IIE Transactions (2008) **40**, 800–819

ISSN: 0740-817X print / 1545-8830 online
DOI: 10.1080/07408170802165880

Appointment scheduling in health care: Challenges and opportunities

DIWAKAR GUPTA[1,*] and BRIAN DENTON[2]

[1]*Graduate Program in Industrial & Systems Engineering, Department of Mechanical Engineering, University of Minnesota, 111 Church Street S.E., Minneapolis, MN 55455, USA*
E-mail: guptad@me.umn.edu
[2]*North Carolina State University, Edward P. Fitts Department of Industrial & Systems Engineering, 111 Lampe Drive, Raleigh, NC 27695, USA*
E-mail: bdenton@ncsu.edu

Received August 2006 and accepted December 2007

Appointment scheduling systems are used by primary and specialty care clinics to manage access to service providers, as well as by hospitals to schedule elective surgeries. Many factors affect the performance of appointment systems including arrival and service time variability, patient and provider preferences, available information technology and the experience level of the scheduling staff. In addition, a critical bottleneck lies in the application of Industrial Engineering and Operations Research (IE/OR) techniques. The most common types of health care delivery systems are described in this article with particular attention on the factors that make appointment scheduling challenging. For each environment relevant decisions ranging from a set of rules that guide schedulers to real-time responses to deviations from plans are described. A road map of the state of the art in the design of appointment management systems is provided and future opportunities for novel applications of IE/OR models are identified.

Keywords: Appointment scheduling, health care operations management, access rules

1. Introduction

The health care industry represents approximately 15% of the gross domestic product of the United States. Health care expenditures are growing at a rate such that the amount of public money needed to finance health care, which currently stands at 45% of all health care expenditures, is expected to double by 2050. Americans who are covered by employer-sponsored plans have experienced double digit increases in premiums as employers shift a greater portion of rapidly rising health care costs to the employees (Economist, 2004). When we add factors such as an aging population, increasing demand for chronic care and strained public and private health care budgets to this mix, it is no surprise that there is a growing pressure on health service providers to improve efficiency.

Appointment scheduling systems lie at the intersection of efficiency and timely access to health services. Timely access is important for realizing good medical outcomes. It is also an important determinant of patient satisfaction. The ability to provide timely access is determined by a variety of factors that include fundamental questions about how many and which types of physical assets and equipment a health system should invest in, how should it allocate resources among multiple sites, how should it staff each clinic or hospital site, what rules best determine which providers and patients receive higher priority access to resources, and how appointments are scheduled. The focus of this article is on appointment scheduling. We do not consider questions pertaining to the size of facilities, equipment and staff, and to resource allocation in multiple-service-site systems. (An example of such issues can be found in Chao *et al.* (2003)).

Scheduled patient encounters include primary and specialty care visits, as well as elective surgeries. In each of these environments, the process of scheduling appointments (assigning a specific time when the patient is scheduled to start receiving care) is different, which we will describe shortly. In addition, there are unscheduled encounters that include walk-ins and urgent or emergency cases. The former, occurring mostly in primary care clinics, can be directed to an alternate facility if the clinic in question is heavily booked. However, urgent specialty care and surgical patients often need to be treated as soon as possible. The goal of a well-designed appointment system is to deliver timely and convenient access to health services for all patients. Appointment systems also smooth work flow, reduce crowding in waiting

*Corresponding author

rooms and allow health systems to honor patient and provider preferences while matching supply and demand.

Each of primary care, specialty care and hospital services have certain unique features that give rise to different challenges for managing appointments. In the primary care setting, the vast majority of patients require services that can be performed within a fixed time length. Therefore, primary care clinics tend to divide available provider time into equal-length time slots such that, by and large, patients' needs can be accommodated in a standard appointment slot. For certain types of visits that require more time, clinics may assign multiple appointment slots. The appointment scheduling problem then reduces to that of finding a suitable match among the available time slots of providers in the clinic, provider prescribed restrictions on how available slots may be filled and patients' preferences for day/time of week as well as for a particular service provider. (An example of provider restrictions on the use of available slots is the limit that many providers place on the number of physical exams or new patients that can be scheduled in any given session.) Still, the problem of matching supply and demand is not easy because different patients have different perceptions of the urgency of their need and different day-of-week and time-of-day preference patterns.

Patient service times in specialty care clinics tend to vary more depending on the patients' diagnoses and other characteristics. Therefore, provider time may not be divided into standard time slots. Moreover, many specialty services require a referral from the primary care physician. In such cases, appointments are booked by the medical assistant of the referring doctor. Bookings may occur at periodic intervals (e.g., at the end of each day). Appointment management for specialty care clinics is further complicated because of two reasons: (i) the need to reserve capacity for urgent appointment requests that must be treated soon after they occur; and (ii) the need to realize high utilization of more-expensive specialists' time.

Scheduling surgical appointments is even more complex. Procedure times are variable, several pre-surgery appointments may be required for necessary medical exams and a variety of service providers/resources have to be simultaneously scheduled in order to deliver the desired services. For example, in addition to the team of surgeons, a properly equipped surgery room, specialized nursing staff and anesthesiologists have to be available at the desired start time. Therefore, surgery scheduling sometimes occurs in two stages. Patients first choose from a menu of available time windows (each may be a week long) during which they prefer to have the procedure performed. The physician's office later confirms a specific day and surgery start time, which we refer to as the appointment. Surgeons typically need to fit all procedures scheduled for a day within a block of operating room time that is assigned for their use. They have preferences with respect to which types of procedures they like to perform on specific days and times of week. Such considerations can further complicate scheduling. Finally, a significant proportion of overall demand in a variety of surgical specialty areas is from urgent patients who need to be treated as soon as possible.

In all three environments, a patient who schedules an appointment faces two types of access delays. *Indirect* (virtual) waiting time is the difference between the time that a patient requests an appointment and the time of that appointment. *Direct* (captive) waiting time is the difference between a patient's appointment time (or his/her arrival time if he/she is tardy) and the time when he/she is actually served by the service provider. In contrast, in a system with no appointments, e.g., when a service provider accepts only walk-ins or urgent cases, patients experience only direct waiting. Indirect waiting is usually orders of magnitude greater than direct waiting. Whereas direct waiting is an inconvenience to the patient, excessive indirect wait can pose a serious safety concern (Murray and Berwick, 2003).

A well-designed appointment system achieves small direct waiting times for unscheduled (especially urgent/emergency) episodes without increasing the direct waiting times of scheduled patients or lowering resource utilization. This is accomplished by specifying various "rules" that determine which types of patients may access available service provider resources at what times. We refer to such rules as access rules and include them within the scope of this article. Access rules also reduce the negative impact of indirect waiting on scheduled urgent appointments by reserving some capacity exclusively for their use.

Appointment systems can be a source of dissatisfaction, both for the patients and for the providers. Patients are impacted by the lack of availability of timely and convenient appointment slots, especially when their need is urgent. Clinicians are impacted by the uncertainty in the number of patient appointments from day to day, and the mix of appointments on any given day. These factors can affect clinicians' earnings as well as their job satisfaction levels. In many instances, clinicians can handle high-priority demand, and variations in case mix, only by stretching their schedules to absorb demand variation—i.e., by shrinking lunch time, pushing back dinner and double booking (working faster). (The soft nature of provider capacity is one of the factors that differentiates health care delivery systems from manufacturing, transportation and logistics systems.) Even with such strategies in place, it is sometimes necessary to reschedule certain booked appointments for non-urgent services in order to take care of urgent demand. Moreover, significant direct waiting time is not uncommon in environments that deal with life-threatening urgent cases. Frequent involuntary changes in appointments and long direct waits can cause dissatisfaction among patients who book in advance.

There are many factors that affect the ability of appointment schedulers to utilize available providers' time efficiently and effectively. Some examples include inter-arrival and service time variability, cancelations and no-shows, patient preferences for a particular day of week, time of

day and for certain physicians, degree of flexibility permitted by the physicians in the use of their time (e.g., some preauthorize double booking at certain times of the day if demand is high, whereas others require the scheduler to call for approval each time), appropriate level of information technology, and a smooth-running call center for managing patient requests. In addition to these issues, we believe that a critical bottleneck lies with the application of Industrial Engineering & Operations Research (IE/OR) models. Such models have the potential to improve appointment scheduling via algorithmic decision support tools, similar to their successful application in other service industries such as the airlines, car rental agencies and hotels (see Talluri and Van Ryzin (2004a)).

We believe that IE/OR decision support techniques can simultaneously reduce costs and improve access to health services. The purpose of this article is to provide a critical survey of the state of the art in modeling and optimization of appointment scheduling problems. We consider both direct and indirect patient delays and provide a synthesis of previous research pertaining to all three scheduling environments mentioned earlier. Moreover, our approach views the scheduling problems arising in the three environments as different application domains within a common underlying modeling framework. We also discuss opportunities for IE/OR community to make significant future contributions toward solving health care appointment scheduling problems. Our focus is on appointment scheduling by which we mean three things: (i) choice of access rules; (ii) encounter start times; and (iii) approaches for handling differences between scheduled and realized supply/demand. We do not consider issues related to the size of facilities, equipment and staff.

This article is organized as follows. In Section 2 we describe the appointment scheduling environments and decisions in primary care clinics, specialty clinics and hospitals. In Section 3 we classify the underlying complexity of scheduling appointments according to four key factors. This helps to categorize and critique the relevant IE/OR literature, which we do in Section 4. Section 5 summarizes several areas that represent opportunities for future research. Finally, in Section 6, we conclude the paper by commenting on the role of Electronic Medical Records (EMR) in particular, and Health Information Technology (HIT) in general, as vehicles for deploying IE/OR-based decision support systems. While some of the above-mentioned discussions focus greater attention on problems that are common to the US health care system, much of the underlying content of this article is generalizable to appointment scheduling issues in any health care delivery system.

2. Scheduling environments and decisions

The ensuing descriptions of the three commonly encountered health care scheduling environments are based on the authors' first-hand knowledge of the systems used by several health service providers. They do not reflect specific practices of any one provider. Furthermore, they do not cover all possible variations found in practice. Instead, our goal is to paint a picture of the typical scheduling environment in each setting.

We focus on three topics—access rules, encounter start times and approaches for handling differences between the scheduled and the realized supply/demand as the day unfolds. (Note that sometimes appointment scheduling is understood to imply only the task of setting encounter start times.) Access rules help sort patients into different priority classes, specify access targets and the amount of reserved capacity for each class, and guide managers' response to the variation between realized and scheduled demand and supply. Encounter start times specify the date and time when service providers and patients are expected to be ready for the examination or procedure. Differences between scheduled and actual demand/supply are common. They can arise as a result of longer than expected service times, provider/patient tardiness, late cancelations and no-shows, and unanticipated urgent/emergency demand. Hereafter, we refer to the latter decisions as *daily scheduling* decisions. Equipment/staff capacity choices and staff scheduling also affect appointment scheduling; they are not considered in this paper.

2.1. *Primary care appointments scheduling*

Historically, primary care practices were the quintessential cottage industry within the system of delivering health care. Physician-owned and -managed clinics were typically served by a single doctor or a small group of doctors, who took care of the medical needs of families from cradle to grave. Appointment systems tended to be manual and *ad hoc*, and physicians often worked variable hours to provide needed service to urgent requests. In recent years, clinics have grown larger. Often, they are a part of an even larger health care delivery system (or network) comprising many primary care and specialty clinics and hospitals. Modern clinics also have on-site lab facilities and X-ray machines for carrying out routine diagnostic tests.

