Springer Proceedings in Mathematics & Statistics

Andrea Matta
Jingshan Li
Evren Sahin
Ettore Lanzarone
John Fowler *Editors*

Proceedings of the International Conference on Health Care Systems Engineering

Springer

Springer Proceedings in Mathematics & Statistics

Volume 61

For further volumes:
http://www.springer.com/series/10533

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of select contributions from workshops and conferences in all areas of current research in mathematics and statistics, including OR and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

Andrea Matta • Jingshan Li • Evren Sahin
Ettore Lanzarone • John Fowler
Editors

Proceedings of the International Conference on Health Care Systems Engineering

Springer

Editors
Andrea Matta
Dipartimento di Meccanica
Politecnico di Milano
Milano, Italy

Jingshan Li
College of Engineering
University of Wisconsin
Madison, WI, USA

Evren Sahin
Laboratoire Génie Industriel
Ecole Centrale Paris
Grande Voie des Vignes
Châtenay-Malabry, France

Ettore Lanzarone
CNR-IMATI
Milan, Italy

John Fowler
Department of Supply Chain Management
Arizona State University
Tempe, AZ, USA

ISSN 2194-1009 ISSN 2194-1017 (electronic)
ISBN 978-3-319-01847-8 ISBN 978-3-319-01848-5 (eBook)
DOI 10.1007/978-3-319-01848-5
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013952915

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This volume features selected and peer-reviewed contributions from the International Conference on Health Care Systems Engineering (HCSE). This conference provides an opportunity to discuss operations management issues in health care delivery systems. The emphasis is on quantitative methods for the analysis, design and management of health care systems.

The participants are faculties, students and medical doctors from several disciplines. The main objective is fostering the collaboration between operations management scientists and clinicians.

Scientists and practitioners have the opportunity to discuss about new ideas, methods and technologies for improving the operation of health care organizations. The event emphasizes the research in the field of health care systems engineering developed in close collaboration with clinicians.

The conference took place in Milan, Italy, between the 22nd and the 24th of May 2013 in the San Raffaele Hospital. A limited number of papers was selected under a double-blind review process. I would like to thank all of the Scientific Committee and the 43 anonymous reviewers for the selection of the works. In total, 24 papers are included in the conference proceedings. Each paper was presented at the conference and discussed with experts from the clinical field.

I would like to express my deep gratitude to our invited speakers, Dr. Gianlorenzo Scaccabarozzi for agreeing to address the conference on the "Emerging Needs and Future Perspectives of Italian Home Care Providers", and Prof. Xiaolan Xie with the topic of "Mathematical Modeling of Healthcare Engineering Problems". Their contributions are perfectly in line with the aim of the conference which tries to show the two facets of the same game: problems and related solutions.

I would like to thank all the speakers, authors and discussants of the papers together their accompanying persons for their participation to HCSE.

I gratefully acknowledge the Organizing Committee: Riccardo Dodi, Nicla Frigerio, Ettore Lanzarone, Elettra Oleari, Giulia Pedrielli, Alberto Sanna and Semih Yalcindag.

My hope is that this conference will serve as a forum for researchers, academics and clinicians in the broad area of health care systems engineering to discuss their most recent research findings and to provide them with opportunities for technology transfer.

Milan, Italy Andrea Matta

Contents

Chapter 1
Home Care Services Delivery: Equity Versus Efficiency in Optimization Models

Paola Cappanera, Maria Grazia Scutellà, and Filippo Visintin

Abstract Home Care Services (HCS) delivery is a quite recent and challenging problem motivated by the ever increasing age of population and the consequent need to reduce hospitalization costs. Integer Linear Programming (ILP) models have been recently proposed in [5] to formulate a very general HCS problem, with the aim at balancing the operator workload. In fact, in Home Care setting "equity" criteria are crucial to guide the decisions. "Efficiency" criteria, i.e., the minimization of the operating costs, are essential as well. The aim of this paper is thus to compare equity criteria versus efficiency criteria in HCS. Preliminary computational results on a set of real instances are presented and analysed. Specifically, two alternative "balancing" objective functions are compared via optimization and simulation, by showing their impact on diverse relevant Quality of Service indicators, including cost indicators.

1.1 Introduction

Nowadays, the ever increasing average age of population, at least in industrialized countries, and the increased costs for the consequently required care, compel the medical care units to offer Home Care Services (HCS) in an attempt to limit costs. Elderly people have in fact varying degrees of need for assistance and medical

P. Cappanera
Dipartimento di Ingegneria dell'Informazione, Università degli Studi di Firenze, Firenze, Italy
e-mail: paola.cappanera@unifi.it

M.G. Scutellà
Dipartimento di Informatica, Università di Pisa, Pisa, Italy
e-mail: scut@di.unipi.it

F. Visintin (✉)
Dipartimento di Ingegneria Industriale, Università degli Studi di Firenze, Firenze, Italy
e-mail: filippo.visintin@unifi.it

A. Matta et al. (eds.), *Proceedings of the International Conference on Health Care Systems Engineering*, Springer Proceedings in Mathematics & Statistics 61,
DOI 10.1007/978-3-319-01848-5_1,

treatment, and it may be advantageous to allow them to live in their own homes as long as possible. In addition, medical treatments carried out at patients home impact favorably on their quality of life. Therefore, HCS are a cost-effective and flexible instrument in the social system.

Interestingly, in Home Care setting the minimization of the operating costs, that is a common objective of the stakeholders (either private or public) providing the service, is not the only objective to be taken into account to guide the Home Care decisions. In fact, another objective typically used in HCS is the balancing of the utilization factor among the operators, where the *operator utilization factor* is the total workload of the operator in the considered planning horizon over his/her maximum possible workload. In order to achieve this objective, one possibility is to maximize the minimum operator utilization factor. Hereafter this balancing objective function will be referred to as *maxmin*. Anyway, an alternative balancing function may be defined, which consists in minimizing the maximum operator utilization factor. This alternative function will be indicated as *minmax*. Both formulations have been proposed in [5]. In the context of assignment decisions in HCS, the *maxmin* criterion has been also investigated in [9].

The aim of this paper is to compare the two balancing objective functions in an extensive way. Specifically, *maxmin* and *minmax* are compared both via an optimization approach and also via a simulation experimentation performed on a set of real HCS instances, by showing their impact on diverse relevant Quality of Service (QoS) indicators. This set of QoS indicators includes the mean operator utilization factor over the considered planning horizon, the corresponding range, i.e. the difference between the maximum and the minimum operator utilization factors, and the daily variation of the operator utilization factor. In addition, in order to provide a hint about the influence of such equity measures on the HCS efficiency, QoS indicators related to the operated service time and the operator travelled time are also investigated.

The results of the preliminary computational experiments are very interesting. In fact, they show that the *maxmin* criterion is able to return more balanced HCS solutions, in the sense that the difference between the maximum and the minimum operator utilization factors is smaller than the one returned by *minmax*. The *maxmin* criterion is also preferable in balancing the operator traveling time and service time. This is true not only by looking at the overall planning horizon, which is a week in our experiments, but also at a daily level. Such stronger equity achievements are obtained for not too high a price in the increased mean operator utilization factor, mean operator service time and mean operator traveling time. On the other hand, the *minmax* criterion appears to be more suitable for the minimization of the operating costs since it always returns solutions with the smaller total travelled time.

The achievements above have been shown first in a deterministic setting, via optimization, and then confirmed by the simulation experiments, where the robustness of the computed HCS solutions against travel time and service time variability has been evaluated.

The plan of the paper is the following. In Sect. 1.2 we introduce the HCS problem, and describe the two alternative objective functions *maxmin* and *min-max* [5]. In Sect. 1.3 we describe the HCS dataset, and present the computational campaign. Then, in Sect. 1.4 preliminary computational results are presented and commented. Observations about future researches conclude the paper.

1.2 The HCS Problem

In this paper we address a relevant optimization problem arising in HCS. Given a planning horizon W, which is a week in the considered experiments, a set of patients with an associated *care plan*, i.e. weekly requests each of them demanding a specific ability or *skill* to be operated, and a set of operators also characterized by a specific skill, the problem asks to schedule the patient's request during the week, assign the operators to the patients by taking into account the compatibility between request and operator skills, and determine the tour each operator has to perform in every day of the week. Each tour must start at the operator's premises and come back to the operator's premises.

Specifically, the *care plan* associated with patient j specifies the type and, for each type, the number of visits required by j in the planning horizon W. Two types of visits are considered: *ordinary requests* (requiring an ability or *skill 1*), and *palliative requests* (requiring an ability or *skill 2*). Accordingly, it is assumed that each operator has skill 1 or skill 2, and that a hierarchical structure of the skills exists, such that an operator with skill 2 can work all the requests, whereas operators with skill 1 can work only requests of skill 1.

In the considered HCS problem the scheduling of the patient requests in W, the operator assignment and the routing decisions are offered through a new modelling device, called *pattern*. We assume in fact that the patient's requests are operated according to a set P of a priori given patterns. Specifically, for each pattern $p \in P$ we define $p(d) = 0$ if no service is offered at day d, while it is $p(d) = 1$ or $p(d) = 2$ if a visit of skill 1 or 2, respectively, is operated according to pattern p on day d. Only one visit per day can be operated. Several pattern generation approaches can be proposed to generate a subset P of patterns rather than considering the entire set of all possible patterns. The motivation to generate a good but limited set of patterns stems from the fact that the cardinality of P influences the size of the resulting optimization models. In this paper we refer to a flow pattern generation approach, which is based on the solution of an auxiliary network flow problem, and which proved to be very effective in selecting a small number of patterns of good quality, according to the results in [5].

Given the input data above, the studied HCS problem consists in assigning a pattern from P to each patient j, so scheduling the requests of j, expressed by his/her care plan, during the planning horizon (*care plan scheduling*), in assigning operators to each patient j, for each day where a request of j has been scheduled (*operator assignment*), and in determining the tour of each operator for each scheduled

day (*routing decisions*). In addressing these three groups of decisions, the skill constraints, that is the compatibility between the skills associated with the patient's requests and the skills of the operators, have to be taken into account as well as other relevant Quality of Service requisites.

Observe that the Home Care context under investigation involves joint assignment, scheduling and routing decisions over W. In the state-of-the-art literature Home Care problems are usually solved in cascade: first the operators are assigned to the patients; second, the schedule of each operator is determined. Some optimization models that extend Vehicle Routing Problem (VRP) formulations have been proposed, but generally they deal with a daily planning horizon. To the best of our knowledge there are only three exceptions [2, 6, 10]. However, no exact approach is proposed there to solve the overall problem, but two-stage solution approaches are presented. In fact, in the literature tailored metaheuristic approaches are usually proposed to solve Home Care problems rather than exact approaches.

On the other hand, as outlined before, here the Home Care problem is solved by jointly addressing assignment, scheduling and routing decisions over W, by suitably generalizing the VRP, and specifically the Skill VRP [3, 4], from which it inherits the skill based structure.

New Integer Linear Programming (ILP) formulations have been proposed in [5] to formulate the stated HCS problem, by considering two balancing objective functions, and preliminary computational results have been reported in a deterministic setting. Let O denote the set of the operators available in W, while O_d be the subset of the operators available on day d, for each d in W. Let t_{ij} denote the traveling time from patient i to patient j along the link (i, j) of the logistics network, with A denoting the link set. Finally, let t'_j be the service time at patient j, and D_t indicate the workday length of operator t. Then the *maxmin* objective function can be defined as follows:

$$\begin{aligned} \max\ & m \\ & D_{td} = \sum_{(i,j)\in A} (t_{ij} + t'_j) \cdot x_{ij}^{td},\ \forall d \in W, \forall t \in O_d \\ & \frac{\sum_{d\in W} D_{td}}{|W| \cdot D_t} \geq m, \qquad\qquad \forall t \in O, \end{aligned}$$

where $|W|$ is used to denote the width of the planning horizon W. The decision variables x_{ij}^{td} take value 1 if the operator t travels along (i, j) on day d, and 0 otherwise. Therefore, D_{td} represents the workload of operator t on day d, expressed as the sum of the service times and the traveling times on day d. m is an auxiliary variable which, in a standard way, is introduced to linearize the objective function. In fact, m estimates from below the utilization factor of each operator, expressed as the weekly workload of the operator over his/her maximum possible workload in W: by maximizing m, then the model maximizes the minimum operator utilization factor. In a similar way we define the alternative balancing objective function *minmax*: in such a case, we minimize the auxiliary variable m which, now, estimates from above the utilization factor of each operator.

The aim of this paper is to enhance the computational results in [5] by performing a deeper comparison of the criteria *maxmin* and *minmax*, also investigating their impact on efficiency indicators. In the simulation setting, the robustness of the HCS solutions returned by the two balancing criteria will be stressed under scenarios of service time and traveling time variability. This will be the subject of the next two sections.

1.3 The HCS Dataset

The real data used in this work have been provided by one of the largest Italian public medical care unit operating in the north of Italy, and they have been already used in [8]. The HCS instances are characterized by a geographical area which comprises five or eight municipalities where patients are located. In regards to the patients, we selected 2 weeks in the time period (2004–2008), i.e. a week in January 2006 (hereafter denoted by *January 2006*) and a week in April 2007 (hereafter denoted by *April 2007*), and we then selected subsets of patients with a care profile in that week. Specifically, for the January 2006 week, patients are 40 or 60, whereas for the April 2007 week, patients are 50 or 80. Patient's demand had been computed by looking at the scheduling implemented by the provider: specifically, for each skill, the requested number of visits in our instances is set equal to the real number of visits performed by operators of that particular skill. This choice is supported by the fact that the provider never used operators with skill different from the skill required by a visit. As already indicated, two skills are considered for operators and patient's requests: *ordinary*, corresponding to skill 1, and *palliative*, corresponding to skill 2. The geographical area under consideration is characterized by 11 operators and a subset of them is selected in our instances according to the number of patients: when the number of patients is 40, 4 operators are chosen; when the patients are 50 or 60, the number of operators is fixed to 5, while for 80 patients 6 operators are selected. In all the instances only one operator of skill 2 (with workday duration equal to 6 h) is selected, while the remaining operators are all characterized by a workday duration of 8 h and skill 1. For a given combination of number of municipalities, number of patients and number of operators, three instances are generated by randomly selecting the desired number of patients among the available patients. The instances are thus identified by a string reporting the following fields separated by a "-" character: the week, the number of municipalities, the number of patients, the number of operators, the instance identifier in the group (i.e. 0, 1 or 2) and the objective function used which can be *maxmin* or *minmax*. As an example "Jan06-5-40-4-0-maxmin" refers to a week in January 2006, 5 municipalities, 40 patients, 4 operators, instance number 0 and *maxmin* objective function. Summarizing, for each of the 2 weeks, 2 values for the number of municipalities are combined with 2 values for the number of patients and for each of these combinations 3 instances are generated, thus giving rise to 12 instances for each week. The resulting 24 instances are run with the 2 alternative objective functions.

