

Reinvention of Community Health Centers: The Movement from Visit to Panel

If we're going to build a healthcare practice or healthcare system that is successful, we have to build this around one fundamental issue: what is it that our patients want, and how can we deliver it? What our patients want is the opportunity to choose a Primary Care Provider, access to that provider when the patient chooses (they don't want to wait for an appointment) and a quality healthcare experience (they don't want to wait at the appointment). Clinical quality is assumed; it is a given. If we can't get our patients into the practice and to their chosen provider without a delay, then we have failed in the fundamental mission. Delays, both for and at an appointment, have significant adverse effects. The longer patients wait, the higher the patient dissatisfaction. The longer patients wait, the higher the staff dissatisfaction, due to all the frustration, rework and redundancy. The longer patients wait, the higher the provider dissatisfaction, due to frustrations in being able to take care of their patients in a timely manner. The longer patients wait, the higher the cost of care, due to the increased rework and redundancy caused by the delay, the necessity to use precious resource to triage and sort the work because of the delay, and finally, the increased no-show rate, which is directly correlated with long waiting times. The longer patients wait, the lower the revenue in our fee-for-service components (although this is counterintuitive, there is a great deal of recent data that supports this). The longer patients wait, the worse the clinical outcomes, whether we're measuring compliance with preventive services, the care and management to patients with chronic illness, or the early detection of serious disease. Thus, the waiting time has tremendous adverse effects on system performance, and improvement in the waiting time will improve system performance in all of these areas. Waiting time improvement is the glue that holds all of this together.

Once the unnecessary waiting time, both for an appointment and at an appointment, is eliminated, then the practice needs to turn its attention to the further refinement of its operations: to care delivery models, to integration of care, etc. Once these operational issues are optimized—that is, we have the right person doing the right work and in the right order—and once the basic infrastructure is built, the practice can

turn its attention to explicitly improving clinical outcomes. Clinical outcome improvement can only be optimized when patients can see their own doctor without delay, and the care team and care delivery model is focused on supporting that relationship between provider and patient. Previous efforts to improve Community Health Centers that focused on “redesign,” which is a euphemism for operational improvement, and improvements in clinical care (the Chronic Care Model) have been done in isolation, without access delay reduction improvements as a platform and have not achieved their full optimization or promise.

In the current Community Health Center environment, there is a perceived contradiction and misalignment of incentive that grows out of the reimbursement methodology. The goal of our patients is simple: “I want the opportunity to choose a Primary Care provider, I want access to the provider when I choose, and I want a quality healthcare experience.” The goal of the organization is, “I want to provide service as my customers desire, and at the same time I need to generate enough revenue to be able to pay for this, in order to stay in business.” Because the current reimbursement methodology is based on cost-based reimbursement, the organizations and Executive Directors falsely believe that what they need to do is to increase the total number of visits into the practice. This is the best way, the feel, to guarantee the revenue in order to stay in business. They then pass this incentive on to the providers, and say to the providers, “Produce 4,200 visits per year, and produce 3 visits per hour.” In light of this directive, and in light of the fact that the providers are reimbursed by salary, the providers’ incentive then is mixed. On the one hand, there is, quite frankly, an incentive for avoidance of workload because of the salary. Secondly, there is a pressure and a directive to increase visits. The providers then, feeling this tension, and after complaining about how their patients are more difficult, will often reluctantly comply by increasing visits. They increase visits by making a patient visit to a provider the final common pathway for all patient issues. For example, if a patient wants a question answered, a medication refilled, a lab test reviewed, etc., this is all done by a visit, rather than other alternative and potentially more satisfying methods of resolution. In this way, the providers can generate their visits and maintain a workload that they can feel comfortably with. The emphasis on the chronic disease model, which stated that

patients with chronic illness need monitoring and follow-up, then can, and has become, an excuse to generate more follow-up visits, because “the chronic care model (CCM) says that I need to bring patients back very frequently.” A closer review of the CCM reveals that monitoring does not equate with provider visit.

Here’s the fundamental contradiction. The Executive Directors and the providers are, in a sense, in collusion to create visits. Visits do not equal value. Visits are a cost in the capitated environment and increased visits actually reduce potential revenue in a fee-for-service environment. On the other hand, visits generate cost-based reimbursement in a cost-based reimbursement environment. Most Community Health Centers have varying levels of these 3 basic payment methodologies as part of their practice. For those with high capitated and fee-for-service components, increasing the visits has an adverse effect on the overall revenues. In addition, as the visits (visits are a manifestation of demand) go up, and many of these visits are clinically and socially unnecessary, then the waiting times rise due to the consequent mismatch of demand for appointment with supply (resource) for appointment, and delays are not addressed. In fact, the best way to guarantee that visits can be generated is by creating a waiting time—a warehouse—where patients are stored temporarily and then trucked over to the practice in order to guarantee a specified number of visits per day. This artificial creation of waiting times has, as I noted above, a tremendous adverse effect on system performance: patient, provider and staff satisfaction is reduced; costs rise, revenues are sub-optimized, and clinical outcomes are adversely affected. Without addressing the delay as the fundamental improvement issue, Community Health Centers will forever be sub-optimized and deliver less-than-satisfying, costly or clinically sub-optimal care.

Transformation. There is a way to resolve this perceived contradiction. If the Community Health Centers commit to building a practice based around what their patients (customers) want: “ I want the opportunity to choose a Primary Care provider, I want access to her when I choose, and I want a quality healthcare experience”, then the delay issue, the access issue, will have to be addressed, since the fundamental patient issue is waiting times. In order to address the delay issue, we need to recognize that the delay represents a mismatch between demand for appointment service and supply of appointment service. So we have to understand that, in order to build a system

without a delay, we need to achieve a balance between the demand for appointments and the supply of appointments, at the organizational level, the practice level and the individual level. Thus, we need to recognize that in order to optimize performance, providers need to be positioned so that the demand for their service does not outstrip their personal supply for that service. The best way to operationalize this is to have providers of care responsible for an equitable panel of patients (the panel is that part of the practice that each provider is responsible for). The demand for appointments from that panel of patients needs to be balanced with the supply of appointments that provider can deliver. In this way, by making a commitment to a panel of patients and a commitment to continuity, that is, seeing those patients, and a commitment to no waiting times, system performance can be optimized. Providers need to have a panel that generates just enough demand and not too much demand. Patients can get what they want, staff can be satisfied, providers can get what they want, costs will go down, revenues will go up, and clinical outcomes can be optimized.

