Operations Research Perspectives 6 (2019) 100110

Contents lists available at ScienceDirect

Operations Research Perspectives

journal homepage: www.elsevier.com/locate/orp

Integrating lean thinking and mathematical optimization: A case study in appointment scheduling of hematological treatments

Alessandro Agnetis[*,a], Caterina Bianciardi[b], Nicola Iasparra[c]

[a] Dipartimento di Ingegneria dell'Informazione e Scienze Matematiche, University of Siena, Via Roma 56, Siena 53100, Italy
[b] Azienda Ospedaliera Universitaria Senese, Siena, Italy
[c] Reply s.r.l., Milano, Italy

ARTICLE INFO

Keywords:
Appointment scheduling
Hematological treatments
Lean thinking

ABSTRACT

This paper addresses the relationship between *lean thinking* and *mathematical optimization*. We discuss the roles of the two approaches, using as a reference case study the appointment scheduling process in a hematological center of a large Italian hospital. We report on how lean tools have been deployed to improve the process, we present a mathematical optimization model and discuss its implementation. Our aim is to show that the joint use of lean thinking and mathematical optimization can disclose large benefits when they are properly integrated in the improvement process. In our case study, simulated experiments point out that the average patient lead time could be decreased by more than 30%.

1. Introduction

Lean thinking (LT) [1] is a managerial philosophy focusing on enhancing efficiency and reducing waste. LT provides a wide scope of tools and methods to carry out an improvement process, including problem framing, people commitment, goal setting, improvement action design and implementation [2,3]. Most of these tools do not require much quantitative elaboration to be laid out. In order to start a lean project, the main requirements include having a clear knowledge of the system organization, the goodwill of facing the problem with a fresh approach, pooling together the individuals' experiences (disregarding hierarchical bounds and biases), and even having the humility of acknowledging the existence of some problems when quantitative data point them out [4].

Indeed, management engineers criticize the fact that when faced with a complex, possibly blurry problem, mathematical analysts come up with huge, complex and mysterious models in which unrealistic assumptions on problem features and the role of the decision maker are made [5,6]. Reading the vast literature on lean implementations and achievements in manufacturing and service companies, one can get the feeling that most problems can be solved with no need of devising complex optimization models [2]. We do not agree with this view. Our point is that LT and MO are two complementary disciplines, having distinct goals and different approaches, but their combined approach – along with other quantitative modeling tools, such as simulation – can disclose huge benefits, not achievable using one approach alone.

As lean intervention typically results in a simpler, smoother and stabler process [7], new improvement opportunities arise. After all, "creating initial process stability" (see [8], Chapter 4) is a fundamental starting point of any lean intervention. These opportunities can be seized by means of *a suitable MO model*, allowing the manager to devise the most appropriate quantitative decisions for driving the system towards the overall goals. In this paper we devote special attention to the latter point, i.e., how an optimization model can enable achieving significant benefits in the context of a lean improvement process. We illustrate our view through a case study concerning the appointment scheduling process in the hematological ward of a large Italian hospital.

The plan of the paper is as follows. In Section 2 we first provide a quick review of LT ideas (Section 2.1) and we briefly assess the literature on optimization models for appointment scheduling in similar environments, i.e., hematological and chemotherapeutic centers (Section 2.2). Then we state the purpose of the study and the research approach we undertook (Section 3). Section 4 contains the case study. After describing the application setting (Section 4.1), in the subsequent sections we illustrate the reengineering process following the typical *plan-do-check-act* (PDCA) paradigm. So, we subsequently focus on data analysis and simulation (*plan*, Section 4.2), process reengineering, which includes devising an optimization model (*do*, Sections 4.3 and 4.4), model results (*check*, Section 4.5) and some implementation issues (*act*, Section 4.6). Finally, some conclusions are drawn in Section 5.

* Corresponding author.
E-mail addresses: agnetis@diism.unisi.it (A. Agnetis), c.bianciardi@ao-siena.toscana.it (C. Bianciardi), iasparra@student.unisi.it (N. Iasparra).

https://doi.org/10.1016/j.orp.2019.100110
Received 13 August 2018; Received in revised form 3 April 2019; Accepted 3 April 2019
Available online 04 April 2019

2. Literature review

2.1. Lean thinking

Lean thinking can be viewed as a process management philosophy which focuses on delivering *value* to the customer, eliminating *waste* in all its forms. Value in a process is defined as the organization's capability to deliver what the customer needs. Lean was originally developed in the automotive industry, namely the Toyota Production System, and it has been divulgated in the western world in the early 1990s (a major role having been played by a famous book by Womack et al. [1]). LT provides a systematic view to processes, and nowadays it is applied not only in manufacturing companies, but also in service industries, including healthcare. LT is based on five simple principles [9]:

1 Define what is *value* to the customer.
2 Identify the *Value Stream Map*, a diagram showing how the value flows through the process, and detect waste in all its forms.
3 Create a continuous flow. The process should flow smoothly and without delays or interruptions.
4 Implement a pull system: the production is driven by customer needs.
5 Pursue perfection: lean implementation is a cultural challenge and not just a single project. It requires continuous effort and broad application throughout the organization.

Despite its origins, Womack et al. [1] claimed that LT could be applied to the healthcare sector as well, in view of its general principles. Many success stories have been reported concerning both private and public healthcare providers. Among the others, Koning et al. [10] show that LT can help healthcare industries in improving quality and costs. Other benefits to various stakeholders are reported by the Institute of Healthcare Improvement [7] and by several other studies such as Fine et al. [11] (improved patient experience, resource efficiency), Mazzocato et al. [12] (improved process understanding and reliability), Ford et al. [13] (improved clinical outcomes). Very significant examples of lean implementations include Virginia Mason Medical Center [14] and Theda Care [15].

However, not all researchers agree that LT is the answer to all problems [16,17]. A criticism is that lean projects often focus on efficiency and cost-effectiveness, less on patient satisfaction [18]. Moreover, the success of a lean intervention may be affected by several issues, including change management and the degree of integration among various lean interventions carried out in parallel [19]. Some studies argue that for a LT approach to be successful, besides the correct application of lean concepts and tools, it is necessary to implement lean as a managerial "culture" including both operational and sociotechnical aspects [20,21]. While one may expect that the debate on lean effectiveness and its possible limits will continue, we want to focus on the specific aspect of the integration between LT and MO, and how it may affect the impact of an improvement process.

Operations research provides analytical methods (such as optimization and simulation models) to represent a real-life setting. Such models are typically designed by scientists who are expert of mathematical methods, but who often have little knowledge of specific real-life problems in which their models may be applied (this is regarded by Sodhi and Tang [5] as a weakness of some OR/MS research). On the other hand, LT emphasizes the importance of regularly visiting the workplace in order to have a direct acquaintance of the actual practices (*Gemba Walk* [22]). So, the correct integration of LT and MO has the potential to provide a wider view of an healthcare service. In particular, to achieve a successful application, the challenge is to ensure an adequate engagement of all stakeholders involved from the very beginning, pursuing skill integration between mathematical analysts and healthcare professionals. In fact, the lean philosophy underlines the need to create such a multidisciplinary team [23].

2.2. Scheduling hematological treatments

Successful applications of LT in hematology are reported in the literature, but these typically focus on its most qualitative aspects [24,25]. Here we want to focus on the improvement process of outpatient appointment scheduling, a problem involving sophisticated quantitative decisions.