When faced with a medical problem, patients often contact their Primary Care Physicians (PCPs) first, with the result that PCPs are sometimes called the gatekeepers of the health care delivery system. In a majority of cases, patients call in advance to book an appointment; however, some do walk in. Some clinics have on-site appointment schedulers. Increasingly, however, appointment booking operations are centralized at a remote call center, which serves many clinics belonging to a health care network.

Physicians divide their available clinic time into appointment slots, which are usually between 15 to 30 minutes long. In addition, providers determine the number of standard slots needed for each category of appointment request. Certain types of appointments, e.g., physical exams, require

multiple slots. Other types, e.g., routine follow-up visits, require a single slot. Providers choose start and end times of their work schedule for each day over a pre-specified period of time (say 4 weeks) several weeks in advance of that time period. They also provide schedulers with any restrictions on how available slots may be assigned to incoming requests for appointments. For example, certain appointment slots are reserved each day for physical exams, which makes it easy for schedulers to find contiguous open slots needed for such appointments. This practice also helps physicians plan their day in advance.

Access rules in the primary care environment reserve certain slots exclusively for certain types of patients. Patient types could depend on medical urgency, type of service requested and on whether or not the patient belongs to the physician's panel. For each physician, his/her panel consists of all those patients who have designated him/her as their PCP. Matching patients with their PCPs is important for continuity (quality) of care and for clinic efficiency because otherwise physicians' end up spending more time reading medical histories of unfamiliar patients.

A recent innovation that has been adopted by many primary care clinics is the Advanced or Open Access system, credited to Murray and Tantau (1999, 2000). In this approach, physicians attempt to accommodate patients' requests for appointments on the day they call. (This is not to be confused with walk-ins who do not call in advance. We discuss walk-ins at a later point in this section.) Future appointments at a time that is more convenient for the patient are also permitted. Service providers vary available capacity to meet each day's demand. The ability of a patient to book an appointment on the day (s)he calls is no longer a function of his/her medical condition. In contrast, clinics that do not offer Advanced Access often employ a triage nurse to assess the urgency of medical need of a caller who requests an appointment without delay. Only those callers whose need is deemed urgent are offered one of several slots reserved each day for urgent requests.

The impetus for adopting Advanced Access comes from the desire to make clinic practices more patient focused, to accommodate faster access for patients with urgent needs and to gain competitive advantage. This approach also eliminates the need for a triage nurse. However, the implementation of Advanced Access systems remains a challenge because of a variety of reasons. First, even when providers work hard to absorb variations in daily demand, their ability to do so is limited. Therefore, it may not be possible to accommodate all appointment requests on the day they originate. This leads to demand spillover to a future day, limiting the clinic's ability to meet that future day's demand. Second, the true demand for same-day service is not captured by the appointments data because it is difficult to tell whether a patient actually preferred to book an appointment on a future date or (s)he did so because a same-day appointment was not available. This makes it difficult for clinic directors to determine approximately how much capacity should be available at the start of each day for that day's demand. Third, in many clinics, different physicians' panel compositions and sizes are significantly different, with the result that some physicians have fewer available slots to accommodate same-day demand.

In addition to deciding how to schedule appointment requests from patients who call in advance, primary care clinics also need to decide how to respond to walk-ins, and any unplanned shortfall in capacity (due to provider illness or emergency) while minimizing their impact on the direct waiting time of patients with scheduled appointments. We call such decisions by clinic managers daily scheduling decisions. Daily scheduling has become increasingly more important as many clinics have adopted Advanced Access systems and because the amount of excess capacity that a provider can make available in response to greater than anticipated demand is limited. Clinic directors recognize that patients' perceptions of urgency of need are an important factor in determining their overall satisfaction with the timeliness of access. Therefore, many are experimenting with alternative ways to accommodate daily scheduling variation, such as pooling provider appointment slots for urgent care, using nurse practitioners and doctors' assistants, and forming provider teams. With a provider team in place, when an appointment with a patient's PCP is not available, (s)he is offered an appointment with a member of the care team who is somewhat familiar with the patient's medical history.

2.2. *Specialty clinic appointments scheduling*

Specialty care clinics are designed to deliver health services that are focused on specific, often complex, diagnoses and treatments. In some cases, multiple medical specialties may be integrated into a group practice which may have several departments, each specializing in a different branch of medicine. Rules governing access to specialists can vary by the medical specialty, as well as by the health network. Certain specialties such as pediatrics and obstetrics are typically designated as open access. This means that patients can call to book an appointment without the need for a referral first. Open access clinics are similar to primary care clinics in terms of appointment scheduling. In fact, it is not uncommon to find a shared call center, which serves all primary care and open access specialty clinics in a health care network. Many specialty clinics do require referrals. In that case, the referring physician is often the patient's PCP and his/her clinical assistant books an appointment for the patient. In many instances, a referral is required only for the first appointment, and the patient is able to directly schedule all subsequent appointments.

Unlike a primary care environment where most services can be performed within a fixed-length appointment slot, specialists' appointment lengths can be highly variable and diagnosis dependent. Different patients have different urgency of need and quick access is critical to realizing good

medical outcomes for urgent cases. In addition, a specialist is a more expensive resource for the health service network, and certain specialty areas have a shortage of qualified physicians. Thus, specialty clinics face the difficult task of simultaneously guaranteeing quick access for high-priority cases and realizing high utilization of the specialist's time. This is accomplished via a combination of access rules that govern advance booking and daily scheduling approaches for managing supply–demand mismatches.

Access rules help clinics determine how much capacity to reserve for each type (or length) of appointment and for future callers with more urgent needs. These rules also determine planned appointment lengths for each diagnosis of the referring physicians. In addition to medical urgency, capacity reservation can also be driven by the need to serve out-of-town patients who are unable to take advantage of near-term appointments on account of travel delays, and to accommodate patient cross-flows among several specialty clinics. A cross-flow occurs when a patient discovers during an appointment with a particular specialist that (s)he needs an appointment with a different specialist in order to complete his/her diagnosis/treatment. Access rules also help improve capacity utilization by making it easier to fit appointments of varying lengths in a provider's daily schedule.

Daily scheduling concerns taking care of deviations from planned clinic time and booked appointments, both of which are common in specialty services. For example, when outpatient clinics are attached to hospitals, specialists may serve as stand-by consultants and providers of emergency care. They may also see patients on short notice to clear them as surgical candidates, with the result that clinic managers face uncertainty in both demand and physician availability. Managing short-term supply and demand imbalance in specialty care environments is particularly difficult in rural or less populated areas. In such cases health care providers may collaborate to pool resources, for instance, by having a rotating specialist-on-call schedule.

2.3. *Scheduling elective surgery appointments*

Surgery may be performed either on an inpatient or on an outpatient basis. In the inpatient setting, patients are admitted to the hospital prior to surgery and assigned a hospital bed. After the scheduled procedure is completed, they return to their room for recovery. Outpatients, on the other hand, arrive at the hospital on the day of surgery. After surgery, they are held until post-operative recovery is complete and then discharged. In many cases home care visits and follow-up appointments are scheduled for additional post-operative care.

As in primary and specialty care environments, there are different degrees of urgency associated with surgical cases. Elective or deferrable surgeries may be scheduled well in advance because there is no need for immediate intervention. Urgent or emergency cases, on the other hand, arise on short notice and the speed of intervention is critical to the patients' potential for recovery. Such cases are not scheduled, but they must be accommodated along with the cases that are scheduled on any given day. In some hospitals one or more Operating Rooms (ORs) are reserved for such cases, whereas in others, slack time is spread across multiple ORs to accommodate unplanned procedures.

It is common for some ORs to be specialized for certain types of surgery. In such cases, certain specialized equipment are dedicated to the OR. In contrast, when ORs are not specialized, equipment may be moved from one OR to another. Some examples of specialized equipment include a pressurized environment for hyperbaric surgery, diagnostic imaging equipment for gastrointestinal endoscopy and cardiopulmonary bypass equipment for coronary interventions. A parallel also exists with respect to nursing staff, who often have highly specialized training. The level of cross-training of staff has a significant effect on the OR managers' ability to generate feasible OR schedules.

Elective surgery scheduling systems come in two varieties: block-scheduling and open (or nonblock) scheduling. Under a block-scheduling system individual surgeons or surgical groups are assigned blocks of time of a particular OR in a periodic schedule (weekly or monthly). The surgeons may book cases into their assigned blocks subject to the condition that the cases "fit" within the block time. Mean surgery durations (obtained from historical records) are typically used to determine whether or not the cases fit. For cases that do not fit, surgeons must request an allowance to overbook. In an open scheduling system, surgeons submit requests for OR time, and an OR schedule is created by the OR manager prior to the day of surgery. The schedule specifies which surgeries are assigned to which ORs and their start times. Hybrid systems are also prevalent. In such cases, either some ORs are block booked while others are left open, or unused block time is made available to other surgeons after a certain deadline, which is set a particular number of days prior to the planned session (Dexter *et al*., 1999).

Access rules for elective surgery are concerned with allocating surgeries to multiple ORs in open (non-block) environments and to allocating OR time among surgical groups in block-booking environments. These rules may depend on many factors including operational costs, the demand for certain surgery types, the degree of urgency and the revenue associated with each surgery type. Sometimes block environments have a two-step process for setting elective surgery appointments. In the first step, a patient selects a window of time (say a particular week) during which the procedure is likely to take place. This decision is largely determined by patient preference, medical urgency and applicable access rules. This exercise is referred to as advance scheduling in the literature (see, for example, Magerlein and Martin (1978) and Blake and Carter (1997)). In the second step, the patient is informed of a particular day and a particular start time for his/her procedure. This is usually specified

by the physician's office several days to weeks in advance of the procedure. Setting start times is sometimes referred to as allocation scheduling. Start times are designated times when the patient and the care giver team (surgeon, anesthesiologist and nurses) are scheduled to be available for the start of the procedure. We give an example of a model that can be used to determine surgery start times in Section 3.4. Finally, it is necessary to ensure that both pre- and post-operative (external) resources are available before and after the scheduled date and time of the surgery. This exercise is sometimes called external resource scheduling (Blake and Carter, 1997).

Daily surgery scheduling decisions are made by an OR suite manager, who is often the Chief Anesthesiologist or the Head Nurse. As actual surgery durations and the number and mix of *add-on* cases are observed, the OR manager must decide how to accommodate the new cases. (Add-on cases are those that are accommodated after the initial OR schedule is generated. They may be urgent cases, but that is not always the case.) Even without add-on cases, daily scheduling is complicated by the fact that delays in staff and equipment availability can affect multiple such procedures simultaneously. Responses to schedule variation may involve moving scheduled surgeries from one OR to another, and delaying or rescheduling previously scheduled surgeries. Ultimately, the OR manager must ensure patient safety by accommodating urgent cases, and strive to close ORs on schedule to minimize overtime costs.

3. Complicating factors

Health service providers struggle to balance supply and demand. Achieving this balance is often difficult on account of the uncertainty in the patient arrival and service times, patient and provider preferences, punctuality, cancelations and no-shows. It is further complicated by the fact that patients' needs for health services have varying degrees of urgency, and the decision-making process is often dynamic, i.e., some decisions about non-urgent patients must be made in advance of having complete information about urgent and emergency demand. Furthermore, the allocation of rewards and costs in health systems are such that the patient's, the physician's and the health system's incentives may not be aligned. This leads to the additional complexity of deciding whose perspective is appropriate when designing appointment systems.

