Process Improvement and Supply and Demand: The Elements That Underlie Integration

Mark Murray

Abstract

Integration relies on a series of key change strategies connected by a fundamental dynamic: system capacity has to match demand or it will ultimately result in expanding delay and system failure. A balance of supply and demand is necessary for successful system performance. If demand and capacity are balanced, then delays are not required.

Integration

Integration as a system strategy has been evolving within health systems across Canada. In Alberta, it was first a way of bringing together services within and between the former system of regional health entities, and now within the province-wide Alberta Health Services amalgamation of the former health regions, the Alberta Mental Health Board and the Alberta Cancer Board. The intent of these integration efforts was described in the former Capital Health region as a focus on "building stronger connections between health services, people and providers to better support people in their care journey and realize the benefits of a regional health system" (Abbott 1999: 13).

These integration efforts identified four "key change strategies" as central to the process of better integrating services and achieving best practice:

1. Providing people-centred care
2. Reducing clinical variance
3. Organizing the care continuum
4. Process improvement

While these four key strategies seem to act as independent perspectives on integration, they are, however, implicitly connected. In order to realize the full potential of integration, it is critical to convert the implicit connection to an explicit one. The fundamental underlying dynamic in healthcare is relatively simple: every day, all day long, and one person or service at a time, we use our system capacity to meet customer demand. We either perform this function well or poorly. While system performance is a choice, matching capacity to demand is not.

Operationally, healthcare is no different than any other flow system where customer demand flows through a series of interrelated, interconnected people or process steps as that demand traverses the system. Each step has a demand, a supply/capacity, an activity and a delay. System performance is assessed or gauged by measuring the delay, either at each step or for the

series of interconnected steps. The measure of delay demonstrates how well our systems function in matching demand to capacity. Permanent mismatch of demand to supply will result in expanding delay and system failure. A balance of demand and supply is therefore required for successful system performance. If demand and capacity are balanced, then we simply do not need a delay. The four key strategies are merely external manifestations of this basic underlying dynamic.

...a key component of people-centred care must focus on meeting customer demand without delay. All efforts to foster people-centred care are meaningless and unachievable if a demand/supply mismatch exists.

Providing People-Centred Care

While this strategy focuses on efforts to optimize patient understanding of and participation in the care journey, a key component of people-centred care must focus on meeting customer demand without delay. All efforts to foster people-centred care are meaningless and unachievable if a demand/supply mismatch exists. There are two critical issues here: First, can the organization actually deliver on the promises to deliver care? Is there enough capacity to accomplish this? And second, if there is enough measurable capacity, can the organization accomplish these tasks without a delay? If measured demand can be balanced and met by a corresponding measured supply, then there is no need for delay. If demand exceeds capacity, there is no solution. Attempts at triage, priority or sorting are misguided efforts to deal with either a real or perceived mismatch, and serve only to degrade system performance. Arbitrary delays just increase cost, increase the risk of "no-show" for the requested service and use up precious supply resources simply to sort the work. Either demand is balanced by supply, or it is not.

At the same time, people-centred care does not mean that the individual customer is always right or that every customer always gets what he or she wants. There are measurable capacity limits for both individual and practice. If those practice or individual limits are exceeded – that is, if the measured demand exceeds the capacity – then the individual or practice simply cannot perform the expected work tasks. Intentionally delaying the work through priority mechanisms does not change this dynamic. "People-centred" means that organizations need to measure and understand those limits and work to improve capacity and capability but, most importantly, make those limits clear and explicit.

People do not want to be presented with false dilemmas such as "you can have quality but you have to wait," or "you can have choice but you have to wait." Organizations will often pose the false choice of choosing your own provider versus a delay. Providers of care, departments or services all have measurable capacity limits, and if the capacity can meet the demand then there is no need, with the exception of predictable supply absence, to tolerate a delay. Promising service to a patient when demand exceeds supply, and using a delay to accomplish this, makes no sense. If demand exceeds supply and promises are made to patients to meet the demand, these are false promises, as some unknown random demand will be neglected. It is not patient-centred to offer a service that exceeds provider, enterprise or system capacity limit. Systems need to solve problems for the many, sometimes at the expense of individual preference.

