Timeliness and Efficiency

Providing Timely Access to Care: What is the Right Patient Panel Size?

Linda V. Green, Ph.D.
Sergei Savin, Ph.D.
Mark Murray, M.D., M.P.A.

Difficulty in obtaining a timely appointment to see a physician is a common problem. In one study, 33% of patients cited "inability to get an appointment soon" as a significant obstacle to care,[1] and the Institute of Medicine has identified "timeliness" as one of the six key "aims for improvement" in its major report on quality of health care.[2]

Primary Care and Advanced Access

For most patients, their primary care physician is their major access point into the health care system. Yet primary care practices often have long waits for appointments and may have difficulty in accommodating patients who have potentially urgent problems. As a result, patients experience delays in treatment and may be seen by someone other than their own physician, potentially leading to adverse clinical consequences, patient dissatisfaction, and loss of revenue for the practice. Large backlogs may require additional staff and resources to deal with patients trying to get appointments for the same day and are often correlated with a high rate of cancellations or "no-shows," which can result in lost income and wasted capacity.[3]

To remedy this problem, some primary care practices have adopted a patient scheduling approach known as *advanced access.* As opposed to a "traditional" system where each physician's daily schedule is fully booked in advance or a "carve-out" model in which a fixed number of appointment slots are held open for urgent cases, the goal of the advanced access approach is to reduce delays by offering every patient a same-day appointment, regardless of the urgency of the problem. The fundamental idea behind advanced access is to "do all of today's work today"

Article-at-a-Glance

Background: Delays for appointments are prevalent, resulting in patient dissatisfaction, higher costs, and possible adverse clinical consequences. A "just-in-time" approach to patient scheduling, called advanced access, has been effective in reducing delays in multiple clinical settings. Offering most patients appointments on the same day requires achieving an appropriate balance between supply of and demand for appointments, but no methods have been previously proposed to determine what this balance should be.

Methods: A measure of balance is termed the *overflow frequency level*—the fraction of days when demand exceeds the average number of appointment slots available. A probability model was developed to estimate this measure for any practice. The model can be used in identifying an appropriate panel size or, conversely, the physician capacity needed to provide timely access.

Results: Delays for appointments will be excessive unless the ratio of the average daily demand for appointments to the average daily capacity is less than one. This ratio's appropriate value is dependent on the desired *overflow frequency level,* which indicates the fraction of days for which physician overtime would be necessary to offer most patients same-day appointments. A table provides suggested panel sizes for a range of practice types, and a spreadsheet file is available on request to help determine panel size or physician capacity in any specific situation.

Conclusion: The simple probability model can be used to improve the timeliness of care while considering the constraints on physicians' working hours.

Patient Backlog When Average Daily Patient Demand Equals the Appointment Capacity

