

Lab Turnaround Time and Delayed Discharges: A Systems-Based Action Research Investigation

Timothy D. Quinn

System Dynamics Group, MIT Sloan School of Management
30 Wadsworth Street, Bldg E53, Suite 360, Cambridge, MA 02142
tdquinn@mit.edu

Jenny W. Rudolph

Health Services Department, Boston University School of Public Health
715 Albany Street, Boston, MA 02118
jrudolph@bu.edu

David G. Fairchild, MD

General Medicine, Tufts-New England Medical Center
750 Washington Street, Boston, MA 02111
dfairchild@tufts-nemc.org

Abstract

This paper reports an action research study in which we applied Edgar Schein's process consultation approach to a cross-functional problem in a large academic teaching hospital. The project task force was charged with investigating a hypothesized effect of poor lab turnaround time on the risk of probable discharges being postponed until the following day, thereby increasing average length of stay and associated hospital operating costs. The tools we used at different stages of our process included group facilitation, interviews, process flowcharts, systems thinking with causal loop diagrams, and what-if analysis with a system dynamics simulation model. Through facilitation of the task force's work, we were able to reorient each constituent group's perspective from a parochial to a systemic view, greatly improving the task force's functioning and chances for successful sustainable improvement.

Introduction

This study was conducted at a large, urban, academic teaching hospital. We limited the scope of our investigation to the medicine and surgical inpatient units (excluding pediatrics and OB-GYN), as these comprised the majority of the hospital's inpatient volume. The hospital's combined inpatient census for these units had remained mostly stable over the past year, between 150 and 180 patients on any given day.

The timing of our study was fortuitous. Complaints about the turnaround time for laboratory tests and a hypothesized effect of this slow turnaround time on delayed discharges had prompted the hospital's administration to create a task force to investigate the problem and provide recommendations for improvement. The task force had not yet met for the first time, nor was its membership yet completely determined, when the hospital accepted our proposal to participate as process consultants (see Schein 1999).

Typical for academic teaching hospitals, many different groups must coordinate their work processes to effectively care for patients. These groups are the nurses, the ancillary services staff, the physicians, and the residents. The work content for each of these groups is incidental to this analysis, but it is important to note that in order to minimize the complexity of the healthcare delivery process, the work processes of each group have evolved to minimize their interdependence (of course, they are still highly interdependent). In this way, communication and task hand-offs become as efficient as possible, given the exigencies of the work environment. Also, the financial management of the hospital is removed from the clinical work being performed.

The implication of the complexity-mitigating work-process design for these groups is that each has a locally-bounded view of hospital operations. Like the proverbial blind men all touching different parts of an elephant, each group has its perceptions and interpretations of different slices of the organization's reality. When problems arise, as in the case of perceived slow lab turnaround times, each group interprets the problem relative to the cues and constraints present in its locality. For this study, we began by collecting these diverse perspectives to characterize the *symptoms* of the underlying problem.

Problem Perspective: Residents

To make diagnoses and clinical management decisions, the residents rely heavily on information from tests and studies, primarily from laboratory tests performed on blood samples drawn by phlebotomy early each morning. The residents were frustrated with a slow morning lab turnaround time and the fault-intolerant nature of the hospital's processes for patient lab tests. Delayed morning test results contributed to waste, inefficiency, and rework in the residents' workflow, and the effect of process breakdowns would 'cascade' through the day, leading to more work later and higher risks to patients. Importantly, the residents feel that lack of timely lab test results contributes to delays in discharging patients. Already working long hours to care for a high volume of patients, the residents resented the fact that they had to make up the shortfall in phlebotomy staffing levels, by drawing blood samples during those periods when there was little or no phlebotomy coverage.

Problem Perspective: Nurses

As the front-line care-providers for patients, the nurses (and other members of the nursing staff) are unhappy with any impediments to properly attending to each patient's changing status and needs. They were dissatisfied with the responsiveness of the phlebotomists for scheduled blood draws, phlebotomy's lack of good communication and coordination with the nursing staff, and the unavailability of resources to draw *stat*¹ lab test samples. Some nurses would draw *stat* labs when their unit's residents (the back-up resource) were unavailable, even though this task was not part of their job function. The nurses also perceive that lack of timely lab test results contribute to delays in discharging patients.

Problem Perspective: Laboratory & Phlebotomy Managers

The laboratory managers feel that their department is understaffed. First, the employees working as phlebotomists also perform the central receiving functions in the laboratory.

¹ *Stat* is short for the Latin word *statim*, which means "at once" or "immediately".

Second, the demands for phlebotomy coverage (types of blood samples drawn and when) throughout the hospital has been steadily increasing without a corresponding increase in staff. Third, units dissatisfied with the current level of service sometimes succeed in ‘stealing’ a phlebotomy FTE to be dedicated to its own needs. As low-paid, low-status, entry-level positions, absenteeism and turnover of phlebotomists are high and filling vacant positions is very difficult in a tight labor market. The phlebotomists feel under constant pressure to work faster, resulting in a focus to rush through their draw lists without attention paid to needs specific to certain units or patients.

Problem Perspective: Hospital Administration

The management team of the hospital is struggling to improve cash flow in the face of several years of operating expenses exceeding revenues by millions of dollars. The hospital has remained solvent only by relying on the cushion provided by its cash reserves—hardly a sustainable position. The requisite belt-tightening under these circumstances has comprised deferring capital investments and trimming operational costs, including staff where it was deemed possible. These hard choices have engendered an environment in which everyone is expected to do the best they can with the prevailing allocation of resources. However, shouldering the burden of budgetary constraints has inevitably created tension among the hospital’s constituent groups: nurses, physicians, residents, and the ancillary services staff. Complaints about insufficient staffing levels by all these groups are common.

Group Building through Action Research

Our intervention in the first task force meeting was comprised of setting an agenda and attempting to create norms for the task force members’ interactions, especially early in the group-building phase (Schein 1999). Specifically, we had to arrest each groups’ tendency to take the problem definition for granted and to jump straight into brainstorming possible ‘solutions’. We made a case for a measured approach that afforded enough time for a systems-based analysis of the problem. We proposed—and the task force accepted—that we would interview each group separately, map out a process flowchart of their respective work processes (as they pertain or are influenced by phlebotomy), and report back with an organizational ‘landscape’ on which we could collectively survey next steps.

