

Advanced Access

Reducing Waiting and Delays in Primary Care

Mark Murray, MD, MPA

Donald M. Berwick, MD, MPP

TWO CARDINAL GOALS OF PRIMARY care are accessibility and continuity of care. Many primary care practices are struggling to achieve these goals, engulfed by seemingly overwhelming demand for patient visits and chaotic procedures for triaging patients into crammed office schedules. Too often, patients are unable to see their own primary care physician in a timely fashion, resulting in delays in care and disruption of patient-physician continuity.

To confront this problem, many large and small primary care practices across the United States and Europe have initiated an approach known as *advanced access*, *open access*, or *same-day scheduling*. Technically, it is now possible to meet expectations for essentially wait-free care. Examples abound of health care organizations and small primary care offices that have reduced delays significantly without adding resources.¹

In this fifth article of the series Innovations in Primary Care, we explain the advanced access model. The accompanying article presents case studies of medical practices instituting advanced access and discusses pitfalls that can derail the innovation.

PROBLEM OF INADEQUATE ACCESS

Mr W had abdominal pain. After meals, a wave of cramps would spread across both lower quadrants. He experienced no diarrhea, vomiting, or melena. Calling his internist's office, he was told that the next available appointment was in 3 weeks; if

Delay of care is a persistent and undesirable feature of current health care systems. Although delay seems to be inevitable and linked to resource limitations, it often is neither. Rather, it is usually the result of unplanned, irrational scheduling and resource allocation. Application of queuing theory and principles of industrial engineering, adapted appropriately to clinical settings, can reduce delay substantially, even in small practices, without requiring additional resources. One model, sometimes referred to as *advanced access*, has increasingly been shown to reduce waiting times in primary care. The core principle of advanced access is that patients calling to schedule a physician visit are offered an appointment the same day. Advanced access is not sustainable if patient demand for appointments is permanently greater than physician capacity to offer appointments. Six elements of advanced access are important in its application: balancing supply and demand, reducing backlog, reducing the variety of appointment types, developing contingency plans for unusual circumstances, working to adjust demand profiles, and increasing the availability of bottleneck resources. Although these principles are powerful, they are counter to deeply held beliefs and established practices in health care organizations. Adopting these principles requires strong leadership investment and support.

JAMA. 2003;289:1035-1040

www.jama.com

his problem was urgent he could go to the emergency department. Mr W liked his internist and hated emergency departments. Once he spent 6 hours in an emergency department and came away with nothing more than a note telling him to contact his physician. Mr W's pain continued for 3 weeks. Barely able to eat, he lost 8 pounds. When the appointment day arrived, his internist found an abdominal mass that proved to be a partially obstructing diverticular abscess.

Waits and delays in health care have been a problem for many years.^{2,3} The 1999 Kaiser Family Foundation survey⁴ of insured adults younger than 65 years found that 27% of people with health problems had difficulty gaining timely access to a clinician. Forty per-

cent of emergency department visits are not urgent. Many take place because of an inability to obtain a prompt primary care appointment.^{4,5} From 1997 to 2001, the percentage of people reporting an inability to obtain a timely appointment rose from 23% to 33%.⁶ In 2001, 43% of adults reporting an urgent condition were sometimes unable to receive care as soon as they wanted.⁷ A 2001 women's health survey found that 28% of

Author Affiliations: Mark Murray & Associates, Sacramento, Calif (Dr Murray); and Institute for Healthcare Improvement, Boston, Mass (Dr Berwick).

Corresponding Author and Reprints: Donald M. Berwick, MD, MPP, Institute for Healthcare Improvement, 375 Longwood Ave, Boston, MA 02215 (e-mail: dberwick@ihi.org).

Section Editor: Drummond Rennie, MD, Deputy Editor.

See also p 1042.

women in fair or poor health reported delaying care or failing to receive care because of an inability to obtain a timely physician appointment.⁸ In its landmark report *Crossing the Quality Chasm: A New Health System for the 21st Century*,⁹ the Institute of Medicine's committee on quality of care in America designated "timeliness" as 1 of the 6 key "aims for improvement" in health care.

