# MODELING TELEMEDICINE SYSTEMS TO EFFECTIVELY ALLOCATE ADMINISTRATIVE AND MEDICAL RESOURCES

by

RYAN AESCHLIMAN

Industrial and Manufacturing Systems Engineering

A FINAL HONORS PROJECT

submitted in accordance with the University Honors Program requirements

Industrial and Manufacturing Systems Engineering Department
College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

May, 2015

Approved by:

David Ben-Arieh

# Abstract

Telemedicine stands as one of the most promising innovations in healthcare. By delivering healthcare through electronic means, doctors can greatly expand the number and geographic diversity of patients they serve. While most telemedicine today focuses on more traditional health care applications, telemental health aims to apply telemedicine principles to mental health services. Telemental systems offer convenient and efficient ways for healthcare providers to provide psychological and psychiatric service to patients in far-flung geographic areas. Unfortunately, these systems can suffer from serious congestion if not well put-together. Maximizing the valuable time of doctors while ensuring short waits for patients should be a primary goal of telemedicine system design.

Telemental health systems come in several varieties such as synchronous, asynchronous, and group telemental health. Each variety offers different system properties and flow behavior.

This paper presents models for synchronous, asynchronous, and group telemental health systems using a discrete-event simulation and examines their properties using that tool. It compares their relative performance in this way. The analysis determines effective system capacity and demonstrates the effect of expanding the number of doctors and patients in the system. Moreover, the results can serve as a tool for healthcare providers seeking to establish telemental health systems.

The models revealed that, given a single specialist, nurse, set of audiovisual equipment, and administrative team, group telemedicine offered the highest capacity of roughly 300 patients at a time. Asynchronous individual systems followed with a capacity of 30, and synchronous individual systems trailed with a capacity of 25. As capacity decreases, however, the configuration's ability to provide patients with one-on-one care rises. The proper selection depends on the needs of the patient and the demands on the provider.

The analysis also extends to some modifications of the original models that remove assumptions made in describing group telemedicine, probe the impact of variance reduction, and examine the maximum number of specialists a single administrative team can handle.

# Table of Contents

# List of Figures

# List of Tables

# Section 1 - Introduction

Telemedicine, "the use of electronic information and communications technologies to provide and support clinical health care when distance separates the participants" (Institute of Medicine, 1996), stands as one of today's most potent innovations to extend the geographic reach of healthcare providers and free doctors from unnecessary waiting. In general, a patient at a "spoke" site and a doctor at a centralized "hub" site use some sort of electronic communication medium to connect for diagnosis, treatment, or treatment maintenance.

Recently, telemedicine has extended beyond its more traditional realms of dermatology, optometry, and even surgery to gain headway with psychologists and psychiatrists (Locatis & Ackerman, 2013). Literature refers to the practice of psychiatry and psychology via telemedicine as telemental health. One paper aims to define telemental health as broadly as possible by describing it as "the practice of mental health specialties at a distance," (Grady et al., 2011).

The field of telemedicine suffers from several problems, however. Many of them, such as the relatively shallow pool of research on its effectiveness and the frequent resistance to implementing telemedicine systems, can be best resolved by healthcare professionals. Other problems, such as difficulties estimating the capacity of telemedicine systems, can be solved easily with engineering tools. Section two provides a detailed discussion of telemedicine.

Many of the weaknesses of telemedicine can be understood through discrete event simulation. This method of analysis models systems as a series of tasks that take probabilistic times to perform and a series of queues that represent people or things waiting for a task to complete. The tool can quickly reveal information about systems that may not be easy to measure in practice. Moreover, analysts can use discrete event simulations to explore the impacts

of changes to system configuration without committing resources to piloting the change in real life.

The bulk of this paper describes models to capture the intricacies of telemental health systems using discrete event simulation. Sections three, four, and five layout models of three different types of telemental health systems. Finally, section six uses the results of the models to compare these three systems and draw conclusions regarding their appropriateness to certain applications.

# Section 2 - The Challenges of Telemedicine

Telemedicine holds enormous potential but poses enormous challenges as well. This reality has prompted many researchers to study the issues telemedicine has and propose entirely novel system configurations to alleviate some of these concerns. This section discusses this literature and lays out the fundamental components of telemedicine systems. Much of this section comes from previous telemedicine research (Aeschliman, 2015).

## 2.1 – Telemedicine Basics

The earliest manifestation of telemedicine arose in the 1870's when a group of drugstores bought telephones to coordinate orders. Telemedicine's contemporary incarnation materialized in the 1960's, however, with the advent of modern telecommunications (Sang Goo, Mun, Jha, Levine, & Ro, 2000).

The primary problem telemedicine aims to solve is one of distance (Rabinowitz, Brennan, Chumbler, Kobb, & Yellowlees, 2008). Many communities in rural or traditionally underserved areas lack access to efficient regional health centers, making access to specialized care a problem. By eliminating travel time, telemedicine can greatly extent the effective range of these specialists. For example, an initiative in rural Alaska provides expert consultations to Health Aids in far-flung communities (Sang Goo et al., 2000). Similarly, a project in rural Kansas provides child psychological services directly to patients via telemedicine (Spaulding, 2010).

From a clinical perspective, telemental heath performs on-par with face to face consultations. In fact, in some cases it may prove superior. Patients with paranoia or PTSD can befit from the "distance" a video screen provides. Moreover, some forms of telemedicine done in the patients' homes may allow the specialist to examine living conditions and make better

diagnoses. Finally, the zoomed-in perspective and ability to pause the video in some situations can allow providers to more accurately read certain facial expressions (Rabinowitz et al., 2008).

Studies show telemedicine can be cost-effective. As one queuing network model suggested, telemedicine makes financial sense when it works well and specialists have extremely valuable time. The model also argued that telemedicine should be avoided when administrative staff costs a comparatively large amount of money, traditional telemedicine proves highly effective, or mistreatment costs loom large. The authors of the study caution that telemedicine should supplement, rather than replace, conventional treatments (Tarakci, Sharafali, & Ozdemir, 2007).

## 2.2 – Areas of Weakness for Telemedicine

The body of literature on telemedicine lacks the depth and validity of other fields, according to some authors. In particular, sample sizes for studies need to expand to more than just pilot cases with a few dozen patients (Rabinowitz et al., 2008). Others researchers call for more studies on the clinical effectiveness of telemedicine and observe a number of reports that merely look at the practice's economic impact (Krupinski et al., 2006).

Rabinowitz (Rabinowitz et al., 2008) contends more research on telemedicine's impact and patient satisfaction in rural areas should be performed. The authors note the unreliable methodology used in many studies to gauge patient satisfaction. Other researchers contend the flood of literature responding to this call tends to measure patient satisfaction with an upward bias (Zhang, McClean, Jackson, Nugent, & Cleland, 2013). In particular, many studies only look at patient satisfaction after the implementation of the system without obtaining baseline data.

Complexity often wards off would-be telemedicine users. A case study of a teledermatology system by Lasierra et. al. notes that the program lacked support from

organizational leadership or buy-in from providers. To make matters worse, the change required more person-hours per case than the original system. The program called for irregular hours on the part of specialists and did not integrate with the hospital's healthcare information system, making the change highly unpopular and ultimately a failure (Lasierra, Alesanco, Gilaberte, Magallón, & García, 2012).

Telemental health can require an uncomfortable change of tactics for providers to make an accurate diagnosis. The ability to re-watch video records and gather extra patient information is offset by many problems including the difficulty in establishing eye contact in a video messaging system (Grady et al., 2011). Grady notes that good-quality equipment can partially alleviate this; unfortunately, this adds to the cost of an already expensive system.

Limited patient knowledge regarding telemedicine also undermines implementation efforts: patient engagement is absolutely critical to the success of a system. Some literature calls this an "alignment with learning processes" and applies it to both patients and providers. The field of telemedicine is advancing rapidly and organizations must make an effort to keep pace (Cegarra-Navarro, Sanchez, & Cegarra, 2012).

Finally, patient-provider rapport building can be negatively impacted by telemental health. Relationships improve treatment quality through open communication (Grady et al., 2011); consequently, telemedicine systems should be designed to facilitate comfortable interactions. Telemedicine poses an additional challenge in this way as well. Expanded geographic ranges for specialists can put them face-to-face with patients outside their "cultural competency."

## 2.3 – Structure of Telemedicine

Telemedicine can be delivered in three different modalities: synchronous individual, asynchronous individual and synchronous group. Each modality has its own pros, cons, and particular system behavior. This subsection discusses each of these modalities.

Synchronous individual telemedicine systems are the most intuitive way to deliver healthcare by distance. They involve simply setting up a video camera or phone call with the provider at a "hub site" and the patient at a "spoke site". Hub sites will typically consist of a large, regional hospital, while spoke sites can be anything from a local hospital, to a clinic, to the patient's home. Telemedicine systems will usually have one hub and a large number of spoke sites.

Synchronous individual systems get their designation from the single patient that meets with a doctor in real, synchronous time. This setup is simple and "looks like" healthcare systems patients and providers may be familiar with. Asynchronous individual telemedicine, on the other hand, offers a very unconventional approach. Real time communication between the patient and provider does not occur in this modality. Instead, the patient collects or generates data at the spoke site and sends it to the specialist at the hub. The specialist can then examine the data at their convenience, prepare treatment recommendations, and send the recommendations back to the spoke site. For this reason, asynchronous telemedicine is often referred to as "store-and-forward" (Yellowlees et al., 2011).

Disciplines like dermatology and radiology work well as store-and-forward telemedicine. In the case of the former, photographs of skin can be easily stored digitally and sent to a hub site. For the latter, the typical radiology interpretation process works like a store-and-forward system by default (Yellowlees et al., 2011).

Telemental health services can be delivered asynchronously as well, as shown by researchers at the University of California-Davis in 2011. The psychologist or psychiatrist first prepares questions for the patient to answer and sends them to the spoke site. A nurse or general practitioner asks the patient those questions and records their responses on camera, sending them back to the specialist. They can make treatment recommendations at their convenience after watching the patient's responses. This process requires more person-hours to complete but also greatly increases the flexibility of the specialist's time. Incidentally, it also aids in record-keeping since hospital workers record the entire interview (Yellowlees et al., 2011).

This modality drastically changes the responsibilities of each stakeholder in the process. Specialists spend less of their time waiting and doing paperwork because administrators perform different tasks and patients take more responsibility for their own treatment (Yellowlees et al., 2011).

Asynchronous telemedicine can offer cost savings over synchronous systems. One study breaks down psychiatric evaluations into three components: "data collection, data analysis, and treatment planning" (Butler & Yellowlees, 2012). In synchronous systems, the psychiatrist performs all three components. The asynchronous telemedicine system leaves data collection in the capable hands of a nurse or general practitioner. Transferring these duties to a lower-cost employee effectively increases the number of patients a single psychiatrist can treat.

Laurant et al. argues that nurses can perform this duty competently and with similar health outcomes (Laurant et al., 2005). Nurses, according to the authors, tend to call for more tests and support services when filling these types of roles; however, the lower number of tests used in psychiatric and psychological evaluations minimizes additional costs imposed by this change. The true cost of asynchronous telemedicine systems depends highly on context. Laurant

et al. also note higher satisfaction scores from patients who receive treatment from nurses if they believe their cases are fairly minor.

The study goes on to note that simply making the change won't instantly free up the specialists' time. Doctors may have trouble delegating their responsibilities to nurses and may end up spending extra time needlessly double checking paperwork. The authors suggest "active steps" to keep doctors using their time as effectively as possible (Laurant et al., 2005).

The differences between synchronous and asynchronous systems extend to how the costs of the system scale with size. Synchronous systems have a higher fixed cost due to their more expensive equipment costs, while asynchronous systems suffer from a steeper variable cost. The breakeven point lies at approximately 250 patients (Butler & Yellowlees, 2012). Consequently, synchronous systems work well for larger health networks while asynchronous systems work better for small, local providers. The hardware cost difference can be surprisingly dramatic. One study indicated that, while synchronous systems require powerful and reliable professional audiovisual equipment, asynchronous systems can perform just as well with "pro-sumer" grade equipment normally targeted at hobbyists (Odor et al., 2011).

Psychologists feel comfortable with the performance of asynchronous systems. One study in California indicated doctors felt confident diagnosing Axis I and Axis II psychological disorders and even recommended treatment in 95% of cases (Yellowlees et al., 2010). That said, the authors of the study still recommend telemental health as only a supplement to, rather than a replacement for, conventional psychology.

Synchronous group therapy is the third modality of telemental health. This system bears resemblance to synchronous individual telemedicine, but with one key difference: more than one patient can receive treatment at the same time. Providers use teleconferencing equipment to

perform a group therapy session in real time with a facilitator in a remote location. This modality requires more time per session in preparatory and follow-up work, but also requires less time per patient to execute. The group therapy environment naturally limits the types of conditions treated to those treatable with group telemedicine; while it can serve as a good environment for combatting alcoholism and post-traumatic stress disorder, more severe disorders require individual attention.

Asynchronous group telemedicine intuitively lacks appeal – it essentially amounts to a group of patients gathering in a room to watch a video of the provider. While researchers may one day stumble into an effective use of this modality, this paper will consider it no further due to its presumed lack of effectiveness.

## 2.4 – Gauging the Effectiveness of Telemedicine

As hospital network leadership keeps a sharp eye on the effectiveness of emerging telemedicine systems, system proponents do well to utilize metrics to measure their program's success. Locatis and Ackerman recommend three "principles" to gauge the effectiveness of telemedicine systems: "congruency, fidelity, and reliability." In other words, telemedicine systems should seem similar to conventional systems, should collect comparable information in terms of quantity and quality, and should lead to the same conclusions (Locatis & Ackerman, 2013).

Patient satisfaction can also be reasonably tracked with a few key metrics (Zhang et al., 2013). These metrics hinge on surveys examining how a system-wide change to telemedicine impacted patient-provider relationships, communication, and the system as a whole.

Other researchers probe the effectiveness of telemedicine software platforms. In particular, "freedom from risk", reliability, and security stand out as essential features of store-

and-forward software systems (von Wangenheim, von Wengenheim, Hauck, McCaffery, & Buglione, 2012). The first and last criteria are straightforward. Reliability refers to the software's ability to perform tasks when needed.

Few studies exist that examine telemedicine systems from a process engineering perspective. In particular, system capacities and employee utilization lack dedicated studies. Consequently, this paper aims to begin filling this void in research.

## 2.5 – Key Roles in Telemedicine Systems

Telemental health systems require several key resources and employees to function. Experience at the New Mexico Veteran's Administration Hospital and literature (Laurant et al., 2005; Tarakci et al., 2007) indicate telemedicine systems typically include eight distinct resources and employees. These resources can be found in the list below.

- **The Patient –** Telemental health systems exist to provide treatment to patients and should focus on their needs. They attend therapy at the spoke site.
- **The Specialist –** This person is a psychologist or psychiatrist who works at the hub site. The system aims to extend their reach to as many patients as possible.
- **The Nurse –** Any capable staff member at the spoke site must attend the patient during sessions to answer questions and react to emergencies. In larger systems, this employee will often be a nurse. Smaller, local clinics may use a general practitioner instead.
- **Hub Site Administrator –** This employee works at the hub site to manage the telemedicine system, enroll patients, prepare the specialist, and generally keep things running smoothly. This paper uses the terms "hub site clinic coordinator" or "hub CC" interchangeably with this position.

- **Spoke Site Administrator –** The employee works at the spoke site and primarily checks patients in, works with the hub site administrator to disseminate information from the hub, and checks patients out while delivering any treatment materials. They will often have many roles at the spoke site in addition to running the telemedicine program; their work coordinating with the hub is absolutely critical to the system, however, and cannot be discredited (Grady et al., 2011). This paper may also refer to this position as the "spoke site clinic coordinator" or "spoke CC."
- **Hub Site Technology Specialist –** Large or complex telemedicine systems may have a separate employee to handle audiovisual equipment setup at the hub site. Smaller systems may have the specialist or hub site administrator establish a connection prior to appointments instead.
- **Spoke Site Technology Specialist –** Many clinics or small hospitals may employ someone to establish software connections with the hub site prior to meetings. Other operations may have the nurse, general practitioner, or spoke site clinic coordinator fulfill this role.
- **The Room –** While the specialist at the hub site will frequently be able to perform treatment from their own office, spoke sites will often have a dedicated room with audiovisual equipment to handle telemedicine sessions. These rooms may be limited in number and can, in some instances, represent a distinct constraint on the system.

