# Resource Allocation Using Queuing Theory In A Walk-In Clinic

Coleen R. Wilder, Valparaiso University, USA

**ABSTRACT**

*The objective of this research is to produce a simple tool to assist health care management in quantifying the tradeoffs between different resource allocations. In many cases, intuition results in an appropriate selection; the quandary, however, is typically over the magnitude of improvement. The problem addressed herein uses queuing theory in the context of a hypothetical walk-in clinic. Different resource allocations are compared on the basis of the expected number of patients in the waiting room. Without comparative numbers, managers are forced to guesstimate the difference in expected queue lengths. Fact-based decisions not only improve quality but give the decision maker a sense of comfort.*

**Keywords:** Queuing Theory; Resource Allocation; Decision Analysis

## INTRODUCTION

Managing a walk-in service is a challenge for any industry and for the healthcare industry, it is an even greater challenge when contagious patients are considered. Healthcare management and patients alike are concerned about exposure to diseases, other than that for which they are currently afflicted while waiting to see a doctor. Certain infections are transmitted by airborne means when an infected individual coughs, sneezes, or simply breathes, which raises the issue of how to best manage the flow of patients through a shared air space – the waiting room.

The extent to which airborne transmission is possible is controversial and the subject of much research. In a four-year study in Oklahoma (Istre, et. al 1987), it was found that 45% of the confirmed measles cases were the direct or indirect result of exposures in the waiting room. In a more recent article (Beggs, et. al 2010), researchers used a Monte Carlo simulation to examine the impact of exposure in waiting areas for a selected set of diseases; namely, tuberculosis, measles, and influenza. Their conclusions showed the greatest threat of infection from exposure in a waiting area was for measles, followed by influenza, then tuberculosis; although the latter was considered negligible. Not surprisingly, it was also found that increasing the exposure time also increased the probability of infection. Whether the threat of exposure is real or not, some patients believe it to be substantial, thereby increasing their stress levels. For this reason alone, it is a worthwhile effort to compare the performance of different strategies affecting patient flow through a common waiting area.

The objective of this research is to construct a simple tool to assist management in assessing the tradeoffs between different policies used to manage patient flow; a tool as simple as a set of tables. These tables are to be used to compare the expected number ( $L_q$) of patients in a queue (waiting room) for each selected policy. Intuition frequently directs us to an appropriate option; the quandary, however, is over the magnitude of improvement. In addition, few decisions are based on numbers alone but are used with qualitative factors to form a convincing case. For example, single queues with multiple servers are more efficient than a separate queue for each service provider. Yet, separate queues are still in use to give people an illusion of control by choosing their own line, a preference supported in a recent Wall Street Journal article (Smith, 2011). Would management be willing to sacrifice efficiency to appeal to patient preferences if the level of improvement was greater by a factor of ten? Or, would management only be willing to sacrifice efficiency if the level of improvement was fractional?

## MODEL

To answer the aforementioned questions, a hypothetical situation is examined in which the management of a walk-in clinic must evaluate the best use of resources during peak flu season. Although the situation is hypothetical, it is a common dilemma faced by many walk-in service providers and is appropriate for any communicable disease. The clinic manager must decide how to assign patients to doctors in an effort to decrease the expected number of patients in a waiting room while giving attention to other priorities, such as doctor workloads. The problem is also expanded to incorporate different productivity rates for each doctor; this feature is reasonable under a variety of scenarios. For example, one possible scenario may use a temporary doctor that would not be familiar with the clinic's policies or logistics and would require more time per patient than the in-house doctor would require treating the same patient.

To simplify the analysis, each proposed queuing option is examined using two classes of patients and two doctors. This limitation is for illustrative purposes and may be relaxed with minimum effort, but is a reasonable design for a small clinic. Three common queuing structures are then compared on the basis of their respective performance; namely, the expected number ($L_q$) of patients in the waiting room. Patients are separated into the two classes based on their sensitivity to influenza. The first group, referred to as the at-risk group, includes the elderly, young, and people with a compromised immune system. Variables that refer to this group are denoted with a subscript of one. The second group, referred to as the general population, includes all patients not belonging to the first group, and associated variables are subscripted with the number two. Let $\alpha_1$ represent the percentage of patients belonging to the at-risk group; the percentage belonging to the general population then is $\alpha_2 = 1 - \alpha_1$. Values for patient arrivals ($\lambda$) and doctor service rates ($\mu$) may be obtained from estimates or historical data; each doctor's workload, is defined as $\rho_1 = \lambda_1/\mu_1$ and $\rho_2 = \lambda_2/\mu_2$.

