Research Article

Quality and Reliability Engineering International

(wileyonlinelibrary.com) DOI: 10.1002/qre.1198

Published online 1 April 2011 in Wiley Online Library

Process Improvement in Healthcare: Overall Resource Efficiency

Jeroen de Mast,[a*†‡] Benjamin Kemper,[a§] Ronald J. M. M. Does,[a¶] Michel Mandjes[b‖] and Yohan van der Bijl[c]**

This paper aims to develop a unifying and quantitative conceptual framework for healthcare processes from the viewpoint of process improvement. The work adapts standard models from operation management to the specifics of healthcare processes. We propose concepts for organizational modeling of healthcare processes, breaking down work into micro processes, tasks, and resources. In addition, we propose an axiological model which breaks down general performance goals into process metrics. The connexion between both types of models is made explicit as a system of metrics for process flow and resource efficiency. The conceptual models offer exemplars for practical support in process improvement efforts, suggesting to project leaders how to make a diagrammatic representation of a process, which data to gather, and how to analyze and diagnose a process's flow and resource utilization. The proposed methodology links on to process improvement methodologies such as business process reengineering, six sigma, lean thinking, theory of constraints, and total quality management. In these approaches, opportunities for process improvement are identified from a diagnosis of the process under study. By providing conceptual models and practical templates for process diagnosis, the framework relates many disconnected strands of research and application in process improvement in healthcare to the unifying pursuit of process improvement. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: capacity analysis; lean six sigma; line balancing; overall resource efficiency; resource management

1. Introduction

Perhaps the first connotation with the topic of healthcare improvement is innovation in medical science, including innovations in treatment protocols, medical equipment, and pharmaceuticals[1]. This paper, however, focuses on the improvement of healthcare by improving its delivery, that is, by improving a hospital's primary patient processes, medical support processes, and nonmedical support processes. Characteristics of these processes, such as their capacity, efficiency, and reliability, determine important performance dimensions of healthcare, such as throughput, patient safety, and waiting times. Ultimately, they have a substantial impact on patient satisfaction, cost, and the quality and timeliness of medical care.

The improvement of processes, a perspective referred to as technical efficiency in the health economics literature[2], is the subject of a discipline that goes back to scientific management around 1900[3], and has resulted in manifestations that are well known in the quality discipline, such as total quality management, theory of constraints, business process reengineering, lean thinking, and six sigma[4–7]. In the recent years, business process management and workflow modeling have become thriving disciplines in information technology[8]. These approaches have been well studied in the academic literature, and tried and tested first in the industry, and later also in service organizations. The recent years have witnessed a growing interest from healthcare in these approaches[9–11].

In the process improvement paradigm, improvement originates in mapping processes and measuring carefully defined quality characteristics and performance metrics. In six sigma, for example, these diagnostic studies are done in the first three phases

[a]*Institute for Business and Industrial Statistics of the University of Amsterdam (IBIS UvA), Plantage Muidergracht 12, 1018 TV Amsterdam, The Netherlands*
[b]*KdV Institute for Mathematics, University of Amsterdam, P.O. Box 94248, 1090 GE, Amsterdam, The Netherlands*
[c]*Deventer Ziekenhuis, N. Bolkensteinlaan 75, 7416 SE, Deventer, The Netherlands*
**Correspondence to: Jeroen de Mast, Institute for Business and Industrial Statistics of the University of Amsterdam (IBIS UvA), Plantage Muidergracht 12, 1018 TV Amsterdam, The Netherlands.*
[†]*E-mail: j.demast@uva.nl*
[‡]*Associate Professor in Industrial Statistics and Principal Consultant at IBIS UvA.*
[§]*Consultant at IBIS UvA.*
[¶]*Professor in Industrial Statistics and Managing Director at IBIS UvA.*
[‖]*Professor in Applied Probability at the University of Amsterdam.*
[**]*Lean Six Sigma blackbelt.*

Copyright © 2011 John Wiley & Sons, Ltd.

of the DMAIC (define, measure, analyze, improve, control) stage structure[12]. This diagnosis of the process reveals improvement opportunities such as

- Optimizing capacity and utilization of staff and equipment, ensuring a smooth workflow with acceptable waiting times, and reducing cost for personnel and equipment.
- Reducing throughput times and waiting times by identifying bottlenecks and iterations in the processes.
- Optimizing or introducing standardized routing through the process, such as introducing sequencing rules, introducing restrictions on the amount of work-in-process as in kanban and CONWIP, or replacing batch-wise work with a single-piece flow discipline.
- Improving a process's reliability and safety by mitigating failure opportunities and making the process more robust.
- Reducing cycle times per task by optimizing work methods and procedures.
- Reducing variability in the process, thereby optimizing utilization and reducing waiting times.

Some of these improvement opportunities are self-evident once the process has been mapped and diagnosed; examples include poorly organized or inefficiently structured work, redundant work, and repeated but avoidable mistakes. Other improvement opportunities are derived from heuristics such as the ones from lean[6], business process reengineering[13], and the theory of constraints[14].

The idea that improvement opportunities follow from a diagnosis of the process under study discerns the process improvement paradigm, dominant in the quality literature, from competing approaches to healthcare improvement, dominant in the OR/MS literature. These OR/MS approaches are based on mathematical and simulation modeling; see, e.g. [15–19]. A substantial empirical basis of applications of process improvement in healthcare is already available; for example[20, 21], and the references therein. Also, there is an expanding literature discussing the techniques and methods for process improvement in healthcare[22–24].

This paper contributes conceptual models for process diagnosis in healthcare, thus facilitating projects according to the business process management, six sigma, lean, business process reengineering, theory of constraints, total quality management, or other process improvement approaches. We propose a class of organizational models, which conceptualize the types of elements that healthcare processes consist. To facilitate their application in process improvement, we associate them to an axiological model, which conceptualizes what constitutes value in healthcare processes. Third, we explicate the connexion between organizational and axiological models by a system of metrics for quantifying process flow. The metrics allow an analysis of the allocation of resources in healthcare processes, and we propose an aggregate metric that we refer to as *overall resource efficiency*. We also demonstrate how the proposed metrics help in *bottleneck analysis*.

This system of models contributes a unifying context and terminology to the methodological development of the field of healthcare delivery improvement. For practitioners, the models may serve as exemplars for diagnosing processes in hospitals, suggesting what to measure, and how to associate these measurements to organizational performance. The overall resource efficiency metric helps in identifying wasted capacity of resources, while bottleneck analysis helps in improving the flow or capacity. We demonstrate this practical value of the work by applying our models in a real improvement project, optimizing a CT scan process.

The presentation of the work has the following structure. Section 2 introduces our system of metrics for quantifying process flow. Section 3 presents a breakdown of workflow into micro processes, tasks and resources (the organizational model). Section 4 links these elements to value by proposing a breakdown of performance indicators (the axiological model). Section 5, finally, demonstrates the use of our models as an exemplar for studying a real healthcare process. We discuss the implications of our work in the Discussion and Conclusions section.

2. Process flow metrics

In the subsequent sections we develop our model for process flow in healthcare. Our model includes a system of metrics for calculating the capacities of resources, tasks, and processes, as well as efficiency factors for each. The calculations resemble the framework of overall equipment effectiveness (OEE) in the manufacturing industry[25, 26]. This framework allows the identification and diagnosis of bottlenecks in the process, the key to improving throughput or reducing waiting times. Further, it allows an assessment of the efficiency of the process, quantifying where resources are wasted. We propose to refer to our framework as overall resource efficiency (ORE).

