016/j.ejor.2016.06.064
0377-2217/

Table 1
The frequencies of papers on OASs optimization studies by year.

Year	Number of papers	Percentage
2003–2005	3	2.7
2006–2008	11	9.9
2009–2011	16	14.4
2012–2016 (June)	81	73
Total	111	100

paper. This will especially be the case if this other review is based on the framework presented here. One can also read a recent review by Günal and Pidd (2010), which surveys discrete-event simulation studies in healthcare, including OASs.

Forecasting techniques are a cornerstone for most OAS decisions (see Ozcan (2009) for a brief discussion of forecasting techniques used in healthcare). Almost all optimization studies use the outputs of these techniques to predict demand, no-shows, and cancellations. Although a number of papers discuss demand forecasting in some forms, and while some OAS papers focus entirely on predicting demand, we do not focus on that aspect in this review, as another full paper would likely be required to comprehensively cover forecasting in OASs.

According to Gupta and Denton (2008), healthcare ASs can be divided into three categories with respect to their environments: 1) primary care, 2) specialty care, and 3) elective (deferrable) surgical care. This review focuses on the first two categories, since the environment of elective surgical care is substantially different. Unlike primary care and specialty care, which are normally provided for outpatients, elective surgical care can be performed either on an inpatient or on an outpatient basis (Gupta & Denton, 2008). Moreover, surgery scheduling requires various human resources (e.g., surgeons, anesthetists, and nurses) and facilities (e.g., pre-operative holding units, operating theaters, and post-anesthesia care units) (Guerriero & Guido, 2011). Most studies related to surgery scheduling focus on staff scheduling and resource planning. For more details, one can refer to comprehensive reviews of studies on operating theater management and surgery scheduling (Cardoen, Demeulemeester, & Beliën, 2010; Guerriero & Guido, 2011; May, Spangler, Strum, & Vargas, 2011).

We restrict our attention to papers written in English and published in 2003 or later, since pre-2003 scientific papers were reviewed by Cayirli and Veral (2003). The search strategy for this study is based on selected databases: *Scopus, Web of Science, and PubMed*, and selected keywords: ("appointment" OR "outpatient appointment") AND ("system" OR "scheduling"). For any article found, the complete reference list was checked, and then the search process continued. Afterward, the above-mentioned restrictions were applied. Table 1 provides a summary of the resulting papers and demonstrates that the number of papers published on OASs continues to grow significantly: 73 percent of papers were published after 2011.

Occasionally, the terminology used in the literature lacks precision. Therefore, a list of terms along with the most commonly used definitions is provided in Appendix A.

Although ASs can be used in any setting where appointment times are scheduled for a set of customers and a service provider (e.g., patients and a medical practitioner, doctors and an operating room, clients and a consulting professional (lawyer or accountant), automobiles and a service center, tractor-trailers and a receiving bay, legal cases and a courtroom, or students and a professor; see Robinson & Chen (2003)), the literature mostly focuses on healthcare ASs (or OASs). The reason may be that these ASs are the most challenging ASs because of their characteristics (e.g., uncertainty, importance, and high demand compared to available capacity). Moreover, many (web-based or installed) appointment scheduling or medical scheduling software products with different features have been developed recently (links to some of these software products can be found in websites, such as capterra.com and softwareadvice.com). There has not been a study that reviews to what extent current academic research in ASs is used to enhance these software products; therefore, this is a potential research area.

In the operations management field, games are a common method to facilitate the learning of complex concepts such as decision making under uncertainty (Griffin, 2007). Classroom games can be used to reveal realistic challenges in ASs for students and beginners in healthcare scheduling; Saure and Puterman (2014) describe such a game. Game development for ASs is a non-traditional avenue of research that may benefit both learning and practice.

The remainder of the paper is organized as follows: Section 2 presents an introduction and overview of OAS problems. Sections 3, 4, and 5 discuss the strategic, tactical, and operational decisions found in the literature, respectively. Section 6 surveys the environmental factors influencing OASs. Section 7 reviews the optimization models and solution methods used in the literature. A discussion and future research directions are provided in Section 8. In addition, a glossary of technical terms, a considerable amount of information on individual papers that were reviewed, and overall summaries are presented in Appendices A, B, and C, respectively.

2. Overview of OAS problems

In this section, the classification of OAS decisions that is used to organize the literature is explained. Second, an overall comparison of OAS decisions is drawn based on this classification. Finally, a review of the objective functions and performance criteria used to evaluate OASs is presented.

2.1. Classification of OAS decisions/settings

Decisions made to design and plan OASs can be divided into three categories: strategic, tactical, and operational. Strategic (or design) decisions are long-term decisions that determine the main structure of an OAS. Tactical decisions are medium-term decisions related to how patients as a whole are scheduled, or how groups of patients are processed. Operational decisions are short-term and are concerned with efficiently scheduling individual patients. Table 2 shows the OAS decisions identified in this study; definitions are provided in Sections 3–5. It should be noted that there are a few unique or rarely used OAS decisions - such as the physician scheme design, the maximum number of appointments in each slot, and the panel composition, which are not listed in Table 2. In Fig. 1, Table 3, and Table B.2 in Appendix B, which summarizes all reviewed papers, these other strategic, tactical, and operational decisions are coded by "SO", "TO", and "OO", respectively.

Previous studies classify decisions in various ways. Cayirli and Veral (2003) classify OAS design decisions into three major groups: appointment rule, patient classification, and adjustments. Hulshof, Kortbeek et al. (2012) present a classification to review the operations research and management science (OR/MS) studies related to planning decisions in healthcare. For ambulatory care services, they list seven key decisions: number of patients per consultation session, patient overbooking, length of the appointment interval, number of patients per appointment slot (i.e., block size), sequence of appointments, queue discipline in the waiting room, and anticipation for unscheduled patients. Wang and Gupta (2011) divide OAS decisions into two categories: clinic profile setup and appointment booking, based on a two-stage process that is usually used in outpatient clinics. These categorizations are used in our proposed classification. Some of the OAS decisions listed in Table 2, particularly at the strategic level, have not been considered in previous

Table 2
Classification of OAS decisions/settings (other possible strategic, tactical, and operational decisions that are not listed in this table are coded by "SO", "TO", and "OO", respectively).

	Decision level	Code	Name (in alphabetical order)
Design decisions	Strategic	S1	Access policy
		S2	Number of servers/resources
		S3	Policy on acceptance of walk-ins
		S4	Type of scheduling
Planning decisions	Tactical	T1	Allocation of capacity to patient groups
		T2	Appointment interval (slot)
		T3	Appointment scheduling window
		T4	Block size
		T5	Number of appointments in consultation session
		T6	Panel size
		T7	Priority of patient groups
	Operational	O1	Allocation of patients to servers/resources
		O2	Appointment day
		O3	Appointment time
		O4	Patient acceptance/rejection
		O5	Patient selection from waiting list
		O6	Patient sequence

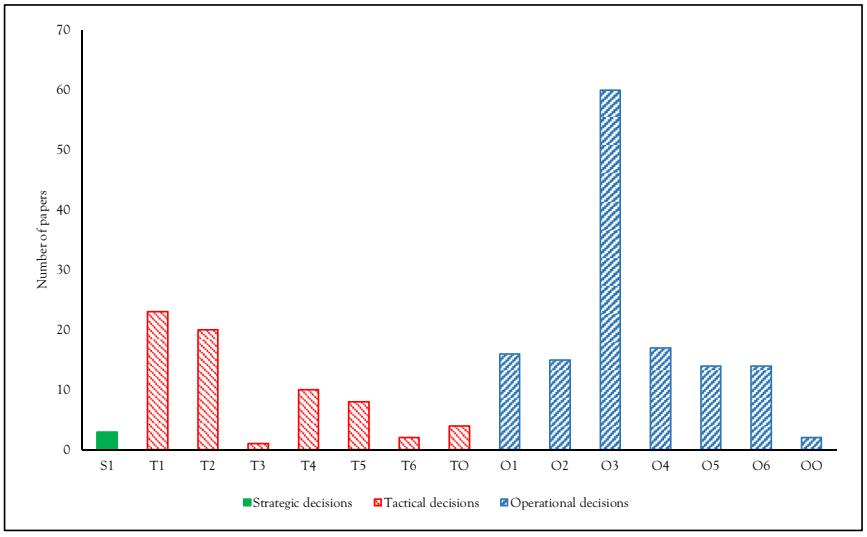

Fig. 1. The frequencies of papers on different decisions in the OAS literature (the decision codes are defined in Table 2, and there is no optimization study on decisions S2, S3, S4, SO, and T7).

literature reviews. Moreover, the hierarchical structure used here is helpful to demonstrate existing gaps in the literature.

In addition, we also catalog papers according to other criteria. There are a number of environmental factors that impact OASs, and several research studies develop methods for mitigating the negative impacts of these factors. The work on these factors is summarized in Section 6. Moreover, the various modeling approaches and solution methods used in prior optimization work are summarized in Section 7.

2.2. Overview of literature

There are only a few studies that compare strategic-level options (all for access policy (S1)) based on optimization models. Almost all optimization studies which focus on tactical and operational decisions assume that strategic decisions are set in advance. In Table B.2 in Appendix B, the settings considered for strategic decisions are given for each paper. As shown in Fig. 1, operational decisions have received more attention than tactical decisions in the OAS literature. At the tactical level, researchers seem most interested in appointment interval (slot), block size, and capacity allocation to patient groups; this is probably because these are the most common tactical decisions in an outpatient clinic, while other tactical decisions are often made on an experimental basis by clinic managers. At the operational level, the literature on the appointment time decision constitutes more than half the studies.

Some papers consider multiple decisions. The proposed models in these studies can be categorized into two classes: integrated models where decisions are considered simultaneously, and sequential models that solve the considered problems sequentially and separately. The distribution of the reviewed multi-decision papers with respect to the type of modeling used and the decisions addressed is depicted in Table 3. No explicitly OR contributions were found that take more than four decisions into account; thus, developing novel multi-decision approaches to better address real-life situations is an open research area. However, methodological limitations will undoubtedly restrict this direction. It should also be noted that a research article cannot be expected to incorporate all of the decisions discussed in this paper for two reasons. First, in an OAS with particular strategic settings, only some tactical and operational decisions are meaningful. To better understand which

Table 3
The frequencies of papers on multiple decisions in the OAS literature.

Decision	Integrated	Sequential	Decision	Integrated	Sequential
T1/T4	1	0	T1/T5	1	0
T4/T5	1	0	T5/T6	1	0
T5/TO	1	0	T6/TO	1	0
T1/O1	1	0	T1/O4	1	0
T1/O5	3	0	T2/O3	14	0
T4/O3	4	0	T5/O1	0	1
O1/O3	5	1	O2/O4	1	0
O2/O5	1	0	O3/O4	5	0
O3/O6	4	1	O4/O5	1	0
T1/O2/O4	2	0	(T2/O3)/O6	2	2
T1/O2/O5	1	0	T1/(T4/O3)	2	1
T4/O3/O6	1	0	T5/O2/O5	1	0
T2/T5/O3	1	0	(O1/O3)/O2	0	1
O1/O3/O4	0	1	O1/O3/O4	2	0
O1/O3/O6	1	0	O2/O3/O4	1	0
O2/O3/O5	2	0	O1/O6/OO	1	0
O1/O2/O3/O5	1	0	O1/O2/O5/OO	1	0

decisions are valid under specific strategic OAS settings, see Table C.1 in Appendix C. Second, a choice made for one decision will sometimes determine another decision. For instance, in the slot scheduling method if the appointment intervals and block sizes are determined, the appointment times are evident when the patient sequence is given or when patients are homogenous.

2.3. Objective functions and performance measures

OAS problems are characterized by three main stakeholder groups: patients, system owners, and staff. These stakeholders often have conflicting objectives (for instance, patient satisfaction and system utilization can be opposing goals); therefore, performance measures often include metrics pertinent to the three groups. Some of these are listed and explained by Cayirli and Veral (2003). Mondschein and Weintraub (2003) also propose an economic framework for a critical analysis of the objective functions used in the AS literature, which is broader than OAS. It should be noted that the existing optimization studies do not differentiate between system owners and staff; however, these groups may have different goals (e.g., system profit versus staff work load).

The most common performance measures used in optimization studies are the patient waiting time, server idle time, system overtime, number of patients seen, and number of rejected (or deferred) patients. In almost all of the studies, the patient satisfaction is measured through waiting time. Although the indirect waiting time (i.e., the time between the patient's call for an appointment and the scheduled appointment time) is frequently recommended as a direction for future research in earlier studies (e.g., Gupta & Denton, 2008), it is taken into account in the objective functions by a few recent papers (e.g., Erdogan, Gose, & Denton, 2015; Liu, Ziya, & Kulkarni, 2010). While other criteria, such as the continuity of care and matching of appointments to patient preferences can be used to evaluate patient satisfaction in addition to time-based measures, they have been studied very little due to the complex calculations required.

To calculate the revenue of an OAS, most optimization studies use the number of patients seen as a surrogate measure. Some studies also include a penalty for deferring patients (e.g., Green, Savin, & Wang, 2006; Kolisch & Sickinger, 2008; Schuetz & Kolisch, 2012).

Many studies assume that the physician is free to leave the clinic after seeing all of the patients scheduled for a given day. Nonetheless, in some cases, the physician is obliged to remain until the end of the predetermined consultation session, which is referred to as a no-golf policy in the literature (Tang, Yan, & Fung, 2014). Under this policy the physician idle time cost increases. In addition, the acceptance of late walk-ins may increase the physician overtime cost (Chen & Robinson, 2014). Tag et al. (2014) and Chen and Robinson (2014) compare their results under the no-golf policy.

In addition to Table B.2 in Appendix B, where the objective functions are presented, Table C.2 in Appendix C depicts the distribution of the papers based on performance measures. In the latter table, performance criteria that are rarely used in optimization studies are classified under the "Other" group. This group of criteria includes measures, such as the number of patients exceeding waiting-time targets (e,g., Castro & Petrovic, 2012; Gocgun & Puterman, 2014), congestion (e.g., Lin, 2015), unfairness (e.g., Turkcan, Zeng, Muthuraman, & Lawley, 2011; Qi, 2016), and continuity of care (e.g., Balasubramanian, Biehl, Dai, & Muriel, 2014; Gupta & Wang, 2008). Qi (2016) proposes a new threshold-based performance measure, called the delay unpleasantness, to describe the dissatisfaction of both patients and physician by taking into account the frequency and intensity of patient waiting time and the physician overtime above predetermined patient and physician thresholds. In a multi-objective setting, she then uses the concept of lexicographic min-max fairness to improve fairness in the AS design. Yan, Tang, Jiang, and Fung (2015) consider service fairness as a constraint that is formulated as the difference between the maximum and minimum average waiting time between intervals.

Almost all of the studies that consider several performance criteria in their modeling use the conventional weighted sum approach to obtain a single objective function. Kemper, Klaassen, and Mandjes (2014), Kuiper, Kemper, and Mandjes (2015), Ge, Wan, Wang, and Zhang (2013), and Vink, Kuiper, Kemper, and Bhulai (2015) use different disutility functions for patient waiting times and server idle times to have more balanced objective functions. It is also worth noting that similar objective functions are used for both tactical and operational decisions; for example, for capacity allocation (tactical) and patient scheduling (operational), it is common to use similar objective functions (e.g., a weighted sum of the waiting time and overtime).

In general, three directions for future research related to objective functions can be proposed. First, the linear relationship between time-based measures and their cost can be relaxed. For example, the cost per unit of waiting time when a patient waits one minute is often not the same as when a patient waits thirty minutes (this is also recommended by Cayirli & Veral (2003)). Second, the Pareto approach, which provides a set of non-dominant (Pareto optimal) solutions, is used in multi-criteria OAS problems by only a few papers (e.g., Castro & Petrovic, 2012; Qu, Rardin, & Williams, 2012), and therefore it can be considered in the future. Third, risk-averse objectives can be used instead of risk-neutral objectives, which are the mean of an OAS performance measure, in order to control the variability of the target performance measures. A few optimization studies use older methods to propose risk-averse objectives, such as the Markowitz mean-variance method (e.g., Mak, Rong, & Zhang, 2014a; Qu et al., 2012) and Von Neumann–Morgenstern expected utility method (e.g., Kemper et al., 2014; Kuiper et al., 2015; LaGanga & Lawrence, 2012; Vink et al., 2015).

3. Strategic decisions

As stated above, strategic decisions are usually regarded as inputs. Only a few optimization studies numerically compare strategic-level options for access policy. Therefore, strategic deci-

sions represent an area of open research questions that can be addressed in the future.

3.1. Access policy

One way to classify patients is to separate them based on scheduled patients versus walk-in patients (walk-ins) – at times the tactical and operational implications are very different for the two. In this section, we consider access policies for scheduled patients; walk-in patients are discussed in Section 3.3. Scheduled patients, who make an appointment before arriving at the clinic, can be divided into two subclasses: pre-scheduled patients, who are scheduled in advance of their appointment days, and same-day (open-access) patients, who are scheduled on the same day that they call for an appointment.

The access policy of an OAS determines which types of appointments (same-day and pre-scheduled) are scheduled. According to the relevant literature, there are three major types of access policies: traditional, open-access (also called advanced access), and hybrid (Robinson & Chen, 2010). A traditional policy books all patients in advance (i.e., all capacity is assigned to pre-scheduled patients). No-show rates are often higher under such a policy, mainly because of the longer indirect waiting time (Qu, Rardin, Williams, & Willis, 2007). An open-access policy, which assigns all capacities to same-day appointments, has been proposed to avoid the negative effects of the high no-show rate under a traditional policy and to serve today's demand today (Murray & Tantau, 2000). However, daily fluctuations in patient demand – which may lead to poor resource utilization – and patient appointment-booking preferences are barriers to the extensive use of open-access policies. A hybrid policy is a combination of both open-access and traditional policies (i.e., accommodating both same-day and pre-scheduled appointments). Other policies are also considered by the literature; for example, a carve-out policy, which reserves a certain amount of capacity for specific procedures (often urgent services) (Wang & Gupta, 2011); and a same-day-or-next-day policy, which books today's patients either on the day they call for an appointment or on the following day (Robinson & Chen, 2010).

Robinson and Chen (2010) compare the performance of open-access and traditional policies when the number of appointments are given, service time is deterministic, and patients are punctual. Their numerical analysis reveals that the open-access policy (and the corresponding same-day-or-next-day policy) significantly outperforms the traditional policy in most cases. Dobson, Hasija, and Pinker (2011) also numerically compare a traditional policy that reserves a fraction of the daily capacity for urgent patients (i.e., a carve-out policy) with an open-access policy; they find that the open-access policy performs better when urgent walk-in patients exist in the system. Gupta and Wang (2008) show that open-access systems perform worse when there is greater variability of same-day demand or greater positive dependence among the same-day demands for different physicians. Patrick (2012) presents a specific hybrid policy in which the ratio of the number of same-day patients to the number of pre-scheduled patients can be dynamically modified in each time period (i.e., consultation session), after today's demand has arrived. He also shows that this access policy, combined with overbooking to mitigate the impact of no-shows, outperforms other access policies. Qu, Rardin, and Williams (2011) develop an optimization model to compare a single-period hybrid policy with a two-period hybrid policy when the number of appointments is given. Their numerical results reveal that the single-period hybrid policy is not worse than the two-period hybrid policy where the objectives are to maximize the expectation of the number of patients scheduled and to minimize the variance of the number of patients scheduled.

3.2. Number of servers/resources

An important step in the design of an OAS is to determine how many servers/resources should be selected to meet patient demand. The major resources that have been considered by previous studies include clinic staff (i.e., physicians, nurses, and other medical specialists), medical equipment (e.g., medical imaging machines), rooms, and beds/chairs (e.g., infusion seat) (e.g., Turkcan et al., 2012; Riise, Mannino, & Lamorgese, 2016). Studies that consider equipment in their modeling make use of different assumptions which are based on the particular environment they are getting data from. Some of these are as follows: different sets of identical machines (Castro & Petrovic, 2012), only one set of identical machines (Gocgun, Bresnahan, Ghate, & Gunn, 2011; Saure, Patrick, Tyldesley, & Puterman, 2012), and different sets with only one machine (Pérez, Ntaimo, Wilhelm, Bailey, & McCormack, 2011; Pérez, Ntaimo, Malavé, Bailey, & McCormack, 2013).

The number of servers and level of resources, especially in specialty clinics, are a key factor in the clinic profit. A high level of resources may increase costs due to under-utilization, but insufficient resources may increase service delays, and thus decrease the quality of care and clinic revenue. Although optimizing the number of service providers (or the level of other resources) can be beneficial for outpatient clinics, this number is given as an input in the mathematical models reviewed. Therefore, this strategic decision would be a worthwhile direction for future research.

3.3. Policy on acceptance of walk-in patients (walk-ins)

Walk-in patients (walk-ins) are patients who arrive at the clinic without an appointment during the consultation session. Two major walk-in patient classes are considered in the literature: urgent and regular (Cayirli & Veral, 2003). The urgent walk-in patients (urgent walk-ins) often need to be treated as soon as possible, whereas regular walk-ins have a lower priority in the system; they are placed in slots left open for potential walk-ins or must wait for no-show slots. It should be noted that walk-in patients are different from same-day (open-access) patients, who are scheduled and served on the same day (for a detailed discussion, see Robinson & Chen (2010)).