Three different options are evaluated:

### Shared Doctor Option

The first option is very typical for walk-in facilities and is used as a benchmark. The usual practice is to ask patients to sign in as they arrive. Patients are then treated in the order in which they arrived by the next available doctor. This method is considered fair by patients and staff alike.

### Dedicated Doctor Option

This option assigns patients to an exclusive doctor depending on their class. At-risk patients have a dedicated doctor which gives them the feeling of specialized care. It should be noted, however, that both groups of patients still share the same waiting room.

### Balanced Doctor Option

This option is an adaptation of the Dedicated Doctor Option. Each patient is assigned a doctor with the exception that the doctor, with the lightest workload may treat a pre-determined portion of patients from the other group, in an effort to balance the workload between the two doctors. The methodology used to determine the percentage of patients allowed to switch doctors is discussed in the Appendix. Switching doctors is, in effect, a managed effort and not an elective effort by the patient. This method is considered fair by the doctors, but patients wanting specialized care may not be happy when asked to switch doctors.

The dependent variable of interest is the expected number of patients ($L_q$) waiting to see a doctor. Each of the aforementioned options requires its own set of calculations to produce their respective values. The Shared Doctor option cannot be calculated using closed form equations. Using standard Kendall notation, this structure is an example of an $M/H_2/2$ system with finite capacity (k); arrivals are exponentially distributed, service times are the combination of two exponential distributions which form a hyper exponential distribution, and two servers (doctors) are available. Standard procedure for solving these systems is to use linear algebra to solve a system of equilibrium equations; this methodology was employed herein but requires two explanations. When a patient arrives and both

 

doctors are available, a dilemma is posed as to which doctor should see the patient. Saaty (1961) suggests several options in this situation. His work is based on using a probability of $\frac{\mu_1}{\mu_1+\mu_2}$ for the first facility and $\frac{\mu_2}{\mu_1+\mu_2}$ for the second which results in the fastest server chosen more often. The practice employed in this research is to assign an equal probability (0.5) to each doctor to support the idea of fairness. The other variable requiring explanation is the capacity (k). The methodology used here is to loop through incremental values of k until the probability of a full waiting room is less than 0.001 which for all practical purposes is zero. The expected number of patients in the waiting room is then derived by using the state probabilities ($P_n$) as follows:

$$L_q = \sum_{n=3}^{k} (n-2)P_n$$

The Dedicated Doctor option is the simplest queuing system to construct; each is defined as an example of an M/M/1 system. Due to their basic nature, the details are not discussed, but it should be noted that the expected time in the waiting room is a weighted average of the two groups while the expected number of patients waiting is a direct sum of the two groups.

The Balanced Doctor option is of particular interest primarily due to the method employed to redistribute the workload. The dependent variables are derived as for the previous options with the exception that the arrival rates for each queue are adjusted (adjustment is described in the Appendix). The adjusted rates are then used in place of the original arrival rates to construct equilibrium equations similar to the techniques used in the Dedicated option.

**RESULTS**

For illustration purposes, an example has been created where 70% of the total patient population belongs to the at-risk group. At-risk patients' arrival rate is estimated at $\lambda_1 = 0.96$ (patients per unit of time). The doctor treating at-risk patients has a service rate of $\mu_1 = 1$ (patient per unit of time). The doctor treating at-risk patients, therefore, has a workload of $\rho_1 = \lambda_1/\mu_1 = 0.96$. General patient's arrival rate is estimated at $\lambda_2 = 0.427$ and the respective service rate is $\mu_2 = 0.667$ with a resulting workload of $\rho_2 = 0.64$. A combined workload is defined as the sum of the individual utilizations expressed as $\rho_c = \rho_1 + \rho_2$.

Table 1 shows the expected number of patients ($L_q$) waiting to see each doctor. The best performance is achieved using the Shared Doctor option, which is expected. The Balanced model is designed so the effective workloads are equal, but this occurs at a cost to the expected number of patients waiting to see a doctor. This increase in queue length is noticeable. For example, when the combined workload is 1, the expected number of patients in the waiting room increases from 0.381 to 1.127 or a 196% increase. The percentage increase in queue length steadily decreases as the combined workload increases. For example, when the combined workload is 1.8, the percentage increase drops to 122%. This relationship may justify a policy to the effect that a Balanced strategy will only be used when the combined workload ($\rho_c$) is estimated to exceed 1.5 or some other level determined by management.

The Dedicated Doctor option has the worst performance, as expected, but offers what patients may perceive as the best quality since at-risk patients are given a specialized doctor. It should be noted that the dedicated option is not feasible when either channel results in infinite queues; these cases are noted as "NS" in Table 1 for not stable. The number of patients in the waiting room increases by 567% when the combined workload is 1.6. Management must decide whether the tradeoff between specialized care is worth the increased exposure.