As action researchers, we were careful to continually evaluate our effect on the organization under study, particularly our influence on the task force’s group dynamics. We recognized the need to gain the trust of all represented groups, to be considered partners in the task force’s activities instead of observers or meddlers from the ‘outside’. Most important, we did not want to create the impression in the minds of any group that we were biased toward or more sympathetic to the worldview of a particular group.

By interviewing each group separately, we created an intimate atmosphere in which we asked members to teach us how they perform their work and to air their frustrations about perceived impediments to efficient workflow. In this manner, we demonstrated that we had not anchored on any presumptions about each group or their work processes and that we were willing to take their worldviews seriously. We were also correct in our belief

that eliciting frustrations uncovered symptoms of deeper structural problems in the organization. This approach echoes Elton Mayo's observation that led to the unraveling of the productivity paradox in the classic Western Electric studies (which uncovered the now-famous Hawthorne effect²):

Workers wished to talk, and to talk freely under the seal of professional confidence (which was never abused) to someone who seemed representative of the company or who seemed, by his [sic] very attitude, to carry authority. The experience itself was unusual; there are few people in this world who have had the experience of finding someone intelligent, attentive and eager to listen without interruption to all that he or she has to say. (Mayo 1949)

Attribution Errors

Common to all groups we interviewed was an understandable preoccupation with the constraints and exigencies of one's own work process, to the exclusion of a fair appreciation for the work processes of others, even when groups are highly interdependent. Particular to this case, the residents and nurses expected phlebotomists to be efficient, flexible to contingent demands, and highly responsive to ad hoc requests. Failure to meet these expectations caused both groups to characterize the phlebotomists as "lazy" and "unresponsive".

We were aware of the documented social psychological phenomenon called the *fundamental attribution error*: the tendency of organizational actors to attribute undesirable outcomes to the perceived character flaws of people rather than to the processes engendered by the organization's structure (Ross 1977). Given this common behavioral tendency, we suspected that the residents' and nurses' beliefs about the phlebotomists were based on the degree to which their expectations for phlebotomy were not met, rather than based on concrete knowledge of the phlebotomists' productivity.

We realized that such attributions impeded effective collaboration on problem analysis and solution formulation. Indeed, the first meeting of our task force began with each group explicitly delineating the boundaries of their institutionalized roles and preparing for what all assumed to be inevitable conflict over assigning blame for organizational dysfunction. Our agenda was to avert such political maneuvering by first addressing the tacit assumptions underlying each group's beliefs.

A critical part of framing the problem and setting the climate for subsequent group interaction was to deconstruct the attribution of "undisciplined phlebotomists" and to replace it with a broader understanding of what the phlebotomist's world looks like. Several weeks prior, the laboratory supervisor had collected data on phlebotomy for two weeks, noting the total number of patient-sample orders and the number of phlebotomists (FTEs) working on each scheduled draw. The data had been graphed on a bar chart, but

² The Hawthorne effect describes an increase in workers' productivity in both the treatment *and* control groups in an experiment in which working conditions were varied. Productivity increased, in all cases and independent from any incremental changes to conditions, simply because the workers in both groups were being observed.

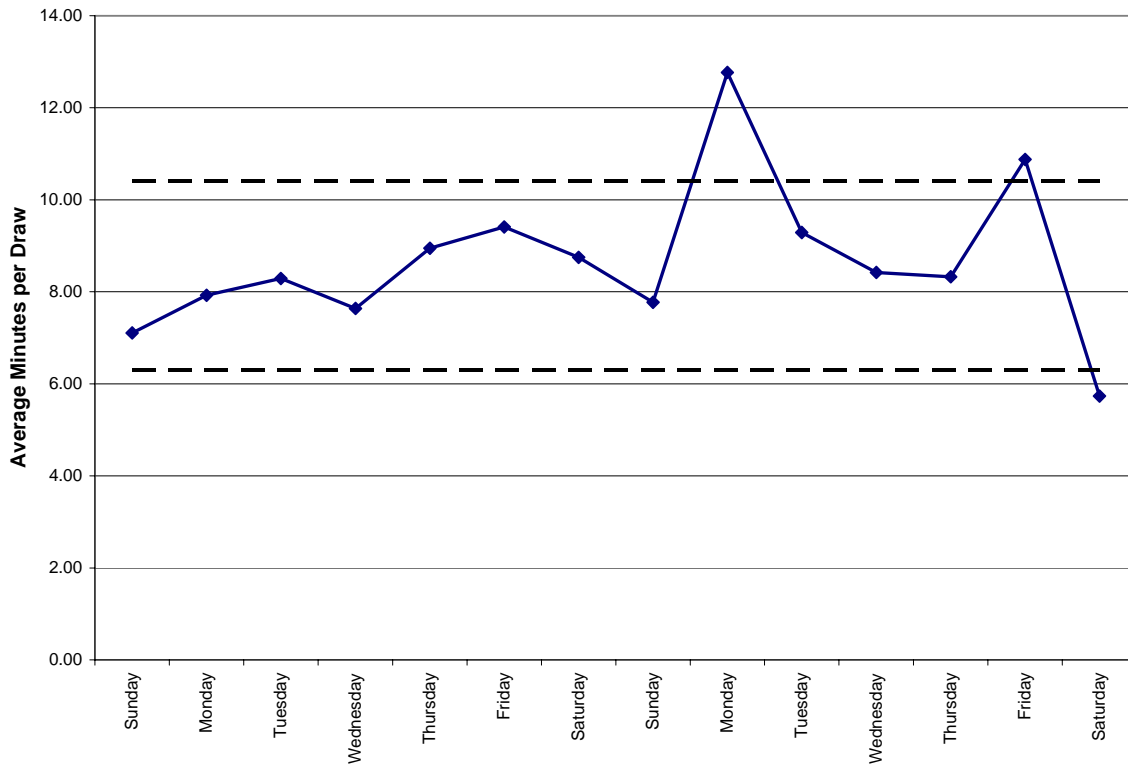
no attempt had been made to analyze them. We obtained the data and focused on the 7:00 A.M. scheduled draw time, because those orders constituted the majority of the daily phlebotomy. The data for this draw time are shown in Table 1 below.