Accepting walk-ins is a way to reduce the negative effects of no-shows (Gupta & Denton, 2008). Simultaneously considering both no-shows and walk-ins improves the productivity of OASs (Begen & Queyranne, 2011; El-Sharo, Zheng, Yoon, & Khasawneh, 2015; Green et al., 2006; Liu & Ziya, 2014; Luo, Kulkarni, & Ziya, 2012; Qu, Peng, Shi, & LaGanga, 2015). Cayirli and Yang (2014) suggest a procedure that adjusts the mean and standard deviation of service times based on the probabilities of no-shows and walk-ins in order to explicitly minimize the disruptive effects of no-shows and walk-ins, which is relevant to any appointment rule.

Recently, Qu et al. (2015) show that admitting all walk-ins is a simple and effective rule in clinics if the walk-in rate is less than 20 percent of the service rate. If the walk-in rate is greater than 20 percent, walk-ins should only be admitted when the number of patients waiting for service plus the expected number of punctual pre-scheduled patients is less than the total capacity of remaining slots.

Cayirli and Gunes (2013) focus on the capacity allocation problem with seasonal walk-ins. They analyze different adjustment policies for a weekly or a monthly walk-in seasonality. Their results show that a combined adjustment addressing both monthly and weekly seasonality outperforms other policies. Koeleman and Koole (2012) study the appointment scheduling problem considering urgent walk-ins that arrive following a non-stationary Poisson process. Their analysis show that it is the best to leave slots open for urgent patients toward the end of the consultation session.

The presence of walk-in patients leads to increased modeling complexity because of the dynamic stochastic arrivals of walk-in patients. This may be the reason that a majority of authors do not include them in their models. Therefore, further optimization studies with walk-ins could be beneficial.

3.4. Types of scheduling

Two scheduling approaches can be considered in OASs: online (i.e., sequential) and offline (i.e., simultaneous). In the offline approach, appointments are scheduled after all requests have arrived, while in the online approach, patients are scheduled immediately upon the arrival of their request (Zacharias & Pinedo, 2014). Pérez et al. (2013) formulate the problem of scheduling patients and resources in three forms: offline, online, and stochastic online. Stochastic online scheduling is an extension of online scheduling, which takes into consideration the potential patient requests that may arise after the current patient's request.

Online systems are more common in practice, while the offline approach has received more attention in the literature (offline systems are easier to model). Kuiper et al. (2015) compare the online and offline approaches with quadratic and linear loss functions. They observe that the online approach favors the server for both loss functions. Some studies use the results obtained from an offline case to examine the corresponding online case (e.g., Zacharias & Pinedo, 2014).

As electronic OASs are developing rapidly (Wang & Gupta, 2011; Weiner et al., 2009), the importance of offline scheduling is growing, since requests are collected via an IT tool (e.g., electronic mail or Web-based portal), and patients are later advised of their appointment times. Online systems have recently been the subject of some studies, but more work, especially considering realistic arrival patterns and future arrivals, could be of value.

4. Tactical decisions

The main aim of problems dealing with decisions at the tactical level (also referred to as the clinic profile setup; see Wang & Gupta (2011)) is to determine the OAS characteristics in order to maximize the resource utilization and the accessibility of care services. This section covers OAS tactical decisions in the next subsections.

4.1. Allocation of capacity to different patient groups

Capacity allocation deals with the problem of how the available capacity should be divided among different patient groups (e.g., first-time or returning patients, and pre-scheduled or same-day patients). The various patient classifications are discussed in Section 6.7. Sometimes these classifications involve patients with differing priorities (discussed in Section 4.7).

Creemers, Beliën, and Lambrecht (2012) present an optimization procedure for a capacity allocation problem to minimize the total weighted waiting time, which is calculated by a bulk service queuing model. Their model allows for time-dependent arrival patterns. Nguyen, Sivakumar, and Graves (2015) propose a network flow approach to determine the optimal allocated capacity between two patient groups: those on their first visit and returning patients.

One of the main tactical decisions in an OAS with a hybrid access policy is how to share the available capacity between same-day and pre-scheduled appointments. Qu et al. (2012) propose a mean-variance model with an efficient solution procedure to determine the percentage of same-day appointments with the objective of increasing the average number of patients seen while also reducing the variability. Qu et al. (2007) demonstrate that the optimal percentage of same-day appointments depends on the ratio of the demand for open-access appointments to the provider's capacity, and the ratio of the show rate of pre-scheduled appointments to the show rate of open-access appointments. Balasubramanian et al. (2014) illustrate that the location of pre-scheduled appointment slots has a significant impact on the number of serviced same-day patients, as well as on the continuity of care in hybrid appointment systems.

Patrick, Puterman, and Queyranne (2008) present a model to dynamically allocate the available daily capacity, considering an arbitrary number of patient groups (called priority classes in their paper) that can differ in terms of viable scheduling window and costs for late booking. Saure et al. (2012) expand this model by adding multiple appointment requests and multiple session durations, and by allowing some appointments during overtime.

There is a strong relationship between OAS problems at the operational level and capacity allocation decisions. For example, the problem of patient selection from a waiting list is dependent on the capacity allocated to each group of patients. A few studies focusing on operational decisions simultaneously consider the capacity allocation problem (Cayirli & Gunes, 2013; Green et al., 2006; Qu, Peng, Kong, & Shi, 2013). Qu et al. (2013) propose a two-phase approach for planning a multi-server OAS. In the first phase, they focus on capacity allocation among different service categories during one week to balance the workload of the providers among consultation sessions. In the second phase, they consider the appointment time problem.

Factors that should be taken into account for this decision are the demand of each patient group, priority level, no-show probability, revenue from each patient group, and preferences of patients and physicians. These have not all been considered simultaneously in prior studies and could offer avenues for future research.

4.2. Appointment intervals (slots)

An appointment interval (slot) is defined as the interval between two successive appointment times (Cayirli & Veral, 2003). Each consultation session is divided into several slots, in which patients are scheduled.

Numerical results presented by Wang (1993) reveal that optimal appointment intervals for homogenous patients have a dome-shaped pattern (i.e., the length of appointment intervals gradually increases toward the middle of the consultation session, and gradually decreases afterward), where service times are independent and identically, exponentially distributed. Denton and Gupta (2003), Robinson and Chen (2003), and Kupier and Mandjes (2015) also present supporting evidence of the dome-shaped pattern. Robinson and Chen (2003) extend this result for the general service time distribution, and Kupier and Mandies (2015) find this result for a two-stage treatment process. Klassen and Yoogalingam (2009) numerically show that the best pattern of appointment intervals with an integer restriction on the appointment interval length is a plateau-dome structure, where the middle appointment intervals are all the same length, creating a plateau.

Luo et al. (2012) show that when the interruption rate is high, the optimal pattern does not have a dome shape, but rather an increasing structure: more specifically, the time between two consecutive appointments is greater for appointments scheduled later in the day. Klassen and Yoogalingam (2013) also focus on the appointment interval problem by considering interruption and lateness. They find that increased interruption level should be managed by having longer appointment intervals, and that an increase in physician lateness should be managed by moving appointments to later in the session. Erdogan and Denton (2013) consider the uncertainty that is caused by service durations and the number of patients seen on each day. Their numerical results indicate a dome-shaped pattern in the absence of no-shows, while no-shows cause the optimal appointment intervals to decrease.

Denton and Gupta (2003) show that the optimal structure of appointment intervals depends on the ratio of the idle-time cost per time unit to waiting-time cost per time unit. For higher values of this ratio, optimal appointment intervals have a dome-shaped structure, but for lower values, the optimal intervals are roughly uniform. Chakraborty, Muthuraman, and Lawley (2013) show that the optimal appointment intervals are dependent on the variation of service time and no-show behavior. Furthermore, they illustrate that a shorter slot length performs better overall because it can enhance the system flexibility. The numerical results of Turkcan et al. (2011) also confirm that shorter slots can increase the total revenue and decrease waiting time. In a congested system with two types of patients (with low and high service-time variability), the optimal schedule often allocates nearly zero time slots to the first few patients (similar to dome and plateau-dome), and inserts a break before switching from a class of patients with higher service-time variability to another class of patients (Kong, Lee, Teo, & Zheng, 2013).

Some authors suggest that adjusting appointment intervals proportionally may offset the negative impacts of patient no-shows (Cayirli & Veral, 2003). In addition, it is shown that optimal appointment intervals depend on patient unpunctuality patterns (Klassen & Yoogalingam, 2014; Tai & Williams, 2012). Patient waiting time and physician idle time can occur when patients arrive earlier than or later than their scheduled appointment times, respectively. Thus, OASs try to reduce these negative effects by choosing appropriate appointment intervals. Klassen and Yoogalingam (2014) illustrate that as the standard deviation of patient unpunctuality increases, clinics will benefit by reducing interval size throughout the schedule. In contrast, as the mean of patient unpunctuality increases, clinics can benefit by increasing interval length slightly. They also find that the increasing-intervals-and-clustering rule, in which the length of appointment intervals and block sizes gradually increase, can improve the system performance when the waiting times of early patients before their appointment times are considered in the objective function.

It should be noted that all of the results above are based on a numerical analysis, and thus there may be conditions under which they are not the best.

Although appointment intervals have a significant effect on system efficiency, many optimization studies in the OAS literature consider the appointment intervals to equal the mean service time. Hence, further integration of the appointment interval decision with other related decisions is a major topic for future research.

4.3. Appointment scheduling window

The appointment scheduling window (scheduling horizon) of an OAS determines how far into the future an appointment can be scheduled (Liu, 2016). This affects the no-show rate, and thus the efficiency of the OAS. When the scheduling window is reduced, indirect waiting time decreases, and consequently the no-show rate decreases. This leads to more efficient use of the clinic capacity. However, an overly restrictive scheduling window may reduce the number of patients seen, thereby leading to a reduction in the clinic revenue. Liu (2016) is the only study that focuses on the effects of the appointment scheduling window on the clinic operational efficiency. He studies a single server system with two types of patients that differ in their no-show probabilities, assuming the service provider knows the type of each incoming patient, and can set different scheduling windows depending on patient type. Generalizable insights are limited, and therefore understanding the operational implications of the length of the scheduling window is an open research area.

4.4. Block size

In general, a block is a set of patients scheduled at the same time, and block size is the number of patients in a block, or the number of patients scheduled at the beginning of a slot (Cayirli & Veral, 2003). This decision is of interest when the patients are homogenous in the slot scheduling method. To determine block sizes, four main types of rules have been studied: 1) Bailey's rule, 2) individual block rule, 3) multiple block rule, and 4) variable block rule. In Bailey's rule, two patients are scheduled in the initial slot, and one patient in all the other slots, leaving the last slot empty. The individual block rule schedules one patient in all slots, and the multiple block rule schedules a fixed batch of patients (often two patients for single-server systems) in all slots. Varying numbers of patients are scheduled in each slot under the variable block rule. The multiple and variable block rules can reduce the negative effect of no-shows, but if the block size is too large, both patient waiting and system overtime costs can increase. Therefore, some papers studies optimal block sizes for multiple and/or variable block rules (Zacharias & Pinedo, 2014; Zeng, Turkcan, Lin, & Lawley, 2010). Zacharias and Pinedo (2014) illustrate that slots with the block size of more than one should be placed at the beginning of the session in the majority of optimal policies. Kong et al. (2013) show that if there is a need to bunch the arrivals of patients together, then it is optimal to bunch the arrivals at the end of the session.

4.5. Number of appointments per consultation session

Studies that consider the decision about the number of appointments per consultation session seek to determine the optimal number of patients that can be scheduled for each consultation session, usually with the goal of minimizing the patient waiting time and provider overtime. LaGanga and Lawrence (2012) present an upper bound for the optimal number of appointments per consultation session. Saure et al. (2012) show that an approximate optimal policy is to book as much demand as possible on the first day of the scheduling horizon.

Clinics may accept more patients than their available capacities to reduce the negative impacts of no-shows and to improve patient access and provider profits (Cayirli & Veral, 2003). This policy is called overbooking in the literature. If the level of overbooking is determined inefficiently, it leads to longer patient waiting times and system overtime (Lin, Muthuraman, & Lawley, 2011; Zeng et al., 2010). Thus, some authors are interested in determining the optimal level of overbooking through the use of mathematical models (e.g., Kim & Giachetti, 2006; LaGanga & Lawrence, 2012; Liu & Ziya, 2014). LaGanga and Lawrence (2012) obtain supporting evidence of numerical results that an overbooking policy can improve the performance of the OAS across a wide range of service environments and cost structures. Liu et al. (2010) declare that the overbooking policy outperforms other strategies when the patient load is relatively low.

Environmental factors (which are discussed in Section 6) have a significant impact on the optimal number of appointments per consultation session. In spite of this, only a few studies focusing on this decision consider some of these factors. Thus, investigating impacts of environmental factors on determining optimal level of this decision is a potential area for future research.

4.6. Panel size

The panel size of a healthcare facility (or a physician) is the size of the population that the facility is committed to provide services for, or the number of potential patients of the facility (Green & Savin, 2008). Only a fraction of these patients, called the calling population, actually demand health care during a given period,

and hence the panel size can be larger than the service capacity, i.e., the number of patients that the facility or physician can serve (Hulshof, Kortbeek et al., 2012). A panel size should be large enough for the facility to be profitable and to benefit from economies of scale (Green and Savin, 2008). Moreover, a panel size should not be so large that the indirect waiting times (and consequently the no-shows probabilities) increase (Liu & Ziya, 2014), or that the level of continuity of care (as an important measure of healthcare service quality) decreases (Murray & Davies, 2007). Recently, Liu and Ziya (2014) study the optimal panel size under two cases where the overbooking capacity is given and where the overbooking capacity is considered a decision variable. They interpret the arrival rate of the patients as a linear function of the panel size. Their characterizations show how optimal decisions depend on patient no-show behavior that varies with the indirect waiting times. Ozen and Balasubramanian (2013) focus on the optimal panel sizes and panel compositions for a set of physicians with different service capacities where there are multiple types of patients. Their objective is to minimize the maximum of the physician overflow frequencies (i.e., the probability that the demand assigned to a physician exceeds his/her capacity). They show that the physicians with the same panel size but different panel composition can have very different overflow frequencies.

Besides optimization studies on panel size, Green, Savin, and Murray (2007) use a model to investigate the link between the panel size and the probability of physician overtime under an open-access policy. Green and Savin (2008) propose a single-server queuing model to determine the effects of no-show rate on the panel size. Their results can be used to provide a guidance on identifying the right panel size for an OAS with an open-access policy.

Note that although a healthcare facility may be able to determine the optimal panel size, in many cases it cannot control the panel size because it does not have complete control over the factors impacting its panel size. In these cases, the panel size cannot be considered a decision variable. For instance, in a publicly-funded healthcare system, any given healthcare provider facility usually must serve all people in its geographical area, often defined as a hospital or clinic being required to serve anyone that shows up and asks for service. In these cases, the optimal panel size can be used as a target for a healthcare facility which can be achieved, for example, by carrying out promotional activities to increase demand or service improvements that increase capacity. There are also a few cases where a healthcare facility or a physician has more control over the panel size. For example, the number of families that are assigned to a family doctor or the set of areas that must be served by a healthcare facility can be controlled where the panel size is a function of the number of families and the set of assigned areas, respectively. In addition, healthcare providers in privatized systems and privately-run clinics in public systems usually have some control over their panel sizes.

Two future directions can be suggested in this context. First, the panel size has an effect on other OAS decisions; hence, considering the panel size problem with other OAS decisions is one possible avenue for future work. Second, the no-show rate is an important factor, and thus its dependence on the panel size should be addressed using more realistic models when determining the optimal panel size.

4.7. Priority of patient groups

Medical facilities with heterogeneous patients usually need to determine priorities for various patient groups. There are two major types of patient priorities, which may be called *hard* and *soft*. For the first, the patient priorities must be met at all times. Actually, hard priorities are applied upon patient arrivals. Such priorities are very popular in the queuing system literature where the customers typically come to the system without setting any appointment. In OASs, hard priorities can be considered for various groups of walk-ins. For instance, in almost all medical centers, the highest priority is assigned to urgent walk-ins that must be served upon their arrival.

Soft priorities are not applied upon patients arrivals and only incorporated in scheduling, deciding appointment day, capacity planning or determining other operational or tactical decisions, in order to put more values on different patient groups who make appointments before their arrivals. A common method to define soft priorities is to assign a different weight to each patient group to reflect its relative importance. In most studies, these weights are represented by waiting-time penalty coefficients (Conforti, Guerriero, & Guido, 2008, Conforti, Guerriero, Guido, & Veltri, 2011; Patrick et al., 2008; Saure et al., 2012). Note that, both hard and soft priorities can be considered simultaneously in an OAS for walk-ins and other patients, respectively. Although priorities have a significant impact on patient waiting time and on resource utilization (Hulshof, Kortbeek et al., 2012), no optimization study has focused on determining patient priorities.

5. Operational decisions

Operational decisions (also referred to as booking decisions; see Wang & Gupta (2011)) are related to the execution of plans at the individual patient level. There are two approaches to determine these decisions in the literature: the rule-based approach (RBA) and the optimization-based approach (OBA). The RBA provides a set of easy-to-implement instructions, possibly in the form of a heuristic; the OBA specifies the optimal level for an operational decision. Although the rules and parameters related to the RBA are determined so that some criteria are improved, the RBA does not guarantee that the best performance will be achieved. In contrast, the OBA aims to obtain the globally optimal solution for the operational decisions under consideration in that study, but often the required assumptions make it less practical, and its implementation in practice is more difficult than the RBA. For this reason, some studies that use the OBA also present rules for practitioners (Balasubramanian et al., 2014; Mak, Rong, & Zhang, 2014b; Oh, Muriel, Balasubramanian, Atkinson, Ptaszkiewicz, 2013; Samorani & LaGanga, 2015). In the following subsections, each of the operational decisions is discussed briefly and its importance is clarified by means of some relevant studies.

5.1. Allocation of patients to servers/resources

The allocation of a patient to a server or a set of resources considerably complicates OAS problems when there are multiple service providers or a limited number of resources.

Service providers in multi-server systems are modeled as multiple identical or multiple different service providers. When modeling primary care clinics, identical servers are often assumed (e.g., El-Sharo et al., 2015; Nguyen et al., 2015; Qu et al., 2013). The most important factors that need to be considered in the patient-physician allocation problem in clinics with different servers are the continuity of care and patient preference for a particular physician. The only study in this category that considers the continuity of care is Balasubramanian et al. (2014), in which the revenue earned when a primary-care provider sees one of his/her own patients is higher than when the continuity of care is broken. Wang and Fung (2014a,b), Wang, Fung, and Chan (2015), and Wang and Gupta (2011) study patient preferences in a patient-physician allocation problem.

Special clinics often simultaneously require multiple resources to serve patients; for example, a chemotherapy center needs oncologist, nurse, and infusion seat (Sevinc, Sanli, & Goker, 2013). There-

fore, appointment systems in these clinics are confronted with multi-resource scheduling problems (e.g., Pérez et al., 2013; Pérez et al., 2011; Riise et al., 2016; Turkcan, Zeng, & Lawley, 2012). Riise et al. (2016) is the only study that considers setup times for resources as well as multiple resources (i.e., physician, room, and equipment), which may be unavailable in periods on each day of the scheduling window. Most studies consider only a single type of resources for modeling OAS problems. For example, Azadeh, Baghersad, Farahani, and Zarrin (2015), Sevinc et al. (2013), and Conforti et al. (2011) consider a set of specific equipment for modeling their problems in special clinics.

It also should be noted that almost all studies on this decision apply the OBA. There are only two papers that propose a rule in addition to their optimization models (Balasubramanian et al., 2014; Wang & Gupta, 2011). None of the studies in this category deals with important environmental factors, such as no-shows, and patient and physician unpunctuality. Hence, a direction for future research would be to investigate the impact of these factors on OAS problems with resource allocation considerations.

5.2. Appointment day

In OASs with traditional and hybrid scheduling polices which have a multi-day scheduling window, the appointment day for each patient must be determined so that constraints on the patient's priority level and waiting-time target (i.e., the time when the patient needs to receive the treatment) are satisfied. The most common approach applied in the literature to tackle this problem is the OBA (e.g. Conforti et al., 2008, Conforti, Guerriero, & Guido, 2010, 2011; Gocgun & Puterman, 2014).

Erdelyi and Topaloglu (2009) study the problem of selecting service jobs which should be scheduled on appropriate days for a general service system using the OBA. They focus on a protection policy, in which a portion of the capacity is protected from lower priority jobs, so as to be available for the higher priority jobs.

Recently, Truong (2015) propose an efficient algorithm to compute the optimal policy for the dynamic assignment of patients to exam days with non-stationary stochastic demand and capacity. Parizi and Ghate (2016) propose a model to dynamically assign patients to exam days in a multi-class and multi-server OAS while considering no-shows, cancellations, and overbooking.

The literature review shows that determining the appointment day alone is not attractive for OR researchers. Normally the authors consider it along with other operational decisions, especially appointment time (e.g., Conforti et al., 2008, 2010, 2011). Incorporating patient preferences for a particular day can further complicate the mathematical model for this problem; there is only one paper considering this. Feldman, Liu, Topaloglu, and Ziya (2014) propose two models (static and dynamic) to decide which set of days to make available for the patients in order to maximize the expected net profit per day. Their models incorporate no-shows and cancellations that are dependent on the indirect waiting time, and embed patient preferences among the different appointment days.