At the same time, this requires a transformation in the Community Health Center perspective. This transformation is the change of incentive for visits (“output equity”- all providers deliver 4200 visits per year) to the incentive to manage a panel of patients (input equity- all providers are responsible for an equitable (not necessarily equal) panel of patients). When a practice or individual providers move from the incentive to manage or deliver a prescribed set of visits to the incentive to manage a panel of patients, the inevitable consequence is reduction of visits. This occurs because the incentive for the providers changes dramatically: when the provider is responsible for a set of patients and the patients can’t wait, the incentive or the advantage for the provider becomes, “How can I comb the future schedule to look for unnecessary visits, how can I do more with each visit, how can I develop telephone treatment protocols, how can I extend visit intervals, and how can I optimize my care team?” Without the incentive to manage a panel, these activities will not occur. The inverse is also true: with high levels of discontinuity, unnecessary visits rise).

When the visits are reduced in this way there are financial consequences. In the capitated component of the practice, since visits are cost, reduction of visits actually saves money over and above operational cost. In the fee-for-service component of the

practice, the visits that are reduced are the low-level RVU visits: patients from one provider who see another provider, and vice versa. The result is that the RVU per visit that remains is much higher than the previous RVU per visit. In addition, when providers are put in a position where they have to see their own patients, and their patients can't wait, what we find is that they do more with every visit, they maximize the efficiency of that visit, and the RVU per visit rises to a greater degree than the loss of revenue from the loss of the lower RVU visits. So consequently, providers see fewer patients and generate more revenue. This increased revenue is over and above the net revenue increase that occurs due to reduction in cost. On the cost-based reimbursement side of the practice, it is clear that less visits equals less revenue. So the revenue in that component falls with less visits. On the other hand, if those visits can be replaced by other needy patients in the community, then this is a break-even. So there will be some visits that are "lost," but when these are replaced by other new, unique, unduplicated patients needing service from the community, this then becomes a break-even. The challenge, of course, for the Executive Directors, is to manage and inspire the providers to see more unique, unduplicated, new patients, so that instead of seeing the same patients over and over again, or seeing other providers' patients and just sending them back to their own provider, the providers need to be managed, or inspired, to see more unique, unduplicated patients. In order to accomplish this, the Executive Directors will have to not only provide that inspiration, leadership and understanding of the overall mission, but they will also have to provide support for those doctors to be able to manage in this new way.

Thus, the reduction in visits that almost inevitably occurs, due to recognition of demand–supply dynamics at the practice and at the individual provider level and due to improved continuity which resonates strongly with our patients' desires, when we address the access issue by reducing waits for service and as we move from focus on visits to focus on panel, will result in a stronger financial position. That position can be fully realized with the consequent reduction in cost, the increased revenue per visits and by replacing "lost" cost based reimbursed visits with new patients.

There are distinct advantages in a transformation to this kind of approach toward management of a panel, not visits:

1. In order to eliminate the delays (improve access) a practice will have to move toward managing a panel, not visits (if we keep doing what we're doing, we're going to keep getting what we've got). And consequently, as the transformation from visits to panel occurs and access improves, then patient, provider and staff satisfaction will improve, costs will go down, revenues will go up and clinical outcomes can be optimized.
2. There are currently approximately 16 million underserved patients in the United States. Only about 12.4 million of these are now currently served by Community Health Centers. With the current staffing, moving from 12.4 million to 16 million is impossible. There are far too many unnecessary visits; there is far too much redundancy. The only way to be able to move in this direction is to streamline the operations and improve the overall aggregate practice panel size. The transformation to this methodology can accomplish that. When the waiting times go down, and the platform for further improvements is built, then projects like "redesign," which focus on improving the care delivery model, can finally be optimized, and clinical care can be optimized. Redesigning operations and attempting to improve clinical care without setting a platform that operationalizes patients being able to see their own provider without delay, is futile.
3. As the visits go down, the Community Health Centers have an opportunity to grow. Not only can they grow into the underserved market, but they can also grow into the other capitated and fee-for-service components, in order to develop a more stable and diversified business structure. With growth, then the visits will rebound. The "goal" of 4200 then becomes a reality, but not as a goal but as an outcome of the correct panel size. The panel will generate 4200 visits and , at the same time, this panel will be larger.

4. As the visits go down and the continuity and responsibility for a panel increases, then the movement toward self-care, self-management, and other alternatives to face-to-face visit methods and structures can be implemented, i.e., group visits, e-mail care, telephone care, care with the nurse, protocols, guidelines, and self-management.
5. As these other alternative forms of care become more predominant, the dependency and paternalism exhibited by visit-based care (all pathways lead to a visit with a provider) will start to melt away. Patients then can become more independent, self-confident, and empowered.

Mark Murray, MD, MPA

Mark Murray & Associates

Reduction of demand in the CHC environment

The common observation in the CHC environment is that while we want to reduce the wait time for appointments and the office cycle time more productive, reducing demand, when we are paid by the visit seems counter productive.

Response:

Actually reducing demand is not counter productive. If the demand per patient is so high as to start to reduce value (churned visits) then this precludes you from either pursuing your mission of serving the underserved or keeps you in continuous and ever expanding BL. The longer patients wait the worse your system performance due to the rework and redundancy necessary to deal with all the wait time issues, due to higher no shows which cause staff rework and unused system capacity (productivity issues) and due to the high resource cost necessary to triage and sort the long waits from the short waits. The delay and waiting times significantly add to the cost of care, and reduce overall system revenue performance.