Day of Week	Orders for Patient-Samples	FTEs	Average Samples per FTE
Sunday	152	6	25.3
Monday	159	7	22.7
Tuesday	152	7	21.7
Wednesday	165	7	23.6
Thursday	161	8	20.1
Friday	153	8	19.1
Saturday	144	7	20.6
Sunday	139	6	23.2
Monday	141	10	14.1
Tuesday	155	8	19.4
Wednesday	171	8	21.4
Thursday	173	8	21.6
Friday	182	11	16.5
Saturday	157	5	31.4

Table 1. Phlebotomy Demand and Staffing Levels for 7:00 A.M. Scheduled Draw

Orders were counted by patient-samples, because multiple laboratory tests for a given patient typically could be processed using the blood collected from a single sample. Thus, the number of patient-sample orders approximates the hospital’s inpatient census for the units considered in this analysis. The number of phlebotomists working the scheduled morning draw fluctuates due to absenteeism and attempts at meeting demand by varying daily resource allocations. From these data, we first calculated the average number of samples drawn per phlebotomy FTE per day (see the last column of Table 1).

Our interviews with the laboratory personnel revealed that on a typical day the phlebotomists began the scheduled morning draw at 7:00 A.M. and completed it by 9:30 or 10:00 A.M. at the latest. Unfortunately, it was not possible for the laboratory supervisor to track the actual start and end times of each phlebotomist on each day. Therefore, we had to make some assumptions to further our analysis. To be conservative, we assumed that each phlebotomist spent a fixed 180 minutes working the scheduled morning draw. By dividing the average samples drawn per phlebotomy FTE into this task duration, we obtained a very rough approximation of phlebotomy productivity (see Figure 1).



Note: The dashed lines indicate the lower and upper control limits of 6.3 and 10.4 minutes, respectively.

Figure 1. Approximated Minutes per Draw

Despite the assumptions made, the phlebotomist’s productivity displays remarkable consistency. The average is 8.7 minutes per draw, with the limits of statistical control³ for this process found to be from 4.1 to 13.2 minutes per draw. However, it appears that certain days are affected more by our assumption of a fixed duration for the scheduled morning draw. By comparing the extreme cases (the outliers) in Figure 1 with the data in Table 1, it is clear that staffing was unusually high on the second Monday, on the second Friday (due to a high census), and unusually low on the second Saturday. Removing these three cases from our calculations yielded an average of 8.4 minutes per draw, with the limits of statistical control from 6.3 to 10.4 minutes per draw (dashed lines shown in Figure 1).

The effect of this analysis on the task force was dramatic. Not only could everyone see the consistency with which the phlebotomists carried out the scheduled morning draw, but the total time required to complete the draw and the average time spent per patient were made salient and ‘accessible’. Only after we presented these results did the laboratory director deliver the *coup de grâce*: the phlebotomists’ average time per draw (8.4 minutes) was significantly lower than the benchmark of 10 minutes per draw recommended by industry consultants. (Given the statistical control of the phlebotomists’ work process, it is likely that their performance will stay below this

³ See Appendix A for an explanation of statistical control.

benchmark, excepting any unusual circumstances or fundamental changes to the work process. This fact can be seen from the upper statistical control limit of 10.4 minutes per draw, just barely over the benchmark of 10 minutes per draw.)

Was phlebotomy in this hospital simply more efficient than the industry average, as reported by the consultants? Clearly not, as evidenced by the complaints about phlebotomy from the nurses and residents. Confronted with the data, these two groups had to revise their conceptions of the problem locus, increasing their willingness to collaborate with the laboratory employees. These data also prompted a reconceptualization by management. First, a plausible story now existed for the basis of the nurses' and residents' complaints: phlebotomists were unresponsive because they were too busy rushing to complete the scheduled morning draw in a 'reasonable' amount of time. Second, perhaps the approved staffing levels for phlebotomy, thought adequate for recent patient volumes and an acceptable budgetary compromise, were not 'adequate' after all. Completing work within a certain time frame as a criterion for adequacy needed revision to include the potentially detrimental side-effects downstream in the patient-care process.

Diffusing the politics of blame and orienting the task force toward the analysis of constraints inherent in the structure of the hospital's processes comprised a significant step toward sustainable organizational improvement.

Problem Framing with Systems Thinking

The average length of stay in this hospital's recent history was high enough above the industry average, controlling for acuity, that hospital administration were concerned about the financial ramifications of further deterioration in this metric due to delayed discharges. A discharge can be delayed when hospital processes impede either the timeliness of clinical decision-making required to 'clear' a patient's discharge or the logistics of the discharge process itself. Delayed discharges are costly for two reasons. First, a patient spending an extra night in the hospital can erode the hospital's margin on that case if reimbursement is on a per-DRG (i.e. per-case adjusted for acuity), instead of a per-diem, basis. Second, demand for inpatient beds in the hospital is high enough that a delayed discharge means the hospital incurs an opportunity cost equal to the foregone margin from a new admission.

Hospital administration prompted the formation of the task force because the residents claimed that the timeliness of lab test results—as a result of lab turnaround time—impeded the timeliness of their clinical decision-making to such an extent that discharges were sometimes delayed. The director of the laboratory denied that this was possible, but hospital administration charged the task force with investigating this hypothesized causal connection. In the first task force meeting, the following causal chain was described verbally by the physicians: a slow average lab turnaround time will constrain and sometimes impede the rate at which clinical decisions can be made (discharge decisions are of particular concern); a slower rate of clinical decision-making increases the risk of a delayed discharge, which, if they happen often enough, will cause the hospital's average length of stay to deteriorate (the measure will increase). The financial link between

average length of stay and the hospital’s profitability was taken for granted by hospital administration.

Consistent with our systems-based approach to the analysis of this problem, we sought to diagram a complete causal loop (Sterman 2000) to explain why the problem of slow lab turnaround times—if actually present and perceived—wasn’t corrected. With the physicians’ and residents’ perspective as a starting point, we solicited the views of the other groups. The nurses confirmed the residents’ story that lab turnaround times can delay decision-making and therefore increase the risk of delayed discharges. The laboratory supervisor explained that, consistent with the phlebotomist productivity analysis, they were doing the best they could with the resources at their disposal, especially the number of FTEs available to be assigned to phlebotomy. Hospital administrators confirmed that hiring had essentially been ‘frozen’ in the face of budgetary pressures.