5.3. Appointment time

An appointment time decision problem deals with finding a specific time when a patient is scheduled to start receiving care, so that a performance criterion is optimized. In an OAS that uses the slot scheduling method (slots are predetermined), when an individual patient calls, he\she is scheduled by assigning him\her to an appointment time (operational level) that corresponds to the start time of a predetermined interval (see Section 4.2 for slots, also called appointment intervals). In OASs that do not use the slot scheduling method (no slots are predetermined), patients are scheduled any time in the consultation session. This can be more efficient than slot scheduling (Chakraborty et al., 2013; Conforti et al., 2010), but it is less attractive in practice because the resulting appointment times have no particular patterns, which makes implementation difficult.

Early studies in the appointment scheduling literature often focus on the RBA. In a primary care clinic, where it is valid to assume that patients are identical in terms of required service time and resources, the RBA can achieve an optimal performance. When patients are homogeneous, the dominant scheduling rule is a combination of two variables: appointment interval and block size, discussed at the tactical level in Sections 4.2 and 4.4, where the first-in-first-served discipline determines the sequence of the patients. Cayirli and Veral (2003) explain these scheduling rules and review pre-2003 papers based on this approach.

In contrast to the RBA, which presents a structure to be used for scheduling patients over a period of time, the OBA is often used to update scheduling in each decision period (for example, daily or even with the arrival of each patient). Most optimization studies on the appointment time problems use the OBA (e.g., Begen & Queyranne, 2011; Castro & Petrovic, 2012; Chen & Robinson, 2014; Erdogan & Denton, 2013; Ge et al., 2013; Kuiper et al., 2015; Mak et al., 2014b). Moreover, some studies use the insights obtained by the OBA to propose an easy-to-implement rule (Zacharias & Pinedo, 2014).

As shown in Table B.2 in Appendix B almost all papers assume that the sequence of patients is known in advance, and only a few papers consider the timing and sequencing problems simultaneously (e.g., Mancilla & Storer, 2012; Oh et al., 2013) or suggest heuristics to determine an improved sequence before finding the appointment times (e.g., Berg, Denton, Erdogan, Rohleder, & Huschka, 2014; Kemper et al., 2014; Mak et al., 2014a,b). Since a complete scheduling system determines both sequence and appointment times, it is expected that future studies will develop appointment scheduling models for simultaneously sequencing and timing heterogeneous patients, which can be complex from a mathematical standpoint because patient sequencing problems are generally NP-hard (e.g., Mancilla & Storer, 2012). Moreover, other research opportunities are sequencing and appointment timing with both random arrivals (i.e., unpunctuality of patients that is discussed in Section 6.1) and random service durations.

Another related issue that may be interesting for future study is to account for the size of the waiting area when scheduling, especially when it is limited. This includes more than just using a queue capacity, possibly including the comfort level of patients and reducing the spread of germs.

It seems that the appointment time decision has been well addressed in the literature, but most proposed models apply simplifying assumptions (such as offline scheduling, punctual patients, single-stage service, and unlimited waiting area capacity). Hence, the appointment time problem with more realistic assumptions is a major area for future study opportunities (see Gupta & Denton (2008) for a detailed discussion).

5.4. Patient acceptance/rejection

In some outpatient clinics, a patient could be rejected to reserve capacity for other patient groups. A key tradeoff in accepting patients is the balance between service provision revenues and costs that are incurred from patient waiting time and provider overtime. Factors that are taken into account are resource availability, capacity allocated to each patient group, and demand of each patient group.

This decision often appears in OASs with online scheduling and is determined together with appointment time or appointment day decisions (e.g., Chakraborty et al., 2013; Samorani & LaGanga, 2015; Wang & Fung, 2014a). As shown in Table B.2 in Appendix B both

approaches (i.e., RBA and OBA) are used to decide which appointment requests to accept. Most of the rules for this decision are obtained from the capacity allocation policy that is determined at the tactical level.

Gupta and Wang (2008) propose heuristics to decide which requests to accept in a multi-server system. Chakraborty, Muthuraman, and Lawley (2010) and Chakraborty et al. (2013) present a stopping criterion for scheduling patients based on the unimodality of their objective functions.

5.5. Patient selection from waiting list

Sometimes there are more patients on the waiting list than the available capacity, even with overbooking. In such a case, the number of patients to serve should be selected in terms of various criteria, such as the capacity allocated to each patient group, the patient waiting time, and the patient priority levels. Some authors assume that the priority initially assigned to each patient does not change, but in the real world the patient's condition may change during that time (e.g., Min & Yih, 2014).

A variation of this decision occurs in some hospital diagnostic facilities (such as magnetic resonance imaging (MRI) and computed tomography (CT)) that serve both inpatients and outpatients. These clinics face the challenge of selecting, in real time, who will be served next when demand comes from three patient groups: inpatients, emergency patients, and prescheduled patients, which based on our terminology could be referred to as regular walk-ins, urgent walk-ins, and scheduled patients, respectively (Gocgun et al., 2011; Green et al., 2006; Kolisch & Sickinger, 2008). This problem can be also considered a type of dynamic capacity allocation problem for a diagnostic facility.

As shown in Table B.2 in Appendix B, most of the studies apply the OBA for this decision; hence, one future direction would be to present easy-to-implement rules for near-optimal selection of patients from the waiting list.

5.6. Patient sequence

A patient sequencing problem deals with determining the order in which a list of patients should be scheduled. Because of the dependence between sequencing and timing problems, all studies in this category also deal with their corresponding appointment time problems.

A few studies seek to determine an appropriate sequence; most of them use the insights obtained from the OBA to propose easy-to-implement rules (e.g., Berg et al., 2014; Mak et al., 2014b; Oh et al., 2013). Erdogan et al. (2015) establish the optimal patient sequence for an OAS with two patients in order to minimize a weighted sum of the makespan and indirect waiting cost. Their numerical results for a similar OAS with more than two patients show that a first-come first-appointment rule is approximately good when all patients have the same waiting costs and service time distributions. The ordered variance (OV) policy – which orders patients in increasing variance – is the most common rule found to produce a good sequence (Mak et al., 2014a). Kemper et al. (2014) and Mak et al. (2014a) derive results that confirm this policy in cases where service time distributions belong to a scale family (e.g., the exponential family) and where only the means and supports of the distributions of service durations are known, respectively. Mak et al. (2014a) also prove the optimality of the OV policy for a robust mean-variance model with any number of jobs under a mild condition. Berg et al. (2014) present supporting evidence that scheduling in increasing order of service time variance and no-show probability is the optimal sequence. Erdogan et al. (2015) recommend the OV policy for heterogeneous patients when no-show probabilities are nearly zero. Qi (2016) shows numerically that the sequencing decisions depend on both the patient and physician thresholds for the threshold-based measure discussed in Section 2.3. When the patient thresholds are relatively small, patients with lower variance and shorter average service time should be served first, and vice versa. Mancilla and Storer (2012) develop new algorithms based on Benders decomposition that perform significantly better than the OV policy, but their run times are too high for implementation in real problems. Recently, Kong, Lee, Teo, and Zheng (2016) show that the optimality of the OV policy depends on the number of patients in the system as well as the shape of service time distributions. This policy is more likely to be optimal if the service time distributions are more positively skewed, but this advantage gradually disappears as the number of patients increases. They also prove that the deterministic variant of the appointment sequencing problem is strongly NP-hard even if both patient waiting-time cost and overtime cost are equal. Zacharias and Pinedo (2014) introduce a new sequencing rule that orders patients according to a single index that is a function of the no-show probability and the weight of each patient.

As mentioned in Section 5.3, due to the inherent complexity of sequencing problems, most optimization studies assume that the sequence of patients is given or is determined by simple heuristic rules. The exact optimal sequencing policy is still unknown in cases where there are three or more patients (Mak et al., 2014b). Therefore, one of the biggest challenges for future research is to find optimal (or near-optimal) solutions to more realistic appointment sequencing problems.

6. Environmental factors

Various external and internal environmental factors are considered in the OAS literature. The complexity added by these environmental factors makes designing and planning OASs an interesting and challenging research field. Table C.3 in Appendix C summarizes which environmental factors are addressed in the OAS literature, and they are discussed in more detail below.

6.1. Patient unpunctuality

Patient unpunctuality is defined as the difference between a patient's appointment time and actual arrival time (Cayirli & Veral, 2003). Thus, unpunctuality includes a patient's earliness or lateness. Although patient earliness can affect OAS planning and appointment scheduling (Klassen & Yoogalingam, 2014), most authors assume that early patients are on their own time, and therefore only consider lateness in measuring performance with patient unpunctuality.

Due to the complexity caused by considering patient unpunctuality in a mathematical model, few optimization studies consider it. Almost all of these studies assume that the unpunctuality rate is independent of appointment time and patient characteristics, and there is some empirical evidence to suggest that this is a valid assumption (Klassen & Yoogalingam, 2014). One issue that faces clinics is how to deal with patients arriving out of the scheduled order; whether an available physician should see a patient that has arrived early or wait for the patient scheduled next (known as the wait-preempt dilemma). Recently, Samorani and Ganguly (2016) solved this problem optimally for two patients; they analytically determine the time intervals when it is optimal to see the early patient and those time intervals when it is optimal to wait. They also extend their analytical method to the n-patient case and provide a software application for clinics to manage this dilemma.

6.2. Physician lateness

Physician lateness can lead to a reduction in clinic efficiency by extending patient waiting times (Klassen & Yoogalingam, 2013). This is a less interesting factor to analyze mathematically than others because it is mostly controllable if the physician arrives on time for the session. However, there are cases where it may be completely uncontrollable, for example, if the physician gets delayed off-site due to hospital rounds or emergencies.

6.3. Interruption

Interruptions may include writing up notes, talking with support staff, or emergency patient arrivals. There are two main types of interruptions: non-preemptive and preemptive. Non-preemptive interruptions occur between patient consultations (e.g., Klassen & Yoogalingam, 2013). Preemptive interruptions occur during patient consultations and can be divided into two categories: preemptive-repeat and preemptive-resume (Krishnamoorthy, Pramod, & Chakravarthy, 2014). In the preemptive-repeat, the treatment procedure repeats from the beginning independently of the earlier service. Although this type of interruption is conceivable in a wide range of medical tests, such as an electrocardiogram test that records the heart's rhythm and activity continuously during the predetermined period, this has not been addressed previously. In the preemptive-resume, the service resumes from the place it was interrupted as the server becomes available again. This type of interruption is applied by Luo et al. (2012) to model the presence of emergency requests that have a higher priority and can interrupt scheduled patient consultations. For a detailed discussion on interruptions in healthcare, the reader is referred to Grundgeiger and Sanderson (2009) and Rivera-Rodriguez and Karsh (2010).

Although this environmental factor has a significant impact on the reliability and performance of an OAS, it has received limited attention in the literature, as shown in Table C.3 in Appendix C. Therefore, developing mathematical models for OAS problems considering interruptions could be a direction for future research.

6.4. Patient no-show and cancellation

One of the major problems that almost all outpatient clinics are confronted with is patient no-shows or last minute cancellations. If patients cancel their appointments far enough in advance so that the scheduler can substitute new requests for canceled appointments, that is not a significant problem for clinics; a few papers deal with this phenomenon (e.g., Liu et al., 2010; Parizi & Ghate, 2016; Schuetz & Kolisch, 2013). Conversely, if patients cancel too late that a new request cannot be substituted, this is the same as a no-show operationally. Therefore, most papers that consider this issue focus on no-shows, including late cancellations. According to Wang and Gupta (2011), based on factors affecting the no-show probability, no-shows can be divided into five categories: homogeneous, wait-dependent, patient-dependent, service-dependent, and time-dependent. In addition to these classes, Samorani and LaGanga (2015) find that weather conditions on the day of the appointment can impact no-show probabilities.

Two strategies are proposed in the literature to manage the negative effects of no-shows: reducing appointment intervals and overbooking. For a detailed discussion, the reader is referred to Cayirli and Veral (2003) and Gupta and Denton (2008). While patient no-shows have been well addressed in the literature, only a few studies have considered heterogeneous no-show probabilities which depend on realistic factors such as indirect waiting time. Hence, more work can be done in this direction.

6.5. Patient preference

Patients usually prefer a specific date and physician over other dates and physicians. These preferences often differ from one patient to another, and may change over time. Wang and Gupta (2011) present a framework for the design of OASs that dynamically learn and update patient preferences, and use this information to improve the clinic revenue. Feldman et al. (2014) use a multinomial logit model to govern the patient preferences. In their model, each patient can choose a day between a set of the days that is optimally suggested by the model. Yan et al. (2015) propose a model that considers patient choice and service fairness simultaneously, where assigning a patient to non-preferred slots increases his/her no-show probability. They prove that their objective function – including the number of patients seen, idle time, overtime, and waiting time – is unimodal when the service time is exponentially distributed. Their numerical results also show that considering patient preferences decreases both overtime and waiting time, but increases the idle time. Wang et al. (2015) show that more revenue can be expected if patient preferences and choices are considered. As shown in Table C.3 in Appendix C, most studies do not include patient preferences. This may be both because this data is difficult to obtain, and because incorporating patient preferences complicates OAS designing and planning.

It is observed that physicians have various preferences (Gupta & Denton, 2008), but no paper was found that focused on physician preferences.

6.6. Random service time

Service times are assumed to be either deterministic or stochastic (Gupta & Denton, 2008). Both of these assumptions can include either homogeneous or heterogeneous patient characteristics. In the presence of deterministic service times, there have to be some other factors, e.g., random arrivals and no-shows; otherwise, the problem is trivial. Almost all of the reviewed papers consider the uncertainty of service times. Most studies assume identical service time distributions for all patients. A recent research trend is therefore to study the heterogeneous case where service time distributions depend on patient or service types. Some optimization studies use attributes of exponential distributions to propose an optimal solution for OAS problems in a reasonable amount of computational time (e.g., Kaandorp & Koole, 2007; Tang et al., 2014; Turkcan et al., 2011). Other distributions used to model service time are reviewed in Cayirli and Veral (2003). Kuiper et al. (2015) approximate the service time distributions with phase-type distributions to efficiently determine an optimized schedule with low computational effort. Chakraborty et al. (2010) study an online appointment scheduling problem under a general service time distribution. They also show that using the Gamma distribution for service times significantly reduces the computation time. Kong et al. (2013) use a distributionally robust model to match the prescribed mean and covariance estimates of the service times for an appointment scheduling problem. Begen and Queyrane (2011) apply a joint discrete probability distribution and show that under a cost rate condition there exists an optimal integer appointment schedule that can be found in polynomial time. Begen, Levi, and Queyranne (2012) assume that the distribution of service times are not known and use the sample average approximation method to tackle the appointment time problem. Mak et al. (2014a) also develop a distributionally robust model by assuming that some moments of the service time distribution are given. Kemper et al. (2014) provide an approach to schedule patients for any convex loss function and any service time distribution.

Due to the fact that primary care providers have some control over the service time for each patient, it may be reasonable to assume that service times are deterministic in order to focus on other complicating factors, such as random arrivals and no-shows (Qu et al., 2015). A direction for future research can be studying the impact of uncertainty resulting from random service times on the results obtained by those studies considering deterministic service times.

It should be also noted that there are some OAS decisions for which the uncertainty of service times has not been studied extensively – such as the capacity allocation, number of appointments in consultation session, patient-to-server allocation, appointment day, and patient selection from a waiting list (e.g., Qu et al. 2007, 2012; Kortbeek et al., 2014; Nguyen et al., 2015; Patrick et al., 2008; Ratcliffe, Gilland, & Marucheck, 2012). Thus, more work can be done in these areas.

6.7. Patient heterogeneity

In the majority of OASs, some patient characteristics are known, such as priority level, consultation time, disease type, treatment type, and even no-show probability. Hence, patients can be classified according to these characteristics. These classifications can be used for prioritizing, sequencing, scheduling patients, and adjusting the appointment intervals (Cayirli & Veral, 2003). Zacharias and Pinedo (2014) also find that the patient heterogeneity and no-show rate have a significant impact on the optimal schedule and should be taken into consideration. Kolisch and Sickinger (2008) find that clustering patients according to their no-show probabilities help to build schedules with better performance.

6.8. Type of appointment required by patients

There are three types of appointments in outpatient clinics, namely, single appointments, combination appointments, and appointment series (Hulshof, Kortbeek et al., 2012). Most OASs described in the literature focus on single appointments, where only one appointment is allocated to each patient. In other situations, medical clinics with several departments or specialty clinics may need to schedule multiple appointments on one or more days. Combination appointments refer to cases where multiple appointments are scheduled for a single patient in one day. This includes multi-stage services, where each stage is scheduled separately. Azadeh et al. (2015) propose a model to determine the optimal sequence of required tests for each patient by considering precedence constraints on some tests in a multi-stage OAS. Multiple appointments on more than one day are considered as series. Modeling appointment series for online systems that need to consider future arrivals can further complicate the mathematical models. Therefore, this is an interesting direction for future work.

6.9. Environmental factors: future research possibilities

It should be noted that adding any environmental factor results in a more complicated mathematical model (see Section 5 in Gupta & Denton (2008) for a detailed discussion), for which developing efficient solution procedures is more challenging.

In addition to the above-mentioned factors that affect an OAS's performance and cause inherent uncertainties, there are other environmental factors that could be studied. The first is disruption; although it has been well addressed in other fields (such as supply chain management, Kleindorfer & Saad, 2005; Snyder, Scaparra, Daskin, & Church, 2006), it has received limited attention in the healthcare optimization literature. The disruptions in OASs can be classified into two main classes. The first often stems from natural disasters, such as earthquakes, floods, and terrorist attacks, which result in very high levels of urgent walk-in demand that disrupt routine planned care delivery. In these situations, appointment times of most pre-scheduled patients are postponed and the OASs must make decisions about which pre-scheduled and which urgent patients should be seen based on available resources. The second class includes disruptions caused by economic or financial crises, social events, or machine breakdowns, resulting in complete stoppage or severely reduced available resources in the clinic. In these situations, the OAS must make decisions about rescheduling pre-scheduled patients and how to manage new requests during the disruption period. Determining the optimal policy in the presence of disrupting factors can be a significant area for future research.

Secondly, all reviewed papers model a non-competitive environment, where each outpatient clinic is assumed to have enough patient demand and where the decisions of clinics do not have any effect on each other. Although this assumption is true in some healthcare environments, it may not hold for privatized systems (as in the U.S.) or luxury healthcare service providers. The most important factors affecting customers' choice in competitive healthcare services are price; quality of services; direct and indirect waiting times; and travel time to the office. Thus, this is an open research area, which could be addressed using game theory (see Allon & Federgruen (2007) for more discussion about competition in service industries).

7. Modeling approaches and solution methods

In this section, the wide variety of modeling approaches and solution methods used in the OAS literature are discussed, in order to define future research directions. A classification for modeling approaches and solution methods is given in Table B.1 in Appendix B. The modeling approaches and solution methods used for each reviewed paper are shown in Table B.2 in Appendix B. In addition, Figs. 2 and 3 provide an overview of how frequently they are used in the literature.

7.1. Modeling approaches

Most approaches to model OAS problems make use of stochastic optimization and stochastic dynamic programming, in order to deal with the uncertainties inherent in OASs.

Stochastic programming allows the user to minimize or maximize an objective function in the presence of randomness. Single-stage and two-stage stochastic programming are the most popular stochastic optimization modeling approaches for OAS problems. To the best of our knowledge, no study has used chance-constrained programming in this context. Single-stage stochastic programming optimizes a problem with a random objective function or constraints where a decision is implemented without subsequent recourse (Begen & Queyranne, 2011; Begen et al., 2012; Luo et al., 2012). Two-stage stochastic programming is often used to formulate an appointment scheduling problem. In the resulting formulations, the first stage decision variables are usually appointment times and the second stage variables are generally auxiliary variables, such as the patient waiting times, server idle time, and system overtime. For instance, Denton and Gupta (2003) present a two-stage stochastic programming model that permits considerable flexibility in modeling different types of cost considerations. Berg et al. (2014) also formulate a model for optimal booking, sequencing, and scheduling of a single stochastic server by using two-stage stochastic programming. A few studies also use multi-stage stochastic programming, such as Erdogan and Denton (2013) who propose a multi-stage stochastic linear programming model with stages defined by patient appointment requests to analyze the structure of online scheduling and to provide insights for practice. In addition, they present a two-stage stochastic linear programming model for an offline scheduling system. Both models

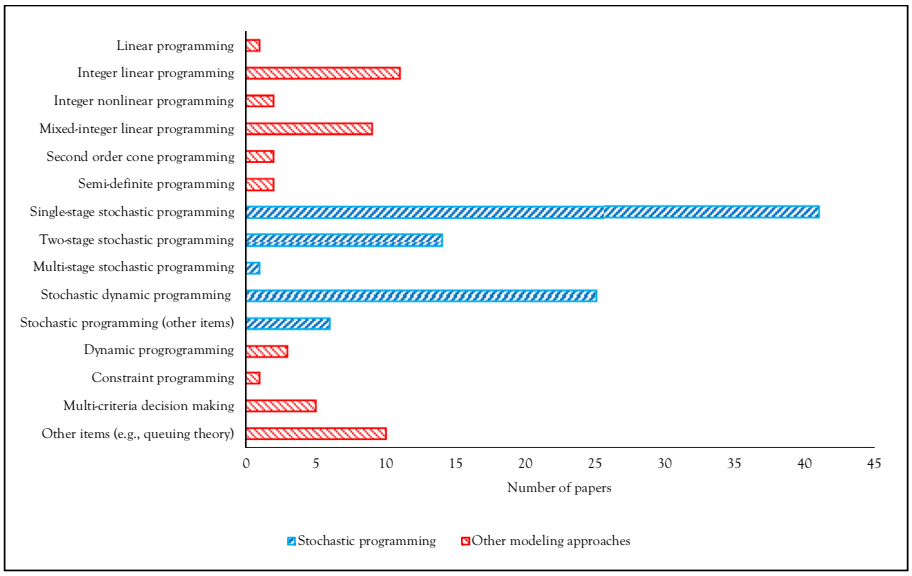

Fig. 2. The frequencies of modeling approaches used in the OAS literature.