In this section we introduce this system of metrics for healthcare processes by considering a single task involving a single type of resource.

2.1. Potential capacity

The effective workload EWL (Table I) is the number of patients to be processed, whereas ETP is the number of patients that is actually processed. For many processes, EWL may momentarily exceed the process's capacity, and therefore, when considered over smaller units of time, ETP<EWL. When considered over a suitably long period of time, workload and capacity are often balanced, and EWL= ETP. One of the stabilizing mechanisms is that long waiting queues tend to deter demand[27]. Another mechanism is staff working overtime until the work is done.

Copyright © 2011 John Wiley & Sons, Ltd.

Table I. Potential capacity and other metrics

Effective workload	EWL	Number of patients to be treated per time unit	Patients/day
Effective throughput	ETP	Number of patients treated per time unit	Patients/day
Total time	TotT	Resource time scheduled for a task	Min/day
Cycle time	CT	Processing and changeover time per patient	Min/patient
# Resources	N	Number of specimens of a type of resource	
Potential capacity	PCap	$=N\times\text{TotT}/\text{CT}$	Patients/day

Table II. Effective capacity and other metrics

Available time	AvT	Time that a resource is actually available for a task	Min/day
Availability	Av	$=\text{AvT}/\text{TotT}$	%
First time right	FTR	Ratio or percentage of jobs done right the first time	%
Nominal workload	NWL	$=\text{EWL}/\text{FTR}$	Patients/day
Nominal throughput	NTP	$=\text{ETP}/\text{FTR}$	Patients/day
Effective capacity	ECap	$=\text{FTR}\times\text{Av}\times\text{PCap}$	Patients/day

Table III. Utilization and idle time

Idle time	IT	$=\text{AvT}-\text{CT}\times\text{NTP}/N$	Min/day
Effective utilization	EUt	$=\text{ETP}/\text{ECap}$	%
		$=(\text{AvT}-\text{IT})/\text{AvT}$	

The cycle time CT is the required resource time per patient, and equals the sum of processing time per patient and changeover times in between patients. Given the total working time per day allotted to the task in question, TotT, and the number N of specimens of a resource, the potential capacity of the resource is $\text{PCap}=N\times\text{TotT}/\text{CT}$.

2.2. *Effective capacity: taking rework and availability into account*

Where TotT is the time that a resource is *budgeted* for a task, AvT (see Table II) is the time that the resource is actually *available* for the primary task (compare a machine's uptime in industry). For physicians and staff, AvT is typically TotT minus time lost to distractions, interruptions, searches for missing equipment, arranging for replacements for defective equipment, and other secondary activities. For equipment, causes of unavailability include being missing, defective, and in maintenance. The percentage of TotT that a resource is actually available, Av, is often below 100%, but in the case a resource works overtime, it can also be above 100%. To avoid confusion, we note that changeover times in between patients are not considered a part of resource unavailability, as they are a part of the patient cycle and included in CT.

Some of the work is not done right the first time, and must be redone; FTR is the percentage of jobs done right the first time (Table II). For each individual patient treated, the number of patient treatments (including double, triple, and more-than-triple counts) is higher. We discern *nominal* and *effective throughput*, and they are related as

$$\text{NTP}=\text{ETP}\times\sum_{k=0}^{\infty}(1-\text{FTR})^k=\frac{\text{ETP}}{\text{FTR}}.$$

For the nominal workload we have

$$\text{NWL}=\text{EWL}+(1-\text{FTR})\text{NTP}=\text{EWL}+\frac{1-\text{FTR}}{\text{FTR}}\text{ETP}. \tag{1}$$

If $\text{ETP}=\text{EWL}$, then Equation (1) reduces to $\text{NWL}=\text{EWL}/\text{FTR}$. Taking rework and availability into account, the effective capacity is typically lower than the potential capacity: $\text{ECap}=\text{FTR}\times\text{Av}\times\text{PCap}$.

2.3. *Utilization and idle time*

The effective utilization EUt (see Table III) is the ratio or percentage of the available time that the process is not idle ($\text{EUt}=(\text{AvT}-\text{IT})/\text{AvT}$), and also, EUt is the percentage of the effective capacity that is used (EUt=ETP/ECap). Idle time can best be calculated ($\text{IT}=\text{AvT}-\text{CT}\times\text{NTP}/N$), rather than measured, as employees adjust their work pace to camouflage overcapacity.

Even in bottlenecks, $\text{EUt}<100\%$ (and thus $\text{IT}>0$), as some idle time is unavoidable due to synchronization losses. Synchronization loss occurs if there is enough work in the system, but the resource has idle time because it is waiting for other resources or patients. Examples of causes of idle time due to synchronization are:

- Tardiness of patients or staff members, no-shows, or last-minute disruptions of the schedule[28].
- Schedules of physicians, rooms, and facilities impeding in utilizing all capacity.
- Variation in cycle times and fluctuations in demand.

Copyright © 2011 John Wiley & Sons, Ltd.

Taking the first two for self-evident, the third point follows from a generally known principle in industrial engineering (see, for example, Hopp and Spearman[29], especially chapters 8 and 9), which states that higher variability (in cycle times, inter-arrival times, outages, quality problems, and other sources) results in lower utilization, unless it is buffered against by keeping work on standby. The high level of synchronization needed to achieve near 100% utilization for all resources is unrealistic, and therefore, a certain percentage of nonutilized capacity is unavoidable. However, a possibly substantial fraction of nonutilized capacity is typically dispensable (the resource's 'overcapacity'), especially in the nonbottleneck resources.

2.4. Diagnostics for process flow improvement

The metrics introduced in the previous sections allow the identification of improvement opportunities, which, in the process improvement paradigm, are identified from process diagnosis. First, we discuss *bottleneck analysis,* the optimization of a bottleneck, which is a resource whose throughput ETP is smaller than its workload EWL. The equation $\mathrm{ETP} = \min\{\mathrm{EWL}, \mathrm{EUt} \times \mathrm{ECap}\}$ suggests two improvement strategies. The first is to improve the bottleneck's capacity. The equation

$$\mathrm{ECap} = \mathrm{FTR} \times \mathrm{Av} \times N \times \mathrm{TotT} / \mathrm{CT}$$

reveals several options:

- Reduce cycle time CT by reducing processing time per patient or changeover times.
- Extend the budgeted resource time TotT.
- Increase the number of resources *N*.
- Improve availability Av by limiting distractions or working overtime.
- Improve the first-time-right ratio FTR.

The second strategy is to improve the bottleneck's utilization EUt. For a bottleneck, all idle times can be assumed attributable to synchronization losses, so better synchronization of patients and other resources with the bottleneck is the key to improvement. Some options include:

(1) Schedule patients so as to build up a buffer of work on standby.
(2) Schedule patients to minimize variation in cycle times (for example time slots with homogeneous patient groups).
(3) Increase the capacities of other resources in the micro process to build up a buffer of work before the bottleneck.
(4) Influence demand to reduce fluctuations in workload, or adjust capacities to match fluctuations in demand.
(5) Reduce tardiness, no-show, cancellations, and other disruptions of schedules.
(6) Improve the reproducibility of the process (standardization and structuring of work, well defined and coordinated routing, and minimal rework and iterations).
(7) Change the order of tasks, eliminate redundant tasks, merge tasks, or modify the breakdown of work into tasks.

