



Achieving rapid and significant results in healthcare services by using the theory of constraints

Gustavo M. Bacelar-Silva, James F. Cox III & Pedro Rodrigues

To cite this article: Gustavo M. Bacelar-Silva, James F. Cox III & Pedro Rodrigues (2024) Achieving rapid and significant results in healthcare services by using the theory of constraints, Health Systems, 13:1, 48-61, DOI: [10.1080/20476965.2022.2115408](https://doi.org/10.1080/20476965.2022.2115408)

To link to this article: <https://doi.org/10.1080/20476965.2022.2115408>




View supplementary material 



Published online: 29 Aug 2022.



Submit your article to this journal 



Article views: 2496



View related articles 



View Crossmark data 



Citing articles: 2 View citing articles 

RESEARCH ARTICLE



Achieving rapid and significant results in healthcare services by using the theory of constraints

Gustavo M. Bacelar-Silva ^{a,b,c}, James F. Cox III ^d and Pedro Rodrigues ^{a,b}

^aDepartment of Community Medicine, Information and Health Decision Sciences, Faculty of Medicine (MEDCIDS-FMUP), University of Porto, Porto, Portugal; ^bCenter for Health Technology and Services Research (CINTESIS), Porto, Portugal; ^cDepartment of Distance Learning, Bahiana School of Medicine and Public Health, Salvador, Brazil; ^dManagement Department, Terry College of Business, University of Georgia, Athens, Georgia, USA

ABSTRACT

Lack of timeliness and capacity are seen as fundamental problems that jeopardise healthcare delivery systems everywhere. Many believe the shortage of medical providers is causing this timeliness problem. This action research presents how one doctor implemented the theory of constraints (TOC) to improve the throughput (quantity of patients treated) of his ophthalmology imaging practice by 64% in a few weeks with little to no expense. The five focusing steps (5FS) guided the TOC implementation – which included the drum-buffer-rope scheduling and buffer management – and occurred in a matter of days. The implementation provided significant bottom-line results almost immediately. This article explains each step of the 5FS in general terms followed by specific applications to healthcare services, as well as the detailed use in this action research. Although TOC successfully addressed the practice problems, this implementation was not sustained after the TOC champion left the organisation. However, this drawback provided valuable knowledge. The article provides insightful knowledge to help readers implement TOC in their environments to provide immediate and significant results at little to no expense.

ARTICLE HISTORY

Received 30 July 2021
Accepted 12 August 2022

KEYWORDS

Theory of constraints; healthcare services management; patient flow; process of ongoing improvement; ophthalmology

1. Introduction

Lack of timeliness and capacity are seen as fundamental problems that jeopardise healthcare delivery systems everywhere. When patients have concerns about their health, timeliness is paramount. However, patients frequently wait for weeks (or even months) for an appointment. Even when patients have an appointment scheduled, they still may suffer from the lack of timeliness in the treatment process. Once at the practice, patients may have to wait long past their designated appointment time to be with the provider.

Patients may suffer from a lack of timeliness derived from a couple of reasons: indirect wait time and direct wait time (Gupta & Denton, 2008). The indirect wait time is the time from when a patient requests the appointment till the designated time of the appointment. The direct wait time is the time from the designated appointment time until the provider sees the patient.

Long indirect wait times have psychological consequences (e.g., stress, desperation) and contribute to worsening medical conditions, even to death (Corley, 2016; Ryu & Lee, 2017). Patients may also have recovered (or in some settings, died) and no longer need the appointment. Treating patients having more advanced

conditions also has financial consequences, as it requires more specialised care resources, which increases the cost.

Merritt Hawkins published a study (Hawkins, 2017) about indirect wait times (time from calling for an appointment till the designated time of the appointment). This study considered the first available slot for a new patient appointment in the US. The reported indirect wait times were incredibly high: primary care physicians had a 54-day wait, cardiology 32 days, dermatology 35 days, obstetrics/gynaecology 23 days, and orthopaedic surgery 15 days. These long wait times are not unique to the US system. In Brazil, results from a recent survey (Conselho Federal de Medicina, 2018) reported the wait time as the most common cause of complaints for 61% of those patients who need surgery, for 56% of those patients who need an imaging exam, and for 55% of those who need an appointment in the Brazilian Unique Health System – Sistema Único de Saúde (SUS). This survey also identified that 45% of the participants had a 6-month wait for an appointment, an exam, or surgery, which is an increasing problem since this number was 29% four years earlier.

Long direct wait times cause patient frustration and may cause late arrivals or the patient not

showing up for future appointments. If patients expect a long direct wait for the provider, then many of them do not arrive on time – they arrive earlier (expecting the doctor may see them in case of another patient does not show up) or later (to make better use of their time other than be in a waiting room).

Long indirect wait times result from poor schedule design, while long direct wait times result from poor schedule execution (Cox, 2021; Cox & Boyd, 2020). Some patients may not show up (because they forgot, or possibly received treatment elsewhere). These situations jeopardise the system and put more pressure on its execution.

Many believe this timeliness problem is caused by a shortage of medical providers. A recent report released by the Association of American Medical Colleges (AAMC) provides a projection of physician shortages of between 54,000 and 139,000 by 2033 in the US (IHS Markit Ltd, 2020). This is comprised of shortages of primary care physicians (the gateway to the healthcare system in many countries) of between 21,400 and 55,200 physicians and in speciality care of between 33,700 and 86,700 physicians. Additionally, the AAMC study reports that other sources (unnamed) suggest that patients receive only 55% of the recommended chronic/preventive services. Authors claim that this shortfall between provided services and recommended services might be due to the time constraints faced by providers when attending to their patients. Additionally, physician demand would increase between 74,000 to 145,000 additional physicians if healthcare access was extended to underserved populations. Similar statistics are available in many other countries. The World Health Organization (2016) estimates a shortage of almost 18 million healthcare workers by 2030, especially in low-income and lower-middle-income countries. These estimates do not even consider an unusual situation, such as the COVID-19 pandemic crisis.

The most common (yet infeasible) solution suggested for this chronic problem is to hire more doctors and nurses, as supported by the timeliness and shortage statistics discussed previously. Based on the statistics, most managers consider this lack of provider resources as the core problem in healthcare environments. It seems to be logical: if we have more doctors, we can treat more patients. However, healthcare organisations cannot hire more provider resources because it would demand more investment and increase their operating expenses above their already too high current levels. It is truly a chronic conflict.

These statistics seem to reflect a significant shortage of physicians but there is an unspoken assumption in this argument. The unspoken assumption is that, currently, physicians are utilised

as effectively as possible. Is this reality? Or can we schedule physician appointments and execute physician schedules more effectively and, thus, create far more capacity without increasing expenses significantly? To schedule patient appointments more effectively, the provider must provide timely appointments for both acute and concerned patients needing treatment. To execute a schedule with more patient appointments without taxing the physician and clinical staff, one must streamline the patient-provider process (how the patient is treated from arriving till departing the provider's practice) by eliminating disruptions to provider utilisation and patient flow.

The direction of the solution for this chronic healthcare conflict may come from a disruptive management philosophy developed in the late 1970s – the theory of constraints (TOC). TOC considers an organisation as a system made of many interdependent resources that work collectively to achieve the organisation's goal. Since there is no such organisation capable of providing infinite throughput, at least one resource will limit the productivity of the whole organisation. This resource that limits the system is the constraint, it is the most important/valuable resource of any organisation because the constraint determines the organisation's overall performance (Goldratt & Cox, 2004).

