ALBERTA AIM
ACCESS · IMPROVEMENT · MEASURES

Specialty Care: Basic Issues

This manuscript has been developed to address basic issues in Specialty Care (SC). There are other more detailed materials available on measurement (demand, supply and activity) as well as appointment types.

Demand Streams

Demand stream is the term that describes distinctly different types of workload that enter the overall SC practice. Workload can be divided into work that directly supports patients (office, procedure, surgery, on-call, hospital, etc.) and work that indirectly supports the patient (administration, teaching, meetings, etc.). Direct support workload is either scheduled or non-scheduled. Scheduled workload demand streams have their own demand, supply, activity and their own delay, which is measured by third next available appointment (TNA). Demand is measured as workload generated. Non-scheduled workload demand streams have their own demand, supply, and activity. There is no appreciable delay for the service, so TNA does not apply. Demand is measured as arrivals.

Balance

To work without a delay, a "department," a practice and individual providers must achieve a balance between demand (patient need) and supply (capacity). First, the supply/capacity must be balanced across the various demand streams, for workload that both directly and indirectly supports patients. Enough capacity must be allocated to each specific type of work. This "balance" is commonly attempted by instinct or intuition. For consistent and accurate balance, demand and supply need to be measured for both the total workload and for each specific demand stream. Different SC practices will divide the work into different demand streams. Occasionally, workloads from some demand streams can be overlapped (the on-call function and the office, for example) but more commonly, since the work is distinctly different and commonly requires a different place, different staff, different equipment or significantly different amounts of time, these streams need to achieve their own balance. Overlapping will usually create workflow turbulence and delay since one of the streams is "going faster" than the others. Having an "on-call" provider for deliveries who works simultaneously in the office will inevitably create turbulence in the office flow due to the interruptions.

Secondly, in addition to the requirement for balance between demand streams, in order to work without a delay, practices and individuals need to achieve a balance within demand streams. Within the office demand stream, for example, balance needs to be achieved between new and return patient appointments. New and return appointments are distinctly different and need separate channels (appointment types). The sum of new plus return demand must equal the capacity or supply allocated to support that demand stream.

Thirdly, in order to work without delay, a practice and individuals need to achieve a linear balance between upstream workload and downstream workload. For example, upstream new patient appointment work needs to balance and link to downstream workload in procedure or the OR. This linkage can be achieved, at least in part, by understanding, influencing and

measuring “octane.” “Octane” is the measure of the richness of the demand mixture. The first level of octane is measured as the ratio of new patient appointments to total patient appointments (new plus return). The second level is measured within the new patient appointments as the ratio of new patients that go deeper into the system (as a procedure or surgical case) to total new patients. Octane levels in both measures ought to be high.

System Delays

Delays are created in systems through four situations:

- If demand exceeds supply, either as total workload, or within any distinct channeled demand stream, then a delay will ensue. These delays will only get worse until “balking and reneging” occur.
- Variation of either demand or supply at any demand stream will result in a delay. While there is some demand variation into SC, it commonly falls within a measurable range and can be managed. New patient demand into SC generally, but not always, arises from upstream referral sources. While there is some variation, this can be predicted and managed, in part by service agreements. Variation in return patient visit rates between providers is commonly seen but can be managed by data, intervention, input equity (requiring all SC providers to manage an equal number of new patients in a given time frame) and by varying the scheduled ratio of new appointments to return. Most healthcare variation, particularly in SC, is due to variable supply. The demand streams compete for the supply that is available. Some demand streams - on-call and hospital - are always supported. Other demand streams – OR and procedure - are supported most of the time. Decisions for allocation of support for these demand streams are made based on opinions and priority for delays for these clinical issues, rather than on data and measurement. The office demand stream is often supported last and, as a consequence, the supply there is highly variable. Any time out of office for any reason has the greatest variation effect on workloads within the office and has minimal to no effect on the other “higher priority” demand streams.
- Triage and prioritization also create delays. Priority creates more channels of work. Again, due to variations in demand and supply, more channels and narrower channels of work will inevitably result in even more delay. Practices will commonly create multiple priorities and channels based on a patient’s clinical condition and, at the same time, are blind to the operational effects of such tactics. Not only does demand variation into narrow channels create delay but, at the same time, the wider range of supply variation creates even more delay. There is minimal delay in the demand streams that are always supported (on-call and hospital), less variation and delay in the streams supported most of the time (OR and procedure) and significant variation in delay in those streams supported as low priority (office). Within low priority streams, if workload is further sub-divided by clinical priority, there is even more variation and delay that results.
- Delays are often intentionally created or allowed in order to buffer variation by reducing the effect of demand variation on the schedule, or in order to maximize revenue or visits. An intentional warehouse of delayed work is not an effective buffer for variation, particularly since most variation is created on the supply side, not the demand side. In addition, an intentional warehouse does not guarantee visits or revenue. In fact, the warehouse buffer

