PROCESS ANALYSIS IN HEALTH CARE

Introduction*

In order to improve the performance of a system that delivers a product or service we need ways to describe a system and its performance that the agents involved will quickly recognize and understand. After the industrial revolution many of the earliest efforts to systematically do this was done by Industrial Engineers (IE). Over time the field of Operations Management (OM) developed by merging this work with additional insights from the field of Economics. Since the early 20th century experts and scholars in OM have collaborated with practitioners and scholars in health care to extend the use of engineering and OM tools to improve the performance of systems that deliver health care services.

The chief objective of this note is to introduce a few of the resulting ideas developed to look at delivery systems involving health care processes. More specifically we aim to introduce elements of Process Management and Process Analysis to create a vocabulary and framework to look at these critical systems in a way that facilitates their improvement. We begin with a brief collection of terminology that we shamelessly borrow from early works in IE and OM.

While these common terms are often not a perfect fit for health care settings, for the most part we will stick with this terminology so that users will find it easy to find supporting documentation in articles and textbooks as needed. However, at each point we will work to position these rather "generic" terms in the health care context.

With this vocabulary in place we consider the most common measurements used in Process Analysis. We close with a few examples to illustrate how these ideas can be applied in a variety of health care settings.

Basic Definitions

Flow Unit: For our purposes, this is the discrete entity that is being altered by the process. Many textbooks and articles use the term "job." This is most likely to be a physical item in production settings, or a customer in service settings. However, it may be a "virtual" item such as a file or data packet. It may also be some other manifestation of the process such as an invoice, application, or collection of records or data points.

Note: One key distinction between production settings and service settings is that in almost all service settings the flow unit holds a special relationship to the customer. In some cases the flow unit is the customer's information. In other settings it is the customer's property. The hardest systems to manage are those in which the flow unit is the customer him/herself, which is often the case in health care.

Flow Time: This is the amount of time each flow unit spends in some segment of a

* Revised May 2019. Chester Chambers prepared this note as a basis for classroom discussion. Please do not duplicate without authorization.

process. This includes processing time for the unit as well as any time spent between steps (waiting time.) In other words, this is the total time spent between 2 defined points in a process. These two points may surround one or more process steps.

Note: Note that a Flow Time may be calculated between any two points in a process. This may span the entire process, or it may only focus on some smaller part or segment.

Cycle Time: When a Flow Time is found using the beginning and the end of a process to define the relevant segment we will call this special case the Cycle Time **(CT)**. Improving CT is a very common objective. Just as with any flow time, Cycle Time includes processing time, waiting time, rework time, inspection times, and transportation times as long as these events take place within the process of interest.

Note: For our purposes the Cycle Time will often be the duration between patient arrival and exit from the system. In some contexts this will be labeled Length of Stay (LOS). However, in other settings the definition can be quite different. Many payers are concerned with the span of time from first visit for a condition and the end of the last follow-up visit. Thus the Cycle may involve multiple episodes of care. The key point is to define Cycle Time relative the process of interest and the goal of the analysis.

Process: A set of "activities" or steps that accept one or more inputs including some flow unit, transform the flow unit in some way, and produce one or more outputs. It is often useful to consider the activities as being the actions of "resources" such as machines or workers. The chronological sequence of activities is specified in the job's "route" through the process. Thus we can say that a process is composed of routes along which units are being transformed through activities that are performed by resources.

Note: The concept of a "route" can be quite complex in health care. Patients in a maternity ward and patients in an Oncology unit will follow completely different routes through the system. The key point is that for a clearly defined population of cases we should be able to specify a series of steps that we anticipate happening when a member of the relevant population arrives.

Throughput: The rate at which flow units actually move through the process. This may be expressed as units per hour, per clinic session, per day, etc.

Note: Keep in mind that throughput is a rate. Consequently, counting the units of analysis along with the durations of time involved are both necessary to define this value. In addition, we are typically concerned with the realized rate – not simply what we wish the rate to be or what it "should" be. In fact much work to improve health care processes boils down to getting the realized throughput rate to approach its theoretical maximum.

Takt Time: The minimum time that can be sustained as the duration between the completion of successive units. It is directly related to the throughput rate. For example, a process with an output rate of 4 units per hour has a takt time of 60/4 = ¼ hr or 15 minutes. Thus Takt Time can be viewed as the inverse of capacity. [Takt Time = 1/Capacity]

Note: We must emphasize the term "sustained" in this definition. In almost any process involving highly skilled labor it may be possible to "rush" a job or two to shorten this time for a while. The key question is often whether this can be sustained for an

entire shift or clinic session. Thus the manager must be careful not to mistake a special case for a useful measure of takt time. Consequently, measurement across multiple cases, sessions, or even weeks is needed to find this value in practice.