In all the generated instances, the traveling times t_{ij} have been computed via Google Maps for the inter-municipalities distances, while they have been set equal to 3 min for the intra-municipalities distances, consistently with the provider indications. Furthermore, according to the medical care unit indications, the service time has been fixed to 30 min (i.e. $t'_j = 30$ min).

The 48 optimization runs have been performed on a AMD Opteron(tm) Dual Core Processor 246 (CPU MHz 1991.060). The solver is CPLEX 12.4 with a time limit of 12 h and a memory limit for the branch and bound tree of 1 GB. On the other hand for the simulation experiments, which are based on a discrete event simulation model integrated with the optimization models, VBA has been used as the integration environment, and Rockwell Arena13 as the simulation platform. Referring to the simulation length, we performed 30 simulations runs for each instance, which is a fairly large number [7]. In total we thus performed 1,410 simulation runs.

1.4 Computational Results

1.4.1 Optimization Results

Tables 1.1 and 1.2 report a comparison between the solutions obtained with the two alternative objective functions in terms of some QoS indicators: (1) the Coefficient of Variation *CV*, which measures the day-by-day variability of the operator utilization factor; (2) *WeeklyST*; (3) *WeeklyTT*; and (4) *WeeklyWT*. The last three indicators refer respectively to the service time, the travelling time, and the workload of the operator in W, over his/her maximum possible workload. Therefore, *WeeklyWT* represents the operator utilization factor. For each quality indicator the mean value computed over the operators and the width of the range between the maximum value of the indicator and the minimum value of the indicator are given respectively as "Mean" and "Range" columns. For the traveling time, the percentage of the total travel time with respect to the total working time is also given in column "All". For all the quality indicators except for *CV*, mean values are given as percentage.

To provide a more formal definition of *CV*, let $\overline{D}_t$ be the average daily workload of operator t, i.e. $\overline{D}_t = \sum_{d=1}^{|W|} D_{td}/|W|$ and denote with $S(D_t)$ the standard deviation of the daily workload of operator t. Then the Coefficient of Variation of operator t is defined as follows:

$$CV_t = \frac{S(D_t)}{\overline{D}_t}.$$

In Tables 1.1 and 1.2, the "Mean" and "Range" columns under the multicolumns *CV* refer, respectively, to the mean value of CV_t over t, and to the difference between the maximum and the minimum of such values over t.

The data in Tables 1.1 and 1.2 correspond to the best solutions given by the optimization solver within the given time limit; these solutions are very close to

Table 1.1 January 2006 – optimization results (values in % except for *CV*)

	CV		WeeklyST		WeeklyTT		WeeklyWT		
	Mean	Range	Mean	Range	Mean	Range	All	Mean	Range
Jan06-5-40-4-0-maxmin	1.06	0.34	18.75	2.50	7.80	2.28	7.87	26.55	0.26
Jan06-5-40-4-0-minmax	1.46	0.53	18.65	0.42	3.90	0.13	3.90	22.54	0.54
Jan06-5-40-4-1-maxmin	0.99	0.69	18.85	4.17	7.95	1.17	7.96	26.80	3.63
Jan06-5-40-4-1-minmax	1.17	1.14	18.85	5.42	6.91	2.46	6.89	25.76	6.79
Jan06-5-40-4-2-maxmin	1.06	0.63	17.92	5.42	7.70	0.92	7.72	25.61	4.63
Jan06-5-40-4-2-minmax	1.15	1.19	17.92	5.42	6.84	0.79	6.82	24.76	6.21
Jan06-5-60-5-0-maxmin	1.09	0.59	21.83	12.92	8.38	1.42	8.35	30.21	12.92
Jan06-5-60-5-0-minmax	1.39	1.07	21.83	17.92	8.84	3.71	8.88	30.68	19.79
Jan06-5-60-5-1-maxmin	1.15	0.88	20.75	18.75	7.78	1.63	7.73	28.53	19.13
Jan06-5-60-5-1-minmax	1.51	1.28	20.75	21.25	7.88	2.54	7.85	28.63	22.46
Jan06-5-60-5-2-maxmin	1.14	0.54	20.92	3.33	7.88	2.88	7.98	28.79	0.54
Jan06-5-60-5-2-minmax	1.37	0.36	20.92	3.33	6.64	3.29	6.76	27.56	0.21
Jan06-8-40-4-0-maxmin	0.97	0.61	20.42	2.92	10.34	2.94	10.42	30.75	0.07
Jan06-8-40-4-0-minmax	1.03	0.96	20.31	1.25	6.92	1.29	6.89	27.23	0.33
Jan06-8-40-4-1-maxmin	0.93	0.84	20.52	17.08	9.33	0.10	9.33	29.85	17.03
Jan06-8-40-4-1-minmax	1.06	0.74	20.52	27.08	8.47	7.17	8.46	28.99	31.96
Jan06-8-40-4-2-maxmin	1.13	0.39	19.06	8.75	7.79	3.42	7.92	26.85	5.33
Jan06-8-40-4-2-minmax	1.25	0.54	19.06	12.50	6.66	3.67	6.76	25.72	11.71
Jan06-8-60-5-0-maxmin	1.04	0.34	22.83	5.42	10.66	1.72	10.71	33.50	3.69
Jan06-8-60-5-0-minmax	1.12	0.87	22.83	6.67	9.11	4.46	9.13	31.94	8.68
Jan06-8-60-5-1-maxmin	1.21	0.44	21.25	11.25	8.64	1.90	8.70	29.89	9.39
Jan06-8-60-5-1-minmax	1.33	0.72	21.25	16.25	8.15	6.21	8.27	29.40	15.63
Jan06-8-60-5-2-maxmin	1.12	0.56	22.42	14.58	10.54	1.25	10.53	32.96	13.94
Jan06-8-60-5-2-minmax	1.10	1.10	22.42	22.08	9.51	7.25	9.50	31.92	25.82

the optimum ones except for instance Apr07-5-80-6-1. Furthermore, computational results on instance Apr07-8-80-6-0 are not reported since *minmax* failed to provide a feasible solution within the time limit.

The main achievements related to this deterministic scenario can be summarized as follows. By considering *WeeklyWT*, i.e. the operator utilization factor, its mean value for the *maxmin* criterion is usually greater than the mean value returned by the *minmax* criterion, although their difference is often small. On the other hand, the range of *WeeklyWT* for the *maxmin* solutions is almost always substantially smaller than the one returned by *minmax*. The same kind of relationship can be observed for the day-by-day variability of the operator utilization factor. For *CV*, this relationship is true also considering the mean values.

Concerning the two main components contributing to the operator workload it is possible to observe that, whereas the mean percentage service time is about the same for the two objective functions, the range of *WeeklyST* is often substantially smaller for the *maxmin* solutions. A similar trend, although in a weaker form, can be observed by considering the range of *WeeklyTT*. However, as expected, the total

Table 1.2 April 2007 – optimization results (values in % except for *CV*)

	CV		WeeklyST		WeeklyTT		WeeklyWT		
	Mean	Range	Mean	Range	Mean	Range	All	Mean	Range
Apr07-5-50-5-0-maxmin	0.53	0.61	25.42	5.83	8.99	4.51	9.09	34.41	1.32
Apr07-5-50-5-0-minmax	0.68	1.14	25.42	4.58	7.73	3.51	7.84	33.15	2.32
Apr07-5-50-5-1-maxmin	0.59	0.29	25.33	2.92	8.13	3.00	8.20	33.47	0.17
Apr07-5-50-5-1-minmax	0.78	1.05	25.33	2.92	4.81	2.00	4.81	30.14	2.88
Apr07-5-50-5-2-maxmin	0.53	0.46	28.25	10.00	9.69	3.47	9.77	37.94	6.53
Apr07-5-50-5-2-minmax	0.58	0.68	28.25	16.25	7.16	5.58	7.21	35.41	18.03
Apr07-5-80-6-0-maxmin	0.55	0.43	33.19	17.92	11.07	2.58	11.07	44.26	16.42
Apr07-5-80-6-0-minmax	0.89	1.10	33.19	25.42	9.22	5.29	9.21	42.41	28.74
Apr07-5-80-6-1-maxmin	0.48	0.37	33.82	4.17	13.01	4.26	13.13	46.83	0.37
Apr07-5-80-6-1-minmax	0.69	0.41	34.03	16.67	10.64	6.92	10.72	44.67	18.17
Apr07-5-80-6-2-maxmin	0.61	0.60	33.47	13.33	11.78	2.46	11.76	45.25	12.28
Apr07-5-80-6-2-minmax	0.87	1.53	33.47	27.08	9.51	9.00	9.53	42.98	32.19
Apr07-8-50-5-0-maxmin	0.55	0.32	27.42	8.33	11.58	1.67	11.61	39.00	6.67
Apr07-8-50-5-0-minmax	0.69	0.50	27.42	8.33	9.47	2.85	9.57	36.89	6.49
Apr07-8-50-5-1-maxmin	0.59	0.72	26.00	7.50	9.55	4.56	9.41	35.55	8.39
Apr07-8-50-5-1-minmax	0.87	1.20	26.00	10.00	8.12	3.63	8.08	34.12	13.03
Apr07-8-50-5-2-maxmin	0.79	0.34	21.33	2.50	8.19	2.54	8.21	29.53	0.08
Apr07-8-50-5-2-minmax	0.82	0.51	21.33	0.42	4.96	1.24	5.00	26.29	0.87
Apr07-8-80-6-1-maxmin	0.62	0.45	33.26	30.83	13.00	3.44	12.90	46.27	33.03
Apr07-8-80-6-1-minmax	0.70	0.58	33.40	55.42	11.02	15.79	11.01	44.43	64.81
Apr07-8-80-6-2-maxmin	0.64	0.62	31.39	27.08	12.55	6.72	12.33	43.94	32.56
Apr07-8-80-6-2-minmax	0.89	0.71	31.39	33.33	10.04	4.67	9.91	41.43	37.79

traveling time spent by the operators during the week is usually smaller in the solutions returned by the *minmax* criterion.

Therefore, as already outlined, in a deterministic setting and for the tested instances, *maxmin* appears to be preferable in balancing the operator percentage traveling time and the operator percentage service time, and therefore the operator utilization factor. This is true not only by looking at the overall planning horizon, but also at a daily level. Such stronger equity achievements are obtained for not too high a price in the increased average quality indicators. On the other hand, *minmax* always returns solutions with the smaller total travelled time for the operators. Therefore, it appears to be more suitable for the minimization of the operating costs, which are measured here in terms of travelling costs.

1.4.2 Simulation Results

The simulation model reproduces the activities of the operators for each day of the week. However, now the travel times t_{ij} and the service times t'_j are realization of random variables. Concerning their randomness, since the provider did not

collect data relevant to service and travel times, we could neither use empirical distributions, nor fit theoretical distributions to real data. Hence, to randomize these times we have multiplied the standard values of the service and travel times by numbers randomly sampled from triangular distributions (called TRIA), according to the formulas below, where N denotes the set of the patients:

$$\tilde{t'}_j = t'_j \bullet TRIA(0.9, 1, 1.1),\ \forall j \in N$$
$$\tilde{t}_{ij} = t_{ij} \bullet TRIA(0.8, 1, 1.5),\ \forall (i, j) \in A.$$

The use of triangular distributions is coherent with the recommendations of [7], who suggest using finite distributions to avoid sampling excessively large and meaningless times. In addition, triangular distributions have been successfully applied by [1] to model travel times in a similar setting.

The simulation experiments have been conducted with a threefold aim:

- To verify whether the randomness of the service and travel times can lead to overtime, and therefore to additional costs for the provider; observe that overtime could happen especially in case of not evenly balanced workload among the operators, during the week and/or across the days;
- To understand if *maxmin* and *minmax* lead to solutions that significantly differ in terms of overtime;
- To determine how the randomness of travel and service times impacts on the quality indicators presented in Sect. 1.4.1.

Referring to the first point, for each of the 23 instances and for both *maxmin* and *minmax*, we have calculated the mean values and the standard errors, across the 30 replications, of the total weekly overtime. Hence, for each instance, we have performed a one-sided independent t-test to ascertain whether the mean value (M), across 30 replications, of the weekly overtime ($AllWeeklyOT$) could be considered significantly larger than zero. In other terms, we have tested the alternative hypothesis $H1 : M(AllWeeklyOT) > 0$ against the null hypothesis $H0 : M(AllWeeklyOT) = 0$. For all these tests we were not able to reject the null hypotheses at a significance level $\alpha = 0.05$. It led us to conclude that for all the instances the overtime is never significantly different from zero, regardless of the objective functions considered. Actually, even in the worst case (i.e. considering the maximum of the individual replication maxima values), the overtime is smaller than 4 min/week. This fact implies that both objective functions allow avoiding undesirable daily workload peaks that would lead to overtimes. It is worth to observe, however, that for the investigated instances the operator utilization factor is rather small (see Tables 1.1 and 1.2), and therefore overtimes may be difficult to emerge. Since the overtime is always very close to zero for the tested instances, the comparison between the overtimes associated with *maxmin* and *minmax* is not meaningful.