Appointment scheduling is a major organizational issue in healthcare delivery, and it can be addressed at various levels. Some authors distinguish between *planning* and *scheduling* levels, where typically the former refers to deciding the days in which each patient should receive treatments (under some aggregate capacity constraint), while the latter deals with the detailed schedule of patients' treatments in a single day [26]. The problem we address in this paper is at the scheduling level.

As observed by Turkcan et al. [26], hematology and chemotherapy administration has progressively shifted from the inpatient to the outpatient setting, with consequent higher patient comfort and cost savings for the hospital. However, this requires the use of accurate scheduling tools, accounting for the specificity of hematological treatments. In particular, if the ward delivers various services, as typical of hematological centers (less so in chemotherapeutic centers), different patients may need to follow different paths throughout the center, and this can make detailed scheduling particularly complex. Mustafee et al. [27] advocate the use of simulation tools to actually deal with the complex patient flows of a hematological outpatient clinic, though they claim that if the model is too detailed, its reusability can be compromised. For this reason, their model does not consider such details as the processing times of each step. Santibanez et al. [28] also use simulation models to validate the use of a dispatching rule to schedule appointments based on their duration and variability, while Wijewickrama and Takakuwa [29] resort to simulation to evaluate various appointment systems, and conclude that performance can largely benefit from adjusting pre-allocated time slots on the basis of patient individual characteristics.

On the other hand, a few studies concern the use of mathematical programming models for appointment scheduling, as we do in this paper. Santibanez [30] uses MILP for chemotherapy outpatient scheduling, considering a single treatment stage of the problem, and focusing on balancing the workload among the nurses. Liang et al. [31] propose an ILP to schedule daily patients accounting for various resource constraints. They assume that the patient mix is known in advance and give appointments accordingly. Le et al. [32] devised a single model to schedule in detail a period of various weeks, solved by a metaheuristic, while Hesaraki et al. [33] use a MO model to create a template for detailed infusion scheduling in a chemotherapeutic center, using a convex combination of makespan and total flow time as objective. All the last three papers assume that each treatment consists of single infusions, hence disregarding other steps through the ward. Hahn-Goldberg et al. [34] consider an incremental approach in which chemotherapy appointments are given upon request, which is close to the setting analyzed in this paper. In the process described, on the basis of a forecasted patient mix, an operator manually allocates time slots to patients, and only when the operator is stuck an optimization routine is run to reschedule the appointments which have not been agreed with the patients. This is also similar to our model, however again the complexity of individual patient treatments is not taken into consideration. Finally, Lamé et al. [35] thoroughly review existing approaches to chemotherapy scheduling models, focusing on the lack of models addressing the coordination between administration center and pharmacy, an increasingly important issue.

3. Methods

The purpose of this study is to demonstrate the possibilities of process improvement using a combination of LT and MO. Such a purpose can be also expressed as an empirical contribution to the investigation of the concept of *multimethodology* (see Mingers and

Brocklesby [36]), i.e., the combination of different methodologies, possibly from different paradigms, to address real-world challenges. Defining a methodology as "a structured set of guidelines or activities to assist people in undertaking research or intervention", we can view LT as a methodology (for problem framing and solving), while MO is usually viewed as a *technique*, i.e., again referring to the terminology in [36], a specific activity serving a purpose in the context of a *methodology* – in this case, the methodology is LT and the purpose is the optimization of the redesigned process. Hence, the combination of LT and MO is eligible for attaining the potential connected with a multi-methodological endeavour, i.e., to "deal comprehensively with a particular intervention" better than what might be attained by LT or MO separately. Moreover, in our study we also make use of *simulation modeling*, as a *technique* which serves validation purposes. While simulation has already been (successfully) integrated in lean healthcare contexts [37,38], less evidence exists in the literature of a similar operation involving MO. However, with reference to the *matrix for multimethodology design* introduced in [36], one may expect that potential exists, as MO especially focuses on *assessment* and *action* in the material domain, while LT is a broad methodology permeating also social and personal "worlds". In fact, LT is often perceived as a methodology for addressing a wide spectrum of activities [12], ranging from problem appreciation to final action. (Incidentally, though the two cycles are not perfectly overlapping, one may argue that there is a close relationship between the *appreciation-analysis-assessment-action* activity classification in [36] and the *plan-do-check-act* cycle).

The logic development of the remainder of the paper can be outlined as follows.

1. After describing the specific problem at hand, we illustrate the improvement process in terms of the *plan-do-check-act* cycle, emphasizing the application of various lean concepts or techniques such as multidisciplinary brainstorming sessions, value stream mapping, streamlining patient journeys, use of simulation, root-cause analysis, *pull* system implementation.
2. We refer to the above scheme to specify the point of the process where mathematical optimization enters the picture (namely, within the *do* phase) and its role. We discuss the conditions under which one can expect the mathematical model to be profitable, namely only after the process has been redesigned to remove causes of waste (such as congestion or variability).
3. Once the whole improvement process has been defined, we focus on system performance, i.e., discuss the impact of process re-engineering on the main KPIs of the system (*check* phase).

A specific collaboration between the hospital and the University of Siena has been activated for this project. More specifically, the hospital participated through its *lean group* (regarded by some researchers as mandatory for the development of a lean project in healthcare [39]). The lean group has been set up to support improvement projects, and it is formed by 6 full-time personnel units having different backgrounds and skills, namely management engineers, physicians and nurses. In the project, the lean group provided specific expertise on the process and the organizational issues, while the University mainly contributed to design and development of the mathematical models and tools.

We view the project as an example of action research [40]. In fact, on the one hand we focused on solving a specific problem (improving patient flow in a chemotherapeutic center), and on the other hand we wanted to evaluate the benefits accruing from the integration between LT and MO, as detailed in Section 5, in order to produce evidence that such a combined approach should be pursued in other similar situations. In fact, the size of the ward and the problems encountered are typical of many medium-size hospitals and hence potentially interesting to researchers and to a large number of operators.

4. Case study and results

4.1. The ward and the problem

In this study we consider an improvement process for the hematology ward of the Policlinico Santa Maria alle Scotte of Siena. The hematology ward deals with the diagnosis and treatment of blood diseases, mainly neoplasms. The average number of yearly admissions is around 10000, for various treatments. The ward operates every day except holidays from 8 a.m. to 3 p.m.

A field survey (administered to 165 patients in 2017, see below) prompted the hospital management to take action to improve patient satisfaction. In fact, top management sponsorship is often a fundamental factor for lean success [41]. The field survey pointed out that the most critical aspects perceived by patients were *environmental comfort* and *long waiting times*, particularly fit for being addressed by a lean approach [12]. We especially focused on the latter issue, aiming at reducing waiting times through a review of appointment scheduling procedures. One should notice however that also environmental comfort would expectedly benefit from shorter patient lead times. In particular, a crowded waiting room forces the front-desk staff to make frequent on-the-spot patient scheduling decisions. The need for such stressful decisions decreases if patient lead times are shortened.