In what follows, we have devoted a separate subsection to four key factors, which we believe to be the key variables that influence the performance of appointment systems. We also identify which types of variability predominate in each of the three delivery environments discussed earlier. We have deliberately left out factors such as cancelations and no-shows. Late cancelations and no-shows can lead to poor resource utilization, lower revenues and longer patient waiting times. Providers often use the openings in their schedule created by cancelations and no-shows to accommodate walk-ins and urgent requests. However, these actions are typically insufficient to replace revenue and utilization shortfall (Moore *et al*., 2001).

Late cancelations and no-shows are important in environments where capacity is tight or where no-shows and cancelations constitute a significant proportion of all appointments. The proportion of appointments affected by late cancelations and no-shows is low for clinics that largely serve patients with private insurance or Medicare patients. In contrast, clinics that serve under/uninsured populations, Medicaid recipients, or patients with mental health issues experience significant no-shows. No-shows are also positively correlated with the amount of time patients have to wait to get an appointment (Dove and Schneider, 1981). We discuss opportunities for future work on appointment scheduling with no-shows and cancelations in Section 5.2.

3.1. *The mapped arrival process*

In each of the three environments described in Section 2, the appointment system is designed for a particular mapping of the actual patient appointment requests to a mapped arrival process. For example, appointment systems for primary care clinics assume that each appointment decision is made when the patient calls with an appointment request. Thus, in this instance the mapped arrival process is the same as the actual arrival process. In contrast, specialty clinic appointments may be booked by the referring physician's clinical assistant at the end of each session. That is, the actual requests during a session are accumulated and for the purpose of designing the appointment system, the mapped process has batch arrivals that occur at regular intervals.

As described in Section 2, surgical appointments may occur in two steps. In the first step, the patient and the provider agree upon a time window during which the procedure might occur. For making these decisions, the mapped arrival process is usually the same as the actual arrival process and the physician's office manager may use aggregate capacity control rules to determine whether or not to book a surgery in a particular week. In the second step, the exact date and time of the procedure is determined. Typically, this occurs after all demand for elective procedures for that time window has been observed. Thus, for the purpose of setting appointment start times, the mapped arrival process consists in a given number of procedures that need to be scheduled during a session (block). In addition to the advantage of having more demand information before choosing the day of surgery and surgery start times, this two-step process provides some insulation against cancelations and delays in scheduling pre-surgery exams and obtaining medical records from other providers.

The mapped arrival process can be classified by the inter-arrival times, the number of arrivals at each arrival epoch and the number of arrival epochs during the booking horizon. The booking horizon refers to the length of time

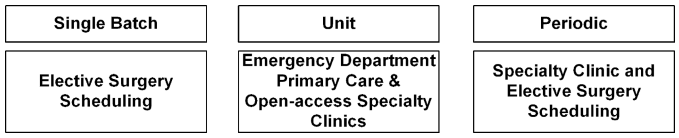

Fig. 1. Arrival process types and delivery environments.

between the opening of bookings for a particular session (block of time) of the provider to the start of that session. Rather than describe all possible variations, we focus in the following on the most common types of mapped arrival processes. The delivery environments in which each type of mapped process is most common are summarized in Fig. 1.

- **The Single Batch Process:** In this situation, appointment scheduling decisions are not made until after observing all demand for a session, with the result that inter-arrival times are irrelevant. The number of cases of each patient-priority class that should be scheduled during a session is determined by access rules. This number is assumed to be known at the time of determining encounter start times, albeit may vary in different instances of the problem. Note that the mix of patients who need to be scheduled during a session is not homogenous—they may have different medical urgencies, require different types of equipment and staff resources and vary considerably in the required service time. The single batch process is commonly assumed when determining elective surgery start times.
- **The Unit Process:** In this case, booking requests are assumed to occur one at a time and at random time epochs. This corresponds to the situation where the mapped process is identical to the actual appointment-request arrival process. Booking requests can be for different types of services and of different urgency levels. The unit arrival process is commonly assumed for designing appointment systems for primary and specialty care environments.
- **The Periodic Process:** The periodic mapped process arises when appointment requests are accumulated over discrete time periods and appointment times are firmed up for all requests over an interval at the end of each period. The inter-arrival times are constant. However, the number and type of arrivals during an inter-arrival period may be random. The periodic mapped process is assumed when scheduling appointments for specialty care clinics and for elective surgeries. Note that the single batch process is a special case of the periodic process. The former arises when the discrete interval covers the entire booking horizon. Still, we treat them as separate categories because the appropriate appointment scheduling models for these two situations are quite distinct.

Illustrative example 1: Next, we present summary statistics from a particular clinic to provide an example of typical arrival process variability in a primary care setting. Requests for appointments arrive throughout the day. However, since only booked calls are tracked, we can obtain statistics only on those calls that resulted in an appointment. We found that the call volume is the highest at the start of the day, and that there is significant day-of-week seasonality; many more calls are received on Mondays. Specifically, the number of daily calls per 1000 patient panel which resulted in a booked visit, after excluding certain types of appointments such as physical exams and follow ups, ranged from 6.68 to 12.17 in a clinic served by ten physicians. Furthermore, the coefficient of variation of daily calls (which captures the day-of-week variability) ranged from 0.34 to 0.49.

We show the arrival process variability for a particular physician from this clinic in Fig. 2. This plot shows the number of appointment requests on Mondays and Fridays by the number of the week; week 1 is the first week of the

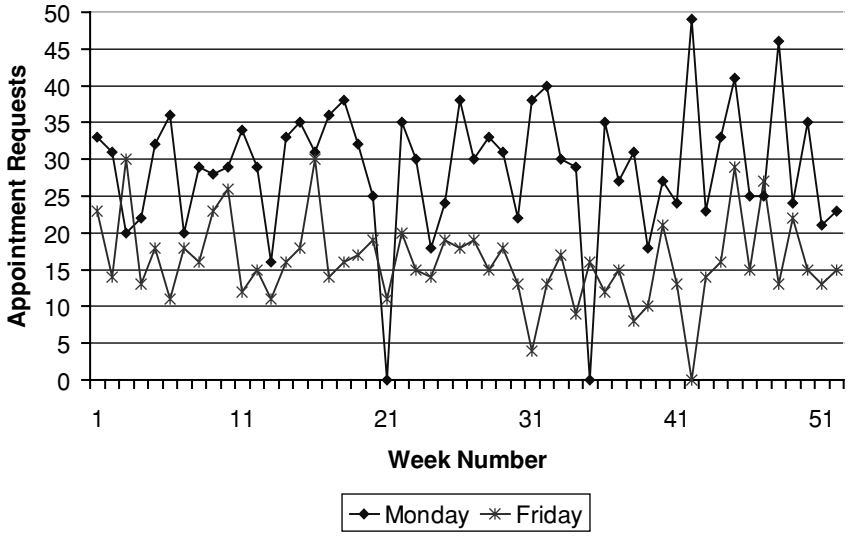

Fig. 2. Day-of-week pattern of appointment requests.

year, and week 52 the last. Note, zero requests occur when a particular Monday or Friday is a holiday. The mean, median and standard deviation of the number of requests are 29.88, 30 and 7.09 on Mondays and similar statistics for Fridays are 16.45, 16.2 and 5.43. In addition to day of week seasonality, call volumes are also affected by annual patterns, e.g., greater demand occurs during the flu/allergy seasons.

A highly variable arrival pattern makes it more difficult to provide timely access, particularly after the effect of patient choices is factored into the problem. Without careful attention to patient-choice patterns, clinics can end up with crossed appointments (two service providers serving patients in each other's panel at a particular time of day), loss of revenue and patient and provider dissatisfaction.

3.2. *The service process*

Service time requirements can be assumed to be either known (deterministic) or random. In some cases, such as for routine follow-up appointments at primary care clinics it may be reasonable to assume that service times are approximately deterministic. This is in part due the fact that PCPs can more easily influence the time they spend with a patient to fit within a fixed length of time. On the other hand, for some types of surgical procedures, service times can vary significantly from one patient to another. This gives rise to three types of modeling scenarios: constant service times, diagnosis dependent service times, and random service times. Random service times can be either identically distributed or diagnosis dependent. The delivery environments in which each type of service process is most commonly found are summarized in Fig. 3.

There are many factors that affect service durations. For instance, the presence of student doctors in a surgery room can increase all service durations. Such factors also make sequential surgery durations correlated. Case loads also affect service duration. Clinicians work faster on days when their calendar is heavily booked as compared to lightly booked days. Similarly, patient attributes such as age, degree of disease progression, cultural background and language fluency (need for an interpreter) can affect service durations.

Illustrative example 2: A key difficulty in setting surgery start times comes from the uncertainty in procedure times. First, the quality of solution is much worse when procedure times are highly random. This means that both patient/provider waiting and use of overtime may be unavoidable. Second, solutions that optimize an expected performance measure are often not implementable. For example, key support staff may not volunteer to work the extra hours needed to complete a day's scheduled procedures. In that case, it is necessary to cancel procedures and reschedule them at a later date on a priority basis, which makes daily scheduling much more challenging.

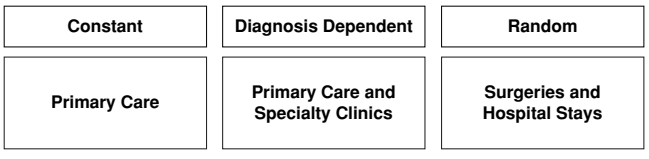

Fig. 3. Service process types and delivery environments.

Procedure times can vary substantially from one patient to another, and from one surgeon to another. Figure 4 shows a histogram of procedure times (not counting changeovers) of two surgeons, labeled A and B, for a common procedure called *left-heart catheterization* at a particular hospital. This is a diagnostic procedure that is used to determine the patient's coronary health and the need for further intervention. The procedure times are in minutes.

The mean and median procedure times are 52 and 47 minutes for Doctor A, and 58 and 50 minutes for Doctor B. The mean procedure times are statistically different (p-value $= 0.007$ for a t-test of equality of means with unequal variances). The 95th percentiles of the procedure time distributions for Doctors A and B are 90 minutes and 102 minutes, respectively. Since in a large number of hospitals, the planned surgery times equal the mean of a few most recent surgeries (by type and provider), the presence of unpredictable cases simultaneously leads to longer direct waits and poor utilization of OR time. In fact, the optimal allowances for surgeries are not uniform even when surgery times are sampled from the same distribution. The optimal allowance depends on the position of a particular procedure in the sequence of procedures performed during a particular OR session (see Denton and Gupta (2003) for details).

3.3. *Patient and provider preferences*

Common examples of patient preferences are as follows. Some patients prefer an appointment on the day they call, or soon thereafter, and the day of the week or the time of the appointment is not particularly important to them. Others prefer a particular day of week and a convenient time. They do not mind waiting for convenience. Patients have different degrees of loyalty toward their designated PCP, or a particular specialist/surgeon. Some book appointments only with a particular provider, even when this leads to an inconvenient appointment time or extra waiting, whereas others switch easily to alternate providers.