I would approach somewhat differently:

The work in improving access and optimizing Primary Care delivery requires strategies to promote continuity and reduce wait time for and at services. As an inevitable consequence of the utilization of these strategies, the number of visits per patient per year will be decreased. This will make your executive director nervous. On the other hand, if you can replace those "lost" visits per patient per year with new visits from new patients, external to the practice, or with new visits from patients with unmet needs from inside the practice—that is, women that need PAP smears, etc—then you can either break even or actually increase the number of total visits into the system, increase your market share, create opportunity for more "mixed portfolio" of patients, pursue your mission of caring for the underserved and improve clinical care. But to do this, doctors have to see more unique patients, and you'll have to inspire them to this mission and support them in this regard

As you can see we may have to reduce demand for visits per patients per year but that does not at all mean that total visits per year will be decreased. Systems that incent visits will get just that: visits. But visits are not customer value nor are visits organizational value either even in pay by the visit systems. Visits are an outcome, not a goal. If we make visits the goal and transmit this visit incentive pressure to the providers we get providers who want a BL since having a BL is the best way, they feel to ensure that they can each day deliver the required number of visits. Having no BL they believe puts them at risk for unused capacity due to risk of low demand. So in order to mitigate the risk of low demand and ensure that the patients that they see are not sick, provider in these types of systems will often opt to have a significant BL of not so sick patients. This, as I mentioned above, is very costly. The key change strategy is to shift the goal focus from visits to value and understand that at the right panel size and with the absolute requirement that providers see their own patients and don't make them wait that they will deliver visits but as outcome of the system,. And, at the same time the patients won't wait, the cost will go down as satisfaction and outcomes improve.

Myth of full utilization

Many practices fear working without a warehouse full of scheduled work. This fear focuses on the risk that working without a delay- without the buffer of the warehouse- may result in unused system capacity. And, at the same time, there exists, in parallel, the false belief that a warehouse of work –delayed work waiting to be processed— enhances practice revenues, because this warehouse provides reassurance against any down variation in demand. For example, if there's minimized waiting time and the demand goes down, there's the risk of unused capacity, and the myth is that a warehouse will prevent or mitigate that risk.

Demand variation does exist in healthcare, so working without a wait and completing all the work each day can result in either activity that exceeds demand (up demand) or in activity less than demand (down demand).

Meticulous contingency planning can mitigate demand variation. For example, bringing planned return visits back early in the day, late in the week, changing the ratio of internal to external demand expectations based on the number of providers present and sharing the work of the absent providers are all strategies that level workloads amongst days and between providers.

In addition, the strategy of having a delay in order to ensure a full schedule simply does not work. The delay increases the likelihood of no shows (reduced activity), increases the likelihood of walk-ins (chaos) and requires the use of resource to triage the work.

Optimizing Primary Care Delivery

A gap exists between current healthcare performance and possibility. This gap exists not only in the delivery of effective and evidence-based clinical care, but also exists in the operational delivery of care itself—how the care is delivered—and in the delivery of satisfying care. This gap exists not only between systems and possibility, but also within systems between sites, and within sites between individuals.

Paraphrasing the Institute of Medicine report, the goal for optimized Primary Care practice delivery is to deliver the best care, in the best way, on time, every time. Achieving the ultimate goal, then, involves a complex choreography and requires successful connection of a number of critical components. Previous attempts to narrow the gap between current performance and possibility in all of these dimensions have failed primarily due to the fact that these efforts have been attempted in isolation, or the failures are due to a lack of understanding or a misunderstanding of the basic overall flow dynamic.

Healthcare systems- both operational and clinical care- are flow systems. In flow systems we match the demand for service to the supply (resource) of service. The most effective, the most efficient and the most satisfying flow systems, demand –supply matching systems, work without a wait. The complex choreography of connected supply components can only go as fast as the slowest step or component. At the heart of healthcare clinical and operation improvement lies the reduction of delay.

Our healthcare systems lack relationship and purpose, seem disconnected from the patients that the system intends to serve, and are characterized primarily by a system or internal focus rather than directed by “patient-centered-ness.” The delivery of clinical care is fragmented, isolated, non-systematic, and from the patient’s perspective is often delivered by luck rather than by intention. The relationship that patients desire, the trust that they want to have, and the comprehensive care that they deserve, is far too often non-existent.

Our healthcare systems are plagued by waits and delays that lead to patient, provider and staff dissatisfaction, increase the cost of care, sub-optimize the revenue, and adversely affect clinical care. When we focus on the reduction of delay we are able to simultaneously improve within all these parameters.

Customers and Value

Despite effort to develop patient-centered care, there is often a failure to operationalize this. Investigations and evidence show that patients desire:

- The opportunity to choose a Primary Care provider
- Access to that provider when the patient chooses
- A quality healthcare experience, which means “don’t make me wait” and “respect my time”
- Up-to-date, evidence-based and quality clinical care

Cost is a strong consideration, but cost most commonly, in this country, is borne not by the individual patient but by intermediaries (government funding, insurance companies, etc).

Delivery of Value

In healthcare, every day, all day long, one patient at a time, one service at a time, we match the demand for service with our supply for that service. We live in a flow system. In flow systems, where supply is matched to the demand, the most successful systems will match supply to demand without a waiting time. The most effective (clinical outcome), the most efficient (the cost/revenue ratio) and the most satisfying systems match supply to demand without a wait. At each step in a flow system there is a demand, there is a supply, and there is a waiting time. The work flows. Our goal as stated above can only be achieved if the work flows smoothly, if there's perfect balance all of the time so that value is delivered perfectly on time, every time. Whether we acknowledge this or understand this, this is the basic and fundamental dynamic that exists in flow systems. Matching the demand to the supply is really not a choice. It is the basic dynamic that exists. We either match the demand to the supply poorly, or we match it well—it's not a choice.

Balance

In flow systems, the balance between the demand and the supply is critical. There has to be a balance between the demand for service and the delivery of supply, at the organizational level, the site level, the department level and the individual level. The balance at the individual level is driven by customer desire. Patients want the opportunity to choose a Primary Care provider, and they want access to her when they choose. For that clinician to be successful, she has to have a balance between the workload demand and her workload ability. This workload balance is measured by the determination of the **panel size**, which is the macro level from which demand for service arises. The panel size or the number of customers within the organization, at the site, at the department and at the individual clinician level has to be balanced against opportunity and ability for service delivery. Identification, determination and measurement of panel size at all of these levels are critical for success. In addition, evidence and studies demonstrate that continuity of care—seeing the same identified and linked provider—results in better clinical care. There is no data or evidence or study that has demonstrated that random service delivery (discontinuity) results in improved or even the same level of clinical care. Thus, the two critical system design elements for success in healthcare delivery are continuity and the reduction of waiting time. Designing systems with these elements as system property will result in improved patient, clinician and staff satisfaction, as well as less cost, higher potential revenues, and improved clinical care and outcome.

