



Continuity: Q and A

Questions on the following topics are posed and answered in this document:

- Continuity and Part-Time Providers
- How to Appoint Patients of Absent Providers
- Continuity and Delays
- Discontinuity Visits

Question: Continuity and Part-Time Providers

Please provide strategies around maximizing continuity in a clinic with FTE's of 70%, 60%, 60%, 40% and 20%. In a way, theoretically, it seems that the maximum continuity achievable would be those percentages for each of the doctors respectively.

Answer:

If a provider is present in the office 50% of the time, for example, and demand arrives randomly, over time, we would expect a hit rate (continuity) of 50%. The approach, though, would have to include a commitment by the doctors to see their own patients when they are present in the clinic.

If we can move continuity above the random expected percentage (the amount of time the doctor is in the clinic), this means that we have a system that moves the work (demand) intentionally to the appropriate provider.

Strategies:

1. See your own when present. To accomplish this, there must be no delays in the practice. If there are delays, sick patients get diverted. So see your own patients and don't make them wait. Practices with less than 60% presence would have to use a carve out model (save some space each day for "sick" patients)
2. Focus on bringing the return visits back to the same provider.
3. Make sure that the panel is right sized. This is absolutely critical. If the panel is too big, the provider cannot see his/her own patients without a delay.
4. Measure continuity correctly (the number of visits my patients made to me divided by the number of visits my patients made to all doctors in the clinic)

In addition, for practices where there could be an FTE that is very small, say 10% or even 20%, that means that provider is in only one day every two weeks for the 10%, and perhaps once a week for the 20%. In this case, you have to question whether that provider can really maintain a panel, and certainly wonder how he/she can provide continuity to that panel. In a practice where that 0.1 or 0.2 FTE is one of many providers, you might distribute their patients among the other providers, and make the very small FTE doctor someone who doesn't have his/her own panel, but rather sees patients of absent doctors as required. For example, if this is a practice of 10 docs, and

three or four are “out” on Thursdays, the very small FTE doctor would be scheduled for Thursdays and would see the patients of the four absent doctors, so that the six doctors that ARE in that day could see their own patients and not have to worry about having to see patients who belong to their absent colleagues. This helps the doctors who are in the office to maintain continuity with their own patients.

In another scenario, you might consider “combining” two doctors with smaller clinical FTEs, to make a “single team.” This could work with two 0.4 FTEs, or even perhaps two 0.3 FTEs. There would have to be some rules though:

- The main rule is that the doctor “team” would have to be accountable for the entire panel of both doctors. Of course the combined panel would have to be the right size for the amount of supply available by combining the two FTEs.
- Combining the panels of two part-time providers also works best if the providers are the same gender.
- The doctors should also be the same type of doctor. You couldn’t combine a pediatric and an internal medicine doctor, as there is no overlap in their practice.
- The two doctors must consider the “combined panel” as their own, and not favor their patients over their “partner’s” patients.
- The two doctors should not be scheduled in the clinic on the same day. They have to distribute their work days across the week so that one is always there, or in the case of adding up to less than a full FTE, their “in” days are distributed across the week (M, T, TH, F or M, W, F).
- The doctors must have a way to communicate with each other relative to critical cases, lab work and hand-offs. The ideal situation would be to have the same nurse or MOA working with both doctors. The nurse or MOA could then be the conduit for communication and continuity of care.

Question: How to Appoint Patients of Absent Providers

In relation to questions about continuity, it would seem that patients who call for an appointment would want to see the same doctor unless it was an urgency. If it is an urgency, then it would seem to me that the benefits obtained from continuity (e.g. knowing the background, history, social situation, etc.) are less critical than the need for urgent care. Moreover, the need to “know” the patient is less critical for urgent conditions - you just treat the sore throat or otitis media. On the other hand, for chronic conditions, it surely makes sense to maintain continuity to avoid the inefficiencies associated with seeing a different provider. For example, seeing someone on an urgent basis for a severe sore throat is different than seeing someone on the same day for adjustment of their blood pressure medication. Could a receptionist sort this out by asking the patient if they can wait a day or two to see their “own” doctor?

Answer:

There are a number of issues at play here.

First, it is difficult to determine in advance just what is “acute” and what is “chronic.”

Second, I always suggest minimizing the delays so that there are real choices for patients. If the provider is present, he/she ought to see his/her own patients no matter what. We assure the provider of a smooth day by having the right panel size and using all the contingencies at our disposal - moving the return appointments to "load-level," using data to predict, post vacation plans, having a team to absorb any daily variation etc. In this light I think it is critical to understand that we already deal with variation - a provider rarely sees exactly the number of visits on the schedule.