These options are based on well-known principles from lean thinking, industrial engineering, and especially the theory of constraints[14, 30]. In particular, options (1)–(4). follow directly from the principle that variability in a process will be buffered against by a combination of work in process, waiting time and excess capacity[29]. Reducing variability, or keeping a buffer of work on stand-by, reduces excess capacity and thus improves utilization. Options (5) and (6) exploit the same principle by eliminating variability. Option (7) is quite general, and comprises the redesign of a process with an eye for reducing propagation of variability through the process, for reducing variability by pooling of variation sources, and for making processes less complex and less interdependent[31].

Besides the optimization of bottlenecks, one could pursue the reduction of wasted capacity in nonbottleneck resources. The *overall resource efficiency* indicates what percentage of a resource's potential capacity is effectively used. It can be broken down into three efficiency factors:

$$\mathrm{ORE} = \mathrm{ETP} / \mathrm{PCap} = \mathrm{EUt} \times \mathrm{FTR} \times \mathrm{Av}.$$

Low percentages show where capacity is wasted:

- Low availability Av: capacity is wasted due to distractions, disturbances and other secondary activities.
- Low first-time-right FTR: capacity is wasted due to rework.
- Low effective utilization EUt: capacity is wasted as idle time.

The last term suggests that it may be possible to discard part of the nonutilized capacity, thus saving on costs or making this capacity available for other purposes. It is in general difficult to determine analytically which fraction of nonutilized capacity can be discarded without consequences for the ETP; the pursuit of near 100% utilization for one resource typically creates substantial synchronization idle times for other resources. One approach is to determine a safe capacity level empirically. The idea is to remove all nonutilized capacity (that is, one reduces the number *N* of resources or the total time TotT until $\mathrm{ECap} = \mathrm{EWL}$, and EUt approaches 100%). This will typically result in a growing queue of work somewhere in the system. By gradually increasing capacity until the queue stabilizes, one determines a realistic need for capacity. Simulation modeling (e.g. Davies and Davies[15] and Jun *et al.*[16]) is a more thorough approach.

Copyright © 2011 John Wiley & Sons, Ltd.

3. Organizational models

The metrics introduced in the previous section are the building blocks for our models for healthcare processes. Our models comprise two types of diagrams. The first type, such as the ones in Figures 1 and 2, has an organizational focus. It models how the work to be done is broken down into micro processes, tasks, and how these tasks are assigned to resources. The second type, such as the one in Figure 3, has a focus on value.

3.1. Micro and macro processes

Considering patient trajectories in healthcare, it is fruitful to discern between two types of processes, which we name *macro* and *micro processes*. The motivation for the distinction is in their decisively different stochastic behavior, underlying structure of influence factors, and functional implications. Macro processes are the end-to-end trajectories that patients follow (see Figure 1). Their dynamics revolve around waiting times (in the order of magnitude of days and weeks) and scheduling efforts. The 'jobs' flowing through the process are patients. The stochastic behavior of the process flow is similar to that of the typical exemplars in queuing theory: random, perhaps Poisson, job arrivals[32]; queues arising from a mismatch between capacity (of staff and facilities) and workload; and amplified by synchronization problems.

The building blocks of the macro processes are the micro processes. The jobs flowing through micro processes can be patients, but also requests for an examination, files that are processed, or other types of jobs. Often, but not always, micro

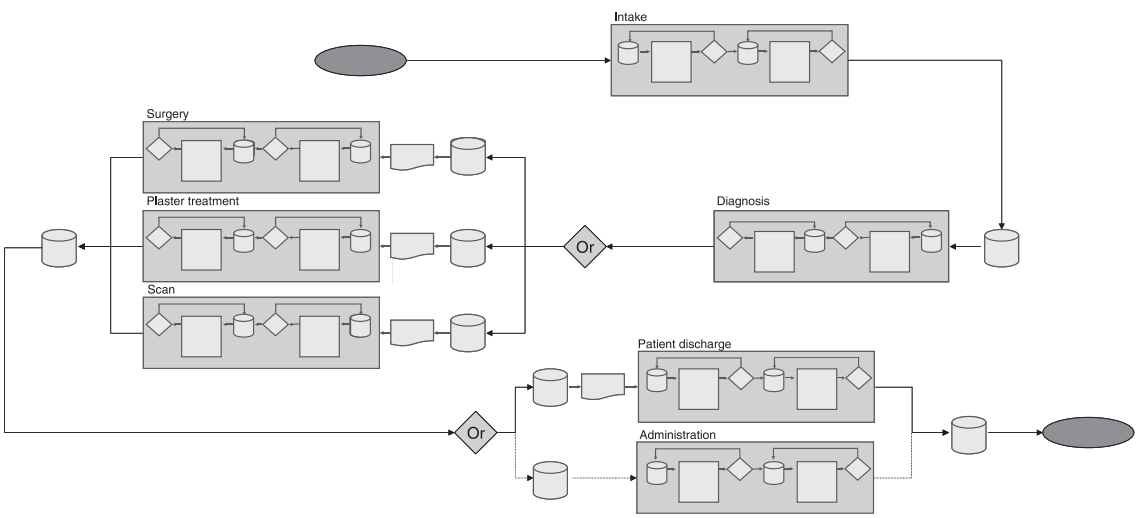

Figure 1. Paradigmatic form of healthcare processes. The figure shows a macro process (end-to-end patient trajectory) involving seven micro processes. The micro processes are often preceded by a scheduling step and a queue, which act as a buffer transforming a (typically Poisson) stream into a scheduled stream

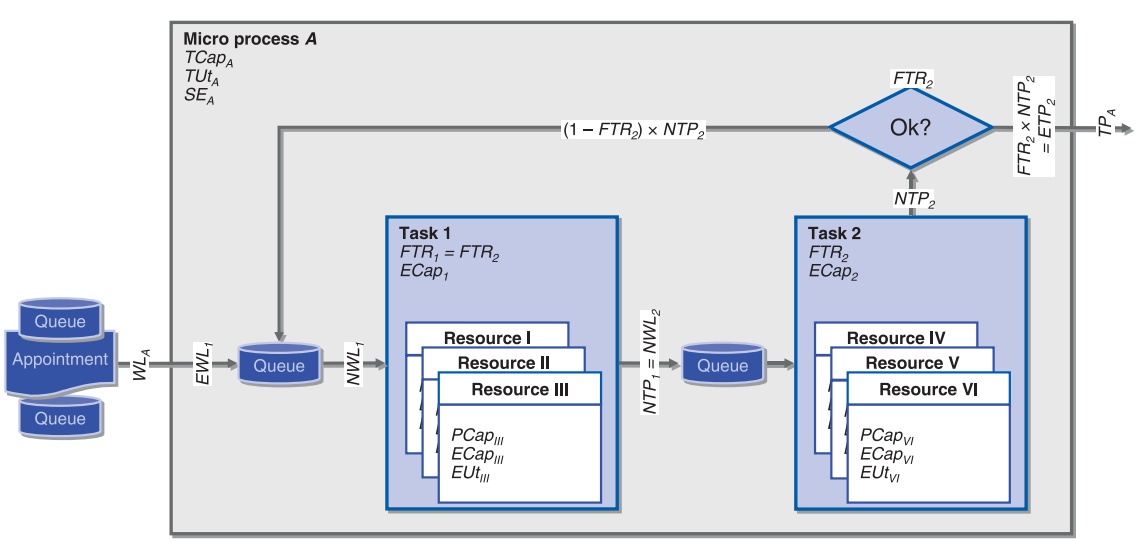

Figure 2. Organizational model of micro processes including the queue before the micro process, the queues between the tasks in the micro process, and metrics for capacity and utilization

Copyright © 2011 John Wiley & Sons, Ltd.