Example: We worked with a primary care practice in the Southern United States. One of the physicians is a woman who was quite "popular." New patients were accepted into the practice based on "choice" and popularity. There was no measurement of physician capacity limit. The refrain from the practice was that this physician was "busy" and "was popular with all of women patients and we are all about being patient-centred." There were no measurements of system performance. Delays were reported as "a long time" and "not very long." There was no formal concept of a panel but, instead, there was a loose and ambiguous "promise" of a relationship. When we finally measured system performance, we found that the physician had a panel size that far exceeded her capacity to complete the work, that delays for her were extended (anywhere from 30 to 360 days), that the practice had initiated an elaborate priority system that they used to bargain with patients and that close to 25% of this physician's patients cut the queue and saw her colleagues. As a consequence, the lowest priority was for women with prevention and surveillance needs. These women were pushed, by priority and false "choice," beyond the recommended threshold for these preventive services. Within this cohort of patients with delayed care, we "discovered" five patients with breast cancer and two with cervical cancer. The physician, of course, stated that "this is not my fault," and "these patients choose to wait." Systems with an inherent mismatch of demand to supply will always fail.

Too often, *patient-centred* is a vague and loose term that describes what the supply or system determines patient needs to be. Supply dictates to demand. This is quite simple: Investigations, surveys and studies all reveal the same concerns. Patients want the opportunity to choose their provider or venue of care; they want to have access to that provider or venue when they choose, not when the system says it is "possible"; and they want a quality healthcare experience. In common colloquium: "let me choose, don't make me wait to enter the system at any point, and don't make me wait at the point of care" (Murray and Berwick 2003; Murray et al. 2003). There are, of course, other considerations in any discussion of people-centred care: the

quality of the care itself, participation opportunities, promises and reliable systems, information and support. But these considerations are not possible without the fundamental underlying balance of patient demand and system capacity.

Reducing Clinical Variance

Variation is a problem in all flow systems. Variation arises from the demand or supply side, whether operational or clinical, and creates flow turbulence. Variation can represent a temporary mismatch of demand and supply, resulting in either unused capacity when supply exceeds demand or a delay when demand exceeds supply. Multiple non-standardized processes are a manifestation of variation. Variation results in multiple smaller channels of work, which in turn increase the risk of mismatch with the consequent unused capacity or delay, and increase the risk of error and the need to repeat demand.

Clinical care variation, manifested by multiple care processes, leads to errors that in and of themselves can be harmful, but from a flow perspective an error represents a demand that has to be repeated

Variation in clinical care functions in a similar manner. Clinical care variation, manifested by multiple care processes, leads to errors that in and of themselves can be harmful, but from a flow perspective an error represents a demand that has to be repeated (Walley et al. 2006).

The antidote then is clear: reduce the variation. Standardized, non-variable clinical care is characterized by a series of interrelated steps or interventions – tests, procedures or treatments – organized in a prescribed sequence in order to achieve an aim of measurable optimized outcome. Both the process as a set of sequential steps, and each individual step itself, require harmonic convergence of a number of critical supply components. The ultimate governor of flow is the patient's physiology. The work cannot move any faster than that physiology. At the same time, the process and the steps can only proceed as fast as the slowest, most delayed of those components. Clinical interventions are all crafted to accelerate, supplement or support the patient's physiology. Hence, underneath the interventions, decisions and treatments, clinical care is subject to the same operational dynamic of demand and supply matching. Clinical care can never be fully optimized unless the demand can be moved to the right supply, right on time. Once the flow dynamic, the matching of the supply and demand, is accomplished, in order to begin to improve clinical care processes, process variation must also be eliminated.

Example: We worked with a specialty care practice in which demand entered the practice as referrals primarily from primary care. While the referral demand exhibited some variation in the volume range of referrals, this variation was analyzed and found through statistical process control methods to be "natural variation" – a variation that is inherent to the system. The only way to deal with natural variation is to flex capacity to meet up- or downswings of demand. On the other hand, an analysis of the office supply demonstrated a wide range of office appointment availability. This variation was found to be artificial, that is, created by intentional actions within the system. The best method to deal with this is to plan. These findings surprised the practice since they thought that the sole source of variation and the cause of the oscillating delays was the variation in demand. Reducing the artificial variation caused by the supply helped the practice keep up with the demand and work with a minimal wait. In this practice, variation in flow created variation in clinical care process. Once the flow variations were eliminated, the practice developed service agreements, which minimized the clinical care process variation and allowed for improvement.