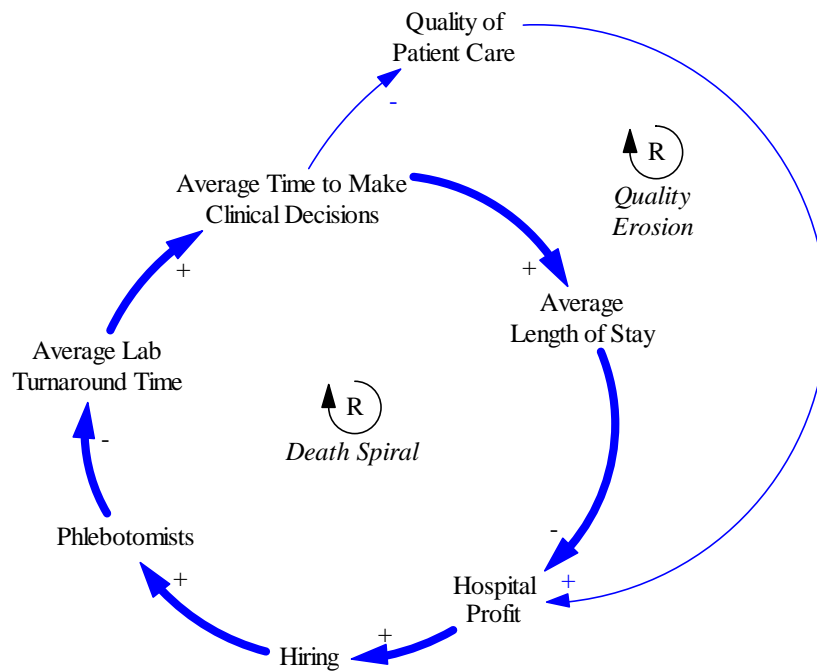


Figure 2. Causal Loop Diagram for Delayed-Discharges Hypothesis

Assembling all groups’ causal chains yields a *feedback loop*⁴, as shown in Figure 2, congruent with our everyday notions of a virtuous or vicious cycle (depending on whether the state of affairs is getting progressively better or worse). Using this technique of systems thinking, we constructed an overarching hypothesis from the perspectives contributed by the separate groups. Examining the diagram, we see that the hospital might be caught in a so-called “death spiral”: low hospital profitability prevents hiring

⁴ See Appendix B for more information on the symbols used in the causal loop diagram and the definition of a feedback loop.

phlebotomists up to a level that would provide for fast lab turnaround times, thereby encumbering the rate of clinical decision-making and increasing the incidence of delayed discharges. The costs, including opportunity costs, associated with delayed discharges further suppress the hospital's operating margin, trapping the hospital in a cycle of low performance and low profitability.

That each group represented on the task force had a piece to contribute to this causal loop hypothesis highlights the importance of accessing everyone's knowledge about their own work process to build a conception of the whole hospital system and how these pieces interact. Second, that each group contributed a small piece to the larger puzzle but didn't have visibility into the 'worlds' of the other groups offers an explanation for why the problem wasn't identified and solved in the past. The hospital's necessary division of labor for patient care creates functional silos across which there wasn't a framework for thinking about or addressing cross-functional problems.

We were lucky that the rough sketch of the causal loop diagram presented in Figure 2 emerged from our work with the task force relatively early in the project. However, it is important to point out that the final result is a distillation of many detailed causal loop diagrams, each built from interviews with the various groups and from analyses of their respective work-process flowcharts. These detailed diagrams described various interdependencies and work characteristics, and served to justify each link in the overall causal hypothesis of Figure 2. Some examples are as follows:

- Delayed lab test results prompt the residents and physicians to formulate conditional patient care plans, resolved when the test results become available. Such contingency planning lowers overall productivity.
- Delays in clinical decision-making increase the amount of time when patients' statuses are uncertain. Such uncertainty increases the risk of adverse events and malpractice exposure.
- Delays in clinical decision-making and hospital processes, if recognized, may lead to lower patient satisfaction.
- The residents must make up the shortfall in any phlebotomy coverage, such as blood cultures, *stat* lab tests, and phlebotomy services overnight or during weekends.
- With staffing levels 'frozen' in the laboratory, any unit (such as oncology) that can successfully argue for a dedicated phlebotomist results in a net decrease in laboratory personnel available for phlebotomy coverage for the rest of the hospital.
- High turnover among phlebotomists may be due to burnout from high work pressure, or from low morale and low job satisfaction due to a lack of recognition (or even negative attributions) from other groups.

Problem Analysis with System Dynamics

Our next steps involved a careful evaluation of the validity of the delayed-discharges hypothesis with respect to the hospital's actual operations. For this stage of the project, we used the analysis tools of system dynamics (Forrester 1999 [1961]; Sterman 2000) to

explore how the structure of the hospital's operations—including the formally defined jobs, process relationships, and “in-use” work practices—contributes to observed phenomena. In particular, we constructed a differential-equation computer simulation model of the phlebotomists' work, the laboratory's processing of blood samples, the ‘flow’ of patients from admission to discharge, and the residents' utilization of lab test results and clinical decision-making. We also incorporated groups' work schedule constraints into the computer model, such as phlebotomist shift changes and the residents' rounds with the attending physicians.

As with most computer simulation models of work processes, we made assumptions about the duration variability of each groups' tasks and then tested the reasonableness of these assumptions by comparing the model's behavior with the actual variability observed in the hospital's day-to-day operations. We calibrated the model to match reality as best we could with the model's level of aggregation. In some cases, the model's behavior exhibited significant departures from the hospital's actual operations. In these instances, we investigated the cause of the disparity in the model, which usually prompted us to clarify our understanding of some operational details with members of one or more of the groups. Then we revised the model structure to reflect these details and thereby achieve a better match between the model and actual workflow.

The computer simulation model was constructed and thoroughly tested in phases. In each phase, we strove to formulate as parsimonious a structure as possible that still accurately reflected the hospital's operations. The results of an early test of the model's simulation of the phlebotomists' morning blood draw, the laboratory's processing of the blood samples, and the residents' use of the test results is shown in Figure 3. These results are for a single, representative day, so the variability in the time required to draw and to process the blood samples, and the time required to use the lab test results, can be considered to “average out” over many days to yield the pattern displayed in Figure 3. The actual pattern for any given day would show task completion times earlier or later, depending on patient volumes and the effects of other workload factors.