One person may be able to perform multiple roles. For example, a patient who connects to a specialist from home effectively plays the roles of patient, nurse, spoke site administrator, and spoke site technology specialist. These roles can also be scaled up to accommodate large systems with dozens of specialists and spoke sites.

## 2.6 – The Telemedicine Process

The “invisible” background processes and administrative tasks critical to any well-run telemedicine process can lead to the system’s downfall if not well-planned. Meaningful analysis, therefore, requires a holistic view of telemedicine systems. Telemental health systems consist of four phases: patient enrollment, session preparation, the session itself, and session follow-up (Aeschliman, 2015). These four major components remain largely the same from system to system, regardless of the resource levels, configurations, and details of a specific instance.

The synchronous individual telemental health process at the New Mexico Veteran’s Administration Hospital gave rise to this model. Their system hinges around several psychologists, psychiatrists, and social workers stationed in the central hospital in Albuquerque. Community-Based Outpatient Centers (CBOCs) across the state, shown in the figure below, host spoke-side telemedicine services in locations more accessible from rural areas. Patients visit them, establish a connection with a provider in Albuquerque, receive their treatment, and check out. The following pages describe this process in-depth.


**Figure 2.1- CBOC Locations in New Mexico and southern Colorado**

**Image Source: (Google, 2015)**

The process begins with enrollment, where the specialist must accept or deny incoming referrals for treatment. Approved referrals proceed to the hub site administrator, who handles the scheduling, details, and preparatory work of patient enrollment. Enrollment occurs once per patient for some set number of appointments. In contrast, the remaining three phases happen weekly for the duration of the patient's enrollment.

The day of any appointment starts with session preparation. The phase begins with the patient's arrival, check-in, and initial paperwork at the spoke site. This requires time from the spoke site administrator as well, who may need to transmit some of this information to the hub site. Both technology specialists must set up the room and audiovisual equipment prior to the

session. Meanwhile, the hub administrator must provide patient information to the specialist who prepares for the upcoming therapy session.

After preparation, the patient and staff move on to the third stage: the therapy session. The patient and nurse/general practitioner meet in the room in the spoke site to converse with the specialist at the hub site.

The end of the session triggers the follow-up phase. The specialist retreats to recommend treatment, write prescriptions, or perform whatever follow-up proves necessary. The nurse or general practitioner answers patient queries while the specialist works and may send any important revelations back to the hub site for consideration. Upon completion of recommendations, the specialist forwards them to the hub site administrator for record-keeping, and the hub CC sends the recommendations back to the spoke site. The spoke site administrator uses these recommendations to provide the patient with treatment instructions and check them out. The patients effectively return to the start of phase two for the next week's appointment. Figure 2.2 on page 23 shows this process.

The store-and-forward model of asynchronous individual telemental health follows a fairly similar system (Butler & Yellowlees, 2012)**.** While not identical to the synchronous model, the asynchronous system contains the same four basic components.

Enrollment phases for asynchronous and synchronous sessions are exactly the same. The specialist approves or denies referrals and then hub site administrators enroll and schedule approved patients.

The session preparation phase begins closer to the session. As before, hub site administrators provide the specialist with relevant patient information. The specialist then generates a series of questions for the patient to answer on-camera based on this information.

The specialist sends this information to the spoke site via the hub site administrator, ending the first half of session preparation. While the first half occurs as much as a week before the session and happens entirely at the hub site, the second occurs the day of the session and happens entirely at the spoke site. The patient arrives at the spoke site, checks in with the spoke site administrator, and the process begins. The spoke site technology specialist (or, more likely, the nurse) prepares the therapy room. This task is simpler than in synchronous telemental health; no internet connection needs to be established to simply record the patient.

The session commences when both the room and patient are ready. The nurse or general practitioner asks the patient questions prepared by the specialist and trains a camera on the patient. Notably, the specialist has no real-time part in any of the proceedings. The session ends when the patient answers all the questions; they may check out for the day afterward.

Soon after the interview, the spoke site administrator sends the recorded responses to the hub site, signaling the start of the session follow-up phase. When they find a good time, the specialist watches the interview and recommends treatment based on the patient's responses. They send these back to the spoke site administrator via the hub site administrator.

Patients may need to return to the spoke site to receive treatment, possibly even meeting with the nurse or general practitioner for detailed instructions. Alternatively, they may be able to receive instructions remotely from their home. Either way, the patient checks out of the system for the week, indicating the end of phase four and the beginning of phase two for the next week. Figure 2.3 on page 24 describes this asynchronous process.

Synchronous group telemedicine systems are nearly identical to synchronous individual systems. The primary difference lies in group processing. Patients check in, check out, and enroll individually, but all other processes happen for the group as a whole. The existence of the group

does not change the resources, steps, precedence of events, or rough timeframe needed to execute a session. Some process times for individual steps may be shorter since the cases considered in group therapy environments frequently lack the complexity of one-on-one therapy cases. Figure 2.4 on page 25 shows this process.

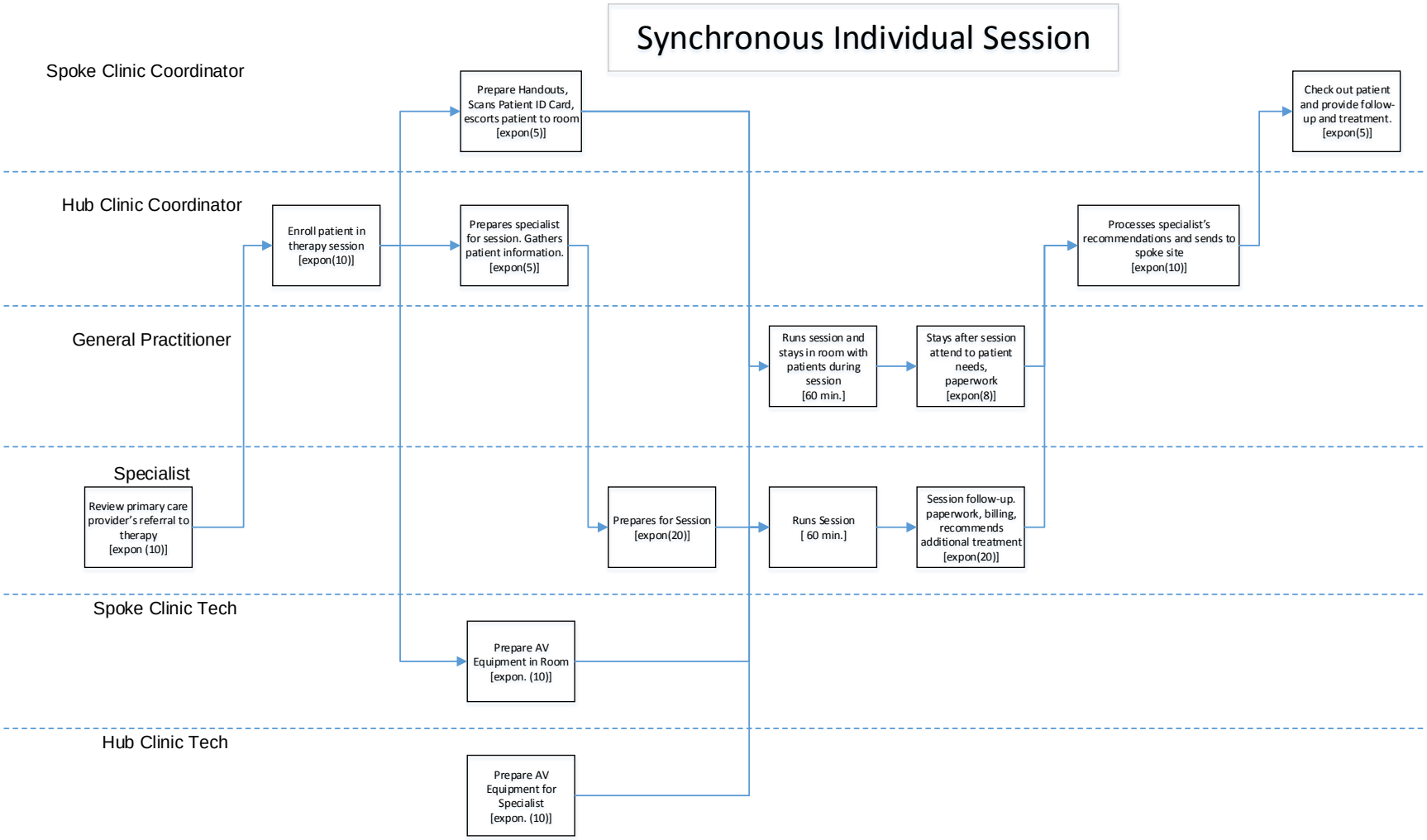


**Figure 2.2-Structure of an individual synchronous telemental health session**

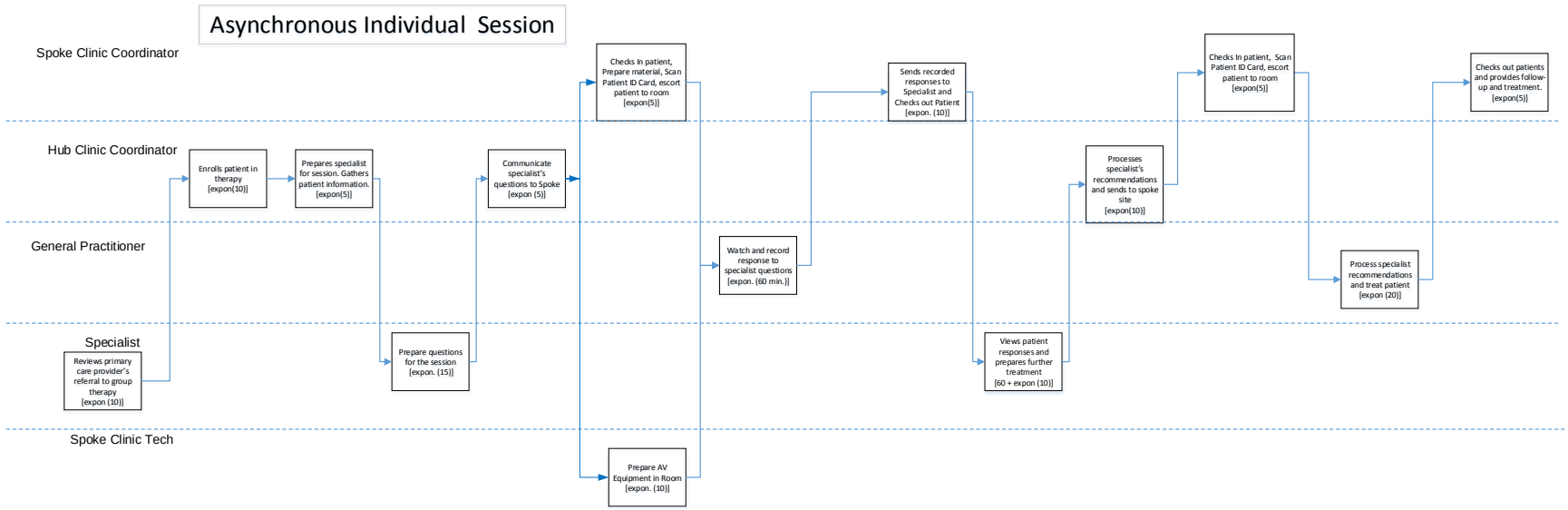


**Figure 2.3– Structure of an individual asynchronous telemental health system**

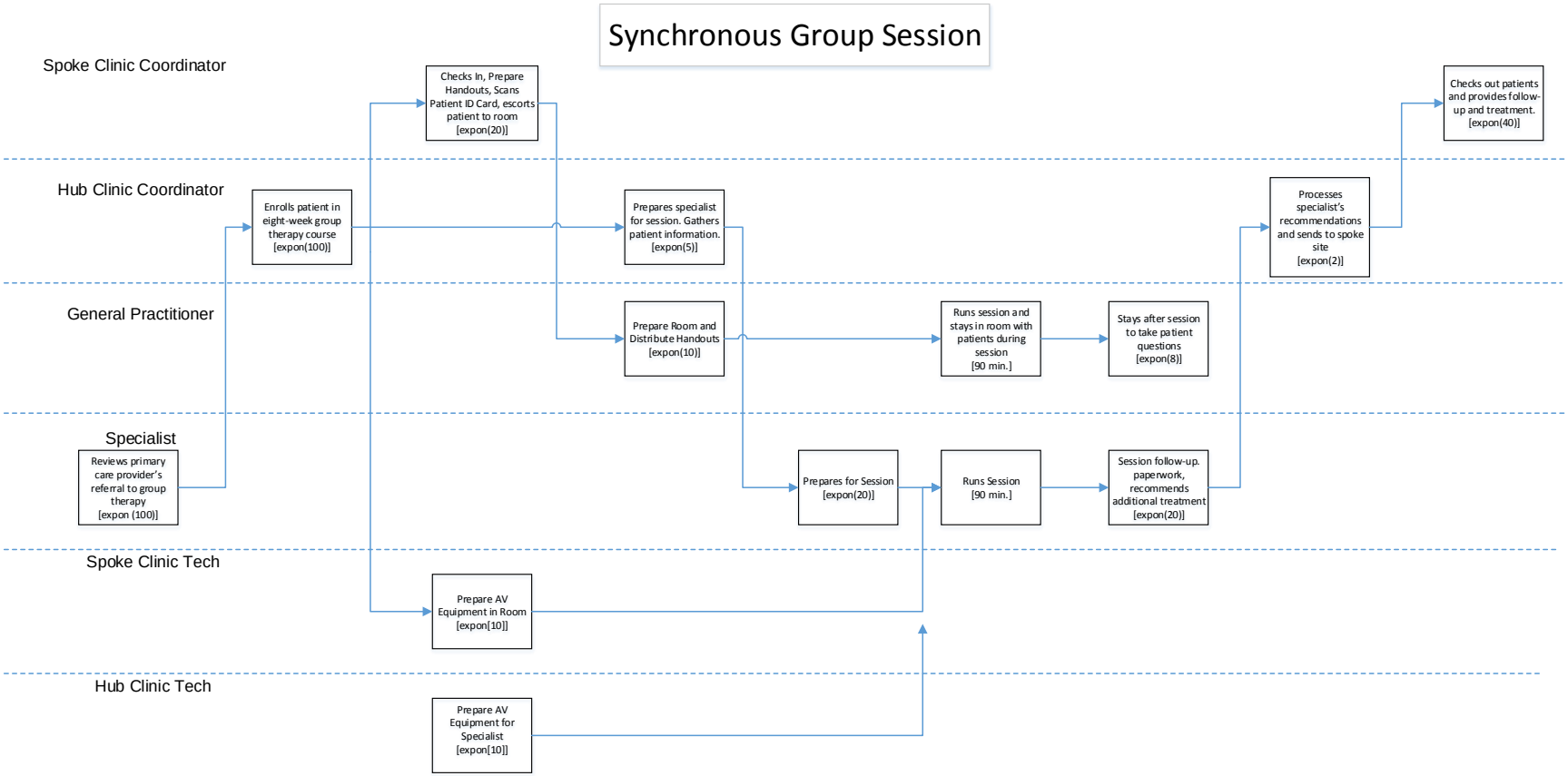


**Figure 2.4 – Structure of a group synchronous telemental health system**

# Section 3 - Simulating Synchronous Individual Telemental Health Systems

Telemental health systems can prove difficult to mathematically describe. Their high number of interconnected actors and interdependent events means simple equations and basic queueing models lack the power to adequately predict their behavior. Simulation offers a better option (Lach & Vazquez, 2004). Medical settings tend to offer the intrinsic structure and clearly defined responsibilities that make modeling them this way both simple and accurate. Moreover, variability can be easily included in the model to capture the unpredictability of treatment times and patient behavior.