Many primary care practices are in a state of disarray because of overfilled appointment books. Receptionists may spend 10 minutes on the telephone negotiating appointment times with patients, causing other patients to wait interminably on hold. As the day progresses, the stack of messages on the physician's desk—each requiring a chart to be pulled and later refiled—grows as patients insist on speaking with their physician. Medical assistants and nurses are mired in telephone triage, attempting to determine the urgency of patients' problems. Patients with urgent-sounding problems are squeezed in during lunch hours and into the early evening. The task of returning telephone calls to patients who were unable to schedule an appointment because of crammed schedules lengthens a physician's seemingly endless day. Upon instructing receptionists to reschedule a patient with diabetes in a month, physicians are called on the intercom and told that no slots are open at that time. One physician described the reception desk as a war zone with patients and receptionists battling over appointment times. It is hard to judge who is more stressed and dissatisfied—patient, receptionist, or physician.

Scientific Foundations of Delay Reduction

Despite widespread beliefs to the contrary, waits, delays, and restricted access are rarely symptoms of inadequate resources. Analysis of most waiting reveals problems in matching physician capacity to offer appointments with patient demand for appointments on a day-to-day basis, rather than an absolute lack of capacity.¹⁰ Using principles from industrial engineering

and queuing theory,¹¹ health care systems can reduce or eliminate delays without adding resources. Although these management principles are rational and straightforward, they are, unfortunately, not always easy to apply in health care settings.

The primary barriers are psychological because the principles run counter to deeply held beliefs about scheduling systems and what can be achieved with the resources at hand. The major barriers are the fear of change and the lack of confidence that existing resources can meet the demand for care. Often, clinicians think that they need their scheduling systems to protect them from looming demand and that they will become overwhelmed if they try to meet each day's demand on that day. They point to their backlogs and long work days to prove it.

Empirically, however, it is usually the case in office practices that absolute supply and absolute demand are well matched, despite strong beliefs to the contrary and notwithstanding the evidence of long queues. In fact, the lengths of the queues in most health care systems remain steady at a given number of weeks or months, rather than growing without limit as they would if the supply were absolutely insufficient.

To understand advanced access, it is helpful first to understand the 2 prevailing alternative approaches to managing health care demand—the traditional model and the carve-out model—and why they fail so often.

Traditional Model: Meet Urgent Demand Now and Meet Nonurgent Demand Later

The traditional model stratifies appointment demand into 2 streams: *urgent* (same-day) and *nonurgent*. It seems logical that a stressed system, fraught with delays, ought to meet the most pressing needs for care quickly, even if doing so requires meeting routine needs sometime in the future. In this model, a patient typically contacts a receptionist, who determines the urgency of the clinical condition and checks with a nurse or

physician for approval to bring the patient into an already full schedule, often by double-booking an appointment slot. If the schedule is saturated, the practice may send the patient elsewhere for immediate help—to an urgent care clinic or emergency department.

Delays are the result of mismatches—usually temporary—between supply and demand. The traditional model worsens that mismatch by reducing supply. In the short term, this system diverts potentially productive time of clinicians, receptionists, and others into the noncare processes of decision making and triage. In the longer term, it diverts staff into the time-consuming process of managing an intentionally created delay, such as making reminder calls or filing and retrieving lists for future appointments.

Moreover, the traditional model artificially increases demand in several ways. First, an urgent visit usually addresses only 1 of the patient's problems (the urgent one), thus forgoing the opportunity to meet several needs in a single visit. Second, diverted patients, sent to other settings for immediate care, often want to see their personal clinician for reassurance and follow-up later on. Clinicians in emergency departments and urgent care clinics quite responsibly tell patients to "contact your own doctor in the morning," thus converting the need for 1 visit into a need for 2 visits. At one site, 49% of patients seeing an unfamiliar physician made such a return visit.¹²

As most clinicians already feel intuitively and as many frustrated and angry patients say out loud to beleaguered front office staff, the traditional model does not work well.

Carve-Out Model: Predict Urgent Demand and Reserve Time to Meet It

Clinicians sometimes seek relief by replacing the traditional model with the carve-out model, which reserves urgent care time in advance. Carve-out models function better than the traditional model but incur their own problems. Some forms of the carve-out model

reserve a supply of urgent care by designating a “triage doctor of the day” or “jeopardy doc.” Patients with urgent care needs can see someone, but usually not their own clinician, which threatens continuity of care and creates the same artificial demand for extra return visits with the patient’s personal physician as the traditional model does. Other carve-out models set aside appointments in each clinician’s schedule, often reserving more carved-out time than is strictly needed to meet the aggregate demand, thus pushing even more nonurgent demand into the future.