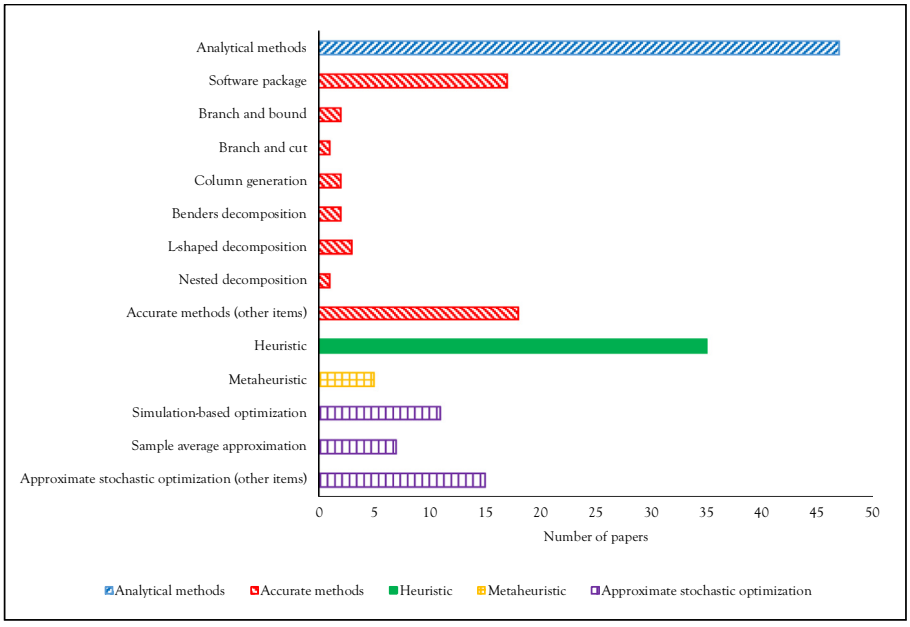

Fig. 3. The frequencies of solution methods used in the OAS literature.

have complete recourse and are solved by decomposition-based algorithms that are proposed based on the structure of the models.

Some decisions in an online OAS – such as allocation of patients to servers/resources, appointment day and time, and patient acceptance/rejection – are determined during the ongoing patient call-in process. Markov decision process (MDP) is the most useful stochastic dynamic programming approach for dealing with this application (Lin et al., 2011). For instance, MDP is used in a chemotherapy appointment scheduling problem (Gocgun & Puterman, 2014), an appointment day problem (Feldman et al., 2014; Liu et al., 2010; Patrick et al., 2008; Saure et al., 2012), a capacity management problem (Gupta & Wang, 2008), and to select patients from a waiting list in each period (Kolisch & Sickinger, 2008; Min & Yih, 2014).

Within this context, queuing theory represents another important tool for managing uncertainty in OAS problems and for calculating parameters and performance measures. The considerable contributions and applications of queuing theory in healthcare systems are described by Green (2006) and Preater (2002). As mentioned in Section 6.5, the exponential distribution is commonly used to model stochastic service times, since that makes the model tractable (Cayirli & Veral, 2003). Therefore, the most common queues studied in the OAS literature are M/M/s models. Recently, other models like D/G/1 have received attention in the literature (e.g., Kemper et al., 2014; Kuiper et al., 2015). Creemers et al. (2012) use the output of a bulk service queuing model as the input of an optimization procedure to determine the optimal allocation of slots over a number of patient classes.

Almost all deterministic models in the OAS literature are formulated by using (mixed-) integer linear programming. Deterministic models are often used to formulate problems at the tactical level that are less affected by the uncertainty caused by random arrivals and random service times. These models are also used in specialty clinics (such as radiotherapy clinics) where service time in each

stage of the treatment procedure is deterministic and no-show probability is close to zero. Typical complications for these models include capacity and due date constraints with multi-resource, multi-stage treatment procedures (Castro & Petrovic, 2012; Conforti et al., 2008, 2010, 2011; Pérez et al., 2011; Turkcan et al., 2012; Wang & Fung, 2014b).

7.2. Solution methods

As mentioned in Section 6, healthcare systems face uncertainty arising from service duration, patient preferences, patient arrivals (walk-ins and scheduled), interruptions, etc. Considering these issues in the formulations often makes OAS problems difficult to solve. Some of these problems have been proven to be NP-hard. Studies that focus on these problems try to propose an efficient algorithm to find a good solution in a reasonable time, see, for instance, Mancilla & Storer (2012). However, there are OAS problems that can be solved efficiently in theory and practice. Such problems can be solved in polynomial time under specific conditions, see, for instance, Begen & Queyranne (2011) and Ge et al. (2013).

To solve an optimization model, several solution methods may be used. These methods can be classified into two main classes: analytical and numerical methods. Analytical methods involve a generic process that uses mathematical principles to obtain provable results for a problem. Although analytical methods have received considerable attention in the OAS literature (as shown in Fig. 3), they are rarely used to generate closed-form optimal solutions. The most common results obtained by these methods are to prove the existence of optimal solutions or some properties of optimal solutions, to compare some given policies in a general setting, or to present bounds on an optimal objective function. These analytical results may be also used to develop efficient algorithms (Chakraborty et al., 2013; Feldman et al., 2014; Gupta & Wang, 2008; Huh, Liu, & Truong, 2013; Koeleman & Koole, 2012; Truong, 2015).

Numerical methods are often developed as a set of iterative procedures to solve complex problems that cannot be easily solved by analytical methods. They can be divided into two main categories: accurate and inaccurate, where the term accurate is borrowed from terminology used in the metrology literature (Menditto, Patriarca, & Magnusson, 2007) (the optimal value can be considered the reference value and the computational method seeking the optimal value can be thought of as a measurement system or method). A numerical method is here referred to as accurate if it finds either an optimal solution in a given time or a perturbed (or near-optimal) solution with a known deterministic error bound, which is a bound on the solution's (relative or absolute) optimality gap. However, an inaccurate method does not provide an error bound and cannot determine the quality of its resulting solution. As shown in Fig. 3, the majority of reviewed papers use inaccurate methods; this is due to the complexity of OAS problems that makes them difficult to solve with an accurate method. Inaccurate methods can be classified into three main groups: heuristics, metaheuristics, and approximate stochastic optimization. There may be exceptional optimization procedures based on one of these three groups with known accuracy, but these methods typically cannot determine the quality of their resulting solutions. Non-exact algorithms are sometimes called heuristic. This definition is much broader than the definition considered here for heuristic algorithms, which are inaccurate methods that do not fall into the other two subclasses: metaheuristic and approximate stochastic optimization.

Sample average approximation (SAA) is a common method in the approximate stochastic optimization class that uses sampling methods to tackle stochastic OAS problems with complex expectation terms (Kim, Pasupathy, & Henderson, 2015). This method is used in Begen et al. (2012) to solve an appointment scheduling problem with the assumptions that service time distributions are discrete, unknown and possibly correlated, and only sample information is available. This study uses the convexity and subdifferential of the objective function to determine a bound on the number of independent samples required to obtain a near-optimal solution with high probability. It also shows that their SAA-based scheduling problem can be solved in polynomial time under mild conditions. Ge et al. (2013) extend the sampling method presented in Begen et al. (2012) to a general class of functions and provide an alternative bound on the number of samples required to obtain a near-optimal solution with an additive error. Mancilla and Storer (2012) prove that their SAA-based sequencing and scheduling problem with two scenarios and different waiting costs is NP-complete by polynomially transforming the problem to the three-partition problem. Then, they present a heuristic solution approach based on Benders decomposition.

Another frequently applied method in the approximate stochastic optimization class is simulation-based optimization that combines optimization and simulation methods to facilitate the search procedure in uncertain complex systems. For instance, Klassen and Yoogalingam (2013) use this method to solve an appointment scheduling problem with physician lateness and interruptions. Schuetz and Kolisch (2012) and Lin et al. (2011) use a simulation-optimization approach to approximate the Bellman equation of their proposed MDP model for accepting and scheduling patients with no-shows.

Heuristics form an important subclass of inaccurate solution methods to solve OAS problems. Although these methods do not guarantee optimal solutions, sometimes they are indispensable in practical because of the presence of complex environmental characteristics and stochastic factors. Balasubramanian et al. (2014) develop a heuristic which blends both access and continuity in its decision process to provide near optimal same-day assignments. Robinson and Chen (2003) propose a simple heuristic by using the structure of the optimal solution for an appointment scheduling problem.

Almost all proposed MDP models, because of their large state spaces, are solved by approximate methods that are based on approximate dynamic programming. Patrick et al. (2008) approximate their proposed MDP model by a linear program with a tractable number of variables and an intractable number of constraints, whose dual can be solved using the column generation method. A similar approach is followed by Saure et al. (2012) and Wang et al. (2015). Qu et al. (2015) present heuristic admission policies by examining the properties of the proposed MDP model and compare the performance of them under different clinic settings. Feldman et al. (2014) and Liu et al. (2010) propose an approximate method based on an MDP model, where the approximate method applies a single step of the standard policy improvement algorithm to an initial good policy.

The other group of inaccurate methods is comprised of metaheuristic algorithms. Although metaheuristics can help to solve problems with more realistic assumptions or in less computational time, there are a limited number of studies that have employed metaheuristic algorithms for obtaining a solution. In most of these studies, a simulation method is employed to compare the solution candidates, and a metaheuristic algorithm is used to guide the search solution space (Qu et al., 2013; Saremi, Jula, ElMekkawy, & Wang, 2013). Recently, Azadeh et al. (2015) propose a lower bound and a genetic algorithm for an appointment scheduling problem considering heterogeneous multi-stage treatments.

The presence of uncertainty in healthcare systems makes it difficult to propose efficient accurate methods for OAS problems, but a number of accurate methods presented in the literature seem to be powerful enough to solve realistic problem instances. For

example, Chen and Robinson (2014) use Benders decomposition to solve the stochastic linear programming model that is proposed for an appointment scheduling problem while considering random and heterogeneous service times (they also develop two simple heuristic solutions). Erdogan and Denton (2013) propose a nested decomposition algorithm that iteratively improves the approximation of each stage's convex objective function by adding supporting hyperplanes to solve a multi-stage stochastic program. Denton and Gupta (2003) also propose an L-shaped algorithm to solve their stochastic model for scheduling patients where service durations are continuously distributed. For a multi-resource appointment scheduling problem, Riise et al. (2016) present an exact method based on a recursive logic-based Benders decomposition, where each subproblem is formulated as an integer linear program.

Some authors identify several important properties of their model and use them to propose an efficient algorithm to find an optimal solution in a reasonable time. For instance, in Begen and Queyranne (2011), the existence of an optimal integer solution for a specific appointment scheduling problem is proven when service durations are integer discrete random variables. It is also shown that the objective function is L-convex under a mild condition (the so-called α-monotonicity condition) on the cost coefficients. Moreover, an algorithm is proposed to obtain the optimal schedule in polynomial time when the service durations are independent and bounded. Ge et al. (2013) extend this work to piecewise linear cost functions with integer break points, and show that under mild conditions their problem can be solved in polynomial time. Kaandorp and Koole (2007) prove that their proposed model for a scheduling problem with exponentially-distributed service times is multi-modular, and a local search algorithm can be used to obtain the global optimal solution where the objective function is a weighted sum of the patient waiting time, server idle time, and lateness. However, their algorithm is of super-polynomial complexity. Their model is extended to capture emergency arrivals and general service times by Koeleman and Koole (2012).

A widely-used solution method is to use software packages (e.g., CPLEX). Indeed, software packages use some algorithms to solve the input models; however, using them is here considered as a distinct solution method to indicate those studies that only focus on modeling their problems without developing any specialized solution method. It should be also noted that although using software packages is generally classified as an accurate method, in some situations software packages may be unable to find an optimal solution because of two main reasons: 1) the limitation of possible run time or PC memory, 2) the complex structure of the input model, for which no efficient algorithm has been developed yet, for example, non-convex models, which cannot be solved by most existing solvers except in a very small scale.

One way to tackle the difficulty of providing a good solution for a complex OAS problem is to intuitively establish a connection between that problem and a well-known problem for which efficient algorithms have been already introduced. For example, Mak et al. (2014b) use a classical serial supply chain inventory problem to develop a mixed-integer second order cone programming approximation for an appointment sequencing problem.

A significant number of studies present a case study or use real data to validate their proposed models as well as to evaluate the applicability of their solution methods. These studies are listed in Table C.4 in Appendix C.

7.3. Research opportunities

The review of optimization modeling approaches and solution methods used in the OAS literature highlights some gaps in this field. Due to the complexity of the problems presented in this context, heuristic and metaheuristic methods have the most potential for solving large-sized problem instances; however, metaheuristics are rarely used for OAS problems.

Most reviewed studies propose a simple basic model with limited constraints and environmental factors. However, in order to better address practical problems, the models need to become more complex, which cannot be solved in a reasonable amount of time unless efficient solution methods are used. Thus, this will be an ongoing challenge as researchers attempt to develop better accurate and inaccurate solution methods. At this point, for most existing OAS problems, no efficient accurate (exact or bounded-error) solution method has been developed yet.

A methodology which may be promising for modeling and solving OAS problems in competitive environments is game theory; no prior work was found using this modeling approach.

The list of modeling approaches and solution methods presented in Table B.2 in Appendix B, together with the points above, provide directions for future research.

8. Conclusions and future research

In this paper, a comprehensive review of optimization studies on outpatient appointment system (OAS) problems is presented, categorizing them using a hierarchical structure at the strategic, tactical, and operational levels. The aims of this work are to provide an overview of OASs in healthcare, to identify aspects that have received less attention in the operations research (OR) literature, and to indicate directions for future research in this field. In this section, some of the major findings are summarized, and Table 4 encapsulates the open research questions that are discussed throughout the paper.

Strategic decisions have a significant effect on modeling and on the practical applicability of the presented solution methods. Such decisions are most frequently treated as inputs into an OAS model, and only a few optimization studies are found that compare strategic-level options based on numerical experiments.

The problems addressed at the tactical level aim to determine the OAS structure. While it seems that determining the optimal level for tactical decisions can increase system performance in the long term, some of these decisions have only received limited attention in the OR literature (e.g., appointment scheduling window).

Operational decisions, especially the appointment time decision, have received the most attention in the OAS literature. Two different approaches are used to determine decisions at the operational level: optimization-based approach (OBA) and rule-based approach (RBA). The OBA is used to develop optimization models to optimally determine operational decisions. However, due to the difficulty of implementing these models in practice, some studies using the OBA also suggest a rule based on the insights obtained from their proposed optimization models. The RBA proposes easy-to-implement rules to obtain good solutions for operational decisions. Optimization models are used to optimally set the parameters of the rules considered by the RBA. Although operational decisions are highly dependent on strategic and some tactical decisions, they are generally modeled separately.

It is found that most optimization studies neglect environmental factors that may further complicate the mathematical models (e.g., patient and physician preferences). As has been noted repeatedly in the previous studies, these factors appear in realistic problems, and hence considering them can help increase the applicability of the resulting models. Thus, this is a highly recommended direction for future research.

Due to the inherent uncertainty of OAS decisions and their sequential property, especially at the operational level, stochastic optimization and stochastic dynamic programming (especially the

Table 4
Summary of future research directions.

Subject category	Future research direction
Strategic level	Comparing possible options for strategic-level decisions, such as access policy and type of scheduling, by analytical methods. Presenting a model for finding the optimal number of servers/resources for an outpatient clinic. Determining optimal (or near optimal) policies on acceptance of walk-ins in different types of OASs. Developing online scheduling models that consider realistic arrival patterns and future requests. Integrating strategic decisions with relevant tactical/operational decisions in a stochastic multi-stage setting.
Tactical level	Presenting provable results for optimal appointment interval patterns in OASs that use the slot scheduling method. Considering the relationship between the capacity allocation and operational decisions in an integrated optimization model. Incorporating the appointment interval decision into determining operational decisions. Investigating impacts of environmental factors on determining the optimal number of appointments in a consultation session. Studying impacts of panel size on optimal levels of other tactical OAS decisions. Presenting optimal rules for prioritizing patient groups. Developing optimization models for the choice of appointment scheduling window. Incorporating no-show rate into determining the optimal panel size under more realistic assumptions. Developing a capacity allocation problem while simultaneously considering demand, priority level, and earned revenue of each patient group, as well as physician preferences.
Operational level	Presenting easy-to-implement rules for near-optimal selection of patients from a waiting list. Studying multi-resource OAS problems with environmental factors, such as no-shows, and patient and physician unpunctuality. Investigating impact of the continuity of care and patient preferences on the patient-to-server allocation. Determining appointment times under more realistic assumptions. Presenting scheduling models for simultaneously sequencing and timing heterogeneous patients. Incorporating both random service times and random arrivals into appointment sequencing and timing problems. Considering a dynamic waiting list in the patient selection. Incorporating patient preferences and wait-dependent no-shows into the appointment day selection. Considering the limitations of the waiting area that can include the comfort level of patients and the spread of germs into scheduling problems.
Environmental factors	Designing and planning OASs in competitive environments, such as privatized systems and luxury healthcare service providers. Adopting more realistic assumptions about environmental factors, such as unpunctual and heterogeneous patients, heterogeneous service times, provider preferences, and heterogeneous no-show probabilities that depend on realistic factors (e.g., indirect waiting time). Modeling multiple appointments in specialty clinics (e.g., appointment series in radiotherapy clinics). Including interruptions (especially preemptive-repeat interruptions) in the existing optimization models. Planning OASs and scheduling appointments in the presence of disruptive factors. Considering interactions between no-shows and walk-ins.
Objective functions and performance measures	Using non-linear structures for time-based cost functions. Applying the Pareto approach to multi-objective models. Including fairness, congestion, and indirect waiting time in objective functions. Using risk-averse objective functions.
Modeling approaches and solution methods	Developing integrated models for multi-decision OAS problems with more than four decisions. Presenting comprehensive models with a hierarchical approach. Modeling OAS problems in a competitive healthcare environment using game theory. Developing optimization models considering changes in patients' behavior based on their previous experience in the OAS. Proposing optimization models with more practical applicability. Presenting good solution procedures in terms of quality, speed, and implementation for nearly realistic models.
Other items	Considering characteristics and restrictions of electronic appointment booking systems. Comparing ASs in traditional and non-traditional care paradigms, such as non-visit (e-consult) care, home care, and team-based care. Presenting new technologies to produce insights from the big data generated with HIT and HER systems. Considering interactions between the AS and other systems of a medical center. Integrating and coordinating several interrelated OASs that share the same patients, health procedures, or resources. Considering impacts of EDs on the ASs of other units, and providing methods to better link EDs and those ASs. Addressing OASs for multi-site care networks and outpatient clinics with multiple consultation rooms for each physician. Developing OASs for facilities that provide group visits, e-visits and home visits. Studying OASs for special healthcare facilitates, such as blood units, medical laboratories, rehabilitation centers, etc. Using off-site public access devices as parts of OASs. Incorporating the real-life complexities of patient flows into OAS optimization studies. Reviewing to what extent current academic research in ASs is used to enhance software scheduling products. Reviewing simulation and case studies in the OAS literature based on the framework presented here. Reviewing forecasting techniques used in the OAS literature.

Markov decision process) models are used most frequently. In most of the proposed models, time-based measures are used to evaluate system performance, and the conventional weighted sum approach is used to tackle the multi-criteria feature of OAS problems.

In the last decade, due to increased interest in more realistic OASs, the mathematical models presented in the literature have become complex and more difficult to solve optimally in an acceptable amount of time. Therefore, finding a good solution procedure in terms of quality and speed for nearly realistic models is an important and challenging issue.

There are a few other research opportunities that do not fit well into the framework presented above; these are discussed briefly below.