Discrete-event simulation can handle the complexity of telemental health systems. The basic premise behind this technique is fairly simple. Discrete-event simulations model the time between individual events using probability distributions. Computers typically use random number generators to produce a time according to a distribution set by the user. To model a process, it checks to see if the resources needed to execute that process are available, marks them as unavailable if they are, and keeps them as such for a randomly generated period of time. A simulation will include numerous processes that compete for resources, and also will track the queues that form as patients wait for a particular process to begin.

While simulations are a virtual environment to model changes to systems, the time distributions for each individual process in a system must be based on reality for the simulation results to carry weight. Preliminary work with the New Mexico VA provided useful estimated process times that serve as the foundation of a simulation model to describe their telemental health system. The simulation aims to measure utilization of the hardware and employees that make telemedicine systems work, as well as predict the maximum capacity of such a system

given a set number of employees. The model works by tracking the flow of a theoretical object through the system. This object merely triggers events and is a "token" with no physical analog. If anything, the token signifies the flow of paperwork through the system. As the paperwork progresses, it summons employees, patients, and resources to aid in its forward march. The simulation measures the amount of work a patient's case gives to each resource as well as the amount of time it takes the system to treat a patient, from the first referral to when they walk out the door after therapy (Aeschliman, 2015).

The simulation requires service time distributions for each task. Due to geographic constraints, measured service times proved impractical to obtain. Instead, the average times presented in the following sections are estimates deemed reasonable by the New Mexico VA. The simulation assumes exponential distributions as a safe choice for most service times due to their convenient properties and moderate levels of variability. Indeed, the exponential distribution is one of the most traditional choices for modeling interarrival times and researchers frequently use it in simulations (Lach & Vazquez, 2004).

This section describes a discrete-event simulation model used to describe the synchronous, individual system utilized by the New Mexico VA Hospital. This model has been previously described (Aeschliman, 2015). Sections four and five propose alternative models for asynchronous and group telemedicine systems with comparable characteristics.

## 3.1 – Discrete Event Simulation Fundamentals

Arena, the simulation software from Rockwell Software used to model this system, bases its logic off of four key "blocks" or actions: queue, seize, delay, and release. Paperwork enters a queue to wait for attention from a resource, seizes that resource when it becomes available, uses or delays that resource for a period of time, and releases that resource when finished. For

example, the spoke site administrator could have a pile (queue) of clinical recommendations from a therapist to distribute one day. The administrator would select the top paper in the stack, work on it for a time, then move on to the next paper in the stack when complete. The paperwork then proceeds through many more queues until finally leaving the system at a dispose block. The simulation basically "sees" this process from the paperwork's point of view.

The model assumes that only one specialist, hub administrator, spoke administrator, spoke nurse or general practitioner, pair of technology specialists, and set of spoke telemedicine equipment can be occupied at any given time. These numbers can be tailored to each spoke site in application. For flexibility, the simulation tracks the time of all the employees, the patient, and the physical room at the spoke site where patients go for therapy.

Resources can be reconfigured or expanded. Reassigning resources to multiple tasks can model systems that have one person fill several roles. For example, in a system where the nurse or general practitioner also sets up the spoke-side audiovisual equipment, all tasks that require the spoke technology specialist would require the nurse or general practitioner instead. Simulation users may also model facilities with multiple sets of resources by simply increasing the number of available resources.

## 3.2 – Synchronous Individual Telemental Health Systems

The most basic and intuitive telemedicine systems fall into the category of synchronous individual, where a single patient at a time meets with a provider in a remote location in real-time. The simulation creates one patient at a time at a specified rate and sees how the system performs over a day (480 minutes). This "day by day" model paints an accurate picture of steady-state operations.

The model begins with a queue representing the referral review process, as shown in figures 3.1 and 3.2. Since this process occurs independently of the actual care delivery, the simulation has a separate series of events modeling them. The model assumes that for every four patients that have appointments on a given day, one patient needs to have their referral reviewed and enrollment performed. A separate "create" block generates these cases one at a time, according to an interarrival time that varies with the number of patients scheduled that day. In fact, the interarrival time on this block is precisely equal to the interarrival rate of regular patients times four. After arrival, the two queues proceed to their own dispose block for the "paperwork" to leave the system. The dispose block ends the process. The referrals do not interact with the appointment execution side of the simulation except to use the specialist and hub administrator's time.

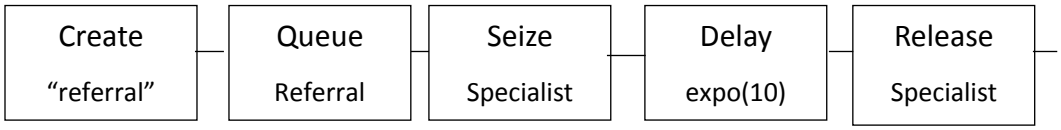


**Figure 3.1- The first part of the independent referral process**

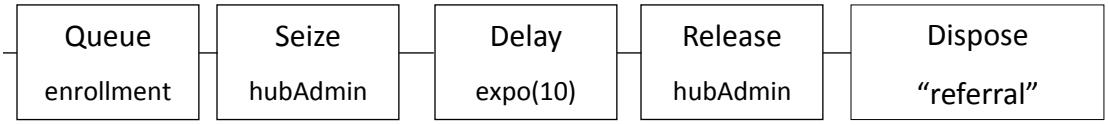


**Figure 3.2– The second part of the independent referral process**

While not connected to the system directly, the enrollment process still requires the time of the specialist and spoke administrator to complete. Neither operation happens instantly, so both delay for time following an exponential distribution with a mean of 10 minutes (henceforth abbreviated as expo(10) minutes) each.

For the primary treatment phase of the telemental health process, the simulation begins with a "create" block that generates patients one at a time. The first patient arrives at time zero

(8:00 AM), and the next arrives at a precise time determined by the desired number of patients scheduled. For example, if the clinic is supposed to serve two patients, the interarrival time is 240 minutes, while a clinic serving eight patients daily would have an interarrival time of exactly 60 minutes.

The output of the create block uses an "alter" block to add one to the number of patients available to seize for appointments. The patient cases proceed to an "alter" block which adds one to the number of "appointment" resources the simulation will attempt to complete. A "duplicate" block then creates four additional copies of the patient's case, allowing the simulation to execute the next four tasks simultaneously.

As patients congregate to the spoke site, they must check in. Each one seizes a patient resource, spoke site staff, and an appointment resource. It delays them all for expo(5) minutes, then releases the patient and spoke site staff for other tasks. It keeps the appointment resource seized; it will not release it until the entire appointment process is complete, effectively keeping the same appointment from happening twice.

The second copy of the original patient goes to the spoke site technology specialist. The copy seizes the employee as well as the AV equipment at the spoke site. The simulation delays for expo(10) minutes to model the time the spoke tech specialist needs to prepare the room at the spoke site and set up the connection to the hub site. At that point, the employee is free to go, but the AV equipment remains unavailable until the completion of the group therapy session. The third copy triggers the hub site tech specialist to spend expo(10) establishing the connection on the other side.

The fourth and final copy travels immediately to a hub site staff person, who takes expo(5) to gather up-to-date information on the patients and transfer that information to the

specialist. The specialist then takes expo(20) researching and preparing for the upcoming session. Figure 3.3 shows this stage of the process.

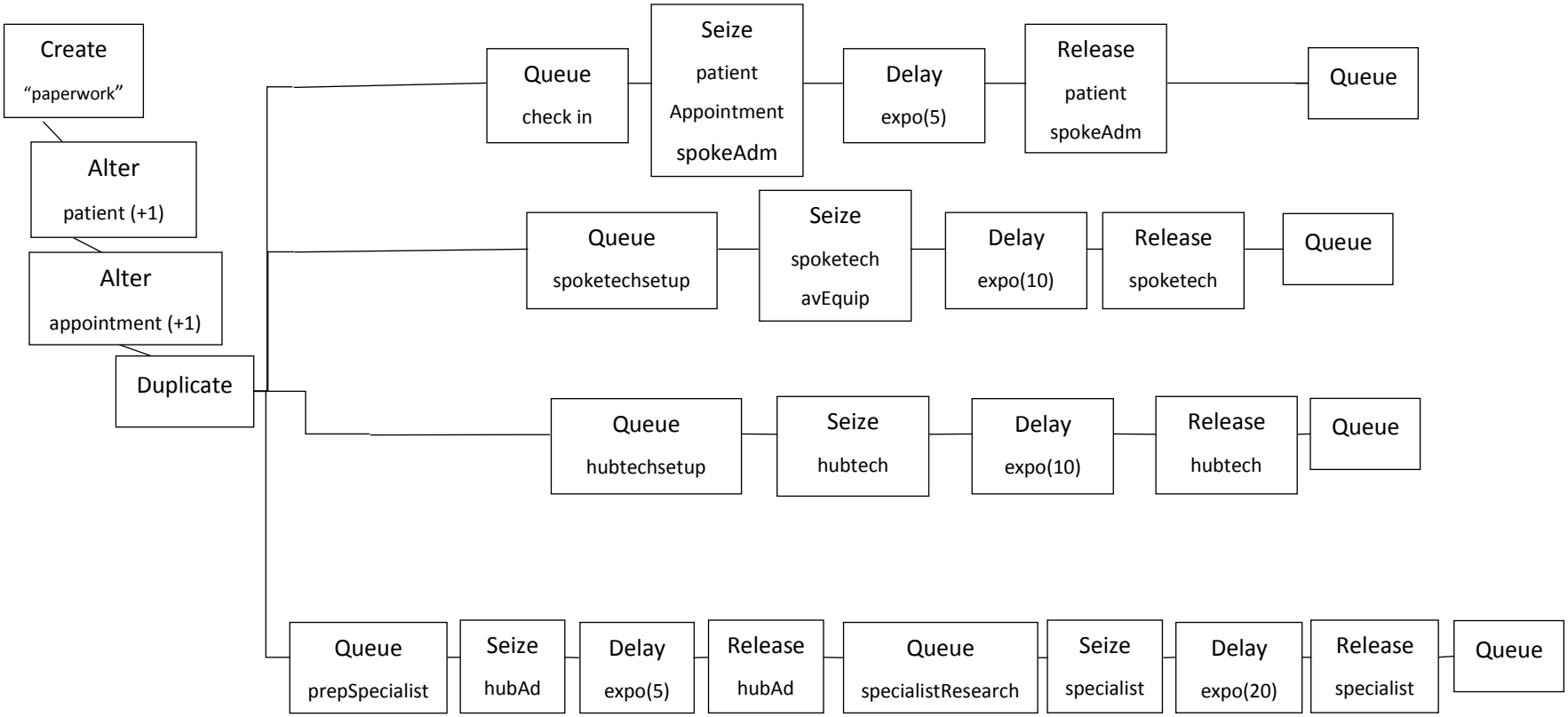


**Figure 3.3– Simultaneous preparatory events for the synchronous individual session**

The four copies empty into queues which feed into a "match" block that bars progress until all four tasks are complete. When that occurs, all four copies proceed at once. Three simply get disposed, but the first, the original, goes on to trigger the start of the group session. The main event requires exactly 60 minutes of simultaneous time from the patient, a general practitioner (GP)/nurse (to sit in on the session and serve as an in-person point of contact for patients), and the specialist (virtually present through a teleconferencing or telemedicine service). At the end of the session, the simulation releases only the AV equipment; the people involved in the session have follow-up work to do. Figure 3.4, below, shows this simple part of the simulation.

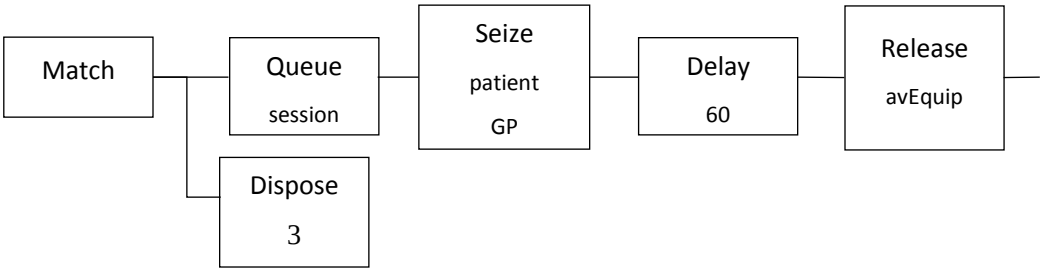


**Figure 3.4 – The completion of the preparatory activities and session itself for the individual synchronous model**

Upon the completion of the session, the general practitioner/nurse takes expo(8) minutes after the session to talk to patients, answer questions, and collect feedback. They send their notes to the specialist when complete. At the same time, the specialist then takes expo(20) minutes to prepare follow-up from the session and recommend further treatment, if necessary. The simulation releases them as they give their follow-up to hub site staff. Figure 3.5 illustrates this.

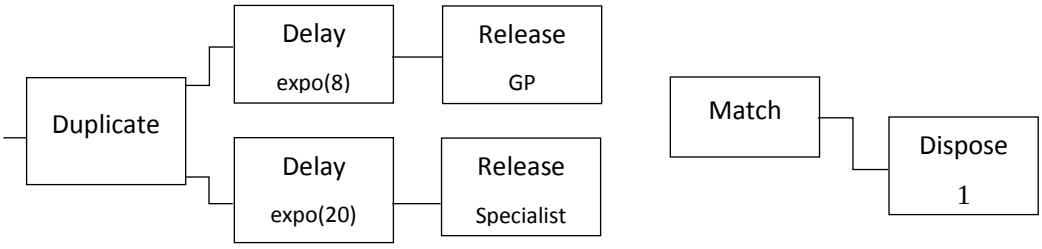


**Figure 3.5– Immediate follow-up from the session in the individual synchronous model**

When the match block allows the simulation to progress (that is, when the specialist and general practitioner complete their follow-up work), the hub site staff takes expo(10) minutes to send that follow-up to the spoke site. Finally, the patient takes expo(5) minutes of a spoke site staff member's time to receive follow-up instructions. Figure 3.6 shows this logic.

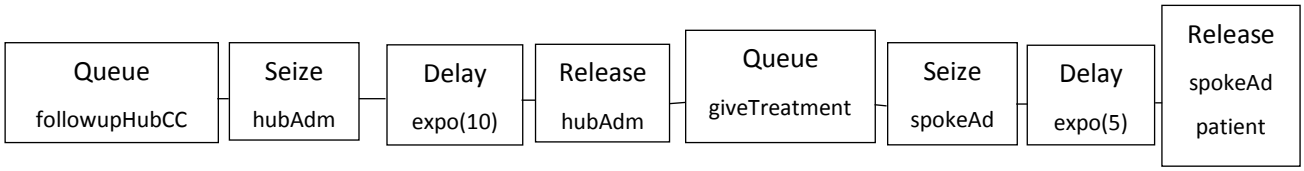


**Figure 3.6– Wrap-up activities for the individual synchronous session**

After this, the patients are free to go. A "duplicate" block makes a copy of each patient and sends it to the "duplicate" block at the start of figure 3.3 via a "delay" block that holds each patient for 2280 minutes. In more intuitive terms, after the appointments, patients wait a week (minus the time of the appointment itself) and go on to their next one. Since the simulation only lasts a day, this means the patients are effectively gone. That said, this construction also grants flexibility for variations of the model that look at larger timeframes. On the other side of the duplicate block, the simulation releases the appointment and permanently reduces the number of available appointments by one. Figure 3.7 shows this part of the simulation.

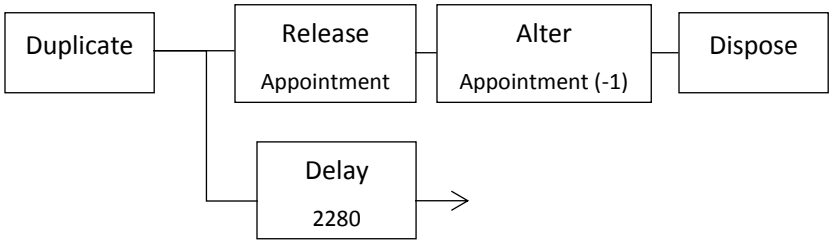


**Figure 3.7 – Final closure logic for the individual synchronous session**

The model runs for 200 replications, each lasting one eight-hour day (480 minutes). At the end of each day, the program resets the queues and statistics; each day is independent of every other day. The table below summarizes the general structure of the model. Figure 3.8, immediately following, shows a screenshot of the simulation model in Arena software to demonstrate how all the model components fit together.