Delivery of Care

Healthcare is a team sport. Evidence, investigation and study have shown that clinicians cannot provide optimal care working in isolation, and particularly working with visit-by-visit approaches. There is far too much knowledge and far too much information. For successful, optimized clinical care, clinicians need others (teams) to share in the care delivery, information systems to monitor, track and measure care delivery, and systematic approaches for optimized care delivery. The planned or chronic care model provides us with a systematic approach for clinical care delivery. The original model de-emphasized the critical importance of the care delivery features. We have modified the basic planned care model to explicitly address

leadership, care delivery systems, and community outreach as system design elements, and reinforced the value of self-care/self-management, decision support, and information technology features as practice elements. This modified planned care model incorporates explicitly the value of continuity and a no-wait culture.

Clinical Care

We can never optimize clinical care- measured by compliance with preventive service guidelines, the management of patients with chronic illness or by measurement of early detection of serious illness- unless we can get patients to their own provider without waiting. At that point we are in position to optimize clinical care. Just as I takes a “system” to optimize flow for the appointment or other service and at that service, it takes a “system” to optimize clinical care.

Process for Improvement

There are four (4) components within the process for improvement:

- Form a team. The people that do the work need to transform the work. the likelihood of successful change occurring from outside announcements or implementation is small compared to the likelihood of successful change occurring when this is accomplished by the people who actually do the work
- Aim. The team needs a focus or an aim. In supply/demand flow systems, the aim or focus is always on reducing the waiting time, because the most effective, efficient and satisfying systems in flow work without a wait.
- Change. If we keep doing what we're doing, we're going to keep getting what we've got, so in order to make improvement we have to change our actions and behaviors.
- Measure. How do we know that the changes we've made have actually resulted in improvement? We need to measure. The focus on measurement is initially on reduction of waiting time both for and at the service. Measurements of demand, which include measurements of macro demand at the practice panel and individual panel levels, and then the demand that arises from that panel, as well as measurements of supply. In addition, we would measure fail-to-keep appointment (defect) rates and continuity (success) rates, as well as measuring clinical care and clinical outcome.

This process is the same process we would use for both operational as well as clinical improvement. The change strategies are different but the process is the same.

Summary

In order to build the most successful systems for clinical care outcome, we need to build a system that can get patients to their identified clinician without waiting. In order to work without a wait, we need to have a balance between the workload and the worker. The first step in optimization of a Primary Care practice, then, is to build the operational foundation that can ensure that patients can see their own doctor without waiting. The second step, then, is to identify and clarify the operational team whose role is to put the clinician in a position to be

successful. Next, a practice develops a clinical profile to determine the needs of its patients, and then develops a clinical care team that responds to those specific care needs. Development of a system that works without waits requires utilization of a set of principles to eliminate the waiting times for and at the service. The development of teams also includes a set of principles, and the delivery of clinical care is also driven by a set of systematic principles. Putting all of this together requires forming a team, setting an aim, making the systematic changes necessary, and measuring the results of those changes.

Questions and answers: Provider productivity

The following question and answer sequence points out some of the very difficult challenges we face in our Community Health Center environment. Since Centers are reimbursed by the visit, there is an overwhelming tendency to focus directly on producing visits as the performance goal. I recognize the challenges here since Centers, for the most part, only make money with visits. At the same time, a focus on visits- making visits the expressed goal,- creates a misalignment of focus, intent, goal and performance for Centers, their patient customers and provider staff. Every day, all day long, one patient or one request at a time, Centers match customer demand to their capacity. Successful performance in these demand supply systems is always assessed by how well did the system match supply or capacity to demand. Visits are really just an outcome. If flow performance improves visits are the inevitable result. While we could increase visits and, at the same time, worsen system flow performance, if we improve flow we cannot help but increase visits.

The questions arose in sequence. The early answers focus on visits. The final answer offers a wider perspective. This answer is meant to be a companion to other papers written previously.

Question: What is the goal for the average # visits per hour and by the year by provider? Are there different expectations for specialty and or mid-level providers?

Answer 1. : We expect productivity for our Family Practitioners and OB/GYN physicians to be 4,200 visits per year and mid-level providers to be 3,150 visits per year.

Answer 2: We have an office goal of 2.4 patients/hour for our FP (non -OB) physicians. This level of productivity pays the bills. We then pay a small bonus to the provider team for visits greater than 2.3. Mid-levels achieve an additional bonus at greater than 1.9 and doctors achieve an additional bonus at greater than 2.7.

Answer 3: We are starting a productivity based compensation plan with a goal for all FP, IM, Peds, and Specialty physicians of 2.5 patients per hour with a graduated increase in compensation for visits greater than 2.2 patients per hour. For midlevels, the goal is 2.0 patients per hour with gradual increases starting at 1.76. We want to incorporate an RVUs per day component.

My response:

These are common responses. The Center is paid by the visit, so the leadership pushes that incentive directly to the providers and sets "standards". However, since the providers are paid by salary there is a misalignment of incentive: with a fixed salary, there is no direct incentive to produce visits. So the production goals are then indexed to salary plus, that is, get a fixed salary but make more if production in visits rises. I wonder if the plus part of the compensation is gained by lowering the base salaries and taking this out of a total compensation pool: put all the salaries in a pot, lower the base and pay higher for more

visits but take the "higher" out of the total fixed salary pot? Or if they take base salaries and add money to it. If the budget was set based on an expectation of compensation from only less than 2.5 visits per hour, then more visits result in more revenue which gives them the excess. If the budget is fixed on higher visit production expectation, something will break. I also find it interesting that they throw in RVU. It is not clear at all how this fits. I expect that adding in an RVU component is meant to reduce churning of low value visits. In addition, setting a standard of visits per hour creates behaviors and incentives to manipulate the hours per day in order to "elevate" the visits per hour. It is not unusual to see providers work for 8 hours, "count" that as 6 in order to achieve a visit per hour standard.