Providers also vary greatly in their practice styles. Some open up more capacity by double booking, working through lunch and working after hours to take care of urgent demand. Others adhere strictly to their daily schedules. Some place few, if any, restrictions on how their available time is used for appointments. Others have strict guidelines for the use of their time. For instance, many physicians restrict the number and timing of physical exams each day. Some surgeons require office visits and pre-operative evaluation for all referrals. Others may do so only for certain

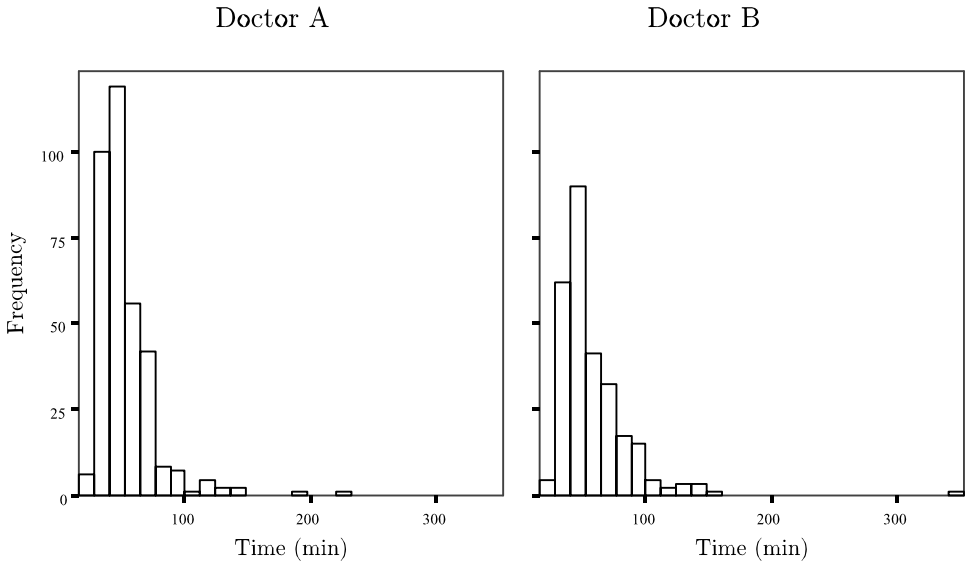

Fig. 4. Procedure times for left-heart catheterization.

cases, based on medical histories. Hospital data show that surgeons prefer to perform surgeries on certain days of the week. For example, some prefer the first half of the week (Monday to Wednesday) and demand for OR time is typically greater earlier in the week.

The presence of patient preferences often implies that the optimal scheduling policies are neither simple, nor easy to implement (see examples in Gupta and Wang (2008)). Furthermore, accommodating preferences can easily make mathematical models of the appointment booking process intractable, which is perhaps one reason why the majority of mathematical models do not include preferences. It is also the reason why the vast majority of appointment booking systems are not automated. Instead, they rely on a human scheduler to work with patients to determine a convenient date and time. However, the criticality of modeling patient and provider preferences varies by the delivery environment, as summarized in Fig. 5.

The ability to model patient preferences is very important in the primary care setting. It is also important to match patients with their PCPs for continuity of care and revenue/cost considerations. In contrast, in some specialty clinics and surgery scheduling environments, provider preferences take on a greater significance, and patient expectations for scheduling flexibility are lower. Patients needs are often associated with a specific episode of care. Whether or not a patient seeks to be paired with a particular provider can vary significantly depending on the nature of the health service. As a result, in some cases, there is greater opportunity in the specialty care and surgery environments to control the match between supply and demand of appointments by pooling supply. With respect to the urgent cases in these environments, such patients are accommodated immediately without regard to the time of the appointment or the on-call physician.

3.4. *Incentives and performance measures*

Design of appointment systems needs to consider the costs and benefits of the various options to the health service network, the physicians and the patients. Unfortunately, the incentives of these groups are not always aligned, which makes it difficult to get buy-in from the different stakeholders on the choice of acceptable objective functions. By and large, the appointment systems in use today benefit the service network and the physician more than the patient. Developing patient-oriented solutions offers exciting new opportunities for research on appointment systems design.

Appointment scheduling problems can be formulated either as cost minimization problems or as revenue (profit) maximization problems. Focusing on the first approach, we illustrate the difficulty associated with choosing an objective function for the problem of determining the start times

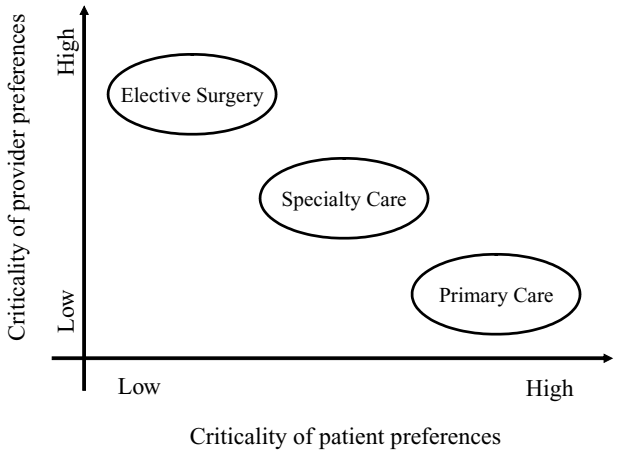

Fig. 5. Criticality of patient and provider preferences and delivery environments.

of n surgeries in a non-block environment. In this example scenario, which is fashioned after the model presented in Denton and Gupta (2003), appointment requests are mapped onto a single batch arrival process, surgery times are random, provider preferences determine the surgery sequence and patient preferences are not modeled. All n procedures must be scheduled and only decisions variables are the surgery start times. These must be chosen to minimize overall cost that has three components: patient direct waiting time, OR idle time, and tardiness with respect to the allotted time for the session. We assume that the patients and the providers are punctual.

Let $\mathbf{Z}$ denote the vector of random surgery durations, $\mathbf{a}$ the vector of scheduled start times, $\mathbf{W}$ and $\mathbf{S}$ the vectors of waiting and OR idle times for a given $\mathbf{a}$ and $\mathbf{Z}$, d the length of the day, and L the tardiness for a given $\mathbf{a}$. For ease of exposition, we also define scheduled surgery times $\mathbf{x}$, which are equivalent to $\mathbf{a}$. In particular, only $n-1$ parameters need to be determined, the ith surgery's scheduled duration $x_i = a_{i+1} - a_i$ for $i = 1, \cdots, n-1$, and $a_1 = 0$. The waiting, idleness and tardiness metrics can be determined according to the following recursive relationships:

$$W_i = (W_{i-1} + Z_{i-1} - x_{i-1})^+, \qquad i = 2, \ldots, n, \tag{1}$$

$$S_i = (-W_{i-1} - Z_{i-1} + x_{i-1})^+, \qquad i = 2, \ldots, n, \tag{2}$$

$$L = \left(W_n + Z_n + \sum_{i=1}^{n-1} x_i - d \right)^+. \tag{3}$$

The burden of costs associated with waiting, idling and tardiness with respect to the session length does not fall equally on the different stakeholders in this problem. OR idling and overtime with respect to a fixed session length matters a great deal to hospital administrators (since hospitals bear the extra cost of staffing ORs), but it may not affect physicians' rewards. Similarly, if the same surgeon performs all procedures in an OR on a given day, patients are typically asked to arrive early and the waiting costs are largely borne by patients. Anesthesiologists care about discrepancy between planned and actual start times, which can affect their remuneration, and introduce variability in the length of their workday.

Assuming linear costs, with $\mathbf{c}^{\mathrm{w}}$, $\mathbf{c}^{\mathrm{s}}$ and c_ℓ denoting per unit costs of waiting, idling and tardiness, the OR manager's problem can be formulated as follows:

$$\min_{\mathbf{x}} \left\{ \sum_{i=1}^{n} c_i^{\mathrm{w}} E[W_i] + \sum_{i=1}^{n} c_i^{\mathrm{s}} E[S_i] + c_\ell E[L] \right\}, \tag{4}$$

where the expectations are over $\mathbf{Z}$. Even though a model-based solution can substantially improve the procedure for setting start times (see Denton and Gupta (2003) for details), it is often difficult for OR managers to get buy-in from surgeons. In effect, this is a problem of arriving at an agreement over what should be the relative magnitude of waiting, idling and tardiness costs. It is not uncommon for accomplished surgeons to command their own OR with on-demand access. This is an expensive option for the health care system, but allows surgeons to optimize the use of their time. Similar problems arise when addressing related questions such as what should be the session length and how many surgeries should be scheduled for each session length. These choices also affect other decisions such as staffing of patient-intake and recovery areas and hospital wards.

Incentive misalignment induced problems can also arise in revenue-based models. Physicians are paid according to a variety of different formulae. Some are on salary, some are paid on a fee-for-service basis with negotiated fees for each service, and some others are compensated on the basis of the value of work performed (called the Relative Value Unit (RVU)). Health service networks also use a combination of guaranteed and production-level (or RVU) based compensation. These compensation schemes affect physician behavior and it is important to accurately model such behavior when designing appointment systems.

4. The state of the art

In this section, we examine the Health Services Research and IE/OR literature on appointment systems. Previous surveys of IE/OR studies on health care applications can be found in Pierskalla and Brailer (1994) and Lagergren (1998). A review of the literature on outpatient appointment systems can be found in Cayirli and Veral (2003), whereas Magerlein and Martin (1978) and Blake and Carter (1997) summarize articles on surgery scheduling.

Studies on the appointment scheduling problem can be categorized based on several criteria. The most straightforward classification is by the type of problem considered, i.e., outpatient or surgery scheduling. Since these problem types generally have different features, many previous review articles have used this approach. Alternatively, we can use the mapping of the actual arrival process to the arrival process used in appointment system design to classify appointment scheduling problems. When the mapped process is a single batch arrival process, such problems are referred to as *static* scheduling problems. This contrasts with *dynamic* problems that typically assume a unit arrival process. Similarly, we can classify studies based on what type of waiting cost is included in the objective function. Most studies consider either the direct waiting cost, or the indirect waiting cost, but not both. Finally, articles on appointment scheduling can be classified on the basis of the solution approach used. Commonly used approaches fall into four categories: heuristics, simulation, queueing theory, and optimization (deterministic and stochastic). Heuristics are compared either in empirical studies or via computer simulation. The empirical approach is more common in the health services research literature than in the IE/OR literature.

Our goal is not to provide a comprehensive review of the literature, since that has been accomplished elsewhere. Rather, we focus on discussing major themes in the

literature based on the complicating factors discussed in Section 3. It turns out that depending on the aspect of the problem being modeled, and assumptions concerning input parameters, the mathematical formulation of problems arising in outpatient scheduling and in surgery scheduling can be quite similar. Another advantage of this approach is that it uncovers the combinations of factors that have not been addressed in the literature. We have left out certain types of problems from this review, e.g., advance and external resource scheduling problems that arise in surgery scheduling environments. Such problems are related to but are not central to the theme of appointment scheduling. Moreover, the ensuing literature review does not deal with cancelations and no-shows. We discuss those in Section 5.2.

We categorize the literature into the following three themes.

Theme A: Single batch arrival process, random service times, no patient or provider preferences, and a cost-based formulation that includes only the direct waiting cost.

Theme B: Unit or periodic arrival process, random service times, no patient or provider preferences, and a cost-based formulation that includes only direct waiting cost.

Theme C: Unit arrival process, deterministic service times, patient and provider preferences, indirect waiting times, and a revenue-based (or cost-based) formulation.

We will next discuss major contributions within each theme one by one. In addition, we will describe relevant work in other application areas that has a bearing on appointment scheduling. Opportunities for future research are described in Section 5.