Health information technology (HIT) and electronic health record (EHR) systems, used by healthcare organizations and encompassing their organizations' appointment systems, are becoming more prevalent due to their important role in improving patient safety, increasing operational efficiency, and reducing costs. The results of operations research/management science (OR/MS) studies in healthcare scheduling contexts will be enhanced if they are able to interface or be implemented with these electronic systems. Nevertheless, most of the recommendations made by optimization studies are at a level that renders them very difficult or impossible to implement in electronic appointment systems. Lack of attention to the characteristics and restrictions of electronic appointment booking systems reduces the applicability of the opti-

mization models, so this would be a worthwhile direction for future research. A few studies that directly take into account electronic appointment booking systems are Feldman et al. (2014), Wang and Gupta (2011), and Zhu, Hou, Wang, and Zhou (2012). A contributing factor to the gap between electronic appointment systems and optimization models may be the users' inability to provide information required to implement OR/MS models.

As another result of the rapid growth of HIT and EHR systems, healthcare organizations are faced with exponential growth in data availability, which can yield invaluable information if properly managed. Referred to as "big data" in the literature, these datasets can be used to calibrate optimization models, although they are difficult to manage using traditional computing technologies. Therefore, focusing on new technologies to produce insights from big data and to implement them in healthcare organizations is a significant area for future research.

Non-traditional approaches to healthcare resulting from HIT and EHR systems are also being evaluated in terms of their cost-effectiveness in a number of outpatient care settings, as they provide faster access to high-quality care, offer enhanced communication between primary care providers and specialists, reduce no-show rates, and decrease unnecessary or inappropriate specialty referrals. Care paradigms, such as non-visit (e-consult) care, home care, and team-based care are generating significant momentum in the delivery of outpatient care. Although there has been some early research in this area by the OR/MS community (e.g., Yuan, Liu, & Jiang, 2015), more studies are needed on how to efficiently manage patient requests and available resources (e.g., physicians and nurses) under these new outpatient care delivery systems. Interestingly, these topics are receiving quite a bit of attention in medical journals (e.g., Popejoy et al., 2015; Jaber, Braksmajer, & Trilling, 2006; Gidwani, Fernandez, & Schlossman, 2012; Greenhalgh et al., 2016; Al-Mahdi, Gray, & Lederman, 2015; Mitchell et al., 2012), which can provide a resource for relevant problems that the OR/MS community could help address. Therefore, an important new research area is to study OASs for group visits (also known as shared medical visits), e-visits, and home visits.

For different healthcare facilitates (HCFs), such as blood stations, medical laboratories, rehabilitation centers, etc., OASs may have different requirements. Hence, OASs for special HCFs providing outpatient services is worth investigating in future study. Moreover, the use of off-site public access devices (OPADs) as parts of OASs is another emerging research area. One may refer to Ahmadi-Javid, Seyedi & Syam (2016) for a comprehensive classification of different types of HCFs, as well as, characteristics of OPADs, which are a new type of HCF.

Another interesting research question is how to integrate or coordinate several interrelated OASs that share the same patients, health procedures, or resources, for example, OASs for the different offices used by a single physician, or OASs for different physicians sharing one facility. Another research opportunity is to investigate OASs for integrated multi-site care networks where physicians' travel assignments and patients' appointment locations are decided (e.g., Li, Kong, Chen, & Zheng, 2016). Some reviewed papers assume that a physician remains in one consultation room. Others assume that the physician travels between multiple consultation rooms, which allows patients to prepare for their consultation before the physician arrives to their room. This is useful for outpatient clinics with long pre-consultation and/or post-consultation activities (e.g., undressing, measuring blood pressure, and cleaning the room). What is lacking in the literature is the modeling of multiple consultation rooms including travel times between them. (Hulshof, Vanberkel et al., 2012) consider this issue and show that having multiple rooms is beneficial if the physician travel time is less than the patient preparation time. With models such as these other interesting questions can also be addressed, such as how many exam rooms are necessary for a given environment, and how nurses or nurse practitioners can be integrated into the clinic flow. Moreover, studying OAS problems in competitive environments using game theory may be another emerging research opportunity.

Patient flows to and from an emergency department (ED) influence other ASs in the hospital or clinic (see a recent review by Saghafian, Austin, and Traub (2015) on patient flow optimization in EDs). Most emergency walk-ins arriving at non-emergency units of the hospital are patients coming from EDs, and they have higher priority than scheduled patients (e.g, Luo et al., 2012). Interestingly, the majority of the patients seen in EDs are not true emergencies, and EDs are often used for non-urgent or routine health services. This results in increased costs and reduced quality of critical care (Wexler et al., 2015). Proposed solutions to guide non-emergency patients away from the ED are education regarding the importance of using a primary care physician, and a referral system which helps direct patients seeking non-urgent care to relevant non-emergency units and assists them to set timely appointments in the other units (Chan et al., 2009; Murnik, Randal, Guevara, Skipper, & Kaufman, 2006; OnceCity Health, 2016). The impacts of EDs on the ASs of other units and solutions for properly linking EDs and the ASs could be addressed in future studies.

Most outpatient clinics, especially specialty care clinics, need to deal with varying patient requirements and complex patient care pathways. However, almost all optimization studies in the OAS literature consider either a single care pathway for all patients or a predetermined care pathway for each patient. The reason may be that patient flows are highly jumbled and unpredictable processes. Incorporating the real-life complexities of patient flows through outpatient clinics into OAS studies has received limited attention in simulation studies (e.g., Rohleder, Lewkonia, Bischak, Duffy, & Hendijani, 2011). Hence, this is a worthwhile future research direction due to the impact of patient flows on OAS performances.

Healthcare facilities make use of a variety of systems in addition to an appointment system, such as a service delivery system, a capacity planning system, a marketing system, a financial system, a physical system, and an information technology system. While all of these systems influence the efficiency and performance of the others, the interactions between the appointment system and other systems have not yet been considered in any of the reviewed studies; therefore, this remains a potential area for future research.

Although the body of literature on OASs is growing, and many of the simpler issues have been addressed, it remains an area with many interesting and complex open research questions.

Acknowledgments

The authors are very grateful to the four anonymous reviewers and the journal Editors, Professor Jean-Charles Billaut and Professor José Oliveira, whose constructive comments and insightful critiques resulted in a much improved paper. We also thank those who provided comments on earlier versions of this paper, which were presented in 2015 MSOM and 2016 POMS annual conferences.

Appendix A. Glossary

Term	Definition
Appointment interval (slot)	The time interval between two successive appointment times; in other words, each consultation session is partitioned into subintervals which are called slots (Cayirli & Veral, 2003).
Appointment time	The start time that an individual patient is scheduled to receive service.
Block	A set of patients scheduled at the same time.
Block size	The number of patients scheduled at the same time; in other words, the number of patients in a block (Cayirli & Veral, 2003).
Cancellation	A situation where a patient cancels his/her appointment far enough in advance of their scheduled time to allow for a new appointment to be substituted.
Consultation session	The time period available for scheduling patients (Cayirli & Veral, 2003).
Direct waiting time	The time between the maximum of the appointment time and arrival time, and the consultation start time (Gupta & Denton, 2008).
Flow time	The total time a patient spends in the clinic (Cayirli & Veral, 2003).
Indirect waiting time	The time between when a patient requests an appointment and the scheduled appointment time (Gupta & Denton, 2008).
Late cancellation	A situation where a patient cancels his/her appointment so late that a new request cannot be substituted, which is the same as a no-show operationally.
No-show patient	A patient who does not show up for his\her appointment, and does not give prior notice.
No-show rate (no-show probability)	The probability of a patient being a no-show.
Outpatient appointment system	A main component of an outpatient clinic that plan and schedule appointment requests to deliver timely and convenient access to health services for all patients (Gupta & Denton, 2008).
Outpatient clinic	A health facility that provides care to patients that do not need an overnight stay (Hulshof, Kortbeek et al., 2012).
Overtime	The positive difference between the desired completion time of the clinic session and the actual end of the service for the last patient (Cayirli & Veral, 2003).
Panel size	The size of the population that a healthcare facility (or a physician) is committed to provide services for, or the number of potential patients of the facility (Green & Savin, 2008).
Pre-scheduled appointment/patient	An appointment/patient that is scheduled in advance of the appointment day (Balasubramanian et al., 2014).
Regular walk-in patient	A walk-in patient who does not require urgent treatment.
Same-day appointment/patient	An appointment/patient that is scheduled in the same day that the patient call for an appointment (Balasubramanian et al., 2014).
Server idle time	The part of the consultation session that the server (or physician) is idle due to lack of patient(s).
Service duration	The length of time a single patient spends with the service provider.
Scheduled patient	A patient who makes an appointment before arriving at the clinic.
Urgent walk-in patient	A walk-in patient who needs treatment as soon as possible (Gupta & Denton, 2008).
Walk-in patient (walk-ins)	A patient who shows up during the consultation session without any scheduled appointment (Gupta & Denton, 2008).

Appendix B. Reviewed papers

Table B.1
Survey descriptive dimensions from computational perspective: modeling approach and solution method; including abbreviations used in Table B.2.

Survey dimension	Code				
Solution method	AM	Analytical method			
	SL	Numerical method	Accurate method (Exact or bounded-error method)	Solving by a general-purpose optimization software package	LINGO
	SC				CPLEX
	SG				GAMS
	SO				Other items
	BB			Branch and bound	
	BC			Branch and cut	
	CG			Column generation	
	BD			Benders decomposition	
	LD			L-shaped decomposition	
	ND			Nested decomposition	
	O			Other items	
	H		Inaccurate method	Heuristic	
	MH-TS			Metaheuristic	Tabu search
	MH-GA				Genetic algorithm
	MH-SA				Simulated annealing
	MH-O				Other items
	S-SBO			Approximate stochastic optimization	Simulation-based optimization
	S-SAA				Sample average approximation
	S-O				Other items
Modeling approach	LP	Linear programming			
	ILP	Integer linear programming			
	INLP	Integer nonlinear programming			
	MILP	Mixed-integer linear programming			
	SOCP	Convex conic programming		Second order cone programming	
	C-SDP			Semi-definite programming	
	PSP	Stochastic programming		Probabilistic (or chance-constraint) programming	
	1-SSP			Single-stage stochastic programming	
	2-SSP			Two-stage stochastic programming	
	M-SSP			Multi-stage stochastic programming	
	MDP			Stochastic dynamic programming	Markov decision process
	SDP-O				Other items
	SP-O			Other items, such as distributionally robust optimization (DRO)	
	DP	Dynamic programming			
	CP	Constraint programming			
	MCDM	Multi-criteria decision making			
	MPDM	Multi-person decision making (game theory)			
	O	Other items, such as queuing theory (QT), graph theory (GT), and network theory (NT)			

Table B.2
Characteristics of reviewed papers (see Section 2.1 and Table B.1 for definitions and abbreviations).

Reference	Tactical and operational decisions	Strategic decisions				Objective: minimize (Min.), maximize (Max.)	Modeling approach	Solution method
		Access policy	Type of scheduling: online (On), offline (Off)	Number of servers/resources: single (S), multiple (M)	Policy on acceptance of walk-ins: allowed (Yes), not allowed (No)			
(Anderson, Zheng, Yoon, & Khasawneh, 2015)	T2	Traditional	-	S	No	Min. costs of waiting time, idle time, and overtime	1-SSP	S-SBO
(Azadeh et al., 2015)	O1/O6/OO (sequence of tests required by patients) (OBA) (Integrated)	-	On	M	No	Min. makespan	MILP	MH-GA
(Balasubramanian et al., 2014)	O1/O3/O4 (main focus on same-day patients) (OBA, RBA) (integrated)	Hybrid	On	M	No	Max. timely access and continuity of care	SDP-O	O/H
(Balasubramanian, Muriel, & Wang, 2012)	T1/O1 (OBA) (integrated)	Hybrid	Off	M	No	Max. revenue of patients seen	2-SSP	AM/S-SAA
(Begen et al., 2012)	O3 (OBA)	Traditional	Off	S	No	Min. costs of waiting time, idle time, and overtime	1-SSP	AM/S-SAA
(Begen & Queyranne, 2011)	O3 (OBA)	Traditional	Off	S	Yes (urgent)	Min. costs of waiting time, idle time, and overtime	1-SSP	AM
(Berg et al., 2014)	T2/O3/O6 (T2/O3: OBA, O6: OBA, RBA) (integrated)	Traditional	Off	S	No	Max. profit (revenue of patients seen – costs of waiting time, idle time, and overtime)	2-SSP	AM/BB/LD/H
(Bikker, Kortbeek, van Os, & Boucherie, 2015)	TO (physician's scheme design)	Hybrid	On	M	No	Min. weighted sum of lower bound of indirect waiting time, and difference between daily supply and demand	MILP	SC
(Castaing, Chon, Denton, & Weizer, 2016)	O3 (OBA)	Traditional	Off	M	No	Min. costs of waiting time and overtime	2-SSP	S-SAA/O
(Castro & Petrovic, 2012)	O3 (OBA)	Traditional	Off	M	No	Hierarchical objectives: Min. weighted # of patients exceeding waiting-time targets; Min. maximum lateness; Min. sum of weighted lateness	MCDM/MILP	H/SC
(Cayirli & Gunes, 2013)	T1/T4 (integrated)	Hybrid	On	S	Yes (regular)	Min. costs of waiting time, idle time, and overtime	1-SSP	S-SBO
(Chakraborty et al., 2010)	O3/O4 (OBA) (integrated)	Open-access	On	S	No	Max. profit (revenue of patients seen – costs of patients overflowing between each two successive slots)	1-SSP	AM/H
(Chakraborty et al., 2013)	O3/O4 (OBA) (integrated)	Hybrid	On	S	No	Max. profit (revenue of patients seen – costs of waiting time and overtime)	1-SSP	AM/H
(Chen & Robinson, 2014)	T2/O3/O6 (T2/O3: OBA, O6: RBA) (T2/O3: integrated, (T2/O3)/O6: sequential)	Hybrid	On	S	No	Min. costs of waiting time, idle time, and overtime	1-SSP	BD/H (for O6)

(*continued on next page*)

Table B.2 (*continued*)

Reference	Tactical and operational decisions	Strategic decisions				Objective: minimize (Min.), maximize (Max.)	Modeling approach	Solution method
		Access policy	Type of scheduling: online (On), offline (Off)	Number of servers/resources: single (S), multiple (M)	Policy on acceptance of walk-ins: allowed (Yes), not allowed (No)			
(Choi & Banerjee, 2016)	T4	-	-	S	No	Min. costs of delay and idle time between blocks	SP-O	H
(Conforti et al., 2008)	O1/O2/O5/OO (appointment shift) (OBA) (integrated)	Traditional	Off	S	No	Max. # of patients seen	ILP	SL
(Conforti et al., 2010)	O1/O2/O3 (OBA) (integrated)	Traditional	Off	M	No	Max. # of patients seen	ILP	SC
(Conforti et al., 2011)	O2/O3/O5 (OBA) (integrated)	Traditional	Off	S	No	Max. # of patients seen	ILP	SC
(Creemers et al., 2012)	T1	Hybrid	On	S	No	Min. cost of waiting time	1-SSP/O (QT)	H
(Denton & Gupta, 2003)	T2/O3 (OBA) (integrated)	Traditional	Off	S	No	Min. costs of waiting time, idle time, and overtime	2-SSP	AM/LD
(Dobson et al., 2011)	T1	Hybrid	On	S	Yes (urgent)	Max. profit (revenue of patients seen – costs of waiting time and urgent patient overflow)	1-SSP/O (QT)	AM/H
(El-Sharo et al., 2015)	O1/O3 (OBA) (integrated)	-	On	M	Yes (regular)	Max. profit (revenue of patients seen – costs of waiting time and overtime)	1-SSP	O
(Erdelyi & Topaloglu, 2009)	T1/O2/O4 (OBA) (integrated)	Traditional	On	S	No	Min. costs of holding and rejecting patients	SP-O	S-O
(Erdogan & Denton, 2013)	T2/O3 (OBA) (integrated)	Traditional Open-access	Off On	S	No	Min. costs of waiting time and overtime	2-SSP M-SSP	- ND
(Erdogan et al., 2015)	T2/O3/O6 (T2/O3: OBA, O6: OBA, RBA) (integrated)	Traditional	On	S	Yes	Min. costs of direct and indirect waiting times, idle times, and overtime	2-SSP	AM/LD
(Feldman et al., 2014)	O2 (main focus on determining appointment days to be available for each patient) (OBA)	Hybrid	On	S	No	Max. profit (revenue of patients seen – cost of overtime)	MDP	AM/S-O
(Ge et al., 2013)	O3 (OBA)	Traditional	Off	S	No	Min. costs of waiting time and idle time	2-SSP	AM/S-SAA
(Geng & Xie, 2016)	O2/O4 (OBA) (integrated)	Hybrid	On	S	No	Max. revenue of patients seen	MDP	AM
(Geng & Xie, 2012)	T1	Hybrid	-	S	No	Min. costs of waiting time, unused contracted time slot, and contracted time slot cancellation	MDP	AM/O
(Geng, Xie, & Jiang, 2011)	T1	Hybrid	On	S	No	Min. costs of idle time and waiting time	MDP	AM/H

(*continued on next page*)

Table B.2 (*continued*)

Reference	Tactical and operational decisions	Strategic decisions				Objective: minimize (Min.), maximize (Max.)	Modeling approach	Solution method
		Access policy	Type of scheduling: online (On), offline (Off)	Number of servers/resources: single (S), multiple (M)	Policy on acceptance of walk-ins: allowed (Yes), not allowed (No)			
(Geng, Xie, & Jiang, 2011)	T1	Hybrid	-	S	No	Min. costs of waiting time, unused contracted time slot, and contracted time slot cancellation	MDP	AM/S-O
(Gocgun et al., 2011)	O5 (OBA)	Traditional	On	M	Yes (urgent, regular)	Max. profit (revenue of patients seen – costs of waiting time and overtime)	MDP	-
(Gocgun & Puterman, 2014)	O2/O5 (OBA, RBA) (integrated)	Traditional	Off	S	No	Min. costs of patient rejection and patients exceeding target dates	MDP	S-O
(Granja, Almada-Lobo, Janela, Seabra, & Mendes, 2014)	O6 (OBA)	Traditional	Off	M	No	Min. costs of makespan and waiting time	-	S-SBO/MH-SA
(Green et al., 2006)	T1/O5 (OBA, RBA) (integrated)	Traditional	On	S	Yes (regular, urgent)	Max. profit (revenue of patients seen – costs of waiting time and patient rejection)	DP	AM/S-O
(Gupta & Wang, 2008)	O4 (OBA, RBA)	Hybrid	On	S/M	No	Max. profit (revenue of patients seen – costs of patient-physician mismatch, patient rejection, and idle time)	MDP	AM/O
(Hahn-Goldberg et al., 2014)	O3 (OBA)	Traditional	Off On	M	No	Min. makespan	CP -	S-O H
(Hassin & Mendel, 2008)	T2/O3 (OBA) (integrated)	Traditional	Off	S	No	Min. costs of waiting time and server availability	1-SSP	O
(Huang, Hancock, & Herrin, 2012)	T2	-	-	S	No	Min. waiting time and idle time	1-SSP	S-SBO
(Huang & Zuniga, 2012)	T4/O3 (based on no-show threshold for each slot) (OBA) (integrated)	-	On	S	No	Min. costs of waiting time, idle time, and overtime	1-SSP	S-SBO
(Huh et al., 2013)	T1/O5 (OBA) (integrated)	Traditional	Off	M	Yes (urgent)	Min. costs of waiting time and overtime	MDP	AM/O
(Hulshof, Boucherie, Hans, & Hurink, 2013)	T1/O5 (OBA) (integrated)	Open-access	On	M	No	Min. weighted # of patients waiting in each queue and in each time period	MILP	SC
(Kaandorp & Koole, 2007)	T4/O3 (OBA) (integrated)	Traditional	-	S	No	Min. costs of waiting time, idle time, and overtime	1-SSP	AM/O
(Kemper et al., 2014)	O3/O6 (O3: OBA, O6: RBA) (sequential)	-	On	S	No	Min. costs of waiting time and idle time	1-SSP/O (QT)	AM
(Kim & Giachetti, 2006)	T5	Traditional	-	S	Yes (regular)	Max. profit (revenue of patients seen – costs of overtime and patient rejection)	1-SSP	O
(Klassen & Yoogalingam, 2009)	T2/O3 (OBA) (integrated)	Traditional	-	S	No	Min. costs of waiting time, idle time, and overtime	1-SSP	S-SBO
(Klassen & Yoogalingam, 2013)	T2/O3 (OBA) (integrated)	Traditional	-	S	No	Min. costs of waiting time and idle time	1-SSP	S-SBO
(Klassen & Yoogalingam, 2014)	T2/O3 (OBA) (integrated)	Traditional	On	S	No	Min. costs of waiting time and idle time	1-SSP	S-SBO

(*continued on next page*)

Table B.2 (*continued*)