Though Dr. Goldratt originally developed TOC to address manufacturing issues, this management philosophy has been effectively used to improve planning (scheduling) and execution in thousands of organisations including for-profit, not-for-profit, and government; service and manufacturing; and very small to Fortune 500 companies. Mabin and Balderstone (2000, 2003) provide the most thorough and comprehensive (but now dated) survey of TOC implementations simultaneously showing significant increases in organisation profits, decreases in lead times, improvements in due date performance, etc. Can TOC be used effectively to produce these same results in healthcare environments?

The purposes of this action research are three-fold:

- (1) Present the TOC methodology and describe how to implement it in a healthcare environment using the five focusing steps (5FS), drum-buffer-rope (DBR), and buffer management.
- (2) Describe and analyse the application of TOC in a healthcare environment providing an action research study of how an ophthalmologist led a TOC implementation to improve his imaging practice in a Brazilian eye hospital using existing resources.
- (3) Provide a direction for a solution to using this approach in more complex healthcare systems.

2. Methodology

This action research study describes and analyses a TOC implementation to improve an imaging practice in a Brazilian eye hospital. The data used in this study came from personal notes of the ophthalmologist (the first author) who implemented TOC in that environment. He collected data for 7 months (including before and after the implementation) from different sources, which included observation, informal interviews (assistants, nurses, and the Head of Retina), and scheduling template documents.

Although this implementation provided a significant improvement, it was not sustained. Therefore, we analysed the reasons why it was not sustained and included recommendations to ensure the sustainability of future TOC implementations.

Three TOC tools were implemented in this action research: the five focusing steps (5FS), drum-buffer-rope scheduling (DBR), and buffer management (BM). Each TOC method is briefly discussed as the implementation is described below. The supplementary file provides details about the TOC methodology.

2.1. TOC overview

The fundamental principle in TOC is that every organisation can be viewed as a system made of many interdependent resources. These resources must work together to achieve the system's goal. TOC acknowledges that one (or very few) resource(s) limits the overall performance of any organisation, otherwise its throughput would have no limits. The limiting resource is the constraint; therefore, it determines the performance of the whole system. For this reason, the constraint is the most relevant resource of any organisation (Goldratt, 1999; Goldratt & Cox, 2004).

In most cases, the initial (or current) constraint is not the strategic constraint (the leverage point of the organisation). Therefore, actions must be taken to move the constraint to its strategic location and increase the throughput of this resource (Ronen & Pass, 2021). Actions to increase productivity at the constraint include more effectively exploiting it and subordinating non-constraint resources to support the constraint.

Next, the organisation must plan and synchronise its productive flow based on the constraint and protect it from uncertainty to ensure maximum performance towards its goal. For example, (Goldratt & Cox, 1986, pp. 178–179), any minute lost at the organisation's constraint is a minute of throughput lost for the entire organisation. In contrast, a minute lost at a non-constraint resource will not dramatically affect the organisation throughput. In fact, non-constraint resources must have a protective capacity (some capacity above the capacity required by the constraint) to

ensure the continued productive flow after a stoppage/delay (Goldratt & Cox, 2004).

2.2. Five focusing steps

The five focusing steps (5FS) is a systematic process to provide focus and significantly improve performance using limited resources. This POOGI is typically adopted to improve the performance of physical constraints (usually equipment, but the constraint can also be a lack of people, a skill set, physical space, and material shortages). The 5FS are:

- (1) IDENTIFY the system's constraint(s).
- (2) Decide how to EXPLOIT the system's constraint(s).
- (3) SUBORDINATE everything else to the above decision.
- (4) ELEVATE the system's constraint(s).
- (5) WARNING! If in the previous steps a constraint has been broken, go back to step 1, but do not allow INERTIA to cause a system's constraint.

Since TOC is unfamiliar to most healthcare academics and practitioners, we provided below a brief overview of each step (the supplementary file contains a detailed description of the 5FS and its steps). Furthermore, in the results section, we provide a detailed description of the actions taken based on these steps. In applying these steps, one must first determine the organisation's goal and necessary conditions placed on the organisation's operations. Given these organisation parameters, the five steps must operate within this environment.

2.3. Drum-buffer-rope (DBR)

DBR is a TOC method developed by Goldratt in the 1980s, before the 5FS, for scheduling and managing the execution of operations when the constraint is an internal resource (Cox et al., 2012, p. 46). The drum (the constraint) is scheduled to process work in a specific sequence based on the negotiated delivery time between the customer and the scheduler and the finite capacity of the constraint. When the market demand is less than what the least capable resource of a system – also called capacity-constrained resource (CCR) – can produce, it means the constraint is in the market and the drum becomes the list of orders (Cox et al., 2012, p. 20; Cox & Spencer, 1997, Chapter 4; Goldratt, 2003, Chapter 15; Schragenheim, 2010; Schragenheim & Dettmer, 2000a, 2000b).

A buffer (of time or stock) protects the constraint from idleness caused by variability at upstream resources (Cox et al., 2012, p. 11; Cox & Spencer, 1997, Chapter 4; Goldratt & Cox, 1986, Chapter 49).

DBR usually has three buffers to protect production against uncertainty: (1) a shipping buffer, which is an estimate of the time to move from the constraint resource to the completion of the order; (2) a constraint buffer, which is the time from the release of raw materials to the constraint; and (3) an assembly buffer, this one is an estimation of the time of those parts that do not go through the constraint, considering from the release of raw materials to a process where those parts are assembled with other parts that came from the constraint (Cox et al., 2012, pp. 8; 28; 110; Cox & Spencer, 1997, Chapter 4; Goldratt & Cox, 1986, Chapters 49, Schragenheim, Dettmer, 2000b, 2000a, Chapter 7).

The rope mechanism is the communication that synchronises the release of the raw materials according to the constraint's consumption. This mechanism protects production preventing the release of raw materials into it at a pace faster than the constraint's consumption. Too much work-in-process (WIP) would lead to unproductive consequences, such as wandering bottlenecks, and considerably increase the lead time (Cox & Spencer, 1997, Chapter 4; Goldratt & Cox, 2004; Goldratt & Fox, 1986, Chapter 49; Schragenheim & Dettmer, 2000b). DBR provided the scheduling methodology for the outpatient schedule of this study but had to be modified for use with appointment times.

2.4. Buffer management

Buffer management is a mechanism used in both the planning and execution phases of TOC applications. It controls the constraint's protection against uncertainty based on the amount of time or stock (in healthcare, e.g., patients or beds) remaining until it is idle (Cox & Spencer, 1997, Chapter 4; Goldratt, 2006, Chapters 20,21; Goldratt & Cox, 2004; Stratton & Knight, 2010). Buffer management has four functions: (1) prioritise tasks according to buffer penetration, (2) identify tasks at risk of delay, (3) provide feedback, and (4) identify major causes of delay.

A buffer is commonly divided into three zones (green, yellow, and red), where each one usually represents 1/3 of the total buffer. A simple colour-coding system similar to traffic lights' colours is used to determine when to take action. Green means everything is running smoothly, continue what you are doing; yellow means an imminent threat to flow or constraint utilisation is approaching (plan accordingly to eliminate the threat); and if the buffer is red, the problem is imminent, address it (implement the plan previously made; Cox et al., 2012, pp. 11,14); black means the constraint is idle, the system is losing throughput. Buffer management works as a proactive control system that eliminates most disruptions to the constraint and patient flow (Cox et al., 2012, p. 11;14;

Cox & Spencer, 1997, p. 4; Goldratt & Cox, 2004; Stratton & Knight, 2010).

