Proceedings of the 2008 Winter Simulation Conference
S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, J. W. Fowler eds.

A SIMULATION STUDY OF INTERVENTIONS TO REDUCE APPOINTMENT LEAD-TIME AND PATIENT NO-SHOW RATE

Ronald E. Giachetti

Dept. of Industrial & Systems Engineering
10555 W. Flagler Street
Miami, FL 33174, U.S.A.

ABSTRACT

A problem in health care is the lengthy waiting time for patients to receive an appointment. Long appointment delays cause patient dissatisfaction with the health care clinic and also has clinical ramifications. Long appointment delays are also found to increase patient no-shows, which further wastes medical resources and leads to a decrease in clinical care. A model of the health care clinic is built to understand the casual relationships in the system contributing to the problem. The model is used to investigate two possible policies. A policy of eliminating multiple appointment types can be effective in reducing appointment delay and as a consequent no-shows. Using data from several clinics, our study also suggests that an effective policy is to segregate habitual no-show patients and double-book them whenever they make appointments. This policy is equally effective as general overbooking without penalizing the entire patient population.

1 INTRODUCTION

Many clinics suffer from two related problems. First, is the long waiting time, which we will call appointment delay, until the next available appointment. Appointment delay is part of what the Institute of Medicine calls timeliness, and was identified as a primary area needing improvement (Institute_of_Medicine 1996). Second, is the high incident of no-shows in many clinics. No-shows are patients who do not appear for their scheduled appointment, thus wasting resources of the clinic by denying the opportunity of using that appointment slot for other patients.

To deal with these problems clinics try various interventions. In some cases the interventions work, in other cases they fail. What is lacking in the literature is a full and clear understanding of the relationships between patient demand, patient behavior, clinic capacity, and clinical policies. Many of the components have been studied individually, but unless models are built of the entire system practitioners will fail to see how their decisions affect other system aspects. Moreover, most analytical and simulation models treat patient behavior as unaffected by clinical policies. The lack of representing human behavior in response to system policy decisions misses significant feedback influencing overall performance.

In this paper, system dynamics simulation is used to model the feedback between clinic interventions and patient demand behavior. We draw upon known behaviors in queueing theory as well as what has been determined from empirical studies of clinical practices. We develop a model that relates patient demand, patient behavior, clinic capacity, and clinic policies to understand how clinical interventions can influence patient behavior. The resulting model is used to evaluate various interventions under some basic assumptions of clinic operations. An analysis of the policies is presented, which is followed by conclusions and suggestions for future research.

2 LITERATURE REVIEW

In the UK and Europe the waiting list for patients to see health care providers is a political and social issue that has received wide attention from the research community. Clearly, the waiting list is a queue, and some studies have analyzed it from this perspective (Worthington 1987). Wolstenholme (1993) uses system dynamics to show how the waiting list grows due to the implementation of a national policy change. Coyle (1984) creates a model to show how admissions policies and treatment durations can cause the waiting list to grow by recycling patients. Van Ackere and Smith (1999) analyze policies and their effect on the waiting list at the national level in the UK; González-Busto and García (1999) do a similar study for Spain. These studies underline the need to consider feedback between policy decisions and their affect on system performance, in this case the length of the waiting list.

In the US, where there is no national health care system, there are far fewer studies of waiting for appointments, hereafter called appointment delay. There is more