A maximum number of patients per day is specified for each possible *therapy type*, based on past experience and the rough-cut production capacity of the ward. Considering that overbooking is allowed, the number of daily booked patients could occasionally reach 50. Table 1 reports the 7 therapy types and how many patients were admitted for each type. Once a patient is admitted, he/she undergoes a sequence of steps or *activities*, depending on the patient's specific medical needs. Patient admissions are concentrated in two blocks, at 8 a.m. and 11 a.m.. Since appointments are given only on the basis of therapy types, no difference is made between first-time visits and returning visits, even though different resources are required in the two cases. Since April 2017, the appointment scheduling procedure has been supported by a computerized system. This has brought many benefits, such as better data storage, but the therapy type-based organization of the process has remained unchanged.

4.2. Process analysis (plan)

The first phase of the overall improvement process is the *plan* phase of the PDCA cycle, starting with a quantitative analysis of the current process. Hence, after problem assessment, a phase of data collection followed.

4.2.1. Data collection

The lean group started by monitoring the process and collecting relevant data by direct observation, with the active collaboration of the healthcare professionals of the ward. In view of the relatively small seasonal variability of the number of patients treated, one month of data collection was deemed sufficient to have a representative sample of data.

Table 1
Therapy types and maximum daily admissions. BMB stands for *bone marrow biopsy*.

Therapy types	Daily admissions
A) Blood sampling and visit	15
B) BMB	3
C) Transfusion	3
D) Monoclonal antibodies	4
E) Intravenous chemotherapy	8
F) Subcutaneous chemotherapy	9
G) Inpatients	5

Table 2
Processing time distributions of various activities (except infusions).

Activity	Distribution
Admission	TRIA(3,11,13)
Blood sampling	TRIA(1,4,8)
After-sampling wait	CONST(90)
Visit	TRIA(3, 20, 40)
BMB	TRIA(3, 19, 35)
Subcutaneous	TRIA(2, 10, 27)

The following types of data were collected.

- *Duration of the activities of the ward*. While the duration of infusions depends on the specific therapy of the individual patient, the duration of all other activities does not depend on the specific patient. Standard statistical analysis showed that the durations of most activities are accurately expressed through triangular distributions, as typical of many situations in which the number of samples is relatively limited [42]. The results are summarized in Table 2. The values reported also include a switching time between one patient and the next.
- *Number of patients for each therapy type*. In the current situation, a maximum number of patients for each therapy type is admitted, as shown in Table 1.
- *Duration of infusive therapies*. For the patients undergoing infusion therapies, the length of stay in the infusion room depends on the individual therapeutic protocol. The infusion time is known at the time of appointment booking, and it is highly deterministic. In the sample day, such durations range from a minimum of 15 to a maximum of 195 min.
- *Resources*. The staff is spread throughout the ward. At patient reception there are 2 nurses until 9:15, and then only one. Blood sampling and subcutaneous therapies are carried out by one nurse for the whole day. There are three physicians for visits and bone marrow biopsy (BMB), but no more than one biopsy can be carried out at a given moment. In the infusion room there are 2 nurses until 9:15 a.m. and 3 nurses afterwards. Patients are accommodated on seven armchairs and one bed. The latter is reserved for patients undergoing monoclonal antibody therapy for the first time, but if there are no such patients, the bed is used for other patients.

As we mentioned in Section 4.1, the field survey showed that patients put significant value in having short lead times. Data collection pointed out that, besides being large on the average, patient waiting times were also highly variable. From the viewpoint of the patients' experience, unpredictable and long lead times are obviously annoying and stressful, and determine general confusion in patients' and personnel flows, as well as resource over- and under-utilization. Long and variable waiting times are among the most common causes of waste in healthcare settings [3,43].

4.2.2. Therapy types vs. paths

As we already observed, one problem with the organization based on therapy types is that even patients requiring the same therapy type may load the system resources in a different way. This aspect had to be appropriately assessed in order to design a realistic simulation of the system. Hence, a fundamental phase was to figure out the set of paths followed by the patients through the ward. All possible treatment sequences required by the patients were carefully tracked down and it was possible to recognize seven distinct *paths*, each defined by a (fixed) sequence of *activities* for each patient (Table 3). For each path, a value stream map has been drawn to assess the process and share knowledge of the problems between the lean group and the healthcare professionals of the ward. Fig. 1 depicts the VSM of path 1 (*blood sampling and visit*), in which the range of observed waiting times is reported. Notice that there is a correspondence between *groups* of therapy types and *groups* of paths, not a one-to-one correspondence between types and paths.

Table 3
Therapy types and paths.

Therapy types	Paths
A) Blood sampling and visit	1. *Blood sampling and visit path*: admission → blood sampling → visit
B) BMB	2. *BMB path*: admission → blood sampling → BMB
C) Transfusion *D*) Monoclonal antibodies	3. *First-time infusions path*: admission → blood sampling → visit → infusion
E) Intravenous chemotherapy	4. *Returning infusion path*: infusion (rarely also → visit)
F) Subcutaneous chemotherapy	5. *First-time subcutaneous path*: admission → blood sampling → visit → subcutaneous therapy 6. *Returning subcutaneous path*: admission → subcutaneous therapy
G) Inpatients	7. *Inpatient path*: visit → admission → blood sampling (rarely also → BMB)

The patients admitted in a sample day were mapped onto paths, obtaining the path mix indicated in Table 4.

4.2.3. As-is system simulation

After collecting all the data, a simulation model of the ward was built using discrete-event simulation software (ARENA). The time horizon on which the simulation is based is 7 h, i.e., the daytime activity of the ward. Arrivals are assumed to be concentrated at appointment times for each therapy type, which is a fairly accurate assumption. Each run simulates the 49 patients of the path mix in Table 4. Runs differ from each other in the duration of the various activities.

For our purposes, the most significant performance indicators are the patients lead time, value-added time (VA time) and percentage of waiting time over the entire lead time. Table 5 reports the average values of such figures (minutes) over 100 simulation runs. The table also reports the average value (over 100 runs) of the maximum lead time experienced by a patient, showing indeed a significant variability.

The results show that patients spend a lot of time waiting. For example, for path 1 (blood sampling and medical visit), an average of 218.52-125.68 = 92.84 min are spent in queue, and 156.67 min for path 5 (patients undergoing subcutaneous therapy for the first time). These values are coherent with the empirical observations carried out during the data collection phase.

4.3. Process reengineering (do)

Following the PDCA cycle, once the problem has been assessed and figured out, the lean group started investigating possible causes. To this aim, *Root Cause Analysis* (RCA) was used. Applying RCA is often an important part of an improvement project. In our study, the RCA method known as *5 Whys* was employed. It helps determining the cause-effect relationships in a problem and can be used whenever the real cause of a problem or situation is not apparent. The 5 Whys method mainly consists of repeatedly (up to five levels) questioning the causes of the observed events, in order to get to the root of the problem. This brainstorming process entails first eliciting various sensible answers to the root question, and then to questions successively generated, so that a *tree* of answers and questions is generated, the leaves of which are then debated and discussed throughout the group. In the end, only the

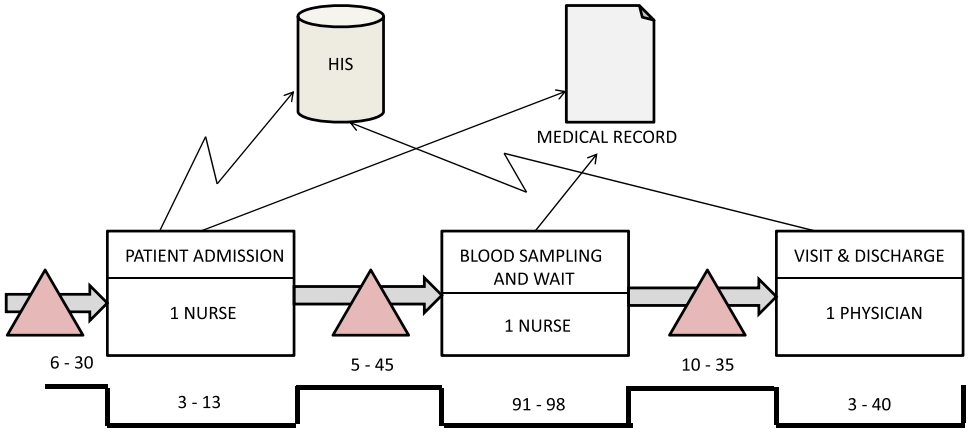

Fig. 1. VSM of the path *Blood sampling and visit*.