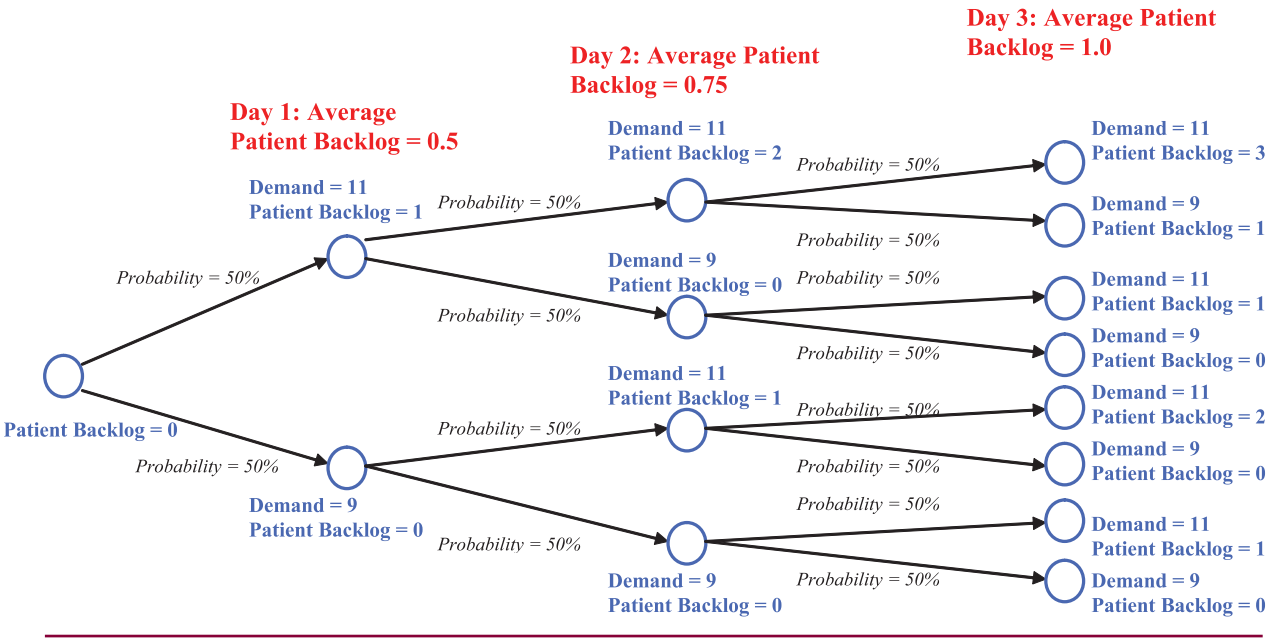

Figure 1. *The example illustrates the growing patient backlog in the case when average daily patient demand equals the appointment capacity.*

so that patients don't have to wait for appointments, practices don't waste capacity holding appointments in anticipation of same-day needs, and patients have a greater likelihood of seeing their own physician. Several success stories have documented the benefits of this approach in both managed care and fee-for-service environments, including dramatically shorter waits, higher levels of continuity of care, less wasted capacity for the practice, and increased patient, staff, and physician satisfaction.[3]

Advanced access can only work if patient demand for visits and physician capacity to see patients are "in balance." Advocates of advanced access have identified several ways in which the number of visits can be reduced, physician time can be better leveraged, and scheduling practices can be streamlined so as to achieve a better balance between supply and demand.[3,4] However, in discussions with practitioners, we have found that questions remain about what constitutes an appropriate balance and, more specifically, what is a "manageable" panel size. The answers to these questions are not obvious and require a quantitative approach.

The Need for "Safety" Capacity

A fundamental feature of patient demand for primary care is its random nature: the actual number of patients requesting care on any particular day will vary around the average daily value, sometimes substantially. It is this inherent randomness that makes it difficult to determine the answers to questions such as: "How large a panel size can be served by a given physician practice?" If not for this variability in demand, the answer would be obvious—the panel size would be the one that made the daily demand for care equal to the daily number of physician appointment slots available. However, with this variability, making supply and demand equal on average would create chronic backlogs for care and waits for appointments that would likely get longer and longer.[5] Although this characteristic of service systems has been known to operations professionals for decades, it may seem counterintuitive. A simple example, illustrated in Figure 1 (above), may help explain this critically important concept.

Consider a primary care practice that has a daily patient demand for appointments that takes on only two

possible values—11 and 9, each with 50% probability. Suppose the maximum number of patients that can be seen each day is exactly equal to 10, so that any "excess" demand must be pushed to the next day that has available appointment slots. Figure 1 illustrates all possible realizations of patient backlog values for a period of three days, assuming we start with no backlog. As shown, the average backlog grows from 0.5 patients at the end of the first day to 0.75 at the end of the second day to 1.0 at the end of the third day. If this exercise is carried out further, the average patient backlog will continue to grow from day to day. This may be surprising because it seems logical to assume that "bad" days, that is, days with a demand of 11, will be balanced out by "good" days, those with only 9 new patient demands.

So why doesn't this balancing out happen? As our simple example shows, when patient demand is less than the appointment capacity, *the extra service capacity cannot be transferred to the next day* to serve future patient demand and is therefore lost. On the other hand, on the "bad" days, when patient demand exceeds service capacity, the unserved demand does not disappear and has to be satisfied in the future. So "good" days cannot clear the backlog created by the equal number of "bad" days. Furthermore, if the demand variability is increased, for example, by adding possible demands of 8 and 12 patients, the average backlog will grow faster.

Thus, if the goal is to provide immediate access to care with a high probability, then the average daily demand for appointments must be *strictly less* than the maximum capacity to see patients. Another way of saying this is that there must be some *safety capacity* relative to demand. Safety capacity, the amount of capacity in excess of average demand, serves as a hedge against demand variability. Without it, a practice will be unable to offer timely care to its patients.