Carve-out models create dysfunctional habits that further impede accurate matching of supply and demand. Office practices using carve-out models often develop vast informal systems to “steal” appointments reserved for patients with urgent problems and other special needs, sometimes administered as an elaborate system of favors. Triage decisions are often wrong since the worried well may get urgent slots while the stoic sick do not.

Under both delayed access models, patients labeled as nonurgent can be described as swimming in a lake. Some of the patients swim quietly as they tread water waiting for their medical visit. Others have needs that cause them to splash around noisily, calling for an earlier appointment, going to the emergency department, requesting specialty consultations and prescriptions by telephone, or showing up as drop-ins.

Advanced Access Model: Do Today’s Work Today

The advanced access model, which attempts to eliminate appointment delay, drains the lake. Under this model, patients calling to see their physician are offered an appointment the same day. Under traditional appointment systems, patients may see an unfamiliar physician even if their own physician is present because the regular physician’s appointment slots are full. The proven benefits of continuity of care are lost.¹³ The new model can improve continuity because all physicians have appointment slots available.

No scheduling system, including advanced access, can work if a physician has too many patients. Patient demand for visits and physician capacity to schedule visits must be in balance. Advanced access can work well even if demand exceeds capacity on a given day, but if demand permanently exceeds capacity, no system will work, neither the traditional model, the carve-out model, nor the advanced access model.

Advanced access rejects the seductive idea of sorting demand into 2 queues, routine and urgent. The primary design objective of an advanced access system is to do today’s work today. In this sense, advanced access uses the same ideas as the methods of *one-piece flow*, *just-in-time engineering*, or *lean thinking* that are now standard in most modern manufacturing and service industries.¹⁴⁻¹⁸

The advanced access model sorts appointment demand by clinician, not by clinical urgency. The crucial question for allocating appointments is simply: “Is your personal clinician here today?” In this model, each clinician manages on a daily basis his or her own patients’ demands for office care, without regard to urgency. Some demand is pushed into the future, including visits for patients who decline the offer of an appointment today, and for patients who are seen today and need to return at some definite future time. Generally, however, these postponements of demand are not defects but deliberate choices.

Queuing theory demonstrates formally how long waits can exist even when adequate supply exists.¹¹ The relevant mathematics is beyond the scope of this article, but, qualitatively, the answer lies in *variation in demand*. An office practice can have enough supply to meet the demand, but because the office carries a 1-month backlog of demand, the average wait for an appointment is 1 month—the match of demand and supply is always a month too late. Furthermore, the matching of supply to demand is a dynamic process, not a static one. A highway can have enough toll booths for the average number of

passing vehicles per minute and yet have long traffic jams at rush hour. The traffic jams do not prove that capacity is overall insufficient but rather that the temporal profile of capacity does not match the temporal profile of demand.

Advanced access tries to close this gap in time between supply and demand for all demand, routine as well as urgent. To do so, it adopts a strategy opposite to that inherent in the other 2 models. Whereas the other 2 models protect current capacity by pushing a large portion of today’s work into the future, the advanced access model protects future capacity by pulling all current work into the present. The definition of *the present* in this strategy depends on the clinical context. For a primary care office, the useful time frame is *today*: “Do today’s work today.” A specialty practice may need to lengthen that time frame: “Do this week’s work this week.” An emergency department or operating room suite may need to shorten it: “Do this hour’s work this hour.” In the best advanced access models, space on the schedule tomorrow is a result of doing today’s work today.

One caveat should be made about doing today’s work today. Many primary care physicians do not work every day. A patient calling to request an appointment with a physician not present that day should be given the choice of seeing another physician today or waiting to schedule an appointment with his or her physician later in the week. The patient can then balance the value of continuity of care against the competing value of immediate access.