Table 4
Path mix in the sample day.

Path	No. of patients
1) Blood sampling and visit	16
2) BMB	3
3) First-time infusion	4
4) Returning infusion	12
5) First-time subcutaneous	3
6) Returning subcutaneous	6
7) Inpatients	5

Table 5
As-is simulation results. All figures are an average over 100 simulation runs. All times are expressed in minutes.

Path	VA time (average)	Total lead time (average)	Wait(%) (average)	Total lead time (maximum)
1	125.68	218.52	42.49%	304.14
2	122.49	138.15	11.34%	151.62
3	154.3	225.98	31.72%	311.08
4	93.75	102.14	8.21%	198.75
5	138.74	295.42	53.03%	352.39
6	22.04	92.59	76.19%	120.8
7	43.48	93.05	53.27%	129.43

most likely systemic cause is selected [44]. As typical of the application of this technique, all members of the lean group contributed to the process. In our case, the outcome of the process was the following chain of questions and answers.

- Q: "Why do patients wait so much?"
  A: "Because resources are not sufficient to let the patients go through"
- Q: "Why are resources all occupied?"
  A: "Because no careful treatment planning has been done"
- Q: "Why don't we plan individual treatments?"
  A: "Because patient arrival times cannot be predicted."
- Q: "Why is it so?"
  A: "Because patient arrivals are concentrated in two time blocks (8 a.m. and 11 a.m.), with no individual appointment."

In conclusion, the analysis revealed that in order to decrease patients' waiting time, it is necessary to have a precise *plan of the daily treatments*, for which a suitable information system can be very useful. To this aim, the access discipline of the patients should be radically changed. More precisely, the current appointment process proceeds in a *push* fashion, as follows.

(*i*) A patient calls the ward, communicating the treatment requested for a certain day;
(*ii*) The operator communicates the time block in which patients requiring the treatment should arrive;
(*iii*) Patients show up during the time block, and are admitted in order of arrival.

For a detailed plan to be worked out, the appointment scheduling process has to be reviewed. Defining countermeasures is the core of the *do* phase in the PDCA cycle, and required a major effort on the lean team. In accordance with lean principles, in order to avoid congestion and make resource utilization smoother over time, the group proposed to change the appointment scheduling policy from *push* to *pull*. In our case this can be attained by means of two distinct, yet complementary actions:

(*i*) Instead of assigning patients to one of two time blocks, each patient is given an *individual appointment*. In this way, patient arrivals can be spread across opening hours.
(*ii*) Action (*i*) is not enough to guarantee a smooth flow, since one might continue giving individual appointments until capacity is congested. The key issue is to account for the *individual requirements* of each patient, in order to anticipate the impact on ward capacity, and determine the appointment accordingly. To this aim the *path* (not just the therapy type) of the patient through the ward should be considered, as well as the expected resource load situation when the patient enters the ward. Note that such load situation depends on all previously scheduled patients (on that day).

The aim of the two above actions is to have a smoother patient flow and shorter in-process waiting times. After careful consideration, it was therefore decided that patients be grouped based on their therapy path rather than their therapy type.

For people who often address theoretical planning and scheduling problems, the idea of assigning an individual appointment to each patient appears highly reasonable, if not trivial. However, this is a good example of gap between theory and practice which is often overlooked. Taking this simple step is perhaps the most demanding aspect of the whole reengineering process. In fact, treatments are booked (by patients or physicians) over time, i.e., requests do not arrive in batches. Implementing individual appointments – whatever the criterion used to determine such appointments – brings about a major change for both clinical staff and patients. The staff will have to deploy the IT infrastructure to manage personalized appointments, while the patients must get accustomed to the new discipline (see the discussion at the end of Section 4.4.2).

The introduction of individual appointments is the main process innovation proposal resulting from the analysis carried out. Only after

this issue is accepted, the problem arises of operating the system *in the best way. Here is where MO enters the picture*. In fact, after a set of patients has already been scheduled in a given day, a method has to be devised for deciding the appointment time of the next patient requesting a treatment. We want to do this so that patient flow is as smooth as possible. Hence, we do not want to overload existing resources at the expense of the patients' waiting times, but rather perform the optimization using a patient's viewpoint. MO makes it possible to translate such a design goal into a precise scheduling objective.

Obviously, it would be possible to resort to very simple heuristic strategies, such as sequentially assigning the resources to the patients requiring them, according to their particular paths and to the workload of the resource that derives from previous appointments. However, as operations research analysts know very well, simple heuristic rules may not optimally exploit the information available on the current patient and on patients already scheduled, and eventually may result in poor solutions. For this reason, an optimization model has been devised to be run at every patient call, according to the following specifications.

- The model computes a schedule for the new patient, accounting for already scheduled patients.
- Appointment times of previous patients have been already communicated to previous patients and hence are kept fixed, but the *internal schedule* of a previous patient may be changed to better accommodate the new patient.
- The model schedules the new patient so that *the overall stay in the ward of all patients is minimized*.
- Once the model is run, the resulting admission time of the new patient is assigned as appointment time to him/her.
- The method must be efficient enough to compute and communicate the appointment to the patient in real time.

4.4. The optimization model

In this section we describe the mathematical model used every time a new patient requests a treatment. Our optimization model is deterministic, and average activity durations are used for this purpose. This means that a suitable validation of the optimization results will have to be carried out (in the *check* phase, Section 4.5).

From a scheduling viewpoint, the problem can be viewed as a *flexible job shop* [45], i.e., a job shop in which patients correspond to jobs and activities to stages. Each stage has a certain *capacity*, i.e., the maximum number of patients that can undergo the same activity at the same time. The values reported in Table 6 derive from resource limitations and were assessed during the data collection phase (Section 4.2.1). More precisely, they correspond to the number of operators devoted to a group of activities, except for the infusion stage, in which the limit is given by the number of armchairs (eight). In fact, in such a stage the three nurses are only involved during patient setup and release, so we disregarded this detail (which would have considerably complicated the model), but we explicitly model this aspect in the simulations. Also, the fact that certain stages (such as blood sampling and subcutaneous therapy) share the same operator(s) can be easily introduced in the model. The latter issue makes the application of standard solution approaches for flexible job shop problems problematic, and an ad-hoc optimization model was devised.

Table 6
Data for the case study. * Admission capacity is 2 only until 9 a.m., then 1. ** Blood sampling and subcutaneous therapy share a single operator. *** Medical visit and BMB share three operators.