Finding the Right Balance Between Supply and Demand

How much safety capacity does any specific practice need? This depends primarily on the desired *overflow frequency level*—the percentage of days when demand exceeds the number of appointment slots for that day. In the example illustrated in Figure 1, the overflow frequency is 50%. The lower the overflow frequency level, the easier it will be to offer same-day appointments by occasional use of physician overtime. Decreasing the overflow frequency level can only be accomplished by increasing the safety capacity. However, more safety capacity also means more days and hours when physicians are not seeing patients. So the "right" level of safety capacity for any given office must be a subjective determination that will likely be based on the trade-off between the revenue associated with seeing more patients and the amount of overtime the practice is willing to undertake to keep patient delays minimal. To evaluate the possible trade-offs, it is necessary to understand the relationship between safety capacity, patient panel size, and overflow frequency.

A Modeling Approach

Safety capacity can be created by either increasing physician capacity or decreasing demand. Physician capacity may be increased by adding appointment slots to the day, and demand might be reduced by using tactics such as greater use of telephone and e-mail and the use of group visits. However, panel size is the major determinant of demand and the prime lever for achieving the right balance between supply and demand.

We have developed a simple quantitative model to help evaluate the trade-offs associated with a given panel size. Since the only objective of the model is to help identify a good balance of overall supply and demand for a given practice, it is not necessary for the model to distinguish between "external" demands, that is, those that are generated by patients' actions, and "internal" demands, that is, those that are the result of the physician's decision to see a patient for follow-up work or chronic care. No matter the source of the demand, all demands must be satisfied in a timely fashion, and doing so requires a panel size that allows for sufficient safety capacity.

The model does not address the "micro-management" issues such as daily scheduling of follow-up visits, dealing with cancellations, or scheduling of physicians' office hours and vacations. Although these are all important factors for the efficient functioning of the practice, they do not significantly affect the best choice of panel size and so are not needed in our "macro" model. We will revisit these issues later in the article.

Although about $^{2}/_{3}$ of all primary care physicians work in group practices,[6] a number of studies[7–10] have

documented the benefits of continuity of care. These observations support the view that a patient should be seen, whenever possible, by his or her physician, and therefore, that a panel should be associated with an individual physician. However, as described later, our approach can easily be extended to allow for determining a panel size for multiple physicians working as a team.

Finding the Right Panel Size

Establishing an appropriate panel size for an existing practice consists of the following six steps:

1. Identifying the current panel size
2. Estimating the daily visit rate per patient
3. Fixing the number of daily appointment slots
4. Calculating the current overflow frequency
5. Setting the target overflow frequency
6. Computing the panel size based on the target flow frequency

Steps 5 and 6 can be done iteratively to identify a desirable trade-off between panel size and overflow frequency.*

1. Identifying the Current Panel Size

In many managed care practices, the patient panel size N_{cur} is simply the number of patients enrolled with a physician. However, in fee-for-service or mixed practices, the number of patients "on file" may be misleading because it is not uncommon to preserve files for patients who may no longer be using the practice's services. In these situations, it has been found that the panel size will be most accurately estimated by calculating the total number of distinct patients seen by a physician in the last 18 months. (Use of a year may underestimate the effective panel size, whereas the two-year count typically produces an overestimated value[4]).

In a multiphysician practice, estimating the current panel size for each physician may be more complicated because a given physician's patient may see another physician if his or her preferred provider is unavailable. Therefore, in these practices, it is important to track for each physician the number of requests for appointments rather than the number of actual visits.

*A spreadsheet file that provides all necessary computations for these six steps is available from the authors by e-mail request.

2. Estimating the Daily Visit Rate per Patient

The most accurate assessment of daily demand requires prospective measurement of the specific appointment dates that patients actually ask for, including walk-ins (external demand), as well as the follow-up visit dates that physicians actually request (internal demand). If prospective data are not available, an estimate can be obtained by examining appointment logs for a recent period of time, for example, 18 months, and counting the number of appointments over that period of time.

Let T be the number of working days for the period of time being examined and let A be the number of patient appointments (or, if available, requests for appointments) for those T days. Then, as shown in Figure 2 (page 215), the daily visit rate per patient p is calculated by dividing A by the product of the number of patients on the current panel, N_{cur} and T: $p = \frac{A}{N_{cur} \times T}$

For example, consider a general/family practitioner with a current panel of $N_{cur} = 2500$ patients who had $A = 6500$ office visits during the last 18 months ($T = 315$ days). For this practice, $p = \frac{A}{N_{cur} \times T} = \frac{6500}{2500 \times 315} = 0.008$ visits/day per patient.