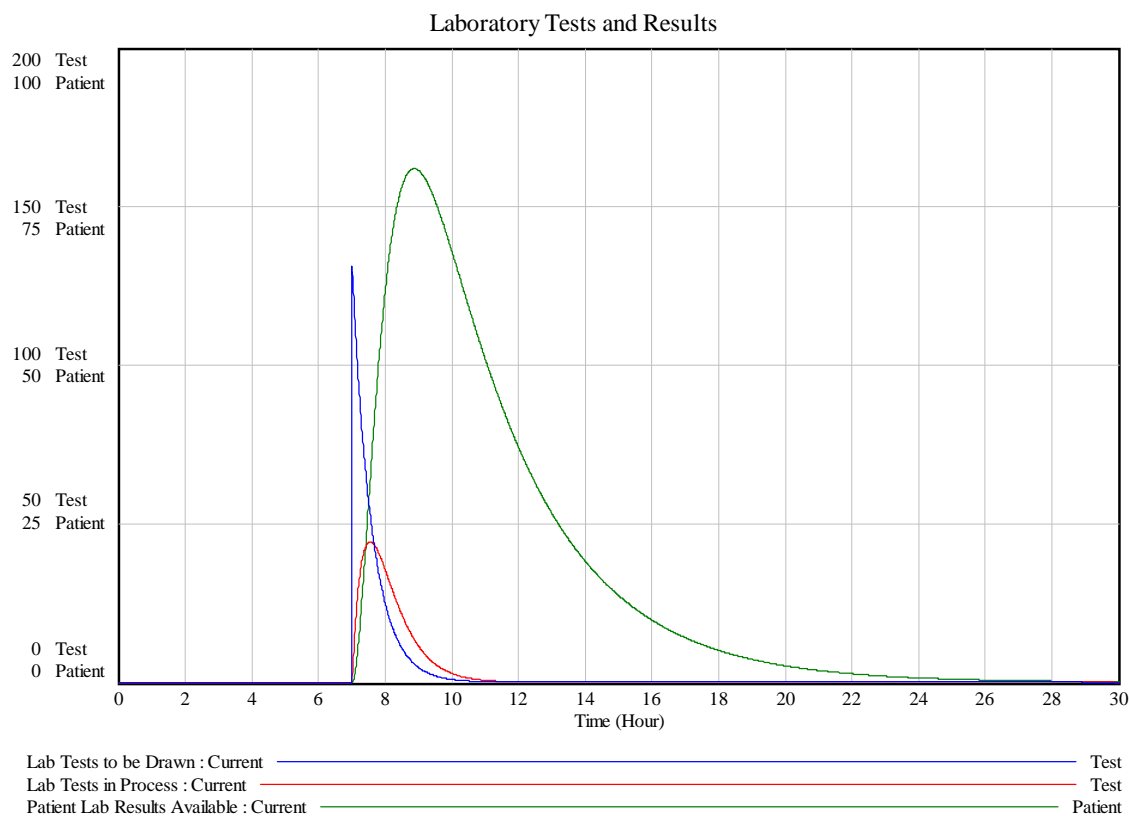


Figure 3. Simulation Results for Morning Blood Sample Collection, Processing, and Usage of Lab Test Results⁵

In the aggregate, Figure 3 provided us with an understanding of the ‘physics’ of the hospital’s lab-test processes: there is a limit to how quickly various groups can accomplish their work. The phlebotomists must make their rounds of the entire hospital, draw the blood samples, and transport the samples to the laboratory. The laboratory must receive the samples; document, prepare, and process them; and validate the results. The rate at which the residents use the lab test results to make clinical decisions about each of their patients is constrained by all demands on their time. Given this lower bound on the efficiency of the lab-test process and the picture of a ‘typical’ work day, Figure 3 demonstrates that delayed discharges, however rare, are bound to happen periodically with an unlucky confluence of factors.

It was useful for us to think of the probability of a discharge being delayed according to James Reason’s Swiss Cheese model of system failure (Reason 1997). Each opportunity for a breakdown in a work process is represented by a slice of Swiss cheese. Even if a breakdown occurs—for example, if a patient’s morning lab test order was requisitioned incorrectly—the error is usually corrected further ‘downstream’; that is, a nurse or

⁵ In the figure, blood samples are measured in units of “tests”, because we assumed that each patient would have one sample drawn per scheduled phlebotomy draw and that all tests ordered for each patient could be lumped together as a single “test”. Simply for convenience in the model, results for this lumped “test” were tracked by the patients to whom the results pertain.

resident will discover the error and take action to correct it as soon as possible. In this way, the hole (possible failure or latent error) in the order requisition slice didn't line up with a hole in another slice (the resident or nurse is too busy to catch the error) later in the process. A complete system breakdown occurs when the holes in all the cheese slices line up. Figure 4 presents another example: a patient can acquire Acute Respiratory Distress Syndrome (ARDS) from the confluence of a number of preventable circumstances.