At the same time, any visit production compensation system always results in bad behaviors by providers. I outlined this in one of my papers. They can't help it. If the incentive is visits, then the behaviors crafted and designed to increase visits (like having a long delay which creates a warehouse of waiting work and like churning which results in high no shows) have adverse system performance effects.

In my view, they miss the point here: that the goal is to get patients in to their own provider without a delay. That behavior will result in the best clinical care, the best satisfaction and the lowest cost. So enterprise incentives ought to be aligned with this goal. That means that the goal is not visits at all but smooth flow. We can get a lot of visits by using long delays, over-booking, and creating provider incentive for visits- but at the same time seriously harm patients. If the goal is flow - match and balance the demand and supply without a wait- then patients have the least risk of error and harm. (what about do no harm?) Visits are an outcome, not the goal. If we want to have more visits, then set an explicit system and provider goal of see your own, don't make them wait, support the providers in this endeavor and set the panel size at the level where the visit "goal" is the inevitable outcome. At the same time, we must recognize that there is a provider capacity limit (much higher than 2.5 visits per hour by the way) and the panels cannot be set at a level so that the resulting demand exceeds that capacity limit. We don't make money when we exceed the limit, we lose money due to the mismatch and delays.

For groups interested in rewarding exemplary behaviors, there are options. If a provider wants to and can manage more visits in a day by forming and keeping a well functioning team, by focusing on the work at hand and by managing the day well, then she can accomplish this by increasing panel. Increased visits are an inevitable outcome of that increase in workload. In this situation I would reward the physician and her team for managing a larger panel.

System performance goals

Every system is perfectly designed to get the results it gets.

Deming

There is a fundamental tension that exists in all healthcare settings that often leads to serious and misguided consequences. This tension has been described as “the myth of 100% utilization” and is manifested in a number of ways. In the hospital Operating Room setting, for example, this tension is manifested by the false belief that the most efficient way to utilize Operating Rooms is to have a full schedule built in advance. This strategy would work perfectly well towards optimizing system performance as measured by throughput (units completed per unit of time), efficiency (cost to revenue ratio) and effectiveness (outcomes) if demand for the OR was matched perfectly with all the supply components needed to successfully complete an OR procedure and if there was no variation in daily or weekly volume, no variation in arrival rates and no variation in case time.

In the clinical outpatient setting this tension is manifested in a similar way by the mistaken belief that pre-filling the entire daily appointment schedule, either far in advance or spontaneously filling any unexpected no show visit with a corresponding unexpected walk-in visit is, again, the most efficient way to work. This pre-filling, use 100% capacity belief occurs in all payment settings. In quintessential fee for service systems (“do more, get more”), the common belief is that the highest net revenues are generated when the schedule is filled in advance. The pre-filling offers the reassurance of continued income. In systems where visits are all paid at essentially the same rate, this myth also exists. The belief here is that the most efficient way to guarantee income, which is a direct result of volume of visits, is to keep the schedule filled. This occurs whether the organization is paid by the visit (Community Health Centers in the US) or the providers themselves are paid by the visit. (the Canadian system). Counter-intuitively, this belief also exists in capitated environments where visits actually represent cost. In many of these environments, and in particular in these environments where the organization is paid by capitation but the providers are paid by salary the belief is that the best way to monitor provider “productivity” is to keep the schedule full and to develop visit productivity standards. In these settings the goal- and this is a misguided goal- is to maximize net revenue and the strategy developed to achieve that goal is to make sure that the schedules are filled to maximum capacity.

There are two fundamental confusions here:

While the goal of revenue maximization is solid and understandable (“no margin, no mission”) this “goal” results from a fundamental misconception and, as such, ought to be viewed as a desired outcome, not a goal. Secondly, the action of filling schedules in advance in order to achieve this desired outcome is not the goal either but a strategy designed to achieve optimal net revenue.

In systems where the stated goal is to optimize revenue using a strategy of pre-filled schedules and 100% full capacity used, there is often, at the same time, a clear recognition that patients want to see the same provider (continuity), that satisfaction of patients and providers alike is improved with that continuity, and that outcomes are improved with that continuity. In many settings there is mixed messages – see your own patients, work towards clinical goals and outcomes and, at the same time, do this while making sure that your schedule is filled to the “productivity standard”. This mixed message can be confusing to providers.

The basic disconnect here occurs due to a lack of understanding of the fundamental dynamic at play in flow systems: in flow systems where we match the demand to the supply, the most effective (outcome) the most efficient (cost to revenue ratio) and the most satisfying systems will match demand to supply without a delay. Successful flow systems focus improvement and goal efforts on the input side, on making the flow work, rather than focusing on the output side. Flow systems see output as an outcome and improvement in output as an inevitable consequence of smooth flow.

The tension between fill the schedule in advance and balance demand to supply without a delay (the fundamental goal in flow systems) gets played out in a number of scenarios:

Manage visits vs. manage a panel

In the “manage visits” approach, the goal, which is measured in visits or uses visits as a surrogate measure for productivity, is to make sure that provider visit productivity is high in relation to capacity. The goal is visits to full capacity. Success is measured by activity (what is done) that either meets or exceeds capacity (supply).