It does make sense, instead, to assess the impact that the time randomness may have on the system performance. We have thus calculated, for all the quality indicators in Sect. 1.4.1, the mean, the standard deviation and the 95 % two-sided

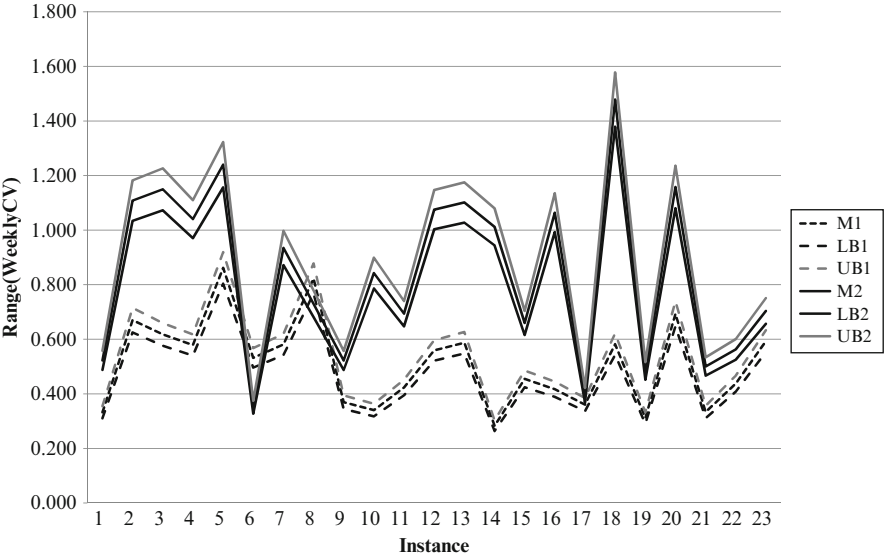


Fig. 1.1 Confidence intervals for the range of daily variability

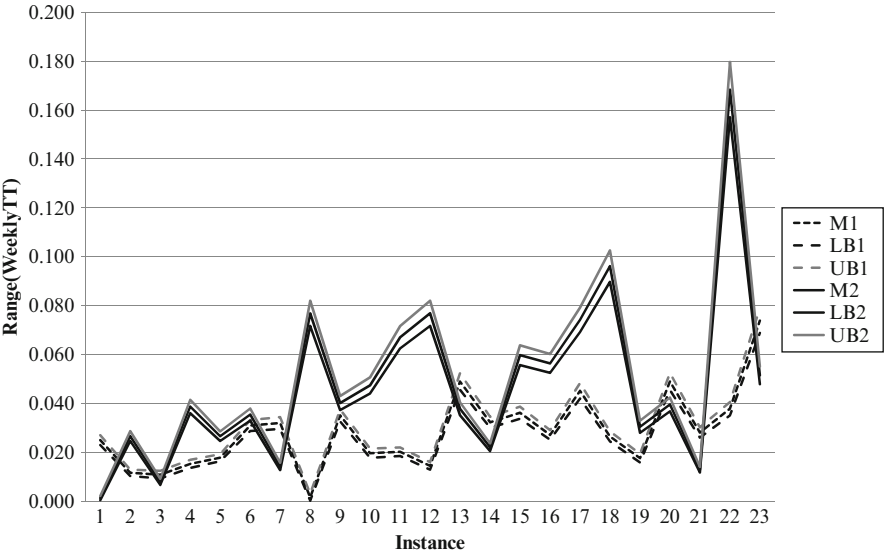


Fig. 1.2 Confidence intervals for the range of *weeklyTT*

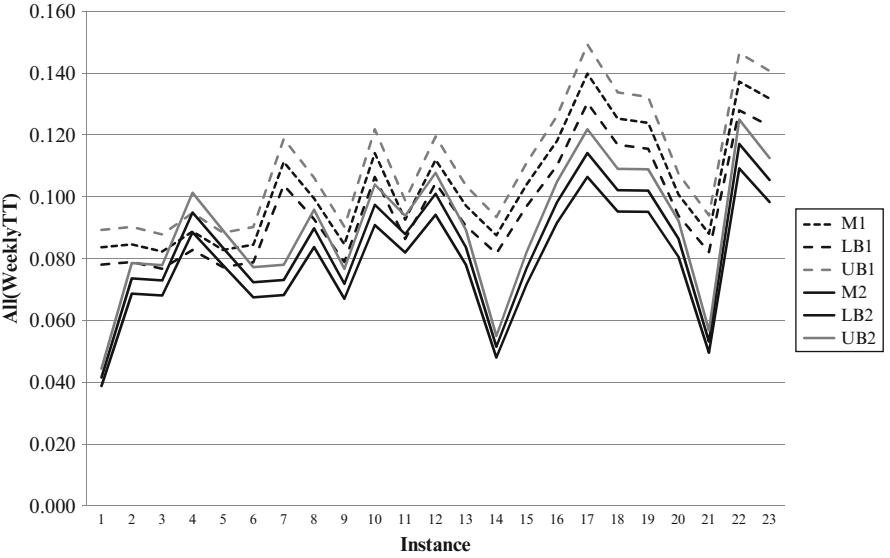


Fig. 1.3 Confidence intervals for the total percentage traveling time

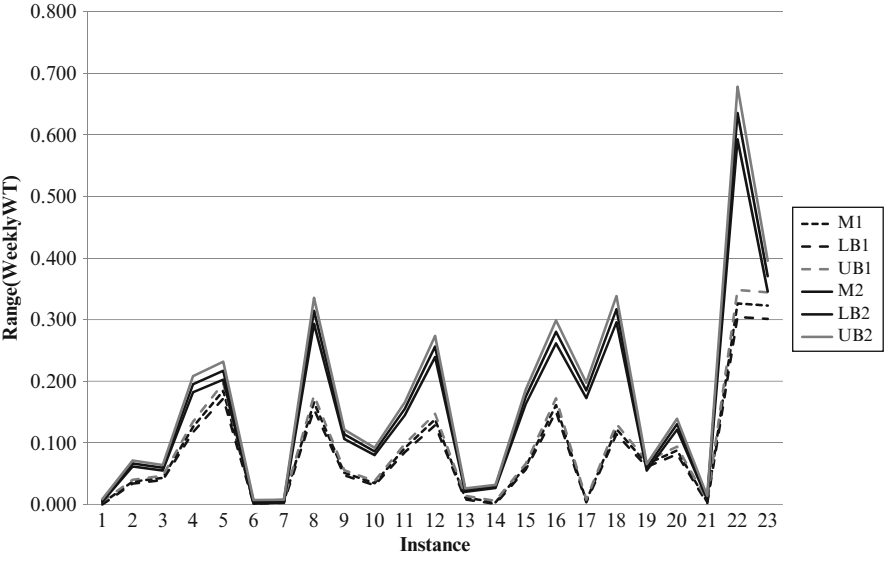


Fig. 1.4 Confidence intervals for the range of operator utilization factor

confidence intervals for the mean. Due to space constraints, hereafter we shall present only the results related to the range of *CV*, *weeklyTT* and *weeklyWT*, by adopting a graphical representation. A graph is provided also for *AllWeeklyTT*. Specifically, each graph refers to one indicator and presents, for each instance and for both the *maxmin* (1 – dashed lines) and *minmax* (2 – regular lines) objective functions: (i) the mean value of the indicator (M); and (ii) the upper (UB) and lower (LB) bound of the confidence intervals for the mean.

Concerning this last point of the simulation study, the main achievement is that, by observing the equity and the efficiency indicators of the system in a stochastic environment, and calculating the confidence intervals for each indicator, the same trend already observed in a deterministic setting appears to be confirmed also in the presence of randomness of travel and service times, as shown in the figures below.

1.5 Future Research

It is worth pointing out that the ones presented in this paper are preliminary results of a study that will be expanded in several ways, especially regarding the simulation experiments. Firstly, the simulation model will be used to assess the robustness of the HCS solutions returned by the optimization models, against the times randomness, in settings characterized by higher resource utilization levels. In these settings, in fact, deviations of the times from their expected values likely cause overtimes and can even prevent operators to ultimate their daily tours. Second, the simulation model will be used to test the output of optimization models developed in context where patients can be visited only in certain time windows. In these contexts, in fact, the times randomness in addition to lead to overtime, can prevent the operators to match their appointments. Finally, the simulation model will be used to study the performance of systems where patient-operator mismatches can occur, thereby determining the need to dynamically reschedule the tours of one or more operators.

References

1. Agnihothri, S.R., Mishra, A.K.: Cross-training decisions in field services with three job types and server–Job mismatch. Decis. Sci. **35**(2), 239–257 (2004)
2. Begur, S.V., Miller, D.M., Weaver, J.R.: An integrated spatial DSS for scheduling and routing home-health-care nurses. Interfaces **27**(4), 35–48 (1997)
3. Cappanera, P., Gouveia, L., Scutellà, M.G.: The Skill vehicle routing problem. In: Pahl, J., Reiners, T., Voss, S. (eds.) Network Optimization, 5th International Conference, INOC 2011, Hamburg, June 2011. Lecture Notes in Computer Science, vol. 6701, pp. 354–364. Springer, Berlin/Heidelberg (2011)
4. Cappanera, P., Gouveia, L., Scutellà, M.G.: Models and valid inequalities to asymmetric skill-based routing problems. EURO J. Transp. Logist (2012). doi:10.1007/s13676-012-0012-y

5. Cappanera, P., Scutellà, M.G.: Joint assignment, scheduling and routing models to home care optimization: a pattern based approach. Technical Report, TR-13-05, Dipartimento di Informatica, Università di Pisa (2013)
6. Jensen, T.S.: Application of metaheuristics to real-life scheduling problems. Ph.D. Thesis. Department of Mathematics and Computer Science, University of Southern Denmark (2012)
7. Kelton, W.D., Sadowski, R.P., Sadowski, D.A.: Simulation with ARENA, vol. 2. McGraw-Hill, Boston (2002)
8. Lanzarone, E., Matta, A.: A cost assignment policy for home care patients. Flex. Serv. Manuf. J. **24**(4), 465–495 (2012)
9. Lanzarone, E., Matta A., Sahin, E.: Operations management applied to home CareServices: the problem of assigning human resources to patients. IEEE Trans. Syst. Man. Cybern. A Syst. Hum. **42**(6) (2012)
10. Nickel, S., Schroder, M., Steeg, J.: Planning for home health care services. Technical Report, Berichte des Fraunhofer ITWM, 173 (2009)

Chapter 2
Redesigning Organ Allocation Boundaries for Liver Transplantation in the United States

Naoru Koizumi, Rajesh Ganesan, Monica Gentili, Chun-Hung Chen, Nigel Waters, Debasree DasGupta, Dennis Nicholas, Amit Patel, Divya Srinivasan, and Keith Melancon

Abstract Geographic disparities in access to and outcomes in transplantation have been a persistent problem widely discussed by transplant researchers and the transplant community. One of the alleged causes of disparities in the United States is administratively determined organ allocation boundaries that limit organ sharing across regions. This paper applies mathematical programming to construct alternative liver allocation boundaries that achieve more geographic equity in access to transplants than the current system. The performance of the optimal boundaries were evaluated and compared to that of current allocation system using discrete event simulation.

2.1 Introduction

Existing studies of organ transplant report various disparities in access to and outcomes in transplantation. Disparities have been found in terms of race, socioeconomic status, insurance type and the location of candidate's residency. While

N. Koizumi (✉) • R. Ganesan • C.-H. Chen • N. Waters • D. DasGupta • D. Nicholas
A. Patel • D. Srinivasan
George Mason University, 4400 University Drive, Fairfax, VA 22030, USA
e-mail: nkoizumi@gmu.edu; rganesan@gmu.edu; cchen9@gmu.edu; nwaters@gmu.edu; ddasgupt@masonlive.gmu.edu; dnichol4@masonlive.gmu.edu; apatelh@masonlive.gmu.edu; dsriniv2@masonlive.gmu.edu

M. Gentili
University of Salerno, Via Ponte Don Melillo, 84084 Fisciano (SA), Italy
e-mail: mgentili@unisa.it

K. Melancon
George Washington University Hospital, 2150 Pennsylvania Avenue, NW, Washington, DC 20037, USA
e-mail: jmelancon@mfa.gwu.edu

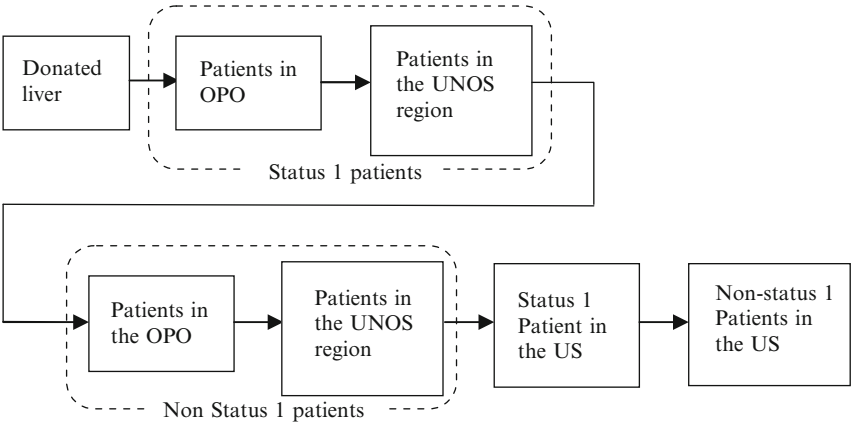
A. Matta et al. (eds.), *Proceedings of the International Conference on Health Care Systems Engineering*, Springer Proceedings in Mathematics & Statistics 61,
DOI 10.1007/978-3-319-01848-5_2,

these disparities tend to coexist, disparity associated with candidates' locations or "geographical disparity" is the first and foremost discussed. Researchers worldwide have repeatedly confirmed that the likelihood of receiving a transplant as well as pre- and post- transplant mortality rates vary significantly from region to region [1–10]. Geographic disparity in transplant access in the US has been a persistent issue ever since organ allocation became a regulated process in 1984 under the National Organ Transplant Act (NOTA). As the most important act in the history of US transplantation, NOTA created the Organ Procurement and Transplantation Network (OPTN), a public-private network of regional organ allocation offices known as Organ Procurement Organizations (OPOs) [6]. NOTA also authorized the Department of Health and Human Services (HHS) to contract with the United Network for Organ Sharing (UNOS) as the only administrative entity to govern the OPTN. At first, all organs were distributed within each OPO's service area (ibid) in order to limit cold ischemia time (CIT), i.e. the interval between organ retrieval and the time of transplantation during which an organ is preserved in a cold perfusion solution. Allocation of organs within each OPO was solely based on the length of time that each candidate had spent waiting for an organ since initial referral. In response to the concern that the waiting time varied significantly by OPO, HHS introduced a new regulation known as the "Final Rule" (42 CFR Part 121) in 1998 to "assure that allocation of scarce organs will be based on common medical criteria, not accidents of geography" (HHS, 1998b) (ibid).

As per the directives of the Final Rule, the allocation mechanism for a number of vital organs has been rectified to address the criterion of medical necessity. For liver allocation, HHS revised the Code of Federal Regulations legislating organ allocation process and, in 2002, the Model for End-Stage-Liver-Disease (MELD) scoring system was launched as a way to prioritize the candidates with a higher medical urgency. Since then, the harvested adult livers had been distributed, in principle, based on the algorithm summarized in Fig. 2.1. Thus the current organ allocation system consists of three hierarchical, geographic levels: the OPOs (a.k.a. the Donor Service Areas), the UNOS regions and the National level.

