Adapting Agile Strategies to Healthcare Service Delivery

Tom Rust[1, 2], Khalid Saeed[2], Isa Bar-On[2], Oleg Pavlov[2]
[1]Veterans Health Administration
[2]Worcester Polytechnic Institute
100 Institute Road
Worcester, MA 01609
(508) 831-5296
tom.rust @ va.gov, saeed @ wpi.edu, ibaron @ wpi.edu, opavlov @ wpi.edu

1. Abstract
Agility is a fundamental characteristic of successful supply chains faced with volatile and unpredictable demand, and has been suggested as a promising new paradigm for improving healthcare delivery. Agility is an organization-wide capability that seeks to improve overall responsiveness to customer demands, synchronize supply to demand, and cope with uncertainty. However, even if many conceptual models of agility are available, extant literature fails to clearly delineate how 'agile' can be applied in healthcare services and to what extent healthcare systems can benefit from these approaches, nor are there comparisons to current healthcare system redesign paradigms. Given the resource constraints in most hospitals, it is useful, if not critical, to develop a good understanding of how, and to what effect, the agile paradigm can be applied in healthcare. We test specific agile operational practices in a simulated healthcare environment using system dynamics modeling, establishing the comparative effectiveness of changes to system structures that promote market sensitivity, demand information sharing, and centralized planning. This study provides healthcare managers and policy makers with concrete guidance to improve system performance through adopting agile practices, and opens a new area for service supply chain management research.

2. Keywords
Healthcare, service delivery, supply chain management, agility

3. Introduction
Healthcare has long been considered to be among the most complex systems in modern society (Longest, 1974), and as healthcare continues to increase in scope and complexity, so will the challenges to manage that system effectively. Present day healthcare delivery is defined by the idea that networks of clinicians, rather than individual clinicians, provide patient care, and that the success or failure of healthcare delivery is ultimately determined by the ability of those clinicians to coordinate their activities. As healthcare increases in complexity, these previously disparate care processes and clinicians become harder to manage and align, resulting in increased risk to patients and inefficient use of system resources. At the same time, increasing competition, expanding public health challenges, and decreasing resources create an increasing necessity to deliver healthcare services in a more efficient way. Hence, being able to understand

and quickly adapt to the ever-changing needs of patients as they move through networks of healthcare providers is crucial to the success of our healthcare delivery system. Ensuring that the proper supply of care can be synchronized to meet the peaks and troughs of demand is clearly of critical importance to providing cost-effective, high-quality healthcare.

The 'old' solutions of reducing costs by cutting back on staff and services are shortsighted at best. Other solutions taken from the manufacturing sector, such as lean or TQM, have yet to yield the scale of improvements predicted. Furthermore, healthcare has been slow to adopt the service supply chain management practices that have benefitted other industries (McKone-Sweet et al, 2005). Research into supply chain principles applied in healthcare settings is still in its infancy; therefore, no operations-level guidelines exist for healthcare managers seeking to improve system-wide service delivery.

To address this gap, this paper explores the use of a relatively new management paradigm in healthcare. Taken from custom manufacturing and service delivery, 'agile' is a set of organization-wide strategies which optimize service delivery in volatile demand environments with highly variable customer requirements. First coined by researchers at the Iacocca Institute at Lehigh University, in 1991, 'agile' focuses on increasing system responsiveness to customers through improved resource coordination and flexibility, by redesigning organizational structures, information systems, logistics processes, and management decision heuristics.

Agile has recently been suggested as a means to improve healthcare service delivery (Vries & Huijsman, 2011), but specific practices or policies to increase 'agility' have not been developed for service chains, including healthcare. Moreover, the comparative effectiveness of individual agile practices is unknown, as are the trade-offs created by individual agile practices on cost, service access, and service quality. While theoretical agile concepts seem perfectly suited for improving the management of complex healthcare organizations faced with inherently variable demand, practical implementation remains challenging.

We seek to determine how agile principles can be operationalized in healthcare redesign efforts to address issues of patient access, service quality, and cost control. The research questions to be answered in this paper are:

- What are key agile operational plans or practices (structural changes to process or information flows or management decision-making) that can be applied or adapted to improve performance of healthcare service delivery chains?
- How do these different agile-derived practices impact cost, quality, and access to services under unpredictable, variable demand?

To this end, we uncover operational plans from agile and service supply chain literature, then using system dynamics modeling, examine the effectiveness of these operations-level changes in

simulation in a generic healthcare service chain. We present our findings as guidelines healthcare managers and policy makers to improve system performance through adopting agile practices. This study opens a new area for service supply chain management research and provides recommendations for future empirical field tests.

The paper is organized as follows. Section 4 briefly details the current trends in increasing healthcare delivery complexity, difficulty with demand and supply synchronization, and resultant service quality and patient safety impacts. In section 5, an 'agile' literature review and summary of current knowledge gaps are presented. Sections 6 and 7 report the methodological approach and formal mathematical conceptualization of the healthcare service chain and agile strategies, respectively. Section 8 is devoted to describing performance measurement and 'base case' simulation analysis. In Section 9 the experimental design is presented and results are discussed. Finally, Section 10 provides the conclusions.

4. Problem Description

With the growing complexity of healthcare, providers are increasingly dependent on sharing care delivery activities with other, specialized healthcare professionals to provide adequate patient care. Patients are now treated in service chains or service networks that combine interventions into serial encounters with specialized providers and link theses encounters into clinical pathways. Moreover, the redesign of hospital services and the implementation of integrated care programs are frequently cited as being critical strategies to decrease resource utilization and improve healthcare quality (Aptel and Pourjalali, 2001). Clearly, from both a theoretical and practical point of view, the health service operations are in the process of changing significantly.

However, the variability and unpredictability inherent to healthcare demand and internal operations render this network approach to care delivery difficult to manage (Li et al, 2002). Individual patient cases are variable and work cannot always proceed according to schedule or plan. New developments in a patient's condition, unexpected diagnostic findings or surprising reactions to medication may call for sudden changes in planned processes with ripple effects throughout the service supply chain. The growing interdependence of healthcare delivery, coupled with pressure to reduce costs and serve greater numbers of patients, makes these delivery chains increasingly difficult to manage and coordinate.

There is also evidence of healthcare service chain generate internal increased demand variability. Similar to the 'bullwhip effect' (Forrester, 1958, 1961; Lee et al, 1997) in manufacturing, research on healthcare service chains has identified structural tendencies toward demand amplification as a key cause of supply chain stress, and leads to reduced access to services (as measured by the distribution of service delivery time), and subsequent degradations in service quality and increasing employee fatigue. Even with significant external variation, internal

variation is clearly introduced by system structure and dynamics. For example, a case study conducted in a 127 bed hospital in Uttar Pradesh, India revealed dynamic system behavior equivalent to the bullwhip effect (Sameul et al, 2010). The bullwhip effect was similarly identified in the a study of a UK hospital: Based on interviews with hospital staff and data from hospital's EHR system, analysis of emergency patient arrivals and discharges revealed amplification of demand variability downstream in the service chain (Walley, 2007). In this case, distortions in demand clearly led to performance degradation, as downstream services reported reduced resource availability and greater probability of exceeding desired utilization and occupancy rates. These are similar consequence to the effects seen in manufacturing systems, where the bullwhip effect has been a suggested cause for increasing stock-outs and higher costs. A study of a large hospital in Australia also directly identifies the bullwhip effect in the patient pathway for elective surgeries (Sethuraman & Tirupati, 2005). The increasing variation in demand for services as elective patients move to downstream clinics creates the need to make more beds available in post-operative care wards than indicated by the initial demand. On peak days, when the bullwhip effect causes the number of elective surgeries to be artificially high, there is a shortage of beds in the patient wards, which restricts the number of surgeries and reducing the theater utilization and hospital throughput on subsequent days. Demand for nursing services is directly affected by higher variability, resulting in higher labor costs. Higher demand variation amplification is also associated with increased dependence on part-time or temp agency staff. Increasing demand variability inside the patient care pathway generally results in greater stress on employees, higher operating costs, and lower hospital revenues.

There is mounting evidence that the US healthcare system has difficulty matching supply of services to patient demand, coordinating transfer of patients between providers in healthcare service chains, and managing demand variability. Each of these issues adversely affects care quality and patient health outcomes. Kane et al (2007) find that the mismatch between resources and peaks in demand is the major source of provider fatigue and reduced quality of care in most healthcare services. With the management systems currently in place, this variation leads to mistakes in care delivery and increased patient safety risks. Specifically, the stresses placed upon a healthcare system by variability have been found to lead to more medication errors, hospital-acquired infections, sicker patients, and are a leading cause of adverse patient outcomes (Needleman et al, 2002; Berens, 2000; Pronovost et al, 1999).

Studies of variation in the patient to provider ratio, a key measure of service supply chain coordination, find that variability is the norm in healthcare services (De Vries et al, 1999). Higher patient to provider ratios have been correlated with increased patient mortality and failure-to-rescue (deaths following complications) rates within 30 days of admission (Aiken et al, 2002). Large, multi-state studies frequently report inverse relationships between the number of nurses per patient and common nosocomial complications, such as urinary tract

infections, pneumonia, thrombosis, and pulmonary compromise (Kovner & Gergen, 1998). Ensuring that the proper supply of care resources can be synchronized to meet the peaks and troughs of demand is clearly of critical importance to providing cost-effective, high-quality healthcare.

The common management practice in healthcare is to accommodate fluctuations in demand with 'mandatory' overtime. Driven by the need to maintain competitive advantage and minimize costs, the common practice in healthcare is to set staff levels equal to the average demand for services as opposed to setting staff to accommodate peak demand (Litvak et al, 2005). Although such staff management strategies help to reduce labor-related costs, this staffing trend leads to the undesirable consequence of care units being increasingly understaffed during periods of peak demand, which limits their ability to match services with patient demand. This results in the use of excessive overtime as a management solution to demand variability. Excessive overtime is a pervasive problem in healthcare; for example, in a national survey of hospital staff nurses, more than one-quarter of respondents reported working unpredictable, 'mandatory' overtime during the 28-day study period (Rogers et al, 2004). A more recent survey of critical care nurses reported that over 60% worked ten or more overtime shifts during the 28-day study period (Scott et al, 2006). This capacity management trend leads to higher turnover rates (some estimates of nurse turnover rates in the US are as high as 20% per year, see Hayes et al, 2012), which leads hospitals to incur excessive training costs and to lower average staff experience levels. Excessive 'mandatory' overtime is also one of the key drivers of increased provider fatigue and error rates, further reinforcing the argument that current healthcare management strategies need improvement, and currently contribute to patient safety risk and deterioration in quality of care.

These reported pressures and adverse feedbacks to care quality all indicate that current service supply management strategies are failing in healthcare. The healthcare sector is far behind other industries with respect to successful service supply chain management. As currently managed, the average healthcare delivery system exposes patients to unnecessary risk and provides sub-optimal use of system resources and personnel. However, with healthcare expenditures currently 18% of GDP and climbing, hospitals cannot return to past practices of setting staff levels based on peak demand; nor, with the near-exponential increase in the number of clinical trials and the medical evidence-base (NIH, 2013), can they effectively simplify care delivery. Healthcare managers need new service management strategies to be able to respond effectively to changes in patient demand and to mitigate the adverse effects of demand variability on patient care.

Other sectors are able to harness the insights developed by industrial supply chain management research, where firms have faced similar challenges of demand variability and the need for increasing supply chain integration. With minimal abstraction, it is possible to align most healthcare service performance improvement or care coordination questions with those from

industrial supply chain management, mostly relating to how a high resource utilization can be matched with a high customer service level. Recent empirical studies show that a significant portion of the costs associated with service chains in the health care sector could be reduced by implementing effective supply chain management principles (Burns, 2000; Dacosta-Claro, 2002; Oliveira & Pinto, 2005). The current discourse in the service supply chain management literature supports the assumption that existing concepts, models and supply chain management practices can be extended to service chain management in health services (Vries & Huijsman, 2011). The healthcare managers should be able to benefit from the lessons learned in the industrial sector.

However, improving healthcare service delivery chains cannot be done by simply transferring product and manufacturing knowledge and models (Ellram et al, 2004; Sengupta, Heiser & Cook, 2006). Service chain management in a healthcare setting is characterized by some unique features, which make it difficult to apply knowledge gleaned from the industrial sector to the healthcare sector in a direct way. The unpredictable, stochastic demand for services, individual patient attributes driving the need for customized services, the inability to maintain physical buffers of finished inventory, the inherent uncertainty in the duration of care processes, and other distinctive characteristics of health service operations impede a straight forward application of industrially-oriented supply chain management practices. In practice, Bohmer (2009, p. 16) finds that "many of the approaches and tools drawn from industrial settings fail to adequately account for the residual uncertainty in medical care or explicitly address the experimental nature of much care." Most manufacturing-based supply chain management paradigms, such as 'lean,' Total Quality Management, or Six Sigma do not function effectively in systems with high levels of inherent process variability and demand uncertainty (Lee, 2004), but it is precisely these context-defining characteristics that cause most of the present difficulties in healthcare service integration and care coordination.

Service chain management in a healthcare context is very much an emerging field, and has not yet identified how to overcome these contextual difficulties, nor has the field identified a service chain management paradigm suited to the healthcare context. Subsequent questions of how service delivery integration and coordination of care systems regarding patient flows and resource management can be best achieved operationally still are a relatively unexplored area of service supply chain management, and starting from this question there are only limited academic studies addressing the challenges unique to the healthcare setting (Vries & Huijsman, 2011). Most service supply chain management research is still theoretical or conceptually-focused as opposed to operational in nature (Sampson & Froelhe, 2006), currently providing little to aid managers in the midst of redesigning their systems and integrating care processes. Healthcare managers face a significant gap in knowledge around the optimal design and management of complex care delivery systems that ensure effective patient care.

If the current trend to integrate patient care through increasingly complex provider networks continues, then matching supply and demand throughout the healthcare service chain will become increasingly difficult. As a result, both patients and providers will suffer. Hospital managers need insight from service management researchers that directly address the problems arising from the variability and complexity of demand within a hospital and coordination issues between healthcare units. They require guidance on decision structures and designs of service chains that create the flexibility necessary for the dynamic nature of health itself and which enhance the effectiveness and efficiency of care delivery in the face of complexity. Service supply chain scholars need to identify and develop a new service management paradigm that accommodates the uncertainty and variability inherent to healthcare, specifically to conduct operations-level research to improve the design and management of healthcare service chains.

5. Literature Review

There is one supply chain management paradigm that does address the context issues which separate most healthcare operations from those in industrial or manufacturing settings. 'Agile' is a manufacturing paradigm, coined by researchers at the Iacocca Institute at Lehigh University in 1991, that describes the strategies they observed as crucial to enterprise success in environments of rapid and unpredictable change (Iacocca Institute, 1991; DeVor et al., 1997). In essence, an agile manufacturing system is one that is capable of operating profitably in a competitive environment of continually and unpredictably changing customer opportunities (Goldman, et al. 1995). Similarly, Gunasekaran (1998) defined 'agility' in manufacturing as the capability to survive and prosper in a competitive environment of continuous and unpredictable change by reacting quickly and effectively to changing markets, driven by customer-designed products and services. An ‘agile’ organization as one that able to compete successfully within a state of dynamic and continuous change (Sarkis, 2001), through efficiently changing operating states in response to uncertain and the changing demands placed upon it (Narasimhan et al. 2006).