978-1-4244-2708-6/08/$25.00 ©2008 IEEE

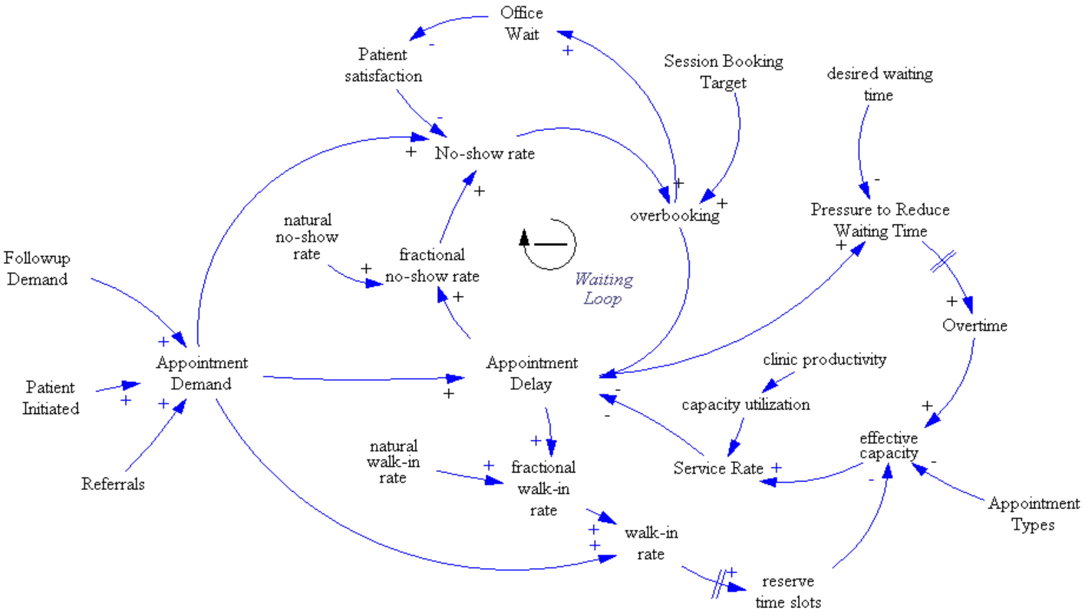

Figure 1: Casual Loop Diagram

emphasis on a related problem of patient no-shows. The reason US researchers concentrate on no-shows is probably due to its effects on clinic profits. There are many factors contributing to patient no-show behavior, the literature has identified demographic factors, income factors, and appointment delay (Murdock, Rodgers et al. 2002). To mitigate the adverse affects no-shows have on clinic operations and revenue, most clinics overbook the schedule. Kim and Giachetti (2006) study overbooking policies, and find most clinics practice what they term "naïve overbooking", which is when clinics simply book extra appointments based on expected number of no-shows. Kim and Giachetti (2006) use the data from several clinics and propose a stochastic overbooking model that considers no-show and walk-in rates by day and physician. One observation of using overbooking in healthcare, is that the penalty the clinic pays when too many patients show up is minimal. This can be contrasted to the use of overbooking in transportation for which there is a significant penalty, such as free flight coupons to passengers who cannot board a flight (Weatherford and Bodily 1992). For this reason, it can be questioned whether overbooking applied in health care will lead to excessive overbooking that deteriorates patient care because the penalty is financially inconsequential.

3 MODEL DEVELOPMENT

In Figure 1 is a casual loop diagram showing the relationships between patient demand, patient behavior, clinic capacity, and clinic policies. The two problems of appointment delay and no-shows are related; appointment delay has been show to affect no-show rates (Galluci, Swartz et al. 2005). Galluci et al. (2005) did an extensive study with five years of data and found a linear relationships between appointment delay and no-show rates in a mental health clinic. Kopach, DeLaurentis et al. (2007) interviewed clinical experts who indicated that appointment delay was significant factor. However, the authors did not have data on appointment delay so they made some reasonable assumptions about patient behavior and used an exponential curve relating appointment delay to no-show rate. In the clinic we studied, the no-show rate was correlated to appointment delay. These studies establish the correlation between no-show behavior and appointment delay; the open question is the nature of the relationship, which may well depend on additional factors such as medical specialty or patient demographics.