**Table 1: Expected Number of Patients Waiting To See Doctor**

| | Shared Doctor | Dedicated Doctor | | | Balanced Doctor | | |
|---|---|---|---|---|---|---|---|
| $\rho_c$ | Total | Total | At-risk Doctor | General Doctor | Total | At-risk Doctor | General Doctor |
| 1 | 0.381 | 1.167 | 0.900 | 0.267 | 1.127 | 0.563 | 0.563 |
| 1.1 | 0.552 | 1.627 | 1.281 | 0.346 | 1.529 | 0.764 | 0.764 |
| 1.2 | 0.789 | 2.295 | 1.851 | 0.443 | 2.071 | 1.036 | 1.036 |
| 1.3 | 1.122 | 3.329 | 2.765 | 0.563 | 2.821 | 1.410 | 1.410 |
| 1.4 | 1.613 | 5.123 | 4.410 | 0.713 | 3.897 | 1.949 | 1.949 |
| 1.5 | 2.368 | 9.000 | 8.100 | 0.900 | 5.531 | 2.766 | 2.766 |
| 1.6 | 3.627 | 24.178 | 23.040 | 1.138 | 8.241 | 4.120 | 4.120 |
| 1.7 | 6.062 | NS | NS | 1.445 | 13.473 | 6.737 | 6.737 |
| 1.8 | 12.337 | NS | NS | 1.851 | 27.378 | 13.689 | 13.689 |

Exposure to various diseases, however, may only be an issue for at-risk patients since they have a greater risk of a disease turning fatal. Table 2 shows the expected number of at-risk patients waiting to see a doctor. The columns labeled "General Doctor" indicate the at-risk patients waiting to the see the doctor assigned to general patients.

**Table 2: Expected Number of At-Risk Patients Waiting to See a Doctor**

| | Shared Doctor | Dedicated Doctor | | | Balanced Doctor | | |
|---|---|---|---|---|---|---|---|
| $\rho_c$ | Total | Total | At-risk Doctor | General Doctor | Total | At-risk Doctor | General Doctor |
| 1 | 0.264 | 0.900 | 0.900 | 0.000 | 0.690 | 0.563 | 0.126 |
| 1.1 | 0.382 | 1.281 | 1.281 | 0.000 | 0.939 | 0.764 | 0.175 |
| 1.2 | 0.546 | 1.851 | 1.851 | 0.000 | 1.270 | 1.036 | 0.234 |
| 1.3 | 0.776 | 2.765 | 2.765 | 0.000 | 1.727 | 1.410 | 0.317 |
| 1.4 | 1.116 | 4.410 | 4.410 | 0.000 | 2.393 | 1.949 | 0.444 |
| 1.5 | 1.639 | 8.100 | 8.100 | 0.000 | 3.391 | 2.766 | 0.626 |
| 1.6 | 2.510 | 23.040 | 23.040 | 0.000 | 5.047 | 4.120 | 0.927 |
| 1.7 | 4.195 | NS | NS | 0.000 | 8.270 | 6.737 | 1.533 |
| 1.8 | 8.537 | NS | NS | 0.000 | 16.787 | 13.689 | 3.098 |

The results for the at-risk patients are consistent with previous results. The expected time ($w_q$) waiting to see a doctor may be calculated using Little's Law as follows: $w_q = L_q/\lambda$.

## CONCLUSIONS

It bears repeating that the primary objective of this research is to produce a simple tool to assist healthcare management in evaluating the tradeoffs between different queuing structures; a small sample of which has been presented here. One way to minimize exposure to airborne diseases is to evaluate the expected queue lengths for various queuing options. All too often, decisions must be made in the absence of data. In these situations, managers must make assumptions about the relative relationships between different options. Without comparative numbers, managers are forced to guesstimate the difference in expected queue lengths. Fact based decisions not only improve quality but give the decision maker a sense of comfort.

A key issue concerns the choice of presentation. Tables are used herein since they are suitable for publication. If a primary objective of this research is to deliver a tool to facilitate decision making, is a table the best format? Tables are considered to be outdated, yet are still used because they require little training. A software application increases the scope of parameters that can be examined, but the algorithms do not produce answers instantaneously. At any rate, it is an issue that has yet to be addressed. Along the same lines, the efficacy of this research needs to be tested using actual data.

Priority queues were not considered in the initial study but would add considerable value. If at-risk patients are a primary concern, why not move them to the front of the line? Since the context of this research pertains to a walk-in facility, patients frequently assume that a first-come first-serve rule is in use. Standard practice, however, is a management decision that can be changed.