For most (but certainly not all) settings of interest here the takt time will be the longest busy time per flow unit for a resource involved in at least one (but often more) activities in a defined process. It is often useful to keep the following fact in mind: if Resource **A** is busy for *y* minutes for every patient, then it does not matter how many other resources we bring to bear – throughput can never exceed 1 unit every *y* minutes unless we do something that affects Resource A.

Bottleneck: The production resource that limits the capacity of the overall process. This is usually the resource with the lowest overall capacity. If each resource is assigned to 1 and only 1 step, then identifying the step with the longest Flow Time also identifies the bottleneck resource. Note that if a resource is involved in multiple steps, then the bottleneck resource may not be synonymous with the longest task. The bottleneck resource may be labor, space, equipment, or material.

Note: It seems safe to say that this is the term on this list must likely to be misused. No matter how many times we stress that the longest step is not necessarily tied to the bottleneck resource we seem to instinctively gravitate to focus on the longest step when we want to speed up a process. As a consequence, it is often easier to focus on the resource with the longest busy time per flow unit. Thus, the discovery of takt time and identification of the bottleneck resource are closely related, and it becomes practical to say that the resource associated with the takt time is the bottleneck resource.

Capacity: The maximum rate of output of a process measured in units of output per unit of time. In other words Capacity is the maximum sustainable level of throughput. The unit of time may be of any length: a second, minute, hour, shift, day, week, quarter, etc. We typically find that Capacity = 1/Takt Time.

Note: Given the definitions in place here it should be clear that to change capacity, we must do something that affects the bottleneck resource. As you work with more complex systems, always keep in mind that the actions of any resource that causes the bottleneck to wait or be idle reduces capacity. For example if a surgeon is waiting for a room to be cleaned or a patient to be moved we are sacrificing capacity. This often happens as part of some effort to reduce cost by focusing on the reductions in the level of some relatively inexpensive resource without recognizing the impact of such changes on the utilization of other resources.

Setup Time: The fixed time involved in processing a job or batch of jobs. Many settings involve time spent preparing to perform a particular task or operation involving a Flow Unit. However, some processes will also have set actions that must take place during or even after a process step, such as data recording, or error checking. Even though these actions do not literally "set up" the job, they may still be fixed amounts of time associated with a job and therefore are best characterized as Setup Time.

Note: Many health care settings have setup times that are quite literal such as positioning a patient for surgery or an x-ray. Other setup times are more subtle and may include actions such as pre-visit charting or post-visit

dictation. Many efforts to analyze patient flows miss these times because they are out of sight or outside the clinic walls. Any setup time that a process needs must be accounted for when calculating Flow Times, Cycle Times or Capacity.

Lead Time: This is identical to Cycle Time when there is no delay between receiving a job and initiating it. However, some settings include a delay between these points in time. In such cases Lead Time adds this delay to Cycle Time and is therefore greater.

Note: The concept of Lead Time is critically related to access to care. In some systems the time between the request for an appointment and the actual visit can be days, weeks, or even months. Thus managing Lead Time may prove to be even more important than managing Cycle Time.

Work-in-Process (WIP): The number of Flow Units within the process boundaries at any given moment in time. This includes units being worked on and those waiting in the system. These units are sometimes referred to as Work-In-Progress.

Note: For many settings of interest here this value will be the same as the Census. For example, when considering an ICU, it is critically important to know how many beds are in use at any point in time.

Turnover: The rate at which all of the Work-in-process in the system is replaced. Thus, Turnover is the inverse of Cycle Time. For example if the Cycle Time in a system is 2 hours, then we expect the WIP turnover to be one turn every 2 hours. Stated differently, all of the WIP in the system right now is "replaced" with new jobs over the next 2 hours, and the Turnover can be stated as once every 2 hours or ½ per hour.

Note: This term is of special interest to those managing "beds" or other critical resources such as examination rooms. One common misconception is that increasing turnover is the same as increasing capacity. This is generally true if resource levels are held fixed but may not be true in more general cases. For example, compare a system with 2 examination rooms and 60 minute cycle times to one with 1 room and a 45 minute cycle time. After 3 hours the first system has processed 6 patients (2 per hour * 3 hours = 6 patients). After 3 hours the second system has processes 4 patients (4/3 per hour * 3 hours). This is part of the reason that we typically want more than one examination room per provider even though it may increase cycle time.

Utilization: Ratio of the amount of a resource actually used over the amount of that resource available. Equivalently, it is the amount of capacity to do work that is actually used divided by the theoretical maximum capacity to do work that is available. Labor utilization is most commonly expressed as actual time spent processing flow units divided by the total time available.