In order to achieve the goal of 100% full schedules, organizations and, in particular, providers, quickly learn that the most effective way to ensure a full schedule of visits is to have patients wait in a warehouse and then aliquot off a full schedule proportion each day. To some extent this strategy might be successful if there was no variation. Variation in volume of demand, variation in “urgency” of demand, and some variation in arrivals- no shows and walk-ins- and variation in supply make this strategy problematic. These types of demand and supply variations characterize healthcare settings. Demand volume variation is dealt with in two ways: first, by appointing those “over demand” patients to another providers “under demand” open schedule. This is done most often by appointing a walk-in from provider A’s panel to replace a no show on provider B’s schedule when provider’s A’s schedule is considered full. Secondly, demand volume variation is dealt with by sending demand volume deeper into the wait time queue. These two actions lead to serious system performance deterioration:

Sending patients deeper into the waiting time increases the warehouse or backlog of work. This backlog raises cost:

- The longer the wait, the higher the cost per visit. This effects the overall net practice revenue

- ❑ The longer the wait, the higher the no show rate. No shows cost staff time and result in unused capacity. That unused capacity is often, as mentioned previously, filled by intentionally over scheduling or by using walk-ins as an indiscriminate random schedule filler
- ❑ The longer the wait the more re-work and redundancy. For example, studies in Call Centers have shown as much as a ten fold increase in call handling time in systems with full schedules and no appointments to offer.
- ❑ The longer the wait, because of the variation in demand urgency, the more the need for triage to determine who can wait and who cannot. Triage uses up precious professional resource.
- ❑ The longer the wait the higher the number of walk-in's. Patients walking in increase the return visit rate by shorten visit lengths and reducing continuity. Walk-in patients also reduce patient satisfaction by increasing office cycle times.

Sending patients from the linked provider to another provider has consequences:

- With discontinuity, we get reduced satisfaction, even when patients “agree” to see the non linked provider
- With discontinuity we lower the revenue per visit and per minute worked
- With discontinuity we adversely affect clinical care and outcome. Studies on clinical care and outcome demonstrate that care and outcome is improved, and can only hope to be optimized, with improved continuity.
- With discontinuity the visit length is extended due to the time necessary to establish rapport, credibility and obtain a history required in the discontinuity visit and not required in the continuity visit. As a consequence, with visit length increase, the number of visits in a day is reduced. So counter-intuitively by trying to fill schedules with un-linked patients in order to ensure high visit rates actually contributes to lowered clinic visit capacity due to longer visit lengths.
- With higher discontinuity we get higher return visits since the usual behavior when patients see a non-linked provider is for that provider to send patients back to their own provider. While this “system churn” may not seem all that bad in a pay by the visit environment, this behavior leads to lower RVU per both visits in the quintessential FFS system and in a pay by the visit environment, this fills the schedule with less than fully useful visits and precludes panel size growth from outside the current practice and precludes opportunity to complete unmet clinical care for current patients.

On the other hand, with a focus on eliminating delay as the overall goal for improvement in flow, a key recognition is that balancing demand to supply not only has to occur at the practice level but more importantly at the individual provider level. Continuity is critical: with better continuity we get better satisfaction, less cost, more revenue per minute and per visit and better clinical care and outcome. Operationalizing continuity is often described as managing the panel. Managing the panel means balancing the demand workload from that panel to the individual provider supply and managing that demand to supply ratio with as short a wait as possible. For Primary Care “as short as possible” most often means balancing the daily demand with the daily supply, that is, “doing all today’s

work today”. Managing the panel then shifts the work focus to the input side- to the panel and away from the output side- the visits. Variation has a tremendous effect here. In systems focused on output- on visits, the goal is to reduce variation in output, particularly to reduce down variation in output (the risk of unfilled appointment slots) and to always reach the productivity standard. So as the output- the visits- the productivity becomes fixed or set as a goal, and the effect of the variation is borne in the number of internal deflections or deflections of workload over the production visit standard into the waiting queue. In effect then the customer absorbs the variation either by being put deeper into the wait time or by being directed away from the linked provider to another provider in order to fill an appointment slot. The consequences of that are outlined above. In a sense then, the goal of visit focused systems is to create “output equity”- equal or fair outputs as measured by equal and standard visits for all providers. This goal is most often seen in environments where the providers are on salary. With a fixed salary the goal is often to fix a production standard against that fixed salary. At the same time, we often see this goal expressed in capitated organizational environments as well where the providers are salaried. This approach constitutes a huge disconnect in incentive and subsequent behaviors between the organization and the salaried providers.

On the other hand, when the goal and incentive is shifted away from “output equity” to “input equity”- that is, panel sizes correlated to time worked, aligned with the twin goals of no delays and continuity (see your own and don’t make them wait) then organizational and provider incentives are aligned and the consequences are far different. Excess visits are reduced, costs are lowered, satisfaction and net revenue rises and outcomes are improved. This shift though from a focus on output and visits to input in panels shifts the burden of variation from the customer having to absorb variation to the individual provider having to absorb the variation. For example, with up or down variation in demand volume, and with a strict prescription to see you own and don’t make them wait, then some days volume of demand and hence activity rises and other days, falls. Of interest here though, is that studies of visit activity in output visit focused systems show that due to the variation in the ratio of no shows to walk-ins that the providers will often have activity (see) more patients on some days and less on others. So while here is great fear of the effects of providers absorbing the variation- and the fears are off too much or too little daily work, this variation is often of less range than what is seen in output focused situations.

As managing the panel shifts the goal from filling the schedule with visits towards a goal of balancing demand and supply and working without a delay, the visits then become an outcome, not the goal. The panel size drives the demand and the activity. So panel can be adjusted to achieve visits as inevitable outcome. The goal for optimization in any flow system is to balance the demand and supply, and work without a wait. Reducing variation is a key strategy to achieve this goal. Systems that make the goal an output goal, based on visit or RVU suffer from a serious and often fatal misreading of the basic fundamental dynamic at play here. Thus production, visit, output goals are not a reasonable option but, in reality, represent a fundamental misreading of the dynamic.

There are a lot of semantic scenarios with which to view this dichotomy: managing visits vs. managing a panel, output equity vs. input equity, the question of who absorbs the variation, patients or providers, or goal vs. outcome. At the end though, no matter what

we call the dichotomy, the final conclusion is the same: when we focus on visits or production as the goal we create cost, cost per visit and system disturbance that overcomes any risk of an unused appointment slot that we might see when we focus on input and panel management. In demand and supply systems, and make no mistake about this, this is what we do in our healthcare systems- we match demand to supply- to achieve optimization of system performance we have to focus on flow. The outcomes in visits, in productivity, in net revenue will inevitably follow.

The fear of too much demand and the fear of too little (Myth of full utilization)

The fear of going to work without a full schedule is real. Some groups fear too much demand and an overwhelmed capacity, while others fear not enough demand and unused capacity. These fears focus on too much or too little in volume of demand as well as too much or too little demand due to variation- temporary mismatches of demand due to daily variations.