2.5. The environment

This study was conducted at an eye hospital located in Salvador (Bahia), a major city in Brazil with a population of approximately 3,000,000. The research site consisted of an imaging practice that performed 2 types of retinal imaging exams: fundus photography and fluorescein angiography (a sequence of fundus photographs after injecting a yellowish-coloured dye – fluorescein – in a vein). An ophthalmologist (the provider) performed and analysed the exams. An assistant supported the provider by performing four different tasks:

- (1) Prepping patients by instilling some eye drops to dilate their pupils.
- (2) Taking prepped patients to the exam room after the previous patient had left.
- (3) Providing patients' paper medical records.
- (4) Getting printed images and reports (in a different room), which were organised into a folder for each patient, brought to the provider for his signature, and finally delivered to the respective patient.

Additionally, to perform fluorescein angiography, a nurse supported the procedure by obtaining venous access, injecting the dye, and observing the patient for a possible adverse reaction while the provider examined the flow of the dye within the retina and recorded it. Figure 1 illustrates the schedule design (a) and schedule execution (b) before implementing TOC.

2.6. What to change?

Before the TOC implementation, the practice normally scheduled 12 regular patients per session and usually had an average of 2 appointment slots for acute patients (those seen by other ophthalmologists who requested a consultation for the same day). The provider was supposed to see the first patient at 14 h. However, patients were still arriving or having their pupils dilated at that time. That happened because the first patient was scheduled for 14 h and even when the patient arrived some minutes earlier, she still needed to have her pupils dilated, a procedure that takes about 20–30 minutes to complete. Since the schedule template consisted of 1 patient every 15 minutes, the second patient was scheduled for 14h15, thus the provider had to wait at least 20 minutes for the first patient. Figure 1.a illustrates the pre-implementation schedule design.

After finishing the first patient, it was common to have two or more patients prepped or being prepped

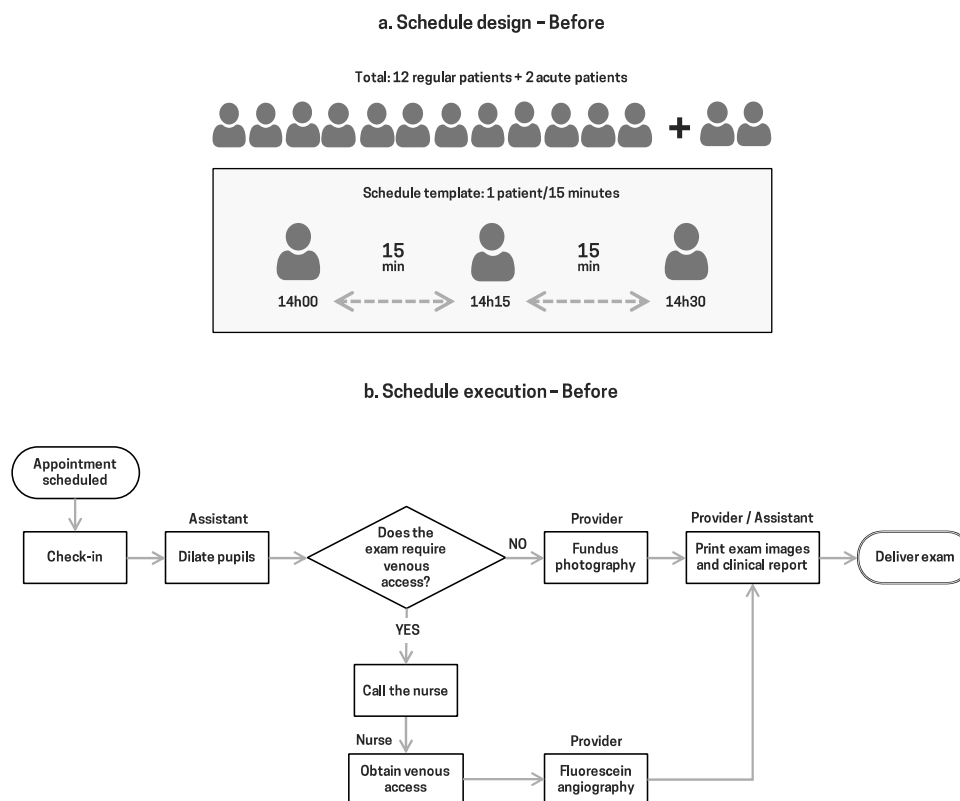


Figure 1. (a) Schedule design and (b) schedule execution before implementing TOC.

(including scheduled and extra patients). However, the provider was frequently idle waiting for the next patient. The assistant rarely noticed when the provider became idle, and the provider did not know who the next patient should be. He had to wait for the assistant to recognise that it was time to take the next patient to the exam room or go after her and ask for another patient. Additionally, when the provider had a prepped patient located in the exam area, the medical record was frequently not available, an example of an incomplete kit – a complete kit includes all the required items (e.g., documents, tools) to accomplish a process before starting the process (Leshno & Ronen, 2001). These situations extended the direct wait time of subsequent patients and extended the session's end.

Performing a fluorescein angiography was the most time-consuming procedure. Since the assistant was not a nurse, she had to call the operating theatre to request a nurse to assist in the procedure. Nevertheless, many times the assistant had taken the patient to the exam room before (or just after) asking for a nurse to support the procedure – another example of an incomplete kit. Usually, there was no nurse immediately available, but even when the nurse arrived soon after the call, the provider still had to wait while she obtained venous access, which could be a time-consuming task due to the patient's characteristics. Performing this exam not only delayed the provider but also delayed subsequent patients and extended the provider's treatment session. Again,

another common cause of extending patients' direct wait time and extending the end of the provider's session.

The increasing indirect and direct wait times reduced throughput and increased operating expenses, which jeopardised the hospital's goal of providing excellent and timely care while staying within its budget. Furthermore, the chaotic environment was jeopardising the two necessary conditions (a secure and satisfying environment for employees and excellent and timely healthcare). The provider, the assistant, and the nurse were always working under time pressures. They rushed to see patients on time and seldom accomplished it. More and more doctors were referring patients to the clinic service because this was the only local hospital that provided the exam result immediately after performing the exam, but the clinic's capacity was unable to meet the increasing demand. As the backlog was increasing fast, management knew it must act before it compromised the clinic's service.

3. Results

The time interval between the decision to change, analysing the environment, and starting the TOC implementation was only one weekend. The provider held a management background (MBA) and had sufficient knowledge about TOC based on

reading and studying several TOC books and articles. He examined the system and developed the implementation plan during the weekend. On Monday, the Head of Retina and Vitreous approved the provider's proposal, and the provider immediately started implementing the plan.

A TOC implementation begins by defining the system (its domain and boundaries), its goal, necessary conditions, and global measurements. Next, the 5FS guides the improvement efforts, and other tools can be added according to necessity, as explained below.

3.1. The domain, goal, and global measurements

The study domain was the ophthalmology scheduling and execution functions, as well as the physical structure available for the ophthalmology imaging exams. The goal of the organisation was “to provide excellent and timely care for an increasing number of patients within budget now and in the future”. There were also two necessary conditions to satisfy, but that had been jeopardised, which were (1) to “provide a secure and satisfying environment for employees now as well as in the future” and (2) to “meet customers’ (e.g., internal and external doctors, patients) needs of providing excellent and timely healthcare”. The hospital measured its performance based on the number of treated patients. Since the no-show and cancellation rates were considerably low, this action research measured performance improvement by the number of appointment slots available per session for regular and acute patients.