‘Agile’ is more than a description of an ideal supply chain. Many manufacturing companies have experienced high costs associated with holding excess inventories as consumer preferences change, or incurring stock-outs and decreased to market share in times of unanticipated demand, and have followed agile principles to re-designed their supply chains to better accommodate such demand volatility, resulting in increased revenues and market share (see Lee, 2004 for discussion). There are academic journals dedicated to the advancement of theory and practice of agility in manufacturing systems (e.g., International Journal of Agile Management Systems). Harvard Business School has published case studies highlighting the principles of agile manufacturing (1991). Since its inception, the agile paradigm has had a profound impact on the design and management of manufacturing systems facing the same problems that healthcare currently faces: the need to integrate and coordinate disparate units in a delivery chain, all in the face of unpredictable, volatile demand.

In an operational sense, agile is a set of strategies that solves the problem of demand uncertainty and variability through increasing system flexibility (Lee, 2004). It encompasses re-design of organizational structures, information systems, logistics processes, and management decision heuristics, all to achieve timely and effective response to rapidly changing demand environments (Christopher & Towill, 2002). Agility involves increasing the capability to quickly identify shifts in market demands or external supply disruptions and execute new, unplanned activities in response (Brown & Bessant, 2003; Prince & Kay, 2003; Sharifi & Zhang, 2001).

There are four commonly agreed upon distinguishing operational characteristics of agile supply chains (Harrison et al., 1999; Christopher, 2000), as outlined in Figure 1. They are highly market sensitive; capacity adjustment decisions are driven by demand information; planning is centralized, not left up to individual units; and processes and performance management are integrated across all units in the chain. Each is discussed in detail below.

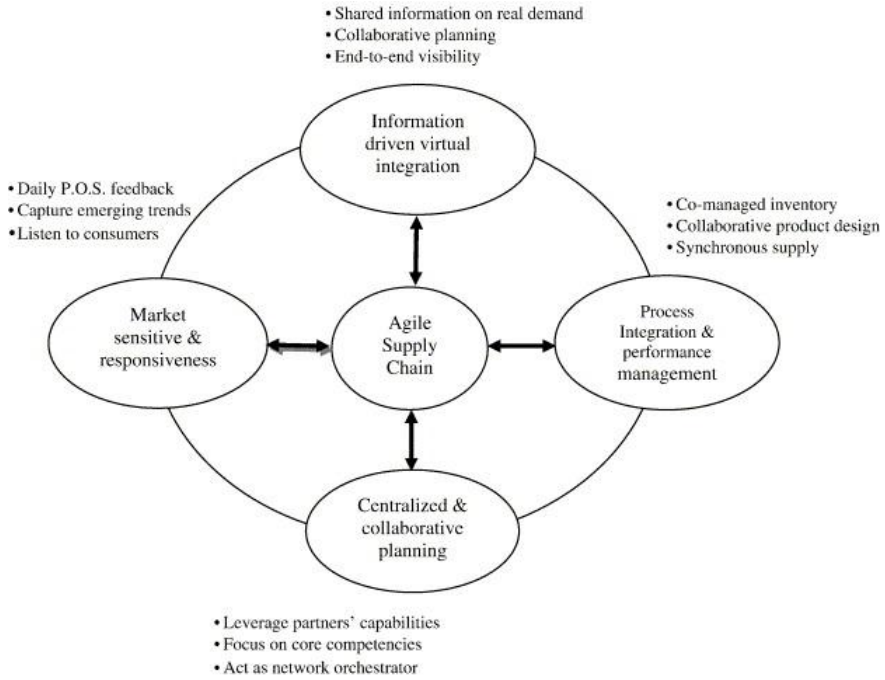

Figure 1. Characteristics of agile supply chain (modified from Harrison et al., 1999).

First, market sensitivity means that the supply chain is capable of reading and responding to real demand in real time. This is a direct contrast to most organizations, where long production times and logistical delays tend to require they be forecast-driven rather than demand-driven. A key operational requirement of agile supply chains is to be able to quickly change production resources to respond to real-time demand data. Agile supply chains often make use of information technology systems to capture and share data on demand quickly throughout the entire supply chain (Harrison et al., 1999).

Christopher (2000) suggests that, to be truly 'agile,' the use of information technology to share data between partners in the supply chain must be taken farther, in effect, creating a virtual supply chain, defined as a supply chain that is information based rather than inventory based. This does not imply simply reducing physical inventory; instead of coordinating flows of physical goods, agile organizations must acquire the capability to coordinate production capacity (Lee, 2004). Shifting to information based supply chain coordination means that all elements in the chain act upon the same data, i.e. real demand, rather than being dependent upon the distorted information that emerges when orders are transmitted from one step to another in an extended chain (Argawal et al, 2007). Without being subsumed under the agile paradigm, research into the phenomenon of demand variation amplification (the 'bullwhip effect') has shown that this operational strategy improves supply chain performance, in terms of reducing risk of stock-outs, reduced 'phantom' demand, and decreased average costs (Croson & Donohue, 2003; Chen et al, 2000).

This shared information between supply chain partners can only be fully leveraged through the third key characteristic of agile supply chains: centralized planning, meaning the collaborative design and implementation of cooperative management structures. Operationally, centralized planning requires disparate units in the production chain to make management decisions and set performance goals to maximize performance of the total chain, not the performance of each individual production unit. Each unit must take the adjacent units into consideration when making production decisions. Central planning creates a shared systems perspective and ensures the appropriate incentives structure is in place to lead to maximizing overall performance (Cannella & Ciancimino, 2010; Cachon & Fisher, 2000). Operationalizing this aspect of agility has been shown to improve visibility of production requirements and reduce the amount of stock (or production capacity) held in anticipation of predicted and often distorted demand (Hewitt, 1999).

This idea of the supply chain as a confederation of partners linked together through a network of continuous collaboration leads to the fourth ingredient of agility. More than cooperation on strategic planning and goal-setting, process integration implies cooperation between production units on production activities themselves. There is a broad assortment of operational plans supporting the concept of process integration in the literature. Examples range from the co-design of new products, so design accommodates both end-user preferences and factory and production constraints; to the most extreme manifestation of direct sharing of production and management resources between production units (Lee, 2004).

Some of these individual operational plans have been tested through the simulation of generic service chains, even if not directly subsumed under the paradigm of 'agile.' Most of these simulation studies inadvertently target market sensitivity. Anderson et al (2005) explore the effects of increased market sensitivity through changing in service capacity adjustment decision-

making times, finding that decreasing decision delays leads to improvement in overall supply chain performance. Lee et al (2009) examine the effect of supplementing demand data with information on the derivative of changes in demand. Their results are mixed, finding that with optimal control schemes it was possible to halve the costs associated with demand variation amplification, but that including derivative-based information could also lead to increasing oscillations in some scenarios.

The other operational strategy tested in simulation is virtual integration, with Anderson and Morrice (2000) assessing the effect of sharing end-customer demand data in real time in a simplified service chain. As is commonly reported in manufacturing settings, they found that incorporating end-customer demand data with local demand data in individual service unit decision making led to increased performance, both in terms of total reduced costs and improved service delivery times. Although not directly discussed, their work also reveals the necessity of centralized planning in service chains. The parameter set that created their lowest-cost, highest-performance scenario would have been unsustainable without centralized planning, as costs were not shared equally across service units. If individual units made decisions only to maximize their own performance, the service chain would never be able to implement this optimal scenario. Cooperating to redistribute the costs and benefits of service delivery redesign seems crucial to the ability to achieve optimal performance.

There is also some anecdotal empirical evidence supporting the use of agile strategies in healthcare service delivery chains. Service chain integration is becoming more prevalent in healthcare, as team-based care models (e.g., the Patient Centered Medical Home, or PCMH) are becoming standardized. As of 2007, an estimated 27% of primary care practices follow some elements of the PCMH model, where disparate elements of the health care system (e.g., subspecialty care, hospitals, home health agencies, nursing homes) and the patient's community (e.g., family, public and private community-based services) are coordinated through a patient's primary care provider (Beal et al, 2007). There are a few reports suggesting that these changes lead to better care quality, reduced errors, and increased patient satisfaction (Rosenthal, 2008), with one recent study of a Seattle health system demonstrating 29% fewer emergency visits, 6% fewer hospitalizations, and total savings of $10.30 per patient per month over a twenty-one month period (Reid et al. 2010). Health services are also moving toward virtual integration as well, with the recent mandate to create health information exchange, which will provide the capability to electronically move clinical information among disparate healthcare information systems (HITECH, 2009). All this suggest that agile strategies have promise in the healthcare context and should be further explored.

The question how to best integrate agile strategies into healthcare is an uncovered field in the area of supply chain management, and has only most recently been suggested. The November 2011 special issue on healthcare of the international journal Supply Chain Management

highlights the need for in-depth research into the strengths and weaknesses of the agile management paradigm in the context of health services. There are clear gaps in knowledge of the application of agile strategies, namely: What are the clearest translations of agile strategies into operational plans applicable and feasible in healthcare service delivery? How do they compare to each other in effectiveness, as defined as the ability to increase total service chain flexibility and mitigate the adverse effects of demand volatility? Do agile strategies need to be implemented as a bundle to be effective, or are they effective at creating service chain flexibility when implemented separately? Answering these questions to further the adaptation and use of agile strategies to healthcare could contribute significantly to the broader field of patient logistics and the improvement of healthcare service management.

6. Methodology/Approach

The literature on supply chain analysis is rich with classifications of methods used to investigate supply chain performance and the effects of demand variation (Riddalls et al, 2000; Kleijnen & Smits, 2003; Dejonckheere et al, 2004; Disney et al, 2004; Geary et al, 2006; Towill et al, 2007; Disney & Lambrecht, 2008). Riddalls et al. (2000) submit that the choice of which methodology is most appropriate is determined by decision-making level under consideration, commonly divided into 1) the local, tactical level for day-to-day decision making, and 2) the implication of strategic design on supply chain performance and overall network functioning. Holweg and Disney (2005) recommend the latter category be analyzed using methods based on the dynamics of the system in question. They recognized three distinct and methodologically independent research domains: continuous time differential equation models, discrete time difference equation models, discrete event simulation systems.

The choice of methodological approach adopted in this paper is based on the need to explore the dynamic interaction effects of various operational plans, and to develop a general understanding of their inherent dynamics when applied in a healthcare service chain. We adopt a continuous time approach, namely system dynamics simulation modeling. The service chain and governing decision heuristics are modeled through first-order nonlinear differential equations. The formal mathematical expressions of the system are reported in the next section. We have not used discrete-event simulation (DES) or stochastic modeling (of variables like ‘patient inflow' or ‘treatment time') because our primary objective is not to quantify numerical results for one specific healthcare delivery chain, but to understand and illustrate to healthcare managers the deterministic behaviors of healthcare delivery systems in general. The use of continuous, as opposed to discrete, flows in the model is a reasonable approximation of the perpetual adjustments (hiring and firing) necessary in the management of service organizations, and is a common method for abstracting these systems in both operations management and supply chain management research (Sethi & Thompson, 2000). For an in-depth discussion of the trade-offs

and appropriate problems for using system dynamics or DES see Tako and Robinson (2009) and Kleijnen (2005).

System dynamics is an appropriate chose for modeling healthcare systems, as it encourages both a systemic view of the interactions of patient flows and information, and a strategic perspective on the management of healthcare delivery systems. System dynamics modeling has been used to address several healthcare related problems and has resulted in about 1500 publications since 1991 (Brailsford 2008). Dangerfield (1999) reviews system dynamics modeling in healthcare and concludes that the method can be used effectively in quantitative ways when based on simulation models. Examples of modeling efforts range from the use system dynamics simulation to analyze the reasons for the failure of management interventions in cardiac catheterization services (Taylor & Dangerfield, 2005), to the improvement of acute patient flows in the UK National Health Service (Lane & Husemann, 2008). With the trend toward care integration across complex networks of activities and specialties, system dynamics offers a rigorous approach for understanding the strengths and weaknesses of that new interconnectedness. A general discussion of the role of system dynamics in analyzing healthcare systems can be found in Taylor and Lane (1998).

The use of system dynamics to understand the dynamics of healthcare service delivery builds on a rich history of developing insights into supply chain management. Forrester (1961), the creator of system dynamics, laid the foundations for the use of the continuous time approach towards the study of supply chain dynamics. Later, the work on 'beer game' simulation (Sterman, 1989) ushered a new era in supply chain management research into understanding how micro-level decision structures, bounded rationality, and misperceptions of system feedback cause macro-level behavior (Cannella & Ciancimino, 2010). Subsequent system dynamics-based research in *service* supply chains has led to knowledge of how the defining phenomena of services, for example, the intangibility of services and the simultaneous interaction of customer and service provider, affect system behavior.

This paper builds on a thread started by Oliva and Sterman's (2001) simulation study of a single-stage service process, as well as Anderson's (2001) analytical model of similar issues. It relates to research on managing service chain dynamics (such as the 'bullwhip' effect), information sharing, and coordination of management decision making (e.g., Lee et al, 1997; Chen, 1998, 1999). Anderson and Morrice (1999, 2000) first consider a multi-stage service system in their model of the mortgage service industry, and start the exploration of the impact of resource acquisition delays on demand variation amplification. In their case study of a European telecom firm, Akkermans and Vos (2003) use a similar model to develop insight into the interdependence between workload, work quality and variation amplification. However, their model is highly context specific and the analysis space is limited by the narrow set of policy options available to the firm's management. Perhaps the most closely related research exists in the linear programing

work Anderson et al. (2005, 2006), which uses a relaxation of a system dynamics model to evaluate the structural causes of, and counter-measures to, demand amplification in a generic service chain. Such uses of system dynamics have been cited as "clear exceptions" to what is normally described as the "forced and unclear" application of supply chain management modeling methods to services (Sampson & Froehle, 2006, p.337).

7. Model Design

The healthcare service chain we model here is an abstract representation of a broad spectrum of possible healthcare delivery networks. The purpose of the model is to capture the essential elements of reality common to most healthcare delivery chains rather than perfectly simulate one specific service. Our delivery chain consists of three stages, the three most clearly defined stages in any patient care event: diagnosis, treatment and recovery, as indicated in Figure 2 (Aronsson et al, 2011). While all steps can be performed by one or several organizations depending on the patient, we represent each stage with a finite workforce capacity handling the different tasks inside a stage, which can represent the organizational separation and specialization among hospitals, or between departments inside a single hospital. The health care service system is modeled in continuous time, and is simulated with Vensim® software. This model is concurrent with previous system dynamics service supply chain models, visualized in Figure 3.

Figure 2. The functional steps in a healthcare process (adapted from Aronsson et al, 2011).