The fear of too much demand

Some practices fear too much demand and fear that when they get the wait time to zero and start the day with unused capacity, that demand will overwhelm them and that they simply will not be able to keep up. Their fear is that each days demand will overwhelm them.

To overcome these fears, these groups need to measure the following:

- practice and individual panel size
- calculate last years visit rate and determine with current performance (panel times visit rate) can they balance that demand with their current supply (provider days per year times patients per provider per day)
- if, in this equation, demand exceeds supply, these groups need to use strategy to reduce demand per patient per year or increase provider visits per day and achieve a balance of the equation
- then, by using data and measurement, these groups need to reduce demand and supply variation and commit to flexing supply as needed in order to maintain daily balance with demand

The fear of too little demand

On the other hand, other practices, primarily where reimbursement is the same for each visit, also fear working without a warehouse full of scheduled work but for an entirely different reason. This fear focuses on the risk that working without a delay (a TNA of zero) - without the buffer of the warehouse- because of the risk of downward variation in demand, may result in unused system capacity and that unused system capacity will result in sub-optimized revenues. And, at the same time, there exists, in parallel, the false belief that a warehouse of work –delayed work waiting to be processed—enhances practice revenues, because this warehouse provides reassurance against any down variation in demand. For example, if there's minimized waiting time and the demand goes down, there's the risk of unused capacity, and the myth is that a warehouse will prevent or mitigate that risk.

The underlying issue here is the issue of variation. If patients became ill every 15 minutes, traveled for 15 minutes to see us and were seen every 15 minutes and that happened 25 times each and every day, we would have no problem. But demand variation does exist in healthcare: there is variation in daily volume, variation in arrival times and variation in handling time. In addition, there is actually more supply variation: each day we have varying amounts of appointment supply. If a practice commits to working without a wait, that is, maintaining a short and fixed waiting time, (completing all the work within a fixed time frame, like a day) then variation in either demand or supply will result in either activity (workload) that exceeds demand (up demand) or in activity less than demand (down demand).

Thus, if the practice commits to high service level, that is, minimal waits , because of variation the practice will have some days of high activity and some days of low activity in relation to the mean (average) activity. On the other hand, if the practice commits to a fixed workload and uses a pre-booked full schedule to try to achieve this, then service levels or wait times will vary.

Practices that fear the risk of unused capacity due to down variation and working with a minimal wait, will use a backlog of work and a full schedule to provide reassurance against that risk. Their belief is that systems work most efficiently when capacity is pre-filled and utilization of that capacity is always at 100%. This belief is incorrect. We call it the "myth of 100% utilization". The strategy of having a delay in order to ensure a full schedule simply does not work.

Let's think this through logically:

1. visualize the wait time like a lake of work. If the height of the lake is stable, then the workload going in (demand) = the workload done (supply - actually supply used or activity) This lake of work is expensive: higher no shows with longer waits, more need to use resource to "triage"(sort) out work when there is a delay and more staff work, frustration, redundancy when there is a wait time.
2. As the height of the lake goes down with BL reduction- doing more work than work that presents itself (activity > demand) then the height of the lake gets closer to the bottom of the lake
3. At some point when all today's work is completed today, the lake of work is virtually eliminated except for the puddles of good BL.
4. We know, though, that the panel- either for the practice or for individuals- will create demand each day. So when we get to tomorrow, there will be demand. Sometimes that demand will fill the schedules perfectly, sometimes the demand will over fill the schedules (up demand variation) and sometimes the demand will not fill the schedules. (down demand variation)
5. Not only is there demand variation but there is supply variation- the size of the lake bed changes- some days we have more providers and some days less- so our lake capacity changes each day.
6. So we have not only demand variation but supply variation as well. The rate at which the lake fills then is dependent on those two variables: a. demand variation and b. supply variation. We can influence the rate of filling by:
 - a. predicting through measurement the amount of external demand
 - b. manipulating the "good backlog"- (internal demand) bringing patients in early in the day, later in the week and by scheduling less returns on days when there are fewer providers (thus, leaving more space for the predicted external demand)
 - c. being flexible in our supply- seeing more patients when the demand is up or supply is down.Keep in mind that we already do this with the ratio of walk-ins to no shows.

Reducing BL and doing the work today without a lake, for the most part does not change your demand. If demand = supply you are just doing the same work but closer to the bottom of the lake rather than at a distance.

There is a choice here and you can measure the consequences of the choice.

1. you can work with a BL- a lake of work that fills the schedules in advance. The consequence is that patients are dissatisfied, staff is dissatisfied, act out, get frustrated, providers are dissatisfied, patients are shifted to non linked providers since the linked provider is full and the non linked has an opening, the cost of care for the practice rises (rework, redundancy, cost of triage, high no shows and poor office flow), the revenues go down (with longer visit lengths due to unhappy patients, due to high discontinuity) and clinical care suffers due to rushed work and poor continuity. This is the choice discussed above: the choice to keep schedules full in advance in an attempt to guarantee no unused capacity. But system performance (high no shows etc.) and provider behaviors (self protection) actually result in highly variable utilization: some days with activity (high walk ins and few no shows) and some days result in low activity (low walk-ins and high no shows)
2. you can work without a BL and take the risk of unfilled appointment slots. But, patients are happier, staff is happier, providers are "happier", the cost of care goes down, the

revenue rises and clinical care improves. Variation in activity does exist but, in our experience, the range of variation is no greater than the range in alternative 1.

In order to see what "works best", I would measure this: what is the cost of care in alternative 1, compare to the cost of care in alternative 2.