Activity	Duration	Capacity
1. Patient entry	0	50
2. Admission	2	2*
3. Blood sampling	1	1**
4. Post-sampling wait	18	50
5. Medical visit	4	3***
6. BMB	4	3***
7. Infusion	Patient-specific	8
8. Subcutaneous therapy	2	1**
9. Exit	–	–

Taking advantage of the fact that the planning horizon is relatively limited (7 h), we devised a time-indexed formulation, as this type of formulations often proves very efficient in solving complex scheduling problems [46]. Moreover, these formulations allow the decision maker to trade modeling accuracy for computational efficiency, as discussed in the next section.

4.4.1. Model formulation

The technical elements of the optimization model can be summarized as follows.

- There are seven different activities (stages), called *real*, plus two *fictitious* activities (having zero processing time), needed to model patient entry and exit (Table 3):
  1. Patient entry
  2. Admission
  3. Blood sampling
  4. Post-blood sampling waiting
  5. Medical visit
  6. BMB
  7. Infusion
  8. Subcutaneous therapy
  9. Patient exit
- In the optimization model, the time is discretized in *time slots*, and each treatment is supposed to last an integer number of slots. A key modeling issue is to define how long should a time slot be. One has to find a tradeoff between the maximization of model significance (which would lead to small time slots) and the minimization of computational complexity (for which fewer variables and hence long time slots would be preferable). After some preliminary experiments, it was decided to adopt the shortest time slot length still compatible with the strict response requirements posed by real-time utilization of the model. Such a length was set to 5 min. This would lead to having a time horizon of 84 slots (equivalent to 420 min = 7 h), in which the paths of all patients must be included. However, in order to accommodate possible random delays, it has been decided to plan for 80 time slots instead of 84, thus keeping a planned 20-min buffer at the end of the day. Moreover, average activity durations were rounded up to the next integer number of slots, which is a further protection against unexpected delays. These choices should significantly decrease the chance that any patient leaves the system after 3 p.m..
- Activity durations and production capacity. For each activity, Table 6 shows the length (number of slots) and the maximum number of patients who can simultaneously perform the treatment.

The model determines the time interval in which each activity must be scheduled, for the current as well as all previous patients, provided that the entry time of the previous patients is kept fixed. The objective of the model is the minimization of the total amount of time spent by patients in the ward (total lead time).

The constraints must correctly represent the operations of the ward. In our case they can be summarized as follows.

- *Constraints on activity duration*. As each activity is assumed to have known, deterministic duration, we must enforce that each activity be completely carried out.
- *Precedence constraints*. The path of each patient is known and the corresponding activities must be performed in the required order.
- *Capacity constraints*. Each activity or set of activities cannot be

provided simultaneously to a number of patients larger than the value indicated in Table 6.

- *Constraints on previously scheduled patients*. The appointment times of the previous patients have been already communicated and therefore cannot be modified, so these are fixed and known in the model. Note that, if this allows a better schedule for the current patient, the schedule of the *intermediate* activities of previous patients may still be changed.

In the mathematical formulation of the problem, we let N denote the set of patients scheduled so far, and n the current patient. We use indices t, i, k for time slots, patients and activities respectively. There are 80 time slots, and $\mathcal{P}_i$ denotes the path of patient i, with $\sigma_i(k)$ denoting the activity following k in $\mathcal{P}_i$. In this model, $x_{tik} = 1$ if patient i is undergoing activity k in time slot t, $s_{tik} = 1$ if t is the first time slot in which activity k of patient i takes place, while $f_{tik} = 1$ if t is the last time slot in which activity k of patient i is performed. Since the appointment time of the *previous* patients is fixed, the values s_{ti1} for all $i \neq n$ are known (denoted by $\tilde{s}_{ti1}$). The duration of activity k for patient i is denoted by d_{ik}. (Except for infusions, such duration is patient-independent.) The continuous variable w_i equals the lead time of patient i. The first and the last activity of all paths are the fictitious entry and exit respectively, such that $d_{i1} = d_{i9} = 0$ for all i. The lead time of patient i is given by the difference between the start times of activities 9 and 1. The objective is to minimize the total lead time, which is equivalent to minimizing non-value-added time.

$$\max \sum_{i=1}^{n} w_i \tag{1}$$

$$\sum_{t=1}^{80} x_{tik} = d_{ik} \quad i \in N, \quad k \in \mathcal{P}_i \setminus \{1\} \tag{2}$$

$$\sum_{t=1}^{80} s_{tik} = 1 \quad i \in N, \quad k \in \mathcal{P}_i \tag{3}$$

$$\sum_{t=1}^{80} f_{tik} = 1 \quad i \in N, \quad k \in \mathcal{P}_i \tag{4}$$

$$s_{ti1} = f_{ti1} \quad i \in N, \quad t = 1, \ldots, 80 \tag{5}$$

$$s_{ti9} = f_{ti9} \quad i \in N, \quad t = 1, \ldots, 80 \tag{6}$$

$$\sum_{i \in N} x_{ti2} \leq 2 \quad t = 1, \ldots, 12 \tag{7}$$

$$\sum_{i \in N} x_{ti2} \leq 1 \quad t = 13, \ldots, 80 \tag{8}$$

$$\sum_{i \in N} x_{ti3} + \sum_{i \in N} x_{ti8} \leq 1 \quad t = 1, \ldots, 80 \tag{9}$$

$$\sum_{i \in N} x_{ti4} \leq 50 \quad t = 1, \ldots, 80 \tag{10}$$

$$\sum_{i \in N} x_{ti5} + \sum_{i \in N} x_{ti6} \leq 3 \quad t = 1, \ldots, 80 \tag{11}$$

$$\sum_{i \in N} x_{ti6} \leq 1 \quad t = 1, \ldots, 80 \tag{12}$$

$$\sum_{i \in N} x_{ti7} \leq 8 \quad t = 1, \ldots, 80 \tag{13}$$

$$\sum_{t=1}^{80} t s_{ti7} \geq 49 \quad i \in N \tag{14}$$

$$\sum_{t=1}^{80} t s_{ti,\sigma_i(1)} \geq \sum_{t=1}^{80} t f_{ti1} \quad i \in N \tag{15}$$

$$\sum_{t=1}^{80} t s_{ti,\sigma_i(k)} \geq \sum_{t=1}^{80} t f_{tik} + 1 \quad i \in N, \quad k \in \mathcal{P}_i \setminus \{1\} \tag{16}$$

$$s_{1ik} \geq x_{1ik} \quad i \in N, \quad k \in \mathcal{P}_i \setminus \{1\} \tag{17}$$

$$s_{tik} \geq x_{tik} - x_{t-1,ik} \quad i \in N, k \in \mathcal{P}_i \setminus \{1\}, t = 2, \ldots, 80 \tag{18}$$

$$f_{tik} \geq x_{tik} - x_{t+1,ik} \quad i \in N, k \in \mathcal{P}_i \setminus \{1\}, t = 1, \ldots, 80 \tag{19}$$

$$w_i = \sum_{t=1}^{80} t s_{ti9} - \sum_{t=1}^{80} t s_{ti1} + 1 \quad i \in N \tag{20}$$

$$s_{ti1} = \tilde{s}_{ti1} \quad i \neq n \tag{21}$$

Constraints in the optimization model have the following meaning.