Note: Do not confuse utilization with productivity. Keeping a resource busy doing work that could be done by a less expensive resource (even though that resource is slower) will almost surely increase utilization of the more expensive resource, but may be counter-productive. The large gaps between the cost rates for medical personal makes this distinction even more salient. One painful lesson for many new managers is that a system with 100% utilization of the most expensive resource is almost certainly not the most productive system.

Process flow diagram: A diagram depicting the activities and flows between them. Most process flow diagrams focus on the physical

movements of flow units in a system. However, information processing may be depicted as well.

Note: In our work, we have found that it is best to begin with very simple diagrams. Many students will argue that such diagrams are incomplete, and they are certainly correct. The fact is that all process diagrams are incomplete. The key is to identify the steps most critical to the process metrics of primary interest. Ultimately, this list may prove to be much shorter than one might think.

Batch size: The number of discrete units that are processed or held as a group. In many cases this is the number of units processed before a unit of a different type can be worked upon. A common rationale for using batch sizes greater than one is that changing from one job type to another requires a new setup to be performed. Again, if this setup is literal, such as positioning a patient for a procedure, then the batch size will almost always be 1. However, in other settings it is advantageous to deliver process outputs as a group. In these instances batch sizes tend to be larger.

Note: Some are surprised that batch processing actually occurs in health care settings. However, we have seen it arise many times in cases involving items including x-rays, lab tests, or drug packaging. Mixing tasks that are done using batch sizes of 1 such as an examination with other tasks that are delivered in larger batches such as delivery of a collection of x-rays can complicate the analysis greatly.

Process Management

The design, management, analysis, and improvement of a process is called Process Management. For our purposes here we have particular interest in three distinct phases: process description, process analysis, and process improvement. These steps are repeatedly undertaken as parts of ongoing Process Management.

1) Process Description: This phase consists of defining and understanding a process, and evaluating its performance. A clear definition of the process is developed by identifying the ownership and purpose of the process, its scope and boundaries, its external customers and the outputs they receive, as well as its suppliers and the inputs that they provide. Customer requirements are identified and the current performance of the process, from both internal and external perspectives is measured. The main goal of this phase is to identify the processes that are performing least well, and that need the most, or the most immediate improvement, so that they can be analyzed further and improved.

   Note: In most of our examples we will focus on only one or two dimensions of performance such as waiting times, capacity or cycle times. Other dimensions of performance may be relevant including patient satisfaction, financial performance, or most importantly health outcomes.

2) Process Analysis: After process description identifies each step, its customer(s), and the related inputs and outputs, this information is typically displayed in a process flow chart. If no specific issue has already been isolated, typical concerns at this point include:

   - Calculation of process capacity
   - Identification of options to increase capacity
   - Calculation of cycle times and/or lead times

- Identification of options to decrease those times
- Calculation of the costs involved in delivering the service
- Identification of means to reduce that cost

Whenever we ask a process manager why the process is not performing as well as we would like, the response is almost invariably "we don't have enough *y*" where *y* some resource. Anticipating this type of response, it is almost always a good exercise to establish what capacity actually is.

Consequently, identifying the bottleneck resource is often the first order of business. The bottleneck resource dictates process capacity, and any effort to change capacity must do so by having some effect on the way the bottleneck functions. In addition, since it is natural to believe that using more resources leads to more output, it is quite common (but not guaranteed) for the bottleneck resource to be the most expensive one as well. Consequently, efforts to manage costs will often center on the productivity of the Bottleneck resource.

When considering the cost of the process we typically focus on the spans of time the each resource is kept busy by the process and some measure of the cost per unit time for those resources. (We will consider this issue in more detail in another reading). In addition, cycle times and delays are often very important because we want to reduce the time between identification of a problem or ailment and some approach to address it. To account for all of these elements we must become familiar with the resources involved and the value added at each step so that any steps that do not add value can be eliminated or improved. The process may sometimes be benchmarked against similar processes at this or some other unit.

Finally, possible actions for performance improvement are developed, and the feasibility of these actions may be reviewed with process participants and customers to develop a plan for process improvement.

Note: At this stage of our work we will typically discuss processes as though activity durations are deterministic. Once these tools are mastered, additional tools can be brought to bear to more formally account for variability. However, we routinely find that analysis based on average values is a good starting point. If a system shows problems before variability is included, it surely will have the same problems (and more) once variability is added.

3) Process improvement: In this phase, the proposed improvement is implemented (on a pilot basis if appropriate), feedback on results is obtained, and an implementation plan may be developed or revised as needed prior to system wide adoption. The process performance should then be continuously monitored and controlled to facilitate ongoing improvement.

Goals of Process Management: Having a clear understanding of the objectives of the organization is an essential prerequisite for conducting process analysis. Otherwise, we do not know what specific aspects of the process need to be analyzed and/or improved. As a general rule "Making things run better" is a concept that everyone welcomes, but it is a little too vague to be of much use as a goal for analysis. This phrase also fails to

adequately capture the idea that for any analysis to be useful it must balance competing concerns. Some agents will have multiple concerns, and different agents interacting with the process can have very different ideas about what "running better" actually means.