Table IV. True capacity, true utilization, and synchronization efficiency

True capacity	TCap	Capacity of a micro process, taking synchronization into account	Patients/day
True utilization	TUt	TUt = TP/TCap	%
Synch. efficiency	SE	SE=TCap/min{ECap}	%

processes are preceded by a scheduling step, in which case arrivals are typically not Poisson-like, but characterized as random variation around a target arrival time plus random no-show[32]. In many micro processes the scheduling ensures that workload does not (substantially) exceed capacity, and as a consequence, the main waiting queue is *before*, rather than *in* the micro process.

Some macro processes are completely polyclinical (outpatient), meaning that all micro processes involved are polyclinical, while others involve a combination of polyclinical and clinical (inpatient) micro processes. Micro processes can be discerned into:

- Primary patient processes: micro processes that have the patient as one of the inputs, such as intakes, diagnostic consults, computed tomography (CT) scans, or surgeries.
- Medical support processes: micro processes that do no not have the patient as an input, such as pathological examinations or sterilization services.
- Nonmedical support processes: services that are not directly related to the patient's primary patient process, such as transport of patients, preparation of meals, or advertisement of staff vacancies.

3.2. *Modeling process flow in the micro process*

In our model, the main organizational building blocks for the micro processes are *tasks* (linked in chronological sequence by routes), *queues* (where jobs, mostly patients, sit idle for some time while no action is performed on them), and *resources* that are involved in tasks. Resources could be staff (such as nurses and operators), physicians, equipment (such as MRI scanners), and other facilities (such as rooms). Note that resources can be allotted to more than one task. The metrics introduced in the previous section can be applied to resources, tasks, and entire micro processes. We differentiate metrics by subscript indices, where resources are numbered $I, II, III, \ldots$; tasks are numbered $1, 2, 3, \ldots$; and micro processes $A, B, C, \ldots$.

In Figure 2, the workload WL_A of the micro process A is the number of patients per day scheduled for the micro process (note that we drop the distinction between nominal and effective workload if there is no rework). There will often be a queue where patients wait before they are scheduled, and there will be a waiting time (also called 'queue' in the figure) until the scheduled time has arrived; both queues are not part of the micro process. Note in particular that patients waiting to be scheduled are not included in WL_A, but WL_A does include emergency workload and walk-ins. Patients enter micro process A when they arrive at the hospital. Arrival times are stochastically distributed around the scheduled times, and the first step in the micro process is again a queue (typically the waiting room).

The throughput TP_A (on the right-hand side of the diagram) is the number of patients per time unit that is actually treated. If the schedule is realistic, this number will typically be equal to WL_A. The effective workload for task 1, $EWL_1 = WL_A$, is augmented with rework, whence the nominal workload NWL_1 is higher. From the potential capacities and availabilities of the resources ($PCap_R$ and Av_R), and the FTR percentage of the task, the effective capacity $ECap_1$ of the tasks can be determined. For task 1, for example,

$$ECap_1 = FTR_1 \times \min\{PCap_I \times Av_I, PCap_{II} \times Av_{II}, PCap_{III} \times Av_{III}\}.$$

One should be careful which FTR percentage to use, depending on the particulars of the rework routes. In the example, failures in tasks 1 and 2 are revealed not until the end of task 2, in which case both tasks must be redone. In this particular setting, therefore, the first-time-right percentage is the same for both tasks ($FTR_1 = FTR_2$). The nominal throughput of task 1 is the nominal workload for task 2; the nominal throughput of task 2, multiplied by FTR_2, gives the effective throughput of task 2 ($ETP_2 = FTR_2 \times NTP_2$).

On micro process level, we define the true capacity TCap (Table IV) to be the maximum throughput that can be achieved (given the current N, TotT, CT, Av and FTR). We have $TP \leq TCap \leq \min\{ECap_1, ECap_2\}$, that is, the micro process's capacity is larger than the throughput, but not larger than the lowest capacity of the tasks that it entails. Since, as explained before, it is unrealistic that all resource utilizations are near 100%, the TCap of a micro process is usually substantially smaller than the lowest of the effective capacities of the tasks. The ratio between the two is the process's synchronization efficiency SE (Table IV). The percentage of TCap which is actually utilized, and therefore results in TP, is the process's true utilization TUt. The TCap (and the related SE) can best be determined empirically, either by experimenting with the real process or a simulation model. Increasing the workload WL until growing queues emerge reveals the process's TCap.

Copyright © 2011 John Wiley & Sons, Ltd.