3. Establishing the Number of Daily Appointment Slots

The average daily supply of appointment slots, C, is determined by the average length of an appointment slot and the average daily number of hours devoted to direct patient care. So if a physician spends an average of 7 hours per day in patient care and appointments are scheduled 20 minutes apart, the daily appointment capacity is $C = 7$ hours $\times$ 3 appointments/hour = 21 appointments. If a practice has a varying number of appointment slots per day during the week, C should be the *average* number of slots per day.

4. Calculating the Overflow Frequency

Consider a practice with panel size N and with daily demand rate p. If each patient request for care is generated independently of any other patient's request, the total daily demand for primary care services on any given day can be modeled as a *binomial* random variable with expectation equal to Np and variance equal to $Np(1–p)$. The binomial random variable with parameters N and $0 < p < 1$ describes the random number of "successes" in N independent trials when the probability of success in any single trial is p. In

the primary care environment, this binomial random variable corresponds to the number of appointment requests that a patient panel of size N generates on a given day. (A more detailed description of the properties of the binomial random variable can be found, for example, in Bertsekas and Tsitsiklis.[11])

Using this model and the number of appointment slots each day C, we can estimate the effect of *any* panel size on the overflow frequency by calculating the probability that the demand for appointments exceeds the supply of slots on any given day, as illustrated by the formula in Figure 3 (page 216). Using this formula with $N = 2700$, $p = 0.008$, and $C = 21$, results in an estimated overflow frequency of 49.4%.

Calculation of the Daily Demand Rate for a Panel of Current Size N_{cur}

Calculating the daily demand rate for a panel of current size N_{cur}

1. Choose an observation period (for example, 18 months) and calculate the number of working days T within this period.
2. Count the number of patient visits, A, over those T days.
3. The daily demand rate for appointments (per day per patient) is

$$p = \frac{A}{N_{cur} \times T}$$, where N_{cur} is the current panel size.

Figure 2. *The calculation of the daily demand rate for a panel of current size N_{cur} is shown.*

6. Computing the Appropriate Panel Size

For a practice that operates five days a week, an overflow frequency of 49.4% implies that to avoid patient delays, the physician will need to see patients during "overtime" more than twice a week on average. It is important to note that the higher the overflow frequency, the greater the average backlog and so the longer the overtime needed to "do today's work today." For the parameters used in the previous example, the average duration of overtime when it occurs can be shown to be more than an hour. It is also important to understand that because overtime frequency is a long-term average, in any given week it could be considerably higher, leading not only to substantial overtime but long backlogs for appointments as well.

So, in our example, the current panel size would need to be reduced to be able to comfortably and consistently offer same-day appointments. This does not mean that the panel size would need to be small enough to lead to a near-zero likelihood of overflow frequency. Infrequent overflows, for example, 5%, 10%, or even 20%, are likely to be small enough that they can usually be handled with occasional and modest levels of overtime and therefore not jeopardize future appointment capacity. On the other hand, the smaller the overflow frequency, the lower will be the average daily utilization of the practice, pN/C. In selecting a target panel size and therefore a target level of overflow frequency, a physician should take into account his or her own tolerance for overtime work—5% (approximately once a month), 10% (once in two weeks), or 20% (once a week).

If the current panel size results in an overflow frequency that is too high, as in our example, a more appropriate panel size can be found by decreasing it and recalculating the overflow frequency using the formula in Figure 3. On the other hand, if the computed overflow frequency is lower than desired, the panel size should be adjusted upward. This process of adjustment and recalculation should be repeated until the overflow frequency computed for the trial value of the panel size is close enough to the desired overflow frequency.

Consider our previous example with initial panel size of 2,700 and assume that the desired overflow frequency level is 20%. Because the computed current value of the overflow frequency (49.4%) is much higher than the target, we might try a panel size of $N = 2000$. Using the Figure 3 calculation for this value of N, we obtain an overflow frequency of 8.8%, which is lower than our target. So on the next iteration, we can try a somewhat larger panel, $N = 2300$, which produces an overflow frequency of 22.8%. Because this is an estimate, this is probably sufficiently close to the target to be considered a good choice. Alternately, continuing iterations, one discovers that for the panel size of $N = 2250$ the overflow frequency comes very close to 20%.