While several changes in allocation rules have been introduced to address discrepancies, transplant researchers still report that a number of key elements that determine equity in transplantation vary significantly depending on the location of a patient. This study thus developed a mathematical programming model to redesign liver allocation boundaries. The optimal boundaries were derived to maximize geographic equity in access to a transplant while maintaining efficiency in outcomes in transplantation. The model was also used to analyze which existing "kidney-only" transplant centers could be activated to improve the current liver allocations. Finally, discrete event simulation was applied to evaluate the performance of the optimal boundaries in comparison to that of the existing boundaries.

The primary data used for the analyses is UNOS's Standard Transplant Analysis and Research (STAR) Dataset that records clinical, administrative, demographic and locational information of over 40,000 adult liver transplant candidates and recipients who appeared on the wait list between 2003 and 2010.



- Status 1 patients refer to those with fulminant liver failure with a life expectancy without a liver transplant or less than 7 days.
- Within each category of patients (i.e. Status 1, MELD scores ≥ 15, MELD score < 15), a liver is offered, in principle, in the descending order of first MELD score and then waiting time.
- Extra points are added to the MELD score for those patients whose blood type is compatible to that of the available liver and those with specific clinical circumstances such as Hepatocellular Carcinoma (HCC).

Fig. 2.1 Current liver allocation system

2.2 Model Description

2.2.1 Mathematical Model

The mathematical programming approach has the twofold objective of: (i) identifying optimal locations for liver transplant centers and (ii) identifying new OPO boundaries that replace existing OPO's boundaries, which are mainly defined by political issues. Two mathematical models are proposed to achieve these objectives. Both models are described next, but, due to the current page limit, we present the mathematical formulation of only the second model.

The first model (Model 1) addresses the problem of: (a) selecting a fixed number p of transplant centers to be opened among a possible set of candidates and (b) associating a subset of donor hospitals (that define the organ acquisition area of the center) and a subset of counties (that define the service area of the center) with each opened transplant center. The model ensures that each donor hospital and each county are associated with exactly one transplant center. Moreover, the distance between a donor hospital and the associated transplant center is such that the corresponding travel time is within the CIT of the organ, and finally, the distance between the centroid of a county and the associated transplant center is not

greater than a predefined maximum threshold. The proposed model is similar to the mathematical model proposed by Bruni et al. [11] in that each selected transplant center is associated with an acquisition area and a service area. However, unlike Bruni's model, we consider an additional set of constraints to ensure that, for each opened transplant center, the ratio between the available organs (coming from the associated acquisition area) and the total number of recipients (coming from the associated service area) is greater than or equal to a fixed threshold α. The objective function of the model is the minimization of the total distance between the set of donor hospitals and the associated transplant centers plus the total distance between the county centroids and the associated transplant center.

Model 2 addresses the problem of clustering a set of transplant centers that are selected for activation (as a result of Model 1) into a predefined number of clusters. Each cluster represents an OPO. The resulting OPOs are defined so that they are balanced both in terms of the supply/demand ratio of organs and in terms of total number of transplant centers that belong to the OPO. The boundary of each OPO is determined by the union of the service areas associated with the transplant centers that belong to the OPO. Hence, one important constraint to take into account when defining the cluster is contiguity of the service areas. To achieve this aim, Model 2 takes a graph $G=(V,E)$ as an input where each vertex $i \in V$ is associated with a transplant center and there is an arc $(i,j) \in E$ between vertex i and vertex j if the corresponding service areas have a common border. Two weights are associated with each vertex i of this graph: w_i and h_i representing, respectively, the total supply and the total demand associated with the transplant center represented by the vertex. A super vertex s is added to the graph and is connected with each vertex of the graph by the set of arcs (s,i), $\forall\, i \in V$. Hence, the resulting graph is such that the total number of vertices is equal to $p+1$ and the total number of arcs depends on the solution returned by Model 1.

Model 2 looks for a spanning tree T_s of G rooted in s such that the total number of children of the root is equal to the total number of clusters that need to be defined. In this way, the vertices of each subtree T_i rooted at vertex i (i.e., one of the children of the supervertex s) represent the set of transplant centers that belong to the cluster. Connection of the subtree ensures contiguity of the service area associated with the cluster. Moreover each subtree is such that the ratio between the sum of the weights w_i associated with the vertices of the subtree and the sum of the weights h_i associated with the vertices of the subtree is greater than or equal to a predefined threshold α. The objective function of the model is the minimization of the maximum number of vertices in each of the resulting subtrees, ensuring that the resulting clusters are also balanced in terms of total number of transplant centers that belong to them.

Let $O=\{1,2,\ldots,l\}$ be the index set of the clusters that need to be defined. Then the proposed formulation is a Miller-Tucker-Zemlin (MTZ) formulation [12] where we considered the following set of variables:

- Variable y_{ik} is a binary variable that is equal to one if vertex $i \in V$ belongs to cluster $k \in O$ and is equal to 0 otherwise;

- Variable x_{ijk} is a binary variable that is equal to one if arc $(i,j) \in E$, that connects vertices i and j in the cluster k, is selected to be in the spanning tree and is equal to 0 otherwise;
- Variable u_i, defined on each vertex $i \in V$, assigns a label to each vertex of the graph. In particular, such a labeling ensures any directed arc that belongs to the optimum spanning tree goes from a vertex with a lower label to a vertex with a higher label.

Hence, variables y_{ik} are used to define the clusters, while variables u_i and x_{ijk} are used to define the final spanning tree.

The resulting Model 2 is the following:

$$\min \max \left(\sum_{i \in V} y_{ik} \right) \tag{2.1}$$

$$\sum_{(s,j) \in E} x_{sjk} = 1 \qquad \forall k \in O \tag{2.2}$$

$$\sum_{k \in O} \sum_{(i,j) \in E} x_{ijk} = 1 \qquad \forall j \in V, j \neq s \tag{2.3}$$

$$\sum_{k \in O} x_{ijk} \leq 1 \qquad \forall (i,j) \in E \tag{2.4}$$

$$x_{ijk} \leq y_{ik} \qquad \forall (i,j) \in E, i \neq s, \forall k \in O \tag{2.5}$$

$$y_{ik} \leq \sum_{(i,j) \in E} x_{ijk} \qquad \forall i \in V, i \neq s, \forall k \in O \tag{2.6}$$

$$u_s = 0 \tag{2.7}$$

$$1 \leq u_i \leq p \qquad \forall i \in V, i \neq s \tag{2.8}$$

$$(p+1)x_{ijk} + u_i - u_j + (p-1)x_{jik} \leq \forall (i,j) \in E, i \neq s, \forall k \in O \tag{2.9}$$

$$\sum_{k \in O} y_{ik} = 1 \qquad \forall i \in V, i \neq s \tag{2.10}$$

$$\sum_{i \in V, i \neq s} w_i y_{ik} \leq \alpha \sum_{i \in V, i \neq s} h_i y_{ik} \qquad \forall k \in O \tag{2.11}$$

$$\sum_{i \in V, i \neq s} y_{ik} \geq 1 \qquad \forall k \in O \tag{2.12}$$

The objective function [1] minimizes the maximum cardinality of the resulting clusters. Constraints [2] ensure that the total number of children of the root s is equal to the total number of clusters that need to be defined. Constraints [3] ensure that each vertex has exactly one entering arc. Each arc can be associated with at most one cluster, which is ensured by constraints [4]. Constraints [11] and [12] are logical constraints linking the binary variables. The spanning tree is defined by the classical MTZ constraints [13, 5]. Constraints [6] ensure that each vertex belongs exactly to one cluster. The structure of the cluster is defined by constraints [14] and [15]. In particular, each cluster must not be empty (constraints [15]) and total supply/demand ratio at each cluster must be greater than or equal to a predefined threshold α (constraints [14]).

Our model extends a handful of studies [11, 14, 16, 17] that investigate optimal boundaries for organ allocation using a mathematical approach. Most previous models [14, 16, 17] are based on a set covering mathematical formulation of which feasible sets are represented by all possible regional configurations resulting from different clusters of OPOs. This approach tends to be computationally very demanding. The MTZ formulation we proposed for Model 2 solves a constrained version of a spanning tree problem. This approach enabled us to solve the problem to optimality through the available commercial solvers, Cplex and Gurobi, in a reasonable amount of time. In this study, all mathematical formulations were coded in AMPL and solved using CPLEX 11 and Gurobi 5.1 on a 2.4GHz Intel Core2 Q6600 processor.

2.2.2 Discrete Event Simulation

A discrete event simulation (DES) was run to evaluate the performance of the boundaries developed by the mathematical model. The key events and the parameters used to frame the simulation were: (i) patient arrival rate; (ii) length of time registered as a transplant candidate; (iii) rates of death and drop-out while waiting for an organ; (iv) rate of candidates receiving a transplant and (v) liver arrival rate. Both livers and patients enter the system with certain characteristics used in "match-run", the process to match a donor to a recipient. Those characteristics included blood type, MELD score and the category and age.

The usefulness of DES in evaluating organ allocation policies/scenarios is already well established [13, 15, 18, 19]. In fact, DES-based simulation software, SAM (Simulated Allocation Model), was developed by the Scientific Registry of

Transplant Recipients (SRTR) and has been used by UNOS to evaluate the impacts of various organ allocation policy alternatives. However, SAM and other existing DES models does not allow for the explicit consideration of geography thereby limiting the simulation of the impacts of boundary changes. Our study developed a simulation model that simulates various allocation boundary scenarios in a more direct and overt manner.

The first task of the baseline simulation modeling was to generate recipient and donor data. As described above, each OPO is comprised of a set of counties, each of which is identified using a unique FIPS code. Each county is characterized by the historical patterns of recipient and donor counts per year and the arrival rates per day of the year, which also follows the historical proportions. Using the historical numbers and proportions from 2003 to 2009, the simulation was able to generate both recipient and donor data for 2010, which was then validated using the actual data from 2010.

The next step was to allocate organs to recipients using the current UNOS and OPO boundaries and using the new OPO boundaries obtained from Model 2. First, candidates waiting as of January 1st 2010 were generated from the actual STAR data. This data was used to initialize the simulation of liver allocation. Livers were then allocated using the current system of allocation in which Status 1 patients were given the top priority followed by patients with MELD > 15 and MELD < 15 (Fig. 2.1). The performance metrics were the waiting time for transplants for status 1, MELD < 15 and MELD > 15 transplant recipients and the geographical disparity measured in terms of the mean squared error, which is calculated as the deviation of the supply/demand ratio of the OPOs from the mean supply/demand ratio. Since there were about 12,000 candidates in the waitlist on Jan 1st 2010 and about 10,000 candidates joined the list in 2010, the supply/demand ratio for 5,000 donors in 2010 is about 0.23. After accounting for death while waiting (12.8 %), the supply/demand ratio is about 0.25 (including both waiting list and new candidates in 2010). The simulation was written and run in MATLAB.

2.3 Results

2.3.1 Results of the Mathematical Model

The Model 1 analysis revealed that opening additional 103 liver transplant centers at kidney-only transplant centers, while ensuring equity in terms of provided service, would marginally increase the efficiency in liver transplantation of the current system. In contrast, the result suggested that opening 61 new liver programs at existing kidney-only transplant centers while keeping 62 of the existing 123 liver transplant centers can substantially reduce waiting time and graft failure.

Model 2 clustered the existing 123 transplant centers into 58 OPO clusters. Figure 2.2 shows the current OPO boundaries in color and the new OPO boundaries

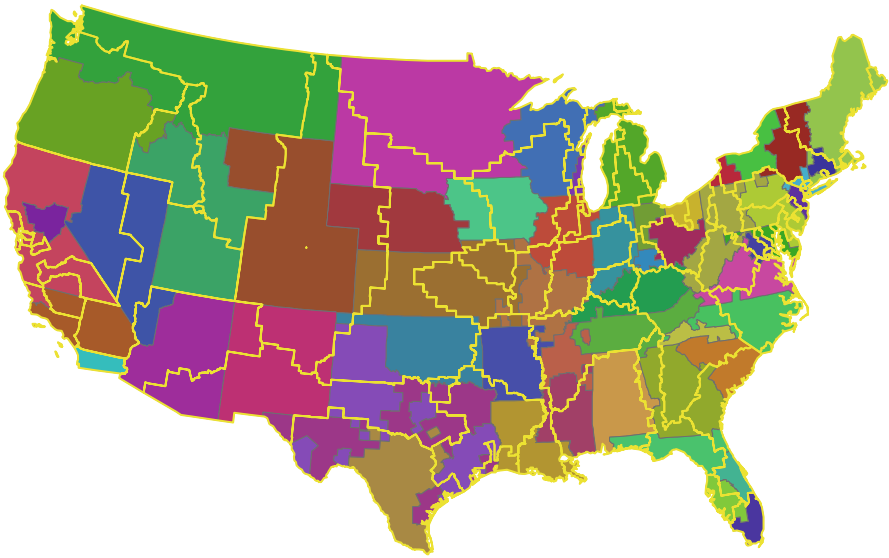


Fig. 2.2 Current (in *color*) and optimal (in *yellow lines*) OPO boundaries

suggested by the model in yellow lines. The resulting boundaries differ considerably from the actual boundaries although, in several OPOs, the boundaries coincide with actual boundaries fairly well.

2.3.2 Results of the Discrete Event Simulation

Figure 2.3 presents the actual and the simulated numbers of recipients per county arranged in ascending order in OPO #12 in 2010, which was randomly picked among other OPO's. The figures show a great deal of similarity, which was verified using the Kolmogorov-Smirnov test for equality of the probability distributions. Likewise, every OPO's recipients and donors were simulated for each county, and the characteristics described above were assigned.

Following the common simulation practice, 30 simulations were run to obtain the performance metrics of the current allocation scheme under each set of the OPO boundaries. Our simulation analysis indicates that it leads to sufficiently tight confidence interval for the estimation. Figures 2.4 and 2.5 show the distributions of the (a) number of counties per OPO, (b) donor counts or supply of liver per OPO, (c) candidate counts or demand of liver per OPO, and (d) supply/demand ratio per OPO for both the current and the new OPO boundaries respectively. Comparison of (b), (c) and (d) in the two figures reveals that, under the current boundaries (Fig. 2.4), there are several OPOs in which the supply/demand ratio was disproportionately higher than that in other OPOs. This is one of the primary causes of geographical

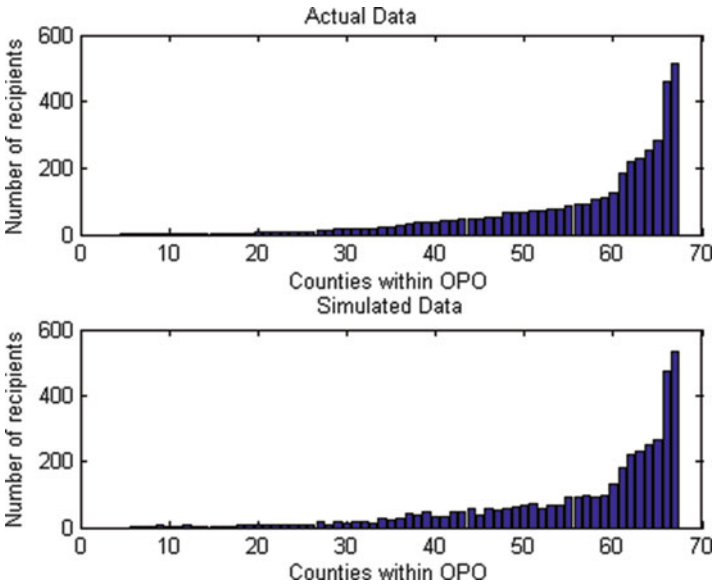


Fig. 2.3 Actual and simulated number of recipients per county for OPO #12 in 2010

disparity. With the new boundaries (Fig. 2.5), the number of instances of such a disproportionate supply/demand ratio is less frequent due to the balancing of supply/demand ratios across OPOs.