**Table 3.1– Tasks and their times in the individual synchronous model**

| **Action** | **Time** | **Actor** | **Assumptions** |
|---|---|---|---|
| Referral Review | 10 min | Specialist | Occurs 25% as often as appointment execution. No patient is turned away. |
| Enrollment in Program/Session | 10 min | Hub Admin | |
| Check-In | 5 min | Spoke Admin, Patient | |
| Room Technology Preparation | 10 min | Spoke Technology Specialist, Room | |
| Hub-Side Video Preparation | 10 min | Hub Technology Specialist | |
| Gather Materials for Specialist | 5 min | Hub Admin | |
| Prepare for Session | 20 min | Specialist | |
| Run Session | 60 min | Patient, Nurse/GP, Specialist, Room | |
| Answer Patient Questions/Follow-Up | 8 min | Nurse/GP | |
| Prescribe Treatment and other Follow-Up | 20 min | Specialist | |
| Send Treatment to Spoke | 10 min | Hub Admin | |
| Check-Out | 5 min | Spoke Admin | |

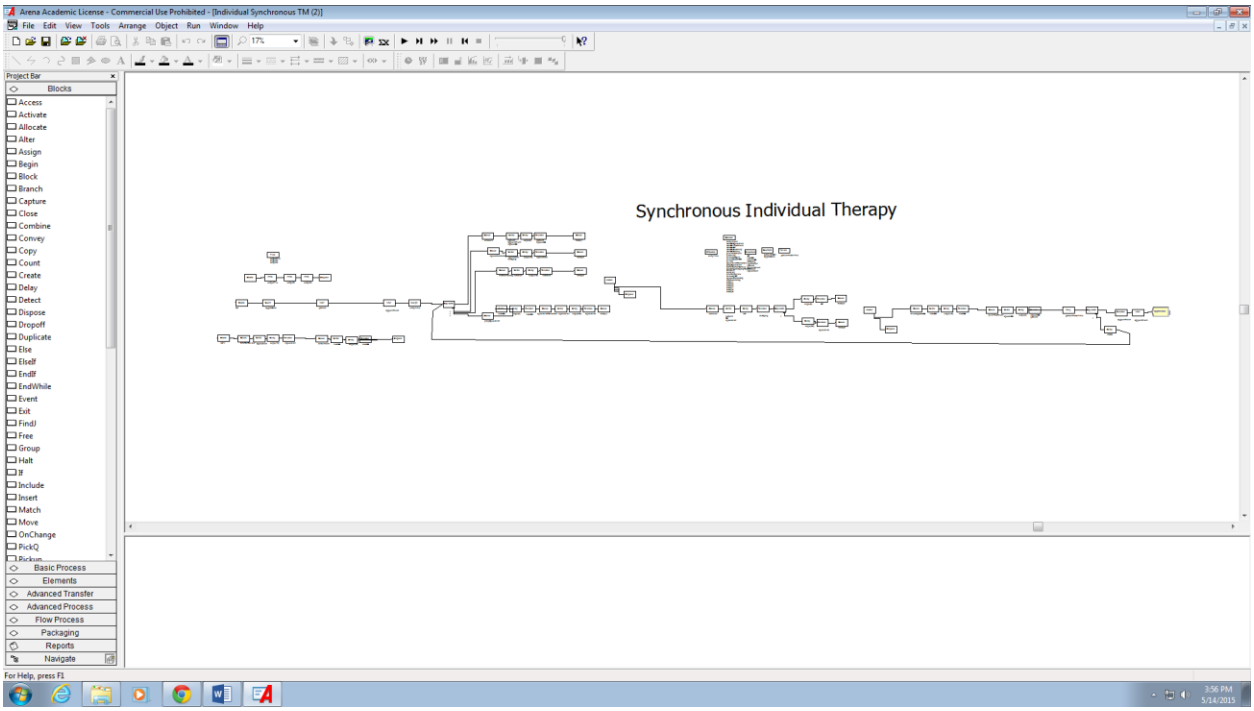


**Figure 3.8 – Individual synchronous model as it appears in Arena**

## 3.3 – Synchronous Simulation Results

Arena performed 200 replications of the individual synchronous telemental health model for each of several varying patient scheduling levels. Essentially the model scheduled some number of potential appointments and tracked how many the available staff could actually fulfill in the allotted time. The simulation tracked the utilization of staff members and physical resources while also gauging the number of appointments that could be completed in that time. Figure 3.9, below, shows how employee and resource utilization changes as the number of enrolled patients increases and table 3.2 provides those numbers directly.

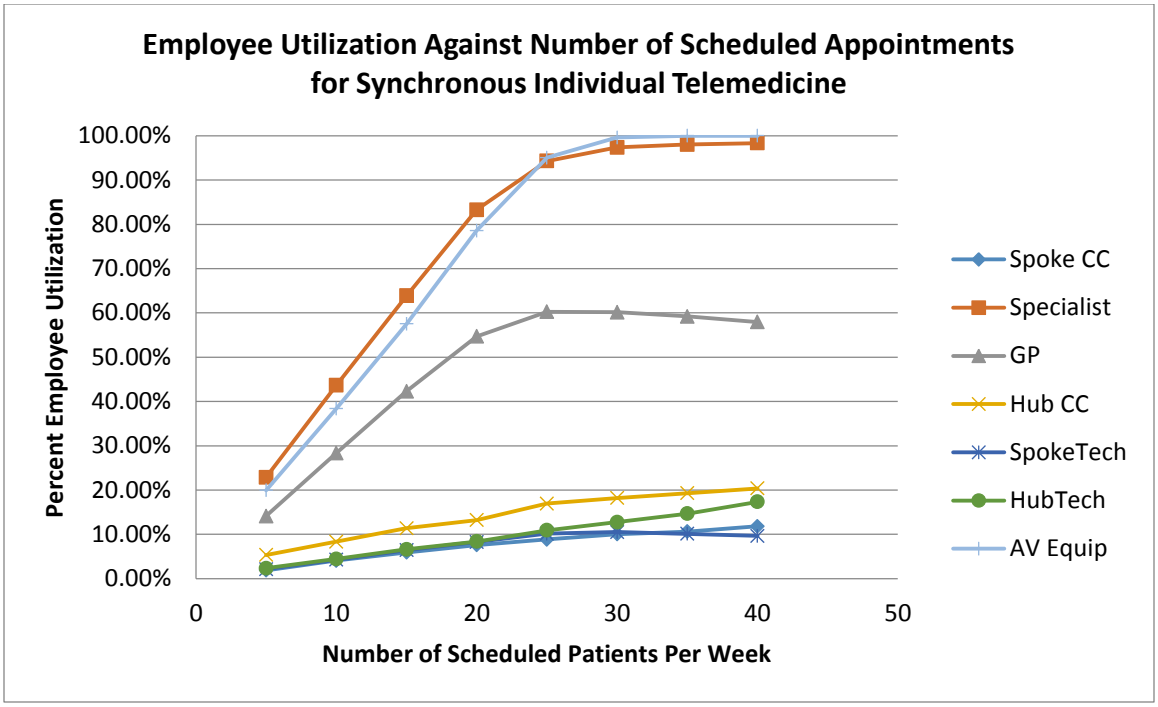


**Figure 3.9– Employee utilization for the synchronous individual model.**

**Table 3.2– Resource utilization for the synchronous individual model**

| # patients | Spoke CC | Specialist | GP | Hub CC | SpokeTech | HubTech | AV Equip |
|---|---|---|---|---|---|---|---|
| **5** | 1.89% | 22.87% | 14.11% | 5.34% | 2.15% | 2.36% | 19.95% |
| **10** | 4.09% | 43.69% | 28.34% | 8.34% | 4.25% | 4.45% | 38.40% |
| **15** | 5.94% | 63.87% | 42.32% | 11.39% | 6.44% | 6.61% | 57.52% |
| **20** | 7.59% | 83.28% | 54.69% | 13.21% | 8.27% | 8.37% | 78.57% |
| **25** | 8.84% | 94.29% | 60.25% | 16.95% | 10.14% | 10.92% | 95.03% |
| **30** | 10.04% | 97.35% | 60.14% | 18.20% | 10.53% | 12.74% | 99.60% |
| **35** | 10.63% | 98.02% | 59.20% | 19.30% | 10.13% | 14.67% | 99.98% |
| **40** | 11.81% | 98.34% | 57.94% | 20.38% | 9.63% | 17.36% | 100.00% |

The AV equipment and specialist approach 100% utilization at 30 scheduled appointments per week. The general practitioner approaches 60% utilization after 25 appointments per week. Synchronous individual telemedicine systems with one specialist have

an effective limit of about 25 appointments per week, or five per day. This leaves roughly three hours per day for non-session activities the specialist needs to complete, such as reviewing referrals, preparing for sessions, and performing follow-up. Figure 3.10, on the next page, details the waiting times for certain processes in this model and shows how it varies as the number of scheduled patients increases.

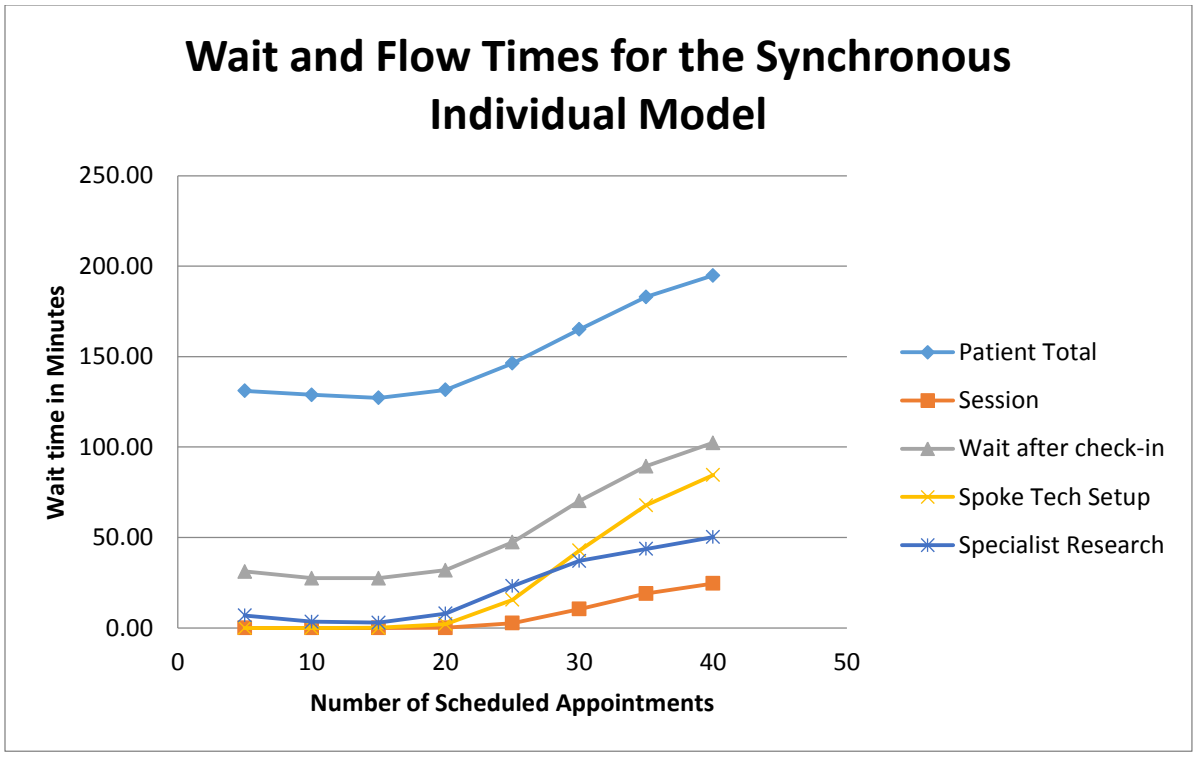


**Figure 3.10 – Wait times for the synchronous individual model by number of scheduled appointments**

In the figure above, "Patient Total" refers to the total time for the patient from their arrival to when they leave the building. "Session", "Spoke Tech Setup", and "Specialist Research" refer to the waiting time to execute those steps of the process. "Wait after check-in" plus "session" shows to how long patients have to wait after checking in before their appointment starts. Note how the wait times remain about constant until the four appointments

per day mark, after which they start rising. The average patient waits for thirty to fifty minutes after checking in before their session begins, but beyond that the simulation processes patients at a fairly reasonable rate. These realistic times help validate the model and indicate it provides a reasonable approximation of similarly configured telemedicine systems.

Table 3.3, below, illustrates the constraints a synchronous system puts on process flow. The numbers in the table represent the average number of people in line ahead of a patient's case upon the completion of a session preparation event. For example, when the system tries to accommodate 25 patients, 0.52 patients at any given time will have checked in and are waiting for the previous session to end. Since sessions last an hour, this time can add up quickly.

**Table 3.3 – Average queue lengths for the session after completion of setup steps**

| # patients | Check In | Spoke Room Setup | Hub Setup | Specialist Prep |
|---|---|---|---|---|
| **5** | 0.07 | 0.05 | 0.05 | 0.01 |
| **10** | 0.11 | 0.09 | 0.09 | 0.01 |
| **15** | 0.17 | 0.14 | 0.14 | 0.02 |
| **20** | 0.27 | 0.21 | 0.23 | 0.03 |
| **25** | 0.52 | 0.29 | 0.46 | 0.05 |
| **30** | 0.92 | 0.25 | 0.86 | 0.20 |
| **35** | 1.39 | 0.19 | 1.32 | 0.43 |
| **40** | 1.85 | 0.14 | 1.76 | 0.67 |

While not perfect, the synchronous individual model appears to offer reasonable performance. However, in today's cash-strapped healthcare industry, reasonable may not be good enough. The next section explores an alternative, asynchronous approach to telemental health.

# Section 4 - Simulating Asynchronous Individual Telemental Health Systems

An asynchronous approach to individual therapy may offer improvements to the synchronous model. By reducing the number of resources that need to be available at once, this model can hopefully cut down on queue times.

In many ways, the asynchronous model works very differently from the synchronous. The asynchronous model has four major components: enrollment, preparation at the hub site, immediate session preparation and follow-up at the spoke site, and final follow-up at the hub site. These components can be executed simultaneously, and the setup of the model reflects this.

## 4.1 – Asynchronous Individual Telemental Health Systems

The simulation lasts for 480 minutes (one day) and starts just like the individual synchronous model. Two "create" blocks, one for session execution and the other for the referral review and enrollment, generate patients according to an interarrival time agreeing with the number of patients on the schedule. The creation of a unit of paperwork also adds a "patient" resource to the simulation using an "alter" block. As before, the referral review and enrollment processes take expo(10) minutes each. The model assumes all patients have their referral accepted. Figures 3.1 and 3.2 in the prior section show roughly how this looks in the simulation.

Individual patient cases go to a "duplicate" block which splits the entire stream into three concurrent processes. After enrollment, an asynchronous session consists of three major phases: preparation, execution, and follow-up. The entirety of any one of these steps should be completed during one day, but all three phases for a single patient need not be completed in the same day. From the simulation's perspective, completing a step in one phase has no bearing on

the ability to complete a step in another phase; for convenience, one can safely assume that the preparation phase represents “tomorrow’s” appointments, the execution phase represents “today’s”, and the follow-up phase represents “yesterday’s”. What does matter, however, is the number of cases at each phase of the process. To stay on pace, each phase has an equal quota of appointments to process. To mark an appointment as complete in the simulation, the program must complete each phase once. Making the three phases independent means that it marks an appointment as “done” each time all three phases have completed a unit of work. While this, strictly speaking, may not represent the same appointment, this difference does not impact the validity of the simulation analysis.