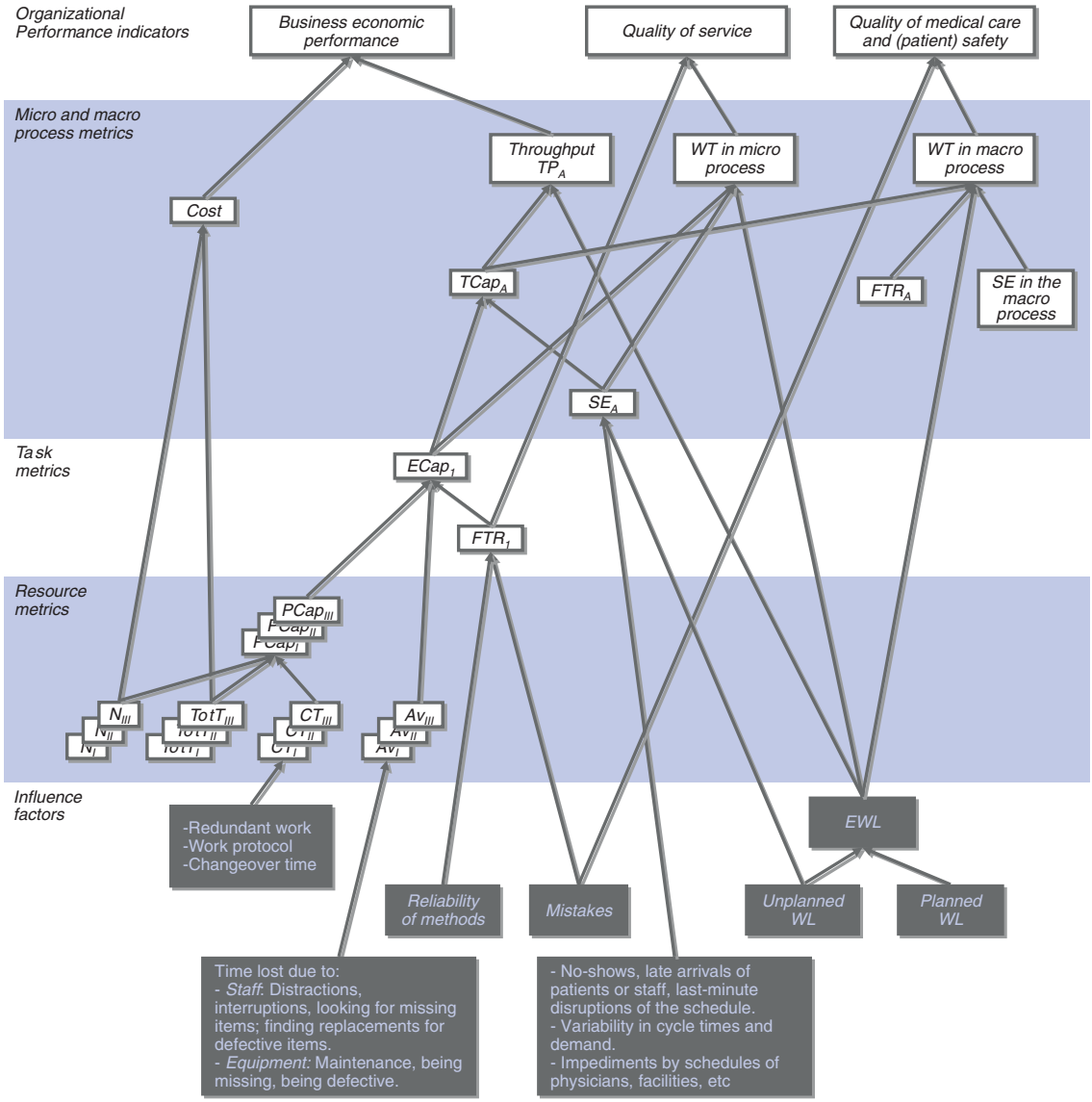

Figure 3. Axiological model for healthcare processes

4. Axiological model and performance metrics

The organizational models in the previous section are complemented by the axiological model in Figure 3. In the downward direction, it shows how organizational objectives relate to the process flow metrics defined in the previous sections, and thus, it helps to translate organizational goals into measurable metrics. In the upward direction, it shows the relevance of process metrics.

The process's flow affects the hospital's *business economic performance* through operational cost (partly determined by the numbers *N* of resources and the resource times TotT that are allocated to a certain task), and the throughput—assuming that the hospital receives payment from the government, insurance companies, or patients themselves proportional to the number of treatments. Throughput depends on the capacities of the micro processes (TCap) and the workload (EWL). TCap is determined by the synchronization efficiency SE and the effective capacities ECap of the tasks in a micro process, whose further breakdown has been explicated in the previous sections.

Quality of service refers to issues that may be an annoyance to patients, but do not jeopardize the patients' health[33]—think of long waiting times in the waiting room, or having to undergo an examination twice because the first time failed. There are numerous factors beyond process flow affecting quality of service, such as courtesy of staff and cleanliness of the facilities, but first-time-right (FTR) percentages of the tasks and waiting times within the micro processes are two process flow issues impacting service quality.

Copyright © 2011 John Wiley & Sons, Ltd.

Under *quality of medical care and* (*patient*) *safety* are understood factors that affect the patient's health and the effectiveness of the medical treatment[34]. Quality of medical care is affected by a few issues in the process's flow, besides of course many factors not related to process flow. In particular, quality of medical care depends on mistakes and errors in the process, which could harm the patient, and by waiting times in the macro process, which could result in treatments being overdue. The latter in turn are determined by the capacities of the micro processes, the workload, the FTR ratios of the micro processes, and the synchronization efficiencies in the macro process (problems arising in matching schedules of patients, physicians, and facilities).

At the bottom of the diagram, we see that cycle times of tasks depend on the work protocol (maybe alternative work procedures are more efficient?), redundant work (maybe some subtasks have no function and can be skipped?), and changeover time (maybe the time in between patients can be minimized?). Availability is influenced by distractions, interruptions, searches for missing items, finding replacements for missing items, and other secondary activities (for staff), and maintenance, being missing, and being defective (for equipment).

5. Real-life example: CT scan process

To illustrate the metrics and models introduced in the previous sections we discuss a micro process in a computed tomography (CT) scan department. The example results from a six sigma project at the Deventer Hospital, a medical teaching hospital in the Netherlands. The measurements were collected during 6 nonsequential days. For each arriving patient the planned arrival time, actual arrival time and start/stop times of all tasks were measured. Also, attributes such as age, type of patient, type of examination, and date of appointment scheduling were recorded. In total, 66 patients treated during polyclinical hours are included in the sample.

A CT scan is a medical imaging method used in the diagnostic phase of a healthcare macro process. The method is part of the branch of medicine called radiology, and the micro process is an example of a primary patient process. The scan process as depicted in Figure 4 has two input streams, a stream from a waiting list of patients scheduled for a scan, and a stream of emergency patients. Emergency patients are handled with priority and scheduled as first patient in line (a nonpreemptive queuing discipline). The scheduled patients are typically scheduled 8am–1pm on workdays, and are treated in the scheduled order. In general, 18 patients are scheduled in time slots of 15 min resulting in 4.5 h of outpatient time slots per day. The remaining 0.5 h are allocated for breaks and emergency patients. During the morning, an average of 1.3 emergency patients arrive, yielding a total workload of $WL_A = 19.3$ patients per day (where a day is understood to be 8am–1pm).

For the outpatient stream, the average waiting time between the appointment and the actual visit is 30.8 days. The waiting time is partly determined by the process's capacity and the patient's flexibility, and partly by the term and priority advised by the specialist. Scheduled patients arrive at the radiology department's reception desk. Upon arrival they are registered and enter the waiting room (average waiting time 7.1 min). When summoned, an outpatient enters a dressing room ('Task 1: (Un)dress').

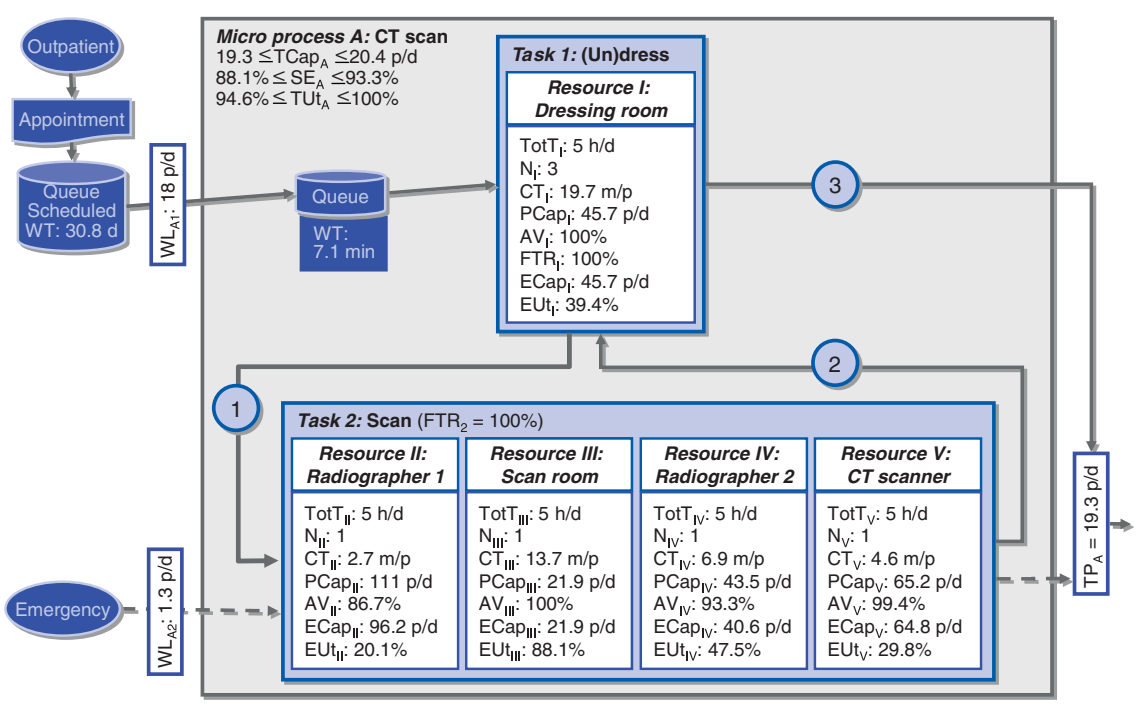

Figure 4. The CT scan primary patient process in its current configuration with two input streams: scheduled and emergency patients

Copyright © 2011 John Wiley & Sons, Ltd.