Table 2.1 indicates that the distribution of the supply/demand ratio is statistically more uniform with the new OPO boundaries. The mean supply/demand ratio is much closer to the total supply/demand ratio of 0.25 under the new boundaries. The standard deviation of the ratios dropped in the new boundary supporting the claim that the new OPOs have a more uniform supply/demand ratio. The mean square error, which is the mean of the squared deviation of errors (error in ratio of OPO i = supply/demand ratio of OPO i – mean supply/demand ratio of all OPOs) was also 15 % less with the new boundaries.

Table 2.1 presents waiting time for a transplant under the current and the new OPO boundaries. Waiting time is presented for each severity category, i.e., for status 1, MELD <15 and MELD >15. As the table shows, mean and median wait time decreased with the new boundaries, most of which is attributable to the wait time among status 1 and MELD < 15 candidates. Mean and median wait time slightly improved for the MELD > 15 category of candidates. However, neither the mean nor the standard deviation was statistically significantly different from that obtained under the existing boundaries. In terms of the number of transplants, MELD > 15 candidates had the highest number of transplants, accounting for about 85 % of the recipients of the 5,000 donors appeared in 2010. One can conclude that the new OPO boundaries are successful in alleviating geographic disparity while reducing wait time significantly among status1 and MELD < 15 candidates.

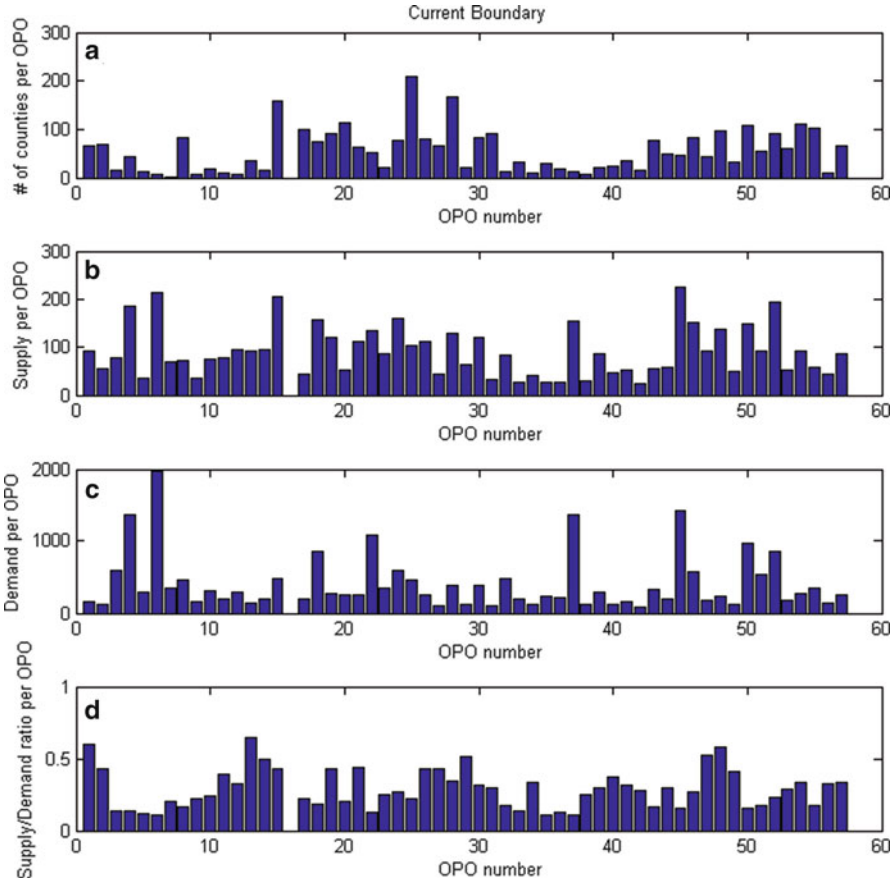


Fig. 2.4 Current boundaries: (**a**) number of counties per OPO, (**b**) supply per OPO, (**c**) demand per OPO, and (**d**) supply/demand ratio per OPO

Figures 2.4 and 2.5 shows that some of the new OPOs are larger containing more counties. The observation corresponds to the map in Fig. 2.2 in which some of the new OPOs are larger, especially in the mid-west region of the US.

2.4 Conclusions

Mathematical programming was used to derive new liver allocation boundaries that maximize geographic equity in access to liver transplant. Our study extended past studies on this topic by introducing MTZ formulation, which enabled us to solve the problem of optimality through the available commercial solvers in a reasonable amount of time. Our study is also different from past studies in that the performance

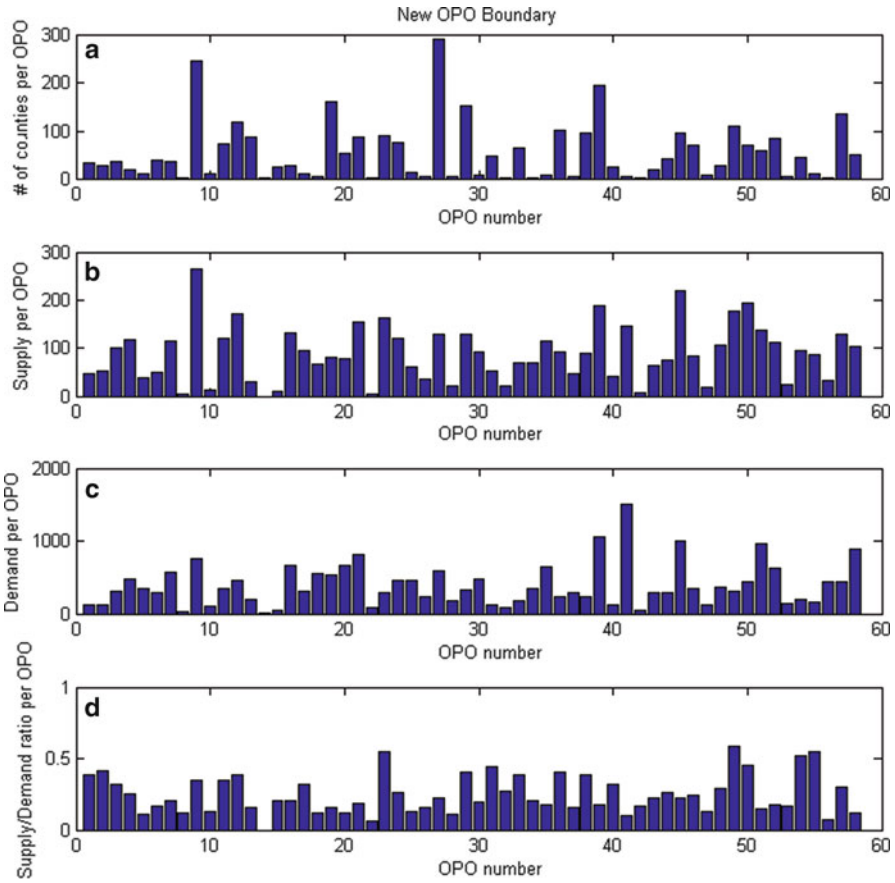


Fig. 2.5 New boundaries: (**a**) number of counties per OPO, (**b**) supply per OPO, (**c**) demand per OPO, and (**d**) supply/demand ratio per OPO

of new boundaries was evaluated dynamically using discrete event simulation. The boundaries derived from the mathematical model differed significantly form the current boundaries and our simulation results confirmed that the new boundaries could achieve a more equal supply demand ratio across OPOs and reduction in waiting time.

We note several directions for future research. First, it would be interesting to explore different equity measure definitions both for Model 1 and Model 2 in order to take into account additional aspects such as those considered by Kong et al. [14] and by Demirci et al. [16]. We are also interested in exploring the possibility of adapting our Model 2 to solve the problem of clustering the OPOs into UNOS regions so that the final allocation system represents the hierarchical system as presently implemented.

Table 2.1 Performance metrics

Performance metric	Current OPO boundaries	New OPO boundaries	Percentage change (%)	Increase/ decrease
Waiting time for Transplant (in days)				
Status 1				
Median	1	1	0.1	Same
Mean	2.3	1.4	39.1	Decrease
Standard deviation	4.8	2.4	50.0	Decrease
MELD<15				
Median	1,139	940	17.5	Decrease
Mean	1,211	1,073	11.4	Decrease
Standard deviation	944	831	12.0	Decrease
MELD>15				
Median	300	278	7.3	Decrease
Mean	508	506	0.4	Decrease
Standard deviation	561	572	−2.0	Increase
Geographical disparity				
Supply/demand ratio among 58 OPO				
Median	0.2700	0.2034	24.7	Decrease
Mean	0.2801	0.2453	12.4	Decrease
Standard deviation	0.1442	0.1330	7.8	Decrease
Maximum	0.6479	0.5803	10.4	Decrease
Mean squared error	0.0204	0.0174	15.0	Decrease

References

1. Ashby, V.B., Kalbfleisch, J.D., Wolfe, R.A., Lind, M.J., Port, F.K., Leichtman, A.B.: Geographic variability in access to primary kidney transplantation in the United States, 1996–2005. Am. J. Transplant. **7**(2), 1412–1423 (2007)
2. Barshes, N.R., Becker, N.S., Washburn, W.K., Halff, G.A., Aloia, T.A., Goss, J.A.: Geographic disparities in deceased donor liver transplantation within a single UNOS region. Liver Transpl. **13**, 747–751 (2007)
3. Brown, K.A., Moonka, D.: Liver transplantation. Curr. Opin. Gastroenterol. **20**(3), 264–269 (2004)
4. Brown, R.S., Lake, J.R.: The survival impact of liver transplantation in the MELD era and the future of organ allocation and distribution. Am. J. Transplant. **5**(2), 203–204 (2005)
5. Ellison, M.D., Edwards, L.B., Edwards, E.B., et al.: Geographic differences in access to transplantation in the United States. Transplantation **76**(9), 1389–1394 (2003)
6. Institute of Medicine Committee on Organ Procurement and Transplantation Policy: Organ Procurement and Transplantation: Assessing Current Policies and the Potential Impact of the DHHS Final Rule. National Academic Press, Washington, DC (1999)
7. Morris, P., Monaco, A.P., et al.: Geographic disparities in access to organ transplantation in France, United States, United Kingdom, Spain and Australia, *Transplantation*. Forum **76**, 1383–1406 (2003)
8. Roberts, J.P., Dykstra, D.M., Goodrich, N.P., Rush, S.H., Merion, R.M., Port, F.K.: Geographic differences in event rates by model for end-stage liver disease score. Am. J. Transplant. **6**, 2470–2475 (2006)

9. Tonelli, M., Kalbfleisch, J.D., Manns, B., Culleton, B., et al.: Residence location and likelihood of kidney transplantation. CMAJ **175**(5), 478–482 (2006)
10. Yeh, H., Smoot, E., Schoenfeld, D.A., Markmann, J.F.: Geographic inequity in access to livers for transplantation. Transplantation **91**(4), 479–486 (2011)
11. Bruni, M.E., Conforti, D., et al.: A new organ transplantation location-allocation policy: a case study of Italy. Health Care Manag. Sci. **9**, 125–142 (2006)
12. Miller, C.E., Tucker, A.W., Zemlin, R.A.: Integer programming formulation of traveling salesman problems. J. ACM **7**(4), 326–329 (1960)
13. Davies, R., Rodrick, P.: Planning resources for renal services throughout UK using simulation. Eur. J. Oper. Res. **105**(2), 285–295 (1998)
14. Kong, N., Schaefer, A.J., Hunsaker, B., Roberts, M.S.: Maximizing the efficiency of the US liver allocation system through region design. Manag. Sci. **56**(12), 2111–2122 (2010)
15. Levine, G.N., McCullough, K.P., Rodgers, A.M., Dickinson, D.M., Ashby, V.B., Schaubel, D.E.: Analytical methods and database design: implications for transplant researchers. Am. J. Transplant. **6**(2), 1228–1242 (2006)
16. Demirci, M.C., Schaefer, A.J., Romeijn, H.E., Robert, M.S.: An exact method for balancing efficiency and equity in the liver allocation hierarchy. INFORMS J. Comput. Spring **24**(2), 260–275 (2012)
17. Stahl, J.E., Kong, N., Shechter, S.M., Schaefer, A.J., Roberts, M.S.: A methodological framework for optimally reorganizing liver transplant regions. Med. Decis. Making **25**(1), 35–46 (2005)
18. Shechter, S.M., Bryce, C.L., Alagoz, O., Kreke, J.E., Stahl, J.E., Schaefer, A.J., et al.: A clinically based discrete-event simulation of end-stage liver disease and the organ allocation process. Med. Decis. Making **25**(2), 199–209 (2005)
19. Thompson, D., Waisansen, L.: Simulating the allocation of organs and transplantation. Health Care Manag. Sci. **7**, 331–338 (2004)

Chapter 3
A Routing Problem for Medical Test Sample Collection in Home Health Care Services

Y. Kergosien, A. Ruiz, and P. Soriano

Abstract Health care organizations are increasingly turning towards home care solutions to provide services to the population under their jurisdiction. Among these services, medical test sample collection (in particular blood and urine) is a highly demanded service by medical doctors. Providing this type of service clearly requires efficient coordination and planning of appointments for patients and healthcare personnel. This management involves the solution of a type of vehicle routing problem that is very complex given the large number and particular nature of the constraints that need to be considered: personnel schedules, patients preferences, maximum transport delay for some blood samples, etc. In this paper, we propose an integer linear programming formulation and different metaheuristics to solve this special vehicle routing problem. Experimental results on randomized instances were performed in order to select the best method to be integrated into a decision support tool and test it in on real data. This study is the result of collaboration with the blood collection service of the Center for Health and Social Service of Laval in the province of Quebec (CSSS-Laval).