The first phase in the model involves preparing for the session at the hub site. The store-and-forward model requires questions be prepared prior to the patient appointment, so cases begin at the hub clinic coordinator. This employee takes expo(5) minutes per session to gather patient information and present it to the specialist. The specialist then needs expo(15) minutes to prepare questions for the patient to answer at a later time. The model assumes the specialist creates an entirely new set of questions for each patient, tailored to them individually. In practice, the specialist may be able to reuse questions, reducing this time. After this, the hubCC must spend expo(5) minutes sending the questions to the spoke site, concluding the preparation phase. Figures 4.1 and 4.2, below, show this part of the simulation.

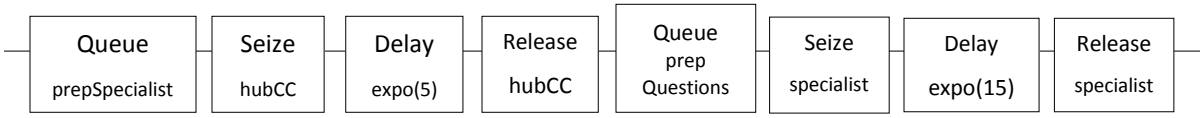


**Figure 4.1 – First part of the asynchronous session preparation phase**


**Figure 4.2 – The hub CC sends session questions to the spoke site**

The second branch leads to session execution at the spoke site, the longest and highest priority phase. Given the choice, the simulation prefers to complete a phase in this part of the process than elsewhere. The phase begins with a "duplicate" block that further splits the process into two concurrent sets of events. To check-in, the simulation requires a patient and spoke site clinic coordinator to spend expo(5) minutes scanning id cards and walking to the appropriate room. At this point, the simulation also reserves an appointment resource to ensure that the same appointment does not happen twice. Unlike the patient and spoke clinic coordinator, the simulation does not release the appointment resource until the entire appointment process is complete. During this time, the spoke site technology specialist sets up the room. This process takes expo(10) minutes and seizes the AV equipment resource. When both the patients and the spoke technology specialist have finished these tasks, a "match" block allows the simulation to progress. Figure 4.3 shows these steps.

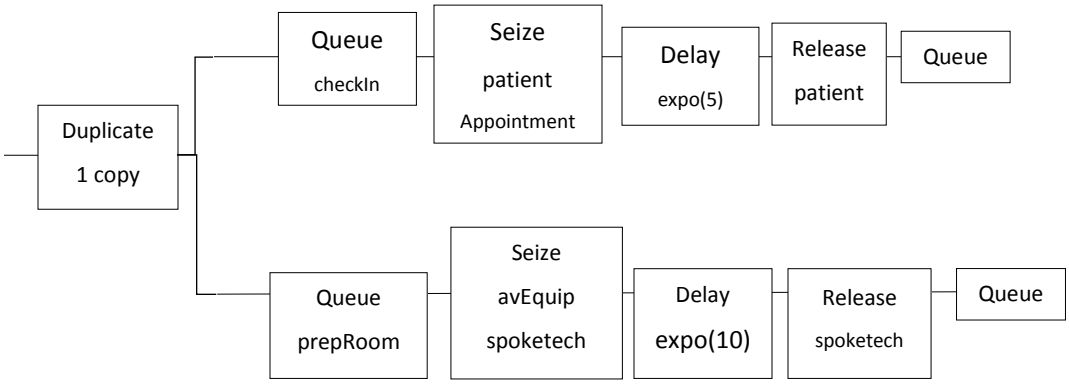


**Figure 4.3 – Patient check-in and room preparation**

With that, the expo(60) minute session begins. Its asynchronous nature means that only the patient, the general practitioner/nurse, and the room (already seized by the technology specialist in the last step) need to be available for the appointment to commence. Note that the model is only asynchronous in the sense that the specialist and patients do not have to be

available at the same time. The asynchronous configuration does not directly use a doctor's time during this phase, making early or late completion more of a possibility as reflected by the probability distribution on the session time.

Upon the group session's completion, the simulation releases the room and equipment, patients, and general practitioner. The less-involved nature of the session negates the need for much paperwork or question time. Rather, the spoke CC takes expo(10) minutes to send the recorded patient responses to the hub site, ending the execution phase. Figure 4.4, below, shows the appointment and immediate follow-up.

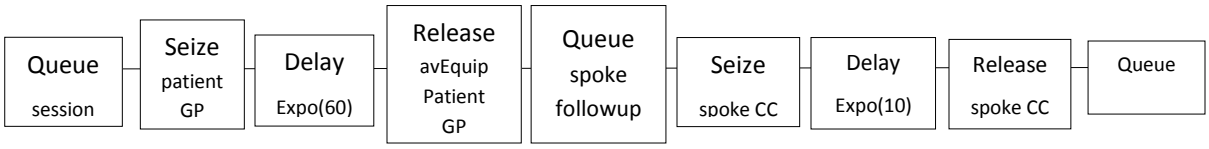


**Figure 4.4 – Appointment and immediate general practitioner follow-up**

Finally, the simulation captures the follow-up phase. Back at the hub site, the specialist reviews the information from the spoke site at their leisure. This involves watching a video of the whole question session and making recommendations for treatment, which would likely take 60 + expo(10) minutes. Figure 4.6, below, shows this part of the model.


**Figure 4.5 – The specialist's review of patient responses**

After the specialist completes their evaluation, the hub clinic coordinator sends any instructions, feedback, or follow-up information for the patients to the spoke clinic coordinator, a process that should take expo(10) minutes. Then the general practitioner/nurse takes expo(20) minutes to deliver this information to the patient and answer any resulting questions. Finally, the spoke clinic coordinator takes expo(5) minutes per patient to complete final wrap-up paperwork.

This could be via e-mail, phone, in-person pickup, or whatever medium is appropriate. Figures 4.6 and 4.7 show the end of the simulation.

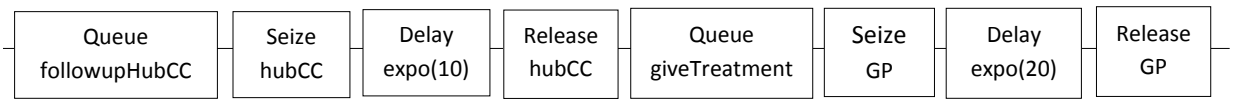


**Figure 4.6 – The hub CC sends the specialist's recommendations and the GP administers them**

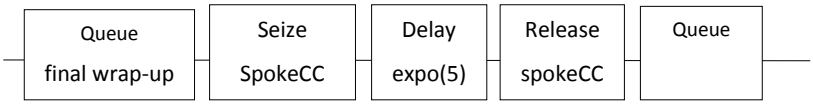


**Figure 4.7 – The spoke CC finishes up final paperwork for the appointment**

Following this, the simulation collects process time statistics, releases an "appointment" resource, decreases the number of appointments available, and disposes of the now-complete patient paperwork.

The table below summarizes the tasks that require completion in an asynchronous group telemedicine system, the times they need for completion, and the assumptions the model makes regarding them. Figure 4.8 illustrates this system in its entirety as it appears in Arena.

**Table 4.1: List of necessary actions for asynchronous group therapy**

| Action | Time | Actor | Assumptions |
|---|---|---|---|
| Referral Review | 10 min | Specialist | Occurs 25% as often as appointment execution. No patient is turned away. |
| Enrollment in Program/Session | 10 min | Hub Admin | |
| Gather Materials for Specialist | 5 min | Hub Admin | |
| Prepare Questions | 15 min | Specialist | |
| Send Questions to Spoke Site | 5 min | Hub Admin | |
| Check-In | 5 min | Spoke Admin, Patient | |
| Room Preparation | 10 min | Spoke Technology Specialist, Room | |
| Record Answers | 60 min | Patient, Nurse/GP, Room | |
| Send Video of Patients | 10 min | Spoke Admin | |
| Watch Patient Video and Prescribe Treatment | 70 min | Specialist | |
| Send Treatment to Spoke | 10 min | Hub Admin | |
| Distribute Treatment to Patients | 20 min | Nurse/GP | Does not require patient to be present; they can leave immediately after session is complete. |
| Final Check-Out | 5 min | Spoke Admin | |

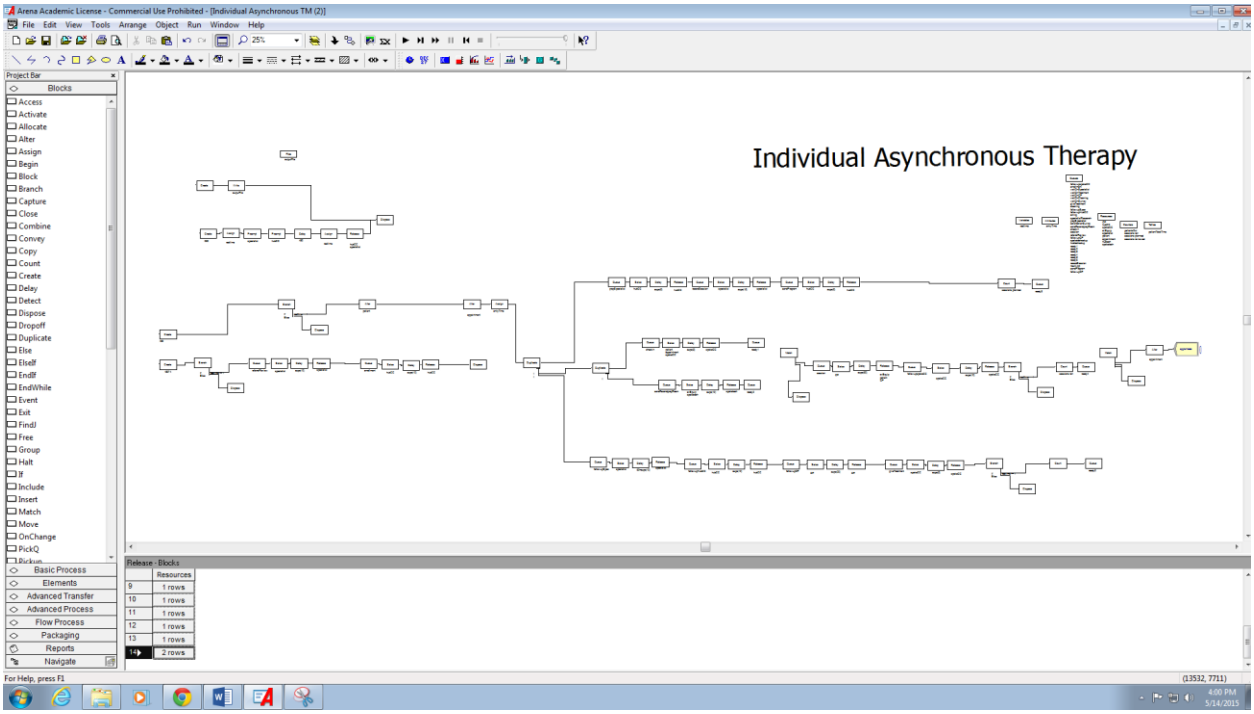


**Figure 4.8 – Individual asynchronous model as it appears in Arena**

## 4.2 – Asynchronous Individual Simulation Results

Running the model for a variety of enrollment rates at 200 replications each generated results similar to the synchronous model. Table 4.2, below, lists these results. Figure 4.9 summarizes them.

**Table 4.2 – Resource utilization for the asynchronous individual model**

| # patients | Spoke CC | Specialist | GP | Hub CC | SpokeTech | HubTech | AV Equip |
|---|---|---|---|---|---|---|---|
| **5** | 4.32% | 17.45% | 16.74% | 4.86% | 1.72% | 0.00% | 15.37% |
| **10** | 8.87% | 35.94% | 35.01% | 9.50% | 3.35% | 0.00% | 31.24% |
| **15** | 12.88% | 53.55% | 49.68% | 13.92% | 5.71% | 0.00% | 44.71% |
| **20** | 15.82% | 71.68% | 62.19% | 18.51% | 7.06% | 0.00% | 57.98% |
| **25** | 18.45% | 87.57% | 73.50% | 20.44% | 9.55% | 0.00% | 73.90% |
| **30** | 20.19% | 97.26% | 79.25% | 23.71% | 9.98% | 0.00% | 82.19% |
| **35** | 21.15% | 99.83% | 82.18% | 24.02% | 11.10% | 0.00% | 86.14% |
| **40** | 22.70% | 100.00% | 83.84% | 23.95% | 12.43% | 0.00% | 88.83% |

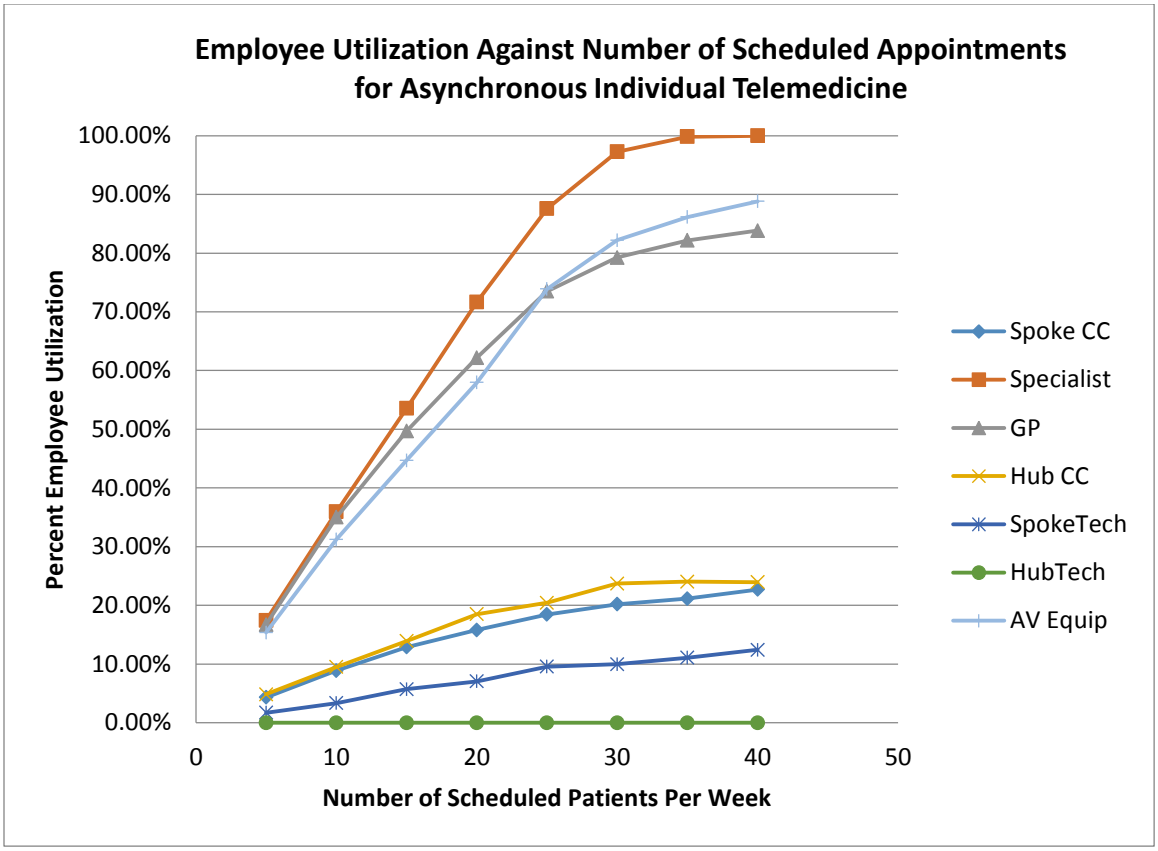


**Figure 4.9 – Employee utilization for the asynchronous group model**

As before, the specialist imposes the most severe bottleneck on the system. They reach their full potential at 25 to 30 appointments per week, or five to six per day. This appears slightly better than the synchronous model. Moreover, the other resources in the system have more free time at higher levels of enrollment. This provides them time to finish paperwork and offers them flexibility to spend a little extra time with patients to provide excellent customer service. While not substantial, the asynchronous system's additional flexibility offers some improvement on the synchronous model in terms of capacity.

# Section 5 - Simulating Synchronous Group Telemental Health Systems

The final modality of telemental health this paper will consider is group telemedicine. Structurally, this model functions similarly to the synchronous individual system aside from one key difference. Patients do not receive treatment one at a time but rather group and wait for a 20-person session to begin. The general principles of process flow indicate that batching prior to a process increases the overall variability in the system and hurts performance (Hopp & Spearman, 2008). However, group models also reduce the time specialists need to spend per patient, reducing the workload per patient of the bottleneck and potentially increasing capacity. The following section details how to simulate and examine the capacity and utilization of one of these systems.