Third, once we eliminate the delays and have real choice, we should let the patient decide: "Your provider is not here. You can see another provider today, or wait until your provider returns."

Question: Continuity and Delays

In our Specialty Care (SC) practice we value continuity. Valuing continuity can have an effect on delays. We think continuity is most important. What should we do?

Answer:

There are a number of issues at play here. We all agree that continuity is "good" and that delays are "bad." We commonly see a tension between continuity and delays. This tension is often resolved by physician opinion and preference (not all bad). There is a tension and I believe that we have to solve this not by alternating competing opinions but by looking at systems operations and making informed choices.

Continuity

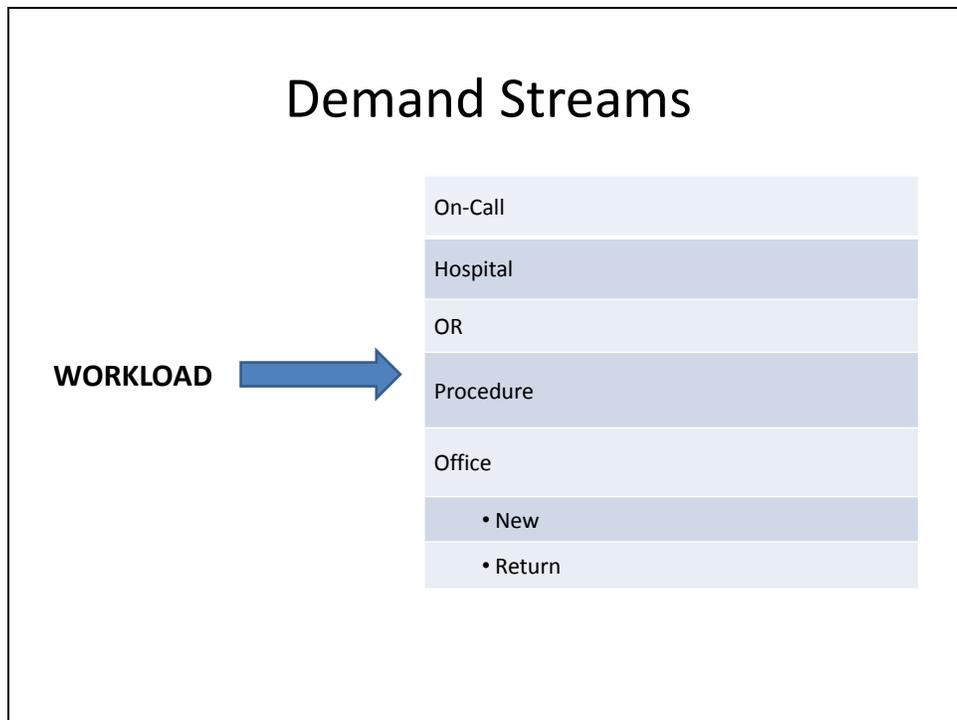
Continuity has, for the most part, been viewed as a system property in SC. Physicians see the same patients as new appointments, as return appointments, as procedure, etc. In SC, the delay for new appointments often acts as a buffer and allows more flexibility to achieve continuity. New patients are referred to individuals within the SC practice and since office supply for new patients is inconsistent, new patients wait various lengths of time to see the selected provider. Office return time, procedure time and OR time are also inconsistent and patients wait varying lengths of time to see their own provider for these services. Continuity comes at the expense of delays. Prioritization contributes to the delay issue. If referrals to individual providers are further stratified, prioritized, triaged or sorted according to acuity, some patients are forced to wait even longer. Inconsistent office supply, unmeasured workloads (demand) coupled with referrals based on "popularity" all contribute to delays. These delays commonly result in patients "jumping the queue" - moving from long delays into the office for a particular physician, to a shorter delay into the ED (on-call function) or hospital for another physician. The desired continuity is lost.