Y. Kergosien (✉)
Université François Rabelais Tours, Laboratoire d'Informatique (EA 2101), Equipe Ordonnancement et Conduite (ERL CNRS 6305), 64 avenue Jean Portalis, 37200 Tours, France
e-mail: yannick.kergosien@univ-tours.fr

A. Ruiz
Département opérations et systèmes de décision, Faculté des sciences de l'administration, Université Laval, Québec (Québec) G1V 0A6, Canada
e-mail: angel.ruiz@fsa.ulaval.ca

P. Soriano
Service de l'enseignement des méthodes quantitatives de gestion, HEC Montréal, 3000 chemin de la Côte Sainte-Catherine, Montréal (Québec) H3T 2A7, Canada
e-mail: patrick.soriano@cirrelt.net

A. Matta et al. (eds.), *Proceedings of the International Conference on Health Care Systems Engineering*, Springer Proceedings in Mathematics & Statistics 61,
DOI 10.1007/978-3-319-01848-5_3,

3.1 Introduction

In recent years Health Care organizations everywhere around the world, but more particularly in developed countries, have increasingly resorted to Home Health Care (HHC) to provide services for the populations under their care. There are several reasons motivating this trend: economic factors, hospital congestion, patient preferences, ageing population, etc. The organizations that provide such Home Health Care services are however faced with difficult management problems resulting from this change in practices. In this paper, we focus on one such problem that arises in the context of a medical test sample collection service at the home of the patients that is provided by the CSSS[1]-Laval (Quebec). This service collects samples from more than 700 patients per week, most of them being blood samples with occasionally other types of samples such as urine, etc. The service involves two types of personnel: clerks and nurses. Clerks centralize all sample collection demands, prepare the planning of the nurses after a scheduling step, and perform other administrative tasks, i.e. printing the labels required for the sample tubes. Then the nurses plan their own routes to collect the test samples at the patients homes. Two types of demands have to be distinguished. The first type deals with sample collection demands for which the specific date when they should be carried out is given by the physician (roughly 70 % of demands are of this type). The second type of demand are those for which there is no specific date given for the collection but rather a time window in which it should be performed (e.g. several days or even weeks). Unfortunately, given the present management process the workload is too large for the service capacity available at the CSSS-Laval, thus the hospital has to hire additional nurses form private agencies on a day to day basis to meet the demand. These extra resources are however quite expensive.

The current organization for managing demands is based on a two phase daily planning process. First the clerks assign the demands that are due to be performed on the next day to the different nurses on the basis of a predetermined geographical decomposition of the area covered by the CSSS-Laval into sectors, each sector being assigned to a given nurse. The nurses then decide which demands of the second type should be added to their task and performed the next day (among those for which their time window overlaps the next day's date). Once these additional demands have been selected, the nurses then manually build their schedule and route for the next day. Two main drawbacks of this planning process easily spring to mind. First, using a predetermined geographical decomposition is far from optimal since the workloads associated with each sector may vary considerably one from another for any given day. Secondly, the scheduling task that the nurses have to perform to generate their routes and schedules is quite difficult as will be detailed later on and particularly so for nurses that are not trained to solve such problems (a large amount of information has to be taken into account, both from geographic as temporal point

[1]Center for Health and Social Services

of view). In addition, after discussions with the nurses and their supervisors, it was estimated that their planning tasks required at least 1 h daily, thus significantly reducing their capacity to collect samples.

The main goal of this study is therefore to develop a methodology capable of solving the routing/scheduling problem faced by the sample collection service of CSSS-Laval efficiently and to propose a planning tool implementing such a methodology in order to automatically compute the route for each nurse. Such a tool would save time for each nurse enabling them to treat more demands and could also improve the quality of the routes, therefore reducing or even eliminating subcontracting costs. The paper is organized as follows. The next section presents a brief literature review. Section 3.3 details the specific problem considered here while Sect. 3.4 describes the proposed solution approaches. Some computational experiments are presented in Sect. 3.5 as well as an overview of the decision tool developed. Finally, concluding remarks and future research avenues are provided in the last section.

3.2 Literature Review

Several studies have focused on decision support tools for home care at strategic, tactical or operational levels. A recent literature review outlining these issues can be found in [3] and focuses on solving assignment, scheduling and routing problems. These types of problems have led to several national or international studies. In [4], the authors study the scheduling and routing of home health care nurses in Alabama and develop a spatial decision support system. They build a heuristic to construct the routes of each nurse, taking unavailability into account. Other problems dealing with staff scheduling/routing in the context of home care are also discussed in [10] and [1]. The home care routing problem can also be generalized as a municipal or communal routing problems [6] taking into account scheduling home care, transportation of the elderly, and home meal delivery. In [7], the problem of routing home health care personnel is studied by considering two types of nurses, part time and full time, with different hourly costs. The authors define a mixed integer programming model taking into account lunch breaks and present a basic heuristic with the objective of minimizing the total cost. A similar problem is tackled in [5] with nurses having different skills. The objective is not simply to minimize the total cost but a weighted sum of the total travel time and several penalties, such as the violation of time windows or of patients preferences. The heuristic developed by the authors for solving this problem is divided into two parts: build a set of patients to be served by each nurse and then find an optimal sequencing for each set of patients. A very similar problem is studied in [8] in a Swedish context, the objective is to minimize the travel time and the waiting time of patients. The authors solve this problem using a set partitioning model with two types of variables (some for assigning a staff member to a schedule, others for assigning a staff member to a visit with a vehicle). A matching approach is used iteratively for finding a solution. They also describe the

development of a decision support system called Laps Care to eliminate the manual planning of home care unit assignments. In [9], the same authors present and discuss some results and experiences from two local government organizations and from the use of Laps Care. They conclude to an improvement of operational efficiency and of the quality of home care services provided to elderly citizens. Other studies [13,18] have investigated the integration of the periodicity of service delivery within this type of problems, i.e. some types of care services for the same patient are required several times over a given period of time. In these cases, the different sessions of care should all be performed by the same professional, if possible, or by the minimum number of different persons in order to ensure better patient monitoring and quality of care. Home health care in times of natural disaster has been studied in [19] and [17] in cooperation with the Austrian Red Cross. The authors propose a solution approach based on variable neighborhood search for the daily scheduling of home health care services. To validate their approach they used real life disaster scenarios. In [16], the authors study a home care rostering and routing problem in which one considers that nurses can use different transportation modes (public transport or car). This multimodal home care scheduling problem is solved using constraint programming to generate an initial solution and several metaheuristics to improve upon it. An exact approach to solve long-term home care scheduling problem using a branch-and-price algorithm is proposed in [11]. The pricing sub-problem consists in generating a 1-day plan for a nurse and the master problem selects the plans to construct a global schedule while considering regularity constraint. The method is able to solve instances up to 44 visits during 1 week. Finally, some recent studies have considered the synchronization of home care personnel visits to patients: [2,14,15] or [12]. These aspects bring a new level of complexity since the routes of each care giver cannot be evaluated independently anymore.

Although the number of studies on home care routing problems continues to increase, there is no published work to our knowledge that can be used directly to solve the problem faced by the sample collection services studied in this paper. This mainly is due to the presence of several characteristics particular to the problem at hand, namely: the existence of a maximum elapsed time for transportation of some types of samples, the fact that nurses routes can include several stops at drop off points where they will leave the samples they have collected up to then (in order for those samples to be quickly transferred by courier service to the laboratory for analysis), and finally, that demands can be subcontracted at a cost to external resources. This study being carried out with the aim of developing tools to solve a particular and difficult type of vehicle routing problem our focus will concentrate on solution methodologies.

3.3 Problem Description

Let N be the set of sample demands and let N_1 be the subset of fixed date demands and N_2 the one with non-fixed dates. Each demand i is characterized by the location of the patient's home, a time window $[e_i, l_i]$ during which the nurse can visit the

patient to collect the samples, an estimate of the processing time p_i required for collecting the samples, a possible requirement for a specific nurse to make the visit to that patient, a subcontracting cost C_i if the demand is performed by external resources, and the list of required tests. For some types of tests included in these lists there is an additional requirement that the time elapsed between the collection of the sample and the moment when the test is performed at the laboratory should not exceed a pre-specified value $DMax_i$ otherwise the test result is useless. These more critical tests essentially concern blood samples and account for about 20 % of the total number of demands. In order to satisfy these requirements, several sample drops off points are scattered all over the city for the nurses to deposit their critical samples. Then a specialized medical courier service collects the deposited samples at each drop off point according to a predetermined schedule and delivers them to the hospital laboratories for analysis. Therefore, the time elapsed before a critical sample is analyzed can easily be computed using the time of collection, the time of drops off and the transportation timetable of that specific drop off point. The nurses must therefore include as many stops at these drop off points as required along their routes so as to make it possible for the critical blood tests to be carried out within the required time. Depending on the number of critical samples in their schedule a nurse's route may need to include several stops at these drops off points (note that in the rest of the paper we will refer indifferently to these as laboratories or drop offs). After collecting a critical sample, a nurse may visit other patients before depositing her critical samples at a drop off point in order to optimize her route. However, the nurses have to visit a drop off point with an appropriate transportation timetable to satisfy the pre-specified maximum elapsed time. The time needed at drop off point i to drop all the samples is noted by td_i. A set M of nurses is available to serve all demands. Nurses, denoted by index k with $k \in M$, are characterized by a work schedule (start time Sta_k and end time End_k), a route starting point H_k (generally the nurse's home) and a route end point B_k (generally the offices of the medical test sample collection service).

3.4 Solving the Routing Problem

The main idea of the proposed solution approach is to decompose the problem into daily routing problem. We will therefore determine the routes of the nurses for only one day at a time (computed the previous day). Obviously, all demands having a fixed date that are due for the next day need to be included in the routes or subcontracted. Then the maximum number of demands of the second type will be included to the routes in order to prevent them from being postponed to another date as much as possible. We considered planning for a longer horizon such as a week or even a month, however two factors made this unattractive. First, a longer planning period complicates significantly the problem to be solved since the number of demands grows quite fast. But even more problematic, is the fact that demands are not known a long time in advance. Indeed, most of them are known around 2 days ahead of time and less which makes planning several days in advance inefficient.

A solution is represented by the sequence of demands performed by each nurse as well as the visits at the drop off points. By solving successively this routing problem day after day, a pseudo load balancing will be achieved. This problem resembles other vehicle routing problems and transportation problems found in the literature and in particular the "*Multi-TSP with time windows and max profits*" but possesses crucial differences, specifically the critical sample aspect. Therefore, to the best of our knowledge, none of the solution approaches that have been proposed in the literature for that problem can be easily adapted to the problem under study here. The home care sample collection routing problem can be modeled as a mixed integer linear program, a formulation of which is presented in the Appendix A.

Since in the real problem setting there are in fact several non-comparable objectives, we optimize a lexicographic objective function based on the following criteria:

- Minimize subcontracting cost.
- Maximize the number of the demand of the second type weighted by the number of remaining days ld_i.
- Minimize the delays with respect to the patients time windows and the nurses work schedules (note that delays are bounded).
- Minimize the sum of total distance traveled.

The experiments we carried out showed that the exact solution of the proposed model by a commercial solver such as Cplex could only be achieved for very small instances. Since computational times rise very fast with the size of the problem and since real life instances in our practical setting are rather large, we decided to concentrate on the development of solution approaches based on meta-heuristics which have proven their efficiency for solving somewhat similar types of problems. We therefore developed a tabu search and a variable neighborhood search. Both meta-heuristics being based on a same set of neighborhood operators, we will first present their common elements and then each general algorithm.

3.4.1 Neighborhood Operators

The algorithms are based on two types of neighborhood operators. The first one is an insertion operator. This operator has three types of possible moves: a demand belonging to a route may be inserted into another route or in another place if moved within the same route, a demand belonging to the set of subcontracted demands may be inserted into a route, and a demand belonging to a route may be removed and added to the set of subcontracted demands. For each type of move, all positions in a route are tested.

The second type of neighborhood operator is a swapping operator. This operator has two types of possible moves: a demand belonging to a route may be exchanged with another one in the same or a different route, and a demand belonging to the set of subcontracted demands may be exchanged with a demand belonging to a route. For the latter case, we consider that it is possible to perform a null exchange in

which no demand is removed from the routes. This was permitted so as to prevent the set of subcontracted demands to always have the same size.

Whatever the neighborhood operator used, only feasible solutions are considered in the neighborhood. Thus, for each move (insertion, deletion or exchange), the resulting solution is tested to check if all constraints are satisfied (time window at patient's home for each demand, maximum transportation delay, etc.). When a demand with critical tests has to be inserted or removed, then the required stop at a drop off point associated with this demand is also inserted or removed (note that a swap is equivalent to performing both an insertion and deletion step). In the case of insertions, the demand is first inserted and then, if the insertion of a drop off point is required, the best location in the route is selected. This best location, according to the objective function, must keep the solution feasible and therefore takes into account the transportation timetables at the drop off points.

After performing these types of moves, the other drop off points of the changed routes may no longer be adequate. Thus, after each change of the current solution, we apply a post-optimization on each route that has changed. This post-optimization procedure consists of two steps. The first step tests if the stop at each one of the drop off points in the route is still necessary, otherwise the drop off point is removed from the route. This step eliminates useless drop off points, e.g. it may be more efficient when two demands with critical tests are close enough in the route to drop off their samples at the same drop off point. The second step consists in changing each drop off point in the route in order to improve the objective function. The main idea is to test the drop off points one by one and try to find one closer than the one presently included while maintaining the solution feasible.

Due to the maximum elapsed time for critical tests and the transportation timetable of each drop off point, the optimal time to perform a sample collection may not necessarily be as soon as the nurse arrives at a patient's location. Indeed, it may be more efficient to wait some time before collecting a sample from a patient in order to achieve a better synchronization with the courier transportation schedule of a drop off point that is closer to the planned route. However, given that a route may have several drop off points it may be quite difficult to evaluate the optimal starting times efficiently. In addition, it is difficult in practice to justify to nurses that they should sometimes stay idle for a given time duration when arriving in front of a patient house before collecting the samples there in order to improve the overall efficiency of their routes. Thus we decided to neglect these possible economies and instead solved the problem with the assumption that each patient is collect as soon as possible (no idle waits).