3.2. Applying the 5FS

The 5FS are applied where a physical constraint controls the system throughput. Since the patient backlog was increasing daily, and requests for last-minute appointments were becoming common, an internal constraint was suspected.

3.2.1. IDENTIFY the system's constraint(s)

The first step was to identify the constraint. The current system measures were based on local efficiencies: workers sequenced and performed their tasks based on their individual efficiency instead of sequencing their tasks with the focus of increasing patient flow and provider capacity and utilisation. In that chaotic environment, the constraint was wandering between the assistant, the provider, and the nurse. No list of priorities existed that the provider and the assistant could use to answer the question: “who is the next patient that I (e.g., provider, assistant, nurse) should attend to?”. After analysing the system, the provider concluded he was the strategic constraint – the scarcest resource based on the

difficulty to increase his throughput (Cox et al., 2012, p. 113). Therefore, all the efforts should focus on improving the patient flow to support the ophthalmologist.

3.2.2. Decide how to EXPLOIT the system's constraint(s)

After defining the provider as the strategic constraint, the decision of how to best utilise the provider's skill and time was made (step 2). This means the provider should never stop due to a lack of prepped patients, missing paper medical records, or waiting for a nurse to obtain venous access – the three most severe disruptions to provider utilisation and patient delays. Therefore, the provider redesigned the whole system with the ophthalmologist fully utilised performing only high-skill-level tasks.

To better exploit the provider, a buffer of fully prepped patients was placed in front of him to protect him from interruptions in the patient flow. This buffer consisted of three chairs located in a small waiting room within the exam room, which was already divided by a frosted glass wall. The buffer accommodated only the first 3 prepped patients from the queue (other patients remained in the regular waiting room). Most patients require between 2 and 5 minutes for a set of fundus photography and 10 to 12 minutes for fluorescein angiography. That means patients on the buffer chairs usually had to wait between 2 and 15 minutes. This number of chairs provided adequate time for the assistant to prep the other patients before filling the buffer – otherwise, the assistant would have to constantly check and transport patients to the exam room (or worst, the provider would leave the room to check who the next patient is and transport the patient himself), as described before the TOC implementation.

This new layout allowed both the provider and the assistant to observe the buffer penetration (number of empty chairs) and take action to fill the provider buffer with more patients. The buffer management rules were simple:

- When prepped patients occupied all three chairs, the buffer was full (green zone) and required no action.
- One empty chair meant a consumed buffer (yellow zone). The assistant should fill it in soon.
- Two empty chairs meant there was only one patient available, and the provider was at risk of being idled (red zone). The assistant had a few minutes to fill the buffer.
- When all three chairs were empty (black zone), the provider was at imminent risk of being idle, jeopardising the system's throughput. The assistant had to stop doing everything and fill the buffer immediately (at least one chair but preferably all).

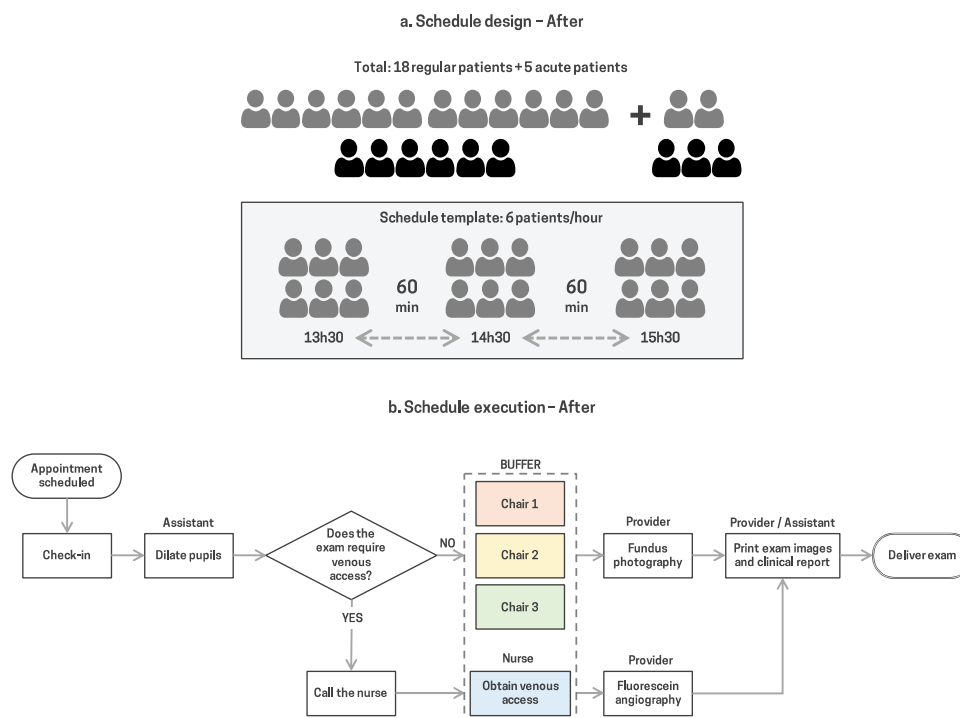


Figure 2. (a) Schedule design and (b) schedule execution after implementing TOC. Patients in black colour represent additional slots available after 4 weeks of TOC (a). A buffer located before the provider (strategic constraint) protects him from uncertainty and maintains patient flow (b).

After (and sometimes during) each session, the provider and the assistant discussed the reasons that led to severe buffer penetration (red or black regions) and how to address them. See Figure 2b.

To communicate the patient sequence to the provider, at the beginning of every session, the assistant printed a list of scheduled patients. She proactively marked when patients arrived, when they were prepped, when they had finished the exam, those who missed their appointments, and added acute patients to the list. In addition, she used this list to obtain the paper medical records in advance and arrange them sequentially for the provider as she filled the buffer (to form a complete kit).

The provider also changed the process to perform fluorescein angiography. Before this exam, the nurse prepped the patients on a specific chair in the small waiting room, where she obtained venous access (see, Figure 2b). The patient only moved to the exam room when fully prepped for that procedure (another complete kit).

These changes allowed the provider (and the assistant) to quickly and visually identify and answer the question “who is the next patient I should attend to?” and the assistant moved that patient to the exam room.

3.2.3. SUBORDINATE everything else to the above decision

Step 3 would ensure the provider would have no interruptions and, as an interesting consequence, alleviate the workload on the assistant.

The schedule template had to change to reflect the patient flow and the provider’s needs. The first change to the schedule template was on the appointments’ time. Traditionally, the schedule represents the appointment time as the moment providers should see their patients with no time offset for patient processing by the assistant. Usually, patients record that moment in their minds as their arrival time (some still consider arriving a few minutes earlier). However, patients need to go through the prepping processes (e.g., check-in, dilate pupils), which may consume a significant time. Even when patients arrive a few minutes earlier, the prepping processes still frequently jeopardise the appointment time. To address this timing problem, the appointment time became the moment patients were supposed to arrive at the practice considering all prepping processes (e.g., check-in, pupils’ dilation, obtaining venous access). This synchronisation of schedules meant that the patients should arrive 30 minutes prior to seeing the provider (the first patient’s appointment time is 13:30, while the provider starts processing fully prepped patients at 14:00).

The other two problems that influenced the schedule design were the statistical fluctuations on the arrival time and the prepping time. Most patients do not arrive on time, some arrive earlier, and some arrive late. The same is true for the time required to prep a patient – some patients (e.g., diabetic patients) require more time to have their pupils dilated than

others. When combined, these two problems cause a significant impact on the schedule execution reducing patient flow. Patients who arrived late and/or required more time to dilate their pupils often had to wait even more because they lost their place in the queue.