The dressing room is occupied during the whole process, for a cycle time of 19.7 min per patient on average. The three dressing rooms are 100% available during the morning shift, resulting in a $PCap_I$ of 45.7 patients per day. The undressed outpatient proceeds to the second task ('Scan'), indicated by route 1 in Figure 4. After the scan, an outpatient returns to his dressing room (route 2). When the patient is dressed again, he or she leaves the CT scan process (route 3). The emergency patients enter the process via the dashed route in Figure 4, directly from their rooms in the emergency department located next to the radiology department. After 'Scan' they return to the emergency department.

Task 2 is facilitated by a scan room and the task may consist of two sub-tasks performed by diagnostic radiographers: a fluid injection for some patients (about 62%) and a CT scan for all patients. Patients that need fluid injection to increase visibility of vital parts in the scan are injected by 'Radiographer 1'. The expected fluid injection time is 4.4 min in total, including after-care when the CT scan is finished (resulting in $CT_{II} = 62\% \times 4.4 = 2.7$ m/p when averaged over all patients). The second sub-task is executed by 'Radiographer 2' and the machine 'CT scanner'. This subtask takes on average 6.9 min of which 4.6 min is CT scan time; the remainder of the Radiographer 2's cycle time consists of transportation time and instruction time.

The expected total time in the scan room is 13.7 min, including a visible fluid injection for some patients, a CT scan, small transportations and instructions. There is no significant difference between the cycle times of outpatients and emergency patients. Given these cycle times we calculate the effective capacities of resources *II,…,V*. For the radiographers, $Av_{II} = 86.7\%$ and $Av_{IV} = 93.3\%$ due to their 15 min' coffee break and interruptions, such as phone calls and incomplete requests. The availability of the CT scanner is 99.4% due to disturbances. The first-time-right percentages of both Tasks 1 and 2 are 100% (some rework and iterations are included in the cycle time of the CT scanner). Consequently, ETP = NTP = 19.3 p/d. The resulting effective capacities are as follows:

- Dressing room: $ECap_I = (300 \times 3/19.7) \times 1.00 \times 1.00 = 45.7$ patients per day (p/d).
- Radiographer 1: $ECap_{II} = 96.2$ p/d (of whom only 62% would need a fluid injection).
- Scan room: $ECap_{III} = 21.9$ p/d.
- Radiographer 2: $ECap_{IV} = 40.6$ p/d.
- CT scanner: $ECap_V = 64.8$ p/d.

The scan room, having the lowest effective capacity, is the constraining resource in the process; it would become a bottleneck if workload increased. Its effective utilization is 88.1%. Most interruptions of the radiographers are taken care of in their idle time, and do not interfere with the utilization of the scan room, but the radiographers' coffee break and a few interruptions make an idle time of 20 min per day unavoidable for the scan room. Therefore, the utilization of the scan room cannot be above $(300-20)/300 = 93.3\%$. Thus, an upper bound for the process's synchronization efficiency is $SE_A \leq 93.3\%$, and $TCap_A \leq 93.3\% \times ECap_{III} = 20.4$ p/d. The current average throughput serves as a proven lower bound, and $TCap_A \geq 19.3$ p/d. A sharper lower bound could, in some cases, be found by taking the highest daily throughput achieved as lower bound. In this case, however, we are afraid that this top day is not representative of the maximum throughput, but rather represents a day with a more than average number of easy patients (patients not requiring fluid injection), and therefore, the top throughput could not be sustained over longer periods. The process's true utilization of $TUt_A \geq 94.6\%$ indicates that, given the current configuration, the process is operating near its maximal throughput. The radiographers and CT scanner, all expensive resources, have fairly low utilizations. For example, the overall resource efficiency of Radiographer 1 is $ORE_{II} = 20.1\% \times 100\% \times 86.7\% = 17.4\%$.

The analysis above helps us to identify constrictions in the performance of the current CT scan process. The improvement effort is focused on improving the true capacity, in order that more patients can be treated per day, and simultaneously improving the utilization of the radiographers and CT scanner. Improving the effective utilization of the scan room gives only limited prospects at improvement; at best, it goes from $EUt_{III} = 88.1$ to 93.3%, improving true capacity by only 1.1 patients per day. Better opportunities are revealed by the equation $ECap_{III} = FTR_{III} \times Av_{III} \times N_{III} \times TotT_{III}/CT_{III} = 100\% \times 100\% \times 1 \times 300/13.7$. The FTR and Av are already perfect (for example, cleaning and maintenance of the scan room are scheduled such that they do not interfere, ensuring that Av is 100%). Opening a second scan room ($N_{III} = 2$) would double the capacity, but then one would also need another CT scanner and possibly more staff. Scheduling longer hours for the service (increasing $TotT_{III}$) would improve $ECap_{III}$, but this is a costly option, as it does not improve the low utilizations of the other resources. It was decided to focus on the cycle time CT_{III}. Following principles from the theory of constraints[14, 30], we spare the bottleneck resource as much as possible. Thus, the scan room is used only for the CT scan itself, moving other tasks (fluid injection and after care) to an area next to the scan room (Figure 5). Further, Radiographer 1 copes with interruptions as much as possible, improving the availability of Radiographer 2 to $Av_{IV} = 95\%$ (the last 5% unavailability due to coffee breaks).

Emergency patients and undressed outpatients either go directly to 'Task 3: Scan', or first to 'Task 2: Injection', executed by Radiographer 1. Radiographer 1 is assigned to both task 2 ($TotT_{II\text{-}2} = 3.1$ h per day) and task 4 ($TotT_{II\text{-}4} = 1.9$ h/d); we have differentiated the process flow metrics by subscripts II-2 and II-4 (Figure 5). The calculations predict effective capacities for tasks 2 and 4 of 58.6 and 57.0 p/d.

With injection, after care and some patient–radiographer interaction taken out, the cycle time for the scan room is expected to fall to 9.3 m/p, increasing its effective capacity to $ECap_{III} = 32.3$ p/d. It is still the bottleneck in the process, and since its effective utilization cannot be more than 95% (due to coffee breaks), the true capacity of the new process is predicted to be below $95\% \times 32.3 = 30.7$ p/d. We propose to schedule 29 outpatients per day, which, combined with an inflow of 1.3 emergency patients per morning, would give a true utilization of at least 98.7%. The extra 11 patients per day (compared to the old setting), or about 2700 patients per year, would generate k€361 extra revenue per year. The CT scanner is still idle for about half the

Copyright © 2011 John Wiley & Sons, Ltd.