(2) each activity must be carried out in a number of time slots equal to its duration.
(3,4) each activity starts and ends exactly once
(5,6) entry and exit are fictitious activities having zero duration
(7)–(13) these constraints account for the limited capacities of various stages
(14) blood transfusions must take place after 12:00 p.m. (from time slot 49 onwards)
(15) the first real activity must start after the patient entered the ward
(16) a real activity ($k > 1$) cannot start before the previous activity on the path is completed
(17,18) define the starting slot of each activity
(19) defines the ending slot of each activity
(20) defines the lead time of each patient
(21) these constraints keep the appointment times of the previous patients fixed. Of course, when the first patient is scheduled, these constraints are omitted.

Notice that constraints (17)–(19), together with (3) and (4), enforce nonpreemption.

4.4.2. Model implementation and solution

The model was run on a 3.2 GHz Intel Core i3 processor with 4GB of RAM, using OPL Studio 6.1 and the ILOG IBM CPLEX 12.2 MILP solver. The optimal solution of a single instance of the model was found in few seconds, a time compatible with the practical use of the model.

In order to validate the method described above, the appointment protocol has been executed on the same data used in the *as-is* simulation described in Section 4.2.3. The experiment consists in randomly ordering the patients in the daily list, and scheduling them one at a time, each time solving an instance of (1)–(21). Obviously, as the number of already scheduled patients grows, the model becomes larger and the CPU time increases accordingly. However, even in the largest instances no more than one minute of CPU time was needed to solve the problem.

From a qualitative viewpoint, the newly computed appointments are completely different from the batch-like appointment policy currently adopted in the ward, and patient waiting times are drastically decreased. However, the model may not accommodate all the patients within the 80-slot time horizon. For example, when it was the turn of patient #31 (path 1) and patient #49 (path 4), the solver was not able to find a feasible solution, so these two patients remained unscheduled. This suggests that 49 patients may not be accommodated with the current resource allocation. This issue is further analyzed in the *check* phase (Section 4.5).

We point out that the model is *flexible*, i.e., it can easily accommodate changes in the data. For instance, even if all activities (except infusions) are assumed to have patient-independent durations, customized values can be used in (2) to account for patients with special needs, hence requiring an expectedly longer medical visit. As another example, the model can be used to perform tactical analysis, allowing to

assess the effect (on total lead time) of deploying one more unit of personnel for a certain activity. This can be done simply by changing the corresponding right-hand side in one of the constraints (7)–(13). Also, if the ILP solver cannot feasibly schedule the current patient, but it is an urgent case, one can simply unfix certain entry times in (21), thus allowing the model to change some previously scheduled appointment (this of course creates some additional burden to the operator). Finally, notice that the model can be applied to any patient mix.

While the use of the optimization model is certainly appealing, since it is conceived to minimize total lead time, one must nonetheless be aware of the model limits and approximations.

One limit of the model is that processing times are considered deterministic, while this may not be the case. In this respect, it is important to recall that activity durations have been somewhat overestimated, but of course uncertainties and disruptions (adverse events, unforecasted delays...) may always occur. The way such unexpected events are handled is not addressed here, since the purpose of the model is only to provide a feasible appointment plan.

Finally, as observed by Peek [47], successful interventions may require behavioral changes to patients, not only to providers. In our case study, the model assumes that each patient punctually shows up at the appointment time. While it may seem reasonable to assume that if a patient is given an appointment for a certain time, he/she will show up exactly at that time, this is not so obvious in practice. Indeed, patients will tend to show up early, for various reasons, such as habits, a conservative attitude (i.e., to decrease the risk of being late), or the belief that showing up early will anyhow accelerate the whole process. These aspects are especially true in a phase of transition from the old to the new management model, though their effect should tend to disappear over the medium term.

4.5. To-be simulation (check)

According to the PDCA cycle, at this point of the process a *check* on the feasibility of the new appointment procedure must be done. In fact, we still must produce evidence that the results anticipated by the mathematical model can indeed be achieved. Given a full-day solution of the optimization model, we therefore setup a new simulation (*to-be*), in which patients are supposed to show up at the appointment times computed by the optimization model. First of all, from the same sample day used in the *as-is* simulation, we considered five different patient booking permutations and hence as many appointment plans generated by solving the mathematical model. For each of such five scenarios, the whole day was simulated 20 times. A synthesis of the final results has been obtained computing the average figures resulting from all 100 simulation runs, and were compared with the results of the as-is simulation.

Table 7 shows the new simulation results. Average patient lead times are drastically reduced for all patients. In absolute terms, the longest waits (almost 10 min on the average) concern patients who get a subcutaneous therapy for the first time (path 5). However, such a waiting time has been decreased by approximately 94%, and it makes up only 4.32% of the lead time. Of course, the main reason for such prospective improvement is the shift from a push- to a pull-type patient flow management strategy, getting closer to the lean objective of attaining the so-called *one-piece flow* (i.e., a flow in which no queues build up between one activity and the next). This is significantly different from the typical *batch-and-queue* logic of most healthcare processes, according to which it is often chosen to gather more patients at the same time for established organizational practices. Another consequence of this logic shift is that lead times are much less variable. While these benefits are mainly due to process streamlining, we observe that MO allows a careful planning of the infusion room, since in principle armchairs are always available by the time a patient requires one of them, and all flow management-related queues are drastically reduced.

Table 7
To-be simulation results. All times are expressed in minutes.

Path	Lead time (average)		Wait time (% on lead time) (average)		Lead time (maximum)	
	to-be	as-is	to-be	as-is	to-be	as-is
1	133.5	218.52	7.82 (5.18%)	92.84 (42.49%)	150.72	304.14
2	127.29	138.15	4.71 (3.7%)	15.66 (11.34%)	146.47	151.62
3	162.83	225.98	8.53 (3.57%)	71.68 (31.72%)	188.83	311.08
4	93.88	102.14	0.13 (0.13%)	8.39 (8.21%)	167.05	198.75
5	148.6	295.42	9.86 (4.32%)	156.67 (53.03%)	169.15	352.39
6	29.9	92.59	7.05 (24.06%)	70.55 (76.19%)	35.15	120.8
7	51.4	93.05	7.92(17.27%)	49.57 (53.27%)	58.15	129.43

However, simulation also highlights the other side of the coin. Out of 100 to-be simulation runs, an average of 6 patients are not able to complete their treatments within 3 p.m.. This is not surprising, since of course streamlining patients' paths may entail some resource idle time. Moreover, the optimization model schedules some patients so that they end their stay in the ward exactly in the final slot (i.e., at 2.40 p.m.), hence some accumulated random variability in the processing times of the last activities may easily determine that the time horizon is exceeded. Indeed, we may expect that such situations actually occur. In one of the 100 simulation runs, as many as 9 patients could not complete the entire path by 3 p.m.. This information can help the managers to correctly size the service, suggesting that in order to ensure completion of 49 treatments within the same day without congestion or without employing additional resources, working hours should be extended.

4.6. Towards implementation (act)

In this study we wanted to synthetically present the integration between LT and MO in the reengineering of an healthcare process, namely appointment scheduling in a hematological ward of a large Italian hospital. Our case study pointed out the benefits of an appropriate use of the two approaches, namely LT for structural process reengineering (shifting from a push strategy for patients' management to a pull strategy for appointments management) and MO to fully exploit the potential of the new process (using optimization software to determine the optimal arrival time for each patient requesting an appointment).