To illustrate how patients flow through this chain, take the example of the care of patients with acute myocardial infarctions (sudden heart attacks). It is a care process that involves several departments inside the hospital and often requires addition rehabilitation services after treatment. There are also clear quality implications of care lead-time, as mortality rates are highly correlated with diagnosis and treatment delays (Gulli et al, 2010). These patients often arrive unscheduled, by ambulance to an emergency department. A diagnosis is made which normally includes lab-tests and X-rays. After diagnosis, the patient is transferred to the cardiac catheterization lab for coronary angioplasty (PCI) or bypass surgery (CABG). After the operation, the patient recuperates in a cardiac care unit. When leaving the hospital there are often extended needs for recovery involving physical therapy and social services, which has to be planned and coordinated with family, physical therapists, and possibly personal care attendants.

7.1 Traditional service supply chain structure

These archetypical functional steps, as outlined by Aronsson (et al, 2011), map directly to our model structure: where boxes represent patient service backlogs, hourglasses represent patient care events (interaction between patient and provider or other staff and resources), and arrows represent the direction of patient flow.

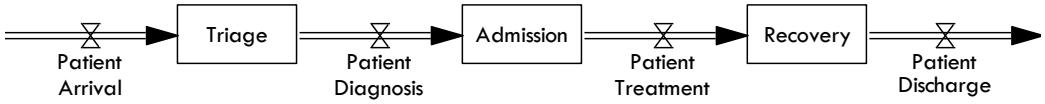

Figure 3. A generic multi-stage healthcare service delivery model

While clinics in our service chain are obviously linked, as the output of one clinic forms the input to the next, each clinic in our model operates autonomously, as management decisions are based only on the information available inside each clinic. Each clinic has sole responsibility for operational performance and control of its own resource actions, i.e., acquiring and releasing workforce. Each clinic requires a separate set of resources to serve its patient backlog; no resources are shared between clinics. Resource sharing may be possible in some healthcare service chains, depending on the specifics of a particular care process, but the high level of specialization and the complexity of healthcare ensure that resource sharing is not the norm. For simplicity, we assume that there are no dropped or lost patients and all of the patient care events in a backlog are eventually concluded.

Each of the three clinics is identical, with a finite capacity for patient care, derived from the number of providers working in that clinic. Each clinic's implicit goal is to keep service performance at a desired level (measured in average service time), while keeping service capacity costs to a minimum. While this structure is far from optimal, it is a realistic representation of the common 'staff to demand' heuristic found currently in most hospitals and health care centers (Litvak et al, 2005).

A more specific stock and flow model of one representative clinic is presented in

Figure **4**, graphically displaying the three control loops fundamental to clinic management: one to prevent number of customers waiting for service from going negative (the Work Availability loop), one representing manager's decisions to add or remove providers from the clinic schedule to balance workforce with demand (the Capacity Management loop), in which is embedded the manager's decisions comparing current workforce with desired workforce to achieve desired service capacity (the Meeting Workforce Goal loop). The formal mathematical details are described below.

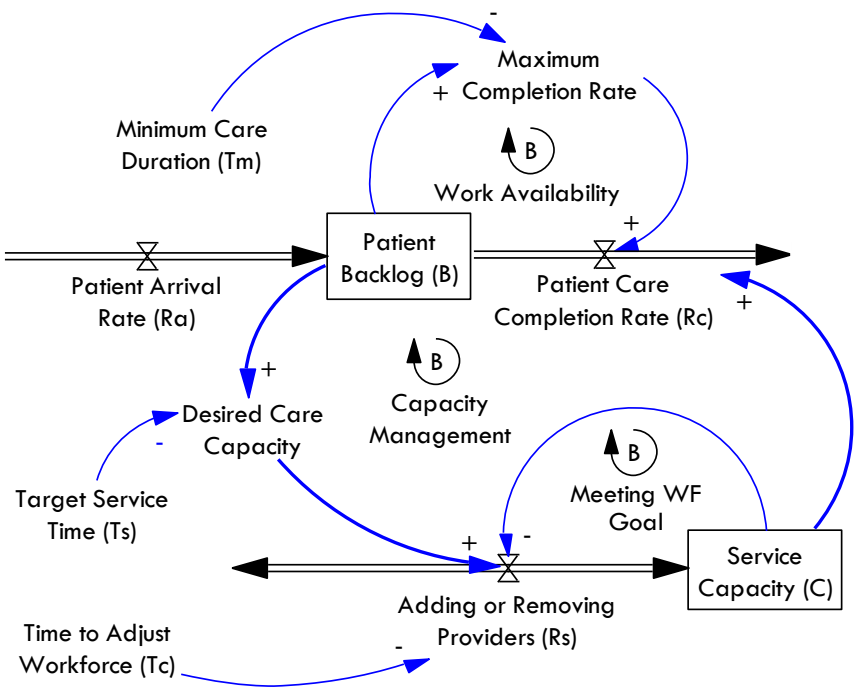

Figure 4. System dynamics model of a single representative care delivery unit; the main capacity adjustment loop is highlighted.

There are two streams of activities in the model. The first one is the uncertain flow patients coming into a clinic as shown in the top portion of
Figure 4. The patient backlog accumulates based on the difference between inflow of demand arrival, R_a , and completion rate, *Rc*. Note that the patient backlog (*B*) is actually the *healthcare work-in-progress*, which is number of patients being diagnosed, treated, or recovering, and has a different meaning than the traditional backlog in an industrial supply chain. The completion of each stage in patient care requires a certain number of resources (personnel, equipment, rooms, etc.) for a certain duration. In this model, we assume that all patients eventually complete each clinic's care process, though some patients can be delayed as accumulated demand backlog due to capacity shortages, as shown in the expression of, R_c , in Equation 1.

$$B(t) = \int_0^t [R_a(t) - R_c(t)dt] + B_0$$

where $B(t)$ = patient backlog at t,
$R_a(t)$= arrival rate of patients at t
$R_c(t)$ = completion rate of patients at t = min[$C(t)$, $B(t)/T_m$]
$C(t)$ = resource capacity available at t
T_m = minimum patient care time
B_0 = initial backlog (patients in-progress) at $t = 0$

Note that patient backlog (*B(t)* and B_0) is expressed in number of patients. Completion rate and resource capacity (R_a , R_e and C) are expressed in the number of patients per day, and completion time (T_m) is expressed in days. There is a minimum completion time even if unlimited resources are available. With the minimum care completion time, adding more resources past a certain level will not reduce the backlog, merely the resource utilization.

The second flow in the model is the flow of workforce as shown in the bottom part of the model in Figure 4. The resource capacity available, *C(t)*, accumulates based on the net capacity discrepancy, *e(t)*, which is defined as the resource capacity needed to care for all patients in the desired time, $C_d(t)$, minus the resource capacity currently available, $C_{t-1}(t)$, as shown in Equation 2.

$$C(t) = \int_0^t \left[\frac{e(t)}{T_c}\right] dt \; + C_0$$

where $C(t)$ = resource capacity available at t
$e(t)$ = the net capacity error = $C_d(t) - C(t-1)$
$C_d(t)$ = resource capacity desired = $[B(t)/ T_s]$
$C(t-1)$ = resource capacity at pervious time step
T_s = target patient care service time
T_c = delay for resource capacity change process
C_0 = initial capacity (providers) at $t = 0$

In the traditional service supply chain, any single clinic only receives information about patient flow from the adjacent upstream clinic. Each clinic calculates the resource adjustment in any given period, $C_d(t)$, on the basis of local data and parameters (service capacity level *C(t)*, delays in the resource adjustment process, T_c) and on the desired service capacity. This desired service capacity is, in turn, modeled as a rate of patients per day, calculated from the current patient backlog and the desired care delivery time. The overall capacity adjustment decision is moderated (divided) by the delays inherent to the resource capacity adjustment process, T_c, which represent delays in decision making and the schedule modification process itself. The denominator of T_c captures the first-order effect of capacity adjustment, and approximates reality for all but extreme target capacity changes. Implicit in these equations is the assumption that one unit of capacity is required to process one unit of patient backlog each day. This assumption can be relaxed by scaling the capacity appropriately. Note that net resource adjustment rate rate, is expressed in the number of patients per day per day, while the units of resource capacities, (C_d, C_{t-1}, C_0) are expressed in units of number of patients per day.

This capacity adjustment policy is derived both from the literature and from interviews conducted with healthcare department chiefs and clinic managers. Anderson (1997) finds an identical stock correction mechanism in custom machine tool industry, Anderson and Morrice (2000, 2001) in a mortgage services company, Akkerman and Vos (2003) in a telecom company, and Anderson et al (2005) in the service supply chains of an oil field development firm. While not an optimal decision heuristic by any means, it is the rule most often used in practice in service supply chain management. Note that the rule in Equation 2 is simpler than the standard stock-adjustment rule found in manufacturing supply chain modeling (Sterman, 2000); we believe this is justified because there is no significant supply line of capacity on order in this system.

Health care service delivery chains are complex and require a great deal of coordination. Based on its use in similarly complex fields, we believe this base structure is both abstract enough to be generalizable across services, and structurally sound enough to yield insights into the dynamics of healthcare service delivery management. We use it as a base upon which to test the adoption of multiple agile-derived operational plans, each explained in detail below.

7.1 Agile operational plans

We identify specific operational plans from the literature on service supply chain management which can be subsumed under the agile paradigm. This assembled set of plans covers all four of the key characteristic of an agile supply chain: market sensitivity, information driven, centralized planning, and process integration. Each is described in detail below; including the structural changes to information flows and management's decision heuristics, along with its mathematical formulation.

7.1.1 Market Sensitivity

Despite its importance in manufacturing supply chain management research, increasing market sensitivity is not directly discussed as such in the service supply chain literature. As a conceptual framework of 'agility' has not yet been universally adopted (Li et al, 2008), there are many meanings of the term market sensitivity depending on context and level of detail under investigation. It could refer to the ability of a manufacturing chain to elicit and respond to patient preferences in new product development (product characteristics), or the ability to perceive, evaluate, and respond to changes in total demand with accurate adjustments to production and inventory quotas (availability), or to identify the relative value of a product in the market and adjust price accordingly (price). In general, market sensitivity is the ability to make swift and appropriate decisions in reaction to changes in demand, in any of its dimensions.

For healthcare delivery, where patients have limited knowledge of the price and relative quality of any given service or provider, the most significant dimension of demand, from a clinic

manager's perspective, is total volume. We narrow our focus through defining market sensitivity as the ability of a service chain to respond quickly to changes in demand for a given service with accurate changes in service capacity in order to maintain an adequate level of service availability. Including the price and service characteristics of patient demand are possible extensions for future work, but are more apposite to the analysis of national healthcare policy than improving healthcare delivery. With this definition, we find two operational plans in the literature that address increasing market sensitivity. Both were first developed in manufacturing and only recently tested (separately) in simulation in service supply chains.

The first focuses on the impact of increasing the speed of capacity adjustment decision making, modeled as a reduction in the service capacity adjustment time, T_c. Analysis of mortgage service simulations (Anderson & Morrice, 2000) suggest that decreasing service capacity adjustment delays is one operational plan available to managers of individual clinics to improve the responsiveness of their clinic to changes in patient arrival rates. In subsequent research, however, they find reducing the equivalent of had mixed effectiveness, improving service delivery (as measured by the variance in average service time), but increasing variance in subsequent stages' capacity stocks and backlogs (Anderson et al, 2005). A reduction in T_c could be achieved through many operations-level changes, from increasing the frequency of information gathering and analysis on the current patient backlog level, to streamlining the HR process for hiring and firing, to improving the quality of training to reduce the training time for new hires, to improving coordination between managers and employees over clinic schedule changes. All would reduce T_c , and make the clinic more sensitive to changes in demand volume.

The second focuses on improving the information used to make the capacity adjustment decision. This operational plan captures emerging trends in demand volume, through including a measure of the change in the patient backlog along with the actual size of the backlog itself. This 'derivative control' is common to physical manufacturing systems (temperature control, velocity control, etc.), and is part of a standard engineering control algorithm called PID (Proportional, Integral, Derivative Control), which is mainly used as a means to minimize the error between a measured value and a target value, given the presence of adjustment delays (Axsater, 1985). Our current decision equation falls under the domain of proportional control, where the rate of capacity adjustment is a proportion of the error between the desired and the actual service capacity. Fundamentally, a PID control algorithm improves basic proportional control in two ways. Integral control creates a system memory of the accumulated error over any period of time when error is being corrected, and adds that error to the correction itself, thus preventing 'steady-state error.' Derivative control increases the correction in response to rising error, thus returning the system to its desired state faster that if the system responded proportion to the error alone. The derivative of the error is a faster signal to clinic managers than the error itself, as the derivative peaks when signal from proportional gap is only at its inflection point.

In simulation, supplementing standard proportional control-base heuristics with integral and derivative control has been found to significantly improve the control of supply chains faced with volatile demand, allowing a reduction in inventory safety stock by over 80% without sacrificing product availability (White, 1999). For discussion of the application of PID in manufacturing settings, see White (1999) and Saeed (2008, 2009). A similar approach was recommended for improving decision-making in strategic management by Warren (2007). The study of the addition of integral and derivative control in service chains is an emerging area of service supply chain research, and has only been explored in simulation. Results from Lee (et al, 2010) indicate that using derivative control to supplement a manager's decision heuristics leads to performance improvement. However, the inclusion of integral control is proposed to not be useful in managing service chains, as the common decision heuristic used produces no steady-state error.

Including derivative control in our clinics' capacity adjustment decision requires changes to model equations. The new equation for net capacity error is shown below.

$$e^*(t) = \left(K_p e(t) + K_d T_d \frac{d[e(t)]}{dt} \right)$$

where $e^*(t)$ = is the error calculated with derivative control
$e(t)$ = the net service capacity error
K_p = gain (sensitivity) constant of the traditional control; normally $K_p = 1$
K_d = gain (sensitivity) constant of derivative control
T_d = derivative time constant

Setting the K_d parameter to zero eliminates derivative control. Changing the parameters for K_p and K_d influences the gain of the traditional and derivative controls, respectively. Adjusting these parameters can be used to optimize the decision equation for a given set of costs.

7.1.2 Information Driven

The importance of demand information in optimizing supply chain management is well known. End-to-end sharing of real-time demand data is one of the common solutions in the supply chain management literature for minimizing the demand amplification (bullwhip) effect inherent to supply chains (Disney & Towill, 2003; Chatfield et al, 2004; Dejonkheere et al, 2004; Shang et al, 2004; Byrne & Heavey, 2006; Kim et al, 2006; Hosoda et al, 2008; Kelepouris et al, 2008; Argawal et al, 2009). Unlike a traditional supply chain, in an 'information driven' system, the information flow consists of both the transmission of stages' orders in the up-stream direction and sharing information on market demand. Each stage remains autonomous and makes decisions on production and distribution independently, but all stages make those decisions on

the basis of shared, global information. This prevents extreme internal demand variation amplification, as each stage now has some understanding of real demand, not just local demand from the downstream stage.

Expanding the data available to supply chain managers at each stage almost always leads to lower costs and fewer stock-outs. The effects of sharing end-customer demand data in industrial supply chains has been tested many times in simulation, under many constraining assumptions, with all findings indicating that sharing end-customer demand data improves performance. There are fewer results from analysis of real world data, but they also re-enforce this finding. For example, Hosoda (et al, 2008) find that sharing 'point-of-sale' data in real-time between a supermarket chain and a soft-drink manufacture reduced the holding and backlog costs incurred by the manufacturer by 8-19%. In a similar study conducted in a Greek retail grocery company, consisting of 250 retail stores and 7 central warehouses, Kelepouris (et al, 2008) find that information sharing results in a 21% reduction in order variability and a 20% reduction in average inventory. These, and other studies, confirm the value of shared information on end-customer demand for mitigating the bullwhip effect and associated costs in physical supply chains.