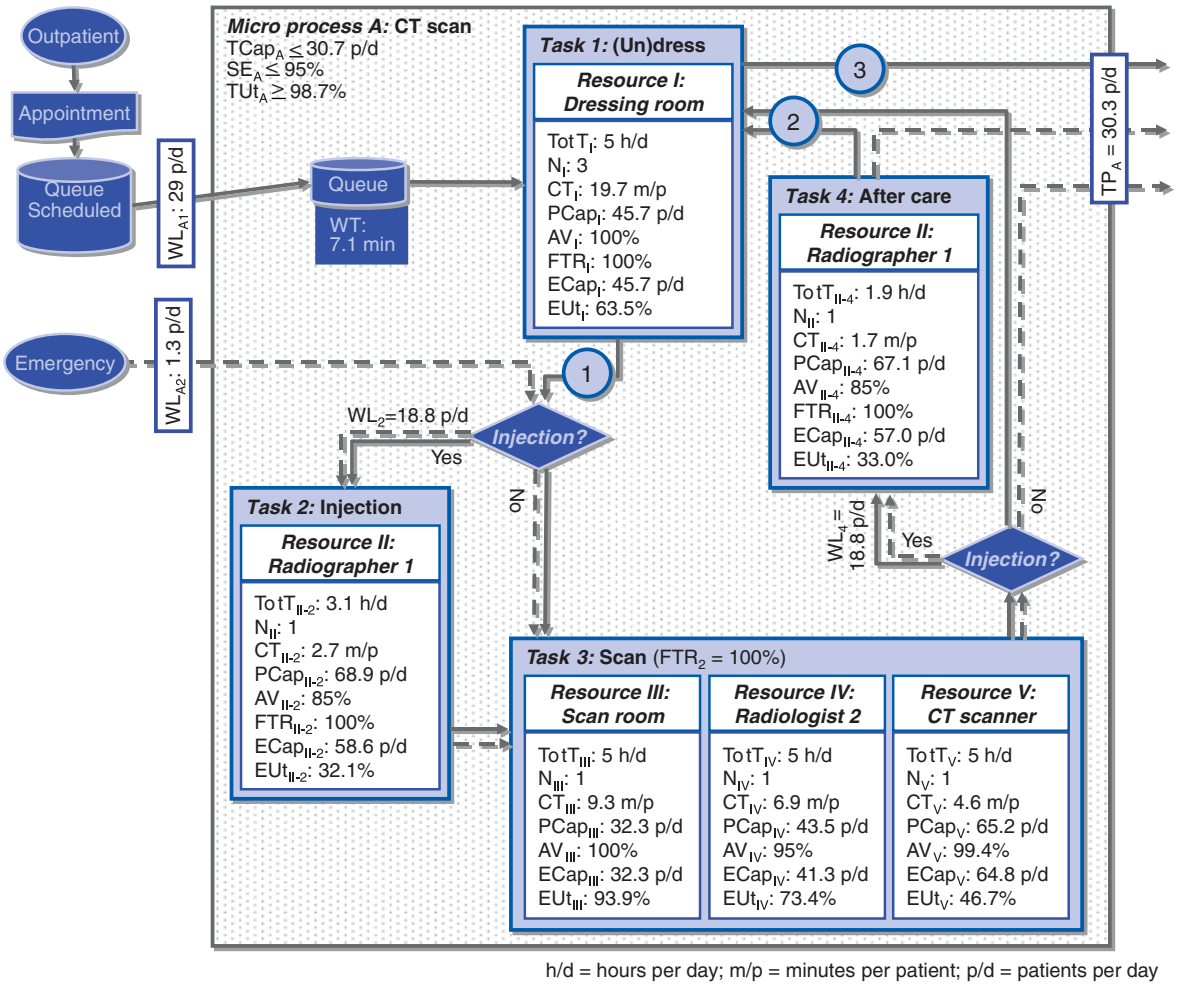

Figure 5. The redesigned CT scan primary patient process with two input streams: scheduled and emergency patients

time (EUt_{IV} = 46.7%), and one could argue that there is still room for improvement, but demand may not be sufficient to utilize more of its capacity. Another direction for improvement is the substantial idle time for Radiographer 1 ($EUt_{II} \approx 32\%$). One could try a scheduling discipline where patients who need a fluid injection are scheduled first. As soon as the last patient needing a fluid injection has been treated, Radiographer 1 is available for other duties (thus reducing his or her $TotT_{II}$).

6. Discussion and conclusions

Process improvement in healthcare is an urgent and important pursuit. This paper's contributions to that pursuit can be summarized as follows.

(1) A system of metrics for quantifying capacities, utilizations, and overall resource efficiency. The system is flexible enough to be of use in the variety of process structures typical for healthcare.
(2) An organizational model which breaks down healthcare processes into macro and micro processes, and the latter into tasks and resources. The model is the basis for the types of diagrams such as in Figures 1, 2, 4, and 5, which we propose as useful instruments in process diagnosis.
(3) An axiological model (Figure 3) which relates general business objectives of hospitals to process flow metrics.

6.1. Managerial implications

The three components mentioned above provide a conceptual framework for understanding and studying process improvement in healthcare in a general context. These components also offer methodological guidance to a project leader responsible for improving processes in a hospital. The presented material has been the basis for training material, which we have integrated in our lean six sigma training curriculum for courses that we teach to professionals in healthcare. The material suggests to project leaders how to make a diagrammatic representation of the process under study, which data to gather, and how to analyze

Copyright © 2011 John Wiley & Sons, Ltd.

and diagnose a process's flow and resource utilization. The proposed diagnostics for bottlenecks and ORE optimization provide guidelines for a methodical exploration of improvement directions. Further, the models offer an instrument for hypothesizing about alternative configurations, and predicting their performance. Finally, they facilitate laying down the specifications for a redesigned process.

6.2. Integration of the work in standard process improvement approaches

The presented models can be readily integrated in currently popular standard improvement approaches, such as the ones mentioned in the introduction. Both in lean thinking and in business process management (BPM) there is an emphasis on diagrammatic modeling of processes. Our type of diagrams, as in Figures 1, 2, 4 and 5, is an alternative, tailored to healthcare processes and the analysis of process flow, to the value stream map in lean thinking[35], and the business process modeling language, unified modeling language, and other standards in BPM[36]. Also six sigma prescribes mapping of processes; further, our axiological model (Figure 3) links on to six sigma's prescription to frame a project's objectives in terms of measurable characteristics named *critical to quality* (CTQ). In fact, Figure 3 represents a generic CTQ-flowdown, see[37], for six sigma projects in healthcare. The figure also places quality, defects and variability, the traditional focal points of six sigma and total quality management, in a coherent breakdown of value in healthcare processes. The ORE system of metrics, finally, facilitates application of the Five Focusing Steps of the theory of constraints[14, 30].

6.3. Directions for future research

An important topic for further study is to develop empirical techniques for determining the metrics proposed in Section 2 for actual processes under study. Most of the presented metrics can be measured by direct observation, and it would be useful to identify methods and equipment which make such data gathering as efficient and reliable as possible, possibly through automation. Some of the metrics cannot be determined in a straightforward manner. For example, although we have made some suggestions for establishing a process's true capacity and synchronizing efficiency, a more thorough study and practical guidelines for setting up such an experiment would be useful.

A second direction for research, also highly relevant in the authors' opinion, and enabled by the models expounded in this paper, is to refine the models for selected generic processes in hospitals. For example, most hospitals have one or more CT scan processes, and by describing a certain number of them in the generic format proposed in this paper, one could compare their organization and performance across hospitals. Eventually, this could result in the identification of standards and best practices.