Organizing the Care Continuum

The flow dynamic discussed extensively above clearly applies to the key strategy of organizing the care continuum. The explosion of healthcare knowledge and customer expectation has made it impossible for the current cadre of clinicians to keep up with workload demands. This mismatch is particularly acute in the primary care setting. Standardization of process, the development of techniques and technologies to share information, and the introduction of multi-disciplinary team approaches to care will be essential to meet these needs (Bodenheimer et al. 2004). These enhancements to care delivery will not be successful unless optimal system performance is guaranteed by a demand/supply balance. With more potential hand-offs, there is an increased risk of error and delay. People view system performance as a sum of the waits. With increased numbers of clinicians and processes involved in the care continuum, paying attention and measuring system performance at every step is critical. Successful care can only proceed at the rate of the slowest step.

The same conditions exist in the acute care hospital setting. Here, however, there are far more hand-offs, far more customized journeys. Much of this work can be standardized and "leveraged" through a multi-disciplinary focus, and all of it can be measured. People moving through these complex systems need to be guided by predetermined "trip plans" that outline the journey, the expectations and the prescribed sequence of events. Measurement in these complex venues is just too great for the isolated human brain and requires more sophisticated tools to gauge, assess, measure and monitor basic system performance. These tools need to measure and display flow of work in real

time, as well as using past behaviours and actions to model and predict future events. The entire continuum needs to be investigated. Individual, isolated solutions will often just move the delay to the next silo or next step and not solve the flow for the customer. For "continuum" improvement, all steps need to work together, which requires a common measurement system: was the customer demand met by system capacity at each step and at the sum of all steps (Bergeson and Dean 2006; Walley et al. 2006)?

Example: Many acute care improvement efforts focus on "fixing" a single isolated part of a flow system. Poor acute care system flow is commonly manifested at the first step – the emergency department (ED). Constraints deeper within the system create a bottleneck, and the work backs up into the ED. One common strategy for reducing the impaction and crowding in the ED is to implement an "express admission unit" (EAU). This is a physical place where patients who have completed their ED evaluation and need to be admitted are sent. EAUs are commonly staffed with personnel from the bed units, and patients are held there until a bed opens. This all sounds fine. The work is moved out of the ED, and the overcrowding in that venue is relieved. But what is the system effect? The EAU acts as a holding tank, drawing resources away from patient care in the next step – the bed and floor. There is another risky hand-off from the EAU to the floor, and the patient's total length of stay (LOS) actually increases. The extension of LOS fills more beds for more days, resulting in an even higher likelihood of more bed constraint.

In the past two decades, a number of improvement strategies that have evolved outside healthcare (primarily in "Industry") have been applied in healthcare settings. These improvement strategies have had the advantage of internal consistency and for the most part have a structure that links aim to change and to measure.

Improving Process Management

Healthcare has struggled for years with improvement. In the past, most improvement efforts were based on anecdote, opinion and "feelings." There was no common unifying philosophy or any consistent method to determine whether the changes proposed or implemented actually resulted in improvement. "Improvement" meant change, but that change was most often an isolated event unconnected to any previous event. The aim or goal was commonly vague and nebulous, and there was only infrequent measurement to assure that the change actually resulted in improvement toward a clear, quantifiable aim.

In the past two decades, a number of improvement strategies that have evolved outside healthcare (primarily in "Industry") have been applied in healthcare settings. These improvement strategies have had the advantage of internal consistency and for the most part have a structure that links aim to change and to measure. These methodologies have been used to address multiple operational processes, including centralization of services such as "central booking"; development of standard processes for admissions, transfers, referrals and discharges; and bed and length-of-stay management, case management and discharge coordination (Nolan et al. 1996).

These improvement methodologies have included:

- **Total Quality Management:** In simple terms, TQM refers to "getting products and services right the first time, rather than waiting for them to be finished before checking for errors."
- **Re-engineering:** Re-engineering is an attempt to break an organization down into component parts and then put it back together in a new and more "efficient" way. All processes are flow-mapped, redundancies are identified and removed, and disparate silo processes are identified and combined. Processes are more important than product: indeed, good products and outcomes should naturally follow good processes.
- **Queuing Methods:** Queuing looks at lines: how demand meets supply. Queuing focuses primarily on static systems where supply is fixed and demand varies, and offers insight on the trade-off between demand and/or supply variation and service levels (delays). While queuing methods tend to focus on retrospective events, more sophisticated queuing methods offer views of how current systems function and offer analysis that can be applied to strategies for improvement.
- **Theory of Constraints:** TOC, using the premise that a system can flow only as fast as the slowest component, offers insights into flow both across systems and through smaller processes within a system.
- **Model for Improvement:** This model, used extensively by the Institute for Healthcare Improvement, is characterized by "Plan, Do, Study, Act" (PDSA) cycles. The model focuses on the connections between aim, change and measure.
- **System of Profound Knowledge:** SoPK, originated by Deming in the 1980s, contends that "quality" equals value for all stakeholders, including society, and that value is defined by these stakeholders. SoPK has four interlocking components: understanding or appreciation of the system (how the parts fit together), understanding of variation (ability to distinguish common from special-cause variation and to act accordingly), theory of knowledge (understanding that knowledge is built on theory and predictions; informa-

tion is not knowledge) and psychology (understanding of people, interactions between people and circumstances).