At this point, the implementation phase (*act*) will close the PDCA cycle. This phase itself requires careful planning since changes will concern not only the patients but also staff and management. It has been observed [48] (in a manufacturing environment, but the concept applies to health services as well) that the reason why certain lean improvement projects fail is related to the lack of knowledge on which changes the management should commit to. In our case, the fact that the arrival of patients is homogeneously distributed throughout the day will imply that the ward has to be always prepared to supply any kind of treatment throughout the opening hours. So, it is mandatory that the management supports these changes in the internal organization of the ward, which in the end should lead to smoother personnel workload over time and lower risk of congestion and related stressful situations (waste reduction). Moreover, as noted in Section 4.4.2, the flexibility of the model (with respect to various issues including ward capacity and patient mix) should help sustaining the improvements over time.

Organizational issues also involve practical tasks. In fact, to make the whole procedure fully operational, it will be necessary to create a computer interface enabling any operator (e.g. nurse) to run the appointment procedure. In order to fix an appointment, the operator should simply specify the patient's path (including detailed information on infusion time) on the interface, run the software and communicate the appointment time to the patient. The transition towards the *pull* appointment planning procedure is expected to be completed by mid

2019. In this phase, new difficulties or overlooked problems may arise, as it is typical of process change implementations. Adjusting to new processes requires some adaptation to both the staff and the patients. However, a knowledgeable use of both LT and MO tools and techniques can help pointing out the right direction and how get the best from it.

5. Conclusions

In conclusion, we summarize here a few considerations suggested by our experience. We draw them on the basis of our single case study, for which, as explained in Section 4.6, implementation is under way. So, rather than stating general principles, we want to underline the aspects, particularly related to the integration of LT and MO, that we believe are likely to occur also in other situations when LT and MO are jointly used.

- *Streamlining the process makes it possible to optimize some KPIs*. In the current situation, the chaotic and random nature of the appointments does not allow any detailed scheduling of the sessions, since it would not be possible to predict which patients are present at different times. Moving from a push to a pull system may enable planning treatments for each patient individually. In turn, this paves the way for a careful scheduling of the appointments, which has the potential for achieving very significant improvements in overall patient lead time (Table 7). Savings are different for different paths, the average decrease in patient lead time over all patients being 34.8% (computed accounting for the number of patients for each path given in Table 1).
  A relevant side-product of the application of a formal scheduling model is the possibility of representing the evolution of the system also through simple visual instruments (e.g. Gantt diagrams), which is extremely valuable for communication and transparency purposes.
- *Planning tools can be used to address both operational and tactical issues*. We showed how the optimization model can be used to address an operational issue, namely scheduling individual patients. However, also *tactical* issues (according to Hulshof et al. [49]) can be addressed. In particular, the joint use of LT and MO tools showed the possibility of crashing lead times *without requiring additional resources*, at the price of a certain decrease in throughput. If such a decreased throughput is deemed unacceptable, the optimization and simulation models developed can help the managers evaluate the impact of tactical decisions on the throughput of the ward. For instance, if overtime is considered, one simply needs to increase the number of time slots (currently 80) in the model. Similarly, the effect of allocating additional resources to an activity can be evaluated adjusting the right hand side of the corresponding constraint(s) in the optimization model, and the corresponding resource parameter in the simulation model.
- *Improving the process brings benefit to multiple stakeholders*. As already observed in Section 4.1, besides improving patient's experience, the minimization of patient lead times positively affects staff working conditions as well, since individual appointments result in a less crowded waiting room. This eliminates the need for making scheduling decisions on the spot (as long as unforeseen events do not occur). This is in accordance with various other experiences reported in the literature, such as [14,50].

As a challenging topic for future research, we view a deeper and more systematic analysis of such the relationship between LT and MO, using the tools of multimethodology, e.g. to better investigate the specific techniques which can be "detached" from a methodology and fruitfully employed in conjunction with another.

Acknowledgment

The authors wish to acknowledge support and help from the Direction of Policlinico Santa Maria alle Scotte throughout the project and the writeup.

References

[1] Womack JP, Jones DT, Roos D. The machine that changed the world. New York: Rawson and Associates; 1990.
[2] Hines P, Found P, Griffiths G, Harrison R. Staying lean: thriving, not just surviving. Lean Enterprise Research Centre; 2008.
[3] Joosten T, Bongers I, Janssen R. Application of lean thinking to health care: issues and observations. Int J Qual Healthc 2009;21(5):341–7. https://doi.org/10.1093/intqhc/mzp036.
[4] Bicheno J, Holweg M. The lean toolbox: the essential guide to lean transformation. Buckingham: PICSIE Books; 2009.
[5] Tang CS, Sodhi MS. The OR/MS ecosystem: strengths, weaknesses, opportunities and threats. Oper Res 2008;56(2):267–77. https://doi.org/10.1287/opre.1080.0519.
[6] Portougal V, Robb DJ. Production scheduling theory: just where is it applicable? Interfaces 2000;30(6):64–76. https://doi.org/10.1287/inte.30.6.64.11623.
[7] Institute for Healthcare improvement. Going lean in healthcare. Cambridge (MS): Innovation Series 2005; 2005.
[8] Liker JK, Meier D. Toyota way fieldbook. McGraw-Hill Education; 2006.
[9] Rooney S, Rooney J. Lean glossary. Qual Prog 2005;38(6):41–7.
[10] Koning HD, Verver JPS, Heuvel JVD, Bisgaard S, Does RJMM. Lean six sigma in healthcare. J Healthc Qual 2006;28(2):4–11. https://doi.org/10.1111/j.1945-1474.2006.tb00596.x.
[11] Fine BA, Golden B, Hannam R, Morra D. Leading lean: a canadian healthcare leader's guide. Healthc Quart 2009;12(3):32–41.
[12] Mazzocato P, Savage C, Brommels M, Aronsson H, Thor J. Lean thinking in healthcare: a realist review of the literature. BMJ Qual Saf HealthCc 2010;19(5):376–82.
[13] Ford AL, Williams JA, Spencer M, McCammon C, Khoury N, Sampson TR, Panagos P, Lee JM. Reducing door-to-needle times using toyota's lean manufacturing principles and value stream analysis. Stroke 2012;43(12):3395–8. https://doi.org/10.1161/STROKEAHA.112.670687.
[14] Nelson-Peterson DL, Leppa CJ. Creating an environment for caring using lean principles of the virginia mason production system. J Nurs Admin 2007;37(6):287–94. https://doi.org/10.1097/01.NNA.0000277717.34134.a9.
[15] Wood D. Taking the pulse of lean healthcare. Healthc Quart 2012;15(4):27–33.
[16] Moraros J, Lemstra M, Nwankwo C. Lean interventions in healthcare: do they actually work? A systematic literature review. Int J Qual Healthc 2016;28(2):150–65. https://doi.org/10.1093/intqhc/mzv123.
[17] Andersen, H., Røvik, K. A., & Ingebrigtsen, T. (a). Lean thinking in hospitals: is there a cure for the absence of evidence? A systematic review of reviews. BMJ Open, 4(1), e003873. 10.1136/bmjopen-2013-003873.
[18] Poksinska BB, Fialkowska-Filipek M, Engström J. Does lean healthcare improve patient satisfaction? A mixed-method investigation into primary care. BMJ Qual Saf 2017;26(2):95–103. https://doi.org/10.1136/bmjqs-2015-004290.
[19] Mazzocato, P., Holden, R. J., Brommels, M., Aronsson, H., Backman, U., Elg, M., & Thor, J. (b). How does lean work in emergency care? A case study of a lean-inspired intervention at the Astrid Lindgren Childrens hospital. Stockholm. Sweden.: BMC Health Services Research. 12:28. 10.1186/1472-6963-12-28.
[20] Hines P, Holweg M, Rich N. Learning to evolve: a review of contemporary lean thinking. Int J Oper Prod Manag 2004;24(10):994–1011. https://doi.org/10.1108/01443570410558049.
[21] Osono E, Shimizu N, Takeuchi H. Extreme toyota: radical contradictions that drive success at the worlds best manufacturer. Hoboken, NJ: John Wiley; 2008.
[22] Castle A, Harvey R. Lean information management: the use of observational data in health care. Int J Prod Perform Manag 2009;58(3):280–99. https://doi.org/10.1108/17410400910938878.
[23] Naidoo P, Smuts B, Claassens M, Rusen ID, Enarson DA, Beyers N. Operational research to improve health services, a guide for proposal development. Desmond Tutu TB Centre, Department of Paediatrics and Child Health, Stellenbosch University; 2013. ISBN 978-0-620-57795-3
[24] Mac Kenzie J, Kassab R, Hong G. Lean management in hematology provides better patient care. Med Lab Observ 2017. www.mlo-online.com
[25] Lamm MH, Eckel S, Daniels R, Amerine LB. Using lean principles to improve outpatient adult infusion clinic chemotherapy preparation turnaround times. Am J Health- Syst Pharm 2015;72(13):1138–46. https://doi.org/10.2146/ajhp140453.
[26] Turkcan A, Zeng B, Lawley M. Chemotherapy operations planning and scheduling. IIE Trans Healthc Syst Eng 2012;2:31–49. https://doi.org/10.1080/19488300.2012.665155.
[27] Mustafee N, Hughes F, Katsaliaki K. Simulation-based study of hematology outpatient clinics with focus on model reusability. In: Jain S, Creasey RR, Himmelspach J, White KP, Fu M, editors. Proceedings of the 2011 winter simulation conference 2011. p. 1178–89. https://doi.org/10.1109/WSC.2011.6147840.
[28] Santibanez P, Chow VS, French J, Puterman ML, Tyldesley S. Reducing patient wait times and improving resource utilization at british columbia cancer agencys ambulatory care unit through simulation. Healthc Manag Sci 2009;12(4):392–407. https://doi.org/10.1007/s10729-009-9103-1.
[29] Wijewickrama AK, Takakuwa S. Designing outpatient appointment systems with patient characteristics: a case study. Int J Healthc Technol Manag 2012;13(1-3):157–69. https://doi.org/10.1504/IJHTM.2012.048953.
[30] Santibanez P. Improving the chemotherapy scheduling process at BCCA vancouver