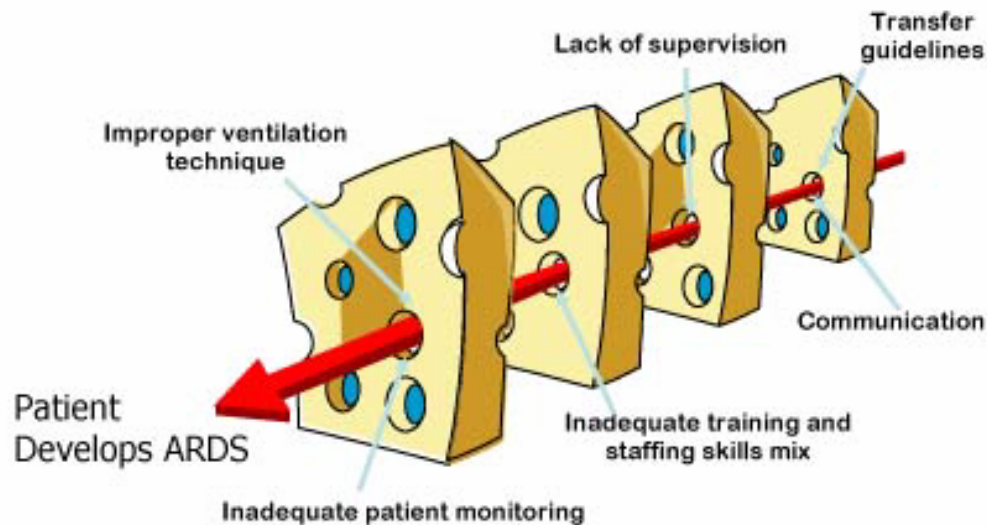


Figure 4. An Example of Reason's Swiss Cheese Model of System Failure⁶

We conducted several iterations of model elaboration, testing, and analysis of simulation results to investigate the effect of various factors on the expected number of annual delayed discharges. Surprisingly, reengineering the laboratory's processing of the blood samples to make test results available sooner—an "obvious" area of focus in early task force meetings—resulted in only a very minor reduction in the probability of a discharge being delayed. Factors that proved important included the phlebotomists' shift schedules and staffing levels, and the timing of residents' rounds with their attending physicians.

Consistent with the perspective of the laboratory's managers, simply hiring more phlebotomists would remove this staffing level variable from the delayed discharge risk equation. We examined this policy in detail by incorporating its effects into our causal loop diagram, as shown in Figure 5. Given the hospital administration's authorized number of positions for phlebotomy, any discrepancy between this authorized number and the number of phlebotomists on the payroll—due to attrition and difficulty in attracting qualified candidates—is resolved by the hospital's hiring process. This process comprises the *balancing* feedback loop⁷ labeled "Workforce Adjustment". However, the hospital's financial position can enter into this loop in two ways. First, low profitability can result in a hiring 'freeze', even if authorized positions are vacant. Second, the hospital's administration seeks the economically efficient number of phlebotomists—

⁶ Figure 4 was taken from Holzmueller et al. 2004.

⁷ For a discussion of balancing feedback, see Appendix B.

where the marginal operational benefit of an additional phlebotomist equals the marginal cost of that extra employee—by attending to signals in the hospital’s accounting.

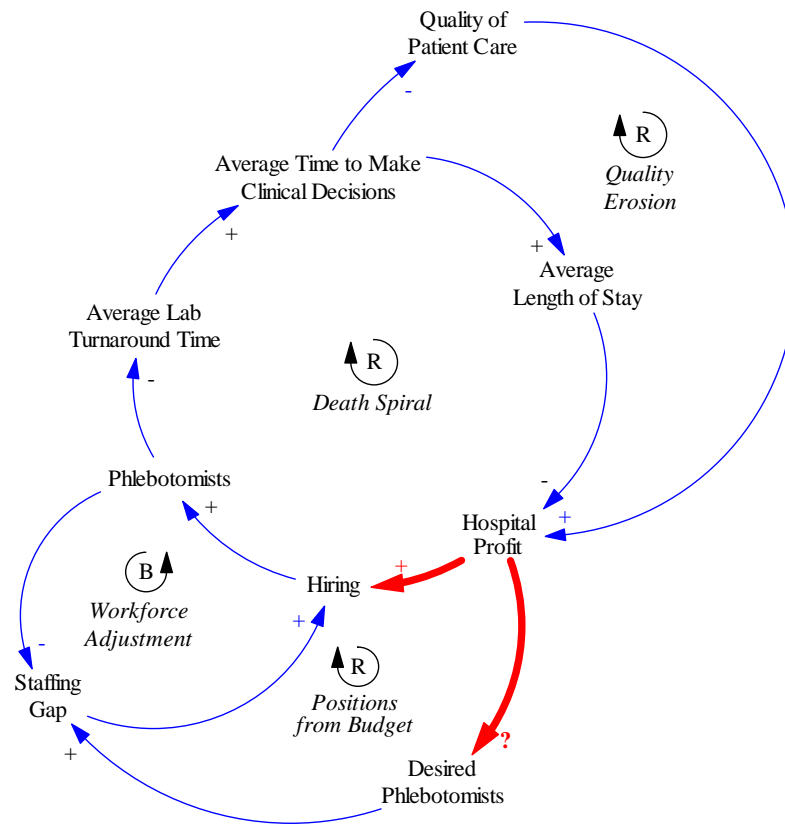


Figure 5. Causal Loop Diagram for the Effect of Management Policies on Phlebotomist Staffing

The problem with neoclassical economic theory in this case, however, is twofold. First, it is difficult, if not impossible, to determine the “marginal operational benefit” of an additional phlebotomist in financial terms. There are simply too many people and processes intervening between the work of the phlebotomists and the hospital’s revenue stream. Second, the operational outcome of a particular staffing level approved by the hospital’s administration may not be fully realized, both actually and cognitively, until quite some time after the decision is implemented. These delays in feedback information about organizational policies are extremely common in complex systems. As John Sterman observed,

Within a causal field, people use various cues to causality, including temporal and spatial proximity of cause and effect, temporal precedence of causes, covariation, and similarity of cause and effect. These heuristics lead to difficulty in complex systems, where cause and effect are often distant in time and space, actions have multiple effects, and the delayed and distant consequences are often different from and less salient than proximate effects—or simply unknown. (Sterman 1994, p. 308)

During our project, the laboratory manager could still hire to fill vacant phlebotomy positions, although he reported having difficulty attracting enough candidates. However, our interviews revealed that the hospital's administration had not adjusted the authorized number of laboratory and phlebotomy personnel for at least the last two years. The administration justified inaction in the face of laboratory requests for more staff on financial grounds. This policy suggests that cost-reduction strategies usually entailed cutting staff to 'minimally adequate' levels. Such an approach is consistent with the theory of bounded rationality (March and Simon 1958): the hospital's administration sets policy based on salient and proximate cues.

Our analysis of the causal loop diagram in Figure 5 presented us with an interesting alternative. Instead of reducing staffing levels to bolster margins when profitability is low, perhaps it is necessary for the hospital to invest *more* resources in certain areas in order to escape from a vicious cycle of underinvestment that reinforces poor performance. This insight prompted the hospital's administration to reconsider their assumptions and motivated our next round of scenario analysis using the simulation model.