- **Six Sigma:** Popularized by Motorola, Six Sigma looks at process, "system" or event; the mean performance of that process, system or event; and the variance in performance and then identifies the standard deviation from the mean, and whether that process is in control and exhibits natural, common-cause variation, or is out of control, exhibiting unnatural, artificial variance.
- **Lean Thinking:** The Lean method identifies the "value stream" from the customer perspective and seeks to eliminate all waste from the system. "Waste" includes waste of time, caused by demand/supply mismatch. Lean has a clear focus on value and on "pull" systems, wherein work is pulled from Step 1 by Step 2 rather than pushed forward from Step 1 into Step 2. Lean seeks perfection in flow across the value stream.
- **Lean/Six Sigma:** Combining the Lean-equals-zero-waste approach and the Six Sigma-equals-zero-variation approach, Lean/Six Sigma creates synergies and a more robust set of change strategies.

At their core, all these improvement strategies indirectly address the same operational reality: how does a system, an organization or a business enterprise successfully match customer demand to system capacity, and, secondly, how is that accomplished with minimal delay? While matching demand to supply is universally implied in all of these strategies, it is not made explicit. This is due to an instinctual knowledge of how things work. Matching customer demand to system capacity and doing so without a delay is considered obvious, and that knowledge is assumed.

Total Quality Management looks at "getting the product right, the first time," which is essentially a demand reduction strategy. The process flow-map component of **Re-engineering** seeks to reduce redundancy and to standardize for reliability. Both these strategies serve to reduce demand and result in improved demand-to-supply match. **Queuing methods** clearly address matching issues and focus primarily on "service level": how service levels deteriorate or delays accumulate due to poor matching. These methods, in addition, explore multiple levels and types of variation – in volume of demand or supply, in arrival rates and in server time – and illustrate the consequences of that variation. **Theory of Constraints** investigates how demand meets supply, either as a series of interrelated steps or at a single point where more than one supply component is needed to successfully complete the process step. TOC addresses customer delays as a result of either a single process delay in a chain of multiple processes, or a supply component delay at a single step. The **Model for Improvement** only obliquely addresses demand and supply, but does utilize many of the other methods with the change and measure components. The **System of Profound Knowledge** not only addresses variation but emphasizes worker knowledge of the process and context. This knowledge starts to make matching demand and supply much more explicit. **Six Sigma** focuses on variation. Variation is a temporary mismatch of demand and supply. A reduction in variation results in a better match and smoother flow. **Lean Thinking** actually maps the flow of demand as that demand moves through supply gates and seeks explicitly to eliminate waste, including the waste of time. Lean emphasizes continuous flow through demand/supply matches at each step, identification of constraint to that flow, error-proofing to reduce demand, and layout optimization and planning. **Lean/Six Sigma** combines these last two methodologies for a more focused view of variation at each step in the "value stream."

While the work in healthcare shares the same basic fundamental dynamic as many other businesses and industries, there is a common misconception that "we are different." This false belief allows healthcare demand/supply matching to escape scrutiny. These improvement methodologies are often applied in healthcare but not at full potential value. The fact that these methods are clearly crafted to investigate efficiencies in matching demand to supply is lost.

While the work in healthcare shares the same basic fundamental dynamic as many other businesses and industries, there is a common misconception that "we are different." This false belief allows healthcare demand/supply matching to escape scrutiny.

The methods are only tools – lenses through which to see how systems perform. The greatest value for these tools comes when the tools are applied in an integrated combination and not in isolation in order to explicitly view demand and supply dynamics.

Healthcare system performance is often measured by revenue, cost, satisfaction or clinical outcome. These are superficial indirect measures of performance. None of these measures can be optimized unless the fundamental issue was met: did the system successfully match customer demand to customer supply? Successful performance in that arena sets the stage for optimization in all the other areas. An organization may perform well in a single isolated area, such as patient satisfaction or revenue generation, but in so doing may sub-optimize overall system performance. One area is "elevated" to the detriment of all other areas. Successful demand/supply balance is the glue that holds all system performance together, and balance is foundational.