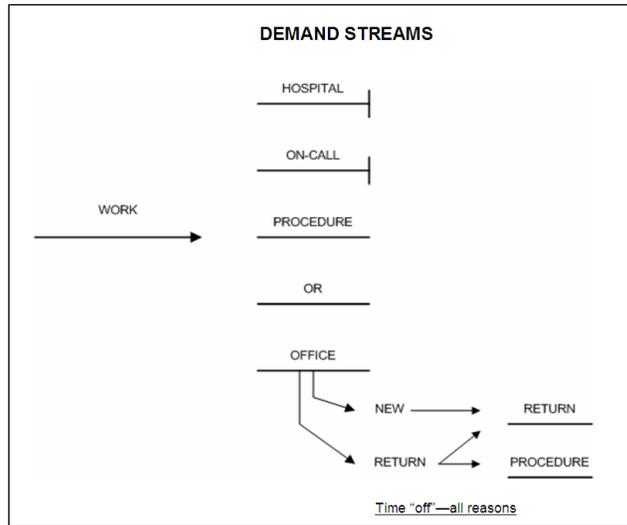
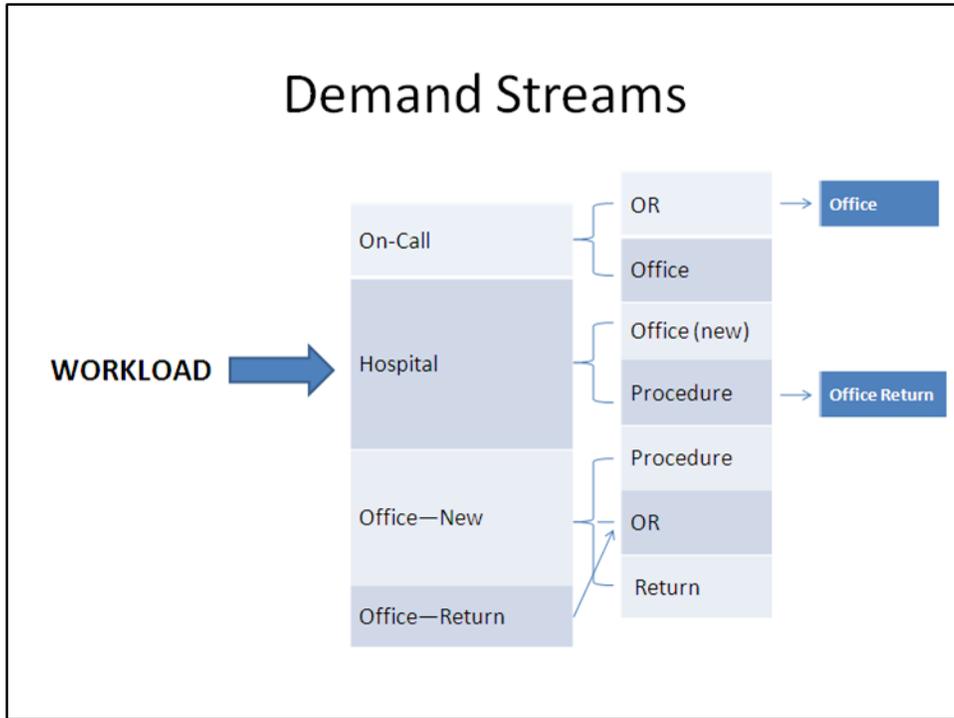
While there is distinct value in establishing a relationship and continuing with it (continuity), the argument for continuity is often made as opinion. Continuity alone and in isolation has operational consequences. If we push for continuity as a desired strategy, we also have to make sure that our system can provide the continuity without the risk and cost of delays. The best systems will have patients see their own physicians and see them with minimal delays. We have to accomplish both goals.

Delays

Work flows **into** specialty care as distinct demand streams. This demand stream work competes with the SC physician for allocated time. Commonly, time is allocated according to instinct and to “necessity,” that is, some of the queues (the on-call function and the hospital function) have to be supported each day while others are supported with varying levels of priority (for example, procedure and OR take precedence over office). Further, within the office there are two competing queues - new and return appointments.

Work flows **through** a system as a series of steps. Each step has a demand, a supply and an activity as well as a delay at the step. My focus is always on the delay. The primitive map below demonstrates the potential delays. Each arrow represents a distinct step and a potential delay. We have to view the workflow from two perspectives: a.) a static look at the competing demand streams and b.) a dynamic look at how one demand stream feeds into another. In the static look, the competing demand streams contain work that is generated at different times. The demand for on-call was generated today. The demand for a new appointment was generated within the last five days, and the demand for procedure was generated from the new appointments seen five days previously. The dynamic view captures the flow from one step to another in sequence.





Patients view our success as, and risk is often determined by, the sum of the delays. Cost is clearly influenced by the sum of the delays. The sheer number of competing demand streams and the amount of “time off” will determine the delay for a new office appointment, if referrals are sent to individual providers.