A typical response to no-shows is to overbook the schedule (Kim and Giachetti 2006). Overbooking increases the number of appointments made in each session, which in turn decreases the appointment delay. However, since the show behavior of individual patients is impossible to predict, the average waiting time in the office increases with overbooking. Waiting time is strongly correlated with patient satisfaction. In a patient satisfaction survey of patients in a Miami clinic the categories patients consistently rated as being most dissatisfied with concerned various types of waiting. We have identified two feedback loops from overbooking to appointment delay. Overbooking directly reduces appointment delay by booking more appointments in each session, but overbooking indirectly increases the no-show rate due to patient dissatisfaction.

Another effect of appointment delay is the propensity of patients to circumvent the scheduling system by walking in without an appointment (Bibi, Cohen et al. 2007). When walk-ins are allowed then many clinics respond to reserving time slots for walk-ins (Kim and Giachetti 2006), which reduces the availability capacity for scheduled appointments, and, consequently, increases the appointment delay for those patients who make appointments. This is especially problematic when the walk-in patient also has a scheduled appointment that they fail to cancel; these appointments are called phantom appointments.

On the capacity side, the clinic management will respond to long appointment delays by increasing capacity. Capacity increases are usually a short-term fix achieved by working overtime to reduce the appointment backlog. Hopp and Spearman (2000) call this the "vicious overtime cycle". Overtime is useful for addressing demand surges but as a long-term strategy, overtime cannot help if the demand is close to or exceeds capacity.

In the model are three demand streams of patient initiated, referrals, and follow-up demand. The reason for segregating demand is that each demand stream can be influenced separately. Additionally, the model allows for walk-ins. In the face of walk-ins we model a delay after which the clinic will set aside appointment slots in anticipation of walk-ins.

3.1 Simulation Model

The casual loop diagram lets us understand the relationships between variables in the model. This diagram was used to build the system dynamics simulation model. The simulation model was built in Venisim 5.0. A portion of the simulation model is depicted in Figure 2. The main state variable is the appointment backlog that is filled by the demand for appointments and is emptied by the service rate. The service rate includes overbooking of appointments. The realized service rate is the service rate minus the no-show rate.

The no-show rate is modeled as an exponential function of appointment delay. We found this to be a reasonable fit for one year of data obtained from the dermatology clinic (R^2 = 0.78). In the second clinic we had fewer data points for different appointment delays, and the model fit was less satisfactory.

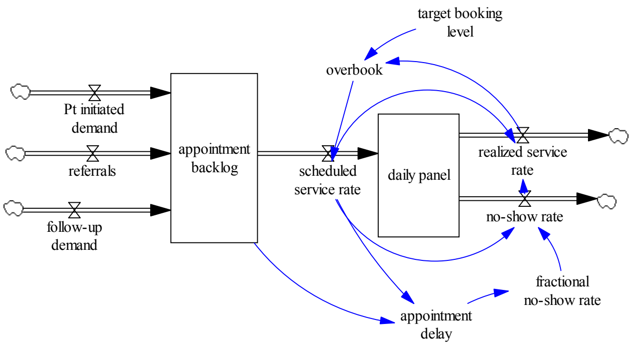

Figure 2: Partial simulation model

3.2 Verification and validation

The model was verified and validated according to the advice provided in (Sterman 2000; Sargent 2004; Goerger, McGinnis et al. 2005). Specifically, the model was verified by tracing through the logic and examining the correctness of all the underlying equations. To validate the model we tested the model assumptions, we ensured face validity in each iteration of model development, we tested the model behavior under different experimental conditions, and we validated all input data. For example, we tested boundary conditions such as what happens when demand is less than service rate? To prevent the appointment backlog from becoming negative in this situation we had to revise the equations governing the service rate such that the service rate is the max of demand rate or the capacity.

We calibrate the model to two clinics. The first clinic is a dermatology out-patient clinic operated by the public hospital in Miami. For this clinic we collected data on patient demand, no-show rates, and service rates. The second clinic is a general practioner clinic run by a health maintenance organization (HMO) for Medicare/Medicaid patients. In this clinic we had patient demand, no-show rates, and daily booking levels. In both cases we were able to use the model and calibrate the equations so as to replication observed system behavior. This shows the model has sufficient fidelity to model actual clinics.