## 5.1 – Synchronous Group Telemental Health Systems

The group model begins by creating 20 patients per class for a set number of eight-week classes. This contrasts with the day-by-day model used for individual telehealth. Group therapy sessions typically last for a set number of appointments, so this more holistic, eight-week model should generate more realistic results than the day-by-day model used for individual systems. As an example of how this system functions, it generates 100 patients to simulate five classes a week for eight weeks. This initial batch is the only time patients are added to the system; more do not trickle-in as time progresses. In other words, all the patients for that period enter the model at the start and just recirculate through the system over the eight weeks.

The patients' cases move to a "queue" block immediately after creation that represents their wait for the specialist to review their referral. In this process, the specialist takes time

following an exponential distribution with a mean of five minutes to determine if each patient is suitable for group therapy. For all the patients in a 20-person group, this roughly adds up to expo(100) minutes; at this stage, however, the simulation still processes patients individually. The team assumes that no patient is turned away at this point. After each patient's case "seizes" the specialist, "delays" them for five minutes, and "releases" them upon completion, the simulation uses an "alter" block to add one to the number of patients available to seize for appointments. Next, hub site staff must spend expo(5) per patient (for a total of expo(100)) minutes enrolling each patient in the eight-week program. The cases proceed to an "alter" block which adds eight to the number of "appointment" resources the simulation will attempt to complete. In other words, the simulation creates eight appointments for each incoming patient. Figures 5.1 and 5.2, display these processes.

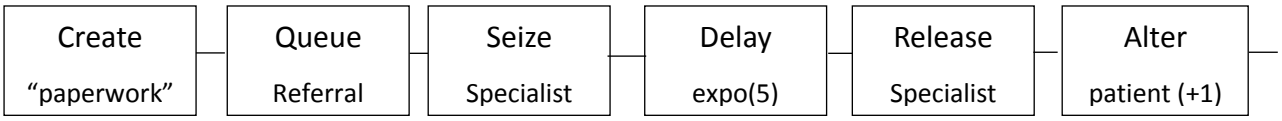


**Figure 5.1 – Patient/paperwork creation and program referral review**

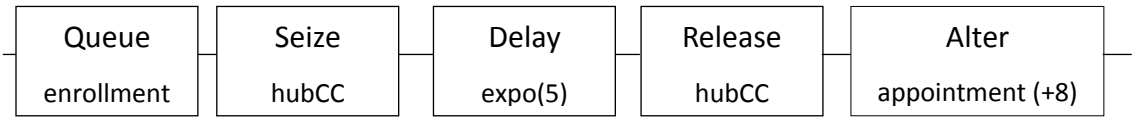


**Figure 5.2 – Program enrollment. HubCC refers to an administrator at the hub site**

The next point in the simulation represents the day of the appointment. It begins with a "group" block, which consolidates the 20 patients into a single unit. A "duplicate" block then creates four additional copies of the group, allowing the simulation to execute the next five tasks simultaneously.

Patients begin by checking in. This process splits the group back into its component patients, and for each one seizes a patient resource, spoke site staff, and an appointment resource. It delays them all for expo(1) (or a total of expo(20)), then releases the patient and spoke site

staff for other tasks. As before, the appointment resource remains seized until the completion of the appointment, preventing the same batch of patients from having two or more appointments in progress. When all 20 patients check in, the simulation groups them again, releases them, and sends them to the appointment.

The second copy of the original group goes to the spoke site technology specialist. The copy seizes both the employee and the room at the spoke site. The simulation delays for expo(10) minutes to model the time the spoke tech specialist needs to prepare the room at the spoke site and set up the connection to the hub site. This delay occurs once for the group rather than for all 20 patients individually. Upon successful setup, the employee may go, but the AV equipment remains unavailable until the completion of the group therapy session. A "delay" block which tells the spoke site technology specialist to wait 15 minutes before beginning precedes this entire process. This delay helps all the tasks occurring simultaneously in the simulation reach completion at about the same time.

The next group copy, after a 15 minute delay, triggers the hub site tech specialist to spend expo(10) minutes establishing the connection on the other side. Again, this occurs for the whole group at once.

The fourth copy travels immediately to a hub site staff person, who takes expo(5) minutes to gather up-to-date information on the patients (as a group) and transfer that information to the specialist. The specialist then takes expo(20) minutes researching and preparing for the upcoming group session. The simulation releases them briefly upon completion so they can attend to other matters prior to the session.

The fifth and final copy waits 15 minutes before instructing the general practitioner/nurse to spend expo(10) minutes setting up the room, distributing handouts, taking attendance, and

doing other preparatory tasks at the spoke site. The model releases them before the session begins. The simulation model of all five pre-session activities can be seen visually in figure 5.3.

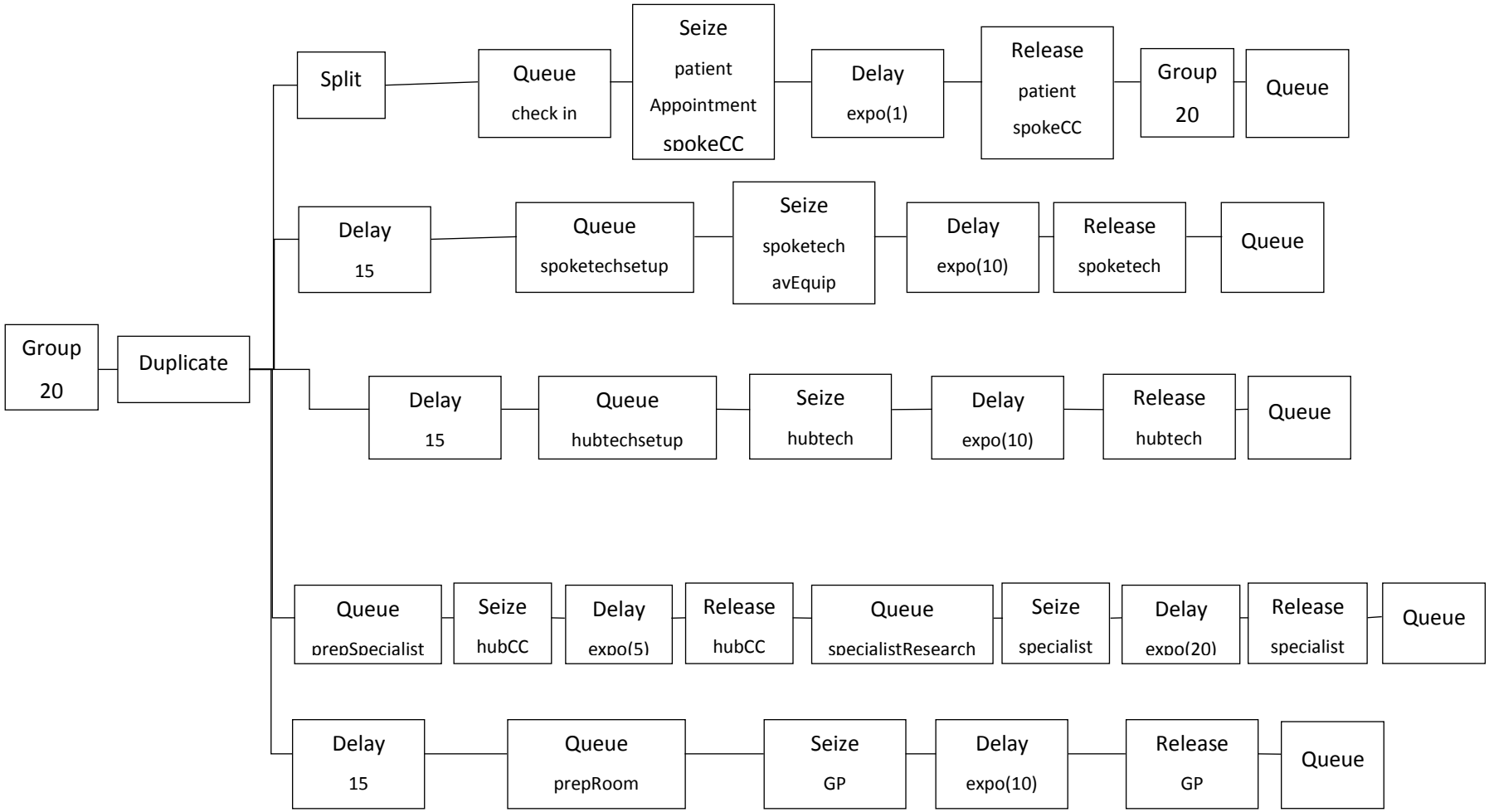


**Figure 5.3 – Simultaneous events that must occur immediately prior to a synchronous group session**

The five copies empty into queues which feed into a "match" block that waits for all five tasks to finish. Upon completion of all the tasks, the model lets all five copies continue, disposes the four clones, and sends the original to begin the session. The main event requires exactly 90 minutes of simultaneous time from 20 patients, a general practitioner/nurse (to sit in on the session and serve as an in-person point of contact for patients), and the specialist (virtually present through a teleconferencing or telemedicine service). At the end of the session, the simulation releases only the AV equipment; the people involved in the session have follow-up work to do. Figure 5.4, below, shows this simple part of the simulation.

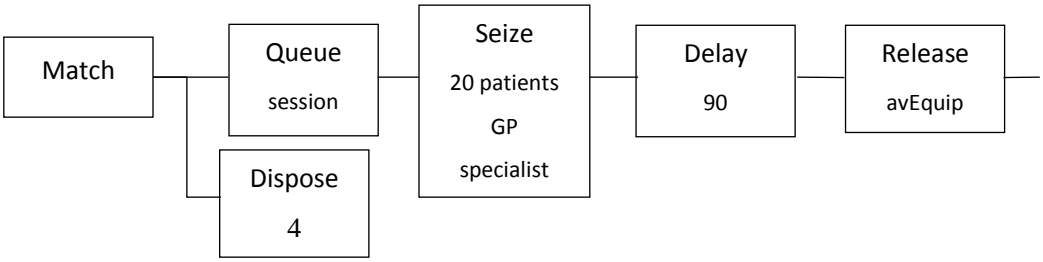


**Figure 5.4 – Group therapy synchronous session**

Upon completion of the session, the general practitioner/nurse takes expo(20) minutes after the session to talk to patients, answer questions, and collect feedback. They send their notes to the specialist and their role in that week's session concludes. A specialist then recommends treatment and prepares follow-up work. This takes expo(20) minutes. The simulation releases the specialist as they give their follow-up to hub site staff. Figure 5.5 illustrates this.

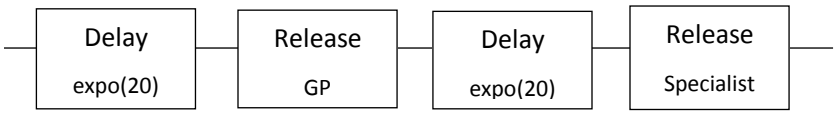


**Figure 5.5- Follow-up from the specialist and general practitioner/nurse**

This model lacks the post-session parallelism the individual model boasts. The less-intensive nature of group therapy minimizes question time and the group format requires the

specialist wait for the general practitioner's observations before finishing follow-up work. Either way, the simulation continues as the hub site staff takes expo(10) minutes to send follow-up to the spoke site. Finally, the group of patients reverts to 20 separate individuals, each of whom take expo(1) minute of a spoke site staff member's time to receive follow-up instructions. Figure 5.6 shows this logic.

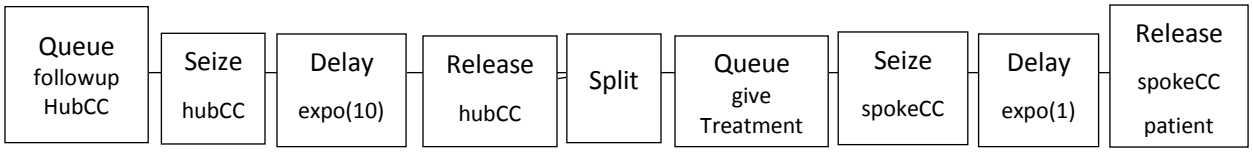


**Figure 5.6 – Distribution of specialist follow-up and treatment information.**

After this, the patients may leave. A "duplicate" block makes a copy of each patient and sends it to the "group" block at the start of figure 5.3 via a "delay" block that holds each patient for 2250 minutes. In more intuitive terms, after each appointment, patients wait a week (minus the time of the appointment itself) and go on to their next one. On the other side of the duplicate block, the simulation releases the appointment resource and permanently reduces the number of available appointments by one. Figure 5.7 shows this part of the simulation.

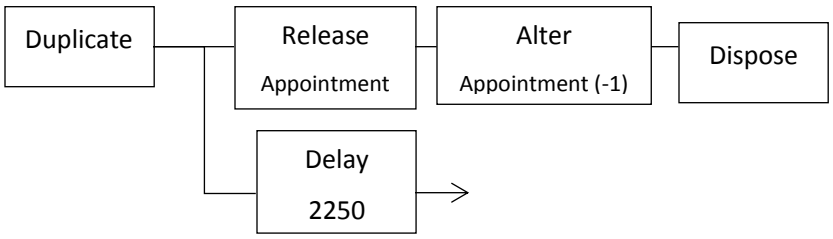


**Figure 5.7 -- Patient recirculation and appointment closure.**

The simulation itself runs for 19,200 minutes, or eight weeks (assuming a standard 40-hour work week). The simulation essentially creates all of its appointments and patients at the start of the eight-week period and attempts to immediately complete as many of them as

possible. The weeklong delay prevents the model from starting repeat appointments too early. The simulation tracks how many times each patient receives service during that eight-week period. Due to the nature of the simulation, appointments can occur over lunch breaks, weekends, and evenings. Hypothetically, a session could start at 4:30 PM one day and end the next at 9:00 AM the next, taking 90 minutes of work-time. While this assumption may result in mild *overestimation* of the number of appointments the system can complete, the difference is fairly small. Additionally, tightening this assumption makes the model needlessly complicated and makes measurement of utilization statistics problematic. The day-by-day individual models only face this problem for lunch breaks, making the impact even less pronounced.

The table below provides a brief summary of the system's components, listing the actions required to execute a telemedicine session and the times needed to do them. Assumptions relevant to each step are listed on the right. Figure 5.8 provides a holistic view of this system as it appears in Arena software.