3.4.2 Algorithms

3.4.2.1 Initial Solution

The initial solution, common to all methods, is built by inserting demands in the routes one by one, according to their priorities, at their best possible position given

the demands already present in the route. Priority is given to the first type of demands – i.e. fixed date – (ordered by decreasing size of the time window available to visit the patient), then to the second type of demand (ordered by the number of remaining days until the end of the time window specified to perform the collection). If a demand cannot be inserted, then it is added to the set of subcontracted/postponed demands. When a demand with critical tests is inserted, a drop off point is also inserted if necessary and the post-optimization procedure is applied (as described above).

3.4.2.2 Tabu Search

The tabu search developed to solve this problem has a classical structure: from an initial solution, the algorithm moves from a current solution to another solution by searching neighborhoods while avoiding those that are currently tabu until some stopping criterion is met. The previous two types of neighborhood operators were tested resulting in two TS variants: TS1 using the Insert neighbourhood and TS2 the Swap neighbourhood. Both variants use as tabu list the list of demands that were recently moved. The list length is fixed which implies that a demand cannot move again during a given number of iterations corresponding to the size of the tabu list. The stopping criterion is a maximum number of iterations without improvement to the best solution found. We also included the classical aspiration criterion that allows moving a demand which is tabu if this move strictly improves upon the best solution found.

3.4.2.3 Variable Neighborhood Search

The general structure of the variable neighborhood search (VNS) used to solve this problem is described below. The algorithm uses all neighborhoods previously defined here above except "Insert a demand belonging to a route into the set of subcontracted demands", since this operator is not necessary.

- $S_{best} \leftarrow$ Generate Initial Solution (same initial algorithm as previously)
- $k \leftarrow 1$
- While a pre-specified number of cycles without improvement is reached, repeat:
  - $S' \leftarrow$ Shake using the kth operators and starting at S_{best}.
  - $S'' \leftarrow$ VND[2] by starting at S'.
  - If $f(S'') < f(S_{best})$ then

[2]The general structure of the variable neighborhood decent (VND) used to solve this problem is classic. At the end of the algorithm, the best solution found cannot be improved with respect to any of the four tested neighborhood operators.

- $S_{best} \leftarrow S''$
- $k \leftarrow 1$

– Else

- $k \leftarrow k+1$

3.5 Computational Experiments and Overview of the Tool

In order to test the proposed algorithms, we generated three groups of instances. The first group is a set of small and rather simple instances composed of 10, 15, 20, and 25 demands, with respectively 6, 9, 12, and 16 fixed day demands. The number of nurses is either 1 or 2, and the number of drop off points ranges from 0 to 2. In the case where the number of drop off points is equal to 0, none of the demands involves critical samples, otherwise the number of demands with critical samples is set to 20 % of the total number of demands. These small instances were generated to enable us to obtain optimal solutions by solving the MILP formulation with the Cplex and then compare them with the ones obtained by the heuristic algorithms. The instances of the second group are instances inspired by the case study with CSSS-Laval and having similar size to the real instances. The number of demands is equal to 150, 175, 200, 225 and 250 with respectively 90, 105, 120, 135 and 150 fixed day demands. For all instances, the number of demand with critical samples is also set to 20 % of the total number of demand, the number of drop off points is set to 5 and the number of nurses varies between 10 and 20. Finally, the last group consists of five real instances that represent a normal week. For these instances we obtained the historical data and decisions made by our partners at CSSS-Laval. This groups of instances is detailed later.

All of the tests were performed on an Intel(R) Core (TM) i7-3610QM CPU running at 2.30 GHz and with 8.00 Gb of RAM. All codes were programmed in C++. Preliminary experiments were carried out on a small subset of the instances in order to determine the strategic values for the heuristic algorithms parameters: the size of the tabu list for the TS1 and TS2 algorithms were set at 65 % of the total number of demands, and the number of iterations without improvement with respect to the best solution were set at 1,000. The number of iteration without improvement for the VNS was set at 500. These two last parameters were selected in order for the algorithms to be able to solve real sized instances in a few minutes so that they could be usable within a practical decision support tool.

Table 3.1 presents the results on the small instances. Each line represents a set of 10 instances. The first three columns indicate the size of the instance: the number of patients, the number of nurses and the number of drop off points/laboratories (recall that if this value is equal to 0 then it means there are no critical samples in that setting). The ILP column represents the percentage of instances for which Cplex found the optimal solution (solution times were limited to a maximum of 1 h).

Table 3.1 Results on the small instances

Set of instances			ILP	Better than ILP		
#Patient	#Nurse	#Laboratory	% optimal	TS1 (%)	TS2 (%)	VNS (%)
10	1	0	100	100	90	70
15	1	0	100	80	100	60
15	2	0	100	60	20	20
20	1	0	100	80	100	50
20	2	0	100	20	30	20
25	2	0	20	80	80	30
10	1	1	100	90	90	90
15	1	1	60	100	100	60
15	2	1	10	90	90	20
20	1	2	0	100	100	70
20	2	2	0	100	100	100
25	2	2	0	100	100	100

In the case where Cplex did not solve to optimality, we used the best solution found by Cplex after 1 h in order to compare the results. The last three columns indicate, for each algorithm, the proportion of instances for which the algorithm found the optimal solution or a better solution than the one returned by Cplex (when optimality was not proven).

These results show that when there are critical samples and therefore the constraint requiring to drop some samples off at a laboratory during a route is present, Cplex is much less successful. The tabu search algorithms seem to be more efficient than the VNS algorithms. They do not always find the optimal solution but are the ones that obtain the best solutions most of the time. It should also be noted that they are extremely fast with solution times below 2 s in general.

Table 3.2 presents the results on the second set of instances. Each line represents 10 instances. The number of patients, the number of nurses and the number of drop off points are indicated in the first columns. To give a better idea about which method is the most efficient, we measured the percentage of instances for which each method found the best solution among all tested methods. The solution times for each method are also given in the Table 3.2. Objective values of each criterion are reported in Appendix B.

One can observe first that TS1 is clearly the fastest of the three algorithms compared, however the quality of the solutions it produces is also clearly inferior. These results show that the insertion operator is less powerful than the swap operator, the solution space is not explored widely enough. Now when comparing the two best performing algorithms, TS2 and VNS, TS2 seems to have the upper hand since on the basis of these results it obtains slightly better results overall and also exhibits shorter running times. The methods could therefore be ranked as follows: TS2>VNS>>TS1. Since the TS2 algorithm seems to be the best compromise between solution quality and computing time, we used this algorithm to integrate within the decision support tool and to perform the tests on the real data.

Table 3.2 Results

Set of Instances			TS1		TS2		VNS	
#Patient	#Nurse	#Laboratory	% best	CPU (s)	% best	CPU (s)	% best	CPU (s)
150	10	5	0	0.9	40	33.9	60	52.4
150	15	5	30	2.5	60	40.8	10	62.6
150	20	5	80	2.9	10	51.2	10	65.7
175	10	5	0	1	40	47.8	60	64.5
175	15	5	0	1.6	70	61.3	30	84.5
175	20	5	60	4.3	10	60	30	91.5
200	10	5	0	1.9	80	60.5	20	73.5
200	15	5	10	2.2	50	71.9	60	104.2
200	20	5	30	6	40	96.2	30	119.9
225	10	5	0	2.8	80	77.2	20	86.1
225	15	5	0	2	20	67.5	80	124.1
225	20	5	20	5.3	60	109.9	20	158.4
250	10	5	0	4.1	100	108.4	0	100
250	15	5	0	2.1	40	101.5	60	146.3
250	20	5	0	5.1	40	126.6	60	187.6

Table 3.3 Results

	Monday	Tuesday	Wednesday	Thursday	Friday
Number of demands	165	151	169	144	122
Number of fixed day demands	128	99	114	101	66
Number of demands with critical tests	37	29	45	40	41
Number of nurses	7	7	6	7	7

The real data was collected from one "normal" work week in December 2011. A total of 751 demands were performed by 7 nurses who worked each the 5 days of the week except for one which was absent on the Wednesday. Some details about the real data are reported in Table 3.3.

The hours during which the nurses usually do their routes and collect the samples are from 7 am until 12 pm. We used the same hours without considering the time that could be saved by using the decision support tool. We therefore simulated the decisions proposed by the proposed algorithm and solution procedure on the data of this same week using the same conditions of work and then compared the results obtained in reality with the ones produced by the TS2 algorithm. Figure 3.1 represents, for each day of the week, the number of demands performed by the nurses and the number of demands that needed to be subcontracted (in red). The histogram on the left corresponds to what was done in reality while the right one illustrates the results obtained using the decision support tool.

The first important observation is that the number of subcontracted demands was equal to 112 for the real week compared to 31 when using the TS2 algorithm. The reduction in subcontracted demands represents 10.8 % of the total number of demands during the week. In the solution proposed by the TS2 algorithm all the

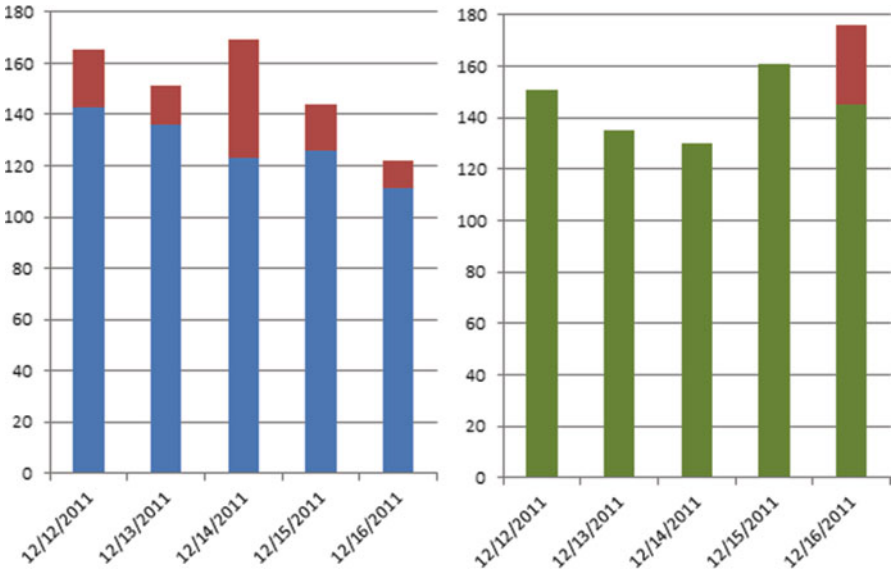


Fig. 3.1 Results on real instances

demands that need to be subcontracted happen on the Friday. This is due to the fact that we are missing some information regarding the second type of demands that were performed that week. Indeed, we did not have the information regarding the time window during which they needed to be performed. We hence considered that they had to be done between Monday and Friday of that week which explains why all such demands that could not be fitted in the routes earlier in the week were concentrated and subcontracted of the Friday. We want to point out that this decision of forcing the decision support tool to perform all second type demands during the week is almost surely a more stringent requirement than what reality imposed since some of those demands may have had end dates for their time windows later than that Friday and could therefore have been postponed to the following week (thus reducing the costs of subcontracting). Finally we also noticed that because of the geographical decomposition, the workload of the nurses varied quite significantly from day to day, resulting in a far less than optimal use of the sample collection service resources. The number of non-worked hours over all nurses is estimated at around 5 % of the total working time.

3.6 Conclusion

In collaboration with a home care medical test sample collection service in Quebec, we have studied a specific vehicle routing problem that arises in the course of their operations. We have developed two meta-heuristics to solve this problem

and integrated the best one in a decision support tool for planning the demands to be performed each day and the routes to be followed by each nurse to carry out the planned sample collections. In addition to a significant gain in time for the planning of nurses' routes, the tool also allows the organization to: increase the traceability of the samples it collects therefore improving the quality of the tests performed, improve the monitoring of the work carried out by the home care nurses, and also makes data entry easier, faster, and less prone to errors. The tool also enables the organization to save about 1 h per day per nurse of clerical work by automating the route production process, which is currently done manually by the nurses themselves. This saved time could therefore be used to increase the collection capacity of the nurses and further reduce the need for subcontracting. On a research perspective, one possibility that seems very appealing is to develop a hybrid algorithm combining the TS2 and VNS algorithms in order to improve further the performance of the decision support tool. A second avenue we are pursuing is to extend this approach to more general home health care settings (e.g. longer planning horizon, continuity of care, multiple care professionals, etc.).

Appendix A

We represent the problem using a complete graph $G(V,E)$. The set of vertices V is obtained as the union of three different sets: all test demands (N), the starting and ending points for each nurse route (D), and all combinations of a laboratory/drop off point with one of its specific pick up times according to its transportation schedule (C). Each laboratory is thus represented by several vertices in the graph, one vertex for each pick up time. The weight of an arc $(i,j) \in E$ represents the travel time d_{ij} for going from vertex i to vertex j. We also introduce parameter HV to represent a large real number and an indicator parameter LT_i for each demand i, that assumes value 1 if the demand i is a critical test, 0 otherwise.

Variables

- $\forall i,j \in V, \forall k \in M : x_{ij}^k \begin{cases} 1 \text{ if nurse k travels from node i to j,} \\ 0 \text{ otherwise,} \end{cases}$
- $\forall i \in N = N_1 \cup N_2 : t_i$: arrival time at patient's home.
- $\forall i \in C, \forall k \in M : t_i^k$: arrival time of nurse k at drop off point with pick up time equal to HR_j.
- $\forall i \in N/LT_i = 1, \forall j \in C, \forall k \in M : aff_{ij}^k \begin{cases} 1 \text{ if i is droped at laboratory/pick} \\ \quad \text{up time j by nurse k,} \\ 0 \text{ otherwise,} \end{cases}$
- $\forall k \in M : Ret_k$: delay of nurse k (overtime work).
- $\forall i \in N : Late_i$: lateness for demand i.