center. University of British Columbia, Sauder School of Business, Center for Operations Excellence; 2009http://www.sauder.ubc.ca/coe. Working Paper
[31] Liang B, Turkcan A, Ceyhan ME, Stuart K, 10.1080/00207543.2014.988891. Improvement of chemotherapy patient flow and scheduling in an outpatient oncology clinic. Int J Prod Res 2015;53(24):7177–90.
[32] Le MD, Nguyen MHN, Baril C, Gascon V, Binh TB. Heuristics to solve appointment scheduling in chemotherapy. The 2015 IEEE RIVF international conference on computing and communication technologies research, innovation, and vision for future (RIVF). 2015. p. 59–64. https://doi.org/10.1109/RIVF.2015.7049875.
[33] Hesaraki AF, Dellaert NP, de Kok T. Generating outpatient chemotherapy appointment templates with balanced flowtime and makespan. Eur J Oper Res 2019;275:304–18. https://doi.org/10.1016/j.ejor.2018.11.028.
[34] Hahn-Goldberg S, Carter MW, Beck JC, Trudeau M, Sousa P, Beattie K. Dynamic optimization of chemotherapy outpatient scheduling with uncertainty. Healthc Manag Sci 2014;17(4):379–92. https://doi.org/10.1007/s10729-014-9268-0.
[35] Lamé G, Jouini O, Stal-Le Cardinal J. Outpatient chemotherapy planning: a literature review with insights from a case study. IIE Trans Healthc SystEng 2016;6(3):127–39. https://doi.org/10.1080/19488300.2016.1189469. 2016
[36] Mingers J, Brocklesby J. Multimethodology: towards a framework for mixing methodologies. Omega 1997;25(5):489–509.
[37] Robinson S, Radnor ZJ, Burgess N, Worthington C. Simlean: utilising simulation in the implementation of lean in healthcare. Eur J Oper Res 2012;219(1):188–97. https://doi.org/10.1016/j.ejor.2011.12.029.
[38] Baril C, Gascon V, Miller J, Cote N. Use of a discrete-event simulation in a Kaizen event: A case study in healthcare. Eur J Oper Res 2016;249(1):327–39. https://doi.org/10.1016/j.ejor.2015.08.036.
[39] Graban M. Lean Hospitals Improving quality, patient safety, and employee engagement. 3rd CRC Press; 2016.
[40] Denscombe M. The good research guide for small scale research projects. Buckingham: Open University Press; 1998.
[41] Creasy T. Factors that lead to success or failure in healthcare projects. Qual Prog 2017:24–9. February
[42] Scherer WT, Pomroy TA, Fuller DN. The triangular density to approximate the normal density: decision rules-of-thumb. Reliabil Eng Syst Saf 2003;82(3):331–41. https://doi.org/10.1016/j.ress.2003.08.003.
[43] Radnor ZJ, Holweg M, Waring J. Lean in healthcare: the unfilled promise? Social Sci Med 2012;74(3):364–71. https://doi.org/10.1016/j.socscimed.2011.02.011.
[44] Serrat O. The five whys technique. In: Serrat O, editor. Knowledge solutions Singapore: Springer; 2017. p. 307–10. https://doi.org/10.1007/978-981-10-0983-9-32.
[45] Al-Hinai N, El Mekkawy TY. An efficient hybridized genetic algorithm architecture for the flexible job shop scheduling problem. Flex Serv Manuf J 2011;23(1):64–85. https://doi.org/10.1007/s10696-010-9067-y.
[46] van den Akker JM, Hurkens CAJ, Savelsbergh MWP. Time-indexed formulations for machine scheduling problems: column generation. INFORMS J Comput 2000;12(2):111–24. https://doi.org/10.1287/ijoc.12.2.111.11896.
[47] Peek CJ. Building a medical home around the patient: What it means for behavior. Fam Syst Health 2010;28(4):322–33. https://doi.org/10.1037/a0022043.
[48] Pearce A, Pons D, Neitzert T. Implementing lean – outcomes from SME case studies. Oper Res Perpect 2018;5:94–104. https://doi.org/10.1016/j.orp.2018.02.002.
[49] Hulshof PJH, Kortbeek N, Boucherie RJ, Hans EW, Bakker PJM. Taxonomic classification of planning decisions in health care: a structured review of the state of the art in OR/MS. Health Syst 2012;1(2):129–75. https://doi.org/10.1057/hs.2012.18.
[50] Braaten JS, Bellhouse DE. Improving patient care by making small sustainable changes: a cardiac telemetry units experience. Nurs Econ 2007;25:162–6.