4.1. *Theme A*

A large number of articles in Theme A deal with a block schedule. In this case, the scheduler is given a session of length t, which is divided into k blocks of usually equal length (each equal to t/k). There are n patients who need to be scheduled during the session. The decision variables are the n_js, the number of patients who should be asked to arrive at the start of each block. Note that $\sum_{j=1}^{k} n_j = n$. There could be a fixed number of service providers, or the number of service providers could change over time. Patients are punctual, do not balk, and receive service in a First-In First-Out (FIFO) manner. The number of patients waiting at the start of a block equals the sum of backlogged patients from previous blocks and new arrivals scheduled for the current block. Service providers idle if all patients assigned to a block are served before the block ends.

A variety of block scheduling regimes have been proposed and studied. On one end of the spectrum, we have a single-block schedule. In this case, $k = 1$ and all n patients are asked to arrive at the start of the session. Clearly, this minimizes service providers' idleness at the expense of patient waiting. At the other end of this spectrum, we have individual appointments (also called sequential block), where $k = n$. Between these two extremes exist a number of possibilities. In a multiple-block schedule, $k < n$ and $n_j = n/k$ for all j, whereas in the modified-block schedule, $n_1 > 1$ patients arrive at the start of the session followed by multiple or individual arrivals at regular intervals, i.e., n_j is either $(n - n_1)/(k - 1)$ or 1 for all $j \geq 2$. Finally, in a variable-sized multiple-block schedule, each n_j may be different.

Several papers in Theme A compare different block-scheduling policies. We begin with a review of empirical studies. Such studies appear in health services research journals. In some protocols a "catch-up" time is allowed for the physician at the end of each block (Heaney *et al*., 1991; Penneys, 2000; Chung 2002). For each choice of the number of blocks and the target number of patients to be scheduled in each block, core metrics are estimated from clinic data and compared. These metrics are patients' waiting times physicians' idle times, and time to clinic completion (also called makespan). Recall that these models do not consider patient choice, different urgency levels and the amount of patient indirect wait. Also, the actual number of arrivals in a block is random and may be less than the target on certain days of clinic operations.

Penneys (2000) compares the hourly block scheduling to the sequential scheduling rule in two clinics, each operating under a different rule, using observed data. He reports that in comparison with sequential scheduling, block scheduling results in the physician being significantly more likely to enter the exam room earlier, increased patient-free time during the day, and the clinic finishing on average 35 minutes earlier, whereas the mean patient wait times remain comparable between the two scheduling rules. Chung (2002) claims that the modified block scheduling approach has improved the bottom line of his practice by 15%. The success of this approach requires leaving physician catch-up time at the end of each hour in order to keep average patient waiting times low (see also Heaney *et al*. (1991) and Dexter (1999)).

In the IE/OR literature computer simulation is often used to study the effectiveness of different heuristics. For example, Bailey (1952, 1954), Welch and Bailey (1952) and Welch (1964) use simulation to study a modified block protocol in which $n_1 = m$ and $n_j = 1$, $\forall\, j \geq 2$. Moreover, the last $n - m$ appointment times are spaced by the mean service duration. Heuristics for assigning individual appointment times to patients have also been explored. For example, Charnetski (1984) considered a heuristic that assigns a service time allowance of $\mu_i + h\sigma_i$ to patient i, where μ_i and σ_i denote the mean and standard deviation of the time needed to serve the ith patient. He experimented with different values of h using a simulation model while assuming that service durations are Normally distributed. Similarly, Ho and Lau (1992) and Robinson and Chen (2003) have used simulation-based techniques to compare the performance

of a variety of heuristic appointment rules. Robinson and Chen (2003) provide an excellent review of heuristics that have been recommended for the class of problems included in Theme A.

White and Pike (1964) use the usual block schedule (k blocks with $n_j = n/k$) and Soriano (1966) compares individual and block scheduling rules with the practice of scheduling two patients at a time, when service durations are identical and gamma distributed. These studies use analytical methods, rather than simulation, to estimate performance metrics. In the two-at-a-time system, the scheduled time between appointments is twice the mean length of a single appointment. Soriano argues that the two-at-a-time system outperforms other block-booking approaches. Other similar studies include Villegas (1967) and Rising *et al*. (1973). Fries and Marathe (1981) use a dynamic programming approach to determine the optimal variable-sized multiple-block schedule. (The bulk of Fries and Marathe (1981) concerns Theme A, but the authors also discuss how their approach can be used to obtain numerical solutions when the number of patients who need to be scheduled is unknown.) Liao *et al*. (1993) allow both static and dynamic choice of n_j over the course of the day as service times are revealed for each block. Liu and Liu (1998) further generalize the approach by allowing doctors to be tardy. This results in a random (but non-decreasing) number of servers being available at the start of each block. Vanden Bosch and Dietz (2000) model patient-class-dependent service times. By and large, the generalizations mentioned above lead to models for which complete analytical solutions are not possible. Therefore, authors propose combinations of simulation, heuristics and approximate solutions.

Another line of research uses optimization models for determining appointment start times for individual appointments, such as those encountered in elective surgery scheduling. This is in contrast to the multiple-block scheduling described above in which block lengths were predetermined and the scheduler chose the number of patients who should arrive at the start of a block. Here, the number of patients in each block is fixed (equal to one), but the length of each appointment interval is chosen optimally. These formulations use an objective function similar to Equation (4) of Section 3.4. In this genre of work, Weiss (1990) and Robinson *et al*. (1996) deal with two and three-patient problems, respectively, which can be solved relatively easily owing to the low dimensionality of the problem. The problem of choosing planned surgery durations is mathematically similar to the problem of setting planned lead times in production systems when production times are random (see Yano (1987) for an instance of this problem and relevant literature).

Wang (1993) considers the case in which patient service durations are exponentially distributed and shows that for this special case the probability density function for patient waiting times is *phase-type*. He then exploits the computational advantages associated with phase-type distributions to find the optimal appointment times. Through numerical examples, he shows that optimal patient allowances have a dome shape, i.e., the optimal appointment lengths are smaller at the start and end of day and longer in the middle of the day. Denton and Gupta (2003) find that this dome-shaped result holds for arbitrarily distributed patient service times. However, the actual shape is strongly affected by the relative magnitudes of per-unit costs of waiting, idling and tardiness. Following these structural results, Robinson and Chen (2003) have suggested heuristics in which the time allotted to a service depends on, among other factors, the relative position of the service in the schedule.

4.2. *Theme B*

Unit or periodic arrival processes are typically considered within the framework of queueing systems. In a typical setting, patients are scheduled to arrive at the clinic at equal (fixed) intervals and the length of this interval is the decision variable. A patient's physical arrival and request for appointment are not differentiated, with the result that all waiting is direct waiting. That is, virtual or indirect wait is not modeled in these studies. Service times are random and come from a common probability distribution. Unless the service provider(s) is (are) idle at the time of patient arrival, patients wait for service in a common queue and the service protocol is FIFO. A cost-based optimization problem is formulated with the goal of minimizing the sum of expected patient waiting cost and the expected service-provider idling costs.

Unit arrivals with m servers give rise to queues of the type $D/G/m$. When a batch of arrivals occur at an arrival epoch, we obtain $D^B/G/m$ queues with B being the batch size. It is clear that these models are quite similar to the individual and multiple-block scheduling protocols discussed in the previous section. In fact, when arrivals are scheduled, and a transient analysis is carried out (this implies a finite number of arrivals, n), the arrival processes in these models are identical to those described in Theme A. That is, they reduce to a single batch arrival process since the total number of patients to be scheduled is known. However, meaningful analysis invariably requires steady-state assumptions. In that case, n is infinite and the two types of models diverge. The models also diverge when arrivals are not scheduled and occur at random time epochs. Examples of such formulations can be found in Mercer (1960, 1973), Jansson (1966), Sabria and Daganzo (1989) and Brahimi and Worthington (1991). We briefly discuss each of these studies next.

Mercer (1960, 1973) does not optimize the appointment system, but presents a queueing-based performance analysis when patients may arrive late, or may not arrive at all. Jansson (1966) considers the problem of choosing the optimal patient arrival interval and the initial number in the queue at the start of operations to minimize total cost. The article provides analysis of costs incurred in both the

transient and the steady state when service times are exponentially distributed. Sabria and Daganzo (1989) relax a key assumption in Mercer's work. Mercer assumes that arrivals always occur in the prescribed order. A customer who fails to arrive by the end of his/her arrival interval balks. Sabria and Daganzo (1989) allow out-of-sequence arrivals. However, customers are served in the planned sequence, which can cause some customers to wait while the server idles. Brahimi and Worthington (1991) report the application of transient analysis of queueing systems to actual outpatient appointment scheduling problems. The arrivals can be scheduled or random and service times are approximated by discrete distributions. They argue that the use of analytical techniques can substantially reduce waiting time without increasing server idleness when compared with the current state of the practice.

As mentioned before, queueing models with random arrivals also diverge from Theme A. A common problem addressed by such models is the problem of allocating medical service capacity among distinct demand themes. Klassen and Rohleder (1996) study the impact of different scheduling rules in simulation experiments when the scheduling environment differs along two dimensions: the mean service time and the potential number of urgent calls. They identify scenarios under which certain rules perform better than others. Gerchak *et al*. (1996) consider the problem of reserving surgical capacity for emergency cases on a daily basis when the same operating rooms are also used for elective surgeries and surgery durations are random. They assume surgery durations to be independent and identically distributed and formulate the problem as a stochastic dynamic program. They show that the optimal amount of capacity to reserve for emergencies is a function of the amount of backlog (queue) for deferrable surgeries. Specifically, less capacity is reserved for emergency arrivals when the queue of patients waiting for deferrable surgeries is longer. They also develop a fast algorithm for finding the optimal number of deferrable surgeries to schedule on any given day.

Models with unit arrivals are used to carry out policy parameter optimization assuming a threshold policy for capacity reservation. Notwithstanding the fact that a threshold policy is not optimal (see Gerchak *et al*. (1996) for details), these articles assume a threshold policy and propose models for performance evaluation. Specifically, the goal is to either study the impact of a given N_i or to compute optimal values of N_is, where N_i is the threshold level for class-i customers. A class-i customer is served if and only if fewer than N_i servers are occupied. Taylor and Templeton (1980) and Schaack and Larson (1986) use queueing models to obtain performance measures such as average utilization and overflow rates for a given set of threshold levels. Kolesar (1970) and Esogbue and Singh (1976) focus on the problem of finding the optimal threshold levels under a linear cost structure. Patient and provider preferences are not modeled in these studies.

4.3. *Theme C*

Gupta and Wang (2008) model a primary care clinic's problem of choosing which appointment requests to accept to maximize its revenue as a Markov Decision Process (MDP). (An equivalent formulation can be obtained in a cost-based setting as well.) They also model patient choices and show that when the clinic is served by a single physician, the optimal booking policy is a threshold policy under a *normal-form* patient-choice model. The class of normal-form choice models is large, intuitively appealing, and includes all of the commonly used models of discrete choice found in the IE/OR and Economics literatures. Membership in the normal-form class requires that the probability that a patient will select a particular available slot does not decrease when fewer other slots are available.

When the clinic has multiple doctors, patients may choose a more convenient time with a doctor other than their PCP. This makes the optimal policy more complicated because patient–PCP mismatch lowers physician and clinic revenues (O'Hare and Corlett, 2004). However, Gupta and Wang show that for each physician, there exist computable upper limits on the number of appointments that a clinic should book. These limits depend on: (i) the total number of booked slots in the clinic at the time of call; (ii) the number of booked slots of the physician with whom an appointment is requested; and (iii) on whether or not the requested appointment results in a patient–PCP mismatch. They use the bounds to develop heuristics for appointment booking control, which are tested and found to perform very well in simulation experiments based on real clinic data.