To address these three major problems, the provider developed a new scheduling template. Appointment times moved forward by 30 minutes considering the time dedicated to the prepping processes (mostly dilating pupils). To guarantee that there was at least one prepped patient when he arrived, the provider scheduled the first 5 patients for the same time (13h30). The provider continued scheduling groups of 5 patients every hour (a wave schedule) – another group of 5 patients arriving at 14h30, and the last group of 5 patients arriving at 15h30 (see, [Figure 2a](#)).

This research developed a new type of buffer in the schedule template: the arrival buffer. Usually, raw material is available in manufacturing, and one must choke the release to avoid unnecessary work-in-process. In the healthcare environment described here, the organisation could not adequately control when the patients arrived. Therefore, the provider scheduled patients in groups of five individuals, protecting the addition of new patients in the system from the uncertainty of arrival.

Having 5 patients scheduled for the first appointment slot may seem to be a questionable decision; after all, what if all 5 patients arrived at the same time? But what are the odds of that happening? Considering a 50% probability of arriving at a determined moment, the probability of all 5 patients arriving together is 3%. Even if that happened, they still need to go through the prepping processes, which take different times for each patient. Thus, scheduling 5 patients was a reasonable measure. Although the patients in the group had the same arrival time (which is different from the appointment time the provider is supposed to attend to the patient), patients arrived at different times (and further demanded different times to be prepped). The overall time difference between patients was not significant to the extent to generate wait complaints, it was quite the opposite, actually – patients spontaneously reported the whole process was faster than they expected.

The provider had sufficient intuition to know he was able to attend to 5 patients (or more) per hour if he did not have to stop doing his work to address the problems reported earlier. This TOC solution immediately increased the number of appointment slots available from 12 to 15 (a 25% increase). These changes kept the provider utilised at the beginning of the sessions and made patients flow through the system once they arrived, reducing individual direct wait times. In addition, buffer management would help to

Table 1. Summary of provider's schedule before and after implementing TOC.

	Scheduled patients	Last-minute patients	Total
Before TOC	12	2	14
After TOC	18	5	23
Difference	6 (50%)	3 (150%)	9 (64%)

assess whether scheduling 5 patients per hour was a good choice or not. If the number of patients waiting before the buffer was too high, then the provider would have to change the scheduling template. However, the reality proved quite the opposite, as later described.

At this moment, the implementation had all the elements to apply DBR and further improve the flow. In this environment, the provider was the drum (the resource that set the rhythm, the throughput) and the assistant would monitor and fill the buffer (and provide the complete kit) according to the drum's pace. The rope was the visual signal of empty chairs, which signalled that the assistant should fill the empty chair with a patient. Furthermore, fluorescein angiography patients moved to the exam room as soon as they were fully prepped by the nurse.

The solution worked immediately. As the days passed, the team progressed down the learning curve and was able to work more effectively, identifying and eliminating any remaining chaos. After a couple of weeks, the provider was frequently idle between patients. Therefore, the provider readjusted his appointment schedule to add more appointment slots, increasing from 2 to 5 acute patients (a 150% increase). The provider slotted in the acute patients as the last patients of each group (1–2 patients per hour), substituting a no-show, or an available slot. Four weeks after implementing TOC, the provider added 3 more appointment slots per session to his appointment schedule, reaching 18 appointment slots (a 50% increase). [Figure 2a](#) illustrates the schedule design after implementing TOC. Overall, the practice increased from attending to 14 patients (12 regular and 2 acute) to 23 patients (18 regular and 5 acute), which represents an addition of 9 patients (a 64% increase) seen using existing resources (at no extra cost), see, [Table 1](#).

3.2.4. *ELEVATE the system's constraint(s)*

At this point, the hospital had created a decisive competitive edge as the only eye clinic in the city that was able to provide same-day results for the exams and had appointment slots available for the short term. However, this competitive edge caused demand to increase, and the hospital decided to move to step 4 (elevate the system's constraint). This step required an investment: the practice bought new equipment to perform the retinal imaging exam on those simpler cases that did not require a doctor to perform.

A different assistant performed these exams, and a provider wrote the reports later.

3.2.5. WARNING! If in the previous steps a constraint has been broken, go back to step 1, but do not allow INERTIA to cause a system's constraint.

This implementation did not have the chance to follow this step. Some months later, the provider who led the TOC implementation left the practice (he moved to another country). Once he left, the rules and procedures developed during the implementation started to fall apart. There were no written TOC job descriptions or documentation, and no new TOC champion was trained to take the lead and sustain the improvement.

3.3. Summary of results

The benefits began immediately and involved patients, staff, and the ophthalmology clinic. Patients were flowing smoothly through the system. The changes reduced both indirect and direct wait times even though demand was increasing. Patients' delays were no longer an issue affecting the workflow, as well as the individual time to dilate pupils.

The provider significantly reduced idleness and was fully utilised to perform only high-skilled tasks and was able to focus on the current patient to provide excellent treatment without interruptions. The assistant kitted everything the provider needed to see the current patient. The allocation of prepped patients on the chairs in front of the exam room (a buffer) significantly prevented the assistant from multitasking and provided her sufficient time to successfully accomplish her job without being rushed. The chair dedicated to venous access virtually eliminated the delay caused by fluorescein angiography. The provider no longer had to wait for the nurse to arrive and prep the patients inside the exam room. Finally, both the assistant and the provider were not exhausted at the end of the session, which started to finish earlier than usual.

More importantly, the quality of care improved after implementing TOC. Since the staff virtually eliminated fighting fires, the provider and the assistant could dedicate more time to each patient, which enabled them to attend to more patients per session, reducing the backlog, and achieving a decisive competitive advantage for the clinic. Ophthalmology revenues increased rapidly and significantly.

4. Discussion

Despite over a thousand of articles written on scheduling and execution, most academic literature addressing healthcare problems and solutions uses mathematical optimisation methods and only focuses

on part of the problem (Ahmadi-Javid et al., 2017; Cayirli & Veral, 2003; Chen et al., 2018; Deceuninck et al., 2018; Fan et al., 2019; Mageshwari & Kanaga, 2012; Rais & Viana, 2011; Wu et al., 2013). Moreover, few of those solutions are implemented in the studied environment, with most of these models being unique to that environment and many not implemented in that environment nor implemented in other environments. In contrast, Lean and Six Sigma have a growing number of actual implementations in real environments but have had mixed results and require significant time and money investments (Chiarini & Bracci, 2013; D'Andreanmatteo et al., 2015; Hallam & Contreras, 2018; Henrique & Godinho Filho, 2020; Poksinska et al., 2017; Stanton et al., 2014).

While TOC has only a small number of implementations in healthcare, a recent study reported that those organisations that appropriately implemented TOC were able to achieve significant results almost immediately with no or negligible investments. (Bacelar-Silva et al., 2022). Note that more details are within the supplementary file to this article.

Why is reporting on a TOC implementation in one department in a large hospital significant? Goldratt's discussion of the complexity of systems provides insight into its importance, Goldratt stated:

It should be noted that the number of constraints does not determine the complexity of a system. Complexity is a result of the number of interactive constraints — constraints that impact each other. To better understand this statement, consider a system containing many constraints where none of them interact with each other. Such a system can readily be dissected into subsystems, each containing only one constraint. These subsystems are the simplest systems. Since the subsystems' constraints do not interact, the performance of the overall system is just the summation of the performance of all the subsystems. Thus, a system which has no interacting constraints is a very simple system, even if it contains a large number of constraints. (Goldratt, 1998, p. 1)

A hospital looks and is quite complex when one examines the many thousand possible patient process flows that exist in this environment (Knight, 2014b). However, many flows, particularly outpatient flows, exhibit the independent subsystem constraint structure described by Goldratt. These subsystems have little interaction with the major flows within the hospital environment. These subsystems can be viewed, analysed, and treated as independent patient-provider structures. Similarly, Mabin et al. (2018) provided an example of a TOC implementation for chemotherapy treatments and reported significant improvements — a reduction of 87% in average wait times and a two-thirds (67%) reduction in nursing staff overtime (and eliminated the need for overtime for all other staff).