The advanced access model must be data driven. Before implementing the reform, practice sites must have a solid understanding of the size of their patient population, level of patient demand for visits, and number of appointment slots available. These data can be calculated using the measures listed in the **BOX** to determine whether demand and capacity are in balance. After achieving same-day appointments, daily measurements of appointment availability—third next available ap-

Box. Evaluation and Monitoring Measures for Advanced Access**Demand**

The number of patient calls for appointments during a day (regardless of when these appointments are scheduled), plus the number of walk-ins, plus the number of follow-up appointments generated by physicians at the practice site. Demand, which is not always easy to measure, can be estimated by keeping daily records.

Capacity

The number of appointment slots per day for each clinician multiplied by the number of clinicians. Capacity can be subdivided according to physician vs nonphysician clinician or family practitioner vs pediatrician vs internist.

Panel Size

Sometimes used to estimate demand. In a capitated primary care practice, it is the number of patients enrolled to that physician. In fee-for-service and mixed practices, it is defined as all patients seen by a physician in the past 18 months (12 months appears to undercount, and 24 months to overcount panel size). Given an average patient panel, not overly weighted with elderly and chronically ill people, about 0.7% to 0.8% of the panel will call for an appointment on the average day.¹⁰ For a panel of 2500, this means 17 to 20 daily calls. For panels with high-risk patients, demand rises markedly—a reality that limits the utility of measuring panel size to predict demand.

Third Next Available Appointment

This statistic is used to measure the number of days a patient has to wait to get an appointment. The third next available physical examination is a sentinel marker. Physical examination is used rather than another appointment type because it is usually the latest scheduled. If access to physical examinations improves, all availability improves. The third appointment is featured because the first and second available appointments may reflect openings created by patients cancelling appointments and thus does not accurately measure true accessibility. This measure is easily obtained, daily or weekly, by the receptionist while counting the number of days until an opening for the third next physical examination appointment is on the schedule.

Future Open Capacity

The number of open appointment slots divided by the total number of appointment slots over the next 4 weeks. In 1 delivery system that used advanced access, some physicians enjoyed a ratio of 80% to 90% while others—those having many patients preschedule appointments—had ratios of only 10% to 15%. Panels with young, healthy patients can achieve 90% of appointment slots open while geriatric and newborn practices will have lower ratios.

Continuity of Care

The percentage of total visits that are visits to the patient's personal physician.

pointment and future open capacity—are required to sustain and institutionalize the reform.

In advanced access models, the proportion of a clinician's personal schedule that is open (ie, with no patients booked in advance) at the beginning of each clinical workday rises substantially. Whereas schedules in the traditional and carve-out models com-

monly have fewer than 10% of appointment slots open at the beginning of the day, with advanced access, the proportion of open slots at the beginning of each workday rises to about 50%. The goal of 50% open appointments may not be achievable for physicians with many elderly patients or many patients with long-term illnesses who require pre-scheduled appointments. However, if

capacity overall exceeds demand, advanced access will succeed even in such circumstances.

By offering all patients an appointment today, the model virtually eliminates the triage function, freeing up personnel for other tasks, and reducing physician interruptions and telephone call-backs. The rate of "no-shows" goes down, avoiding the logjams that are the result of intentionally overbooking appointments in anticipation of a high rate of missed appointments.¹⁹ Hearing about advanced access for the first time, some physicians think that patient demand for appointments will become insatiable, creating more and more work each day. In fact, systems implementing advanced access have found that patient demand decreases simply because patients are more often able to see their own clinician.¹⁰

MAKING ADVANCED ACCESS WORK

The tactics of advanced access are broadly of 3 types: (1) reduce the time interval between when the demand is presented and when it is met; (2) appropriately reduce, or shape, the demand; and (3) appropriately increase the supply, especially the bottleneck supply. To implement the system, most medical practices must make 6 specific changes.

Balance Supply and Demand

Advanced access improves the allocation of supply to demand by making better predictions of both and then acting according to the predictions. This requires that practices measure supply and demand at a level of precision unfamiliar in most clinical units. Some systems try to measure demand retrospectively: "How many patients have we seen on Tuesday mornings?" Unfortunately, retrospection measures supply, not demand, because both the traditional and carve-out models force the natural demand into artificial streams. Historical patterns of encounters are inaccurate surrogates for true, underlying demand. To be accurate, demand

must be measured prospectively; it requires inquiry and record-keeping about what appointments patients actually ask for (external demand), and what follow-up appointments clinicians actually request (internal demand).