creates more rework and redundancy, more no shows (reneging) and, because of the intentional wait, requires more resource to triage. That resource is used up in sorting and prioritizing (deciding who goes “first”) and therefore cannot be used in “doing the work.” This approach is commonly described as the “myth of 100% utilization,” a term that describes the false belief that maximum efficiency is gained by filling the schedule in advance.

Measurement of Delays

Workload (demand) into SC is commonly divided into new and return appointment types. Special value is created in SC by seeing new patients. Delay into the practice is measured as:

- Third next available appointment (TNA) for both new and return patients for all individual providers. This measure is highly dependent upon supply (presence in the office) and variation due to competing demand streams.
- The average TNA for the practice. This measure buffers some of the individual variation among providers.
- The TNA for any provider within the practice. While this measure is valuable in practices with referrals to individuals, it has the greatest utility is practices that pool the referrals. Coupled with planning for consistent practice supply (not necessarily individual provider supply) to meet predicted new patient demand, and (as needed) shifts of the ratios of new to return appointments, pooling of referrals within the department can lead to minimal delays for all new patients. All these strategies need to work in concert in order to achieve significant reductions in delay. Measuring the first TNA new patient appointment with any interchangeable provider reflects the value of these changes.

If a department or practice has individual providers present for most days in the week (that is, over half), TNA by individual provider is the appropriate measure. On the other hand, if a department has minimal office supply presence (each provider is in the office less than 50% of the time) and referrals are sent to individual providers within that department, and TNA is used, then the delays are highly variable. The "average" helps buffer some of the variation but the delays are high, variable and often past the threshold goal. In this situation, the best approach is to pool referrals to the department or practice, avoid referrals to individual providers, and measure the first TNA appointment to any provider in the department or practice. The practice looks for the first TNA open with any provider. Without the necessity to appoint to any specific individual, appointing this way and measuring will reflect the true delay for any new patient into the practice. By spreading out the new appointment capacity over the week, appointment capacity can be provided within the desired threshold goal. For this measure to be applicable, referrals must be pooled.

Variation

Variation, either between or within demand streams, will cause a delay. Variation between demand streams is primarily a supply variation issue and variation within demand streams is both a demand and a supply issue.

The delay for an appointment, within the office demand stream, is due primarily to inconsistent (variable) office supply.

The most critical delay into the office is the delay for new patient appointments. This is where unique and special value is created. These appointments compete with return appointments for capacity on the schedule. Demand for new patients may vary but is predictable and rarely varies outside a common range. If the schedule has a rigid ratio of new to return appointments but the supply of appointments varies, the number of new appointments will vary widely with the number of providers scheduled in the office. As a consequence, since new patient appointment supply is has a fixed ratio but is dependent on the number of providers, the wait time for new patient appointments will vary in conjunction with the number of providers in the office.

Providers in OB-GYN, for example, will commonly support – either alone or with some combination - a number of distinct demand streams: OR, ASC, office, on-call (may include deliveries, ED support, hospital work), hospital, procedure and other. Some of these streams are supported 100% of the time (for example on-call or hospital) without ever being neglected. Other streams can be neglected infrequently but have wait time limits, usually informal and instinctual (OR and ASC). Still other streams - office in particular – are felt to be lower priority and are allowed to be neglected. Hence, there is far more sporadic support of the office demand stream and, as a consequence, while demand continues unabated, the supply variation results in delays. In addition, if the ratio of new and return visits is not constantly adjusted, the delays for new patients into the office rapidly accumulate.