3.3 Model Evaluation

The presence of persistent appointment delays can be explained by the following hypotheses:

1. demand is greater than capacity
2. high variation in demand and/or capacity

A common belief held in health care is that demand is greater than capacity. According to queueing theory, when demand is greater than capacity the appointment delay would grow to infinity. Since, in most clinics the appointment delay is a constant value then either the demand is less than capacity, or demand is somehow reduced by other means. We hypothesize several dynamic system behaviors that hold this in check. No-show rate and balking rate are functions of the appointment delay. As appointment delay increases the no-show rate and balking rate increase, which effectively reduces the patient demand. Figure 3 shows the system dynamic model in which the traffic density, defined as the arrival rate divided by the service rate is p = 1.1 and p = 1.7. The system reaches a steady-state level such that the no-show rate and balking balances the patient demand.

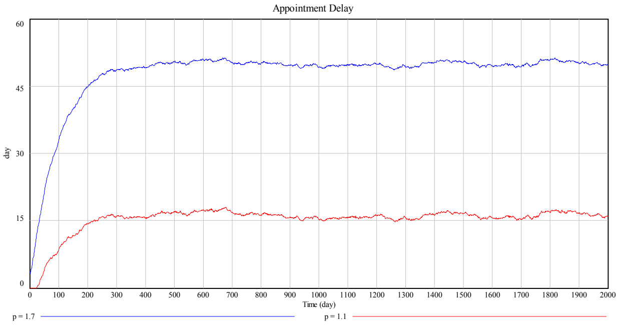

Figure 3: Appointment delay

Although the system reaches steady-state there are two implications. First, of course, is that a large part of the patient population is not being well-served as evidenced by the no-shows. Second, is the steady-state for many clinics is reached at a undesired high-level for the appointment delay, and the realized number of patients seen is less than the clinic capacity. A result of this is many clinics will overbook the schedule.

3.4 Policy Interventions

In this section we evaluate several policy interventions that can be tried to reduce appointment delay.

1. reduce the number of appointment types.
2. individual double-booking
3. Single queue in a region for multiple providers

Each policy intervention is modeled in the base model.

3.4.1 Reduce the Number of Appointment Types

Many clinics specify multiple appointment types. When patients request an appointment, the receptionist determines the appointment type and schedules the appointment in the slots allocated only for that appointment type. Murray et al. (2003) recommend reducing the number of different appointment types. This is an application of the pooling principle identified in queueing theory, which has been shown to reduce the average delay in other application areas such as call centers (Mandelbaum and Reiman). We use the data from a dermatology clinic that defined 21 separate appointment types shown in Table 1. Each appointment type was scheduled only on certain days and times of the week, effectively limiting capacity for each appointment type.

Table 1: Appointment types and their availability by day of the week and session (morning or afternoon)

	M	**T**	**W**	**R**	**F**
AM	3,9,12	3,5,10,12, (21)	3,9,12, (21)	3,12	3,5,9, 12
PM	1,3,4,11 12,14,15,16,17, 19,20	1,3,4,11, 12,14,15, 16,17,19, 20	1,3,4,11, 12,14,15, 16,17,19, 20	1,3,4,11, 12,14,15, 16,17,19, 20	2,3,4, 8,12, 13,17,18

To model the number of different appointment types we create a separate demand stream for each appointment type with a different random number seed. Then capacity is allocated to each demand stream with a separate stock for appointment delay. The results indicate, as expected, that there is unused capacity because when capacity is allocated to a particular demand type, then it is left unavailable for other demand types. Comparing performance for the multiple appointment types versus a single appointment type shows improvement in appointment delay from 27 days to 8 days on average.