For example, if all other things are equal, lower waiting times are preferred. This is particularly true from a patient's perspective. Of course it is also true that if all other things are equal, most patients feel that more face time with the physician is also preferred. The obvious catch is that increased face time essentially guarantees that throughput will be reduced and waiting times will rise. Thus, even if we only consider users of the system, we have competing objectives that must be balanced.

Given the economic realities facing health care providers today, a reduction in delivery cost holding all other things equal is preferred. Again, if something is done to reduce cost, some other aspect of performance will also be affected. If resources are viewed as having a cost per minute, then reducing time spent on each flow unit should reduce cost. However, there will also be some relationship between time spent and the quality of the experiences created for both the patients and the providers. Thus we have competing objectives even if we only consider managers of the system and tradeoffs must be recognized to guide our efforts.

If we are looking for a good place to start in thinking about this collection of issues, it is quite often advisable to consider time as the primary metric of interest. Obviously, waiting times are easy to calculate and all will agree that less is preferable to more. If reducing or controlling cost is important and labor involves a significant cost per unit time then minimizing cycle time or lead time is likely to be an attractive goal. If the objective is to increase the use of a system and users are sensitive to lead times, then a focus on such times may be a good place to start. When other issues such as quality are an obvious concern, reducing the times between when an error occurs and when it is found, and between when an error is found and when it is corrected, will almost always leave us in a better position. In all of these instances, a focus on measurements of time is a natural starting point.

Three Key Process Measures: Throughput, Work in Process, and Cycle Time

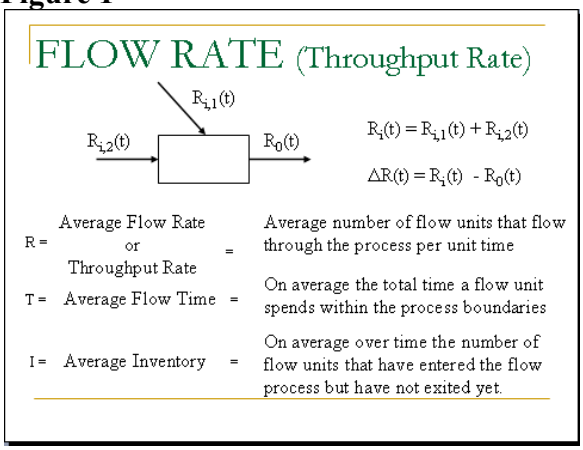
The definitions provided so far are listed one at a time. However, it should be clear that many pairs of metrics are related to each other in rather formal ways. Takt time is related to capacity, Cycle time if related to turnover, etc. It turns out that three particular values are related in a critical way as well.

We often find it useful to speak in terms of Flow as in "we want to improve patient flow." Let us be a bit more formal on this point. When we consider the flow rate at a specific point in time, we call it the instantaneous flow rate and denote it by R(t). For example, if we focus on the flow through entry and exit points of the process at time t, we can denote the instantaneous total inflow and outflow rates through all entry and exit points, respectively as $R_i(t)$ and $R_o(t)$.

Figure 1 shows a simple system with 2 inflows and 1 outflow at time t. Hopefully, it is clear that whenever the total inflows are greater than the total outflows over any span of time, the WIP is rising, and whenever the total outflows exceed the total inflows, WIP is falling.

A stable process is one in which the average inflow rate is the same as the average outflow rate. When we have a stable process, we can refer to either the average inflow or outflow rate as the average flow rate or the Throughput of the process. For ease of exposition, we generally refer to Throughput simply as **R** since it reflects a balancing of rates over time.

Figure 1



We have already noted that WIP is the number of flow units within the process boundary at any given time. Even if the inflows and outflows vary over time, if we have a stable system it is useful to think of the average WIP in the system simply as **I.**

For most settings discussed here we are particularly interested in the average Cycle Time of the system or process. We note that if the system is stable in the sense that the inflow and outflow rates are nearly equal over some reasonable amount of time, then the average Cycle time, **CT** will be stable as well.

Note: These terms are rather generic, but let us underscore the connections to health care settings again here. For many cases of interest to us Throughput will be synonymous with the rate of Patient Arrival, WIP will be synonymous with the Census, and Cycle Time will be synonymous with Length of Stay. These terms are simply setting specific expressions of the same ideas.

Little's Law

With these definitions in place we are able to state a major result describing the connections among them known as Little's Law. Even though the pattern of flows through the system and the corresponding inventory levels can vary greatly over time and can be extremely complex we can use this relationship to gain some insights regarding system behavior. Specifically, Little's Law states that,

$$\mathbf{WIP = R * CT}$$

Again, **WIP** will often be the census in our examples; **R** refers to the throughput rate (patients per session, claims per hour, phone calls per shift, etc.), and **CT** is cycle time which for us of often Length of Stay.