Examples Based on Other Data

Table 1 (page 217) shows the patient panel sizes (and attained utilizations) for a "typical" general and family

Calculation of the Overflow Frequency

Calculating the overflow frequency

If the daily patient demand is modeled as a binomial random variable with parameters N (panel size) and p (demand rate), the probability that the number of patients will exceed the number of available slots C (overflow frequency) can be calculated as :

$$\textbf{Overflow frequency} = 1-(1-p)^N - \sum_{k=1}^{C} \frac{(N-k+1)(N-k+2)\times\ldots\times N}{1\times 2\times\ldots\times k} p^k (1-p)^{N-k}$$

where k is the index of summation.

This expression can also be rewritten as

$$\textbf{Overflow frequency} = 1-(1-p)^N - \frac{N}{1}p(1-p)^{N-1} - \frac{N(N-1)}{1\times 2}p^2(1-p)^{N-2}$$
$$- \frac{N(N-1)(N-2)}{1\times 2\times 3}p^3(1-p)^{N-3} - \ldots - \frac{N(N-1)(N-2)\ldots(N-C+1)}{1\times 2\times 3\times\ldots\times C}p^C(1-p)^{N-C}$$

Figure 3. *The calculation of the overflow frequency—the fraction of days when demand exceeds the average number of appointment slots available—is shown.*

practitioner (on average, 1.575 annual visits per patient,* according to Murray and Berwick[4]) and a "typical" pediatrician (on average, 1.98 annual visits per child according to the 2002 NAMCS[6]), which would result in an overflow frequency of 5% (approximately, once a month), 10% (twice a month), or 20% (once a week). The 2002 NAMCS reported that the average duration of the "face-to-face" part of the office visit is 16.1 minutes for general/family and pediatrics practices, 18.1 minutes for obstetrics/gynecology (OB/GYN) practices, and 20.0 minutes for internal medicine practices. In our calculations we considered appointment intervals of 20 minutes. Under the assumption of an 8-hour workday (for a 5-day work week this roughly corresponds to the 40.2 hours spent by a family physician on direct patient care or patient-related service during a complete week of practice[12]), this results in 24 daily appointment slots. Because the actual daily appointment capacity is likely to be somewhat lower, we also consider a daily capacity of 20 appointment slots. The calculations were performed using the formula in Figure 3 under the assumption of 210 work days per year. This value, in our estimate, is a good representation of the annual number of work days for a large number of primary care practices.

* Although the National Ambulatory Medical Care Survey Series (NAMCS) 2002 survey reports the total number of annual visits to general and family practitioners in the United States (215,466,000), the annual visit rate per patient is not easy to estimate because we could not find reliable statistics on the number of people who actually use (or even have) a primary care physician. The rate of 0.761 annual office visits per person, reported in NAMCS 2002 survey, was obtained by dividing the total number of visits to general and family practitioners by the entire size of the United States population (283,135,000), taken from 2000 U.S. Census data. Clearly, using this value would result in a gross underestimation of actual patient visit rates. The rate we use (1.575 annual visits per patient) is calculated on the basis of the assumption of 210 annual in-office days and on the assumption (used in Murray and Berwick[4]) that in an average patient panel not overly weighted with elderly and chronically ill patients, 0.07%-0.08% of patients will request a visit on an average day. We note that this estimate is somewhat higher than the 0.05% figure used by Smoller (Smoller M.: Telephone calls and appointment requests: Predictability in an unpredictable world. *HMO Practice* 6:25–29, Jun. 1992.)

Adjusting Supply for a Fixed Panel Size

Although the above analyses addressed the issue of determining panel size, the same approach can be used to determine appropriate physician capacity for a given panel size. For the case where continuity of care is considered important, capacity will be the number of daily appointment slots needed by a single physician to handle the proportion of the panel that represents his or her patients. This can be done by using the binomial formula in Figure 3 to assess the overflow that would result from each possible alternative and choosing the minimum number of slots that keeps the overflow within a "tolerable" limit. In a multiphysician setting where continuity is not considered critical, the daily capacity would be the number of slots per day for each physician multiplied by the number of physicians, and the analysis would be done using the possible alternatives as before. It is important to note that the total panel size that can be handled by a group practice in which continuity of care is not considered paramount will likely be significantly larger than the sum of the individual panel

Table 1. Panel Sizes (Capacity Utilizations) for Different Parameter Values, Primary Care Type: General and Family Practice and Pediatrics*

	General and Family Practice		Pediatrics	
	Daily Appt. Slots = 24	Daily Appt. Slots = 20	Daily Appt. Slots = 24	Daily Appt. Slots = 20
Overflow frequency = 5%	2321 (73%)	1879 (70%)	1848 (73%)	1496 (70%)
Overflow frequency= 10%	2515 (79%)	2053 (77%)	2002 (79%)	1635 (77%)
Overflow frequency = 20%	2765 (86%)	2279 (85%)	2200 (86%)	1813 (85%)

* Appt., appointment

sizes of the individual physicians in the practice if the goal is that patients see their preferred physician with high probability.

