# Queues with Scheduled Arrivals in Health Care
# Literature Review and Research Plan

John Gilbertson

March 8, 2016

## 1 Literature Review

Health care providers are under a great deal of pressure to improve service quality and efficiency (Goldsmith 1989 [**Goldsmith**]). There is a large body of literature studying the potential of appointment systems to reduce patient waiting times, and waiting room congestion. Fomundam and Herrmann (2007) [**Fomundam**], and Cayirli and Veral (2009) [**Cayirli**] provide comprehensive surveys of research on appointment scheduling. There is a fundamental trade-off in appointment policies. If patients are scheduled to arrive close together, they experience long waiting times. However, if appointment times are spread further apart, the doctor's idle time increases.

Most of the papers on scheduled arrivals in health care can be classed into two categories. Those that design algorithms to find good schedules, and those that evaluate schedules using simulation. While simulation studies can easily model complicated patient flows, queuing models often provide more generic results than simulation (Green 2006 [**Green**]).

The foundation paper on modeling queues with scheduled arrivals is Bailey (1952) [**Bailey**]. Bailey proposes that customers' waiting times can be reduced without a significant increase in doctor's idle time. The Bailey rule, which is commonly referenced in literature, is that patients should be scheduled to arrive at fixed intervals with two patients scheduled to arrive at the start of service. Bailey found that a great deal of time wasted by patients could be reduced without a signficant increase in the doctor's idle time. Under the Bailey rule, patients with late appointments will wait longer than those with early appointments. This lack of uniformity might be undesirable due to issues of fairness.

Pegden and Rosenshine (1990) [**Pegden**] extend on Bailey's paper. They present an algorithm to iteratively determine the optimal arrival times for $n$ patients that need to be scheduled. The optimal arrival times are those that minimise a weighted sum of the expected patients' waiting time and the expected doctor's idle time. Pegden and Rosenshine prove that their objective function is convex for $n \leq 4$, thus their algorithm finds the optimal schedule. While they conjecture that the objective function is convex for $n \geq 5$, it hasn't been proven.

Stein and Côte (1994) [**Stein**] apply Pegden and Roshenshine's model to obtain numerical results for situations with more than three patients. The optimal times between successive patients become near constant as $n$ grows. This is the often observed dome-shape. Optimal appointment intervals exhibit a common pattern where they initially increase towards the middle of a session, and then decrease. Stein and Côte simplify the model by requiring the intervals between arriving patients to be held constant. This realistic restriction (used commonly in the literature) makes the model more easily applicable in practice without significant altering the results.

Stein and Côte apply queuing theory results to solve the model for the optimal arrival interval assuming the queue reaches its steady state distribution. This assumption greatly reduces the computation required. However, in practice, it is common to find services that never reach steady state. Babes and Sarma (1991) [**Babes**] attempted to apply steady state queuing theory, but found their results tended to be very different from those observed in real operation.

These key papers by Bailey, Pegden and Rosenshine, and Stein and Côte provide the basis for a more realistic exploration of health care systems. DeLaurentis et al. (2006) [**Delaurentis**] point out that patient no-shows can lead to a waste of resources. Mendel (2006) [**Mendel**] incorporates the probability of a patient

not showing up into the model presented by Pegden and Rosenshine. Unsurprisingly, no-shows lead to lower expected waiting times for patients who do show up.

The presence of walk-ins (regular and emergency) can disrupt a schedule. Gupta et al. (1971) [**Gupta**] propose a system where non-routine requests are superimposed on top of routine scheduled requests. Fiem et al. (2007) [**Fiems**] investigate the effect of emergency requests on the waiting times of scheduled patients. Fiem models a system with deterministic service times and discrete time. Despite this research, Cayirli and Veral suggest that walk-ins are neglected in most studies. Further research could investigate their effect on optimal arrival times.

Mondschein and Weintraub (2002) [**Mondschein**] observe that the majority of the literature assumes that demand is exogenous and independent of patients' waiting times. These papers assume the total number of patients $n$ is fixed and independent of waiting times. The vast majority of servers are now private (including medical servers), so face competitive environments. Mondschein and Weintraub thus present a model where demand depends on the patients' expected waiting time.

Simulation is a useful tool to analyse the effectiveness of appointment policies. Kao and Tung (1981) [**Kao**] use simulation to compliment their results obtained from queuing theory. Ho and Lau (1992) [**Ho**] study the performance of eight different appointment rules under different scenarios. They find that no rule will perform well under all circumstances.

Case studies can test the real world applications of an appointment system. While they lack generalisation, they are necessary to compliment the theoretical research. Rockart and Hofmann (1969) [**Rockart**] show individual block systems lead to more punctual doctors and patients, and less no-shows. Walter (1973) [**Walter**] indicates that the simple grouping of inpatients and outpatients results in a substantial improvement in doctor's idle time.

Unfortunately, Cayirli and Veral lament that despite much published work, the impact of appointment systems on outpatient clinics has been limited. Doctors are often unwilling to change their old habits. O'Keefe (1985) [**O'Keefe**] had their proposed appointment system of classifying patients rejected by staff. Huarng and Lee (1996) [**Huarng**] were unable to implement their system due to staff resistance. Bennett and Worthington (1998) [**Bennett**] found their recommendations weren't implemented successfully. Future research must attempt to develop models that will be accepted and implemented in real health care services.

# 2 Research Plan

The first part of my research should be to replicate the results given in some of the key papers. Pegden and Rosenshine, and Stein and Côte give clear explanations of their algorithms. Their results should be easily replicatable using the Newton-Raphson method (in Matlab) to numerically minimise the objective functions.

Afterwards, it would be good to use a simple simulation set-up to investigate the quality of their solutions in comparison to well known heuristics (e.g., the Bailey rule). While patients expected waiting times might be at a minimum, it would be interesting to investigate the variability of waiting times. From the point of view of fairness, comparing the distribution of waiting times for patients with appointments at different times would be important.

From here, it would be interesting to explore some 'what if' questions. What happens to waiting times / idle times if the estimate of the mean service time is slightly off? What happens to a system if the scheduler has accidentally double booked two patients at the same appointment time? What is the effect of patient earliness / companions on congestion in the waiting room? What happens if the doctor arrives late at the start of service? There are many questions that could be explored here.

The next step would be to extend the model proposed by Pegden and Rosenshine. Mendal has already incorporated no shows into the model, thus that should be a simple extension. In addition, including a fairness measure into the objective function would be important. These studies assume a linear relationship between waiting cost and waiting time, which is obviously not true in practice. Another extension would be to adjust the objective function to investigate a quadratic relationship (for example).

One could give the scheduler flexibility in the number of patients $n$ required to be serviced. By running the numerical optimisation several times it would be possible to find the $n$ that minimises the objective function, which is hopefully not as small an $n$ as possible.

It would be very desirable to include walk-ins and unpunctual patients in the model. Moreover, the system should incorporate the reneging behaviour of walk-ins who forego the service if the queue is too long.

Another approach would be to model the queue as a Markov Decision Process. At each state, there are $i$ patients currently waiting in the queue and $N$ patients still to be scheduled. The action $a$ is to decide when to schedule the next patient. Assume exponential service times, one doctor and punctual patients who always show up. The aim is to find the $a^*$ that minimises the total cost (i.e., cost of expected patients' waiting time and doctor's idle time). Can possibly solve using Belmann's equation.