3.4.2 Individual Double-booking

Overbooking is a population-based policy, in that it overbooks regardless of which patients make appointments on a given day. If a sizeable proportion of the no-shows can be identified as being caused by habitual offenders, then a policy could be to double-book only those patients. This changes overbooking from a population-based policy to an individual-based policy. We mined the data for a single clinic that had three physicians over a one-year duration and found the results in Figure 4 indicating that the majority of patients always show for their appointments. The simulation model was modified to emulate a policy of double-booking those patients who missed two or more appointments during the year. To do this we two separate patient populations, one for regular patients and one for habitual no-show patients. The policy has little affect on the actual appointment delay. The policy did reduce the in-office waiting time of the patients. It is hypothesized that less in-office waiting time would increase patient satisfaction and reduce the no-show rate, but we do not have sufficient data between these variables to infer the magnitude of the reduction in the no-show rate.

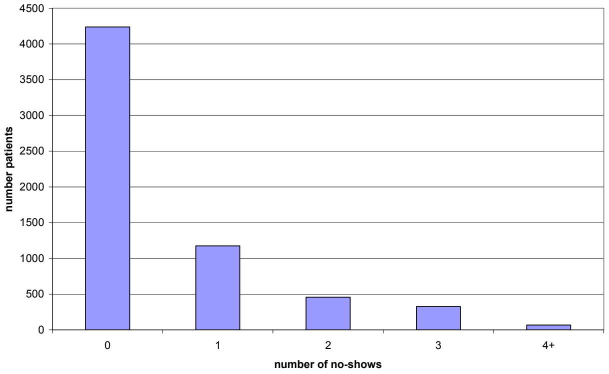

Figure 4: Distribution of show behavior for patients

3.4.3 Single queue in a region for multiple providers

This strategy is another application of the pooling principle. In theory a single queue will have shorter expected waiting time than having multiple queues. To evaluate this strategy we consider the scenario when there are three physicians in a region. These were modeled as three parallel servers all served by a single queue. When demand is such that p = 1.1 the expected waiting time decreases from 14.69 days to 6.15 days. When demand is such that p = 1.1 the expected waiting time decreases from 50.11 days to 19.74 days. These savings are as expected when pooling multiple physicians. The problem for implementation is twofold. First, how to coordinate the queues for multiple physicians and overcome barriers erected due to different insurance plans, fee schedules, and other barriers. Second, this strategy diminishes patient continuity-of-care, which is considered important to health outcomes by the medical community.

4 CONCLUSION

The purpose of the simulation models were to understand quantitatively the performance of a health care clinic and to investigate various interventions that clinics can use to reduce appointment delay and no-show behavior. As Dangerfield (1999) remarks, the value of the system dynamics casual diagrams were once overlooked, but serve a role in describing circular cause-effect relationships and how they influence system behavior. In the case of the health care clinic we identify the feedback between appointment delay and no-show behavior as contributing to overall clinic performance. The casual loop diagram was used to build the simulation model. The model was calibrated using data from an existing clinic in which the model was able to replicate the appointment delay experienced by the clinic. This demonstrated the model usefulness to evaluate the policies under actual clinic scenarios.

The first policy reviewed was to reduce the number of different appointment types. We suggested this is an application of the pooling principle, and reduces the variability in capacity and eliminates any wasted appointment slots that are reserved for a particular appointment type that goes unused. Dramatic improvements are possible by using this policy.

The second policy evaluated was overbooking. Many clinics react to no-show behavior by overbooking. Overbooking is a population-based policy that in a way penalizes patients who show-up on time to their appointments since overbooking can increase office waiting times. We find this less than a satisfactory policy because it is applied to the entire patient population and therefore penalizes good patient behavior as well as bad patient behavior. Instead we recommend a policy of segregating the habitual no-show patients and double-booking them only. The policy change has minimal affect on appointment delay, but it results in less in-office waiting time for the majority of patients, which should improve patient satisfaction because the waiting time is consistently listed as a problem in patient surveys.