Hopefully it is clear why this relationship must hold for a stable system. If arrival rates are always greater than exit rates, the system will accumulate an infinite amount of WIP which is not physically possible. The only way that the outflow rate can always be greater than the inflow rate is to be left with a negative number of flow units in the system which is also impossible. Consequently, over some reasonable time frame the inflow rate and output rate must equal Throughput = **R**. The average time in the system is **CT**, and the average census must be the product of these two values, thus **WIP = R * CT.**

One useful analogy for Little's Law is to think of it as a manifestation of the "conservation of mass." Something is flowing into a vessel (our system) and something is flowing out. The differences

between what came in and what went out must be what remains within the vessel. If a system begins the day empty, patients arrive at a rate of 10 per hour and stay in the system for 2 hours, then it is easy to see why the average census must be 10 per hour * 2 hours = 20 patients.

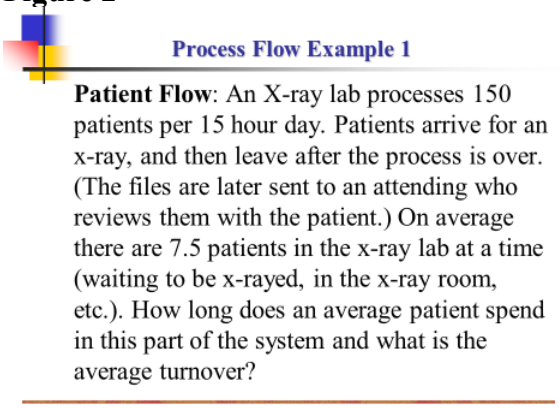
Another analogy is to recall the old lesson from physics that "Rate times Time Equals Distance." In Little's Law we have a Throughput Rate, a Cycle Time, and a Census or inventory level that is analogous to distance. Thus Rate * Time = Distance becomes **R * CT = WIP.**

For settings of interest here, the power of Little's Law often lies in the fact that its application allows us to "bend" the definitions. For example, we may have a pharmacy that sees 20 customers per hour on a typical day between 8 and 11 AM, but then receives 50 customers per hour around lunch time from 11 AM to 1 PM.

This system may be fairly stable during the first block of time, while its WIP inventory (meaning the customers in the system) is likely to be growing rapidly during the second block (between 11 AM and 1 PM). It is also common for the arrival rate to be sharply lower for later hours, say from 1 – 4 PM and then to pick up again when customers get off of work after 4 or 5 PM. Clearly this system is not perfectly stable from minute to minute or hour to hour. However, it is still useful to think about average rates during low and high demand periods, as well as average rates over periods in which the system starts and ends in a similar state.

To generate additional insights into the usefulness of Little's Law let us consider a few examples in more detail.

Figure 2



Note that 150 patients per 15 hour day must be 150/15 = 10 patients per hour. This system is empty at the start of the day and is emptied at the end. 150 patients flow into and out of this facility on a typical day. Thus it is useful to define the Throughput **R** as 10 patients/hr. This does not guarantee that exactly 10 patients arrive every hour, or any hour for that matter. We also see that we are thinking of the flow units as patients in this example. In other instances it would be more useful to think about the x-rays being prepared because some patients will have many more x-rays than others. Thus we see that the flow units may differ for the same system depending on the focus of the analysis.

Given this context the average **WIP** level is the average number of patients in the clinic. This is given as 7.5. Again, this does not mean that there are always 7.5 patients in the clinic, since we hopefully do not have any half bodies walking around our hospital. This is only an average, but it is useful to consider **WIP** = 7.5.

We would like to know about how long the average patient spends at the facility in such cases. Little's Law tells us that

WIP = R * CT.

This clearly implies that **CT = WIP/R** and we can estimate this time as,

$$\frac{7.5 \text{ Patients}}{10 \text{ Patients/Hr}} = 0.75 \text{ Hr} = 45 \text{ min}$$

Students often ask, why can't I simply count or measure each of these things directly, or why do I need an equation when I can simply count? Part of the answer to this question is that our data based on counting is often not nearly as reliable as we would like to believe. For example, it is common to have check-in times for patients in a clinic that are fairly reliable. However, for patients that do not schedule a follow-up visit, they sometimes simply leave the system after a visit. In this case Checkout times are not reliable and direct measurement of cycle times breaks down. If we are willing to assume that the arrival rate equals throughput then we may be able to periodically count patients in the system to estimate average WIP and find CT indirectly using Little's Law.