Caseloads

There is a limit to provider capacity. Direct patient work competes with indirect work. Direct patient work contains a number of specific competing demand streams (office, OR, procedure, on-call, hospital, etc.). Depending on the amount of capacity allocated to the competing demand streams, the office workload has a capacity limit. This limit can be measured. The caseload (number of patients that an individual provider can manage) can be determined by the following equation:

> **Caseload X Expected Visits per Patient per Year = Number of Days Worked in the Office X Number of Scheduled Appointments per Day**

In this equation demand (on the left) equals supply (on the right) There are methods to determine caseload in practices (see articles on Panel). Expected visits can be determined from the last 12 months of practice activity. Supply metrics can be determined by practice analysis. Any of the four variables can be isolated. For example, the optimum number of return visits can be determined if the other three variables are known.

The four variables in this equation can be altered or managed with a multitude of strategies, for example, service agreements, use of midlevel providers, and creating efficiencies in office throughput.

Each individual provider has a specific caseload limit that is also determined by the above equation. Taking on more work than the equation indicates will result in poor system performance. Initially, the over-demand is allowed to wait in an expanding backlog, then the delays become costly and risky, and finally the demand, either by default or by intention, gets moved either to another provider or to another entry point. If the delays for office appointments reach a specific critical point, work commonly flows into the ED where it is, by default, pushed to another provider or admitted to the hospital and presents as hospital demand. The demand

does not simply disappear. There is a common and false belief that a long delay creates a chilling effect on referred demand. Actually a long delay stimulates “just in case demand.” For example, studies have shown that the longer the wait into orthopedics, the more “just in case” referrals are sent.

Caseloads and Pooled Referrals

Sending referral work to specific individual providers, rather than pooling referrals, can result in a number of problems:

- Each provider has his or her own caseload limit and individually directed referrals can create demand/supply mismatches and result in delays and risk.
- Referrals to individual providers can result in delays due to the variable office presence (supply) of the provider. In SC, office supply is highly variable, due to a multitude of competing demand streams.
- Referring to individual providers is often seen as “freedom” or “choice” for the referring doctors, but if a specialist’s caseload exceeds his or her capacity to see patients, that freedom and choice is a falsehood. In addition, the premise that a practice will let patients choose to wait for an individual provider if demand exceeds capacity is deceptive. Demand either exceeds capacity or not. This is an objective measure, not a subjective opinion. If demand and capacity are balanced, a delay is not necessary. If demand exceeds capacity then a lengthening delay will always ensue. The “choice” about waiting to a preferred doctor is nonsensical.
- Referrals to individual providers are often seen by these individuals as a validation of their “popularity.” This is a dangerous situation. If referral demand exceeds capacity then that individual, despite being “popular,” simply cannot do all the work. Making patients wait may give a temporary respite but is not a long-term option since the delay will lengthen, putting these referred patients at risk. Resorting to “let the patients decide” is deceptive since patients do not have all the correct information to make an informed decision.

Thus, sending work (demand) to individual providers without appropriate measures will inevitably result in delays. A better approach is to measure demand for new patient appointments to the practice as a whole, and pool the referrals (interchangeable providers), develop a daily or weekly schedule that ensures provider office presence, make adjustments to the daily or weekly scheduled ratio of new to return appointments in order to keep up with measured demand, and send the work to the first available provider in the pool. There is a caseload limit for both individual providers and the practice. This is objective and measurable. Pooling referrals allows for load leveling of work, helps groups respond to fluctuations in caseload due to patient “graduation” and, most importantly, reduces variations in demand and supply. Sending work to individual providers results in boluses of office demand, whereas pooling referrals levels out the demand fluctuations and results in far less delay. If office supply can be smoothed by ensuring consistent office presence, and through pooling, the effect of an individual provider’s absence can be eliminated and new patient delays can be minimized.