The usefulness of sharing end-customer demand data has not been empirically examined in service supply chains, but has been explored in simulation. Anderson and Morrice (2000, 2001) test the effect of sharing information on real demand with each stage in their mortgage services chain, finding that adding this information to each stage's decision heuristic does improve performance. However, the use of end-customer demand data can create a trade-off between the level of variation in patient backlog and in service capacity. Anderson (et al, 2005) assert that the stages in a service chain can always decrease their backlog variation by paying more attention to end-customer demand rather than local backlog, but only up to a certain point, beyond which capacity variation will start to increase. Thus, information sharing can only lead to limited improvements before creating a direct trade-off between customer service (variation in patient wait times) and personnel costs (variation in capacity). The precise tipping points are determined by the parameters of a process (for discussion, see Anderson et al, 2006). The common strategy advocated for service supply chain scholars is to use a mix of both, rather than completely relying on one or the other. Determining the optimal weights for both types of data in management decision depends on the cost structure in a given service chain.

We change the 'capacity management' loop to include data on initial patient demand, supplementing local patient backlog data. The first term represents the degree to which the target capacity relies on the end-customer demand rate. The second term denotes how the target capacity depends on the magnitude of the local backlog, $B_i(t)$ and the target service care time. The first term represents the service capacity needed to meet end customer demand at time t and the second term represents the capacity required to guarantee that, on average, the orders not yet

met in the local backlog will not be delayed longer than an acceptable amount of time (i.e., the service delay). The weighted sum of these two terms determines target capacity.

$$C_d(t) = \text{resource capacity desired} = \left[R_a(t)a + \frac{(B_i(t)(1-a)}{T_s}\right]$$

Where $R_a(t)$ = arrival rate of patients in the first clinic in the service chain at time t

$B_i(t)$ = local patient backlog

T_s = target patient care service time

a = the relative weight of end-customer demand in the desired capacity calculation. We assume that $0 \leq a \leq 1$.

This modification to management's decision heuristic can be supplemented with previous agile operational plans for increasing market sensitivity, both decreasing capacity adjustment time and including derivative control.

7.1.3 Centralized planning

The literature on service supply chain management commonly defines 'centralized planning' as a system where decisions are made to maximize efficiency and performance of the total chain, as opposed to decisions being made locally to maximize the performance of individual stages (Anderson et al, 2006). In formal mathematical terms, the control policies for adjusting capacities in all stages are determined simultaneously by optimizing a single objective function for the supply chain. In most services, the optimal level of coordination between stages in a service chain is not obvious, nor are the appropriate methods to create that coordination, i.e., by supply contact or direct ownership (Holweg, et al, 2005).

Confounding factors, like tighter integration leading to organizational diseconomies of scale (Zenger, 1994) or the loss of market share due to shifting brand differentiation strategies, may outweigh any gains in operational improvements from increased coordination. Separate from studies of information sharing, limited empirical research in the service supply chain management literature on the impact of transitioning to centralized planning has been reported. Many studies in manufacturing collaboration and centralization report high degrees of difficulty of integrating external collaboration with internal production and inventory control (Cachon & Lariviere, 2001; Stank et al, 2001). Anderson (et al, 2006) provide anecdotal evidence from the oil-field development industry, where firms with centralized planning are found to be no more competitive or successful than firms with individually managed stages.

Simulation studies of service supply chain centralization are also few, and contain inconclusive results. While highly abstract, special case, linear models have been developed which show centralization leading to improved performance (Anderson et al, 2006), the common strategies for moving toward centralized service chains have been shown to have adverse effects on performance. Anderson (et al, 2005) assert that the default centralization strategy for service

chains is to move stages toward uniform decision making (in terms of the type of information used, management's decision rules themselves, and target performance measures, such as service delivery times). Changing these decisions is the least complex way to implement 'global' supply chain policies, particularly if the stages are inside the same firm. However, moving away from idiosyncratic decision strategies to more uniform decision making inadvertently results in worse performance than if decision strategies had not been aligned (Anderson et al, 2005). Their simulation research suggests that adopting a single management decision heuristic (modeled as identical capacity adjustment times and target service delivery times for all stages) across the entire service chain is actually leads to significant increases in variation in both demand for services and capacity adjustment. Centralization is a difficult strategy to implement effectively in service supply chains, as seemingly benign actions can generate unforeseen adverse consequences.

Other examples illustrate how optimizing delivery in a service supply chain through centralization is not simple or intuitive. Under the simplifying assumptions of a linear relaxation of a dynamic optimization model, Anderson (et al, 2006) find that while transitioning to centralized decision making usually leads to increased operational efficiencies for the total chain when compared to local decision making, improvements are not shared equally between supply chain stages. Centralized planning usually decreases backlog and capacity variation overall, but when measured in isolation, the first stage is almost always worse off than before. The use of a single optimization equation to govern both stages almost always results in improvements in the performance of the second stage, but at the cost of decreased performance of the first stage. Depending on the cost and pricing schemes of the services offered in each stage, centralized decision making could lead to increased costs Anderson (et al, 2006). For example, they conclude that if the first stage has a sufficiently higher cost structure (both of holding excess backlog and/or cost of changing capacity) than the second stage, centralized control of capacity adjustment is of no direct benefit. They conclude that centralized planning may improve total chain performance in some situations and under some limiting assumptions, but it is difficult to achieve in practice, with a high possibility of being counter-productive.

Based on these works, we believe the most promising manifestation of the agile concept of centralized planning is the creation of an 'unbalanced' service chain, where each stage follows a different management decision heuristic (modeled as differing capacity adjustment times and target service delivery times). While varying service targets and capacity adjustment processes is by no mean an optimization, exploration of such policies could lead to simple, straightforward guidance for healthcare service chain managers. To test this strategy, we run three sets of simulations, where we vary 1) target service times, or T_s; 2) capacity adjustment delays, or T_c; and 3) both simultaneously. Operationally, these would be time and resource intensive policies to implement: changing target service times in an individual clinic directly affects service quality and resource requirements; changing capacity adjustment times might involve negotiation with

national accreditation bodies, state review boards, internal HR committees, union representation, etc. In order to keep these changes somewhat inside the realm of possibility and comparable to our other policy experiments, we keep the total capacity adjustment and service delivery times constant for all simulations. Thus, while any one clinic may alter their parameters, the sum of these parameters across the service chain will remain constant. It is also important to note that the parameter adjustments we use to manifest these operational changes to decision making can easily accompany the other agile operational plans identified in previous sections.

7.1.4 Process Integration & Performance Management

Supply chain integration has been described as the 'holy grail' of supply chain improvement (Holweg et al, 2005). It is widely accepted that creating a totally seamless, synchronized supply chain will lead to increased responsiveness and lower inventory costs. Jointly creating the common practices for "information sharing, replenishment, and supply synchronization … is essential to avoid the costly bullwhip effect that is still prevalent in so many sectors" (Holweg et al, 2005, p.180). However, in the light of the complexity of today's global supply chains, most firms find it is hard to reap the full benefits from their efforts of integrating with their supply chain partners. Only a few individual success stories have been reported in the industry sector; mainstream implementation within these industries has been much less prominent than expected. In practice, the issue of how to benefit from process integration and how to use performance management to improve capacity utilization and inventory turnover is still not well understood, nor even well defined (Lapide, 2001).

There are many reasons complete integration remains elusive to most firms. The right approach for any firm depends on the supply chain context, in terms of geographical dispersion of retailers and supplier plants, complexity of distribution networks, and constraints on production modifications, as well as in terms of product characteristics and demand patterns (Holweg et al, 2005). Also, there are many different possible strategies to pursue to integrate a supply chain, and most steps toward complete integration, from information sharing to adopting uniform decision rules and service targets, are costly to implement, provide unequal benefit to each stage, and have high potential for generating adverse effects. While the promises of improved performance generated by each strategy are real, actually achieving successful implementation is rare, and achieving those improvements is rarer still.

Most of the supply chain improvement strategies found in the literature and discussed in this paper could not occur without integration of some kind. If the decision to adopt any of these strategies was driven solely by the benefits that would accrue naturally to each stage, then none of them would ever be adopted. For example, out of many possible scenarios, the lowest cost strategy identified in a simulation of the mortgage service industry (Anderson & Morrice, 2000), is where each stage only uses end-customer demand to make capacity adjustment decision. While by far the most efficient supply chain structure overall, the first stage bears all the burden

of demand volatility, while all the benefits of information sharing go only to the downstream stages. Such a 'raw deal' would never arise without the integration of these stages through the creation of additional structures to redistribute the overall benefits of information sharing more equally between supply chain partners. More recent studies suggest this is the norm, that sharing information will only improve performance of downstream stages, never the first stage (Anderson et al, 2006). The same dynamic occurs with strategies to promote efficiency through centralized planning, where no matter what the cost schemes, fundamentally, the benefits of centralized planning do not accrue evenly across all stages in a service chain.

Obviously, it is difficult to encourage each stage to participate in these different improvement strategies when local incentives differ so dramatically. This shows how crucial integration is to achieving efficient supply chain operations. The successful implementation of any of the other improvement strategies discussed requires finding and implementing an incentive scheme to compensate each stage appropriately. Supply chain simulation is an important tool in the design of such integrative incentive structures.

To explore the impact and importance of the agile strategy of supply chain integration, we focus on the need for altering performance measures to promote and sustain these policies and, with a generic cost structure, how efficiency gains must be redistributed to ensure that these policies are actually beneficial to each stage, not just overall. We determine the change in performance caused by each strategy for each stage, and use these findings to describe the necessary redistribution scheme so all stages would be willing to participate. The integration of incentive structures and performance management is key to achieving operational efficiency.

8. Simulation Analysis

In this section, we discuss our selection of performance measures and how they compare to general performance measures previously developed for the evaluation of service chains. Next, we present base case simulations that, consistent with the literature, establish demand variation amplification as an inherent system behavior.

8.1 Performance measures

The most common measures of supply chain simulations are of backlog and capacity variation (for discussion, see Anderson et al, 2006). These most clearly reveal the extent of inherent demand variation amplification, the 'bullwhip effect,' in a supply chain, and quantify the effects of mitigation strategies. It is possible to associate costs with each, but these can be very different depending on specifics of supply chain and stage in question. Not all variation creates cost equally. In a healthcare service chain, costs between service chains and between individual clinics vary considerably. For example, an increase in the post-surgery patient backlog would cost a hospital thousands of dollars per day, as patients took up more hospital beds and attendant

care; whereas an increase in the backlog of patients in the ED waiting room would cost almost nothing. Obviously, the impact to patient is also very different. In the first case, there is probably a null effect, with increased risk of nosocomial complications countered by increased attention; while an increased backlog in the second case clearly has a detrimental effect on patient health. The same is true for service capacity variation: hiring and training a personal care attendant incurs very different costs than hiring and training a pediatric neurosurgeon. Reducing service capacity also incurs some costs, quantitatively with possible severance pay and qualitatively through reduced morale with remaining staff. However, to keep our simulation results generalizable, we do not associate a cost measure with variation, only reporting averages and standard deviations of both backlog and service capacity. It should be noted that adding cost equations to each stage in the model is easily done, if context specific cost data are available.

We also use service time as a measure of performance, which is a common measure of both general service quality and healthcare quality (Parsuraman et al, 1988). Instantaneous average service delivery times for each clinic are calculated based on Little's Law, as the quotient of the current backlog of patients by the rate at which the clinic completes its service, and summed to generate the total average service time. We report both the average and standard deviation of service time for each clinic and the chain as a whole.

$$\textit{Average Patient Service Time} = {}^{B(t)}\Big/_{R_c(t)}$$

where $B(t)$ = patient backlog at t
$R_c(t)$ = completion rate of patients at t

The final measure we consider when evaluating the impact of agile strategies on service chains is the patient to provider ratio. Like service time, this is another measure of care healthcare quality. Healthcare services research has linked the ratio of patients to providers, and the subsequent employee stress and fatigue, to increased error generation, patient safety risk, and reduced overall care quality (Kane et al, 2007; Robertson & Hassan, 1999). Higher patient to provider ratios have been correlated with increased patient mortality, failure-to-rescue (deaths following complications), urinary tract infections, pneumonia, thrombosis, and pulmonary compromise (Aiken et al, 2002; Kovner & Gergen, 1998). While not a precise measure of service quality, it is easily comparable across stages and service chains and could be easily modified to provide more setting specific indications.

However, patients and service capacity are not directly comparable in our model, as they are measured in different units. We convert the measure of patients to that of service capacity, through comparing it to the target service delivery time. This adjusted measure, normally called 'workload,' now represents the ratio of the service capacity necessary to see the current backlog of patients within current standard of care and the service capacity currently available. This

measure of stress and patient safety risk can be modified to suit any service delivery system, and is accepted as a general measure of service supply chain stress and a main contributing factor to reduced service quality and increased rework. This ratio was first proposed by Akkermans and Vos (2003), and is similar to the measures of 'schedule pressure' found in system dynamics workforce models (Lyneis & Ford, 2007).

$$\textit{Normalized Workload} = \left[\left(\frac{B(t)}{Ts}\right) / C(t)\right]$$

where $B(t)$ = patient backlog at t,
T_s = target patient care service time
$C(t)$ = resource capacity available at t

To illustrate how this measure is used, assume a 10% increase in patient demand makes the workload measure triple from 0.1 to 0.3, this suggests severe demand variation amplification, but it also indicates that the system is not put under serious pressure because the workload is still well below 1.0, where 1.0 indicates that demand for services and current service capacity are in equilibrium, and thus all current patients can be seen within the desired service time.

This measure provides an instantaneous measure of provider stress and system flexibility, and is useful for evaluating behavior over the course of a simulation. However, to facilitate comparison of stress and flexibility across multiple simulations, we must condense this behavior into one number. Based on a technique common to control theory (White, 1999), we use the sum of the absolute difference between equilibrium and actual workload generated in all clinics. This accumulated error (the difference between desired and actual workload) stores the history of behavior over the entire simulation, resulting in a less volatile and clearer picture of how different policies affect performance over time, not just at one moment in time.

This measure also provides an estimate of overall system flexibility. If the ratio of patients to providers is often not balanced, then the system is not able to effectively and efficiently address changes in demand for services with changes in service capacity. For example, when workload is high, there are more patients waiting than there are necessary providers to diagnose, treat, and care for them in a timely manner, indicating that the system was not able to successfully respond to the initial increase in demand. The same is true when workload is below 1: the system has more resources than it needs to be able to provide the standard level of care.