Data regarding the census can be more complex than one might think. Counting bodies in examination rooms is a good estimate of WIP for those rooms but may not be such a good estimate if we consider the clinic as a whole because many bodies in waiting areas are not patients. They may be spouses, parents, friends, etc. In these cases we may need a different way to estimate this value. If check-in and check-out times are reliable then we can figure out throughput and CT values, and use Little's Law to fill in the gap about WIP.

In other cases, patient no-shows and lateness distorts our measurement of arrival rates. In these instances it may be useful to have a different way to account for it. Let us consider another simple example that does not involve a clinic setting.

Figure 3

Process Flow Example 2

Job Flow: The Travelers Insurance Company processes 10,000 claims per year. The average processing time is 3 weeks. Assuming 50 weeks per year, what is the average number of claims "in process"?

Here we have output over a yearly period (10,000 claims per year.) It is useful to think of this as 200 claims per week, given 50 working weeks per year. We are also given a lead time of 3 weeks and we may feel confident using this as a measure of cycle time. Thus if we think of the number of claims within the process boundaries as WIP we can use Little's law to argue that the average number of claims in the process must be,

$$I = \frac{10{,}000 \; claims}{50 \; weeks} * 3 \; weeks = 600 \; claims$$

Figure 4

Process Flow Example 3

Job Flow: John Doe has a sample sent to Made-up Labs for processing. Made-up Labs processes an average of 5,000 tests per week. The typical inventory of items to be processed is 250. What is the expected cycle time for John's test and what is the turnover for the system?

In this example, we want to speak about the CT for an individual lab test, but only have

aggregate data, (which is not unusual). We should also see that Little's Law is not entirely sufficient to address the problem. However, we have more than enough data to deal with the question. First, we know that average **WIP** = 250 tests, and **R** = 5,000 tests/week. Thus we see that **CT = WIP/R** = 250/500 = 0.5 weeks, and we can say that a typical test stays in this system for roughly 3.5 days.

We also know that CT is the inverse of turnover. Since CT is ½ week, we can conclude that turnover is the inverse of that value or twice per week.

Process Analysis - Putting the Pieces Together: Instrument Preparation at Madeup Surgery Center

Madeup Surgery Center is a small volume facility that performs a single procedure (with minor variations) on a number of patients on each day (5 days per week.) There are frequent no-shows for this outpatient procedure so the number of jobs will vary slightly from day to day. The list of instruments used for this procedure has been standardized to increase efficiency and consistency. The instruments used each day are delivered to the sterilization unit each evening to be processed so that they can be re-used the following day. In simplest terms, the instruments go through a process that can be described in 4 major parts:

1. Decontamination
2. Assembly and Packaging
3. Sterilization
4. Distribution

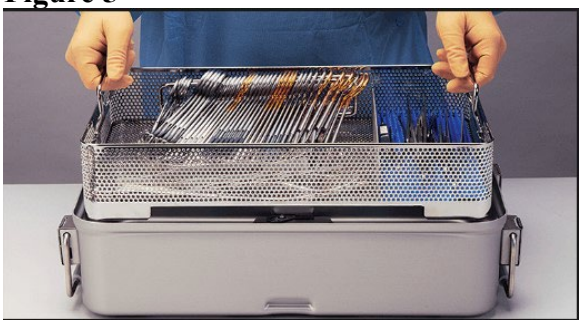
We can think of each of these 4 parts as consisting of 1 or more smaller steps. A more complete description of the process is to list these smaller steps as shown below:

1. Receiving
2. Scrubbing & Soaking
3. Cart and Load De-contaminator
4. Decontamination
5. Remove and re-pack
6. Sterilizer
7. Cart for Pick-up
8. Pick-up

This small unit employs 2 dedicated workers who we will call Albert, and Barry. Albert has been working in this unit longer than Barry. Consequently, Albert makes $30 per hour while Barry makes $20. We can explain the steps (in this admittedly stylized example) as a simple sequence.

In Step 1 items are received in small bins corresponding to a single procedure. Each bin is labeled for tracking. Barry receives the bin, records the relevant data, and positions the bin near a decontamination sink. This is a short step that takes 7 minutes per bin.

Figure 5



In Step 2 Albert visually inspects the instruments by the sink, scrubs those that have visual signs of contamination, and then loads the items onto a tray that is designed to facilitate the next step. Barry is a bit squeamish and is therefore not a good fit for this task. Fortunately, Albert has no such problems and enjoys this work. This step takes 15 minutes per bin.

In Step 3 Barry loads the tray(s) into the Turbo-11 Decontamination Unit. This machine is akin to a very high pressure dish washer. The water pressure is roughly 10 times greater than a typical car wash. This particular machine handles 1 tray at a time. Loading the machine takes 5 minutes per bin. Larger machines are available that can hold up to 4 trays at a time.