**Table 5.1 – Tasks and their times in the group synchronous model**

| Action | Time | Actor | Assumptions |
|---|---|---|---|
| Referral Review | 5 min (100 total) | Specialist | Occurs once per patient. No patient is turned away. |
| Enrollment in Program/Session | 5 min (100 total) | Hub Admin | Occurs once per patient per session. Each session lasts for eight weekly meetings an can accommodate 20 patients. Contact is made with all patients the first time, and all accept treatment. |
| Gather Materials for Specialist | 5 min | Hub Admin | |
| Prepare for Session | 20 min | Specialist | |
| Check-In | 1 min (20 total) | Spoke Admin, Patient | Per patient, not per group |

| | | | |
|---|---|---|---|
| Room Preparation | 10 min | Nurse/GP, Room | |
| Tech Setup at Spoke | 10 min | Spoke Tech Specialist | |
| Tech Setup at Hub | 10 min | Hub Tech Specialist | |
| Execute Session | 90 min | Patient, Nurse/GP, Room, Specialist | No technical problems arise. In practice, technology specialists need to be on-hand to repair lost connections and fix hardware/software issues. |
| Follow-Up Talks With Patients | 20 min | Patient, Nurse/GP | |
| Follow-Up and Recommend Treatment | 20 min | Specialist | |
| Send Treatment to Spoke | 10 min | Hub Admin | |
| Distribute Treatment to Patients | 1 min (20 total) | Spoke Admin, Patient | Per patient |

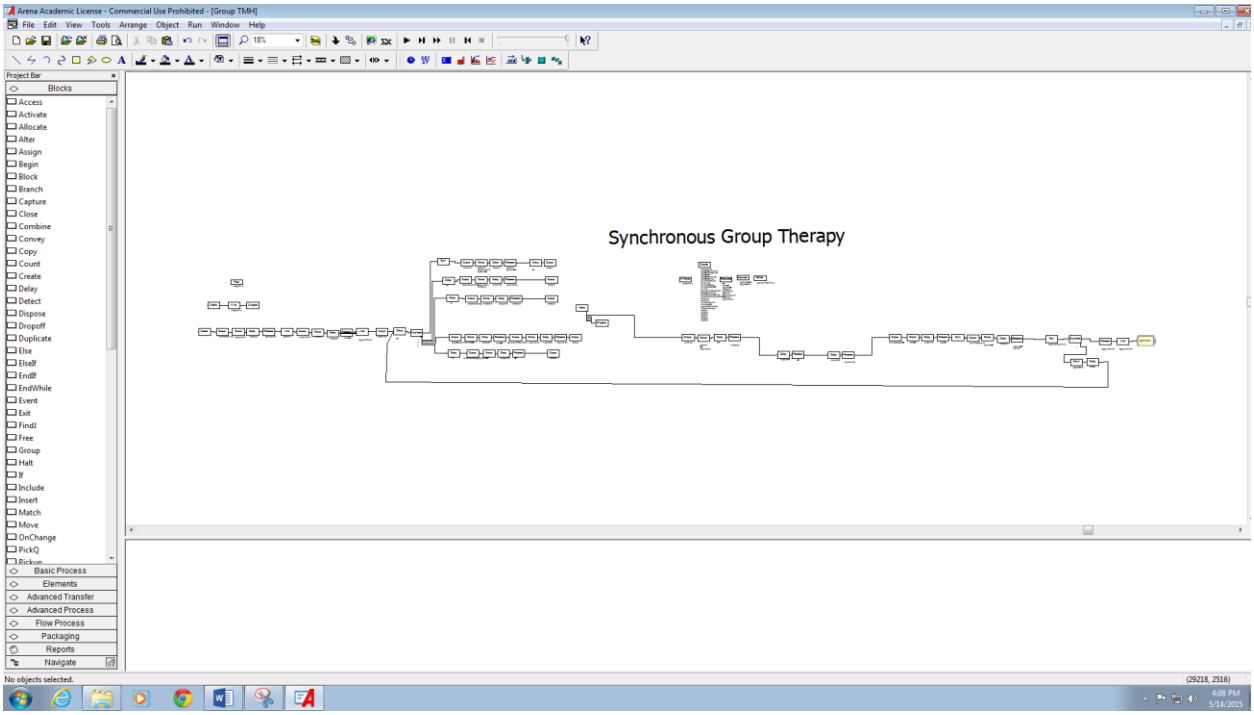


**Figure 5.8 – Synchronous group model as it appears in Arena**

## 5.2 – Synchronous Group Simulation Results

Like the other two models, the table and chart below show the percent utilization of different resources varies as the number of appointments per week increases. While the individual models track the total number of patients, this model tracks the total number of sessions. To find the number of patients served, one must simply multiply the number of sessions by twenty.

**Table 5.2- Percent utilization of employees for varying numbers of scheduled sessions in the synchronous model**

| # sessions | Spoke CC | Specialist | GP | Hub CC | SpokeTech | HubTech | AV Equip |
|---|---|---|---|---|---|---|---|
| **1** | 1.66% | 6.73% | 4.90% | 1.15% | 0.41% | 0.40% | 4.79% |
| **5** | 8.31% | 33.73% | 24.87% | 5.57% | 2.22% | 2.03% | 28.77% |
| **10** | 15.60% | 63.98% | 46.85% | 10.94% | 4.07% | 3.86% | 56.91% |
| **11** | 16.97% | 68.92% | 50.73% | 11.96% | 4.33% | 4.27% | 62.49% |
| **12** | 18.35% | 74.76% | 55.03% | 13.14% | 4.55% | 4.41% | 67.29% |
| **13** | 19.61% | 80.69% | 58.97% | 14.05% | 4.86% | 4.77% | 73.10% |
| **14** | 20.94% | 85.24% | 62.47% | 15.13% | 5.18% | 5.34% | 78.66% |
| **15** | 21.99% | 90.78% | 66.27% | 16.02% | 5.48% | 5.53% | 85.24% |
| **16** | 23.08% | 94.45% | 69.09% | 16.78% | 5.78% | 5.66% | 90.79% |
| **17** | 23.83% | 97.08% | 70.84% | 17.58% | 5.71% | 5.94% | 95.44% |
| **18** | 23.93% | 98.56% | 71.66% | 18.40% | 6.03% | 6.22% | 98.19% |
| **19** | 24.13% | 99.11% | 71.58% | 19.02% | 5.97% | 6.06% | 98.98% |
| **20** | 24.05% | 99.30% | 70.66% | 19.21% | 5.77% | 6.06% | 99.21% |
| **25** | 23.81% | 99.31% | 68.71% | 21.89% | 5.69% | 6.03% | 99.23% |
| **30** | 23.48% | 99.23% | 66.76% | 24.05% | 5.58% | 6.38% | 99.23% |

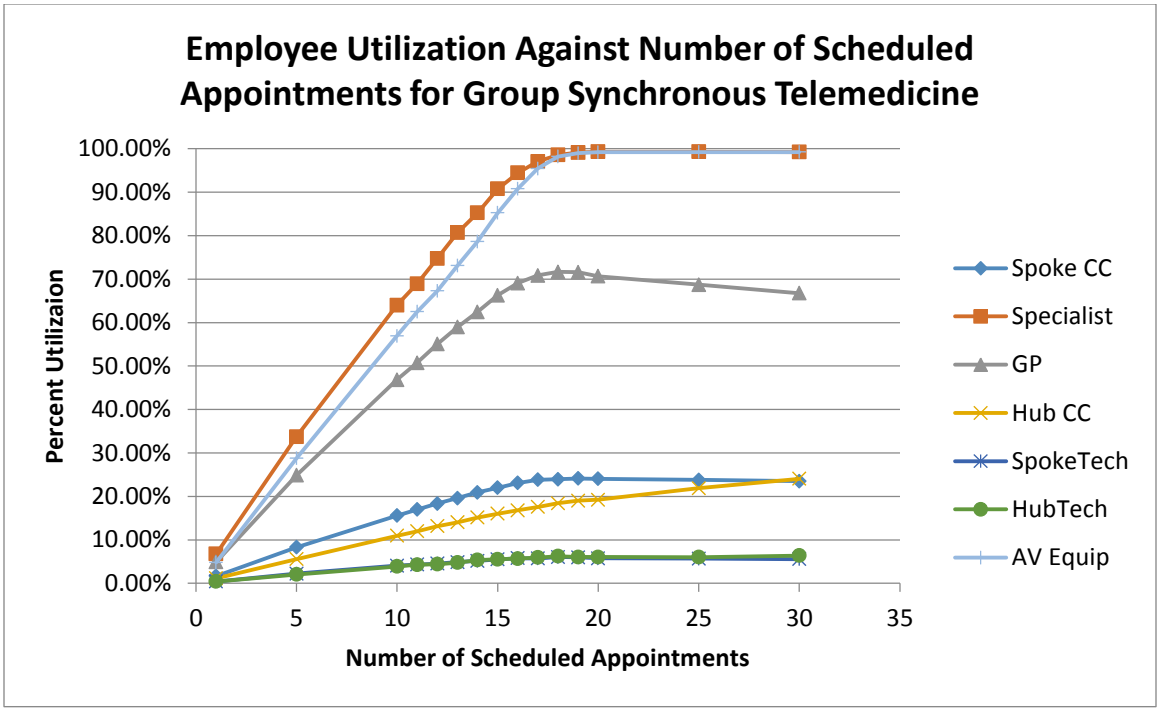


**Figure 5.9 – Employee utilization for synchronous group therapy**

The maximum reasonable capacity for a group model lies around 15 sessions per week, or 300 patients. This model can offer care to a much higher number of patients than the individual models; however, group therapy may not be appropriate for all patients. System choice should ultimately depend on the size of the provider and the demand of their patient base.

# Section 6 - Model Extensions

The three models discussed previously offer tremendous power in regard to modeling their respective systems. However, one of the main benefits of simulation modeling lies in their ability to easily test potential improvements to system configuration. This section discusses three possible changes to the models discussed previously and shows their impact on the system. The first anchors appointment times in the group synchronous model to account for latecomers. The second examines the impact of reducing the variance of service times in the individual synchronous model. Finally, the last shows the impact of increasing the number of doctors available in the individual synchronous model.

## 6.1 - Non-Floating Appointments for Group Synchronous Sessions

One issue with the initial models is that they allow the actual appointment time to "float". Rather than occurring at a set scheduled time, they start whenever everyone is ready to go. This assumption is not unreasonable for individual models, but is problematic for group sessions where people may arrive late.

To tighten this assumption, the simulation detaches the actual session from the preparatory and follow-up work. A "create" block triggers an appointment at some specified interval. This triggers two "preempt" blocks that seize a general practitioner and specialist. Stronger than a normal "seize" block, "preempt" actually makes the specialist or general practitioner drop what they are working on to execute the session. Upon completion of the 90-minute session, the specialist and general practitioner return to what they were working on and pick up where they left off.

While this solution more accurately models the sessions themselves, it does so at the expense of accurate precedence. Theoretically, the model could complete a session's follow-up

work without the actual session occurring. More likely, a session could begin without its associated preparatory work being completed or follow-up work could begin before the session starts. The model does include one failsafe. To move to the follow-up state, all preparatory work must be complete and the specialist and general practitioner must be available. A session cannot proceed to follow-up if there exists a session in progress, since the specialist and general practitioner are busy interacting with patients.

The table below and figure 6.1 show the results of this model when ran for 200 replications and varying numbers of sessions.

**Table 6.1 – Resource utilization for the set appointment time group synchronous model**

| # sessions | Spoke CC | Specialist | GP | Hub CC | SpokeTech | HubTech | AV Equip |
|---|---|---|---|---|---|---|---|
| **5** | 4.68% | 68.06% | 61.95% | 4.18% | 1.14% | 1.30% | 55.56% |
| **10** | 2.43% | 98.03% | 91.39% | 6.01% | 0.50% | 0.88% | 94.35% |
| **15** | 2.07% | 99.99% | 91.24% | 8.43% | 0.23% | 0.96% | 98.32% |
| **20** | 2.30% | 100.00% | 88.32% | 11.05% | 0.17% | 1.07% | 97.37% |

Notably, for any number of scheduled appointments greater than one per day, the simulation fails to complete even a single appointment per day consistently. While the appointments themselves certainly occur, the follow-up or preparatory work does not see completion, causing the simulation to fail to mark the appointment as done. Reducing the flexibility of the model in this way significantly hurts performance and reduces the maximum reasonable capacity to just 100 patients.

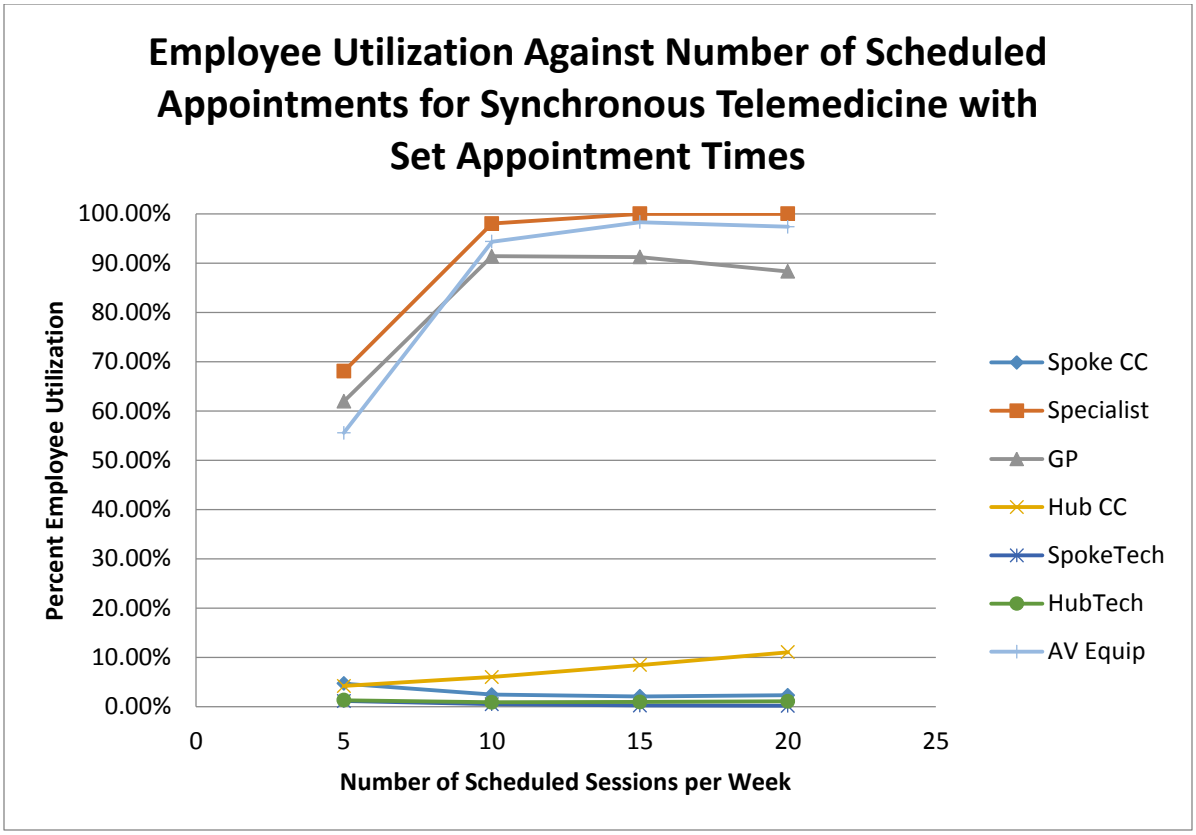


**Figure 6.1 - Employee utilization for the synchronous group set appointment model**

## 6.2 - Reduced Variance Individual Synchronous Model

The variance underlying the exponential distribution may be more than what is experienced in the highly scheduled world of health care. Adjusting the probability distributions that guide the timing of the different steps in the telemedicine process may result in improved system performance. The following changes are logical ways to reduce the variance of several steps in the model:

- The "referral review" and "enrollment" steps follow a normal distribution with a mean of ten minutes and a standard deviation of two minutes (abbreviated as norm(10,2)) each.
- Pre-session research by the specialist takes a normally distributed time with a mean of twenty minutes and a standard deviation of five minutes, instead of expo(20).
- The follow-up by the specialist likewise takes norm(20,5).

- The technology setup steps take norm(10,1).
- The "give treatment" and "prep specialist" steps take norm(5,1).

The table and figure below show the resource utilizations for 200, 480-minute runs.