Re-Pooling Referrals

With the exception of those practices with direct referral to procedure (which can be measured and scheduled), most patients who get procedures (either investigations or surgical procedures) are initially evaluated as new patient appointments and then internally referred into a second appointment. The linkage between the initial appointment and the subsequent visit for procedure represents the third level of balance required for flow (the first balance is between demand streams and the second within demand streams). In order to eliminate delay in this linear two-step process between appointment and procedure, some practices recognize the dependent relationship between these two steps: the octane levels (ratio of new to total and the ratio within new of those patients who progress on to procedure) within the office determine the amount of procedure capacity needed in order to keep up with the workflow. In some practices, primarily those without large numbers of competing demand streams (and dilution of direct patient work by non-direct work), a consistent linkage can be made between office and procedure. In dermatology, for example, office work and octane levels can be used as a link to procedure within a determined time threshold. Patients can be seen in the office within five days and then seen again within five days for their procedure. Within this “simple system” there are still variables - in referred new patient demand, in short term return rates, and, of course, supply scheduling variability. With five-day buffers both at the front end and in between the two steps, much of this variation can be absorbed with planning and adjustment. However, in more complex systems, with wider ranges of demand variation, return visit rates and, in particular, highly variable supply presence due to competing demand streams (not only non-direct patient work competing with direct patient work but OR, on-call, hospital, competing with office and procedure) linkage between initial new appointments (or return appointments) and subsequent procedure is more difficult. In some practices, continuity (same provider) for appointment and procedure is critical. However, in others, commonly where the procedure is a task or test done with an unconscious patient, this linkage may not be critical. In these practices it is possible to re-pool the referrals, that is, unlink the appointment from the procedure to allow for more flexibility in procedure capacity. In gastroenterology for example, by pooling referrals for new patients and “disallowing” referrals to individuals, some practices, are able to provide initial new appointment capacity with an interchangeable provider within the desired time frame (usually five days) and then by re-pooling the procedures are able to achieve linkage and services within the time frame goal (usually another five days) for the procedure as well. Re-pooling allows for more flexibility in capacity and in scheduling since individual offices no longer have to be directly linked from appointment to procedure. A provider scheduled in procedure will have a schedule of patients who have been scheduled no longer than five days previously. The linkage then shifts from individual providers to functions. It is important to note that re-pooling is not applicable in all settings.

Priority

Many complex systems, in attempting to deal with a perceived overwhelming amount of demand, use priority, triage or sorting mechanisms to divide and subdivide that demand. While prioritization segregates demand streams (one stream is more important than another), prioritization is also used to segregate appointment workload within the office, surgery or procedure streams. Segregation into priority lines is based primarily on opinions, but with some “data” about “reasonable medical delays” for the gamut of clinical conditions. Each segmented

priority line has it's own recommended wait time threshold. This approach presents serious and fundamental flaws. While some clinical conditions are indeed more time sensitive than others, the flaws are operational in nature:

- Total system or department/practice capacity has to equal total system or department/practice demand. Without this balance, and in particular, if demand in any of the distinct demand streams exceeds supply, setting a wait time threshold or standard is meaningless, since that mismatch will create an inevitable delay. It makes no sense then to set thresholds for various clinical conditions unless the measured demand is balanced with capacity.
- Prioritizing demand into various sub-streams with standardized waiting time goals or thresholds will require the use of sophisticated triage. These priority decisions are often quite complex and, as a consequence, are commonly made by high-level clinical resources (physicians). This need for triage uses up capacity (to sort the work) that could have been used to do the work. As a consequence, demand will continue and supply will be diluted due to resource required for triage, for training and for updating. Thus, supply needs to exceed demand in order to make this work.
- Each case or event needs to be reviewed against some pre-set criteria. Often information on which to make the priority decision is missing and resource is either spent in gathering the information or the work is sent back - creating a delay - in order to get the correct information.
- Prioritization will inevitably result in mistakes and errors. These errors will occur at the point of referral (inappropriate or inadequate referral) as well as at the level of triage. Even "perfect" information is useless in a changing clinical environment. Errors lead to a desire for even more prioritization, more triage and more inquiry, which will create more channels, standards and lines - which in turn will create a higher likelihood of delay.
- To ensure that the priority decisions are made correctly, a filter or extra step is commonly added to inspect the work for appropriateness and to ensure correct routing. This extra step adds delay and also consumes resources.
- Prioritization will increase the likelihood of no-shows and unused capacity in the lines with longer thresholds. Some of these patients will balk or renege and quit the line, while others will cut in line (or their primary care referring doctor will help them cut in line) and they'll find entry into the system through a shorter portal (through the on-call function or into the hospital, often through the ED). These activities will increase the chaos, rework, phone calls and inquiries in the lines, and increase the no-shows.
- Multiple priorities and entry criteria will lead to exaggeration and fabrication of symptoms on the part of the referring providers in order to get their patients to the front of the line more quickly. This increases the re-work and redundancy in the referral system and increases the likelihood that referring providers will call the specialists personally to advocate for their patients who are "exceptions."
- Workload and resource (demand and supply) need to be measured at both the system and individual stream level. Even if average demand equals average supply, variation will create a wait time. Prioritizing the workload (demand) within each of the streams based on clinical