Reference	Tactical and operational decisions	Strategic decisions				Objective: minimize (Min.), maximize (Max.)	Modeling approach	Solution method
		Access policy	Type of scheduling: online (On), offline (Off)	Number of servers/resources: single (S), multiple (M)	Policy on acceptance of walk-ins: allowed (Yes), not allowed (No)			
(Koeleman & Koole, 2012)	T4/O3 (OBA) (integrated)	-	-	S	Yes (Urgent)	Min. costs of waiting time, idle time, and overtime	1-SSP	AM/O
(Kolisch & Sickinger, 2008)	O5 (OBA)	Traditional	On	S	Yes (regular, urgent)	Max. profit (revenue of patients seen – costs of waiting time and patient rejection)	MDP	O
(Kong et al., 2013)	T2/O3 (OBA) (integrated)	Traditional	Off	S	No	Min. costs of waiting time and overtime	2-SSP/SP-O (DRO)/C-SDP	SO
(Kong et al., 2016)	O6	-	Off	S	No	Min. costs of waiting time and overtime	1-SSP	AM/O
(Kortbeek et al., 2014)	T5/TO (maximum number of appointment in each slot) (integrated)	Hybrid	On	S	Yes (regular)	Min. # of rejected patients	SDP-O (Markov reward model)/O (QT)	H
(Kuiper et al., 2015)	T2/O3 (OBA) (integrated)	Traditional Open-access	Off On	S	No	Min. costs of waiting time and idle time	1-SSP/O (QT)	SO
(Kuiper & Mandjes, 2015)	T2/O3 (OBA) (integrated)	Traditional	Off	M	No	Min. costs of waiting time and idle time	1-SSP/O (QT)	AM/SO
(LaGanga & Lawrence, 2012)	T4/T5 (integrated)	Traditional	-	S	No	Max. profit (revenue of patients seen – costs of waiting time and overtime)	1-SSP	AM/H
(Liang & Turkcan, 2015)	O1/O3 (OBA) (sequential)	Traditional	Off	M	No	Min. waiting time and overtime (for nurse assignment problem) Min. overtime and excess workload of nurses (for patient scheduling problem)	MCDM/MILP	O/SC
(Lin, 2015)	O1/O3/O6 (OBA) (integrated)	Traditional	Off	M	No	Min. costs of waiting time, overtime, and congestion	MIP	H
(Lin et al., 2011)	O3 (OBA)	Open-access	On	S	No	Max. profit (revenue of patients seen – costs of waiting time and overtime)	MDP	AM/S-O
(Liu, 2016)	T3	Hybrid	On	S	No	Max. profit (revenue of patients seen and doing ancillary tasks – costs of patient rejection)	1-SSP/O (QT)	AM
(Liu & Ziya, 2014)	T5/T6 (integrated)	Hybrid	On	S	Yes (regular)	Max. profit (revenue of patients seen – system costs)	1-SSP/O (QT)	AM
(Liu et al., 2010)	O2 (OBA, RBA) (integrated)	Hybrid	Off	S	No	Max. profit (revenue of patients seen – costs of booking patients)	MDP	S-O
(Luo et al., 2012)	T2/T5/O3 (OBA) (integrated)	Traditional	-	S	Yes (urgent)	Max. profit (revenue of patients seen – costs of waiting time and overtime)	1-SSP	AM/SO
(Mak et al., 2014a)	T2/O3/O6 (T2/O3: OBA, O6:OBA, RBA) (T2/O3: integrated, (T2/O3)/O6: sequential)	Traditional	Off	S	No	Min. costs of waiting time and overtime	SP-O (DRO)/C-SDP/SOCP/LP	AM

(*continued on next page*)

Table B.2 (*continued*)

Reference	Tactical and operational decisions	Strategic decisions				Objective: minimize (Min.), maximize (Max.)	Modeling approach	Solution method
		Access policy	Type of scheduling: online (On), offline (Off)	Number of servers/resources: single (S), multiple (M)	Policy on acceptance of walk-ins: allowed (Yes), not allowed (No)			
(Mak et al., 2014b)	O3/O6 (O3: OBA, O6: OBA, RBA) (integrated)	Traditional	Off	S	No	Min. costs of waiting time and idle time	SOCP/1-SSP	AM/H/S-SAA
(Mancilla & Storer, 2012)	O3/O6 (OBA) (integrated)	Traditional	Off	S	No	Min. costs of waiting time, idle time, and overtime	2-SSP	S-SAA/H
(Min & Yih, 2014)	O5 (OBA)	Traditional	Off	S	No	Min. costs of patient rejection, waiting time, and overtime	MDP	AM/S-O
(Muthuraman & Lawley, 2008)	O3 (OBA)	Open-access	On	S	No	Max. profit (revenue of patients seen – costs of waiting time and overtime)	1-SSP	AM/H
(Nguyen et al., 2015)	T1	Hybrid	Off	M	No	Min. maximum required capacity	MILP	BC
(Oh et al., 2013)	O3/O6 (OBA, RBA) (integrated)	Traditional	Off	S	No	Min. costs of waiting time and idle time	2-SSP	S-SAA/SC
(Ozen & Balasubramanian, 2013)	T6/TO (panel composition) (integrated)	Hybrid	-	M	No	Min. maximum physician overflow frequency	INLP	AM/H
(Parizi & Ghate, 2016)	T1/O2/O4 (OBA) (integrated)	Hybrid	On	M	No	Max. profit (revenue of patients seen – costs of patient rejection, waiting time, and overtime)	MDP	S-O
(Patrick, 2012)	T1/T5 (integrated)	Hybrid	Off	S	No	Max. profit (revenue of patients seen – costs of overtime, idle time, indirect waiting time, and switching plan)	MDP	O
(Patrick & Puterman, 2007)	T1	Hybrid	-	S	Yes	Min. # of unused slots	1-SSP	SO
(Patrick et al., 2008)	T1/O2/O5 (OBA) (integrated)	Traditional	Off	S	No	Min. costs of patient rejection, booking patient, and waiting time	MDP	S-O
(Peng, Qu, & Shi, 2014)	T1/T4/O3 (OBA) (integrated)	Hybrid	On	S	Yes (regular)	Min. costs of waiting time, idle time, and overtime	1-SSP	MH-GA/S-SBO
(Pérez et al., 2013)	O1/O3 (OBA) (integrated)	Traditional Open-access	Off On	M	No	Min. waiting time cost Max. # of patients seen	ILP 2-SSP	- H
(Pérez et al., 2011)	O1/O3 (OBA) (integrated)	Traditional	On	M	No	Max. # of patients seen	ILP	H
(Qi, 2016)	O3/O6 (OBA) (integrated)	Traditional	Off	S	No	Min. vector of delay unpleasantness measures for patients and physician based on lexicographic order	MCDM/SP-O (DRO)	SC/O

(*continued on next page*)

Table B.2 (*continued*)

Reference	Tactical and operational decisions	Strategic decisions				Objective: minimize (Min.), maximize (Max.)	Modeling approach	Solution method
		Access policy	Type of scheduling: online (On), offline (Off)	Number of servers/resources: single (S), multiple (M)	Policy on acceptance of walk-ins: allowed (Yes), not allowed (No)			
(Qu et al., 2013)	T1	Traditional	Off	M	No	Min. difference of service times between any 2 sessions	MILP	SG
	T4/O3 (OBA) (T4/O3: integrated, T1/(T4/O3): sequential)					Min. costs of waiting time, idle time, and overtime	2-SSP	S-SBO/MH-GA
(Qu et al., 2015)	O4 (only walk-ins)/O5 (OBA, RBA) (integrated)	-	-	S	Yes (regular)	Min. costs of waiting time, idle time, and overtime	MDP	AM/H
(Qu et al., 2012)	T1	Hybrid	-	S	No	Min. variability in # of patients seen Max. # of patients seen	1-SSP/MCDM	AM/O
(Qu et al., 2007)	T1	Hybrid	-	S	No	Max. # of patients seen	1-SSP	AM/O
(Ratcliffe et al., 2012)	T1/O4 (OBA) (integrated)	Hybrid	Off	S	No	Max. profit (revenue of patients seen – overtime cost)	SDP-O	AM/O
(Riise et al., 2016)	O1/O2/O3/O5 (OBA) (integrated)	Traditional	Off	M	No	Min. a linear function that depends on selected activities and their scheduled days	INLP/ILP	BD
(Robinson & Chen, 2003)	T2/O3 (OBA, RBA) (integrated)	Traditional	Off	S	No	Min. costs of waiting time and idle time	2-SSP	AM/O
(Samorani & Ganguly, 2016)	O6 (for unpunctual patient) (OBA)	-	-	S	No	Min. costs of waiting time and idle time	1-SSP	AM/H
(Samorani & LaGanga, 2015)	O2/O3/O4 (OBA, RBA) (integrated)	Hybrid	On	S	No	Max. profit (revenue of patients seen – costs of waiting time and overtime)	2-SSP	CG/H
(Saremi et al., 2013)	O3 (OBA)	Traditional	Off	M	No	Min. waiting time cost and makespan	MILP	MH-TS/S-SBO
(Saure et al., 2012)	T5/O2/O5 (OBA) (integrated)	Traditional	Off	M	No	Min. costs of patient rejection, waiting time and overtime	MDP	S-O
(Savelsbergh & Smilowitz, 2016)	O3 (OBA, RBA)	Traditional	Off	S	No	Min. aggregate probability of patients being in an uncontrolled health state	ILP	CG/BB/H
(Schuetz & Kolisch, 2012)	O4 (OBA, RBA)	Traditional	On	S	No	Max. profit (revenue of patients seen – costs of waiting time, overtime, and patient rejection)	MDP	S-O
(Schuetz & Kolisch, 2013)	O4 (OBA, RBA)	Traditional	On	S	No	Max. profit (revenue of patients seen – costs of patient rejection and overtime)	MDP	O
(Sevinc et al., 2013)	T5	Traditional	Off	M	No	-	-	H
	O1 (OBA) (sequential)					Max. clinic utilization	ILP	H
(Tai & Williams, 2012)	T2	-	-	S	No	Min. waiting time and idle time	O	H
(Tang et al., 2014)	T2/O3 (RBA) (integrated)	-	-	S	No	Min. costs of waiting time, idle time, and overtime	1-SSP/O (QT)	AM/SO
(Truong, 2015)	O2 (OBA)	Hybrid	Off	S	Yes (urgent)	Min. costs of waiting time and utilization	DP	AM/O

(*continued on next page*)

Table B.2 (*continued*)

Reference	Tactical and operational decisions	Strategic decisions				Objective: minimize (Min.), maximize (Max.)	Modeling approach	Solution method
		Access policy	Type of scheduling: online (On), offline (Off)	Number of servers/resources: single (S), multiple (M)	Policy on acceptance of walk-ins: allowed (Yes), not allowed (No)			
(Tsai & Teng, 2014)	O3 (OBA)	Traditional	On	M	No	Max. profit (revenue of patients seen – costs of waiting time and overtime)	1-SSP	H
(Turkcan et al., 2012)	O1/O2/O3 (OBA) (O1/O3: integrated, (O1/O3)/O2: sequential)	Traditional	Off	M	No	Min. costs of treatment delays, overtime, and idle time Min. makespan	ILP	SC/H
(Turkcan et al., 2011)	O3/O4 (OBA) (integrated)	Open-access	On	S	No	Min. difference between # of patients in system at beginning of each slot Min. difference between waiting time for patients arriving at each slot Max. # of patients seen	MCDM/SP-O	AM/H
(Vandaele, Van Nieuwenhuyse, & Cupers, 2003)	TO (group size)	-	-	S	Yes (urgent)	Min. weighted patient lead time	1-SSP/O (QT)	H
(Vink et al., 2015)	T2/O3 (OBA) (integrated)	Traditional	Off	S	No	Min. costs of waiting time, idle time, and overtime	1-SSP	H
(Wang & Fung, 2014a)	O1/O3/O4 (OBA) (integrated)	Open-access	On	M	No	Max. revenue (dependent on patient satisfaction levels)	MDP	S-O
(Wang et al., 2015)	O1/O3 (main focus on offering a candidate set of times and physician to patients) (OBA) (integrated)	Traditional	On	M	No	Max. revenue per day (dependent on # of patients seen)	DP	S-O
(Wang & Fung, 2014b)	O1/O3 (OBA) (integrated)	Traditional	Off	M	No	Max. profit (revenue of patients seen – patient rejection cost) Min. degree of matching between patient preferences and arrangements	ILP	SC
(Wang & Gupta, 2011)	O1/O3/O4 (OBA, RBA) (integrated)	Hybrid	Off	M	No	Max. profit (revenue of patients seen – cost of insufficient allocated same-day capacity)	SDP-O	AM/H

(*continued on next page*)

Table B.2 (*continued*)

Reference	Tactical and operational decisions	Strategic decisions				Objective: minimize (Min.), maximize (Max.)	Modeling approach	Solution method
		Access policy	Type of scheduling: online (On), offline (Off)	Number of servers/resources: single (S), multiple (M)	Policy on acceptance of walk-ins: allowed (Yes), not allowed (No)			
(Wiesche, Schacht, & Werners, 2016)	T1	Traditional	-	S	Yes	Min. # of appointment slots (to have as much capacity as possible for walk-ins)+# of patients shifted from morning to afternoon session	MILP	SO
(Yan et al., 2015)	O3/O4 (OBA) (integrated)	Traditional	On	S	No	Max. profit (revenue of patients seen – costs of waiting time, overtime, and idle time)	1-SSP	AM/O
(Zacharias & Pinedo, 2014)	T4/O3/O6 (heterogeneous patients) (RBA) (integrated)	Traditional	Off	S	No	Min. costs of waiting time, idle time, and overtime	1-SSP	AM/H
	T4/O3 (homogeneous patients) (OBA) (integrated)		On					
(Zeng et al., 2010)	T1/T4/O3 (OBA) (integrated)	Traditional	Off	S	No	Max. profit (revenue of patients seen – costs of waiting time and overtime)	1-SSP	AM/H
	O3/O4 (OBA) (integrated)	Open-access	On				-	H
(Zhu et al., 2012)	O3 (OBA)	Traditional	Off	S	No	Max. sum of priority values on assigned slots Max. relative preference among different patients and requested slots Max. degree of matching between patient preferences and arrangements	MCDM/ILP	H

Appendix C. Additional summary tables

Table C.1
Valid tactical and operational decisions under specific strategic settings – For a cell whose content is N/A, the decision given in the related row is generally not applicable (N/A) under the strategic setting given in the column.

Decision level	Code	Decision	Strategic decisions								
			Access policy			Number of servers/resources		Policy on acceptance of walk-ins		Type of scheduling	
			Traditional	Open access	Hybrid	Single	Multiple	Yes	No	Online	Offline
Tactical decisions	T1	Allocation of capacity to patient groups									
	T2	Appointment interval (slot)									
	T3	Appointment scheduling window		N/A							
	T4	Block size									
	T5	Number of appointments in consultation session									
	T6	Panel size									
	T7	Priority of patient groups									
Operational decisions	O1	Allocation of patients to servers/resources				N/A					
	O2	Appointment day		N/A							
	O3	Appointment time									
	O4	Patient acceptance/rejection									
	O5	Patient selection from waiting list		N/A						N/A	
	O6	Patient sequence									

Table C.2
Performance criteria in reviewed papers.

Performance criteria	References
Patient waiting time	(Anderson et al., 2015; Begen et al., 2012; Begen & Queyranne, 2011; Berg et al., 2014; Castaing et al., 2016; Cayirli & Gunes, 2013; Chakraborty et al., 2013; Chen & Robinson, 2014; Creemers et al., 2012; Denton & Gupta, 2003; Dobson et al., 2011; El-Sharo et al., 2015; Erdogan & Denton, 2013; Erdogan et al., 2015; Ge et al., 2013; Geng & Xie, 2012; Geng, Xie, Augusto, & Jiang, 2011; Geng, Xie, & Jiang, 2011; Gocgun et al., 2011; Granja et al., 2014; Green et al., 2006; Hassin & Mendel, 2008; Huang et al., 2012; Huang & Zuniga, 2012; Huh et al., 2013; Kaandorp & Koole, 2007; Kemper et al., 2014; Klassen & Yoogalingam, 2009; Klassen & Yoogalingam, 2013; Klassen & Yoogalingam, 2014; Kong et al., 2016; Koeleman & Koole, 2012; Kolisch & Sickinger, 2008; Kong et al., 2013; Kuiper et al., 2015; Kuiper & Mandjes, 2015; LaGanga & Lawrence, 2012; Liang & Turkcan, 2015; Lin, 2015; Lin et al., 2011; Luo et al., 2012; Mak et al., 2014a, b; Mancilla & Storer, 2012; Min & Yih, 2014; Muthuraman & Lawley, 2008; Oh et al., 2013; Parizi & Ghate, 2016; Patrick et al., 2008; Peng et al., 2014; Pérez et al., 2013; Qu et al., 2013; Qu et al., 2015; Robinson & Chen, 2003; Samorani & Ganguly, 2016; Samorani & LaGanga, 2015; Saremi et al., 2013; Saure et al., 2012; Schuetz & Kolisch, 2012; Tai & Williams, 2012; Tang et al., 2014; Truong, 2015; Tsai & Teng, 2014; Vink et al., 2015; Yan et al., 2015; Zacharias & Pinedo, 2014; Zeng et al., 2010)
Server idle time	(Anderson et al., 2015; Begen et al., 2012; Begen & Queyranne, 2011; Berg et al., 2014; Cayirli & Gunes, 2013; Chen & Robinson, 2014; Choi & Banerjee, 2016; Denton & Gupta, 2003; Erdogan et al., 2015; Ge et al., 2013; Geng & Xie, 2012; Geng, Xie, Augusto, & Jiang, 2011; Geng, Xie, & Jiang, 2011; Gupta & Wang, 2008; Huang et al., 2012; Huang & Zuniga, 2012; Kemper et al., 2014; Kaandorp & Koole, 2007; Klassen & Yoogalingam, 2009; Klassen & Yoogalingam, 2013; Klassen & Yoogalingam, 2014; Koeleman & Koole, 2012; Kuiper et al., 2015; Kuiper & Mandjes, 2015; Mak et al., 2014b; Mancilla & Storer, 2012; Oh et al., 2013; Patrick, 2012; Patrick & Puterman, 2007; Peng et al., 2014; Qu et al., 2013; Qu et al., 2015; Robinson & Chen, 2003; Samorani & Ganguly, 2016; Tai & Williams, 2012; Tang et al., 2014; Turkcan et al., 2012; Vink et al., 2015; Yan et al., 2015; Zacharias & Pinedo, 2014)
Overtime	(Anderson et al., 2015; Begen et al., 2012; Begen & Queyranne, 2011; Berg et al., 2014; Castaing et al., 2016; Cayirli & Gunes, 2013; Chakraborty et al., 2013; Chen & Robinson, 2014; Denton & Gupta, 2003; El-Sharo et al., 2015; Erdogan & Denton, 2013; Erdogan et al., 2015; Feldman et al., 2014; Gocgun et al., 2011; Huang & Zuniga, 2012; Huh et al., 2013; Kaandorp & Koole, 2007; Kim & Giachetti, 2006; Klassen & Yoogalingam, 2009; Koeleman & Koole, 2012; Kong et al., 2013; Kong et al., 2016; LaGanga & Lawrence, 2012; Liang & Turkcan, 2015; Lin, 2015; Lin et al., 2011; Luo et al., 2012; Mancilla & Storer, 2012; Mak et al., 2014a; Min & Yih, 2014; Muthuraman & Lawley, 2008; Parizi & Ghate, 2016; Patrick, 2012; Peng et al., 2014; Qu et al., 2013; Qu et al., 2015; Ratcliffe et al., 2012; Samorani & LaGanga, 2015; Saure et al., 2012; Schuetz & Kolisch, 2012; Schuetz & Kolisch, 2013; Tang et al., 2014; Tsai & Teng, 2014; Turkcan et al., 2012; Vink et al., 2015; Yan et al., 2015; Zacharias & Pinedo, 2014; Zeng et al., 2010)
Number of patients seen	(Balasubramanian et al., 2012; Berg et al., 2014; Chakraborty et al., 2010; Chakraborty et al., 2013; Conforti et al., 2008, 2010, 2011; Dobson et al., 2011; El-Sharo et al., 2015; Feldman et al., 2014; Geng & Xie, 2016; Gocgun et al., 2011; Green et al., 2006; Gupta & Wang, 2008; Kim & Giachetti, 2006; Kolisch & Sickinger, 2008; LaGanga & Lawrence, 2012; Lin et al., 2011; Liu, 2016; Liu & Ziya, 2014; Liu et al., 2010; Luo et al., 2012; Muthuraman & Lawley, 2008; Parizi & Ghate, 2016; Patrick, 2012; Pérez et al., 2013; Pérez et al., 2011; Qu et al., 2012; Qu et al., 2007; Ratcliffe et al., 2012; Samorani & LaGanga, 2015; Schuetz & Kolish, 2012; Schuetz & Kolisch, 2013; Tsai & Teng, 2014; Turkcan et al., 2011; Wang et al., 2015; Wang & Fung, 2014b; Wang & Gupta, 2011; Yan et al., 2015; Zeng et al., 2010)
Number of rejected patients	(Erdelyi & Topaloglu, 2009; Green et al., 2006; Gupta & Wang, 2008; Gocgun & Puterman, 2014; Kim & Giachetti, 2006; Kolisch & Sickinger, 2008; Kortbeek et al., 2014; Liu, 2016; Min & Yih, 2014; Parizi & Ghate, 2016; Patrick et al., 2008; Saure et al., 2012; Schuetz & Kolisch, 2012; Schuetz & Kolisch, 2013; Wang & Fung, 2014b)
Other items	(Azadeh et al., 2015; Balasubramanian et al., 2014; Bikker et al., 2015; Brailsford & Vissers, 2011; Choi & Banerjee, 2016; Castro & Petrovic, 2012; Chakraborty et al., 2010; Dobson et al., 2011; Erdelyi & Topaloglu, 2009; Erdogan et al., 2015; Geng & Xie, 2012; Geng, Xie, & Jiang, 2011; Gocgun & Puterman, 2014; Granja et al., 2014; Gupta & Wang, 2008; Hahn-Goldberg et al., 2014; Hassin & Mendel, 2008; Hulshof et al., 2013; Kim & Giachetti, 2006; Liang & Turkcan, 2015; Lin, 2015; Liu & Ziya, 2014; Liu et al., 2010; Liu, 2016; Nguyen et al., 2015; Ozen & Balasubramanian, 2013; Patrick, 2012; Patrick et al., 2008; Qi, 2016; Qu et al., 2013; Qu et al., 2012; Riise et al., 2016; Saremi et al., 2013; Savelsbergh & Smilowitz, 2016; Sevinc et al., 2013; Truong, 2015; Turkcan et al., 2012; Turkcan et al., 2011; Vandaele et al., 2003; Wang & Fung, 2014a; Wang & Fung, 2014b; Wang & Gupta, 2011; Wiesche et al. 2016; Zhu et al., 2012)

Table C.3
Environment factors in reviewed papers.