**Table 6.2 - Resource utilization for the reduced variance individual synchronous model**

| # patients | Spoke CC | Specialist | GP | Hub CC | SpokeTech | HubTech | AV Equip |
|---|---|---|---|---|---|---|---|
| **5** | 1.98% | 23.00% | 14.11% | 5.15% | 2.09% | 2.11% | 18.87% |
| **10** | 4.06% | 43.88% | 28.34% | 8.21% | 4.16% | 4.19% | 36.58% |
| **15** | 6.02% | 64.70% | 42.40% | 11.30% | 6.25% | 6.27% | 54.31% |
| **20** | 7.80% | 85.49% | 56.36% | 13.58% | 8.35% | 8.35% | 73.88% |
| **25** | 9.18% | 99.66% | 64.32% | 17.35% | 10.47% | 10.47% | 96.97% |
| **30** | 10.17% | 99.94% | 60.90% | 18.75% | 10.50% | 12.51% | 99.98% |
| **35** | 11.40% | 99.98% | 58.79% | 20.05% | 10.40% | 14.63% | 100.00% |
| **40** | 12.11% | 99.98% | 57.76% | 20.37% | 10.33% | 16.67% | 100.00% |

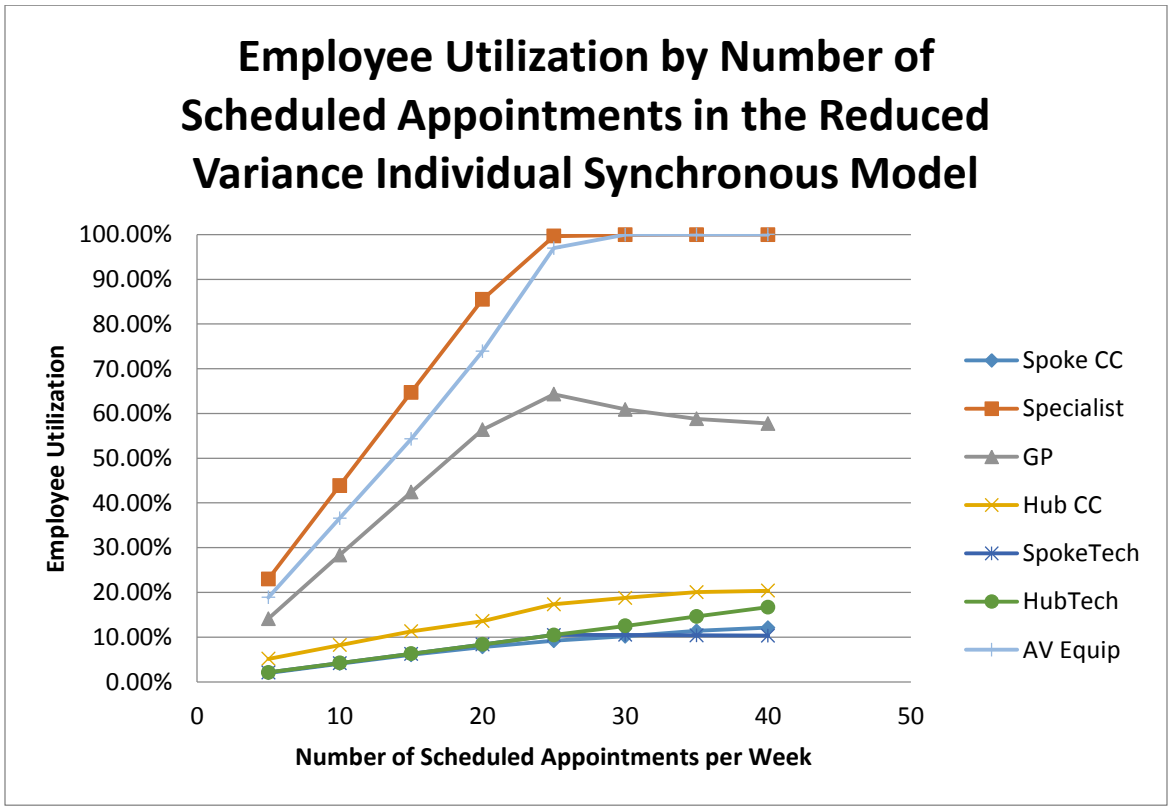


**Figure 6.2 – Employee utilization for the individual synchronous reduced variance model**

The model reaches capacity at approximately 20 appointments per week, or four per day. Surprisingly, this model performed more poorly than the original, despite the reduced variance. The likely cause of this comes from the specialist using extra time after short cases to complete follow-up and preparatory work. The results may also be simply a quirk of how the simulation software measures utilization. Either way, the model indicates that massive expenditures to reduce the variance of employee service times are not worth the cost.

## 6.3 – Increasing Doctors Individual Synchronous Model

The specialist, general practitioner, and room are consistently the most utilized resources in a telemedicine process. Increasing these resource levels certainly will increase capacity, but keeping administrative staff at the same level can pose its own bottleneck. A series of simulation runs, detailed below, reveal how many specialists and general practitioners a single set of administrators can handle, essentially forming an independent unit of people whose whole job is telehealth.

The simulation schedules 80 appointments per week to ensure that every employee has ample work to do. It assumes one patient walks in the door every 30 minutes. Based on the utilization results in table 6.3, for each two additional specialists and rooms, the simulation adds one extra general practitioner.

The simulation indicates that no matter how many specialists and general practitioners get assigned to the administrative staff, the administrative staff never reaches 100% utilization. In fact, the staff reaches their respective steady states (with a maximum of about 50%) with just three specialists and two general practitioners to work with. The specialist, room, and general practitioners, on the other hand, have steadily decreasing utilization as their numbers swell. The simulation reaches its maximum output of just under 12 sessions per week when five specialists

and three general practitioners are scheduled. The following table and figure show the findings of this simulation in depth.

**Table 6.3 – Utilization of employees as administrative staff are assigned more doctors**

| # specialists, # general practitioners | Spoke CC | Specialist | GP | Hub CC | SpokeTech | HubTech | AV Equip |
|---|---|---|---|---|---|---|---|
| **1,1** | 19.71% | 99.07% | 41.77% | 28.88% | 7.74% | 32.93% | 100.00% |
| **2,1** | 21.92% | 84.61% | 86.52% | 36.30% | 15.85% | 33.56% | 96.81% |
| **3,2** | 26.85% | 84.72% | 80.68% | 46.31% | 27.51% | 33.76% | 91.91% |
| **4,2** | 27.44% | 69.21% | 87.52% | 46.37% | 31.51% | 34.16% | 82.95% |
| **5,3** | 29.51% | 60.73% | 66.68% | 49.42% | 32.56% | 33.75% | 57.97% |
| **6,3** | 29.56% | 50.61% | 66.93% | 49.29% | 32.87% | 33.70% | 48.09% |
| **7,4** | 29.39% | 43.54% | 50.46% | 50.34% | 32.69% | 33.56% | 40.61% |
| **8,4** | 29.39% | 38.10% | 50.46% | 50.34% | 32.69% | 33.56% | 35.54% |

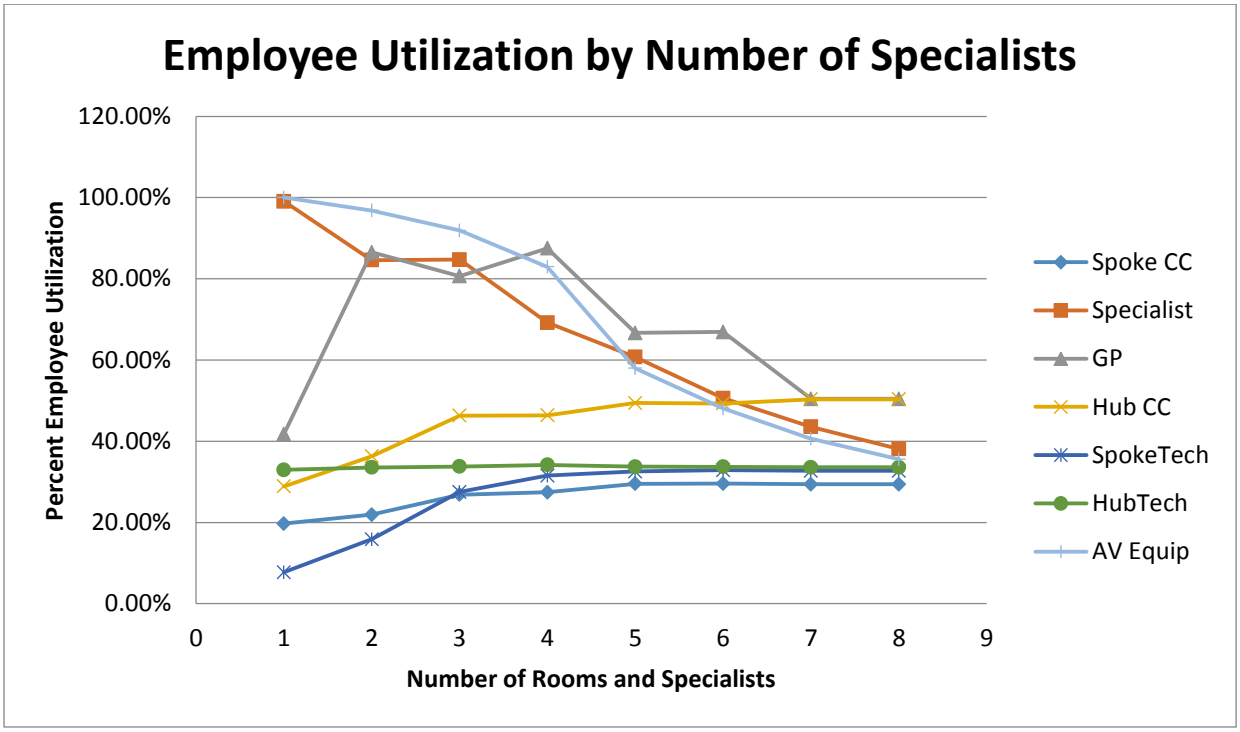


**Figure 6.3 – Employee utilization as administrative staff are assigned more doctors**

# Section 7 - Conclusion

Telemedicine offers unrivaled potential to solve some of today's most pressing challenges in care allocation and delivery. By expanding the geographic range of providers with specialized knowledge, telemedicine can positively impact the lives of hundreds of thousands of people in rural communities. Telemental health in particular can provide otherwise unavailable psychiatric and psychological services to distant patients.

Despite its potential, telemental health systems face several challenges. Telemental health systems suffer from a lack of large samples in many of its proponent studies, concerns regarding its clinical effectiveness in comparison to face-to-face consultations, and difficulties in implementation when supplanting existing systems. Beyond that, their complexity can make them challenging to understand and obfuscates the overall impact of changes within the system. They confound simple modeling tools and call for versatile techniques like discrete-event simulation.

Telemental health systems come in three modalities that each offer their own benefits and drawbacks. Synchronous individual, asynchronous individual, and synchronous group systems all feature different costs, different setup procedures, different levels of care, and different system dynamics.

Implementation of discrete-event simulations on Arena, from Rockwell Software, revealed the capacities, utilization, and overall behavior of each of these modalities. The simulations compared systems that included one of every type of resource and employee for varying numbers of patients. A synchronous individual system can handle roughly five appointments per day. An asynchronous individual system can handle five to six per day. A synchronous group model can accommodate three appointments per day, for a total of 60

patients. If sessions meet weekly, this means the systems can respectively enroll 25, 30, and 300 patients at any given time. As capacity increases, however, the amount of individual attention each patient can receive generally decreases as well. Care providers must balance the needs of the patient with the system's ability to provide one-on-one care.

Refinements of the models indicate that reducing the flexibility of group appointment times reduces the effective group capacity to two appointments per day. Further analysis also revealed that reducing the variance of treatment times in the individual synchronous model has no major impact. Finally, when scaling up systems, a single set of staff for a synchronous individual system can effectively handle three specialists and two general practitioners before requiring additional administrative assistance.

This field offers excellent opportunities for future research. Studies using more powerful modeling tools to describe telemedicine systems could potentially reveal general trends and best practices for configuring telemedicine systems. Moreover, the models described in this paper can be easily modified to test new process configurations, resource levels, or service parameters.

Discrete-event simulation provides a powerful tool for understanding the nuances of telemental health systems. Application of these models to a wide range of use cases can predict system capacity and bottlenecks without having to implement difficult changes. By understanding telemedicine systems, healthcare providers can use them to the best of their ability and extend psychiatric and psychological services farther than they have ever gone before.

References

Aeschliman, R. (2015). Modeling and analysis of telemental health systems with petri nets.

Butler, T., & Yellowlees, P. (2012). Cost analysis of store-and-forward telepsychiatry as a consultation model for primary care. *Telemedicine and E-Health, 18*(1), 74-7. doi:10.1089/tmj.2011.0086

Cegarra-Navarro, J. -., Sanchez, A. L. G., & Cegarra, J. L. M. (2012). Creating patient e-knowledge for patients through telemedicine technologies. *Knowledge Management Research & Practice, 10*(2), 153-63. doi:10.1057/kmrp.2011.47

Google. (2015). Map of CBOCs in New Mexico. Retrieved from https://www.google.com/maps/search/CBOC+in+New+Mexico/@34.1662325,-106.0260685,7z/data=!3m1!4b1

Grady, B., Myers, K. M., Nelson, E., Belz, N., Bennett, L., Carnahan, L., . . . Voyles, D. (2011). Evidence-based practice for telemental health. *Telemedicine and E-Health, 17*(2), 131-148. doi:10.1089/tmj.2010.0158

Hopp, W. J., & Spearman, M. L. (2008). *Factory physics* (3rd ed.). Long Grove, IL: Waveland Press.

Institute of Medicine. (1996). In Field M. J. (Ed.), *Telemedicine: A guide to assessing telecommunications in health care*. Washington, DC: National Academy Press.

Krupinski, E., Dimmick, S., Grigsby, J., Mogel, G., Puskin, D., Speedie, S., . . . Yellowlees, P. (2006). Research recommendations for the American telemedicine association. *Telemedicine and E-Health, 12*(5), 579-589. doi:10.1089/tmj.2006.12.579

Lach, J. M., & Vazquez, R. M. (2004). Simulation model of the telemedicine program. *Proceedings of the 2004 Winter Simulation Conference,* , 2012-17.

Lasierra, N., Alesanco, A., Gilaberte, Y., Magallón, R., & García, J. (2012). Lessons learned after a three-year store and forward teledermatology experience using internet: Strengths and limitations. *International Journal of Medical Informatics, 81*(5), 332-43. doi:10.1016/j.ijmedinf.2012.02.008

Laurant, M., Reeves, D., Hermens, R., Braspenning, J., Grol, R., & Sibbald, B. (2005). Substitution of doctors by nurses in primary care. *Cochrane Database of Systematic Reviews,* (2), CD001271. doi:10.1002/14651858.CD001271.pub2

Locatis, C., & Ackerman, M. (2013). Three principles for determining the relevancy of store-and-forward and live interactive telemedicine: Reinterpreting two telemedicine research reviews and other research. *Telemedicine and E-Health, 19*(1), 19-23. doi:10.1089/tmj.2012.0063

Odor, A., Yellowlees, P., Hilty, D., Parish, M. B., Nafiz, N., & Iosif, A. (2011). PsychVACS: A system for asynchronous telepsychiatry. *Telemedicine and E-Health, 17*(4), 299-303. doi:10.1089/tmj.2010.0159

Rabinowitz, T., Brennan, D., Chumbler, N., Kobb, R., & Yellowlees, P. (2008). New directions for telemental health research. *Telemedicine and E-Health, 14*(9), 972-976. doi:10.1089/tmj.2008.0119

Sang Goo, L., Mun, S. K., Jha, P., Levine, B. A., & Ro, D. (2000). Telemedicine: Challenges and opportunities. *Journal of High Speed Networks, 9*(1), 15-30.

Spaulding, R. (2010). Cost savings of telemedicine utilization for child psychiatry in a rural Kansas community. *Telemedicine and E-Health, 16*(8), 867-871. doi:10.1089/tmj.2010.0054

Tarakci, H., Sharafali, M., & Ozdemir, Z. (2007). Optimal staffing policy and telemedicine. *Association for Information Systems - 13th Americas Conference on Information Systems, AMCIS 2007: Reaching New Heights, 1*, 226-232.

von Wangenheim, C. G., von Wengenheim, A., Hauck, J. C., McCaffery, F., & Buglione, L. (2012). Tailoring software process capability/maturity models for telemedicine systems. *18th Americas Conference on Information Systems 2012, AMCIS 2012, 3*, 2472-2480.

Yellowlees, P., Odor, A., Parish, M., Iosif, A., Haught, K., & Hilty, D. (2010). A feasibility study of the use of asynchronous telepsychiatry for psychiatric consultations. *Psychiatric Services, 61*(8), 838-40. doi:10.1176/appi.ps.61.8.838

Yellowlees, P., Odor, A., Patrice, K., Parish, M. B., Nafiz, N., Iosif, A., & Hilty, D. (2011). Disruptive innovation: The future of healthcare? *Telemedicine and E-Health, 17*(3), 231-234. doi:10.1089/tmj.2010.0130

Zhang, S., McClean, S. I., Jackson, D. E., Nugent, C., & Cleland, I. (2013). Patient satisfaction evaluation of telemedicine applications is not satisfactory. *XIII Mediterranean Conference Onf Medical and Biological Engineering and Computing, 41*, 1140. doi:10.1007/978-3-319-00846-2_282