 

## AUTHOR INFORMATION

**Coleen R. Wilder**, Valparaiso University, College of Business. Professor Wilder received her Ph.D. in Management Science from Illinois Institute of Technology (2010). She also has an MBA from the University of Chicago (1995) with concentrations in Finance and Operations Management. Her undergraduate degree is in Mathematics Education from Indiana University (1978). Ms. Wilder also worked in manufacturing for 20 years and real estate for 5 years. E-mail: Coleen.Wilder@valpo.edu

## REFERENCES

1. Beggs C B, Shepherd S J and Kerr K G (2010). Potential for airborne transmission of infection in the waiting area of healthcare premises: stochastic analysis using a Monte Carlo model. *BMC Infections Diseases*, 10(1) 247-254.
2. Gross D, Shortle J F, Thompson J M and Harris C M (2008). *Fundamentals of Queuing Theory Fourth Edition*. John Wiley & Sons, Inc: Hoboken, New Jersey.
3. Istre G R, McKee P A, West G R, O'Mara D J, Rettig P J, Stuemky J and Dwyer D M (1987). Measles Spread in Medical Settings: An Important Focus of Disease Transmission? *Pediatrics*, 79(3) 356-358.
4. Saaty T L (1961). *Elements of Queuing Theory With Applications*. Dover Publications, Inc: New York.
5. Smith Ray A (2011). Find the Best Checkout Line. *Wall Street Journal*, Dec. 8, 2011: D.1.
6. Wilder, Coleen R. (2010). The Queuing Theory of Two-Populations with Lane Switching. Doctoral Dissertation. Chicago, IL: Illinois Institute of Technology.

**APPENDIX**

Before equilibrium equations are developed, system parameters must be revised to reflect their effective rates. Let ω equal the percentage of at-risk patients that switch to the general doctor (or vice versa). The effective arrival rates become:

$$\lambda_{1e} = (1-\omega)\lambda_1 \qquad and \qquad \lambda_{2e} = \omega\lambda_1 + \lambda_2$$

where $\lambda_{1e}$ is the effective or actual arrival rate of patients to the doctor assigned to treat at-risk patients and $\lambda_{2e}$ is the same for the doctor assigned to treat general patients. The proportion of total arrivals to each facility is revised as follows.

$$\alpha_{1e} = \frac{\lambda_{1e}}{\lambda_{1e}+\lambda_{2e}} = \frac{(1-\omega)\lambda_1}{(1-\omega)\lambda_1+\omega\lambda_1+\lambda_2} = \frac{(1-\omega)\lambda_1}{\lambda_1+\lambda_2} = \frac{(1-\omega)\lambda_1}{\lambda} = (1-\omega)\alpha_1$$

$$\alpha_{2e} = \frac{\lambda_{2e}}{\lambda_{1e}+\lambda_{2e}} = \frac{\omega\lambda_1+\lambda_2}{\lambda} = \omega\alpha_1 + \alpha_2$$

The effective utilization factors are then revised with the new arrival rates as follows:

$$\rho_{1e} = \frac{\lambda_{1e}}{\mu_1} = \frac{(1-\omega)\lambda_1}{\mu_1} = (1-\omega)\rho_1 < 1 \rho_{2e} = \frac{\lambda_{2e}}{\mu_2} = \frac{\omega\lambda_1+\lambda_2}{\mu_2} = \omega\frac{\lambda_1}{\mu_2} + \rho_2 < 1$$

$$\rho_{ce} = \rho_{1e} + \rho_{2e} = (1-\omega)\rho_1 + \omega\frac{\lambda_1}{\mu_2} + \rho_2 < 2$$

where $\rho_{1e}$ is the effective utilization for the doctor assigned to treat at-risk patients and $\rho_{2e}$ is the same for the doctor assigned to general patients.

The objective for switching lanes is to balance the work load between the two doctors. In order to determine the percentage of at-risk patients that must switch lanes, the two effective utilizations are set equal to each other and the resulting equation is solved for ω. If the final numerator is negative, ω is set to zero.

$$\rho_{1e} = \rho_{2e}$$
$$(1-\omega)\rho_1 = \omega\frac{\lambda_1}{\mu_2} + \rho_2$$
$$\rho_1 - \omega\rho_1 = \omega\frac{\lambda_1}{\mu_2} + \rho_2$$
$$\rho_1 - \rho_2 = \omega\frac{\lambda_1}{\mu_2} + \omega\rho_1$$
$$\rho_1 - \rho_2 = \omega\left(\frac{\lambda_1}{\mu_2} + \frac{\lambda_1}{\mu_1}\right)$$
$$\frac{\rho_1 - \rho_2}{\left(\frac{\lambda_1}{\mu_2} + \frac{\lambda_1}{\mu_1}\right)} = \omega$$

 