condition will create multiple channels of work. Each of those channels of work will demonstrate either demand or supply variation. These variations will create temporary mismatches in supply and demand in all the various channels. The more and narrower the channels of work, the higher the likelihood of variation and delay. This variation in demand or supply will be extremely difficult to manage. The greatest effect of variation will be on the shorter, more critical, time-sensitive lines:

  - For example, if a gynecology practice prioritizes office demand into seven lines with seven delay thresholds at 26, 12,8, 6,43 and 2 week margins, patients will be scheduled at the margin or limit of those thresholds. Although demand into the two week line (shortest time frame, highest priority) is predictable, that demand is not fixed and it will vary within a predictable range. As variation occurs into the two-week line, there is little flexible capacity to manage “up demand” (variation above the predicted or above the capacity of the line). Delays then occur in the two-week line. However, in order to prevent delays, the practice then has to either overbook into the two-week line or steal capacity from one of the already pre-scheduled lines by overbooking there or canceling from those lines. Stealing supply or capacity from another line will cause that line to exceed its threshold. As an additional consequence, more staff resources are needed, more triage and priority is needed, and the likelihood of delay for those patients deemed to be the most critical actually increases. Ironically, the patients with the highest likelihood of having a delay past their threshold are not those patients with the lowest priority, but those patients with the highest priority. With multiple queues accompanied by set thresholds for each queue, demand variation will guarantee delays in the line with the highest priority. In queuing theory, if supply is fixed (and in SC systems supply is not fixed - supply is highly variable) and demand varies with common natural demand variation, trying to balance more than two lines with set, pre-determined thresholds is impossible. It is thus operationally impossible to achieve the goals set forth by the priority scheme.

- While even normal commonly expected demand variation in an environment of fixed set supply and wait time thresholds will lead to delays, if supply variation is also considered, then delays are virtually guaranteed particularly into the office, which is commonly considered lower priority against other competing demand streams (OR, on-call, hospital and procedure). These competing demand streams are always scheduled and some of them are always scheduled preferentially. The office then bears the brunt of supply variation due to time out of office for any reason. Priority, along with office supply variation, is a dangerous combination.

While it is evident that some clinical conditions are more time dependent than others, priority depends on correct and consistent sorting decisions, requires a high-level resource to do the sorting, assumes (incorrectly) that the clinical condition does not change, assumes or ignores the requirement of overall demand/supply balance, and fails to understand that, unless a measurable slack capacity is introduced into the system, the sickest patients (the ones with the highest priority) are the least likely to meet threshold standards. While slack capacity is built into the “on-call” function, slack is inevitably considered waste when introduced into other streams.

Wait time issues are primarily operational issues, not clinical issues. “Medically reasonable” prioritization sounds reasonable. After all, some patients are sicker than others and deserve to

be seen sooner. However, prioritization as such is built on fantasy- the fantasy that demand equals supply, that all the priority queues will contain exactly the same amount of workload as the previously allocated supply and that there will be no variation. This fantasy hopes that not only will demand exhibit no or little variation, but assumes that there is no supply variation. Most priority focuses on clinical conditions managed within the office or scheduled procedure setting. In any given time frame - a day, a week or a month - the amount of supply variability is staggering. The variation in either demand or supply will have the greatest effects on those with the shortest or the longest delay thresholds respectively. The patients with the highest perceived risk will have the highest likelihood of not being seen within goal. Priority systems are designed to fail and are designed never to achieve their proposed outcomes. As such, it can be expected that practices and providers will find methods to accommodate, to work around and to cloud measures of system performance, to avoid judgment and scrutiny since they simply cannot achieve these goals.