We expected that a large number of phlebotomists would greatly reduce the risk of delayed discharges. We were surprised when the simulation model showed this not to be the case. It was true that an incremental increase did lower the overall risk, but adding more phlebotomists above a certain threshold exhibited drastically diminishing returns. Similarly, small *decreases* in phlebotomy staffing led to disproportionately large increases in delayed discharge risk. We discovered the reason for this asymmetric effect was the way in which the hospital *as a system* responded to such a change.

From our interviews with the residents, we learned that they are the resources responsible for making up any shortfall in phlebotomy coverage. These shortfalls include draws of blood samples for tests needed *stat*, repeat attempts for unsuccessful draws by a phlebotomist, making up shortfalls in phlebotomy staffing due to absenteeism, or draws outside of scheduled draw times (e.g. overnight or on weekends). The residents making up the phlebotomy shortfall is another example of a balancing feedback process, shown in Figure 6 as the "Make Up Shortfall" loop. Much like squeezing a balloon, constricting the maximum workflow in a particular area, such as phlebotomy, only increases the amount of work to be done elsewhere by another group. The nurses we interviewed reported that it was hospital policy that drawing blood samples was not part of their job function, although many nurses would draw samples "as a favor" if the residents were busy attending to other patients' more urgent care problems and couldn't draw blood in a timely manner.

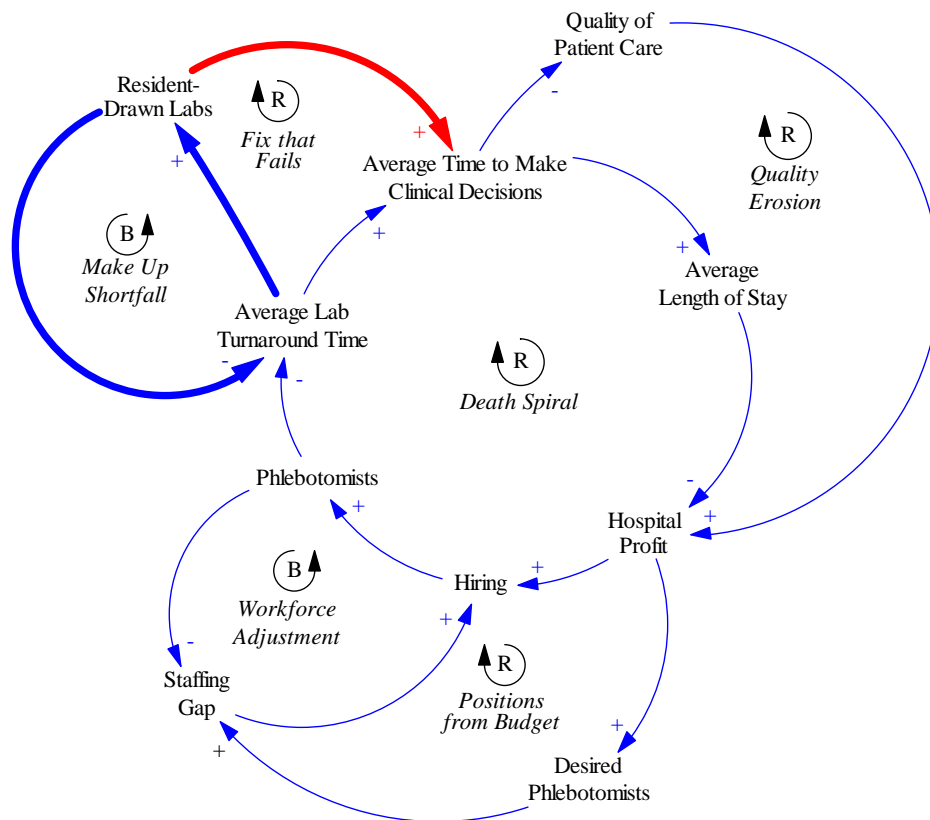


Figure 6. Causal Loop Diagram for Resident Workload

Therefore, the hospital’s system response to a bottleneck in the phlebotomists’ work capacity was to shift the burden to the residents, with significant ramifications. In addition to other potential problems caused by hospital work processes, increasing the residents’ workload causes the average time the residents take to make clinical decisions to deteriorate. In system dynamics, this effect occurs often enough in analyses of organizations that it is considered a “system archetype” (Senge 1990), usually called “Fixes that Fail” (the critical link is denoted by the red arrow in Figure 6).

Because the task force was formally charged with examining the “lab turnaround time problem”, our simulation model lumped all organizational factors that affect the probability of a discharge being delayed—except the availability of discharge lab test results—into a single variable for the rate of clinical decision-making. Of course, this aggregation was a drastic simplification of many detailed work processes of the residents and other hospital employees, but this modeling choice allowed us to examine the effect of changes to phlebotomy separated from all other aspects of patient care. If we had been asked to expand our project’s scope to include an investigation of these other processes, then our model’s level of detail would have been insufficient.

We used the simulation model to analyze the dynamics of the residents’ work process under various conditions. These conditions included changes in phlebotomy staffing levels; changes to the residents’ and phlebotomists work schedules, including the timing

of the residents' rounds with their attending physicians; and hypothetical improvements to other hospital processes that would increase the residents' efficiency.

As illuminated by Eliyahu Goldratt's theory of constraints (Goldratt 1984), what the simulation results showed was that eliminating a throughput bottleneck at the phlebotomists' morning draw would only shift the bottleneck somewhere else in the system—in this case, to the residents' rate of clinical decision-making. However, the overall risk of delayed discharges wasn't attenuated substantially because the average time required to make clinical decisions is still part of the "Death Spiral" causal loop. Therefore, faster lab turnaround time is a necessary, *but not sufficient*, condition for better hospital performance. This insight prompted a major reorientation in the task force's thinking about problems of a *systemic* nature, including this one. Even though divisions of labor prove effective for coordinating complex work such as patient care, a reductionist orientation can be harmful when organizational changes made in one silo ramify to the detriment of other groups elsewhere in the organization and to the organization as a whole.