8.2 Base Case

We run an initial simulation to reveal dynamic behaviors inherent to the system. Key parameters in the model are initially set to the relative equivalents of those found in other generic service supply chain models. The total desired patient service time is 15 days, spread evenly over each clinic (T_s is 5 days); the time to add or remove a provider from any clinic roster is four times as

long (T_c = 20 days). The model starts in equilibrium, where exogenous demand is a constant rate of 10 patients per day, and each clinic is staffed with the exact number needed to meet that demand in the target service time. Thus, initially, there is no variation in backlog or service capacity, and average service time is equal to desired service time. Desired capacity and actual service capacity are equal, therefore our performance measure, workload, is 1.0 in each service clinic. This scenario represents the traditional healthcare service delivery system, where local managers control the workforce at each clinic using only information on their local clinic backlog and provider productivity. In this base case scenario, the system is disturbed from that equilibrium by a minimal level of demand uncertainty, a one-time 10% increase in patient demand.

The model structure clearly generates the demand variation amplification effect, as expected. The results outlining the effect of variation in demand for services on each clinic over a one year period are contained in Figure 5. These oscillations are mirrored in the clinic performance measures. For example, the 10% increase in demand causes clinic workload to peak at 13.4%, 14.5%, 20.0%, in the first, second, and third clinic, respectively. Patient service time averaged over the course of the simulation do not vary significantly between clinics (as would be expected in a return to equilibrium), but the variation is significant, with service times error peaking at 0.67, 0.72, and 1.0 days. This finding compliments previous healthcare service delivery research (Walley, 2007; Sethuraman & Tirupati, 2005), with has identified increasing downstream variation common to both service rates and patient backlogs.

The amplification effect arises from delays in demand signaling and the limited information and bounded rationality of individual clinic managers. As each clinic transfers demand to subsequent clinics, they unknowingly magnify variation as patients move up the service chain, creating system stresses proportionally much larger than the initial increase in demand. Clinics in this model are highly compartmentalized; they share no data on capacity adjustment, patient backlog or service quality. This lack of coordination and information on the other stages is the chain represents the typical organization of care both between healthcare organizations and inside hospitals.

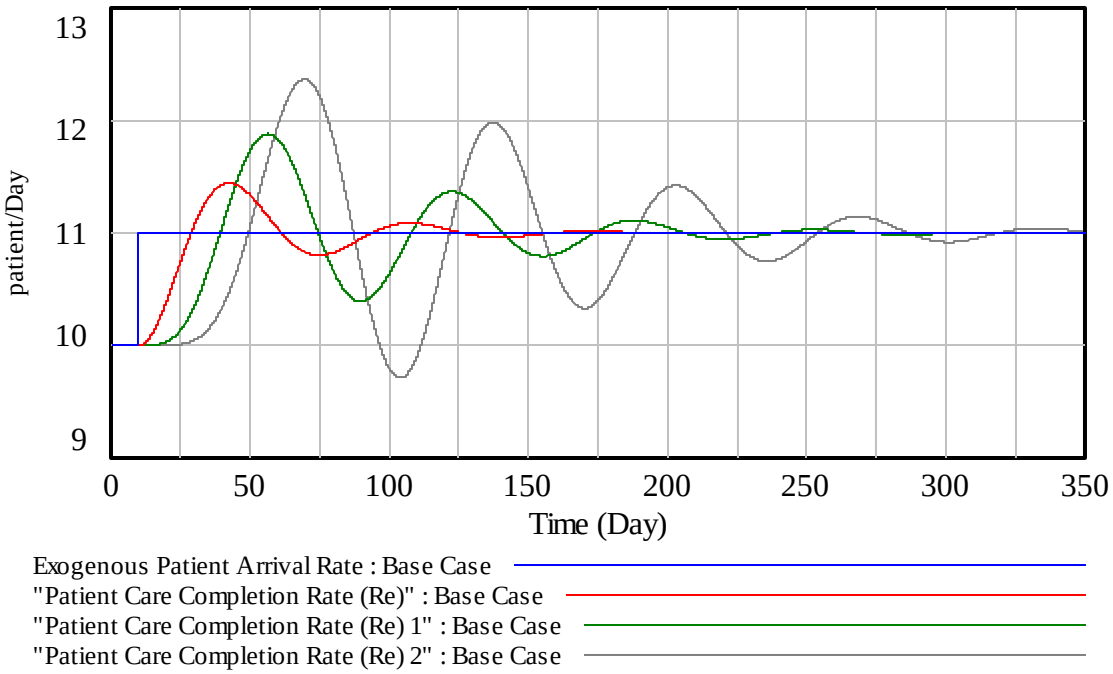

Figure 5. Base case analysis of individual clinic service rates resulting from by an instantaneous 10% increase in demand in week 10 (demand rate changed from 10 patients per day to 11 patients per day). Simulation was conducted in Vensim® software version 6.0 using Runge-Kutta integration methods, with a time step of 0.0625 days.

Given this 'global' perspective on system behavior, the decision rules used in the model clearly lead to unintended adverse effects. After the initial demand disturbance, it takes the clinics between six months and over a year to realign the supply of services with demand, creating intense variation in patient service times, clinic workload, and care quality. These results, while somewhat dependent on model parameters, suggest that the traditional organizational structures governing the management of services do not provide the necessary flexibility to synchronize service supply with fluctuating patient demand.

9. Exploration of Agile Strategies

Increasing service chain flexibility is crucial to synchronizing clinical resources with patient demand, and thus the ability to provide cost effective, quality healthcare. In this section, we test whether each previously identified agile operational strategy improves supply chain flexibility, compared to the base case. While the literature suggests that all should improve performance, mitigate the 'bullwhip effect,' and improve system flexibility, not all have been examined in dynamic simulation, and none of them have been systematically evaluated against the others in an identical setting, facing an identical demand pattern. These simulations will provide an understanding of their general compare effectiveness in modifying the behavior of service chains. They also answer questions on whether or not a service chain requires improvement in

all four characteristics to become 'truly agile,' as proposed by Christopher (2000), or if some characteristics and operational plans are equally impactful on their own. We also discuss the feasibility of implementation of each operation strategy, specifically the need for an incentive structure to compensate for possible reductions in the performance of individual clinics. Finally, we discuss how each change in model structure creates these new behaviors and the generalizable implications for healthcare managers.

9.1 Market sensitivity

The first set of scenarios explores the impact of reduced service capacity adjustment time, T_c. Changes in this parameter can represent any operation changes that directly affect the speed of management decision making. Decreasing the capacity adjustment time has been shown to mitigate the 'bullwhip effect' in dynamic simulations of service chains (Anderson & Morrice, 2000, 2001; Anderson et al, 2005). Reducing this parameter means that each clinic now responds proportionally faster to any change in demand, rendering each clinic more market sensitive, and thus allowing less patient backlog to accumulate. Reducing the ability of the clinic to accumulate unwanted patient backlog is key to reducing downstream demand amplification.

We also explore the effects of the addition of derivative-based information into each clinic's decision heuristic. Adding this signal to the information used in the base case is shown to improve clinic performance and reduce costs associated with demand variation (Lee et al, 2010). The derivative is a faster signal of changes in the patient backlog than simply the measures of the backlog itself; by definition, the derivative peaks when patient backlog is only at its inflection point and still rising.

All of these operational strategies to improve market sensitivity have effect predicted: all yield improvements over base case. However, not all have an equal impact on performance, as shown in Figure 6. The key finding in this set of simulations is that the addition of derivative control appears to be the most effective operational strategy of the set for improving system performance. All strategies reduce the amplitude of variation, but decreasing capacity adjustment time increases the oscillation frequency, while the addition of derivative control returns the system to equilibrium faster than any other strategy in this set. Furthermore, doubling K_d, the weight clinic managers give to the information on the derivative in their decision equation, further improves the performance under the derivative control strategy. Such adjustments are more common in highly measurable systems such as manufacturing, and while more difficult in the healthcare setting, this simulation sheds light onto the improvements made possible by 'fine tuning' management's capacity control heuristics.

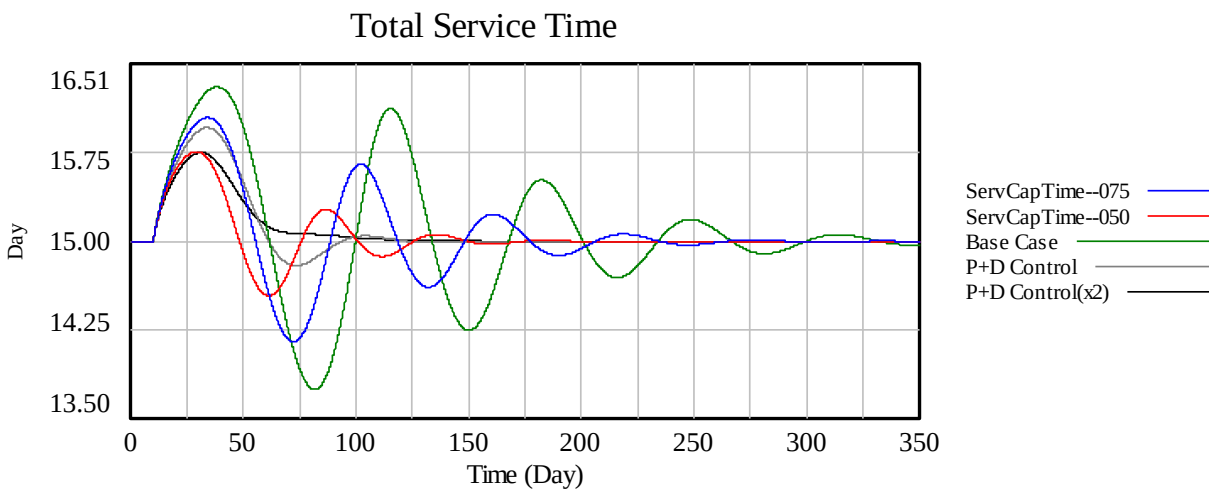

Figure 6. Resultant service times from market sensitivity simulation runs.

Behavior of total service delivery time (sum of each clinic's instantaneous average wait time, as computed by Little's Law) in response to an instantaneous 10% increase in demand in week 10. Figure includes five different scenarios for the model parameter and formulae governing market sensitivity. Simulation was conducted in Vensim® software version 6.0 using Runge-Kutta integration methods, with a time step of 0.0625 days.

In terms of total service chain performance, a greater than 50% reduction in the service capacity adjustment time is required to create the same benefits as basic derivative control (see Figure 7). Basic derivative control generates a 78% decrease in workload error over the course of the simulation and a 63% decrease in service time variation, compared to the base case. One could infer these results to indicate that both methods are equally useful, but they are not equally cost-effective. Derivative control is by far easier to implement. Including information on the derivative of the patient backlog in a manger's decision could be done with a simple spreadsheet, while changing the service capacity adjustment time would require intense effort in HR process redesign. Achieving a 75% reduction would be difficult for most healthcare clinics, with 50% being practically impossible.

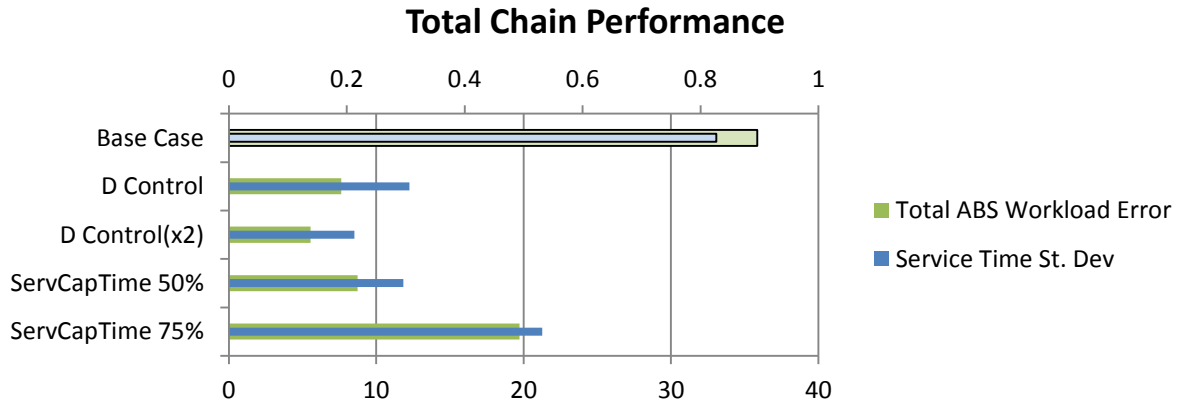

Figure 7. Resultant service supply chain performance under various market sensitivity strategies.

Behavior of total accumulated absolute workload variation and variation in total average service times (sum of each clinic's instantaneous average wait time, as computed by Little's Law) in response to an instantaneous 10% increase in demand in week 10. Figure includes five different scenarios for the model parameter and formulae governing market sensitivity. Simulation was conducted in Vensim® software version 6.0 using Runge-Kutta integration methods, with a time step of 0.0625 days.

All individual clinics incur some benefits under each market sensitivity strategy (see Figure 8). Downstream clinics benefit more from any market sensitivity strategy than upstream clinics, as the effects of dampened demand variation are cumulative, and because downstream stages initially incurred more of the burden of inherent demand variation amplification. In terms of variation in patient backlog and service capacity, derivative control leads to slightly better outcomes than a 50% reduction in service capacity adjustment times in all clinics. Under the derivative control scenario, the final clinic has 46% less variation in capacity and 63% less variation in its patient backlog, compared to -42% and -62% change under the $0.5T_c$ scenario. These results reveal that actual implementation of either strategy can be accomplished without the creation of additional incentives or a benefit redistribution structure. While downstream clinics do benefit from the adoption by upstream clinics, there is no need create further incentives to encourage any clinic to participate.

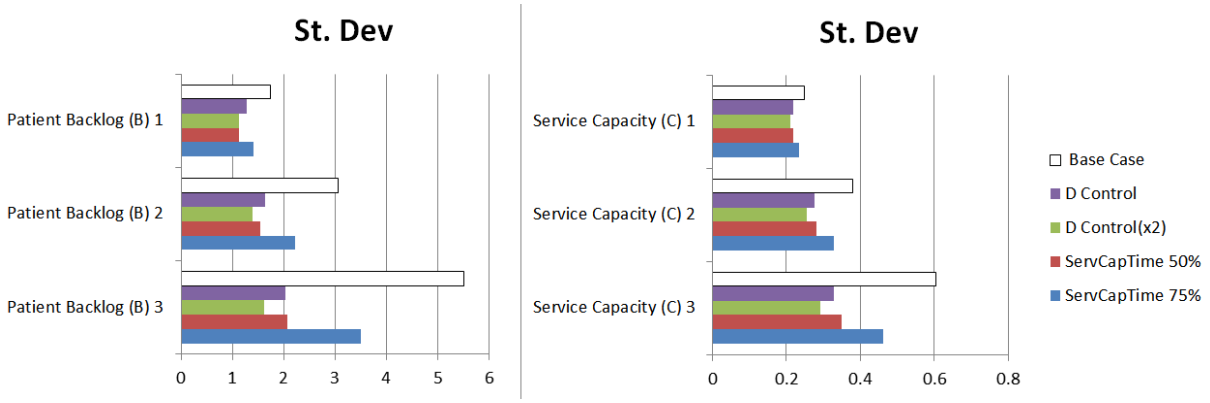

Figure 8. Variation in individual clinic backlogs and capacities following market sensitivity strategies.