The solution, of course, is to minimize the number of priority lines and pull the desired delay to a threshold underneath the highest priority (with the possible exception of the “immediate” groups, most commonly managed by the on-call function). In this way, the various proportions of work do not change, all the work is done with a short delay, and with more capacity due to the load leveling feature of less lines, and due to a merger of multiple narrow channels into a wider channel, slack capacity is built in and up or down variations can be absorbed over that wider channel.

Sub-Specialists

Many SC practices have mixed supply (combinations of both general providers and sub-specialists). If the sub-specialty work is distinct and managed distinctly, then measurement, balance and improvement focus on the sub-specialty alone. On the other hand, in mixed practices, measurement is critical. Because of the narrow supply channel it would be rare to have a sub-specialist working within a general practice have exactly the ideal caseload or ideal demand. Demand and caseloads commonly either exceed or do not meet sub-specialty capacity. If the demand exceeds capacity, demand reduction and supply enhancement strategies are necessary. If demand is less than capacity, the practice ought to measure the sub-specialty demand and caseload and use the general work to load level and fill in the capacity gap. Pooling referrals is of immense benefit here since that pooling allows for measurement, reflection and conscious decision on directing the general workload. Sub-specialty demand can be measured to preferentially fill the sub-specialty schedule and then general workload can be used as the load-leveler.

Planning for Return After Time Off

In order to work with minimal delay, demand and supply need to balance. SC practices are fraught with variation. While there is some office demand variation, it is predictable and can be managed by service agreements, changing ratios and pooling referrals. Office demand variation is passed through to the other competing duties (OR, procedure) deeper in the flow system. This variation can be managed by continuous monitoring of “octane levels” and shifting the amount of the procedure or OR service that is linked to the office entry demand.

The largest ranges of variation are not, however, seen on the demand side but on the supply side. Vacation, Continuing Medical Education, meetings and other diluting factors are all superimposed on the constantly shifting schedule. Additionally, all the competing demand streams are not supported equally. Some of the streams are always "covered" (the hospital and on-call functions for example) while other streams (the office in particular) can seem almost an afterthought. Hence, because of the uneven support, much of the supply variation effect is shifted to the office demand stream and wide ranges of office supply result. As a consequence of uneven office supply, commonly accompanied with rigid schedule ratios of new to return appointments, uneven, variable workloads are sent deeper into the system, amplifying the variation as the workload moves.

Because of the variation, it is critical to manage all forms of "time off," both due to vacation, etc. and participation in or commitment to other duties.

Post Vacation Planning

The classic approach for standard post-vacation planning is a modification of the PC approach:

- As soon as a provider has been approved with vacation or any time off:
    - Block that week(s) the provider will be absent, and at the same time, block the afternoon sessions of the week provider returns.
    - Leave open the morning sessions of the week provider returns. These can be used/booked before he/she goes and while he/she is away. This capacity could be used for new or for return patient demand. The ratio is dependent on the practice capacity needed at that time to keep up with new patient demand.
    - When the provider leaves on vacation, open half of the afternoon sessions of the week he/she is to return, leaving half of the afternoon sessions closed. This capacity could be used for new or for return patient demand. The ratio is dependent on the practice capacity needed at that time to keep up with new patient demand.
    - When the provider returns, open up the remaining half (¼ of the total schedule) of the afternoon sessions, either one day at a time or all together. This capacity could be used for new or for return patient demand. The ratio is dependent on the practice capacity needed at that time to keep up with new patient demand.

While the number of other SC providers in the office when the vacation provider returns may have an effect on the size of the "carve out" (held appointment space) due to less need for new patient capacity for the returning provider, the need for frozen capacity is generally less in SC than in PC:

- Because most return patients are pre-booked as they leave SC, the returns are more "controlled."
- There are fewer walk-ins in SC and most of these are deflected to the on-call function.
- When patients call back for a return appointment there is some discretion in timing and they can be scheduled into the future.

The above comments do not address the distinction between new and return appointments upon return. Decisions about those ratios are also critical. In addition to the classic post vacation planning described above, there are other more sophisticated approaches. Some of these have been addressed previously and will be reviewed here in light of specific post vacation concerns:

Pooling Referrals for New Patients

When the provider is gone, demand for new patients continues unabated and is measurable. Pooling the new patient referral demand and scheduling enough provider new patient capacity each day to keep up with the measurable demand allows the SC practice to direct the new demand to providers with new patient capacity (appointments available) within a five-day goal or threshold. To keep up with new patient demand, the SC practice is required to first ensure enough provider presence and second to change the ratio of new to return capacity on the schedule.