Figure 6

In Step 4 the Turbo-11 washes the instruments for a programed time of 45 minutes. The tray is automatically pushed out of the machine when the step is completed.

In Step 5 Albert inspects each instrument and packs them into a clean bin following a precise pattern for placement, and loads them into a sterilizer. This takes 10 minutes per tray/bin.

Figure 7

In Step 6 the sterilizer (which is essentially a fancy steamer) runs through a standardized cycle of 40 minutes. The model in place can handle either 1 or 2 bins at a time and ejects the load automatically upon completion.

In Step 7 Barry loads the bins onto a clean cart and records information on a paper form and into the IT system which tracks the location of every bin in the system. This takes 10 minutes and Barry can process either one or 2 bins during this time span. If he had to handle 3 or 4 bins it would take 20 minutes.

Finally, in Step 8, Barry hands control of the cart of bins to an orderly who picks them up to be delivered to the OR suite. This takes only 5 minutes regardless of how many bins are involved.

Let us apply the basic tools of process analysis to address a few simple questions.

1) If the system is empty when a single bin arrives how long will it take the unit to process a bin?

2) What is the bottleneck resource, and what is the Takt Time in minutes?

3) What is the capacity of this process?

4) How many bins can the unit process in a night, assuming that the technicians operate for only five hours?

5) How much of Albert and Barry's valuable time will it take to process a single bin? What is the labor cost in $/bin if no other labor is accounted for?

6) What is the capacity of this process if one of the workers is out sick or on vacation? What is the takt time on such days?

7) The hospital is considering leasing a larger decontamination machine; the Turbo 22 which can process 2 bins at a time. If this happens what is the capacity of this modified process?

Let's consider each of these questions in turn. The first question simply asks for the cycle time, **CT**. Since this is a serial system, meaning the steps progress directly from first to last, this is simply the sum of the time spent at each step so

CT = 7 + 15 + 5 + 45 + 10 + 40 + 20 + 5
= 137 minutes.

We can consider questions 2 and 3 together by considering each of the main resources in turn. Albert and Barry are each labor resources while the De-contaminator, and Sterilizer are the main equipment resources. Albert is involved in steps 2 and 5. Barry is needed for steps 1, 3, 7, and 8. The Turbo 11 is used in steps 3 and 4. The Sterilizer is used only in Step 6. Based on this data we can calculate how much time is required from each resource to process one bin.

Albert is needed for 15 + 10 = 25 minutes. Therefore, his capacity is (60 min/hr)/25 min/bin = 2.4 bins/hr.

Barry is needed 7 + 5 + 10 + 5 = 27 minutes and his capacity is (60 min/hr)/(27 min/bin) = 2.22 bins/hr.

Considering the equipment involved we see that the Turbo 11 is needed for 5 + 45 = 50 minutes/bin, and its capacity is (60 min/hr)/(50 minutes/bin) = 1.2 bins/hr. The Sterilizer is used only in Step 6 for 40 minutes and its capacity is 1.5 bins/hr.

In this simple serial process we can say that the takt time is:

Maximum(25, 27, 50, 40) = 50 minutes

corresponding to the busy times for Albert, Barry, Turbo 11, and Sanitizer respectively. Thus the Takt time is 50 minutes, the Bottleneck resource is Turbo 11, and system capacity is 1 job every 50 minutes. Recall that capacity is the inverse of takt time. If we write 50 minutes as 50/60 we see that capacity is 60/50 or 1.2 bins per hour.

For Question 4 we can use the fact that the system capacity is 1.2 bins/hr. If the system runs for 5 hrs/night, the system capacity is 5 * 1.2 = 6 bins per night. Note that there is a special assumption being made here. This implicitly assumes that carts can start the process and stop mid-process at the end of the shift or continue when another shift of workers takes over. This assumption may be reasonable for many "back-office" operations such as cleaning. However, if this is not a reasonable assumption then this calculation is not correct. Many settings such as clinics have fixed start and stop times and a job cannot be split across 2 sessions. For these cases we have to be explicit about how such instances will be handled.

For question 5 we can use some of the work we did earlier. The processing of a single bin occupies Albert for 25 minutes and Barry for 27 minutes. Assuming that the hourly costs provided include any relevant overhead we have $30 per hour or $0.50 per minute for Albert and $20 per hour or $0.33 per minute for Barry. Thus the total labor cost is

25 * 0.50 + 27 * 0.33 = $21.50.

Notice that if one of the workers is unavailable for any reason the remaining worker has to perform 25 + 27 = 52 minutes of direct labor. We also see that the takt time for this setting becomes:

Maximum(52, 50, 40) = 52 minutes,

and the direct labor becomes the bottleneck resource. Thus for the setting laid out in Question 6 the identification of the bottleneck resource is key.