Knight has significant TOC implementation experience in several hospital environments internationally. His novel, *Pride and Joy* (Knight, 2014a) is a consolidation of his knowledge based on these implementations but in conveying this message he has little “how-to” knowledge for one to go back to their environment and implement TOC. A second and quite different approach is provided by Strear and Sirias (2020) in their book, “Breaking the Bottleneck: Fixing patient flow for better care (and a better bottom line)” which provides an implementation approach starting at the emergency department (ED) and progressing from one bottleneck department to the next, etc. to improve patient flow and bottleneck capacity through the hospital system. This research provides a third approach to implementing TOC in a hospital environment. It is applicable to hospital departments that have little to no contact or interaction (the department is primarily an independent patient flow) with the remaining departments in the hospital.

A critical point in implementing the five focusing steps that frequently goes unnoticed is that, in implementing steps 2 (decide how to exploit) and 3 (subordinate), one is fundamentally changing the structure and mindset of work within the organisation. In traditional management, workers try to do their best to perform their specific tasks: perform and sequence the tasks to minimise their time and effort. In TOC the constraint offloads non-constraint level tasks to others, in our example, the provider performs only provider skill-level tasks. In TOC, the non-constraints do not try to do the best they can at their jobs but, in contrast, try to assist the constraint by having everything available to the constraint so he/she can perform their work, buffer the constraint from interruptions, etc. In our example, the staff ensures that patients flow smoothly to and from the provider in the patient-provider process, the provider is never idled by missing items or fully prepped patients, etc. These changes represent a paradigm shift in how work is structured and performed in organisations.

Time is a critical factor in the healthcare environment. Lack of timeliness affects not only patients but also the staff. When a lack of timeliness exists, the staff oscillate between idleness (and filling idleness with multitasking) and rushing to catch up, which increases their stress level and prevents them from properly following clinical guidelines to provide the best medical care.

The implementation described and analysed in this article enabled the clinical staff to significantly increase the number of patients treated using the same resources available and without overtime. Thus, one may question whether the increase in patients per hour would have jeopardised the quality of medical care, transforming it into a manufacturing process. The opposite is true, TOC enabled an increase in

the number of patients per hour and allowed the provider to dedicate more time to each patient. Although it may seem to be a contradiction, the answer is that TOC enabled the staff to better manage patient flow, significantly reducing the time wasted in firefighting. This new way of working involved fewer interruptions, therefore the quality of care naturally improved as the provider could dedicate more time to each patient. Besides increasing the quality, the provider could treat more patients per hour. In contrast to the traditional academic approach of devising highly sophisticated mathematical models to identify the “best compromise solution”, the TOC solution ultimately improved provider utilisation and capacity, timeliness (decreased indirect and direct wait time), quality of patient care, and staff satisfaction, and department revenues.

Similarly, Cox et al. (2016); (2014); (2012) implemented TOC in a large family medicine clinic in the USA and attained a 40% increase in patients attended (appointment slots increased 20%, no-show and late cancellation rates of 20+% reduced to 2%, and summer vacant appointment slots reduced from around 20% to 0%), added an average of 4 slots/provider for four providers (+\$400/provider/day), and providers were able to see all acute patients the same day they contacted the practice – being able to generate successively increases in revenue (21% in 2011, and 29% in 2012) during an economic recession.

4.1. Limitations and lessons learned

Still, some limitations exist. First, in this action research, the provider was the champion of this implementation effort and provided little to no education and training in the TOC tools to supervisors and subordinates in the department or other departments. Once the provider left hospital employment, the implementation deteriorated. This study did not have access to the details of what happened afterwards. Therefore, it is difficult to determine why diligence in following the steps fell away after the provider left, which consequently led to a drop in performance. However, we suppose that the local optima mindset and the lack of TOC education were determinants. This situation contrasts with Mabin et al. (2018), which describes a successful TOC implementation where the champion also left after some time. Despite the absence of TOC practitioners or champions, the organisation was able to sustain the changes and benefits eleven years after the TOC implementation.

To address these issues, the provider should have taken two actions. First, the provider should have trained other employees and supported them in implementing TOC in other areas across the hospital. Second, the provider should have documented the provider’s

and staff's new TOC job descriptions, as recommended by Cox (2021), Cox and Boyd (2020), and Mabin et al. (2018). If the provider had taken these two actions, the outcome may have been more favourable.

Deterioration of this ophthalmology exams performance was also accelerated by the purchase and introduction of new technology, a non-mydratic fundus photography device, which could be operated by a trained medical assistant but still required a doctor to report the findings (a task done afterwards, like other healthcare services in the city). This new equipment might warrant being the new strategic constraint (step 5 of the 5FS) and thus require performing the 5FS again, changing the policies, procedures, and rules for managing the constraint and patient flow. However, with the champion gone, an analysis was never undertaken.

In retrospect, the provider would have taken a different approach to sustain the improvement and further improve the gains. These actions would be related to providing TOC education and changing the local optima mindset. All those people involved in patient flow (e.g., schedulers, receptionists, assistants, nurses, and providers) would have to understand how local optima jeopardises the global optima and how to use TOC to develop improvement solutions by themselves.

Frequently, an unspoken assumption is blocking the ability to understand why an improvement is possible. The unspoken assumptions here concern the differences between traditional and TOC management philosophies, and they must be verbalised to fully understand why significant improvements are achieved when implementing TOC in a traditional management environment.

In traditional management environments, management tries to address the organisation's problems without considering the cause-effect relationships between them (assumption: problems can be addressed individually without understanding their causal relationships to each other). In addition to management solving these organisation's problems separately, they also consider any local improvement as a global improvement (assumption: a local impact translates directly to a global impact). In contrast, in the TOC environment, managers identify the causal relationships of problems and seek to address the core problem (assumption: there are relatively few causes that result in many undesirable effects – inherent simplicity – e.g., cause and effect exist in all environments and must be identified and addressed in fully understanding the environment; Goldratt & Goldratt-Ashlag, 2010). While not utilised in this research the third POOGI, the change question sequence provides a process for logically evaluating complex environments to identify the core problem, a win-win solution, and an implementation plan based on cause-and-effect logic.

Traditional management's major efforts are to save money everywhere, thus reducing the cost per product, e.g., cost per patient, in the healthcare environment. In contrast, TOC focuses on achieving the organisation's goal while meeting the necessary conditions of satisfied customers and satisfied employees – becoming an ever-flourishing company (Cox et al., 2012, p. 52). In healthcare, the goal is usually to provide excellent and timely healthcare within budget (non-profit organisations) or at a profit (for-profit organisations). To achieve the goal of any organisation, one must improve throughput (the number of goal units produced) at the constraint (e.g., the provider) and subordinate the utilisation of non-constraint resources to the constraint (e.g., patients' flow at non-constraints) – assumption: the constraint determines the throughput of the system.