4.4. *Related work*

The problem of allocating service capacity among several competing customer classes, who arrive randomly over a period of time, has been studied in diverse applications including airlines, hotels and car rentals. In particular, airline Revenue Management (RM) has been studied particularly well; see McGill and Van Ryzin (1999) and Talluri and Van Ryzin (2004b) for detailed reviews.

Whereas capacity reservation is also an important aspect of health care access management, there are important differences that make it difficult to simply "tweak" existing models to fit the needs of the health care industry. For example, of the various models suggested for airline RM, comparisons with the Expected Marginal Seat Revenue (EMSR) model (see, Belobaba (1989)) help to highlight the complexity of the health care scenario. In the two fare-class EMSR model, the optimal policy reserves a certain number of seats exclusively for higher fare-class customers. The lower fare-class customers are not allowed to book once a certain number of seats (equal to the booking limit) have been sold. The consumers' trade-off is between buying a cheaper and inflexible ticket versus a more expensive ticket with greater flexibility. In the health

care applications, the patients' choice function has more elements. A patient chooses a particular service provider (which determines service quality and clinic revenue), a particular day of week (service delay) and a particular appointment time (convenience). Prices are not used to control access to clinic capacity. Instead, clinics and physicians place restrictions on the availability of different slots to different types of patients. These features and the need to accommodate urgent demand make access control more difficult in health care applications; see Talluri and Van Ryzin (2004a) for recent efforts to model consumer choice in airline RM models.

Another important aspect of health services is the necessity to take care of urgent demand. In a typical service industry setting, if a particular demand cannot be satisfied, there is an economic penalty to the service provider and it is reasonable to assume that excess demand is lost, possibly served by other service providers. In the health care setting, more urgent demand must be satisfied. As mentioned earlier, capacity is often soft and service providers can vary available capacity to a certain degree by working faster, double booking and working extra hours. However, there are limits to how much extra capacity can be made available by using such approaches. At the same time, the need to respect patient choices puts limits on the extent to which capacity can be pooled to take care of peak demand. Many service providers belong to large health care networks with urgent and emergency care facilities. These arrangements help to take care of unplanned and high-priority demand, but the costs to the health network are higher. All these features make appointment scheduling more challenging in the health care setting.

Many decades of research has been devoted to the problem of scheduling a finite number of jobs on one or more parallel or sequential machines (see Leung (2004) for a collection of recent articles). Variations include sequence-dependent changeover times, resource-use constraints, and both earliness and tardiness penalties. The vast majority of this literature considers deterministic scheduling problems, although in recent years, stochastic versions of these problems have also attracted interest. While this literature has some elements in common with the problem of appointments scheduling, there are key differences in the characteristics of the two problem scenarios. For example, in job scheduling problems, the jobs (raw material kits) are assumed to be available at the time of choosing a processing schedule, or in some cases at a known release time. The release (equivalently appointment) time itself is normally not considered a decision variable in such problems. Furthermore, the nature of performance measures is typically quite different between appointment scheduling and machine scheduling problems. For instance, jobs do not accrue a waiting cost while they wait to be processed unless they are tardy with respect to a due date. Also, machine capacity is typically assumed fixed, and it can neither speed up, nor work on two jobs in parallel when demand is high.

Dynamic demand arrivals are typically not modeled in the machine scheduling context. Instead, the vast majority of work on machine scheduling assumes that the number of jobs to be scheduled is known. This amounts to the assumption of a single batch arrival process in the terminology introduced earlier in this paper. Thus, at best the machine scheduling literature is relevant for a subset of problems encountered in the health care setting. Finally, even though a certain pairing of jobs and machines may be preferred on account of setup and processing time efficiencies, jobs do not exercise their choice in this matter. This minimizes the possibility of independent and competing incentives which patients and providers often have in the health care setting. On account of these differences, research on machine scheduling does not translate directly to the health care setting. In fact, it is the authors' hope that this paper will encourage new modeling effort to address the specific needs of the health care industry.

5. Open challenges

In this section we describe a series of open research problems that either relate directly to the choice of access rules, appointment start times and response to unplanned deviations, or indirectly affect the efficiency of appointment systems.

5.1. *Indirect patient waiting*

The vast majority of the literature we reviewed earlier in this article considers direct waiting and seeks to achieve a satisfactory balance between patients' direct waiting and providers' utilization during a service session. In reality, appointment systems affect both direct and indirect waiting times. However, modeling indirect waiting is challenging for a variety of reasons. First, unlike direct waiting for which the end of the session is a natural termination of the planning horizon, indirect waiting problems are more realistically modeled as infinite-horizon problems. Second, since patients attempt to find a suitable appointment time with one of several desired service providers over one of several future days that might be acceptable, the scheduling decisions made on a certain day for a particular physician are coupled with those of other days and other physicians. There is no obvious decomposition that can be applied to simplify the problem in a manner analogous to the airline seat booking control problems (Lautenbacher and Stidham, 1999). Furthermore, admission control models for queues (see, e.g., Stidham and Muckstadt (1985)) are also inappropriate because upon admission, customers are not necessarily served in the order of arrival at the earliest time that the server becomes available. Rather patients pick an appointment with a particular provider at a particular time in the future. Finally, an additional complicating factor is

the set of access constraints imposed by provider preferences for the sequence and mix of cases each day.

It would be valuable to develop access planning models that consider both aspects of patients' waiting in surgery and specialty care appointments scheduling. (Note that direct wait is not a serious problem for many primary care clinics because most services can be completed in a standard appointment length. Clinics that overbook to mitigate the impact of cancelations and no-shows are an exception to this rule. We discuss such issues in Section 5.2.) However, modeling both types of waiting times can be a difficult problem. For example, with a unit arrival process, the required model needs to consider two time scales: one for requests for appointments that arrive continuously over time, and another for available future sessions (daily or weekly) that are updated periodically. At the end of each planning period, the expired session falls off from one end of the appointment book (typically between 90 to 180 days) and a new session appears at the other end. The presence of different time scales makes such problems difficult to solve analytically.

One possible approach might be to develop two separate but linked models, not unlike the two-step sequential approach used by some hospitals to schedule surgeries. The first model would use a day (or week) as a unit of time and determine aggregate amounts of capacity to make available (at the subspecialty level) for each future period in the appointment book based on the current state of appointments and an estimate of direct waiting times generated by the realized mix of booked appointments. The second model would determine the optimal sequence and appointment start times after accounting for each new request that is accepted for each future service-provider/OR session. A new estimate of direct waiting times will be developed, which will allow the first model to be updated at regular intervals. Although the implied decomposition, and the sequential determination of start times, is not necessarily optimal, a version of this approach is already used by some service providers. Thus, in addition to being tractable, the results of this approach may be relatively easy to implement.

5.2. *Late cancelations and no-shows*

The medical and IE/OR literature on late cancelations and no-shows (hereafter referred to as no-shows) falls into the following four categories: (i) articles that focus on estimating no-show rates and identifying correlations between no-show rates and patient characteristics (e.g., Dove and Schneider (1981)); (ii) articles that use clinic data to document the time and money effects of no-shows in outpatient clinics (e.g., Moore *et al*. (2001)); (iii) articles that use models to estimate the effects of no-shows (e.g., Bailey (1952) and Ho and Lau (1992)); and (iv) articles that study the use of overbooking to mitigate the impact of no-shows (e.g., Kim and Giachetti (2006) and LaGanga and Lawrence (2007)). Health care management journals also contain a variety of suggestions for reducing no-shows, e.g., sending postcard reminders to patients. However, these approaches do not remove the fundamental obstacles faced by patients who miss appointments. Examples include lack of transportation, day care and inability to get time off from work without losing pay.

Evidence suggests that absentee patients tend to be younger, male, of lower socioeconomic status (with either state medical assistance or no insurance), divorced or widowed and have a record of missed appointments (Moore *et al*. (2001) and Lacy *et al*. (2004)). Whereas the vast majority of articles examine clinical data to study correlations between no-show rates and patient characteristics, Lacy *et al*. (2004) report a study in which they interview patients to identify reasons why patients miss appointments. They identify three major reasons for no-shows—discomfort experienced during the appointment, patient perception that the health care system disrespects their time and beliefs, and patient misconceptions about the consequence of missed appointments. Discomfort can be caused by diagnostic tests, blood draws, needles and the fear of bad news. Perceptions about disrespect are caused by long indirect and direct waiting times and the need to get one or more referrals before being able to access the right service provider such as a specialist. Finally, patients often underestimate the consequences of missing a medical appointment believing that the time would be used by the busy service provider to catch up during the day.

For the setting described in Theme A of Section 4, it is possible to include the effect of no-shows on system costs and optimal appointment lengths by modifying the procedure time distributions. Procedure times are either zero with the no-show probability, or a positive random variable otherwise. However, this approach is not suitable for Themes B and C because the number of patients who would need to be scheduled in a session length is unknown at the time of booking an appointment. Therefore, for such cases, especially in outpatient clinic settings with predetermined appointment lengths, IE/OR literature has focused on identifying the optimal overbooking rules.

Kim and Giachetti (2006) model a session of a clinic, and estimate no-show rates, unit costs of unfilled appointment slots, patient direct wait, and provider overtime, and the probability that the clinic would be able to fill slots in excess of its nominal capacity. They use these data to calculate the total cost for different values of overbooking levels and identify the optimal overbooking level. They also compare their model to a simpler one in which the number of overbooked appointments equals the expected number of no-shows. They report that their method results in significantly higher revenue. Kim and Giachetti (2006) do not model each appointment individually. In reality, patient waiting times and provider overtime are affected by the times when overbooked patients are scheduled to arrive and the appointment times of patients who do not show up. This feature of health care appointments is also what sets

the underlying overbooking problem apart from its analog in the transportation setting; see Zhao and Zheng (2001) for an example of the latter.

LaGanga and Lawrence (2007) carry out a computer simulation study to estimate the provider overtime and patient waiting time. They do model individual patient appointment times but their model concerns a single provider with deterministic service times, fixed no-show rate and a target overbooking level that equals the ratio of the number of appointment slots and the mean show-up rate. They argue that overbooking can lead to greater throughput without significantly higher waiting times and clinic finish times.

There are opportunities for making a significant contribution to this topic in future studies. For example, although evidence suggests that longer direct waiting times are a key reason why patients tend to miss appointments, none of the overbooking models have considered a linkage between suggested overbooking targets and the possible increase in no-show rates that could be caused by greater direct waiting times. Similarly, overbooking in multiple provider settings and the benefits of decreased indirect waiting times (because more patients are scheduled per session) are not considered. The interaction between no-shows and walk-ins presents another opportunity for future contributions. Since walk-ins occur at random time epochs and a patient is not deemed a no-show until several minutes past the scheduled appointment time (the authors have encountered a threshold of 10 minutes in several clinics), managing both walk-ins and no-shows is a particularly difficult challenge.

5.3. *Patient-specific resource allocation*

The third class of problems involves patient-specific resource allocation. Patients that have the same diagnoses can nevertheless have significantly different resource requirements. In fact, the length of a surgical procedure is correlated with a variety of known patient characteristics. For example, the time it takes to perform an endoscopy procedure depends on factors such as the presence of polyps; the discovery of one or more polyps requires a biopsy which lengthens the procedure time. Age is a significant risk factor for colon cancer; thus it is correlated with longer procedure times.