Environment factor	References
Patient unpunctuality	(Klassen & Yoogalingam, 2014; Samorani & Ganguly, 2016; Schuetz & Kolisch, 2012; Tai & Williams, 2012)
Physician lateness	(Klassen & Yoogalingam, 2013; Klassen & Yoogalingam, 2014)
Interruption	(Klassen & Yoogalingam, 2013; Luo et al., 2012)
Patient no-show (or late cancellation)	
Homogeneous	(Anderson et al., 2015; Chakraborty et al., 2013; Feldman et al., 2014; Green et al., 2006; Hahn-Goldberg et al., 2014; Hassin & Mendel, 2008; Kaandorp & Koole, 2007; Kim & Giachetti, 2006; Klassen & Yoogalingam, 2009; Klassen & Yoogalingam, 2014; Koeleman & Koole, 2012; Kolisch & Sickinger, 2008; Kortbeek et al., 2014; LaGanga & Lawrence, 2012; Luo et al., 2012; Qu et al., 2015; Tang et al., 2014; Zacharias & Pinedo, 2014)
Patient-dependent	(Begen & Queyranne, 2011; Berg et al., 2014; Chakraborty et al., 2010; Chen & Robinson, 2014; El-Sharo et al., 2015; Erdogan & Denton, 2013; Green et al., 2006; Huang & Zuniga, 2012; Kemper et al., 2014; Kolisch & Sickinger, 2008; Lin et al., 2011; Liu, 2016; Muthuraman & Lawley, 2008; Parizi & Ghate, 2016; Peng et al., 2014; Qu et al., 2007; Ratcliffe et al., 2012; Savelsbergh & Smilowitz, 2016; Schuetz & Kolisch, 2012; Schuetz & Kolisch, 2013; Tang et al., 2014; Tsai & Teng, 2014; Yan et al., 2015; Zacharias & Pinedo, 2014; Zeng et al., 2010)
Wait-dependent	(Huang & Zuniga, 2012; Liu & Ziya, 2014; Liu et al., 2010; Parizi & Ghate, 2016; Patrick, 2012; Qu et al., 2012; Samorani & LaGanga, 2015)
Service-dependent	(Qu et al., 2013)
Time-dependent	(Samorani & LaGanga, 2015; Savelsbergh & Smilowitz, 2016)
Cancellation	(Feldman et al., 2014; Geng, Xie, & Jiang, 2011; Geng & Xie, 2012; Liu et al., 2010; Parizi & Ghate, 2016; Schuetz & Kolisch, 2013)
Patient preference	(Feldman et al., 2014; Gupta & Wang, 2008; Pérez et al., 2011; Savelsbergh & Smilowitz, 2016; Wang & Fung, 2014a; Wang et al., 2015; Wang & Fung, 2014b; Wang & Gupta, 2011; Yan et al., 2015; Zhu et al., 2012)
Random service time	(Anderson et al., 2015; Begen et al., 2012; Begen & Queyranne, 2011; Berg et al., 2014; Castaing et al., 2016; Cayirli & Gunes, 2013; Chakraborty et al., 2010; Chakraborty et al., 2013; Chen & Robinson, 2014; Choi & Banerjee, 2016; Creemers et al., 2012; Denton & Gupta, 2003; Erdogan & Denton, 2013; Erdogan et al., 2015; Ge et al., 2013; Granja et al., 2014; Hassin & Mendel, 2008; Huang et al., 2012; Huang & Zuniga, 2012; Kaandorp & Koole, 2007; Kemper et al., 2014; Klassen & Yoogalingam, 2009; Klassen & Yoogalingam, 2013; Klassen & Yoogalingam, 2014; Koeleman & Koole, 2012; Kong et al., 2013; Kong et al., 2016; Kuiper et al., 2015; Lin et al., 2011; Liu, 2016; Liu & Ziya, 2014; Luo et al., 2012; Mak et al., 2014a, b; Mancilla & Storer, 2012; Min & Yih, 2014; Muthuraman & Lawley, 2008; Oh et al., 2013; Peng et al., 2014; Qi, 2016; Qu et al., 2013; Robinson & Chen, 2003; Saremi et al., 2013; Schuetz & Kolisch, 2012; Tang et al., 2014; Tsai & Teng, 2014; Turkcan et al., 2011; Turkcan et al., 2012; Vink et al., 2015; Wang & Gupta, 2011; Yan et al., 2015; Zeng et al., 2010)
Patient heterogeneity	(Azadeh et al., 2015; Balasubramanian et al., 2014; Balasubramanian et al., 2012; Bikker et al., 2015; Castro & Petrovic, 2012; Chakraborty et al., 2010; Chakraborty et al., 2013; Chen & Robinson, 2014; Conforti et al., 2008, 2010, 2011; Creemers et al., 2012; Dobson et al., 2011; Erdelyi & Topaloglu, 2009; Erdogan & Denton, 2013; Erdogan et al., 2015; Geng & Xie, 2016; Geng, Xie, & Jiang, 2011; Gocgun et al., 2011; Gocgun & Puterman, 2014; Gupta & Wang, 2008; Hahn-Goldberg et al., 2014; Huang et al., 2012; Huang & Zuniga, 2012; Huh et al., 2013; Hulshof et al., 2013; Klassen & Yoogalingam, 2009; Koeleman & Koole, 2012; Kolisch & Sickinger, 2008; Liang & Turkcan, 2015; Lin, 2015; Lin et al., 2011; Liu, 2016; Liu & Ziya, 2014; Mak et al., 2014a, b; Mancilla & Storer, 2012; Min & Yih, 2014; Nguyen et al., 2015; Oh et al., 2013; Ozen & Balasubramanian, 2013; Parizi & Ghate, 2016; Patrick, 2012; Patrick & Puterman, 2007; Patrick et al., 2008; Peng et al., 2014; Pérez et al., 2013; Pérez et al., 2011; Qi, 2016; Qu et al., 2013; Qu et al., 2015; Qu et al., 2012; Qu et al., 2007; Ratcliffe et al., 2012; Riise et al., 2016; Samorani & LaGanga, 2015; Saremi et al., 2013; Saure et al., 2012; Schuetz & Kolisch, 2012; Schuetz & Kolisch, 2013; Sevinc et al., 2013; Tang et al., 2014; Truong, 2015; Tsai & Teng, 2014; Turkcan et al., 2012; Turkcan et al., 2011; Vandaele et al., 2003; Wang & Fung, 2014b; Wang & Gupta, 2011; Wiesche et al. 2016; Yan et al., 2015; Zacharias & Pinedo, 2014; Zeng et al., 2010)
Type of appointment required by patients	
Single	(Anderson et al., 2015; Balasubramanian et al., 2014; Balasubramanian et al., 2012; Begen et al., 2012; Begen & Queyranne, 2011; Berg et al., 2014; Castaing et al., 2016; Cayirli & Gunes, 2013; Cayirli & Veral, 2003; Cayirli & Yang, 2014; Chakraborty et al., 2010; Chakraborty et al., 2013; Chen & Robinson, 2014; Choi & Banerjee, 2016; Creemers et al., 2012; Denton & Gupta, 2003; Dobson et al., 2011; El-Sharo et al., 2015; Erdelyi & Topaloglu, 2009; Erdogan & Denton, 2013; Erdogan et al., 2015; Feldman et al., 2014; Ge et al., 2013; Geng & Xie, 2016; Geng & Xie, 2012; Geng, Xie, Augusto, & Jiang, 2011; Geng, Xie, & Jiang, 2011; Gocgun et al., 2011; Gocgun & Puterman, 2014; Green et al., 2006; Gupta & Wang, 2008; Hahn-Goldberg et al., 2014; Hassin & Mendel, 2008; Huang et al., 2012; Huang & Zuniga, 2012; Huh et al., 2013; Hulshof et al., 2013; Kaandorp & Koole, 2007; Kemper et al., 2014; Kim & Giachetti, 2006; Klassen & Yoogalingam, 2009; Klassen & Yoogalingam, 2013; Klassen & Yoogalingam, 2014; Kong et al., 2016; Koeleman & Koole, 2012; Kortbeek et al., 2014; Kuiper & Mandjes, 2015; LaGanga & Lawrence, 2012; Liang & Turkcan, 2015; Lin et al., 2011; Liu, 2016; Liu & Ziya, 2014; Liu et al., 2010; Luo et al., 2012; Mak et al., 2014a; Mak et al., 2014b; Mancilla & Storer, 2012; Min & Yih, 2014; Mondschein & Weintraub, 2003; Murray & Davies, 2007; Muthuraman & Lawley, 2008; Ozen & Balasubramanian, 2013; Parizi & Ghate, 2016; Patrick, 2012; Patrick & Puterman, 2007; Patrick et al., 2008; Peng et al., 2014; Qi, 2016; Qu et al., 2013; Qu et al., 2015; Qu et al., 2012; Qu et al., 2007; Ratcliffe et al., 2012; Riise et al., 2016; Robinson & Chen, 2003; Samorani & Ganguly, 2016; Samorani & LaGanga, 2015; Saure et al., 2012; Schuetz & Kolisch, 2012; Sevinc et al., 2013; Tai & Williams, 2012; Tang et al., 2014; Truong, 2015; Tsai & Teng, 2014; Turkcan et al., 2012; Turkcan et al., 2011; Vandaele et al., 2003; Vink et al., 2015; Wang & Fung, 2014a; Wang et al., 2015; Wang & Fung, 2014b; Wang & Gupta, 2011; Wiesche et al. 2016; Yan et al., 2015; Zacharias & Pinedo, 2014; Zeng et al., 2010; Zhu et al., 2012)
Combination	(Azadeh et al., 2015; Bikker et al., 2015; Castro & Petrovic, 2012; Hahn-Goldberg et al., 2014; Kuiper et al., 2015; Lin, 2015; Oh et al., 2013; Pérez et al., 2013; Pérez et al., 2011; Saremi et al., 2013; Schuetz & Kolisch, 2013)
Series	(Conforti et al., 2008, 2010, 2011; Granja et al., 2014; Nguyen et al., 2015; Savelsbergh & Smilowitz, 2016; Sevinc et al., 2013; Turkcan et al., 2012)

Table C.4
Papers presenting a case study or using real data for numerical experiments.

Classification	References
Papers with a case study	(Azadeh et al., 2015; Berg et al., 2014; Bikker et al., 2015; Castaing et al., 2016; Conforti et al., 2008, 2011; Erdogan & Denton, 2013; Gocgun & Puterman, 2014; Granja et al., 2014; Hahn-Goldberg et al., 2014; Huang et al., 2012; Kim & Giachetti, 2006; Kong et al., 2013; Lin, 2015; Ozen & Balasubramanian, 2013; Peng et al., 2014; Qu et al., 2013; Riise et al., 2016; Saremi et al., 2013; Saure et al., 2012)
Papers using real data	(Castro & Petrovic, 2012; Cayirli & Gunes, 2013; Feldman et al., 2014; Geng & Xie, 2016; Geng & Xie, 2012; Geng, Xie, Augusto, & Jiang, 2011; Gocgun et al., 2011; Green et al., 2006; Gupta & Wang, 2008; Klassen & Yoogalingam, 2013, 2014; Kolisch & Sickinger, 2008; Liu, 2016; Liu & Ziya, 2014; Liu et al., 2010; Mancilla & Storer, 2012; Nguyen et al., 2015; Oh et al., 2013; Patrick & Puterman, 2007; Pérez et al., 2013; Pérez et al., 2011; Qi, 2016; Qu et al., 2015; Samorani & Ganguly, 2016; Samorani & LaGanga, 2015; Schuetz & Kolisch, 2012, 2013)

References

Ahmadi-Javid, A., Seyedi, P., & Syam, S. A survey of healthcare facility location. *Computers & Operations Research*. doi:10.1016/j.cor.2016.05.018.

Al-Mahdi, I., Gray, K., & Lederman, R. (2015). Online medical consultation: A review of literature and practice. In *Proceedings of the 8th Australasian Workshop on Health Informatics and Knowledge Management (HIKM 2015)* (pp. 97–100).

Allon, G., & Federgruen, A. (2007). Competition in service industries. *Operations Research, 55*(1), 37–55.

Anderson, K., Zheng, B., Yoon, S. W., & Khasawneh, M. T. (2015). An analysis of overlapping appointment scheduling model in an outpatient clinic. *Operations Research for Health Care, 4*, 5–14.

Azadeh, A., Baghersad, M., Farahani, M. H., & Zarrin, M. (2015). Semi-online patient scheduling in pathology laboratories. *Artificial Intelligence in Medicine, 64*(3), 217–226.

Bailey, N. T. (1952). A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting times. *Journal of the Royal Statistical Society. Series B (Methodological), 14*(2), 185–199.

Balasubramanian, H., Biehl, S., Dai, L., & Muriel, A. (2014). Dynamic allocation of same-day requests in multi-physician primary care practices in the presence of prescheduled appointments. *Health Care Management Science, 17*(1), 31–48.

Balasubramanian, H., Muriel, A., & Wang, L. (2012). The impact of provider flexibility and capacity allocation on the performance of primary care practices. *Flexible Services and Manufacturing Journal, 24*(4), 422–447.

Begen, M. A., Levi, R., & Queyranne, M. (2012). Technical note - A sampling-based approach to appointment scheduling. *Operations Research, 60*(3), 675–681.

Begen, M. A., & Queyranne, M. (2011). Appointment scheduling with discrete random durations. *Mathematics of Operations Research, 36*(2), 240–257.

Berg, B. P., Denton, B. T., Erdogan, S. A., Rohleder, T., & Huschka, T. (2014). Optimal booking and scheduling in outpatient procedure centers. *Computers & Operations Research, 50*, 24–37.

Bikker, I. A., Kortbeek, N., van Os, R. M., & Boucherie, R. J. (2015). Reducing access times for radiation treatment by aligning the doctor's schemes. *Operations Research for Health Care, 7*, 111–121.

Brailsford, S., & Vissers, J. (2011). OR in healthcare: A European perspective. *European Journal of Operational Research, 212*(2), 223–234.

Cardoen, B., Demeulemeester, E., & Beliën, J. (2010). Operating room planning and scheduling: A literature review. *European Journal of Operational Research, 201*(3), 921–932.

Castaing, J., Chon, A., Denton, B., & Weizer, A. (2016). A stochastic programming approach to reduce patient wait times and overtime in an outpatient infusion center. *IIE Transactions on Healthcare Systems Engineering, 6*(3), 111–125.

Castro, E., & Petrovic, S. (2012). Combined mathematical programming and heuristics for a radiotherapy pre-treatment scheduling problem. *Journal of Scheduling, 15*(3), 333–346.

Cayirli, T., & Gunes, E. D. (2013). Outpatient appointment scheduling in presence of seasonal walk-ins. *Journal of the Operational Research Society, 65*(4), 512–531.

Cayirli, T., & Veral, E. (2003). Outpatient scheduling in health care: a review of literature. *Production and Operations Management, 12*(4), 519–549.

Cayirli, T., & Yang, K. K. (2014). A universal appointment rule with patient classification for service times, no-shows, and walk-ins. *Service Science, 6*(4), 274–295.

Chakraborty, S., Muthuraman, K., & Lawley, M. (2010). Sequential clinical scheduling with patient no-shows and general service time distributions. *IIE Transactions, 42*(5), 354–366.

Chakraborty, S., Muthuraman, K., & Lawley, M. (2013). Sequential clinical scheduling with patient no-show: The impact of pre-defined slot structures. *Socio-Economic Planning Sciences, 47*(3), 205–219.

Chan, T. C., Killeen, J. P., Castillo, E. M., Vilke, G. M., Guss, D. A., Feinberg, R., & Friedman, L. (2009). Impact of an internet-based emergency department appointment system to access primary care at safety net community clinics. *Annals of Emergency Medicine, 54*(2), 279–284.

Chen, R. R., & Robinson, L. W. (2014). Sequencing and scheduling appointments with potential call-in patients. *Production and Operations Management, 23*(9), 1522–1538.

Choi, S. S., & Banerjee, A. A. (2016). Comparison of a branch-and-bound heuristic, a newsvendor-based heuristic and periodic Bailey rules for outpatients appointment scheduling systems. *Journal of the Operational Research Society, 67*(4), 576–592.

Conforti, D., Guerriero, F., & Guido, R. (2008). Optimization models for radiotherapy patient scheduling. *4OR - A Quarterly Journal of Operations Reserach, 6*(3), 263–278.

Conforti, D., Guerriero, F., Guido, R., & Veltri, M. (2011). An optimal decision-making approach for the management of radiotherapy patients. *OR Spectrum, 33*(1), 123–148.

Conforti, D., Guerriero, F., & Guido, R. (2010). Non-block scheduling with priority for radiotherapy treatments. *European Journal of Operational Research, 201*(1), 289–296.

Creemers, S., Beliën, J., & Lambrecht, M. (2012). The optimal allocation of server time slots over different classes of patients. *European Journal of Operational Research, 219*(3), 508–521.

Denton, B., & Gupta, D. (2003). A sequential bounding approach for optimal appointment scheduling. *IIE Transactions, 35*(11), 1003–1016.

Dobson, G., Hasija, S., & Pinker, E. J. (2011). Reserving capacity for urgent patients in primary care. *Production and Operations Management, 20*(3), 456–473.

El-Sharo, M. D., Zheng, B., Yoon, S. W., & Khasawneh, M. T. (2015). An overbooking scheduling model for outpatient appointments in a multi-provider clinic. *Operations Research for Health Care, 6*, 1–10.

Erdelyi, A., & Topaloglu, H. (2009). Computing protection level policies for dynamic capacity allocation problems by using stochastic approximation methods. *IIE Transactions, 41*(6), 498–510.

Erdogan, S. A., & Denton, B. (2013). Dynamic appointment scheduling of a stochastic server with uncertain demand. *INFORMS Journal on Computing, 25*(1), 116–132.

Erdogan, S. A., Gose, A., & Denton, B. (2015). On-line appointment sequencing and scheduling. *IIE Transactions, 47*(11), 1267–1286.

Feldman, J., Liu, N., Topaloglu, H., & Ziya, S. (2014). Appointment scheduling under patient preference and no-show behavior. *Operations Research, 62*(4), 794–811.

Ge, D., Wan, G., Wang, Z., & Zhang, J. (2013). A note on appointment scheduling with piecewise linear cost functions. *Mathematics of Operations Research, 39*(4), 1244–1251.

Geng, N., & Xie, X. (2016). Optimal Dynamic Outpatient Scheduling for a Diagnostic Facility with Two Waiting Time Targets. *IEEE Transactions on Automatic Control*. doi:10.1109/TAC.2016.2523882.

Geng, N., & Xie, X. (2012). Optimizing contracted resource capacity with two advance cancelation modes. *European Journal of Operational Research, 221*(3), 501–512.

Geng, N., Xie, X., Augusto, V., & Jiang, Z. (2011). A Monte Carlo optimization and dynamic programming approach for managing MRI examinations of stroke patients. *IEEE Transactions on Automatic Control, 56*(11), 2515–2529.

Geng, N., Xie, X., & Jiang, Z. (2011). Capacity reservation and cancellation of critical resources. *IEEE Transactions on Automation Science and Engineering, 8*(3), 470–481.

Gidwani, N., Fernandez, L., & Schlossman, D. (2012). Connecting with patients online: E-visits. *Consulting report prepared for the US Department of Family and Community Medicine Academic Health Center*.

Gocgun, Y., Bresnahan, B. W., Ghate, A., & Gunn, M. L. (2011). A Markov decision process approach to multi-category patient scheduling in a diagnostic facility. *Artificial Intelligence in Medicine, 53*(2), 73–81.

Gocgun, Y., & Puterman, M. L. (2014). Dynamic scheduling with due dates and time windows: An application to chemotherapy patient appointment booking. *Health Care Management Science, 17*(1), 60–76.

Granja, C., Almada-Lobo, B., Janela, F., Seabra, J., & Mendes, A. (2014). An optimization based on simulation approach to the patient admission scheduling problem using a linear programing algorithm. *Journal of Biomedical Informatics, 52*, 427–437.

Green, L. (2006). Queueing analysis in healthcare. In R. W. Hall (Ed.), *Patient flow: Reducing delay in healthcare delivery* (pp. 281–307). Springer.