Behavior of variation in patient backlogs and service capacities in response to an instantaneous 10% increase in demand in week 10. Figure includes five different scenarios for the model parameter and formulae governing market sensitivity. The time horizon for each simulation is 350 days. Simulation was conducted in Vensim® software version 6.0 using Runge-Kutta integration methods, with a time step of 0.0625 days.

9.2 Information Sharing

The second set of scenarios explores the impact of including real-time, end-patient demand data in the decision equation at each clinic. Including this information means that each clinic now responds to both changes in their local patient backlog and to the patients that have started the care process but have not yet arrived at their clinic. Increasing the visibility of real demand throughout the supply chain has been shown to mitigate the 'bullwhip effect' in real-world manufacturing chains (Holweg et al, 2005). Including initial patient demand in individual clinic's capacity adjustment decision making has also improved performance of service chains in dynamic simulations (Anderson & Morrice, 2000, 2001; Anderson et al, 2005). However, implementing this strategy in the real world requires a costly change in operations: including the installation of IT infrastructure to collect and transmit the data in real time, and training managers on how to incorporate these new data into their decision heuristics. Also, the relative weight given to each source of information is both difficult to intuit correctly and fundamentally important to determining overall performance. It should be noted that initial patient demand is a different type of information than managers used in the base case: it is an instantaneous rate of patient arrival, as opposed to a stock of patients waiting (or, depending on the formulation, the stock of patients in process).

We test two different versions of the information sharing strategy, one where information on initial patient demand completely replaces local backlog in the capacity adjustment decision ($a = 1$), and another, more reasonable version, where managers use a mix of local backlog and initial patient demand ($a = 0.5$). In the runs described, all clinics use the same relative weight parameter (we did run experiments varying these weights between clinics, but those runs did not yield significant system improvement or insight into model behavior). For the chain as a whole, both versions of the information sharing strategy reduce workload and service time variation under most scenarios, as shown in Figure 9.

The mixed information version leads to significant improvements in mitigating workload variation, but basing capacity adjustment decisions fully on initial patient demand (i.e., not using local information at all) appears to yield even more improvement. Under the condition $a = 1$, total accumulated workload error is 84% lower than the base case; service times in this version are also more controlled under most scenarios. These results suggest that using initial patient demand in place of local demand increases system flexibility, yielding a faster and more accurate response to demand variation than the traditional decision structure.

We explore the validity of the strategy of only using initial patient demand in all clinics by testing it in multiple versions of our generic service chains, where all clinics are no longer identical. The variation in individual clinic parameters provides a more realistic and representative simulation of the complexity seen in actual healthcare service chains. In these scenarios, each clinic is portrayed as conducting a different care processes, requiring different

target clinic service times. Also, each clinic manages their service capacity differently, with each clinic subject to a different capacity adjustment decision time. Target service time is set at either 2, 5, or 8 days (labeled as L, M, and H for 'low,' 'medium,' and 'high' values); while capacity adjustment time varies between 7, 20, and 33 days. Despite these changes, each scenario remains comparable to the other strategy evaluation simulation runs because total service time and total capacity adjustment delays for the overall chain remain constant ($\sum T_s = 15$ and $\sum T_c = 60$, for all scenarios).

Even under these more realistic conditions, where all clinics are not identical, making decisions solely with initial patient demand often yields better outcomes than either the mixed information strategy or the base case. Only in one simulation run, where the first clinic's parameters produce low market sensitivity, as the capacity adjustment time is set to 33 days, did this strategy lead to a worse outcome, specifically a service time variation 73% worse than the base case. In cases when the first clinic is equally or more responsive to changes in demand than the base case, overall system performance improved. Results are described in Figure 9.

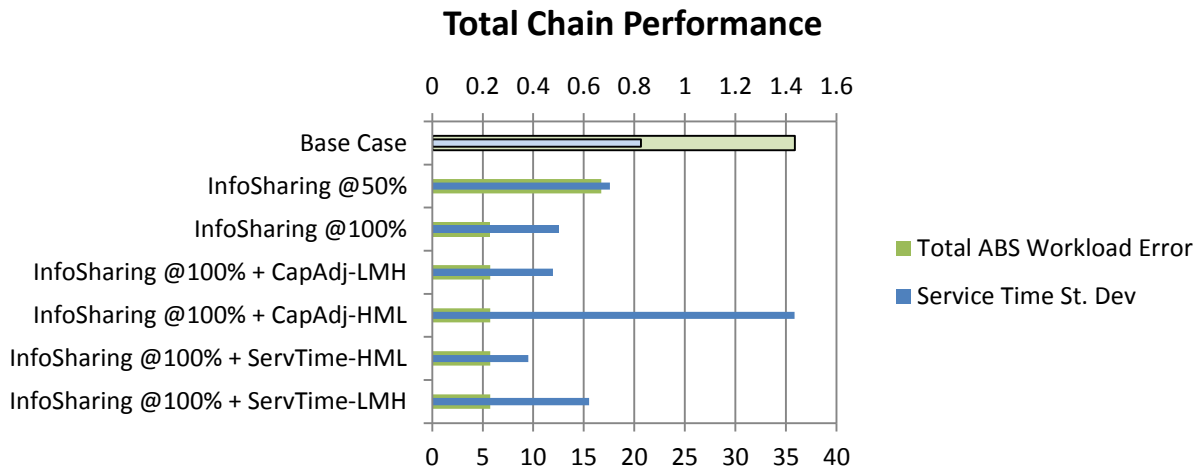

Figure 9. Resultant service supply chain performance under various information sharing strategies.

Behavior of total accumulated absolute workload variation and variation in total average service times (sum of each clinic's instantaneous average wait time, as computed by Little's Law) and in response to an instantaneous 10% increase in demand in week 10. Figure includes seven different scenarios for the model parameter and formulae governing information sharing under differing assumptions of clinic characteristics (parameters delta and tau varied between 2, 5, and 8 days and 7, 20, and 33 days respectively). The time horizon for each simulation is 350 days. Simulation was conducted in Vensim® software version 6.0 using Runge-Kutta integration methods, with a time step of 0.0625 days.

This variation in service chain performance can be explained by examining the behavior of individual clinics, as shown in Figure 10. Not all clinics are affected equally by the extreme reliance on initial patient demand information. A detailed analysis reveals that most of the predicted benefits of relying on initial patient demand are generated in the model by the complete elimination of variation in downstream clinics patient backlogs. This is more than the complete elimination of the amplification of variation, or 'bullwhip effect,' it is the complete elimination of any variation whatsoever. However, these impressive outcomes only occur in the improbable scenario where all clinics have identical capacity adjustment times. More realistic scenarios with variable decision making practices between clinics indicate that these results would not be produced in real-world healthcare service chains, where maintaining identical management decision making heuristics and HR processes across the entire chain is highly unlikely.

These uneven outcomes across clinics render implementation of the information sharing strategy difficult. In almost all scenarios, the first clinic in our service chain incurs more variation in patient backlog when including initial patient demand in their management decisions than they would without. This leads to more variation in service times and service quality levels. Even when demand information is combined with local backlog information, such as the $a = 0.5$ scenario, the first clinic is still subject to an increase in backlog variation. Any implementation of information sharing strategies would require a drastic benefits redistribution mechanism to compensate the first clinic for the use of demand information.

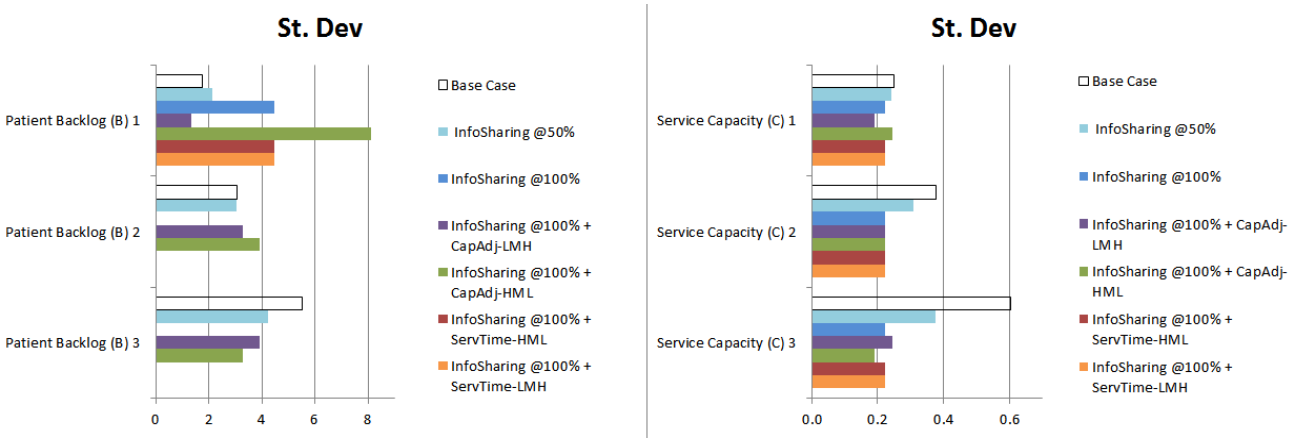

Figure 10. Variation in individual clinic backlogs and capacities following information sharing strategies. Behavior of variation in patient backlogs and service capacities in response to an instantaneous 10% increase in demand in week 10. Figure includes seven different scenarios for the model parameter and formulae governing information sharing under differing assumptions of clinic characteristics (parameters delta and tau varied between 2, 5, and 8 days and 7, 20, and 33 days respectively). The time horizon for each simulation is 350 days. Simulation was conducted in Vensim® software version 6.0 using Runge-Kutta integration methods, with a time step of 0.0625 days.

The only case where patient backlog variation in the first clinic declines is when increasing market sensitivity (through a reduction in T_c) overcomes the detrimental effects of initial patient demand data. These simulations indicate that market sensitivity of the first clinic is a significant factor in determining the performance of the overall chain. When first clinic's ability to respond to changes in demand is low, the variation created in the first clinic outweighs the benefits of increased market sensitivity in downstream clinics. For example, in the *ServTime-HML* scenario, decreasing market sensitivity in the first clinic is (by an increase T_c by 65%, to 33 days) results in a 367% increase in overall service time variation over the base case (as shown in Figure 9), even though opposing parameter changes in downstream clinics rendered them more responsive to the increased demand fluctuations generated by the first clinic. This set of scenarios exposes that replacing local backlog data with initial patient demand data is not a feasible solution for improving service chain performance. Any differences between clinics in management practices and decision heuristics will eliminate the benefits indicated in previous simulation studies (Anderson & Morrice, 2000).

Moreover, while the information sharing strategy does reduce variation in some scenarios, it does so by creating a more significant problem. The complete use of initial patient demand rate in management decisions leads to steady state error in patient service time, as shown in Figure 11. Completely ignoring local data results in a dangerous scenario, where individual clinic managers are blind to the impact of delays in capacity adjustment on patient service times. This consequence, no matter what the possible benefits from reduced variation, is not acceptable in healthcare service chain management.

This steady-state error results from only using a proportional control based on the rate of demand, which cannot keep track of the error built up over the time period when a correction is being made. The 100% initial patient demand decision heuristic will synchronize a clinic quickly to any new demand rate under many clinic configurations and parameter sets, but once supply is again matched with demand, there is no information retained on the accumulation of error that has developed in the interim. Without including this information on clinic backlog, as would occur in the initial decision heuristic, these patients are never accounted for, and their impact on performance is never corrected.

Under the scenario of the 100% use of initial patient demand rate, more variation in demand would lead to more accumulated error remaining unaccounted for. If a system is experiencing demand variation around a steady mean, this ‘ignored’ error would cancel itself out, but if demand fundamentally increases or decreases, steady state error will necessarily occur. With the likelihood of a steady mean demand rate a near impossibility, relying solely on initial patient demand to make clinic capacity adjustment decisions is not a realistic option for healthcare services.

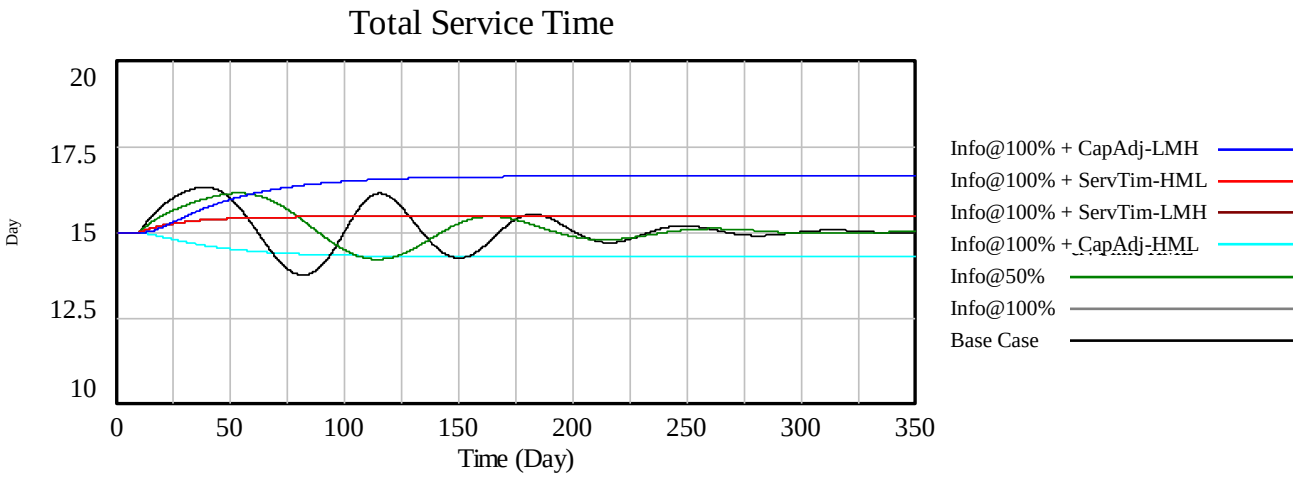

Figure 11. Resultant average patient service times from information sharing simulation runs. Behavior total service time (sum of instantaneous average service times, as calculated by Little's Law, for each clinic) in response to an instantaneous 10% increase in demand in week 10. Figure includes seven different scenarios for the model parameter and formulae governing information sharing under differing assumptions of clinic characteristics (parameters delta and tau varied between 2, 5, and 8 days and 7, 20, and 33 days respectively). The time horizon for each simulation is 350 days. Simulation was conducted in Vensim® software version 6.0 using Runge-Kutta integration methods, with a time step of 0.0625 days

9.3 Coordinated planning

The final set of scenarios explores the impact of centralized planning, specifically the use of a systems perspective to determine clinic decision heuristics and target service times to maximize supply chain performance. These changes are meant to maximize performance for the total chain, even if individual clinics generate worse performance. From past simulation research, we know that service chains should not move toward decision synchronization, as services with identical stages show worse performance than services with varied stages (Anderson et al, 2005). We explore the implications of this finding to determine if healthcare services can operationalize increasing decision and service target variation to improve overall performance.