The results show that the two policies can work, but in the face of a severe mismatch between capacity and demand no policy can completely eliminate the appointment delay. Reducing the number of appointment types reduces appointment delay by increasing availability capacity and reducing variability in capacity. Use of double-booking improves in-office waiting time, and consequently patient satisfaction.

Future work should investigate the impact of what economists call supplier-induced demand. There is a significant body of evidence that the more health care services available then the greater the demand. Consequently, policies to increase supply should consider the impact on demand, which cannot be modeled independently of supply. Additionally, when insurance pays for the service, but the patient or physician determines whether to consume the service, then the model should include how these actors interact to influence demand. Inclusion of the phenomenon will add to our understanding of how capacity decisions will affect demand and system performance.

REFERENCES

Bibi, Y., A. D. Cohen, et al. 2007. Intervention program to reduce waiting time of a dermatological visit: Managed overbooking and service centralization as effective management tools. *International Journal of Dermatology* 46: 830-834.

Coyle, R. G. 1984. A systems approach to the management of a hospital for short-term patients. *Socio-Economic Planning Science* 18: 219-226.

Dangerfield, B. C. 1999. System dynamics applications to European health care issues. *Journal of the Operational Research Society* 50: 345-353.

Galluci, G., W. Swartz, et al. 2005. Impact of the wait for an initial appointment on the rate of kept appointments at a mental health center. *Psychiatric Services* 56(3): 344-347.

Goerger, S. R., M. L. McGinnis, et al. 2005. A validation methodology for human behavior representation models. *JDMS* 2(1): 5-17.

González-Busto, B. and R. García 1999. Waiting lists in Spanish public hospitals: A system dynamics approach. *System Dynamics Review* 15: 201-224.

Hopp, W. and M. Spearman. 2000. *Factory Physics*, McGraw-Hill.

Institute_of_Medicine. 1996. *Primary Care: America's Health in a New Era*. Washington DC, Institute of Medicine, National Academy of Sciences.

Kim, S. and R. Giachetti. 2006. A stochastic mathematical appointment overbooking model for healthcare providers to improve profits. *IEEE Transactions on Systems, Man, and Cybernetics -- Part A: Systems and Humans* 36(6): 1211-1220.

Kopach, R., P. DeLaurentis, et al. 2007. Effects of clinical characteristics on successful open access scheduling. *Health Care Management Science* 10: 111-124.

Mandelbaum, A. and M. I. Reiman On pooling in queueing networks. *Management Science*.

Murdock, A., C. Rodgers, et al. 2002. Why do patients not keep their appointments? Prospective study in a gastroenterology outpatient clinic. *Journal of the Royal Socity of Medicine* 95: 284-286.

Murray, M., T. Bodenheimer, et al. 2003. Improving timely access to primary care. *Journal of the American Medical Association* 289(8): 1042-1046.

Sargent, R. G. 2004. Validation and verification of simulation models. *Proceedings of the 2004 Winter Simulation Conference*. Washington DC.

Sterman, J. 2000. *Business Systems Dynamics*.

van Ackere, A. and P. C. Smith. 1999. Towards a macro model of National Health Service waiting lists. *System Dynamics Review* 15: 225-252.

Weatherford, L. R. and S. E. Bodily. 1992. A taxonomy and research overview of perishable-asset revenue management: yield management, overbooking and pricing. *Operations Research* 40(5): 831-844.

Wolstenholme, E. F. 1993. A case study in community care using systems thinking. *Journal of Operational Research Society* 44: 925-934.

Worthington, D. J. 1987. Queueing models for hospital waiting lists. *Journal of the Operational Research Society* 38: 413-422.

AUTHOR BIOGRAPHY

RONALD GIACHETTI is an Associate Professor in the Department of Industrial & Systems Engineering at Florida International University in Miami, FL. His research interest are in operations research, enterprise systems, and health care. His email is <giachetr@fiu.edu> and his webpage is <http://web.eng.fiu.edu/ronald/> where additional information can be found.