The 5FS are based on cause-effect relationships and provide a sequence to immediately focus improvement efforts to achieve global impact. However, not following the sequence prescribed by the 5FS may explain why improvement initiatives that jump straight to step 4 (e.g., hiring more doctors and nurses) often fail to achieve expected results, as illustrated by Han et al. (2007) and why healthcare costs continue to rise.

Traditional managers strive for cost reduction everywhere (assumption: local cost reduction translates directly into organisation profits). Additionally, in Lean environments, managers reduce waste everywhere (assumption: reducing waste everywhere translates directly into organisation profits). In contrast, TOC managers focus on achieving the organisation's goal by improving throughput (the number of goal units produced) at the constraint. Since the first publication of the business novel *The Goal* (Goldratt & Cox, 1984), TOC has grown from a production to a holistic management philosophy, with numerous tools applicable in various functional areas to support ongoing organisational improvement. The supplementary file provides more details about TOC.

4.2. Future work

The TOC approach described in this article provides the direction of a solution to implementing TOC in complex healthcare environments. These complex healthcare systems comprise multiple subsystems, each containing only one constraint. Given this system and subsystems' characteristics, the system (e.g., a hospital) can be decomposed into simple, independent patient-provider flow where the constraints and near constraints have no interaction (e.g., the case study described in this article). Therefore, these subsystems can be analysed simply by applying the 5FS and BM. The remaining interactive constraint system should be approached using the Strear and Sirias

(2020) approach. Both our and the Strear and Sirias approaches should be refined based on future research. Furthermore, future work would include the implementation of TOC on a larger scale, including an environment with multiple independent and dependent providers. This research would include other relevant measurements, such as the impact of TOC on the patient waiting time (both direct and indirect).

5. Contributions

This article describes a TOC implementation providing the steps followed to successfully address the existing problems. Furthermore, this study is a clear demonstration that it is possible to achieve significant improvement in a few days using existing resources by applying TOC. Although the TOC implementation described in this article successfully addressed the practice problems, progress was not sustained. However, this drawback provided valuable knowledge to highlight the importance of TOC education and documentation of the new processes in the organisation as a fundamental step of any implementation.

The approach taken in this TOC implementation is simple, inexpensive, and easy to implement. The case study illustrates how to implement the 5FS, DBR, and buffer management, by providing examples to follow and avoid. Researchers, healthcare professionals, managers, and policymakers can use this article as a concise and actionable framework to develop a plan to significantly and rapidly improve healthcare services at little to no cost.

This paper introduces a new type of buffer, the arrival buffer. This new buffer protects the workflow against unexpected patient delays. In a manufacturing environment, one can easily control the release of raw materials by using DBR (or S-DBR). However, healthcare environments are different, patients (the raw materials) are not in a warehouse waiting to be released and processed. Healthcare services are subject to the uncertainty of delayed arrivals of their patients, and some services may benefit from protection against it.

Obviously, one cannot repeat all the actions taken in this TOC implementation in a different healthcare environment without making at least a few changes, e.g., adding, deleting, and modifying actions. However, the methods and tools provided here are generic enough to serve as a framework to support clinical and scheduling staff in a successful improvement initiative. The participation of those involved in the system will be paramount because they are familiar with the processes and have sufficient intuition to provide actions to support each of the 5 focusing steps.

The approach described in this article can be used to gain knowledge and experience in implementing TOC at independent subsystems (e.g., patient-provider structures) within large complex hospital systems. The described TOC techniques can be easily and quickly implemented, and provide significant improvements in organisation performance immediately, as described in many case studies (Bacelar-Silva et al., 2022). If similar healthcare organisations implemented these techniques, then global healthcare would improve almost immediately. We have ample healthcare providers in many countries; the problem is not a “shortage of providers”; the problem is the mismanagement of existing providers and not knowing how to effectively manage the constraint resource in a healthcare organisation.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the FCT – Fundação para a Ciência e a Tecnologia as part of a PhD scholarship [PD/BD/129829/2017]. This PhD scholarship was funded by the European Social Fund and national MCTES funds. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of this manuscript.

ORCID

Gustavo M. Bacelar-Silva  <http://orcid.org/0000-0002-0740-5532>

James F. Cox III  <http://orcid.org/0000-0002-0469-9731>

Pedro Rodrigues  <http://orcid.org/0000-0001-7867-6682>

References

- Ahmadi-Javid, A., Jalali, Z., & Klassen, K. J. (2017). Outpatient appointment systems in healthcare: A review of optimization studies. *European Journal of Operational Research*, 258(1), 3–34. <https://doi.org/10.1016/j.ejor.2016.06.064>
- Bacelar-Silva, G. M., Cox, J. F., & Rodrigues, P. P. (2022). Outcomes of managing healthcare services using the theory of constraints: A systematic review. *Health Systems*, 11(1), 1–16. <https://doi.org/10.1080/20476965.2020.1813056>
- Cayirli, T., & Veral, E. (2003). Outpatient scheduling in health care: A review of literature. *Production and Operations Management*, 12(4), 519–549. <https://doi.org/10.1111/j.1937-5956.2003.tb00218.x>
- Chen, Y., Kuo, Y.-H., Fan, P., & Balasubramanian, H. (2018). Appointment overbooking with different time slot structures. *Computers & Industrial Engineering*, 124, 237–248. <https://doi.org/10.1016/j.cie.2018.07.021>