We manifest this idea through changes to clinic parameters T_c and T_s. To illustrate the possible feedback effects, consider the example of an increase in T_c, which results in the clinic now responding proportionally slower to any change in demand, rendering that clinic less market sensitive, and thus allowing more patient backlog to accumulate. Similarly, reducing T_s results in a clinics needing to maintain more staff for the same level of demand, which allows patients to wait less on average to complete services in that clinic. Analogous to the 'more realistic' scenarios in the previous section, we maintain a fixed $\sum T_c$ and $\sum T_s$ for the service chain as a

whole, thus there are no improvements in overall standard of care or market sensitivity. These parameter changes isolate the effects of a redistribution of HR resources (changing T_c allows clinics to make and execute capacity adjustment decisions faster) and service delivery standards (changing T_s directly affects a clinic's average service time).

Decision and service standards de-synchronization has a minimal effect on the behavior of total service time. There is a minor improvement in workload error, compared to the base case, confirming results from previous studies (see Figure 12).

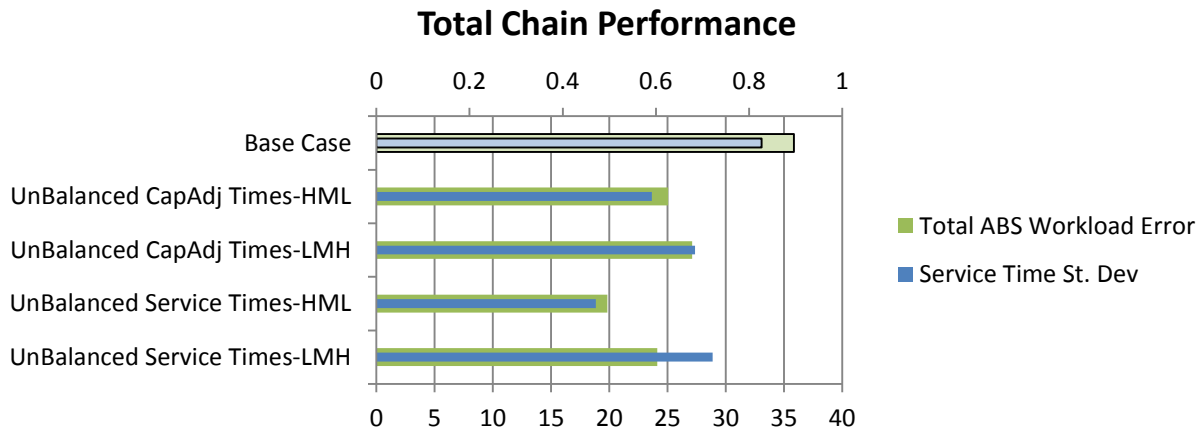

Figure 12. Resultant service supply chain performance under various coordination strategies. Behavior of total accumulated absolute workload variation and variation in total average service times (sum of each clinic's instantaneous average wait time, as computed by Little's Law) in response to an instantaneous 10% increase in demand in week 10. Figure includes five different scenarios for the model parameter and formulae governing service chain coordination (parameters delta and tau varied between 2, 5, and 8 days and 7, 20, and 33 days respectively). Simulation was conducted in Vensim® software version 6.0 using Runge-Kutta integration methods, with a time step of 0.0625 days.

Simulation results suggest that maintaining relatively longer average service times in the first clinic in the service chain yields the best performance of these decision and performance standard de-synchronization scenarios. Changes to the parameter governing target service time (T_s) alters the desired patient backlog level implicit in clinic managers' decision heuristics, affecting the average size of the buffer each clinic maintains against demand variation. By setting performance standards higher in the final clinics of the service chain, the clinics at the beginning of the service delivery chain are allowed lower relative efficiency and larger patient backlogs, when compared to the balanced strategy. This provides the first clinic with more patient demand buffer, so less internal demand variation amplification is passed on to subsequent clinics. This 'front-loaded' service supply chain buffers the entire service chain from external

demand variation, resulting in less variation amplification overall. All else equal, keeping patients concentrated at the beginning of a service delivery chain better accommodates demand fluctuation, resulting in less system stress in any clinic. Holding relatively more patients at the front of the care process (and fewer in the later clinics, to maintain an equivalent total number) also leads to less workload variation than any other distribution.

Changes to capacity adjustment time have a similar effect, but less pronounced than the de-synchronization of service targets. This occurs because the size of T_c is relative small compared to T_s. If the T_c:T_s, ratio were larger, then unbalancing capacity adjustments would have more impact on service performance. Re-designing HR processes to affect capacity adjustment rates could be an important operational improvement if applied in a service chain where T_c:T_s is relatively small.

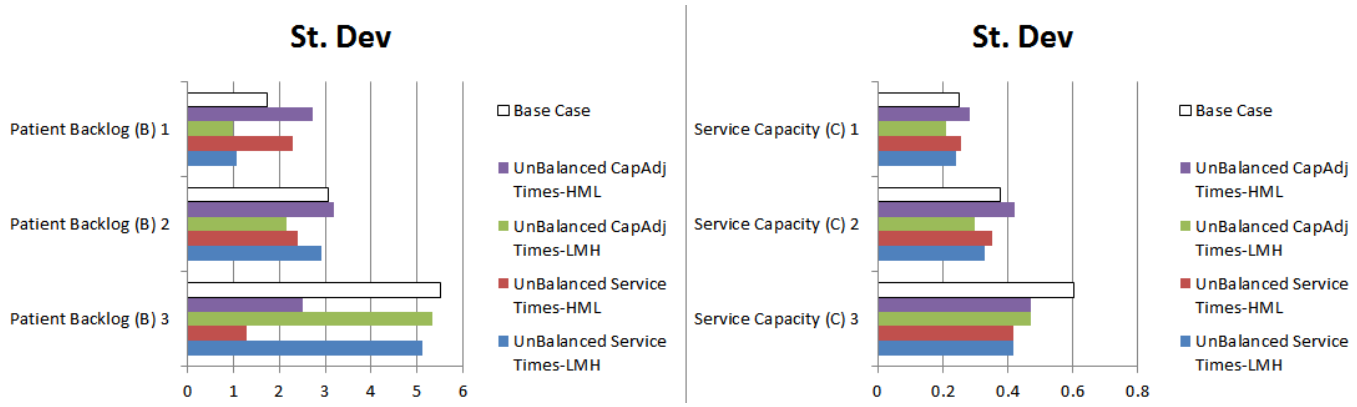

Figure 13. Variation in individual clinic backlogs and capacities following centralization strategies.

Behavior of variation in patient backlogs and service capacities in response to an instantaneous 10% increase in demand in week 10. Figure includes five different scenarios for the model parameter and formulae governing service chain coordination (parameters delta and tau varied between 2, 5, and 8 days and 7, 20, and 33 days respectively). Simulation was conducted in Vensim® software version 6.0 using Runge-Kutta integration methods, with a time step of 0.0625 days.

Overall, these simulations suggest that using centralized planning to redistribute service targets and capacity adjustment times has limited impact on service performance, compared to other possible agile strategies. Furthermore, the fundamental characteristics of individual clinics in actual healthcare service supply chains limit the ability of managers to alter these service targets.

There also exist complex interaction effects between operational strategies to alter these parameters. Tests combining both parameter changes across the service chain (still maintaining $\sum T_c = 60$ and $\sum T_s = 15$) reveal that the impact of changing one set of parameters is strongly influenced by the distribution of the other set, as shown in Figure 14. These experiments reveal a general conclusion that clinics with low desired patient backlog levels should maintain high market sensitivity, as they have minimal buffer against variation in demand, so they must be able

to change service capacity quickly to minimize the accumulation of error. In these runs, the strategy of 'front-loading' the service chain, where the first clinic maintains the largest patient backlog (the *ServTimes-HML* scenario), still produces the best performance; however, it can also produce nearly the worst performance if capacity adjustment times have the opposite distribution (the *CapAdj-LMH* scenario, where the first clinic is the most market sensitive). The opposite strategy, of 'rear-loading' the service chain (where both *ServTimes* parameters are distributed *LMH*), also produces little workload error if matched with a similar distribution of capacity adjustment times. These seemingly contradictory results further reduce the ability to make simple guidelines for unbalancing service chains.

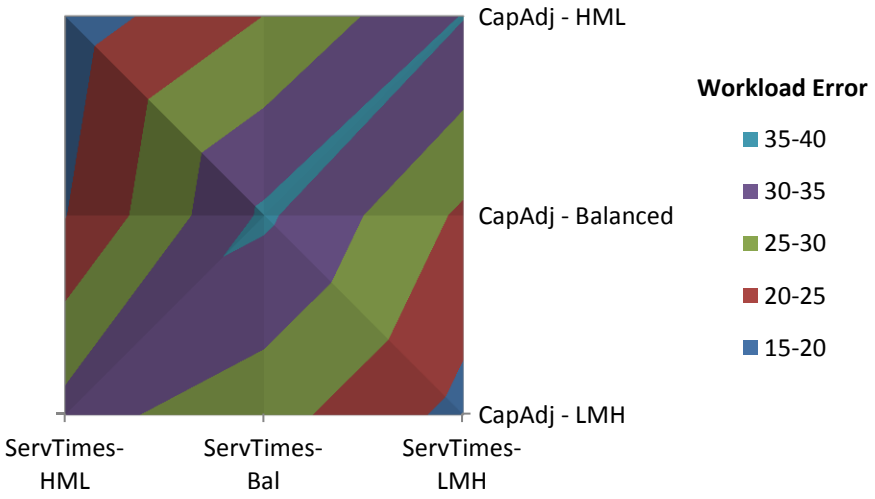

Figure 14. Resultant workload error following clinic parameter re-distribution. Behavior of total accumulated absolute workload variation in response to an instantaneous 10% increase in demand in week 10. Figure includes 25 different scenarios for the model parameter and formulae governing service chain coordination (parameters delta and tau varied between 2, 5, and 8 days and 7, 20, and 33 days respectively). Simulation was conducted in Vensim® software version 6.0 using Runge-Kutta integration methods, with a time step of 0.0625 days.

Attempting to actively alter variation in clinic standards to increase performance should not be made without first understanding the relative market sensitivity of each clinic in the service chain, and vice versa. This centralization strategy presents no simple solutions. Considering the diversity of requirements for any healthcare clinic, the difficulty clinic managers will have assessing the relative value of T_s and T_c for all clinics in a healthcare chain, and the complexity of service delivery in the real world that is not included in this model, this strategy is probably the least useful of the agile operational strategies tested so far.

9.4 Results summary

This series of operational simulations leads to multiple conclusions. The first key finding is that agile strategies do not need to be implemented together to produce significant results. Promoting

individual agile characteristics appears to be an effective improvement strategy for service delivery chains.

Second, under these simplifying assumptions, improving market sensitivity is the most effective agile strategy for improving performance in service chains, as shown in Figure 15. The specific operational plan of introducing derivative-based controls into managers' decision heuristics yields the most improvement in service quality and reduction in cost drivers. This operational change also mitigated internally produced demand variation amplification, the 'bullwhip effect,' more than any other agile operation plan. Furthermore, the inclusion of derivative information improved both overall system performance and that of individual clinics, thus requiring no extra benefits redistribution mechanism to encourage the adoption of this strategy.

The third key finding is that implementation of operational plans to increase a service chain's ability to be 'information driven' can unintentionally produce significant adverse effects. While the sole use of initial patient demand in capacity adjustment decisions appears to be a promising strategy, basing this 'feed forward' proportional controller on the exogenous variable of patient demand leads to steady-state error in key performance metrics. These simulations expose the importance of including endogenous variables in each clinic's control decisions. However, a blended information approach, which is a much more likely implementation in real-world service chains, is not as effective at controlling the bullwhip effect and minimizing the patient safety risk and care delivery costs created by those demand fluctuations than strategies to promote market sensitivity.

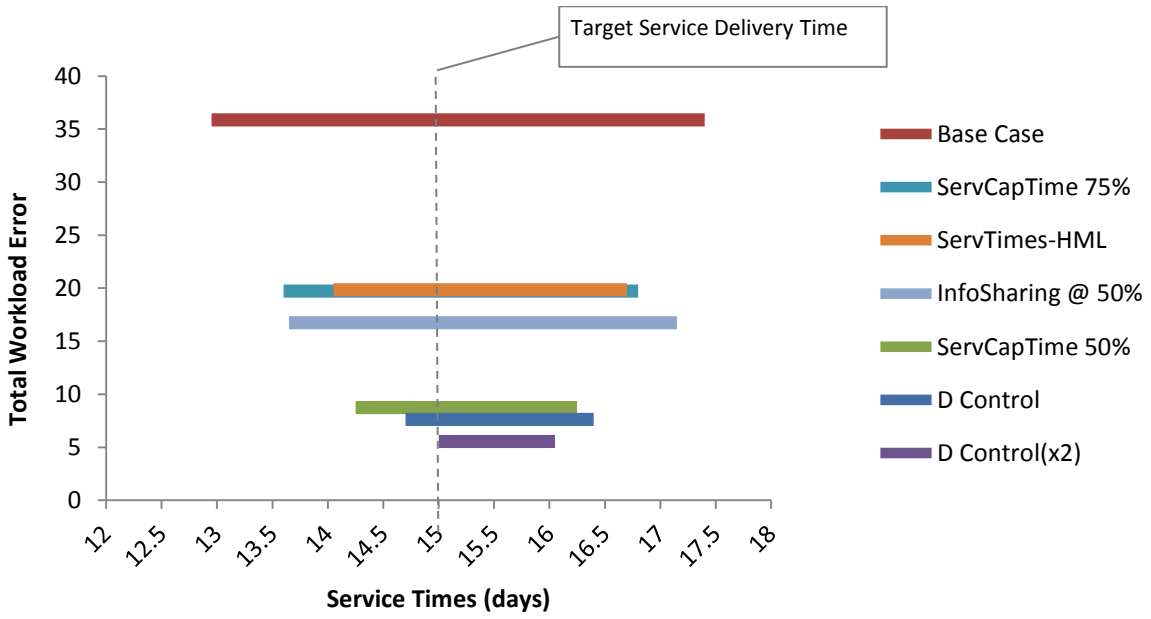

Figure 15. Comparative effectiveness of selected agile operational plans in response to a one-time 10% increase in demand.

The experiment includes seven different scenarios for the model parameter governing market sensitivity, information sharing, and service chain coordination. Chart reports the total range of average service times (computed by Little's law) and accumulated absolute workload variation summed over all clinics for each scenario, in response to an instantaneous 10% increase in demand in week 10. The time horizon for each simulation is 350 days. Simulation was conducted in Vensim® software version 6.0 using Runge-Kutta integration methods, with a time step of 0.0625 days.

10. Conclusion

To date there has been limited success in making system-wide service supply chain management improvements in healthcare (McKone-Sweet et al, 2005; Vries & Huijsman, 2011). We still face significant challenges designing and implementing cost-effective, and at the same time flexible, healthcare systems which increase the availability of scarce service resources and improve patient access to care. Past efforts applying management strategies developed in manufacturing settings have led to little sustained improvement (Joosten et al, 2009). Defining structural differences between services and manufacturing necessitate the adoption of new strategies more suited to the challenges facing healthcare operations and management.

This paper provides a structured assessment of the impact of one possible new strategy, 'agile,' on service performance in simulated healthcare delivery chains. In doing so, we bridge the supply chain and healthcare management literatures and establish 'agile' as a new area of study for service supply chain management research. Our research objectives were to develop a set of operational plans from the literature on 'agility' and service supply chain management and assess the impact of these agile-derived plans in a generalized healthcare service chain. The knowledge gained was to provide healthcare managers with useful guidelines for redesigning service delivery.