Planning the Return From Time Off

In most SC practices, prior to leaving for time off, providers spent disproportionate amounts of time in the on-call function in order to “make up for lost call” and within the OR or procedure functions in order to make sure that patient delays for these activities don’t extend past the comfort threshold due to the pending time off. As a consequence, the office is neglected, demand continues and delays mount up. When a provider returns from time off, he/she commonly returns to the on-call function or to the OR, and not to the office. The office delays continue to increase. The initial return ought to be to the office so that providers can manage the delays there, discover OR or procedure cases, send them deeper into the flow system, and then concentrate on supporting the non-office functions.

Schedule Ratio of New to Return Patients

Most SC office schedules have rigidly determined ratios of new patient time against return patient time. Occasionally these ratios are not displayed on a single schedule but are manifested as the ratio of new patient clinic to return patient clinic.

Specialty Care Recommendations

Thus, in order to achieve a goal of short delays for all patients and demand streams and, in particular for new patients, the best SC strategies are:

- Measure total system capacity and demand across all demand streams (all types of work), and achieve a balance between that demand and the total capacity. Demand streams (distinctly different types of work) include office, procedure, surgery/OR, hospital, on-call, etc.
- Measure demand, supply, activity (what is done) as well as the delay for each individual demand stream, and work to achieve a balance
- Measure demand, supply and activity, as well as the delay within each demand stream. For example, the office demand stream contains new work and return work. The total of that

demand must equal the supply or capacity allocated to that work or an ever-expanding delay will ensue.

  - Make sure that there is enough provider capacity each day or each week in order to keep up with measured new patient demand.
  - Pool the new patient referrals. This allows the practice to avoid delays that would ensue due to referrals to individuals.
  - Constantly alter the ratio of new to return appointment slots to keep up with new demand and to overcome the variation in daily or weekly provider supply.
  - Focus on scheduling the office as priority both in general and, in particular, for post time off scheduling. This may require some shifting of OR and procedure scheduling and some batching or bunching of the on-call function. Commonly on-call is scheduled first in order to make it equitable and spread it out. Secondly, OR or procedures are scheduled in a rigid fashion and then the office "fills in" whatever time is left. Attempting to reverse or at least alter this scheduling priority may be necessary for improved flow. This allows the practice to respond to measurable demand and create more stable flow and linkage from office to OR or procedure.
  - When the post time off provider returns to the office, new patients can be scheduled in order to "prime the pump" and feed the OR or procedure. With pooling of referrals, the new patient appointment slots can be opened within the five-day window. In so doing, the returning provider gets a bolus of new work but due to pooling, the patients do not have to wait beyond the threshold. This office bolus then can be linked to the future OR or procedure schedule by studying the relative "octane" of the office workload. If the octane is known, the amount of procedure or OR time can be calculated, scheduled and linked to the scheduled office time.

- Minimize the number of lines, queues and priorities.
- Reduce the wait time goal for everything to below the lowest clinical threshold. This allows load-leveling of all the various queues, and provides far more flexibility in scheduling.
- Flex the capacity to keep up with demand. This can be done by pooling referrals to avoid individual mismatches and can also buffer the effect of supply absence, changing the ratio of new appointments to return appointments in the office, and making sure there is enough provider new patient capacity in order to keep up with predicted and measured demand.
- Use "secondary pooling" (pooling of procedures) whenever possible. Secondary pooling can provide even more buffer for the linkage of office to procedure. A practice loses continuity (a choice) between the initial office visit and the secondary procedure but gains better flow due to wider channels.
- Carve out post time off capacity as capacity for new patient appointments and open this schedule time with only a five day window or less. In this way, new patient demand can enter within five days and feeds the downstream OR and procedure demand streams. The OR and procedure can be linked to the office by exploring the expected octane levels within the office practice. For example, if half the new patients go to the OR or procedure and they take twice as long, then a practice needs the same amount of procedure time as office time.