Green, L. V., & Savin, S. (2008). Reducing delays for medical appointments: A queueing approach. *Operations Research, 56*(6), 1526–1538.

Green, L. V., Savin, S., & Murray, M. (2007). Providing timely access to care: What is the right patient panel size? *The Joint Commission Journal on Quality and Patient Safety, 33*(4), 211–218.

Green, L. V., Savin, S., & Wang, B. (2006). Managing patient service in a diagnostic medical facility. *Operations Research, 54*(1), 11–25.

Greenhalgh, T., Vijayaraghavan, S., Wherton, J., Shaw, S., Byrne, E., Campbell-Richards, D., et al. (2016). Virtual online consultations: Advantages and limitations (VOCAL) study. *BMJ open, 6*(1), e009388.

Griffin, P. (2007). The use of classroom games in management science and operations research. *INFORMS Transactions on Education, 8*(1), 1–2.

Grundgeiger, T., & Sanderson, P. (2009). Interruptions in healthcare: Theoretical views. *International Journal of Medical Informatics, 78*(5), 293–307.

Guerriero, F., & Guido, R. (2011). Operational research in the management of the operating theatre: A survey. *Health Care Management Science, 14*(1), 89–114.

Günal, M. M., & Pidd, M. (2010). Discrete event simulation for performance modelling in health care: A review of the literature. *Journal of Simulation, 4*(1), 42–51.

Gupta, D., & Denton, B. (2008). Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions, 40*(9), 800–819.

Gupta, D., & Wang, L. (2008). Revenue management for a primary-care clinic in the presence of patient choice. *Operations Research, 56*(3), 576–592.

Hahn-Goldberg, S., Carter, M. W., Beck, J. C., Trudeau, M., Sousa, P., & Beattie, K. (2014). Dynamic optimization of chemotherapy outpatient scheduling with uncertainty. *Health Care Management Science, 17*(4), 379–392.

Hassin, R., & Mendel, S. (2008). Scheduling arrivals to queues: A single-server model with no-shows. *Management Science, 54*(3), 565–572.

Huang, Y. L., Hancock, W. M., & Herrin, G. D. (2012). An alternative outpatient scheduling system: Improving the outpatient experience. *IIE Transactions on Healthcare Systems Engineering, 2*(2), 97–111.

Huang, Y., & Zuniga, P. (2012). Dynamic overbooking scheduling system to improve patient access. *Journal of the Operational Research Society, 63*(6), 810–820.

Huh, W. T., Liu, N., & Truong, V. A. (2013). Multiresource allocation scheduling in dynamic environments. *Manufacturing & Service Operations Management, 15*(2), 280–291.

Hulshof, P. J., Boucherie, R. J., Hans, E. W., & Hurink, J. L. (2013). Tactical resource allocation and elective patient admission planning in care processes. *Health Care Management Science, 16*(2), 152–166.

Hulshof, P. J., Kortbeek, N., Boucherie, R. J., Hans, E. W., & Bakker, P. J. (2012). Taxonomic classification of planning decisions in health care: A structured review of the state of the art in OR/MS. *Health Systems, 1*(2), 129–175.

Hulshof, P. J., Vanberkel, P. T., Boucherie, R. J., Hans, E. W., van Houdenhoven, M., & van Ommeren, J. K. C. (2012). Analytical models to determine room requirements in outpatient clinics. *OR Spectrum, 34*(2), 391–405.

Jaber, R., Braksmajer, A., & Trilling, J. S. (2006). Group visits: A qualitative review of current research. *The Journal of the American Board of Family Medicine, 19*(3), 276–290.

Kaandorp, G. C., & Koole, G. (2007). Optimal outpatient appointment scheduling. *Health Care Management Science, 10*(3), 217–229.

Kemper, B., Klaassen, C. A., & Mandjes, M. (2014). Optimized appointment scheduling. *European Journal of Operational Research, 239*(1), 243–255.

Kim, S., & Giachetti, R. E. (2006). A stochastic mathematical appointment overbooking model for healthcare providers to improve profits. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 36*(6), 1211–1219.

Kim, S., Pasupathy, R., & Henderson, S. G. (2015). A guide to sample average approximation. In M. C. Fu (Ed.), *Handbook of Simulation Optimization* (pp. 207–243). Springer.

Klassen, K. J., & Yoogalingam, R. (2009). Improving performance in outpatient appointment services with a simulation optimization approach. *Production and Operations Management, 18*(4), 447–458.

Klassen, K. J., & Yoogalingam, R. (2013). Appointment system design with interruptions and physician lateness. *International Journal of Operations & Production Management, 33*(4), 394–414.

Klassen, K. J., & Yoogalingam, R. (2014). Strategies for appointment policy design with patient unpunctuality. *Decision Sciences, 45*(5), 881–911.

Kleindorfer, P. R., & Saad, G. H. (2005). Managing disruption risks in supply chains. *Production and Operations Management, 14*(1), 53–68.

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*(1), 14–30.

Kolisch, R., & Sickinger, S. (2008). Providing radiology health care services to stochastic demand of different customer classes. *OR Spectrum, 30*(2), 375–395.

Kong, Q., Lee, C. Y., Teo, C. P., & Zheng, Z. (2016). Appointment sequencing: Why the smallest-variance-first rule may not be optimal. *European Journal of Operational Research, 255*(3), 809–821.

Kong, Q., Lee, C. Y., Teo, C. P., & Zheng, Z. (2013). Scheduling arrivals to a stochastic service delivery system using copositive cones. *Operations Research, 61*(3), 711–726.

Kortbeek, N., Zonderland, M. E., Braaksma, A., Vliegen, I. M., Boucherie, R. J., Litvak, N., & Hans, E. W. (2014). Designing cyclic appointment schedules for outpatient clinics with scheduled and unscheduled patient arrivals. *Performance Evaluation, 80*, 5–26.

Krishnamoorthy, A., Pramod, P., & Chakravarthy, S. (2014). Queues with interruptions: A survey. *TOP, 22*(1), 290–320.

Kuiper, A., Kemper, B., & Mandjes, M. (2015). A computational approach to optimized appointment scheduling. *Queueing Systems, 79*(1), 5–36.

Kuiper, A., & Mandjes, M. (2015). Appointment scheduling in tandem-type service systems. *Omega, 57*, 145–156.

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*(5), 874–888.

Li, Y., Kong, N., Chen, M., & Zheng, Q. P. (2016). Optimal physician assignment and patient demand allocation in an outpatient care network. *Computers & Operations Research, 72*, 107–117.

Liang, B., & Turkcan, A. (2015). Acuity-based nurse assignment and patient scheduling in oncology clinics. *Health Care Management Science, 19*(3), 207–226.

Lin, C. K. Y. (2015). An adaptive scheduling heuristic with memory for the block appointment system of an outpatient specialty clinic. *International Journal of Production Research, 53*(24), 7488–7516.

Lin, J., Muthuraman, K., & Lawley, M. (2011). Optimal and approximate algorithms for sequential clinical scheduling with no-shows. *IIE Transactions on Healthcare Systems Engineering, 1*(1), 20–36.

Liu, N. (2016). Optimal choice for appointment scheduling window under patient no-show behavior. *Production and Operations Management, 25*(1), 128–142.

Liu, N., & Ziya, S. (2014). Panel size and overbooking decisions for appointment-based services under patient no-shows. *Production and Operations Management, 23*(12), 2209–2223.

Liu, N., Ziya, S., & Kulkarni, V. G. (2010). Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. *Manufacturing & Service Operations Management, 12*(2), 347–364.

Luo, J., Kulkarni, V. G., & Ziya, S. (2012). Appointment scheduling under patient no-shows and service interruptions. *Manufacturing & Service Operations Management, 14*(4), 670–684.

Mak, H. Y., Rong, Y., & Zhang, J. (2014a). Appointment scheduling with limited distributional information. *Management Science, 61*(2), 316–334.

Mak, H. Y., Rong, Y., & Zhang, J. (2014b). Sequencing appointments for service systems using inventory approximations. *Manufacturing & Service Operations Management, 16*(2), 251–262.

Mancilla, C., & Storer, R. (2012). A sample average approximation approach to stochastic appointment sequencing and scheduling. *IIE Transactions, 44*(8), 655–670.

May, J. H., Spangler, W. E., Strum, D. P., & Vargas, L. G. (2011). The surgical scheduling problem: Current research and future opportunities. *Production and Operations Management, 20*(3), 392–405.

Menditto, A., Patriarca, M., & Magnusson, B. (2007). Understanding the meaning of accuracy, trueness and precision. *Accreditation and Quality Assurance, 12*(1), 45–47.

Mitchell, P., Wynia, M., Golden, R., McNellis, B., Okun, S., Webb, C. E., et al. (2012). *Core principles and values of effective team-based health care*. Washington, DC: Institute of Medicine.

Min, D., & Yih, Y. (2014). Managing a patient waiting list with time-dependent priority and adverse events. *RAIRO-Operations Research, 48*(1), 53–74.

Mondschein, S. V., & Weintraub, G. Y. (2003). Appointment policies in service operations: A critical analysis of the economic framework. *Production and Operations Management, 12*(2), 266–286.

Murnik, M., Randal, F., Guevara, M., Skipper, B., & Kaufman, A. (2006). Web-based primary care referral program associated with reduced emergency department utilization. *Family Medicine, 38*(3), 185.

Murray, M., & Davies, M. (2007). Panel size: How many patients can one doctor manage. *Family Practice Management, 14*(4), 44–51.

Murray, M., & Tantau, C. (2000). Same-day appointments: Exploding the access paradigm. *Family Practice Management, 7*(8), 45.

Muthuraman, K., & Lawley, M. (2008). A stochastic overbooking model for outpatient clinical scheduling with no-shows. *IIE Transactions, 40*(9), 820–837.

Nguyen, T. B. T., Sivakumar, A. I., & Graves, S. C. (2015). A network flow approach for tactical resource planning in outpatient clinics. *Health Care Management Science, 18*(2), 124–136.

Oh, H. J., Muriel, A., Balasubramanian, H., Atkinson, K., & Ptaszkiewicz, T. (2013). Guidelines for scheduling in primary care under different patient types and stochastic nurse and provider service times. *IIE Transactions on Healthcare Systems Engineering, 3*(4), 263–279.

OnceCity Health. (2016). Initiating ED care triage phase one. Retrived from http://www.onecityhealth.org/initiating-ed-care-triage-phase-one/ April 2016.

Ozcan, Y. A. (2009). *Quantitative methods in health care management: Techniques and applications* (2th edition). John Wiley & Sons.

Ozen, A., & Balasubramanian, H. (2013). The impact of case mix on timely access to appointments in a primary care group practice. *Health Care Management Science, 16*(2), 101–118.

Parizi, M. S., & Ghate, A. (2016). Multi-class, multi-resource advance scheduling with no-shows, cancellations and overbooking. *Computers & Operations Research, 67*, 90–101.

Patrick, J. (2012). A Markov decision model for determining optimal outpatient scheduling. *Health Care Management Science, 15*(2), 91–102.

Patrick, J., & Puterman, M. L. (2007). Improving resource utilization for diagnostic services through flexible inpatient scheduling: A method for improving resource utilization. *Journal of the Operational Research Society, 58*(2), 235–245.

Patrick, J., Puterman, M. L., & Queyranne, M. (2008). Dynamic multipriority patient scheduling for a diagnostic resource. *Operations Research, 56*(6), 1507–1525.

Peng, Y., Qu, X., & Shi, J. (2014). A hybrid simulation and genetic algorithm approach to determine the optimal scheduling templates for open access clinics admitting walk-in patients. *Computers & Industrial Engineering, 72*, 282–296.

Pérez, E., Ntaimo, L., Malavé, C. O., Bailey, C., & McCormack, P. (2013). Stochastic online appointment scheduling of multi-step sequential procedures in nuclear medicine. *Health Care Management Science, 16*(4), 281–299.

Pérez, E., Ntaimo, L., Wilhelm, W. E., Bailey, C., & McCormack, P. (2011). Patient and resource scheduling of multi-step medical procedures in nuclear medicine. *IIE Transactions on Healthcare Systems Engineering, 1*(3), 168–184.

Popejoy, L. L., Stetzer, F., Hicks, L., Rantz, M. J., Galambos, C., Popescu, M., Khalilia, M. A., & Marek, K. D. (2015). Comparing aging in place to home health care: Impact of nurse care coordination on utilization and costs. *Nursing Economics, 33*(6), 306.

Preater, J. (2002). Queues in health. *Health Care Management Science, 5*(4), 283.

Qi, J. (2016). Mitigating delays and unfairness in appointment systems. *Management Science*. doi:10.1287/mnsc.2015.2353.
Qu, X., Peng, Y., Kong, N., & Shi, J. (2013). A two-phase approach to scheduling multi-category outpatient appointments – A case study of a women's clinic. *Health Care Management Science, 16*(3), 197–216.
Qu, X., Peng, Y., Shi, J., & LaGanga, L. (2015). An MDP model for walk-in patient admission management in primary care clinics. *International Journal of Production Economics, 168*, 303–320.
Qu, X., Rardin, R. L., & Williams, J. A. S. (2011). Single versus hybrid time horizons for open access scheduling. *Computers & Industrial Engineering, 60*(1), 56–65.
Qu, X., Rardin, R. L., & Williams, J. A. S. (2012). A mean–variance model to optimize the fixed versus open appointment percentages in open access scheduling systems. *Decision Support Systems, 53*(3), 554–564.
Qu, X., Rardin, R. L., Williams, J. A. S., & Willis, D. R. (2007). Matching daily healthcare provider capacity to demand in advanced access scheduling systems. *European Journal of Operational Research, 183*(2), 812–826.
Rais, A., & Viana, A. (2011). Operations research in healthcare: A survey. *International Transactions in Operational Research, 18*(1), 1–31.
Ratcliffe, A., Gilland, W., & Marucheck, A. (2012). Revenue management for outpatient appointments: joint capacity control and overbooking with class-dependent no-shows. *Flexible Services and Manufacturing Journal, 24*(4), 516–548.
Riise, A., Mannino, C., & Lamorgese, L. (2016). Recursive logic-based Benders' decomposition for multi-mode outpatient scheduling. *European Journal of Operational Research, 225*(3), 719–728.
Rivera-Rodriguez, A., & Karsh, B. T. (2010). Interruptions and distractions in healthcare: Review and reappraisal. *Quality and Safety in Health Care, 19*, 304–312.
Robinson, L. W., & Chen, R. R. (2003). Scheduling doctors' appointments: Optimal and empirically-based heuristic policies. *IIE Transactions, 35*(3), 295–307.
Robinson, L. W., & Chen, R. R. (2010). A comparison of traditional and open-access policies for appointment scheduling. *Manufacturing & Service Operations Management, 12*(2), 330–346.
Rohleder, T. R., Lewkonia, P., Bischak, D. P., Duffy, P., & Hendijani, R. (2011). Using simulation modeling to improve patient flow at an outpatient orthopedic clinic. *Health Care Management Science, 14*(2), 135–145.
Saghafian, S., Austin, G., & Traub, S. J. (2015). Operations research/management contributions to emergency department patient flow optimization: Review and research prospects. *IIE Transactions on Healthcare Systems Engineering, 5*(2), 101–123.
Samorani, M., & Ganguly, S. (2016). Optimal sequencing of unpunctual patients in high-service-level clinics. *Production and Operations Management, 25*(2), 330–346.
Samorani, M., & LaGanga, L. R. (2015). Outpatient appointment scheduling given individual day-dependent no-show predictions. *European Journal of Operational Research, 240*(1), 245–257.
Saremi, A., Jula, P., ElMekkawy, T., & Wang, G. G. (2013). Appointment scheduling of outpatient surgical services in a multistage operating room department. *International Journal of Production Economics, 141*(2), 646–658.
Saure, A., Patrick, J., Tyldesley, S., & Puterman, M. L. (2012). Dynamic multi-appointment patient scheduling for radiation therapy. *European Journal of Operational Research, 223*(2), 573–584.
Sauré, A., & Puterman, M. L. (2014). The appointment scheduling game. *INFORMS Transactions on Education, 14*(2), 73–85.
Savelsbergh, M., & Smilowitz, K. (2016). Stratified patient appointment scheduling for mobile community-based chronic disease management programs. *IIE Transactions on Healthcare Systems Engineering, 6*(2), 65–78.
Schuetz, H. J., & Kolisch, R. (2012). Approximate dynamic programming for capacity allocation in the service industry. *European Journal of Operational Research, 218*(1), 239–250.
Schuetz, H. J., & Kolisch, R. (2013). Capacity allocation for demand of different customer-product-combinations with cancellations, no-shows, and overbooking when there is a sequential delivery of service. *Annals of Operations Research, 206*(1), 401–423.
Sevinc, S., Sanli, U. A., & Goker, E. (2013). Algorithms for scheduling of chemotherapy plans. *Computers in Biology and Medicine, 43*(12), 2103–2109.
Snyder, L. V., Scaparra, M. P., Daskin, M. S., & Church, R. L. (2006). Planning for disruptions in supply chain networks. *Tutorials in Operations Research, 2*, 234–257.
Tai, G., & Williams, P. (2012). Optimization of scheduling patient appointments in clinics using a novel modelling technique of patient arrival. *Computer Methods and Programs in Biomedicine, 108*(2), 467–476.
Tang, J., Yan, C., & Fung, R. Y. (2014). Optimal appointment scheduling with no-shows and exponential service time considering overtime work. *Journal of Management Analytics, 1*(2), 99–129.
Truong, V. A. (2015). Optimal advance scheduling. *Management Science, 61*(7), 1584–1597.
Tsai, P. F. J., & Teng, G. Y. (2014). A stochastic appointment scheduling system on multiple resources with dynamic call-in sequence and patient no-shows for an outpatient clinic. *European Journal of Operational Research, 239*(2), 427–436.
Turkcan, A., Zeng, B., & Lawley, M. (2012). Chemotherapy operations planning and scheduling. *IIE Transactions on Healthcare Systems Engineering, 2*(1), 31–49.
Turkcan, A., Zeng, B., Muthuraman, K., & Lawley, M. (2011). Sequential clinical scheduling with service criteria. *European Journal of Operational Research, 214*(3), 780–795.
Vandaele, N., Van Nieuwenhuyse, I., & Cupers, S. (2003). Optimal grouping for a nuclear magnetic resonance scanner by means of an open queueing model. *European Journal of Operational Research, 151*(1), 181–192.
Vink, W., Kuiper, A., Kemper, B., & Bhulai, S. (2015). Optimal appointment scheduling in continuous time: The lag order approximation method. *European Journal of Operational Research, 240*(1), 213–219.
Wang, J., & Fung, R. Y. (2014a). Adaptive dynamic programming algorithms for sequential appointment scheduling with patient preferences. *Artificial Intelligence in Medicine, 63*(1), 33–40.
Wang, J., & Fung, R. Y. (2014b). An integer programming formulation for outpatient scheduling with patient preference. *Industrial Engineering & Management Systems, 13*(2), 193–202.
Wang, J., Fung, R. Y., & Chan, H. K. (2015). Dynamic appointment scheduling with patient preferences and choices. *Industrial Management & Data Systems, 115*(4), 700–717.
Wang, P. P. (1993). Static and dynamic scheduling of customer arrivals to a single-server system. *Naval Research Logistics, 40*(3), 345–360.
Wang, W. Y., & Gupta, D. (2011). Adaptive appointment systems with patient preferences. *Manufacturing & Service Operations Management, 13*(3), 373–389.
Weiner, M., El Hoyek, G., Wang, L., Dexter, P. R., Zerr, A. D., Perkins, A. J., James, F., & Juneja, R. (2009). A web-based generalist–specialist system to improve scheduling of outpatient specialty consultations in an academic center. *Journal of General Internal Medicine, 24*(6), 710–715.
Wexler, R., Hefner, J. L., Sieck, C., Taylor, C. A., Lehman, J., Panchal, A. R., Aldrich, A., & McAlearney, A. S. (2015). Connecting emergency department patients to primary care. *The Journal of the American Board of Family Medicine, 28*(6), 722–732.
Wiesche, L., Schacht, M., & Werners, B. (2016). Strategies for interday appointment scheduling in primary care. *Health Care Management Science*. doi:10.1007/s10729-016-9361-7.
Yan, C., Tang, J., Jiang, B., & Fung, R. Y. (2015). Sequential appointment scheduling considering patient choice and service fairness. *International Journal of Production Research, 53*(24), 7376–7395.
Yuan, B., Liu, R., & Jiang, Z. (2015). A branch-and-price algorithm for the home health care scheduling and routing problem with stochastic service times and skill requirements. *International Journal of Production Research, 53*(24), 7450–7464.
Zacharias, C., & Pinedo, M. (2014). Appointment Scheduling with No-Shows and Overbooking. *Production and Operations Management, 23*(5), 788–801.
Zeng, B., Turkcan, A., Lin, J., & Lawley, M. (2010). Clinic scheduling models with overbooking for patients with heterogeneous no-show probabilities. *Annals of Operations Research, 178*(1), 121–144.
Zhu, H., Hou, M., Wang, C., & Zhou, M. (2012). An efficient outpatient scheduling approach. *IEEE Transactions on Automation Science and Engineering, 9*(4), 701–709.