There are opportunities for patient-specific resource allocation in specialty services as well. For example, chronic diseases such as arthritis and diabetes tend to have their onset later in life, and both are also influenced by gender. Therefore, resource planning at a tertiary care facility, where patients travel significant distances to receive health services, may benefit from considering the attributes of the population of patients whose visits are scheduled in a particular future week. A higher than average age, or an atypical mix of male versus female patients would affect the optimal balance of resources that should be planned.

How can patient-specific information be used to improve resource allocation decisions? We propose future research focusing on: (i) the discovery of risk factors such as age, gender, body mass index and co-morbidities (multiple complicating diagnoses that might influence the medical outcomes for each individual patient), followed by clustering of patients into different classes based on anticipated capacity utilization and urgency; and (ii) the development of multi-class scheduling and capacity reservation models that account for the variability among classes in patients' needs and resource requirements. The development of such models offers the potential to reduce waiting time and simultaneously increase patient throughput and provider utilization.

5.4. *Patient preferences*

The fourth class of problems concerns patient choices. Virtually all IE/OR models of appointment systems ignore patient scheduling preferences. In reality, patients do exercise choices available to them. Incorporating patient preferences results in models that are mathematically complex and computationally challenging. These issues are further complicated by the fact that patient choice patterns are difficult to glean from available time-stamp data. Specifically, computerized data records track the date and time of each patient's request for appointment and the date and time of the appointment. However, the intervening steps are not recorded. That is, these data do not capture which slots the patient preferred more than the one (s)he actually picked, from the set of choices that are offered. Therefore, it is necessary to develop dynamic models of patient preferences from longitudinal appointment booking data, without increasing the data collection burden. This can be accomplished by using heuristic rules to first classify patients into different choice-based categories and subsequently using dynamic learning algorithms (assuming non-stationary environments) to update and refine this classification scheme. A similar approach can also be used for developing individualized treatment and wellness programs based on clinical data and physician inputs.

Knowing volumes and preference characteristics of patients in each category—e.g., choice with respect to same-day versus future appointment, an appointment with the PCP versus another service provider and morning versus afternoon appointment—the clinic can better match capacity to demand by using strategies such as staggering physicians working hours, making more slots available in the afternoon and encouraging subpopulations without a time-of-day preference to take appointments when the clinic anticipates smaller demand from other subpopulations. Mathematical models can also help develop better rules for protecting the right number of slots for urgent (same-day) requests while maximizing the clinic's revenue. Such decision problems can be modeled within the framework of a MDP model (see Gupta and Wang (2008) for an example).

5.5. *Incentive-based modeling*

The fifth class of problems concerns incentive-based modeling. Much of the previous literature deals with cost-based models. As we argued in Section 3.4, the burden of these costs falls unequally upon the health service network, the service providers (physicians) and the patients, with the result that it is difficult to find consensus about the relative weights of different types of costs. An incentive-based model presents a more natural setting. However, it too can be mathematically challenging due to the presence of multiple objectives and the possibility of both cooperative (bargaining) and non-cooperative approaches to resolving differences. Whereas the economics and the recent IE/OR literatures on supply chain management contain many examples of the use of game-theoretic models to capture different incentives, there are relatively few examples that pertain to health care.

Incentive-based models could be used to determine the lengths of OR block times that should be assigned to each subspecialty (group practice). A key difficulty that OR managers face when making such decisions, either on a periodic basis or whenever new OR capacity is added, is that the true demand for OR time in each subspecialty is unknown. OR managers can use incentives whereby surgeons, who are likely to have more accurate demand information, are encouraged to report the true demand. OR managers can use a principal-agent framework with asymmetric information (examples of this approach can be found in Fudenberg and Tirole (1991), Salanié (1997), and Bolton and Dewatripont (2005)) to model the problem of choosing incentives. The OR managers' problem is that of finding a menu of OR block times and corresponding payment functions for each subspecialty such that it is in the best interest of the surgeons to choose the overall best capacity allocation, i.e., one that maximizes the combined benefit to the hospital, the surgeons and the patients.

5.6. *Scheduling in highly constrained environments*

The sixth class of problems concerns health care delivery environments that are highly constrained. For example, outpatient surgery centers must deliver many services in a specific sequence including patient check-in, nurse intake, surgical preparation, surgery, recovery and check out (Cayirli and Veral, 2005, 2006). Each step requires availability of one or more people resources, such as clinical assistants, nurses, anesthesiologists and surgeons, as well as physical resources, such as operating rooms, diagnostic devices, surgical tools and other equipment. Good outcomes in surgery require that a particular sequence of activities be delivered in a short period of time with minimal waiting at each stage. However, significant uncertainty in the duration of services leads to challenging appointment scheduling problems. The high fixed cost of resources puts pressure on facility managers to schedule high volumes of patients each day, whereas uncertainty in service duration creates resource-use conflicts that are exacerbated by tight schedules. When these factors are added, the environment is one in which it is difficult to determine optimal appointment times for the start of patients' treatments. The problem is further compounded by the occurrence of patient no-shows, tardiness of providers and staff absences.

In order to develop effective strategies for dealing with highly constrained scheduling scenarios, it is often necessary to first quantify the economic impact of different flexibility enhancing choices assuming that all subsequent scheduling decisions are made optimally. Such choices could include redundancy (hospital purchasing multiple copies of critical equipment), cross-training of staff and investment in flexible equipment (e.g., multi-functional beds that can be used for different patient types). The task of quantifying the benefits of these strategies in highly constrained scheduling environments seems well suited for future modeling efforts.

5.7. *Health system design*

The seventh class of problems concerns the design of health care delivery systems. Health care networks realize that the demand for different health services are correlated. For example, patients with chronic diseases have a need for coordinated packages of care—a patient with diabetes may need regular access to a PCP along with specialists such as endocrinologists, cardiologists and neurologists. Therefore, the health care industry is experimenting with different approaches for grouping services and designing multispecialty service facilities. In contrast, the extant IE/OR health care literature has tended to focus separately on each delivery environment. Therefore, the non-traditional modes of delivery discussed below have emerged without the benefit of formal models for planning and coordinating access.

Some primary care clinics are experimenting with the introduction of specialists on site, such as a psychiatrist or a dermatologist. Also, at the primary care level, the concepts of retail health services and concierge service are emerging side by side. For example, some service providers cater only to a limited number of routine diagnoses. Patients can walk-in and expect short waiting times, but no attempt is made to ensure that patients can consult the same provider at each visit. Minute Clinics (www.minuteclinic.com) are an example of such clinics (see Freudenheim (2006) and Phelps (2006) for recent news stories). At the other end of the spectrum, some clinics offer concierge service. Patients enroll with a PCP and pay a fixed fee up front. In return, they get fast personalized service from their PCP who also bills them (or their insurers) for each visit. An example of this type of clinic is the Park Nicollet Clinic in Minneapolis (Haeg, 2002).

The development of new technology and less invasive surgical procedures (e.g., laproscopic surgery) has shifted a large volume of surgeries from the inpatient to the

outpatient setting in recent years. Some surgical services are being offered via a new delivery system called Ambulatory Service Centers (ASCs). ASCs perform elective surgeries in an outpatient setting that can be completed safely with minimal supporting resources. More complex surgeries that require inpatient care and other supporting services (e.g., multi-specialty surgeon teams) are performed at hospitals. In the UK, certain routine elective procedures, such as hip and knee replacements without co-morbidities, are performed at special diagnosis and treatment centers. Such centers benefit from economies of scale and focus (Gupta, 2005). The effects of introducing focused treatment centers on hospitals' case mix has been studied (Bowers and Mould, 2005). However, with the availability of different organizational choices at the primary, specialty and surgical care settings, network models are needed to determine who should provide what services, and how to coordinate access to service providers. Such efforts may require the modeling of a region's population with realistic medical state-transition models to estimate the demand for services and the impact of different network design choices.

5.8. *Education*

The dissemination of IE/OR research to the health care community is an important challenge that has not been addressed well. IE/OR models have been successfully used for improving the efficiency of service systems in many industries including airlines, hotel chains, car rental agencies and natural gas and power. However, the same degree of success has not yet occurred in the health care industry, leading to lack of awareness and skepticism about the potential benefits of IE/OR methodologies.

Physicians occupy top management positions in many health care systems. Typically they have undergraduate degrees in disciplines such as biology, chemistry and psychology. It is much less common for physicians to have training in mathematics and engineering. Moreover, the research methodology that health care providers are most familiar with is the statistical testing of hypotheses through randomized control trials. The methodology of casting decision problems in mathematical models is neither familiar, nor well understood. Therefore, there is a need to disseminate engineering research and case studies of successful implementation of IE/OR methods in high-impact health care journals. It is also important to develop educational offerings of IE/OR methods for health care professionals including short-courses, workshops and course offerings in medical and graduate programs at academic medical centers.

6. Concluding remarks

In this paper, we summarized key issues in designing and managing patient appointment systems for health services. This was intended to clarify the level of complexity encountered in the health care environment. We provided a taxonomy of complicating factors, which made it easier to summarize the contributions of previous research in this area. We exposed open research areas and opportunities for future work.

It is our position that existing models in the manufacturing, transportation and logistics areas cannot be easily "tweaked" to fit the health care environment, and that this, in part, accounts for the lack of adoption of these models in the health care setting. In fact, new models are needed to address health-care-specific issues, such as the soft nature of capacity, the modeling of patient and provider preferences, the stochastic and dynamic nature of multi-priority demand and the need to recover from deviations. Moreover, different modes of organizing health services delivery, as well as technology-led changes in practice norms, provide new opportunities in the area of health services network design.

Some experts see investments in EMR and HIT infrastructure as the key to improving quality and efficiency, and reducing costs of health care delivery systems (Office of the National Coordinator for Health Information Technology 2006). These thoughts were echoed by President Bush in his 2006 State of the Union address, when he said "We will make wider use of electronic records and other health information technology to help control costs and reduce dangerous medical errors" (The Washington Post January 31, 2006). Whereas data availability is necessary for successful calibration of IE/OR models, and EMRs make it easier to implement algorithms for improving access, HIT on its own does not offer a complete solution. Analytical tools are needed to convert data into information, and subsequently, information into smart decisions. IE/OR models can also help inform the designers of health care information systems about what types of data are needed to support future operational decisions.

Acknowledgments

The authors gratefully acknowledge Mr. John Osborn of the Systems & Procedures department at Mayo Clinic, Rochester, MN, for his helpful comments on an earlier version of this article. The authors are also grateful to two anonymous referees and the Editor-in-Chief, Professor Candace Yano, for their help in improving the manuscript. Diwakar Gupta's effort on this project was supported in part by grant CMMI-0620328 from the National Science Foundation. Brian Denton's effort was supported in part by grant CMMI-0620573 from the National Science Foundation.

References

Bailey, N. (1954) Queuing for medical care. *Applied Statistics*, **3**, 137–145.

Bailey, N.T.J. (1952) A study of queues and appointment systems in hospital outpatient departments, with special reference to waiting times. *Journal of the Royal Statistical Society*, **B14**, 185–199.

Belobaba, P.P. (1989) Application of a probabilistic decision model to airline seat inventory control. *Operations Research*, **37**, 183–197.

Blake, J.T. and Carter, M.W. (1997) Surgical process scheduling: a structured review. *Journal of the Society for Health Systems*, **5**(3), 17–30.

Bolton, P. and Dewatripont, M. (2005) *Contract Theory*, The MIT Press, Cambridge, MA.

Bowers, J. and Mould, G. (2005) Ambulatory care & orthopaedic capacity planning. *Health Care Management Science*, **8**(1), 41–48.

Brahimi, M. and Worthington, D.J. (1991) Queuing models for outpatient appointment systems: a case study. *Journal of the Operational Research Society*, **42**, 733–746.