To fulfill the research objectives, we describe and test three sets of agile-based operational strategies, focusing on key characteristics of supply chain agility: increasing market sensitivity, the use of real-time demand information, and centralized planning. We assess the impact to the service system based on three criteria: variation in the stocks of patients and service capacity in each clinic (to expose the ability of each strategy to mitigate the bullwhip effect), average patient service time, and provider workload. These are measured both at the local (single clinic) and systemic level (total service chain).

We determine that agile is a valuable strategy for increasing system flexibility and mitigating internally caused demand variation. Scenarios show improved system performance from both the patients' and providers' perspective, with agile-based operational modifications leading to

reduced variation in service times, improved service quality, and the potential for decreased costs.

Of the agile characteristics under study, increasing market sensitivity led to the most improvement, with the specific operational plan of supplementing manager's traditional decision making heuristic with derivative control resulting in superior performance. Study results indicate that demand volatility can be effectively controlled in healthcare by applying derivative control to the resource adjustment decision. The addition of derivative control reduces the oscillation of patient backlogs and the discrepancy between demand and service capacity created by the simplistic feedback control methods commonly used in healthcare. The application of derivative controls in service chains could refute the conundrum identified by Anderson (et al, 2005), that there is generally a trade-off between policies that improve service quality by reducing backlog variability and those that reduce personnel costs by reducing capacity variability. We find that the addition of derivative control can effectively accomplish both.

In practice, derivative controls could be used to dampen oscillations resulting from any capacity management decision, from resource acquisition, release and write down of capital investments, to hiring and workforce training. Even the most basic derivative control should lead to a sizable improvement in synchronization of service resources with demand and, if implemented in all clinics, significantly mitigate the bullwhip effect. The addition of derivative-based controls is relatively simple to implementable in real-world service chains. Optimizing these control equations could further improve cost, utilization and stability of workforce management in healthcare, if reliable and timely data were available.

Study results also indicate that investing in IT systems to share demand data between clinics might not as useful in healthcare as predicted from the research done in manufacturing sectors. Our evidence runs counter to the common supposition on the value of information sharing strategies in healthcare, as summed up by Baltacioglu (et al, 2007, p121) as "effective management of healthcare supply chain is only possible via the implementation of effective information and technology management systems."
In our simplified service delivery chain, the use of initial patient demand rates either has less impact on performance than agile practices which increase market sensitivity, or leads to significant disruptions in service times and the alignment of clinic incentives. While possibly beneficial when viewing the chain as a whole, relying on initial patient demand does not appear to be an appropriate strategy for all clinics.

Generating results that are in direct disagreement with commonly held supply chain management beliefs could easily be attributed to the abstract nature of our simulation model. Effective information and technology management systems may address key issues and feedbacks that we have decided not to include, such as links between service delivery times and patient health, or

between provider workload, service quality, and rework. These information systems may also be useful in managing details on individual patient demand and provider characteristics not allowed by the mathematical underpinnings of our model. Other critiques of this research could be leveled at the applicability of our results to inform decisions in real-world healthcare chains, as our exogenous demand pattern is undoubtedly not representative of the usually stochastic demand pattern in healthcare. Each of these shortcomings deserves the attention of further research.

Our abstract model is a first step toward understanding and informing the application of agile strategies in healthcare. These results provide only the most general guidance on where agile-derived efforts to improve service delivery will yield the most return. Healthcare managers are still 'on their own' to adapt these recommendations to their unique care settings and service delivery chains. Future work should be directed to examining the validity of our findings under the constraints inherent to different service settings, both in simulation of specific healthcare service chains and empirically in pilot implementation projects in real-world clinics. Another thread of future research centers on derivative control. To truly develop useful guidelines for implementing an agile systems approach in healthcare, the ability of healthcare managers to use derivative control in individual clinics must be empirically evaluated. Case studies of the effectiveness of derivative control in 'noisy' real-world service chains would undoubtedly shed light onto important implementation challenges. A second piece of this thread would be exploring opportunities for 'tuning' proportional and derivative-based control decisions, based on data quality and availability and on the bounded rationality of clinic managers. Increasing knowledge in these areas together will support the creation of effective, flexible service chain management that suits the dynamic nature of health itself, and hopefully will lead to enhanced effectiveness and efficiency of healthcare operations.

References:

Aiken LH, Clarke SP, Sloane DM, et al: Hospital nurse staffing and mortality nurse burnout, and job dissatisfaction. *JAMA* 2002; 88:1987–1993

An Examination of Poorly Performing U.S. Department of Veterans Affairs Regional Offices, 112th Cong. No.112-16 (2011) (testimony of Jon Runyan).

Anderson E, Morrice D and Lundeen G (2005). "The physics of capacity and backlog management in service and custom manufacturing supply chains" *System Dynamics Review,* Vol.21 No.3, 217-247.

Anderson, E. G. (2001). The nonstationary staff-planning problem with business cycle and learning effects. *Management Science, 47*(6), 817-832.

Anderson, E.G. and Morrice, D. (2000), "A simulation game for service-oriented supply chain management: does information sharing help managers with service capacity decisions?", *Production and Operations Management,* Vol. 9 No. 1, pp. 44-55.

Argawal, A., Shankar, R., Tiwari, M., (2007), "Modeling agility of supply chain", Industrial Marketing Management, Vol. 36, pp. 443–457.

Aronsson H., Abrahamsson M., Spens, K., (2011),"Developing lean and agile health care supply chains", Supply Chain Management: An International Journal, Vol. 16 Iss: 3 pp. 176 - 183.

Baltacioglu, T., Ada, E., Kaplan, M.D., Yurt, O., & Kaplan, Y.C. (2007). A new framework for service supply chains. *The Service Industries Journal,* 27(2), 105–124.

Beal A, Doty M, Hernandez S, Shea K, Davis K (2007). "Closing the divide: how medical homes promote equity in health care. Results from The Commonwealth Fund 2006 Health Care Quality Survey". New York: The Commonwealth Fund.

Berens M.J.: Nursing mistakes kill, injure thousands. Cost cutting exacts toll on patients, hospital staffs. *Chicago Tribune,* Sep. 10, 2000, p. 20

Brailsford SC (2008). System Dynamics: what's in it for health care simulation modelers? In: Mason SJ, Hill R, Monch L and Rose O (eds). *Proceedings of the 2008 Winter Simulation Conference*, Miami, FL, pp 1478–1483.

Brown, S., Bessant, J., (2003). The manufacturing strategy-capabilities links in mass customization and agile manufacturing—an exploratory study. International Journal of Operations and Production Management 23 (7), 707–730.

Christopher, M. (2000), "The agile supply chain", Industrial Marketing Management, Vol. 29 No. 29, pp. 37-44.

Christopher, M., & Peck, H. (2004). Building the resilient supply chain. *International Journal of Logistics Management*, 15(2), 1-14.

Christopher, M., & Towill, D. R. (2002). Developing market specific supply chain strategies. *International Journal of Logistics Management, The*, *13*(1), 1-14.

Christopher, M., Towill, D., (2001), "An integrated model for the design of agile supply chains", *International Journal of Physical Distribution & Logistics Management* 31 (4), 235–246.

Dangerfield BC (1999). System dynamics applications to European health care issues. *J Opl Res Soc* 50: 345–353.

DeVor, R., Graves, R. and Mills, J. (1997), "Agile manufacturing research: accomplishments and opportunities", IIE Transactions, Vol. 29 No. 10, pp. 813-23.

Disney, S.M. and Towill, D.R. (2003), "On the bullwhip and inventory variance produced by an ordering policy", *Omega*, Vol. 31 No. 3, pp. 157-67.

Ellram L, Tate W and Billington C. (2004) Understanding and managing the services supply chain, *Journal of Supply Chain Management*: A Global Review of Purchasing and Supply, 40(4), pp.17–32.

Forrester JW (1961). Industrial Dynamics. MIT Press: Cambridge, MA.

Forrester JW (1958). Industrial dynamics: a major breakthrough for decision makers. *Harvard Business Review* 36(4): 37–66.

Goldman, Steven L.; Nagel, Roger N.; Preiss, Kenneth, (1995), Agile Competitors and Virtual Organizations, Van Nostrand Reinhold, New York, New York.

Gunasekaran, A. (1998), "Agile manufacturing: enablers and an implementation framework", *International Journal of Production Research*, Vol. 36 No. 5, pp. 1223-47.

Harrison, A., Christopher, M. and Van Hoek, R. (1999), "Creating the agile supply chain", School of Management Working Paper, Cranfield University, Cranfield.

Health Information Technology for Economic and Clinical Health (HITECH) Act, Title XIII of Division A and Title IV of Division B of the American Recovery and Reinvestment Act of 2009 (ARRA), Pub. L. No. 111-5, 123 Stat. 226 (Feb. 17, 2009).

Hewitt, F. (1999), "Supply or demand? Chains or pipelines? Co-ordination or control?", Proceedings from International Symposium in the Information Age, Florence, pp. 785-90.

Holweg, M. (2007). The genealogy of lean production. *Journal of Operations Management*, 25(2), 420-437.

Iacocca Institute (1991), 21st Century Manufacturing Enterprise Strategy. An Industry-Led View, Vol. 1/2, Iacocca Institute, Bethlehem, PA.

Joosten T, Bongers I, and Janssen R. (2009). Application of lean thinking to health care: issues and observations. *International Journal for Quality in Health Care* 21 (5): 341-347.

Kane, R. L., Shamliyan, T. A., Mueller, C., Duval, S., & Wilt, T. J. (2007). The association of registered nurse staffing levels and patient outcomes: systematic review and meta-analysis. *Medical care*, *45*(12), 1195-1204.

Kleijnen JPC (2005). Supply chain simulation tools and techniques: A survey. Int J Simul & Process Model 1(1&2): 82–89.

Kovner CT, Gergen PJ: Nurse staffing levels and adverse events following surgery in U.S. hospitals. *Image* 1998; 30:315–321

Lane DC and Husemann E (2008). System dynamics mapping of acute patient flows. J Opl Res Soc 59: 213–224.

Lee HL, Padmanabhan V, Wang S. (1997). The bullwhip effect in supply chains. *Sloan Management Review* 38: 93–102.

Lee, H. (2004), "The Triple-A Supply Chain," *Harvard Business Review*, Oct.

Lee, Y. M., An, L., & Connors, D. (2009). Application of Feedback Control Method to Workforce Management in a Service Supply Chain. *Service Science*, *1*(2), 77-92.

Litvak E, Buerhaus P, Davidoff F, et al. (2005). "Managing Unnecessary Variability in Patietn Demand to Reduce Nursing Stress and improve Patient Safety," *Journal on Quality and Patient Safety*, 31:6 330-338.

Lyneis, J. M., & Ford, D. N. (2007). System dynamics applied to project management: a survey, assessment, and directions for future research. *System Dynamics Review*, 23(2-3), 157-189.

Narasimhan, R., Swink, M. and Kim, S. (2006), "Disentangling leanness and agility: an empirical investigation", Journal of Operations Management, Vol. 24 No. 5, pp. 440-57.

Needleman J, Buerhaus P, Mattke S, et al: Nurse-staffing levels and the quality of care in hospitals. *N Engl J Med* 2002; 346: 1715–1722

Prince, J., Kay, J., (2003). Combining lean and agile characteristics: creation of virtual groups by enhanced production flow analysis. International Journal of Production Economics 85, 305–318.

Pronovost PJ, Jenckes MW, Dorman T, et al: Organizational characteristics of intensive care units related to outcomes of abdominal aortic surgery. *JAMA* 1999; 281:1310–1317

Reid R, Coleman K, Johnson E et al. (May 2010). "The group health medical home at year two: cost savings, higher patient satisfaction, and less burnout for providers". *Health Affairs* 29 (5): 835–43.

Robertson RH, Hassan M. Staffing intensity, skill mix and mortality outcomes: the case of chronic obstructive lung disease. *Health Serv Manage Res*. 1999;12:258 –268.

Rogers A, Hwant, W, Scott L, Aiken L, Dinges D (2004), "The Working Hours of Hospital Staff Nurses and Patient Safety," *Health Affairs*, 23:4, 202-212.

Rosenthal, T., (2008), "The Medical Home: Growing Evidence to Support a New Approach to Primary Care", *Journal of the American Board of Family Medicine*, Vol. 21, No. 5, pp. 427-440 .

Saeed, K. (2009), "Can trend forecasting improve stability in supply chains? A response to Forrester's challenge in Appendix L of Industrial Dynamics." *System Dynamics Review*. 25(1): 63-78

Saeed, K. (2008), "Trend Forecasting for Stability in Supply Chains." *Journal of Business Research*. 61(11): 1113-1124

Sampson S & Froehle C. (2006). Foundations and implications of a proposed unified services theory. *Production and Operations Management*, 15(2), 329–343.

Sarkis, J. (2001), "Benchmarking for agility", Benchmarking: An International Journal, Vol. 8 No. 2, pp. 88-107.

Scott LD, Rogers AE, Hwang WT, et al (2006). "Effects of critical care nurses' work hours on vigilance and patients' safety." *Am J Crit Care* 2006; 15:30–37

Sethi SP, Thompson GL. 2000. *Optimal Control Theory: Applications to Management Science and Economics*. Kluwer: Boston, MA.

Sethuraman, K., & Tirupati, D. (2005). Evidence of bullwhip effect in healthcare sector: causes, consequences and cures. *International Journal of Services and Operations Management, 1*(4), 372-394.

Sharifi, H., Zhang, Z., (2001), "Agile manufacturing in practice: application of a methodology", International Journal of Operations and Production Management, Vol. 21 No. 5/6, pp 772–794.

Sterman JD (2000). Business Dynamics: Systems Thinking and Modeling for a ComplexWorld. MIT Sloan School of Management: Irwin/McGraw-Hill.

Sterman JD. 1989. Modeling management behavior: misperceptions of feedback in a dynamic decision making experiment. *Management Science* 35: 321–339.

Tako AA and Robinson S (2009). Comparing discrete-event simulation and system dynamics: Users' perceptions. J Opl Res Soc 60: 296–312.

Taylor K and Dangerfield B (2005). Modeling the feedback effects of reconfiguring health services. *J Opl Res Soc* 56: 659–675

Taylor K and Lane D (1998). Simulation applied to health services: opportunities for applying the system dynamics approach. *Journal of Health Services Research and Policy* 3: 226-232.

Ventana Systems, Inc. 2012. Vensim Version 6.0. 60 JacobGates Road, Harvard, MA 01451.

Vries, J., Huijsman, R. (2011), "Supply chain management in health services: an overview", *Supply Chain Management: An International Journal*, Vol. 16 Iss: 3 pp. 159 - 165

Walley, P. (2007). Managing variation through system redesign. *International Journal of Healthcare Technology and Management*, *8*(6), 589-602.

Warren, K., (2007). "Strategic Management Dynamics", John Wiley and Sons.

White, A.S., 1999, Management of Inventory Using Control Theory. *International Journal of Technology Management*, 17: 847-860.