Measuring supply is easier than measuring demand. How much clinician time is available? What units of service can clinicians provide? Knowing actual demand and supply then enables practices to reduce the gaps between the 2 by reallocating return appointments in a way that smoothes out the overall flow of demand. Physicians will have to adjust their schedules such that, for example, more physicians are available on Mondays, the day of the week when patient demand is the highest.

Work Down the Backlog

Most office practices are too jammed up doing last month's work today to be able to adopt immediately the principle of doing today's work today. To achieve advanced access, the practice must eliminate the backlog that has accumulated due to the strategy of pushing demand into the future.

Backlog reduction is not an ongoing feature of advanced access; it is a 1-time, up-front step to clear the books for the new system. Eliminating the backlog involves no magic; it requires the temporary tactic of doing more work each day than is generated internally or externally for that day. Temporarily adding capacity (through extra sessions, locum tenens, or extended work hours) is the most straightforward approach. For example, a practice that sees 50 patients per day can eliminate a month-long backlog of future appointments by temporarily seeing an extra 25 patients per day for a period of 2 months. Other strategies for eliminating a backlog include using alternative forms of supply such as telephone calls, e-mail interactions, or outside referrals. No matter what tools are used, eliminating the backlog inevitably requires a bolus investment of resources, strong leadership support, and rational incentives.

Reduce the Number of Appointment Types

As noted, in advanced access models, the key question for managing demand is: "Is the patient's personal clinician present today?" If yes, the patient is seen today. If no, the patient can, according to his or her preference, be seen today by someone else or by the personal clinician on a future date. Some practices establish 2 appointment lengths, developing a simple cue to distinguish between shorter and longer appointments; the schedule is kept simple and flexible by combining 2 short appointments when a long appointment is needed.

Develop Contingency Plans

Contingency plans are needed to keep supply and demand in balance on a daily basis, despite inevitable variations in either. Scheduling algorithms that work well most of the time will fail under special circumstances; such circumstances require plan B.

Demand for appointments surges at such times as back-to-school physicals, influenza season, the day after Thanksgiving, and the days following clinician vacations. Although these surges could interfere with doing today's work today if they are not well managed, clinical units can predict most demand surges, and respond by restricting pre-scheduled appointments and increasing clinician capacity on those days. Supply can vary, too, even more than demand, due to provider vacations, illnesses, and absences to attend professional meetings. Whether the increased mismatch is due to increased demand or reduced supply, clinical practices need to adopt innovative methods, such as telephone prescriptions, systems for handling the end of the day, dividing the work of absent physicians, and expanding the use of nonphysician clinicians such as nurse practitioners, in closing these temporary gaps.

Reduce and Shape the Demand for Visits

Reducing and shaping demand can also facilitate better matching of supply to

demand. Among the most powerful demand reduction strategies in ambulatory health care is continuity of provider. The experience of many primary care practices achieving advanced access demonstrates that the total demand for visits falls when patients see their own, personal clinicians.

Other strategies for reducing demand include maximizing the effectiveness of each visit by covering multiple issues at 1 sitting (some practices call this *max-packing*), using the telephone or e-mail instead of visits to respond to patients' questions and to do follow-up care, developing group medical visits,²⁰ and extending the intervals between return visits for patients with long-term illness. Many return-visit intervals became conventional years ago, without evidence of their effect on clinical outcomes.²¹ Providing patients and families with home-care educational and reference materials²² may safely and effectively help reduce the number of unnecessary requests for care.