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

Cayirli, T.E. and Veral, H.R. (2005). Comparison of two approaches to patient classification in appointment system design. Decision Sciences Institute Proceedings, San Francisco, CA, 16191–16196, USA.

Cayirli, T.E. and Veral, H.R. (2006) Designing appointment systems for ambulatory care services. *Health Care Management Science*, **9**(1), 47–58.

Chao, X., Liu, L. and Zheng, S. (2003) Resource allocation in multisite service systems with intersite customer flows. *Management Science*, **49**(12), 1739–1752.

Charnetski, J. (1984) Scheduling operating room surgical procedure with early and late completion penalty costs. *Journal of Operations Management*, **5**, 91–102.

Chung, M.K. (2002) Tuning up your patient schedule. *Family Practice Management*, **8**, 41–45.

Denton, B. and Gupta, D. (2003) Sequential bounding approach for optimal appointment scheduling. *IIE Transactions*, **35**, 1003–1016.

Dexter, F. (1999) Design of appointment systems to minimize patient waiting times: a review of computer simulation and patient survey studies. *Anesthesia and Analgesia*, **89**, 925–931.

Dexter, F., Macario, A., Traub, R.D., Hopwood, M. and Lubarsky, D.A. (1999) An operating room scheduling strategy to maximize the use of operating room block time: computer simulation of patient scheduling and survey of patient preferences for surgical waiting time. *Anesthesia and Analgesia*, **89**, 7–20.

Dove, H.G. and Schneider, K.C. (1981) The usefulness of patients' individual characteristics in predicting no-shows in outpatient clinics. *Medical Care*, **XIX**(7), 734–740.

Economist. (2004) A survey of health care finance. **372**(8384), 3–14.

Esogbue, A.O. and Singh, A. (1976) A stochastic model for an optimal priority bed distribution problem in a hospital ward. *Operations Research*, **24**(5), 884–898.

Freudenheim, M. (2006) Attention shoppers: low prices on shots in the clinic off aisle 7. *The New York Times*, May 14, 2006.

Fries, B. and Marathe, V. (1981) Determination of optimal variable-sized multiple-block appointment systems. *Operations Research*, **29**, 324–345.

Fudenberg, D. and Tirole, J. (1991) *Game Theory*, The MIT Press, Cambridge, MA.

Gerchak, Y., Gupta, D. and Henig, M. (1996) Reservation planning for elective surgery under uncertain demand for emergency surgery. *Management Science*, **42**(3), 321–334.

Gupta, D. (2005). Capacity planning and patient flow management. Operations Research for Health Care Delivery Systems, World Technology Evaluation Center, Inc., Baltimore, MD.

Gupta, D. and Wang, L. (2008) Revenue management for a primary care clinic in the presence of patient choice. *Operations Research*, (to appear).

Haeg, A. (2002) Top-shelf health care—if you have the money. Minnesota Public Radio June 24. Available at http://news.minnesota.publicradio.org/features/200206/24_haega_concierg%ecare, downloaded on December 3, 2005.

Heaney, D.J., Howe, J.G. and Porter, A.M. (1991) Factors influencing waiting times and consultation times in general practice. *British Journal of General Practice*, **41**, 315–319.

Ho, C.-J. and Lau, H.-S. (1992) Minimizing total cost in scheduling outpatient appointments. *Management Science*, **38**(12), 1750–1762.

Jansson, B. (1966) Choosing a good appointment system: a study of queues of the type (*D*, *M*, 1). *Operations Research*, **14**, 292–312.

Kim, S. and 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*, **36**(6), 1211–1219.

Klassen, K.J. and Rohleder, T.R. (1996) Scheduling outpatient appointments in a dynamic environment. *Journal of Operations Management*, **14**, 83–101.

Kolesar, P. (1970) A Markovian model for hospital admission scheduling. *Management Science*, **16**(6), 384–396.

Lacy, N.L., Paulman, A., Reuter, M.D. and Lovejoy, B. (2004) Why we don't come: patient perceptions on no-shows. *Annals of Family Medicine*, **2**(6), 541–545.

LaGanga, L.R. and Lawrence, S.R. (2007) Clinical overbooking to improve patient access and increase provider productivity. *Decision Sciences*, **38**(2), 251–276.

Lagergren, M. (1998) What is the role and contribution of models to management and research in the health services? A review from Europe. *European Journal of Operational Research*, **105**, 257–266.

Lautenbacher, C.J. and Stidham, S. (1999) The underlying Markov decision process in the single-leg airline yield management problem. *Transportation Science*, **33**, 136–146.

Leung, J.Y.-T. (2004) *Handbook of Scheduling: Algorithms, Models and Performance Analysis*, Chapman and Hall/CRC, New York, NY.

Liao, C., Pegden, C.D. and Rosenshine, M. (1993) Planning timely arrivals to a stochastic production or service system. *IIE Transactions*, **25**, 63–73.

Liu, L. and Liu, X. (1998) Dynamic and static job allocation for multiserver systems. *IIE Transactions*, **30**, 845–854.

Magerlein, J.M. and Martin, J.B. (1978) Surgical demand scheduling: a review. *Health Services Research*, **13**, 418–433.

McGill, J.I. and Van Ryzin, G.J. (1999) Revenue management: research overview and prospects. *Transportation Science*, **33**, 233–256.

Mercer, A. (1960) A queueing problem in which the arrival times of the customers are scheduled. *Journal of the Royal Statistical Society*, **22**, 108–113.

Mercer, A. (1973) Queues with scheduled arrivals: a correction simplification and extension. *Journal of the Royal Statistical Society*, **35**, 104–116.

Moore, C.G., Wilson-Witherspoon, P. and Probst, J.C. (2001) Time and money: effects of no-shows at a family practice residency clinic. *Family Medicine*, **33**(7), 522–527.

Murray, M. and Berwick, D.M. (2003) Advanced access: reducing waiting and delays in primary care. *Journal of the American Medical Association*, **289**, 1035–1040.

Murray, M. and Tantau, C. (1999) Redefining open access to primary care. *Managed Care Quarterly*, **7**(3), 45–55.

Murray, M. and Tantau, C. (2000) Same-day appointments: exploding the access paradigm. *Family Practice Management*, **7**(8), 45–50.

Office of the National Coordinator for Health Information Technology (2006) Value of HIT. Available at http://www.dhhs.gov/healthit/valueHIT.html, downloaded on March 21, 2006.

O'Hare, C.D. and Corlett, J. (2004) The outcomes of open-access scheduling. *Family Practice Management*, **11**(2), 35–38. Available at http://www.aafp.org/fpm/20040200/35theo.html, downloaded on December 3, 2005.

Penneys, N.S. (2000) A comparison of hourly block appointments with sequential patient scheduling in a dermatology practice. *Journal of the American Academy of Dermatology*, **443**, 809–813.

Phelps, D. (2006) Quick-clinic competition heats up in metro area. *The Star Tribune*, May 25, 2006.

Pierskalla, W.P. and Brailer, D.J. (1994) Applications of operations research in health care delivery, in *Handbooks in OR & MS, vol. 6*, Pollock, S.M., et al. (eds.), Elsevier Science B.V., Amsterdam, The Netherlands.

Rising, E., Baron, R. and Averill, B. (1973) A system analysis of a university health service outpatient clinic. *Operations Research*, **21**, 1030–1047.

Robinson, L.W. and Chen, R.R. (2003) Scheduling doctors' appointments: optimal and empirically-based heuristic policies. *IIE Transactions on Scheduling and Logistics*, **35**, 295–307.

Robinson, L.W., Gerchak, Y. and Gupta, D. (1996) Appointment times which minimize waiting and facility idleness. Working paper, McMaster University, Hamilton, Canada.

Sabria, F. and Daganzo, C.F. (1989) Approximate expressions for queuing systems with scheduling arrivals and established service order. *Transportation Science*, **23**(3), 159–165.

Salanié, B. (1997) *The Economics of Contracts: A Primer*, The MIT Press, Cambridge, MA.

Schaack, C. and Larson, R.C. (1986) An *n*-server cutoff priority queue. *Operations Research*, **34**, 257–266.

Soriano, A. (1966) Comparison of two scheduling systems. *Operations Research*, **14**, 388–397.

Stidham, C.R. and Muckstadt, J.A. (1985) Optimal control of admission to a queueing system. *IEEE Transactions on Automatic Control*, **30**(8), 705–713.

Talluri, K. and Van Ryzin, G. (2004a) Revenue management under a general discrete choice model of consumer behavior. *Management Science*, **50**(1), 15–33.

Talluri, K. and Van Ryzin, G. (2004b) *The Theory and Practice of Revenue Management*, Kluwer, Boston, MA.

Taylor, I.D.S. and Templeton, J.G.C. (1980) Waiting time in a multi-server cutoff-priority queue, and its application to an urban ambulance service. *Operations Research*, **28**(5), 1168–1188.

The Washington Post (2006) President Bush's state of the union address, January 31, 2006. Available at http://www.washingtonpost.com/wp-dyn/content/article/2006/01/31/AR20060% 13101468.html, downloaded on February 3, 2006.

Vanden Bosch, P.M. and Dietz, D.C. (2000) Minimizing expected waiting in a medical appointment system. *IIE Transactions*, **32**(9), 841–848.

Villegas, E.L. (1967) Outpatient appointment system saves time for patients and doctors. *Hospitals J.A.H.A.*, **41**, 52–57.

Wang, P.P. (1993) Static and dynamic scheduling of customer arrivals to a single-server system. *Naval Research Logistics*, **40**, 345–360.

Weiss, E.N. (1990) Models for determining estimated start times and case orderings in hospital operating rooms. *IIE Transactions*, **22**, 143–150.

Welch, J. (1964) Appointment systems in hospital outpatient departments. *Operations Research Quarterly*, **15**, 224–237.

Welch, J. and Bailey, N. (1952) Appointment systems in hospital outpatient departments. *The Lancet*, 1105–1108.

White, M. and Pike, M. (1964) Appointment systems in outpatient clinics and the effect of patients' unpunctuality. *Medical Care*, **2**, 133–145.

Yano, C.A. (1987) Planned leadtimes for serial production systems. *IIE Transactions*, **19**(3), 300–307.

Zhao, W. and Zheng, Y. (2001) A dynamic model for airline seat allocation with passenger diversion and no-shows. *Transportation Science*, **35**, 80–98.

Biographies

Diwakar Gupta is a Professor and Director of Graduate Studies for the Industrial & Systems Engineering Graduate Program in the Department of Mechanical Engineering at the University of Minnesota. He received his Ph.D. in Management Sciences from the University of Waterloo and taught at the McMaster Universitys DeGroote School of Business before joining the University of Minnesota. His research and teaching interests are in the area of stochastic models for health care delivery systems, production & inventory systems and supply chains management. He serves as a Co Editor-in- Chief of the *Flexible Services and Manufacturing* journal and as a Departmental Editor for *IIE Transactions*.

Brian Denton is an Assistant Professor at North Carolina State University in the Edward P. Fitts Department of Industrial & Systems Engineering. Previously he has been a Senior Associate Consultant at the Mayo Clinic in the College of Medicine, and a Senior Engineer at IBM. His primary research interests are in optimization under uncertainty as it relates to industry applications in health care delivery, medical decision making, supply chain planning and factory scheduling. He completed his Ph.D. in Management Science, his M.Sc. in Physics, and his B.Sc. in Chemistry and Physics at McMaster University in Hamilton, Ontario, Canada.