Achieving the Right Balance

The analyses provided can be easily modified for any particular physician practice. This requires that data be collected to accurately assess both supply and demand. As Murray and Berwick point out,[4] historical visit data may be misleading because they measure activity that may be less than actual demand if a practice has experienced lost and deferred demands. Therefore, it is important that demand be measured prospectively. In doing so, both weekly and seasonal patterns should be considered to identify times of particularly high (and low) levels of demand.

For example, there may be several months each year with particularly high demand because of flu season. In this case, accurate records of demand are important to estimate a seasonally adjusted patient visit rate per day, which can then be used in the binomial model to help identify capacity needs during these times. Physician supply can then be adjusted accordingly if part-time physicians are available. Vacation times and other activities should be scheduled during lower-demand seasons and days if possible to ensure sufficient capacity during higher-demand times. In addition, it is important to identify the fraction of the demand that can be managed to offset the variability in the unscheduled demand.[3] Patients who need follow-up appointments should be scheduled early in the day on lower demand days and/or during lower demand times of the year. Of course, in any given practice, there will be constraints on both physician and patient scheduling and the above guidelines are just that—goals to work toward. To the extent that they can be followed, daily delays for appointments will be reduced.

Conclusion

Ensuring timely access to medical care is an important goal for any physician practice and advanced access requires some specific guidance in achieving it. However, the variability inherent in the demand and delivery of health care makes it impossible to determine specific answers to questions about panel size or, conversely, physician practice size by using guesswork or intuition. In this article, we have described a simple probability model that can be used to supplement the qualitative approach of advanced access to make major improvements in the timeliness of care while considering the constraints on physicians' working hours. J

Linda V. Green, Ph.D., is Armand G. Erpf Professor and **Sergei Savin, Ph.D.,** is Associate Professor, Columbia Business School, Columbia University, New York City. **Mark Murray, M.D., M.P.A.,** is Principal, Mark Murray & Associates, Sacramento, California. Please address correspondence to Linda V. Green, lvg1@columbia.edu.

References

1. Strunk B.C., Cunningham, P.J.: Treading Water: Americans' Access to Needed Medical Care, 1997-2001. Washington, D.C.: Center for Studying Health System Change, Mar. 2002. http://www.hschange.com/CONTENT/421/?words (last accessed Feb. 6, 2007).
2. Institute of Medicine: Crossing the Quality Chasm: A New Health System for the 21st Century. Washington, D.C.: National Academy Press, 2001.
3. Murray M., Tantau C.: Same-day appointments: Exploding the access paradigm. Fam Pract Manag 7:45-50, Sep. 2000.
4. Murray M., Berwick D.M.: Advance access: Reducing waiting and delays in primary care. JAMA 289:1035-1040, Feb. 26, 2003.
5. Hall R.: Queueing Methods for Services and Manufacturing. Englewood Cliffs, N.J.: Prentice Hall, 1991.
6. Hing E., Cherry D.K., Woodwell D.A.: National Ambulatory Medical Care Survey: 2002 Summary. Advance Data from Vital and Health Statistics; No. 346. Hyattsville, MD: National Center for Health Statistics, 2004.
7. Christakis D.A., et al.: The association between greater continuity of care and timely measles-mumps-rubella vaccination. Am J Public Health 90:962-965, Jun. 2000.
8. Becker M.H., Drachman R.H., Kirscht J.P.: Continuity of pediatrician: New support for an old shibboleth. J Pediatr 84:599-605, Apr. 1974.
9. Gill J.M., Mainous A.G.: The role of provider continuity in preventing hospitalizations. Arch Fam Med 7:352-357, Jul.–Aug. 1998.
10. Gill J.M., Mainous A.G. III, Nsereko M.: The effect of continuity of care on emergency department use. Arch Fam Med 9:333-338, Apr. 2000.
11. Bertsekas D.P., Tsitsiklis J.N.: Introduction to Probability. Boston: Athena Scientific Publishing, 2002.
12. American Academy of Family Physicians: About Us: Table 14. Average number of patient contact hours per week by family physicians, May 2005. http://www.aafp.org/online/en/home/aboutus/specialty/facts/14.html (last accessed Feb. 6, 2007).