With shorter visit lengths due to better continuity, net revenues can be the same with less visits in alternative 2. In addition, with alternative 2, if the practice absorbs variation, some days schedules will over fill and some days they will under fill. Look at the activity- the end of the day, not the beginning. Somehow we are reassured if the schedule is full at the beginning of the day and ignore the fact that often that schedule is not filled at the end due to high no shows. This is acceptable, I suppose, because we could not control it. But we get nervous if the schedule is not filled at the beginning and remains unfilled at the end. This, somehow, is "our fault" whereas the initial scenario is not. This is not about fault or a false sense of under control. Look at the activity- the supply used- but look over time, not one day at a time. Alternative 2 will result in less cost and, in quintessential fee for service (do more, get more) and in capitated payment systems, will result in higher net revenues. In pay by the visit systems, there will be less visits in alternative 2. This is due to the reduction of the waste of discontinuity. With better continuity and the incentives that grow from providers seeing their own and not making them wait (incentives to do more with each visit, to extend visit intervals, to use teams etc) will result in less visits per patient per year. At the same time this reduction in visits per patient per year can be a survival strategy if panel leads to demand that exceeds supply or offers the opportunity to grow the practice. This growth can be accomplished in 2 ways: external growth- new business with new patients or internal growth- discovering the current patients with un- met clinical needs and through visits managing those needs.

The key to all of this is measurement: what is the practice and individual panel size, what is the predicted visit rate (we can get last years rate quite easily). This then can tell us if we have enough patients to fill the lake. Second, see the work over time. If we look at a day at a time and the day overfills or under fills we can easily get panicked about isolated events. Third, measure the net revenue. We often forget the cost of alternative 1. Fourth, learn how to manage variation so that the workload is not too much or too little. This requires measuring, predicting from past activity, planning and being flexible. This also requires the avoidance of over concern when for one day the workload drops.

The fear of working without a wait thus has a common origin: knowledge that if a practice commits to working with a minimal wait, demand or supply variation can result in too much daily work or too little workload. Meticulous contingency planning can mitigate demand and supply variation. For example, bringing planned return visits back early in the day, late in the week, changing the ratio of internal to external demand expectations based on the number of providers present and sharing the work of the absent providers are all strategies that level workloads amongst days and between providers. Provider supply can be planned in advance and can be flexed as needed.

Visits are fixed, waits are variable

The panel size issue is tricky-

1. the size of the panel changes each month -some patients leave, some join
2. the acuity changes- some get sicker, some get well
3. some providers do more with each visit, some less
4. some providers bring patients back more frequently, some less frequently

What we used to do at Kaiser- and this is common- is to try to take the patient and provider variability out of this by requiring a productivity standard. In our case, this was 25 visits per doctor per day. We got close to meeting this- most providers saw about 23-24 visits per day. And we could pre-rate bonuses based on visits. We finally came to our senses when we realized that first off, Kaiser is capitated so visits are cost. Each visit cost us money so we were incenting absolutely the wrong thing.

Besides that though and deeper than that we realized that we were creating a culture and behaviors that were counter productive and dangerous to patients. Because we had a fixed point for visits -25- we shifted the variability clearly to providers and patients:

1. panels were extremely variable: some providers had large panels and 25 visits while others had small panels and 25 visits
2. visit rates were extremely variable: some providers patients returned 9 times a year, others 2.1 times. Incidentally the patient satisfaction was not higher with the 9 times providers.
3. Continuity was variable- some providers saw their own patients, some did not.
4. satisfaction was highly variable and not correlated to panel size of visits. But was correlated to continuity
5. Cost per provider was extremely variable- some providers cost us 10 times more than other doctors. But the visits were the same !
6. Clinical care was highly variable- preventive screening rates and management of patients with chronic illness was not correlated to visits, to panel size, to satisfaction or visit rates. It was correlated to continuity.
7. Waiting times- which correlated to satisfaction, cost and clinical care (short waits improved all these measures) - were highly variable as well

So we said: **see your own** - which improved clinical care, satisfaction and cost - and **don't make them wait** - which improved clinical care, cost, satisfaction. But to do that we had to give up a focus on visits. Then in order to make sure that the providers had enough work to do (and in your case , enough visits to pay for the enterprise-) we had to monitor panel size.

We told the providers that we would not watch visits any longer, but that they had to see their own and their own could not wait. Visits then were a tool to get the work done and, counter intuitively, the visits either stayed the same or went down. In that environment, less visits became the reward for doing a good job with the panel. How can we reconcile this with a pay by the visit based system?

1. First, not all patients are pay by the visit- there is a mixed portfolio of customer types. The reduction of visits makes sense in capitation and in the low value visits in RVU based systems
- 2 we can raise panel and visits will rise. One correlates clearly with the other
3. To achieve a balance of just enough visits to pay the rent, but not so much as to burn out the doctors or destroy incentive, we need to a. inspire the providers to the mission of the practice and

b. give the providers the support needed to accomplish this.

Ultimately we need a clear partnership between enterprise and providers in order to walk that fine line.

I think it is tough to open and close panels. Panel size does change monthly, so I think we need to either re-calculate each month or have an automatic report from the provider field that gives us a monthly running panel. Then we know who is rising, who is falling and who is the same.

I think it is valuable as well to see provider variability: not just in panel but in visits, in visit rate etc. This helps to normalize and standardize behaviors and helps us discover the best way to practice. Does it help diabetic care to have 10 visits per year? Does it hurt outcome to have 1 visit per year?

I also see a problem if we can see that 2 providers are clearly over-paneled and one is not and all the new patients go to the one with lowest panel. But this creates an expanding wait time for these new patients. Actually what I would do there is to get "non-provider" type work away from the over-paneled providers, get them some small space to see new patients and at the same time, eliminate the new patient appointment limit for the under-paneled provider. (the new patient wait time/list is in part due to a restriction on the number of new patients permitted on the schedule template) And I would get rid of the wait time for all patients. There is a BL here. All these issues are moving at once so it takes some time, some patience, some persistence and a committed plan.

The no shows are an issue:

1. NS's are directly related to wait time for an appointment
2. NS's are related to loyalty
3. NS's are related to the trade off of patient need and expected experience- if I have to wait 4 hours, I can't do it.

With regards to "budgetary restrictions"- are you required to pay a set base salary? if you don't, will you be adversely effected in recruitment?

Do you have a required budget per provider or is this discretionary? Could you have a set salary at 80% of current and then "allow" increases above the current 100%? anything over 100% has to be paid for by reduced cost or increased revenue.

Is there any way to get some not so perfect data now prior to the management system?

Mark