Given the benefit of hindsight, it is tempting to ask the question of whether we couldn't have simply arrived at these insights from more carefully interviewing the residents, who, as Dr. Paul Batalden⁸ has observed, "live in the fault lines of health care systems and give voice to what life is like there." We believe this possibility to be highly unlikely. Despite the fact that the residents collectively constitute the nexus of most of the coordination required for patient care, they consistently reported how they had "no time to think about the big picture because we're too busy running around trying to get all the work done." For example, the residents we interviewed had not considered the inherent risk trade-off between waiting for an overdue lab test result before taking action or acting sooner without the benefit of the most up-to-date information on a patient's status.

The implications for the task force were clear. Certainly, they had to address the problem regarding the maximum rate at which all the morning blood samples could be drawn. But more importantly for the hospital's overall performance, they recognized that limiting the scope of their interventions to the laboratory area alone would not achieve the sustainable improvement sought by the hospital's administration. After addressing the morning blood draw bottleneck, the next bottleneck—the efficiency with which the residents go about their duties and the hospital's systems that support them—will cry out for attention.

Conclusion

We were able to successfully suspend the task force's initial tendency to assume they understood what the problem really was and jump straight to brainstorming 'solutions'. In this way, we also averted any bias toward 'firefighting'; that is, addressing only the symptoms of the problem in the short term and ignoring the long-term consequences of leaving the root cause unaddressed. This success also highlights the point that the most important aspects of the organization's work processes are often hard to measure: the

⁸ Paul Batalden, MD is the program director of the Dartmouth-Hitchcock Leadership Preventive Medicine Residency Program at Dartmouth Medical School and the Senior Vice President of Health Professional Development for the Institute for Healthcare Improvement in Boston, MA.

hospital tracked average length of stay assiduously, but had no useful way to assess resident workload or productivity rates.

We were able to identify the ‘high-leverage’ points for improvement; in particular, that lab test information available earlier must be acted on earlier to be of benefit, both financially and for the quality of patient care. Process improvements cannot be targeted at a single area, but must be made with the entire patient care process in mind. Such a holistic view is also important in light of the fact that work processes are also subject to social evolution; they grow up around constraints that are perceived to be fixed (at least in the near future). In the case of this hospital, the timing of the residents’ rounds with their attending physicians affected many other work processes, yet no one suspected that this relatively minor aspect of their work schedules had major ramifications for the rest of the workday.

This project demonstrated to the task force the value in taking a systems-based approach to solving organization-wide problems. When asked about our participation as process consultants to structure the problem-solving approach, a physician member of the task force remarked that our tools of systems thinking and system dynamics produced insights that would not have emerged from discussions alone. This observation is important in light of the fact that people tend to interpret problems based on their worldview (even upper management, charged with oversight of the whole organization, tend to think only in financial terms) and usually have not developed a framework for systemic thinking and problem-solving.

Appendix A

Statistical control is a means to characterize the predictability of a process. If a process displays statistical control, then it is reasonable to expect that future measurements of the process will fall within limits set three estimated standard deviations away from the process mean.

The test for statistical control does *not* assume any particular distribution of the data, at the expense of being able to compute a probability for Type I error. Tchebyshev's inequality ensures that *every* distribution will have a three-standard-deviation coverage of *at least* 88.9%, but for distributions of process data in practice, this coverage is almost always above 96% (see Wheeler 2000). Therefore, the choice of three standard deviations above and below the process mean is an *economic* choice, selected to minimize costs associated with investigating the root cause of an outlier (Shewhart 1980 [1931]).

The diagnostic power of the test for statistical control requires a standard deviation estimator that is insensitive to outliers. For this reason, a within-group estimator is used. For data such as those analyzed in this study, the within-group estimator is based on the dispersion of the absolute-value differences between successive process measurements. The within-group standard deviation is then compared with the observed dispersion of (between-group) process measurements to determine statistical control. For details and an exceptional, friendly treatment, see Wheeler 1995.

Appendix B

Arrows indicate the direction of causality. Signs ('+' or '-') at arrowheads indicate the polarity of relationships: a '+' means that an increase in the independent variable, all else being equal, causes the dependent variable to increase (or a decrease causes a corresponding decrease); a '-' means that an increase in the independent variable, all else being equal, causes the dependent variable to decrease (or a decrease causes a corresponding increase).

Causal loops come in two varieties. A *reinforcing* feedback loop (usually labeled with an 'R'), as shown in Figure 2, corresponds to the notion of a virtuous or vicious cycle: increases in a variable anywhere in the loop get amplified as you move around the loop. A *balancing* feedback loop (usually labeled with a 'B') is one in which the behavior of the modeled system tends toward some goal. For example, the hiring process to bring the number of phlebotomists on staff up to the number approved by management is a balancing feedback process. For more information, see Sterman 2000.

References:

- Forrester, J. *Industrial Dynamics*. Waltham, MA: Pegasus Communications. 1999 [1961].
- Goldratt, E. *The Goal: A Process of Ongoing Improvement*. Great Barrington, MA: North River Press. 1984.
- Holzmueller, C., P. Pronovost, and R. Branson. How can we learn from incidents?
Critical Connections. 3(1), 2004.
- March, H. and H. Simon. *Organizations*. New York: John Wiley and Sons. 1958.
- Mayo, E. "Hawthorne and the Western Electric Company" in *The Social Problems of an Industrial Civilization*. Routledge. 1949.
- Schein, E. *Process Consultation Revisited: Building the Helping Relationship*. Reading, MA: Addison-Wesley. 1999.
- Senge, P. *The Fifth Discipline*. New York: Doubleday. 1990.
- Shewhart, W. *Economic Control of the Quality of Manufactured Product*. Milwaukee, WI: Quality Press. 1980 [1931].
- Sterman, J. *Business Dynamics: Systems Thinking and Modeling for a Complex World*. Boston: Irwin McGraw-Hill. 2000.
- Sterman, J. Learning in and about complex systems. *System Dynamics Review* 10(2-3), 1994, p.291-330.
- Reason, J. *Managing the Risks of Organizational Accidents*. Brookfield, VT: Ashgate. 1997.
- Ross, L. "The intuitive psychologist and his shortcomings: Distortions in the attribution process" in L. Berkowitz (ed.), *Advances in Experimental Social Psychology*, vol. 10. New York: Academic Press. 1977.
- Wheeler, D. *Advanced Topics in Statistical Process Control*. Knoxville, TN: SPC Press. 1995.
- Wheeler, D. *Normality and the Process Behavior Chart*. Knoxville, TN: SPC Press. 2000.