Increase the Effective Supply, Especially of Bottleneck Resources

In the routine flow of health care, as in all systems, a single rate-limiting constraint or "bottleneck" determines system throughput. Identifying and optimizing the use of the bottleneck resource helps reduce delays. In ambulatory care, the bottleneck resource is almost always the time of clinicians. The advanced access model attempts to maximize efficiency by transferring from physicians tasks that can be done by someone else. This means giving higher levels of responsibility, under well-designed guidelines, to staff other than physicians and nonphysician clinicians. Some practices supplement the primary physician's care with the care of others on the team—nurses, health educators, or medical assistants—for return visits in which the patient-physician relationship is well established and the treatment plan clearly outlined.²³

CONCLUSION

Long waiting times for care are so familiar to patients and clinicians that

most have become numb to them or have given up hoping for anything different. Waits, delays, and deferred access to care seem inevitable. They are not. Within current resource constraints, planned, rational changes in the way health care systems manage supply and demand can achieve major improvements in the timeliness of care, with no increase in the burden of

work for clinicians and others, in much the same way that organizations outside of medicine have improved the flow of work. As the accompanying article describes, many primary care practices have made those changes, some successfully, although not without effort and important local adaptations. These practices reap large dividends in the form of more satisfied

patients, less stressed staff, and levels of timeliness and clinical continuity that they had not thought possible. In the management of patient flow, as in the management of disease, planning and rational system design, used well, can solve problems.

Disclaimer: Mark Murray & Associates assists medical practices to institute advanced access and other practice improvements.

REFERENCES

1. Nolan TW, Schall MW, Berwick DM, Roessner J. *Reducing Delays and Waiting Times Throughout the Healthcare System*. Boston, Mass: Institute for Healthcare Improvement; 1996.
2. Goitein M. Waiting patiently. *N Engl J Med*. 1990;323:604-608.
3. Waiting patiently [letter]. *N Engl J Med*. 1991;324:335-337.
4. *National Survey of Consumer Experiences With Health Plans*. Menlo Park, Calif: Kaiser Family Foundation; June 2000.
5. Cunningham PJ, Clancy CM, Cohen JW, Wilets M. The use of hospital emergency departments for non-urgent health problems: a national perspective. *Med Care Res Rev*. 1995;52:453-474.
6. Strunk BC, Cunningham PJ. *Treading Water: Americans' Access to Needed Medical Care, 1997-2001*. Washington, DC: Center for Studying Health System Change; March 2002.
7. Greenblatt J. *Access to Urgent Medical Care, 2001*. Rockville, Md: Agency for Healthcare Research and Quality; 2002. Statistical brief No. 08.
8. *Women's Health in the United States: Health Coverage and Access to Care*. Menlo Park, Calif: Kaiser Family Foundation; May 2002.
9. Institute of Medicine. *Crossing the Quality Chasm: A New Health System for the 21st Century*. Washington, DC: National Academy Press; 2001.
10. Murray M, Tantau C. Same-day appointments: exploding the access paradigm. *Fam Pract Manag*. 2000;7:45-50.
11. Hall R. *Queueing Methods for Services and Manufacturing*. Englewood Cliffs, NJ: Prentice Hall; 1991.
12. Lippman H. Same-day scheduling. *Hippocrates*. February 2000;49-53.
13. Starfield B. *Primary Care*. New York, NY: Oxford University Press; 1998.
14. Ohno T. *Toyota Production System*. Cambridge, Mass: Productivity Press; 1988.
15. Spear K, Bowen HK. Decoding the DNA of the Toyota production system. *Harvard Bus Rev*. 1999;77:96-106.
16. Womack J, Jones D, Roos D. *The Machine That Changed the World*. New York, NY: HarperPerennial; 1991.
17. Womack J, Jones D. *Lean Thinking*. New York, NY: Simon & Schuster; 1996.
18. Goldratt E, Cox J. *The Goal: A Process of Ongoing Improvement*. Great Barrington, Mass: North River Press; 1992.
19. Singer IA. *Advanced Access: A New Paradigm in the Delivery of Ambulatory Care Services*. Washington, DC: National Association of Public Hospitals and Health Systems; 2001.
20. Noffsinger EB, Scott JC. Understanding today's group visit models. *Group Practice J*. 2000;49:48-58.
21. Schwartz L, Woloshin S, Wasson J, et al. Setting the revisit interval in primary care. *J Gen Intern Med*. 1999;14:230-235.
22. American College of Physicians. *Complete Home Medical Guide*. New York, NY: DK Publishing; 1999.
23. Patel V, Cytryn K, Shortliffe E, Safran C. The collaborative health care team: the role of individual and group expertise. *Teach Learn Med*. 2000;12:117-132.

Knowledge shrinks as wisdom grows.
—Alfred North Whitehead (1861-1947)