Question 7 presents a slightly different scenario because the resource in Step 4 can handle 2 bins at a time. Note that if only 1 bin is in process at a time nothing has changed. However, if 2 bins are to be processed we have a slightly different system.

In the original process design where a flow unit was a single bin takt time was 50 minutes. If this process is used twice to process 2 bins, it is as though we have an "order size" of 2 units.

If the Turbo 11 is used for a job of this size we see that the total time needed includes 2 iterations of steps 1, 2, 3, 4, and 5. However, we only have one iteration of steps 6, 7, and 8 because the sterilizer can handle two bins at a time, and Barry's busy time does not change for steps 7 and 8 as long as the order size does not exceed 2 bins. Thus, we have busy times for each resource of:

(15 * 2 + 10 * 2) = 50 min for Albert,
(7 * 2 + 5 * 2 + 10 +5) = 39 minutes for Barry,
(5 * 2 + 45 * 2) = 100 min for Turbo 11, and
(40) = 40 min for the Sterilizer

Each of Albert's tasks have to be replicated to handle 2 bins. Thus Albert's busy time doubles when the load doubles. Two of Barry's tasks must be replicated but the other two do not so his busy time increases but does not double. The Turbo 11 must run through 2 cycles to handle 2 loads so its busy time doubles. The sterilizer can handle either 1 or 2 bins/trays at a time so its busy time has not changed.

Given this process design takt time becomes:

Maximum (50, 39, 100, 40) = 100 minutes.

Thus we can say that the capacity is now 1 order every 100 minutes. However, since the order size has changed we need to note that this is still 2 bins every 100 minutes or 1 bin every 50 minutes. Thus capacity has not changed at all.

It is instructive to consider CT for this new arrangement. Looking at Steps 1 through 8 we now have:

CT = (14 + 30 + 5 + 90 + 20 + 40 + 10 + 5)
= 214 minutes.

This is clearly much greater than the value of 137 minutes calculated for question 1. However, note that while this value is greater it is not twice as great. This happens because some of the activity times do not double when the order size is doubled. This is one reason we always have to be careful to be clear on whether we wish to change CT or Capacity as they are not at all the same things.

Now let us consider the introduction of the Turbo 22 that can handle 2 trays at a time. What changes is that Step 4 does not have to be replicated to handle 2 bins and the busy time for the Turbo 22 to process 2 bins stays at 55 minutes (5 * 2 + 45), compared to 100 minutes for the Turbo 11. Now the takt time becomes:

Maximum (50, 39, 55, 40) = 55 minutes.

Takt time for orders of 2 bins becomes 55 minutes per order and capacity is now 1 order every 55 minutes.

Moving from the Turbo 11 to the Turbo 22 we have "apparently" doubled the capacity of this resource. We add the word "apparently"

here because in actuality the useful capacity of this resource is defined by the structure of the system.

The Turbo 11 was the bottleneck resource when the order size was 1 cart. It had a busy time of 50 minutes, which was greater than any other resource. When the order size rises to 2 carts, and we use the Turbo 22 we have a busy time of 55 minutes. However, the capacity of the system does not double. The takt time of this new construct is now 55 minutes. Why does the capacity of the system not double, even though we doubled the capacity of the bottleneck resource? The answer to this question is that this resource has be set up by another resource and we did not double the capacity of that resource. Barry loads the Turbo 11 or Turbo 22. Note that for Step 3 both Barry and the Sterilizer are involved. The fact that resources must be combined to complete a step implies that capacity cannot double unless the ability of each resource doubles for that step.

Increasing the capacity of a bottleneck resource increases the capacity of the entire system, but this only works until some other resource becomes a constraint. Thus we see that doubling the capacity of a resource does not necessarily double system capacity even when that resource was the bottleneck.

We can take this insight a step further. Imagine if the sterilizer became twice as fast, and Step 6 could be done in 20 minutes instead of 40. At first glance this may seem like a good idea. However, since the sterilizer is not the bottleneck resource, increasing its "apparent" capacity has no impact on system capacity at all.

Key Take-Aways

To facilitate the more general application of these terms and ideas, let us collect many of the insights involved into a short list.

- A process is a collection of activities, performed by resources, which alter inputs to create desired outputs

- Efforts to "improve" processes need a coherent definition of what improvement means. These definitions can center around process metrics such as lead time, cycle time, work-in-process, throughput, or capacity

- Process Management involves description of a process, defining what we mean by improvement, focusing on metrics consistent with that idea, coming up with ways to improve those metrics, and implementing these changes

- Process metrics are inter-related meaning that changing the level of one metric will always have some impact on at least one other.

- The first steps in process analysis is typically developing a process map and the relevant metrics will typically involve the measurement of times

- The impact of changing any element of a process must be evaluated in light of its role in the process as a whole

- Basic tools of process management can give us a good starting point in our efforts to bring about system improvement.