References

1. Djellal F, Gallouj F. Mapping innovation dynamics in hospitals. *Research Policy* 2005; **34**(6):817–835.
2. Palmer S, Torgerson DJ. Economic notes: Definitions of efficiency. *British Medical Journal* 1999; **318**(7191):1136.
3. Wren DA. *The History of Management Thought* (5th edn). Wiley: Chichester, U.K., 2005.
4. Goldratt EM. Computerized shop floor scheduling. *International Journal of Production Research* 1988; **26**(3):443–455.
5. Hammer M. Reengineering work: Don't automate, obliterate. *Harvard Business Review* 1990; **68**(4):104–112.
6. Womack JP, Jones DT. *Lean Thinking: Banish Waste and Create Wealth in Your Corporation*. Free Press: Florence, IT, 2003.
7. De Mast J, Does RJMM, De Koning H. *Lean Six Sigma for Service and Healthcare*. Beaumont Quality Publications: Alphen aan den Rijn, NL, 2006.
8. Van der Aalst W, Van Hee K. *Workflow Management: Models, Methods, and Systems*. MIT Press: Cambridge, MA, 2004.
9. Locock L. Healthcare redesign: Meaning, origins and application. *Quality and Safety in Healthcare* 2003; **12**(1):53–58.
10. Gowen CR, Mcfadden KL, Hoobler JM, Tallon WJ. Exploring the efficacy of healthcare quality practices, employee commitment, and employee control. *Journal of Operations Management* 2006; **24**:765–778.
11. Langabeer JR, DelliFraine JL, Heineke J, Abbass I. Implementation of lean and six sigma quality initiatives in hospitals: A goal theoretic perspective. *Operations Management Research* 2009; **2**:13–27.
12. De Koning H, De Mast J. A rational reconstruction of six sigma's breakthrough cookbook. *International Journal of Quality and Reliability Management* 2006; **23**(7):766–787.
13. Reijers HA, Mansar SL. Best practices in business process redesign: An overview and qualitative evaluation of successful redesign heuristics. *Omega* 2005; **33**:283–306.
14. Rahman S. Theory of constraints: A review of the philosophy and its applications. *International Journal of Operations and Production Management* 1998; **18**(4):336–355.
15. Davies R, Davies HTO. Modelling patient flows and resource provision in health systems. *Omega* 1994; **22**(2):123–131.
16. Jun JB, Jacobson SH, Swisher JR. Application of discrete-event simulation in health care clinics: A survey. *Journal of the Operational Research Society* 1999; **50**(2):109–123.
17. Lane DC, Monefeldt C, Rosenhead JV. Looking in the wrong place for healthcare improvements: A system dynamics study of an accident and emergency department. *Journal of the Operational Research Society* 2000; **51**(5):518–531.
18. Green LV. Queueing analysis in health care. *Patient Flow*: *Reducing Delay in Healthcare Delivery*, Hall RW (ed.). Springer: New York, NY, 2006; 281–307.
19. Rohleder TR, Bischak DP, Baskin LB. Modeling patient service centers with simulation and system dynamics. *Health Care Management Science* 2007; **10**:1–12.
20. Does RJMM, Vermaat MB, Verver J, Bisgaard S, Van den Heuvel J. Reducing start time delays in operating rooms. *Journal of Quality Technology* 2009; **41**(1):95–109.
21. Bisgaard S. *Solutions to the Healthcare Quality Crisis: Cases and Examples of Lean Six Sigma in Healthcare*. ASQ Quality Press: Milwaukee, WI, 2009.
22. Plsek PE. Systematic design of healthcare processes. *Quality in Health Care* 1997; **6**:40–48.
23. Does RJMM, Vermaat MB, De Koning H, Bisgaard S, Van den Heuvel J. Standardizing health care projects. *Six Sigma Forum Magazine* 2006; **6**(1):14–23.

Copyright © 2011 John Wiley & Sons, Ltd.

24. Hall RW. *Patient Flow: Reducing Delay in Healthcare Delivery*. Springer: New York, NY, 2006.
25. Ljungberg O. Measurement of the overall equipment effectiveness as a basis for TPM activities. *International Journal of Operations and Production Management* 1998; **18**(5):495–507.
26. Nakajima S. *An Introduction to TPM*. Productivity Press: Portland, OR, 1988.
27. Worthington D. Hospital waiting list management models. *Journal of the Operational Research Society* 1991; **42**(10):833–843.
28. Kim SC, Horowitz I. Scheduling hospital services: The efficacy of elective-surgery quotas. *Omega* 2002; **30**:335–346.
29. Hopp WJ, Spearman ML. *Factory Physics* (3rd edn). McGraw-Hill: Boston, MA, 2008.
30. Davies J, Mabin VJ, Balderstone SJ. The theory of constraints: a methodology apart?—A comparison with selected OR/MS methodologies. *Omega* 2005; **33**(6):506–524.
31. Skinner W. The focused factory. *Harvard Business Review* 1974; **52**(3):113–121.
32. Alexopoulos C, Goldsman D, Fontanesi J, Kopald D, Wilson JR. Modeling patient arrivals in community clinics. *Omega* 2008; **36**:33–43.
33. Kenagy JW, Berwick DM, Shore MF. Service quality in health care. *Journal of the American Medical Association* 1999; **281**(7):661–662.
34. Donabedian A. The quality of medical care. *Science* 1978; **200**(4344):856–864.
35. Braglia M, Carmignani G, Zammori Z. A new value stream mapping approach for complex production systems. *International Journal of Production Research* 2006; **44**(18):3929–3952.
36. Aguilar-Savén RS. Business process modelling: Review and framework. *International Journal of Production Economics* 2004; **90**:129–149.
37. De Koning H, De Mast J. The CTQ flowdown as a conceptual model of project objectives. *Quality Management Journal* 2007; **14**(2):19–28.

Authors' biographies

Jeroen de Mast obtained a doctorate in Statistics from the University of Amsterdam. Currently, he works as a principal consultant at the Institute for Business and Industrial Statistics (IBIS UvA), and as associate professor at the University of Amsterdam. He has coauthored several books about Six Sigma, and is recipient of the ASQ Feigenbaum, Brumbaugh and Nelson awards, as well as the ENBIS Young Statistician Award. He is a senior member of ASQ, and associate member of the International Academy for Quality.

Benjamin Kemper holds a degree in Econometrics and Operations Research from the University of Amsterdam, the Netherlands. He is a consultant at IBIS UvA, and is working on a PhD project that focuses on process flow optimization in service networks.

Ronald J. M. M. Does obtained a PhD in Mathematical Statistics from the University of Leiden. From 1981 to 1989, he worked at the University of Maastricht, where he became the Head of the Department of Medical Informatics and Statistics. In that period his main research interests were medical statistics and psychometrics. In 1989 he joined Philips Electronics as a senior consultant in Industrial Statistics. Since 1991 he is Professor of Industrial Statistics at the University of Amsterdam. In 1994 he founded IBIS UvA, which operates as an independent consultancy firm within the University of Amsterdam. The projects at this institute involve the implementation, training and support of Lean Six Sigma, among others. His current research activities are the design of control charts for nonstandard situations, the methodology of Lean Six Sigma and healthcare engineering.

Michel Mandjes obtained a PhD in Operations Research and Applied Probability from the Vrije Universiteit, Amsterdam. After having worked as a Member of Technical Staff at KPN Research (Leidschendam, the Netherlands), and Lucent Technologies/Bell Laboratories (Murray Hill, NJ, United States), and several academic positions, he is now a full professor in Applied Probability at the University of Amsterdam. His research focuses on queueing theory and stochastic operations research, predominantly applied in the design of communication networks, but also in finance/risk, as well as in the production and service systems. He is the author of the recently published book 'Large Deviations for Gaussian Queues' (Wiley Online Library, 2007).

Yohan van der Bijl has worked for 10 years as a radiological technologist at the Deventer Hospital and University Medical Center Groningen. Trained by IBIS UvA in Lean Six Sigma methodology, he carried out and supported various process improvement projects in healthcare. Currently, he works as a Master Black Belt at the Deventer Hospital.

Copyright © 2011 John Wiley & Sons, Ltd.