The most successful system performance improvements will

be achieved when an organization can integrate components of all these improvement methods. But to accomplish this integration, all these approaches need to be combined into a unified whole, where matching demand to supply is explicit rather than implicit.

If demand into any system exceeds the capacity of that system, the system will fail. That mismatch will inevitably lead to expanding delays and ineffective and short-sighted attempts, like priority and triage, to solve the mismatch. Failure to understand this basic dynamic, a focus on change without a context, coupled with efforts to improve isolated components of a larger system, leads to sub-optimization. Some examples:

- Emergency room: The ER is the first demand/supply step into a much larger interconnected system. Mismatch of excessive demand compared to supply deeper in the system can result in gridlock in the ER. Emergency room improvement efforts, utilizing many of the improvement methods discussed above, often focus on making changes in the ER alone: the initiation of bedside registration to reduce steps, the development of "fast track" for the not-so-sick, and the implementation of an express admission unit – a place to park patients who have completed the ER journey but have no bed. While these changes improve the flow and efficiencies of the ER, the demand and the delay are just sent further downstream. Bedside registration reduces the time of the initial process step, but patients just wait longer for the physician. The EAU moves the work to a parking lot, requires new staff and actually serves to extend the length of stay, which worsened the gridlock.
- Central triage: The workload referral hand-off between primary care and specialty care has been arbitrary and fraught with customization, informality and variation, resulting in inevitable dissatisfaction, errors and delays. The development of a central triage unit to manage the workflow by creating a single standardized entry point and process is an attempt to reduce the variation, dissatisfaction and error. At the same time, the incorporation of formal "priority" as an inherent part of the new process actually maintains a high number of distinct channels of work, resulting in a higher likelihood that the "second sickest" queue will be delayed past the recommended threshold. Even though some significant improvements are achieved, neglecting to "see" that the creation of more priority queues will result in more error and delay actually serves to continue to sub-optimize overall system performance.

All the improvement methodologies listed and discussed above contain strategies crafted toward three potential objectives: reduce demand, increase or enhance supply, or create a more effective match of demand to supply, primarily through the reduction of variation. Successful organizational improvement utilizes all or any of the strategies, linking them through the integrated lens of explicit demand/supply matching.

Successful integration then requires a linkage of all the various methods used as a framework to guide improvement work. In addition, successful integration requires linking the four pillars that frame the integrating services initiative: people-centred care, reduction of variation, a focus on the care continuum and improvement in process management. In order to integrate these approaches into a unified whole, the fundamental dynamic of matching demand to supply must be made explicit. HQ

References

Abbott, J. June 2, 2008. *Clinical Integration: Capital Health's Journey and Early Lessons.* Presentation at the National Healthcare Leadership Conference, Edmonton, Alberta. Retrieved September 2, 2009. <http://www.healthcareleadershipconference.ca/assets/Abbott%20Presentation.pdf>.

Bergeson, S.C. and J.D. Dean. 2006. "A Systems Approach to Patient-Centered Care." *JAMA* 296(23): 2848–51.

Bodenheimer, T. and K. Grumbach. 2004. "Can Health Care Teams Improve Primary Care Practice?" *JAMA* 291(10): 1246–51.

Murray, M. and D. Berwick. 2003. "Advanced Access: Reducing Waiting and Delays in Primary Care." *JAMA*; 289: 1035–40.

Murray, M., T. Bodenheimer, D. Rittenhouse and K. Grumbach. 2003. "Improving Timely Access to Primary Care: Case Studies of the Advanced Access Model." *JAMA* 289: 1042–6.

Nolan, T., M. Schall, D.M. Berwick and J. Roessner. 1996. *Reducing Delays and Wait Times throughout the Healthcare System.* Boston, MA: Institute for Healthcare Improvement.

Walley, P., K. Silvester and R. Steyn. 2006. "Managing Variation in Demand: Lessons from the UK National Health Service." *Journal of Healthcare Management* 51(5): 309–22.

About the Author

Mark Murray, MD, MPA, is a Principal at Mark Murray and Associates, and a Technical Advisor at Idealized Design of Clinical Office Practices (IDCOP), Institute for Healthcare Improvement in Sacramento, CA. For 19 years he worked at Kaiser Permanente in Sacramento, CA, where he held various administrative positions. He has also worked as a consultant with multiple organizations and types of organizations in US and internationally. He is widely published and is recognized as an international authority on the development of access systems in health care.