- Chiarini, A., & Bracci, E. (2013). Implementing lean six sigma in healthcare: Issues from Italy. *Public Money & Management*, 33(5), 361–368. <https://doi.org/10.1080/09540962.2013.817126>
- Conselho Federal de Medicina. (2018). *Pesquisa Datafolha: Dobram queixas por tempo de espera*. Portal Médico. https://portal.cfm.org.br/index.php?option=com_content&view=article&id=27725:pesquisa
- Corley, J. (2016, May 21). *The Global Health Care Crisis No One Is Talking About*. HuffPost. https://www.huffpost.com/entry/the-global-health-care-cr_b_10074262
- Cox, J. F. (2021). Using the theory of constraints' processes of ongoing improvement to address the provider appointment scheduling system execution problem. *Health Systems*, 10(1), 41–72. <https://doi.org/10.1080/20476965.2019.1646105>
- Cox, J. F., & Boyd, L. H. (2020). Using the theory of constraints' processes of ongoing improvement to address the provider appointment scheduling system design problem. *Health Systems*, 9(2), 124–158. <https://doi.org/10.1080/20476965.2018.1471439>
- Cox, J. F., Boyd, L. H., Sullivan, T. T., Reid, R. A., & Cartier, B. (2012). *The theory of constraints international certification organization dictionary* (2nd ed.). McGraw-Hill Education. <http://www.tocico.org/?page=dictionary>
- Cox, J. F., & Robinson, T. M. (2012). The use of TOC in a medical appointment scheduling system for family practice. *TOCICO Conference 2012: 10th Annual Worldwide Gathering of TOC Professionals*, 10. <https://www.tocico.org/page/12ConfVid9>
- Cox, J. F., Robinson, T. M., & Maxwell, W. (2014). Applying the “theory of constraints” to solve your practice’s most vexing problem. *Family Practice Management*, 21(5), 18–22. <https://www.aafp.org/pubs/fpm/issues/2014/0900/p18.html>
- Cox, J. F., sIII, Robinson, T. M., & Maxwell, W. (2016). Unconstraining a doctor’s office. *Industrial Engineer*, 48(2), 28. <https://www.iise.org/industrialengineer/Details.aspx?id=40935>
- Cox, J. F., & Spencer, M. S. (1997). *The Constraints Management Handbook (The CRC Press Series on Constraints Management)*. CRC Press.
- D’Andreamatteo, A., Ianni, L., Lega, F., & Sargiacomo, M. (2015). Lean in healthcare: A comprehensive review. *Health Policy*, 119(9), 1197–1209. <https://doi.org/10.1016/j.healthpol.2015.02.002>
- Deceuninck, M., Fiems, D., & De Vuyst, S. (2018). Outpatient scheduling with unpunctual patients and no-shows. *European Journal of Operational Research*, 265(1), 195–207. <https://doi.org/10.1016/j.ejor.2017.07.006>
- Fan, X., Tang, J., Yan, C., Guo, H., & Cao, Z. (2019). Outpatient appointment scheduling problem considering patient selection behavior: Data modeling and simulation optimization. *Journal of Combinatorial Optimization* 42(4): 677–699. <https://doi.org/10.1007/s10878-019-00487-x>
- Goldratt, E. M. (1998). Chapter 5: How complex are our systems? In *Essays on the theory of constraints*. Great Barrington, MA: North River Press.
- Goldratt, E. M. (1999). *Theory of constraints: What is this thing called and how should it be implemented?* North River Press.
- Goldratt, E. M. (2003). *Production the TOC Way (with simulator)* (Revised ed.). The North River.
- Goldratt, E. M. (2006). *Haystack syndrome: Sifting information out of the data ocean*. North River Press.
- Goldratt, E. M., & Cox, J. (1984). *The goal: Excellence in manufacturing* (First ed.). North River Press Inc.
- Goldratt, E. M., & Cox, J. (1986). *The goal: A process of ongoing improvement* (Revised ed.). North River Press.
- Goldratt, E. M., & Cox, J. (2004). *The goal: A process of ongoing improvement—20th anniversary edition* (3rd ed.). North River Press.
- Goldratt, E. M., & Goldratt-Ashlag, E. (2010). *The choice (revised edition)*. North River Press.
- Gupta, D., & Denton, B. (2008). Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions*, 40(9), 800–819. <https://doi.org/10.1080/07408170802165880>
- Hallam, C. R. A., & Contreras, C. (2018). Lean healthcare: Scale, scope and sustainability. *International Journal of Health Care Quality Assurance*, 31(7), 684–696. <https://doi.org/10.1108/IJHCQA-02-2017-0023>
- Han, J. H., Zhou, C., France, D. J., Zhong, S., Jones, I., Storrow, A. B., & Aronsky, D. (2007). The effect of emergency department expansion on emergency department overcrowding. *Academic Emergency Medicine*, 14(4), 338–343. <https://doi.org/10.1197/j.aem.2006.12.005>
- Hawkins, M. (2017). *2017 Survey of physician appointment wait times*.
- Henrique, D. B., & Godinho Filho, M. (2020). A systematic literature review of empirical research in lean and six sigma in healthcare. *Total Quality Management & Business Excellence*, 31(3–4), 429–449. <https://doi.org/10.1080/14783363.2018.1429259>
- IHS Markit Ltd. (2020). *The complexities of physician supply and demand: Projections from 2018 to 2033*. Association of American Medical Colleges. <https://www.aamc.org/system/files/2020-06/stratcomm-aamc-physician-workforce-projections-june-2020.pdf>
- Knight, A. (2014a). *Pride and joy* (1st ed.). Linney Group Ltd.
- Knight, A. (2014b). Keynote address: Improving global healthcare with the theory of constraints. *TOCICO Conference 2014: 12th Annual Worldwide Gathering of TOC Professionals*, 12.
- Leshno, M., & Ronen, B. The complete kit concept: Implementation in the health care system. (2001). *Human Systems Management*, 20(4), 313–318. 313(6). <https://doi.org/10.3233/HSM-2001-20404>
- Mabin, V., & Balderstone, S. J. (2000). *The world of the theory of constraints: A review of international literature*. St. Lucie Press.
- Mabin, V., & Balderstone, S. J. (2003). The performance of the theory of constraints methodology: Analysis and discussion of successful TOC applications. *International Journal of Operations & Production Management*, 23(6), 568–595. <https://doi.org/10.1108/01443570310476636>
- Mabin, V., Yee, J., Babington, S., Caldwell, V., & Moore, R. (2018). Using the theory of constraints to resolve long-standing resource and service issues in a large public hospital. *Health Systems*, 7(3), 230–249. <https://doi.org/10.1080/20476965.2017.1403674>
- Mageshwari, G., & Kanaga, E. G. M. (2012). Literature review of patient scheduling techniques. *International Journal on Computer Science and Engineering*, 4(3), 397–401. <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.639.6978&rep=rep1&typepdf>
- Poksinska, B. B., Fialkowska-Filipek, M., & Engström, J. (2017). Does lean healthcare improve patient satisfaction? A mixed-method investigation into primary care. *BMJ Quality & Safety*, 26(2), 95. <https://doi.org/10.1136/bmjqs-2015-004290>

- Rais, A., & Viana, A. (2011). Operations Research in Healthcare: A survey. *International Transactions in Operational Research*, 18(1), 1–31. <https://doi.org/10.1111/j.1475-3995.2010.00767.x>
- Ronen, B., & Pass, S. (2021, June 22). Where Should the Constraint be and What to Do About it? TOCICO 2021 International Conference Video Proceedings. 19th Annual TOCICO International Conference, Online. <https://www.tocico.org/page/2021RonenPass>
- Ryu, J., & Lee, T. H. (2017). The waiting game—Why providers may fail to reduce wait times. *New England Journal of Medicine*, 376(24), 2309–2311. <https://doi.org/10.1056/NEJMp1704478>
- Schrageheim, E. (2010). *From DBR to Simplified-DBR for Make-to-Order (Chapter 9)*. J. F. Cox & J. G. Schleier Jr. Theory of Constraints Handbook. McGraw-Hill.
- Schrageheim, E., & Dettmer, H. W. (2000a). *Manufacturing at warp speed: Optimizing supply chain financial performance*. St. Lucie Press ; APICS.
- Schrageheim, E., & Dettmer, H. W. (2000b). Simplified Drum-Buffer-Rope – a Whole System Approach to High Velocity Manufacturing (Goal Systems International). <https://www.goalsys.com/books/documents/S-DBRPaper.pdf>
- Stanton, P., Gough, R., Ballardie, R., Bartram, T., Bamber, G. J., & Sohal, A. (2014). Implementing lean management/six sigma in hospitals: Beyond empowerment or work intensification? *The International Journal of Human Resource Management*, 25(21), 2926–2940. <https://doi.org/10.1080/09585192.2014.963138>
- Stratton, R., & Knight, A. (2010). Managing patient flow using time buffers. *Journal of Manufacturing Technology Management*, 21(4), 484–498.
- Strear, C., & Sirias, D. (2020). *Smash the bottleneck: Fixing patient flow for better care (and a better bottom line)*. Health Administration Press.
- World Health Organization. (2016). *Global strategy on human resources for health: Workforce 2030*. WHO Press. <https://apps.who.int/iris/bitstream/handle/10665/250368/9789241511131-eng.pdf?sequence=1>
- Wu, X. D., Khasawneh, M. T., Hao, J., & Gao, Z. T. (2013). Outpatient scheduling in highly constrained environments: A literature review, 19th International Conference on Industrial Engineering and Engineering Management: Assistive Technology of Industrial Engineering, 14 June 2013. In Qi, E., Shen, J., Dou, (eds). Springer, Berlin, Heidelberg. p.1203–1213. doi:10.1007/978-3-642-38391-5_127