If the number of competing queues is high or the proportional amount of time spent in some necessity queues is high (lots of on-call and hospital time, for example, in smaller settings), the office setting often “loses.” That loss, which is a supply variation, results in delays of work into the office since while supply varies, demand continues unabated. We can, however, mitigate these

delays, particularly for new patients, by making distinct appointment types for new and for return patients, by pooling the new patient referrals (allocating workload of new patients based on proportionate amount of time in the office and minimizing “popularity”), by flexibility in scheduling (eliminating set days for set activities), by measuring and scheduling enough new patient capacity to meet new patient demand each week and by adjusting the ratios of new to return patients on the schedule in order to keep up with the new demand. Referrals sent and divided by “popularity” will always result in delays and in mismatches for the popular physicians, particularly in smaller settings. All of these changes and strategies have to be informed by measurement of demand with a comparison to capacity and by an understanding that demand for office simply cannot exceed office capacity. Prioritization will not solve this sort of mismatch but will actually ensure that system performance deteriorates. Practices may have to use strategies to reduce demand: Service Agreements (that define the work, packaging, and “graduation plans”), a review of the variable return visit rates, and employing supply enhancing strategies (making sure physicians are doing physician work). Changing the way we respond to demand also requires changing the way we measure delay.

Linkage of new patient office capacity to Operating Room or procedure time can minimize the delays into the second step. New patient office time ought to be linked (by an octane measurement of what percent of new patients seen go on to Operating Room or procedure) to the Operating Room or procedure capacity. We need to avoid, for example, scheduling a physician who is leaving on a month long vacation with new patients.

Some practices further mitigate the delays for procedure and even for returns by “pooling” returns or by pooling procedures. An OB group in Calgary pools new patients and then re-pools ALL returns. So patients, for the most part, see a new physician with each visit. There are fewer delays. This is a choice, primarily driven by the fact that deliveries are pooled, and they want to ensure that all patients see the physician who will do their delivery at least once. With 12 physicians and 10 visits they do not always achieve this aim. At the same time, this group only does deliveries and office so there is not a great deal of demand stream (queue) competition. So, they pool for new patients, re-pool for returns and re-re-pool for deliveries. This choice values short delays over strict continuity.

Tension between continuity and delays is particularly difficult in complex practices. “Complex” is not an opinion. Complex practices are characterized by multiple competing demand streams to support, or by practices with a high proportion of the work devoted to supporting the immediate demand streams – on-call and hospital. These are generally hospital-based practices with a small number of physicians.

Thus the goal is actually twofold: get patients to the same provider and minimize the delays. These goals can be achieved but only if we focus on operations and build systems to first address the delays and then to deliver continuity.

Question: Discontinuity Visits

In a situation where the patient sees a physician other than his/her own, how should the “non-principal” physician handle these patients in terms of appointment time? If it takes, on average, seven minutes more to see a patient you are not familiar with, should there be longer appointments for these “other” patients? If so, how could that ever be built into a schedule? And it violates the principle of minimizing appointment types. Or do the patient and “non-principal doctor” stick to the standard appointment length and get short-changed by seven minutes for the therapeutic part of the appointment?

Answer:

We do want to minimize appointment types. However, we want to divide up the work from the absent providers and divide it equitably across all providers who are present, rather than just “punishing the one with the most “openings.” In Primary Care (PC), providers manage a permanent stable panel of patients over a long period of time. While some of the workload is pre-booked (“good backlog”), most or at least a significant portion of the daily work is generated spontaneously from that panel population each day. Due to provider absences and due to the way the workload materializes from that patient population, practices can expect a predictable amount of work generated each day from patients of absent providers. This may require having an appointment type for patients of absent providers, and the length of this appointment type may be slightly longer than a regular appointment (whatever we decide). In addition, there is great value in creating a diary where the provider who sees the patient informs the linked provider of the planned actions. This reduces some of the adverse effects of discontinuity.

In Specialty Care (SC), providers, in general, manage patients for a temporary period of time (caseload). In addition, workload is not generated spontaneously as in PC. For the most part the work is planned and channeled into the practice as either new or return appointment types. There is some spontaneous work generated each day by patients with an acute exacerbation of a current condition. Dependent on the entry point (ED, office, hospital) this work is most commonly managed by the on-call function or primarily by the linked provider in the office. Overall, there is less discontinuity in most SC practices due to the nature of how work materializes, the planned aspect of a significant part of the work, and a clear carve out function (on-call) for acute workload. The need, then, for a distinct appointment type for “work of the absent provider” is minimal.