Constraints

$$\forall i \in N: \quad \sum_{k \in M} \sum_{j \in V/i \neq j} x_{ji}^k \leq 1 \tag{3.1}$$

$$\forall i \in V/D, \forall k \in M: \quad \sum_{j \in V/i \neq j} x_{ji}^k = \sum_{j \in V/i \neq j} x_{ij}^k \tag{3.2}$$

$$\forall k, k' \in M/k \neq k': \quad \sum_{i \in V/i \neq B_k} x_{iB_k}^k = 1 \quad ; \quad \sum_{i \in V/i \neq B_{k'}} x_{iB_{k'}}^k = 0 \tag{3.3}$$

$$\forall k, k' \in M/k \neq k': \quad \sum_{i \in V/i \neq H_k} x_{H_k i}^k = 1 \quad ; \quad \sum_{i \in V/i \neq H_{k'}} x_{H_{k'} i}^k = 0 \tag{3.4}$$

$$\forall k \in M: \quad \sum_{i \in C} x_{H_k i}^k = 0 \tag{3.5}$$

$$\forall i \in C, \forall k \in M: \quad \sum_{j \in C/i \neq j} x_{ij}^k = 0 \tag{3.6}$$

$$\forall i \in N, \forall k \in M \backslash I_i: \quad \sum_{j \in V/i \neq j} x_{ij}^k = 0 \tag{3.7}$$

$$\forall i \in N, \forall j \in N/i \neq j: \quad t_i + p_i + d_{ij} \leq t_j + HV. \left(1 - \sum_{k \in M} x_{ij}^k\right) \tag{3.8}$$

$$\forall i \in N, \forall j \in C, \forall k \in M: \quad t_i + p_i + d_{ij} \leq t_j^k + HV.(1 - x_{ij}^k) \tag{3.9}$$

$$\forall i \in C, \forall j \in N, \forall k \in M: \quad t_i^k + td_i + d_{ij} \leq t_j + HV.(1 - x_{ij}^k) \tag{3.10}$$

$$\forall i \in N, \forall k \in M: \quad x_{H_k i}^k.(Sta_k + d_{H_k i}) \leq t_i \tag{3.11}$$

$$\forall i \in N, \forall k \in M: \quad t_i + p_i + d_{iB_k} \leq End_k + Ret_k + HV.(1 - x_{ij}^k) \tag{3.12}$$

$$\forall i \in C, \forall k \in M: \quad t_i^k + td_i + d_{iB_k} \leq End_k + Ret_k + HV.(1 - x_{ij}^k) \tag{3.13}$$

$$\forall k \in M: \quad Ret_k \leq Max_{Ret_k} \tag{3.14}$$

$$\forall i \in N: \quad e_i. \sum_{k \in M} \sum_{j \in V/i \neq j} x_{ji}^k \leq t_i \leq l_i + Late_i \tag{3.15}$$

$$\forall i \in N: \quad t_i \leq (l_i + Max_{Late_i}). \sum_{k \in M} \sum_{j \in V/i \neq j} x_{ji}^k \tag{3.16}$$

$$\forall i \in N: \quad Late_i \leq Max_{Late_i} \tag{3.17}$$

$$\forall i \in N/LT_i = 1: \quad \sum_{k \in M} \sum_{j \in C} aff_{ij}^k = \sum_{k \in M} \sum_{j \in V/i \neq j} x_{ji}^k \tag{3.18}$$

$$\forall i \in N/LT_i = 1, \forall j \in C, \forall k \in M: \quad aff_{ij}^k \leq \sum_{l \in V/l \neq j} x_{lj}^k + \frac{t_j^k - t_i}{HV} \tag{3.19}$$

$$\forall i \in N/LT_i = 1, \forall j \in C, \forall k \in M: \quad aff_{ij}^k \leq \sum_{l \in V/l \neq i} x_{li}^k \tag{3.20}$$

$$\forall i \in N/LT_i = 1, \forall j \in C, \forall k \in M: \quad HR_j.aff_{ij}^k - t_i \leq DMax_i \tag{3.21}$$

$$\forall i \in N/LT_i = 1, \forall j \in C, \forall k \in M: \quad t_j^k \leq HR_j + HV(1 - aff_{ij}^k) \tag{3.22}$$

Constraints (3.1) indicate that each node can be visited at most once. Constraints (3.2) ensure the continuity of routes. The start and end points of each route are set through Constraints (3.3)–(3.5). Constraints (3.6) are not essential but can reduce the number of variables by prohibiting two consecutive drop off points. Constraints (3.7) specify the possible requirements/preferences for specific nurses for a patient. The traveling times between two points are represented by the Constraints (3.8)–(3.10). The work schedules are taken into account through Constraints (3.11)–(3.13) with a delay (overtime work) that is allowed but bounded by Constraints (3.14). Constraints (3.15) and (3.16) define the time windows of each demand. Lateness is also allowed but bounded by Constraints (3.17). Constraints (3.18)–(3.20) ensure that if a critical test is performed by a nurse, the sample has to be dropped off during the route of that nurse but after it has been collected. Constraints (3.21) and (3.22) imply that the time elapsed between the collection of a critical sample and the moment when it is tested at the laboratory does not exceed a pre-specified length of time $DMax_i$.

Objective Function

$$Min \left(\alpha \sum_{i \in N} Late_i + \beta \sum_{k \in M} Ret_k - \theta \left(\sum_{i \in N_2} \frac{1}{ld_i} \left(\sum_{k \in M} \sum_{j \in V/i \neq j} x_{ji}^k \right) \right) \right.$$
$$\left. + \Omega \left(\sum_{i \in N_1} C_i \left(1 - \sum_{k \in M} \sum_{j \in V/i \neq j} x_{ji}^k \right) \right) + \vartheta \sum_{k \in M} \sum_{i \in V} \sum_{j \in V/i \neq j} d_{ij}.x_{ij}^k \right)$$

The weights ($\Omega >> \theta >> \alpha \geq \beta > \vartheta$) were chosen in order to have a lexicographic optimization of the following criteria:

1. Minimize subcontracting cost.
2. Maximize the number of the demand of the second type weighted by the number of remaining days ld_i.
3. Minimize the delays with respect to the patients time windows and the nurses work schedules (note that delays are bounded).
4. Minimize the sum of total distance traveled.

Appendix B

Table 3.4 Results TS1

Set of instances			TS1			
#Patient	#Nurse	#Laboratory	Number of subcontracted demands	Number of postponed demands	Sum of delays	Total travel distance
150	10	5	9	47.2	279.7	1,632.4
150	15	5	3.2	6.4	207.4	2,149
150	20	5	3	2	19.4	1,786.3
175	10	5	19.1	58.1	308.7	1,604.7
175	15	5	3.1	26.1	337.8	2,238.8
175	20	5	3	2.6	80.6	2,203.4
200	10	5	27.6	70.2	300.8	1,560.1
200	15	5	6.6	42.9	381.6	2,272.7
200	20	5	4	5.4	215.2	2,620
225	10	5	35.9	79.9	328.4	1,479.7
225	15	5	11	64.6	433.8	2,307.6
225	20	5	4.1	15.1	348.3	2,691.9
250	10	5	46.6	91.1	332.4	1,477.8
250	15	5	18.8	80.5	452.9	2,275
250	20	5	5.1	37.6	437.3	2,761.6

Table 3.5 Results TS2

Set of instances			TS2			
#Patient	#Nurse	#Laboratory	Number of subcontracted demands	Number of postponed demands	Sum of delays	Total travel distance
150	10	5	6.8	42.1	278.9	1,529.7
150	15	5	3.1	6.8	216	2,107.4
150	20	5	3	2.2	21.6	1,933.5
175	10	5	14.6	54.7	318.1	1,492.2
175	15	5	2.3	22.4	315.1	2,152.9
175	20	5	3	2.9	86.5	2,316.5
200	10	5	21.6	68.1	319.3	1,440.5
200	15	5	4.9	37.4	386.6	2,161.4
200	20	5	3.8	6.3	192.2	2,576
225	10	5	31.4	81.7	342.5	1,453.1
225	15	5	8.7	58.5	410.5	2,166
225	20	5	3.6	14.7	334.2	2,631.7
250	10	5	40.5	91.9	379.9	1,388.2
250	15	5	14.1	73.8	479.1	2,143.2
250	20	5	4.9	33.6	412	2,666.3

Table 3.6 Results VNS

Set of instances			VNS			
#Patient	#Nurse	#Laboratory	Number of subcontracted demands	Number of postponed demands	Sum of delays	Total travel distance
150	10	5	6.4	44.8	239	1,530.6
150	15	5	3.2	9.6	142	2,070.1
150	20	5	3	2	24.4	1,871.5
175	10	5	14	59.6	258.1	1,507.6
175	15	5	2.4	25.6	263.9	2,148.3
175	20	5	3	3.5	84.8	2,263.9
200	10	5	24	75	272.2	1,535.2
200	15	5	4.5	41.8	324.3	2,152.4
200	20	5	3.8	9.1	112.7	2,422.1
225	10	5	35.9	86.6	309	1,546.4
225	15	5	7.9	63.6	343.6	2,189.5
225	20	5	3.7	16.9	243.4	2,578.8
250	10	5	51.5	96.5	271.3	1,554.5
250	15	5	12.9	79.9	387.5	2,155.1
250	20	5	4.6	34.8	334.1	2,629.6

References

1. Akjiratikarl, C., Yenradee, P., Drake, P.R.: PSO-based algorithm for home care worker scheduling in the UK. Comput. Ind. Eng. **53**, 559–583 (2007)
2. Bachouch, R.B., Guinet, A., Hajri-Gabouj, S.: A model for scheduling drug deliveries in a French homecare structure. In: International Conference on Industrial Engineering and Systems Management, Montréal, 13–15 May 2009
3. Bashir, B., Chabrol, M., Caux, C.: Literature review in home care. In: 9th International Conference of Modeling, Optimization and Simulation, Bordeaux, 06–08 June 2012
4. Begur, S.V., Miller, D.M., Weaver, J.R.: An integrated spatial decision support system for scheduling and routing home health care nurses. Interfaces **27**, 35–48 (1997)
5. Bertels, S., Fahle, T.: A hybrid setup for a hybrid scenario: combining heuristics for the home health care problem. Comput. Oper. Res. **33**, 2866–2890 (2006)
6. Bräysy, O., Dullaert, W., Pentti, N.: Municipal routing problems: a challenge for researchers and policy makers? In: Bijdragen Vervoerslogistieke Werkdagen 2007, pp. 330–347. Nautilus Academic Books, Zelzate (2007)
7. Cheng, E., Rich, J.L.: A home health care routing and scheduling problem. Technical Report TR98-04. Department of Computational And Applied Mathematics, Rice University, Houston (1998)
8. Eveborn, P., Flisberg, P., Rönnqvist, M.: LAPS CARE-an operational system for staff planning of home care. Eur. J. Oper. Res. **171**, 962–976 (2006)
9. Eveborn, P., Rönnqvist, M., Einarsdóttir, H., Eklund, M., Lidén, K., Almroth, M.: Operations research improves quality and efficiency in home care. Interfaces **39**(1), 18–34 (2009)
10. Fahle, T.: Production and transportation planning modeling report. Report, University of Paderborn (2001)
11. Gamst, M., Jensen, T.S.: A branch-and-price algorithm for the longterm home care scheduling problem. In: Klatte D, Lüthi HJ, Schmedders K (eds.) Operations Research Proceedings 2011. Springer, Berlin/Heidelberg (2012)

12. Kergosien, Y., Lenté, C., Billaut, J.C.: Home health care problem: an extended multiple traveling salesman problem. In 4th Multidisciplinary International Conference on Scheduling: Theory and Applications, Dublin, 10–12 Aug 2009
13. Nickel, S., Schröder, M., Steeg, J.: Mid-term and short-term planning support for home health care services. Eur. J. Oper. Res. **219**, 574–587 (2012)
14. Rasmussen, M.S., Justesen, T., Dohn, A., Larsen, J.: The home care crew scheduling problem: preference-based visit clustering and temporal dependencies. Eur. J. Oper. Res. **219**, 598–610 (2012)
15. Redjem, R., Kharraja, S., Xie, X., Marcon, E.: Routing and scheduling of caregivers in home health care with synchronized visits. In: 9th International Conference of Modeling, Optimization and Simulation, Bordeaux – France, 06–08 June (2012)
16. Rendl, A., Prandstetter, M., Hiermann, G., Puchinger, J., Raidl, G.R.: Hybrid heuristics for multimodal homecare scheduling. In: CPAIOR, Nantes, pp. 339–355 (2012)
17. Rest, K.D., Trautsamwieser, A., Hirsch, P.: Trends and risk in home health care. J. Humanit. Logist. Supply Chain Manag. **2**, 34–53 (2012)
18. Steeg, J., Schröder, M.: A hybrid approach to solve the periodic home health care problem. In: Operations Research Proceedings, pp. 297–302. Springer, Berlin/Heidelberg (2007)
19. Trautsamwieser, A., Gronalt, M., Hirsch, P.: Securing home health care in times of natural disasters. OR Spectr. **33**, 787–813 (2011)

Chapter 4
A Two-Stage Approach for Solving Assignment and Routing Problems in Home Health Care Services

Semih Yalçındağ, Andrea Matta, Evren Şahin, and J. George Shanthikumar

Abstract Human resource planning in Home Health Care (HHC) services is a critical activity that may also affect the quality of the delivered care. The assignment of the patient to operators together with their routing in the served territory are relevant problems that service providers have to deal with on a daily frequency. These problems can be either solved with a two-stage approach or with a simultaneous approach. The simultaneous approach enables to hold both assignment and routing decisions at the same time, however solving this problem is computationally difficult. The two-stage approach is the easier way of solving the assignment and routing problems, but an estimation of travel times is required to properly decompose the simultaneous approach into the two stages. This paper presents a new method to estimate operator travel times based on the Kernel Regression technique. Estimation is made on the basis of the operator travel times observed from previous periods. Numerical results based on realistic problem instances show that the proposed estimation method performs better than the classical Average Value method and that the whole approach is promising to construct realistic schedules.

S. Yalçındağ (✉)
Dipartimento di Meccanica, Laboratoire Génie Industriel, 20133 Milen, Italy

Politecnicodi Mileno, Ecole Centrale Paris, 92 295 Châtenay-Malabry Cedex, France
e-mail: semih.yalcindag@polimi.it; semih.yalcindag@ecp.fr

A. Matta
Dipartimento di Meccanica, Politecnico di Milano, 20133 Milan, Italy
e-mail: andrea.matta@polimi.it

E. Şahin
Laboratoire Génie Industriel, Ecole Centrale Paris, 92 295 Châtenay-Malabry Cedex, France
e-mail: evren.sahin@ecp.fr

J.G. Shanthikumar
Krannert School of Management, Purdue University, West Lafayette, IN 47907, USA
e-mail: shanthikumar@purdue.edu

A. Matta et al. (eds.), *Proceedings of the International Conference on Health Care Systems Engineering*, Springer Proceedings in Mathematics & Statistics 61,
DOI 10.1007/978-3-319-01848-5_4,

4.1 Introduction

Home Health Care (HHC) service is an alternative to the conventional hospitalization and consists of delivering medical, paramedical and social services to patients at their homes. The development of the HHC concept can be attributed to ageing of populations, social changes in families, increase in the number of people with chronical diseases, improvements in medical technologies, advent of new drugs and governmental pressures to contain health care costs [8]. The goal is to help patients to improve or keep their best clinical, social and psychological conditions.

Human resource planning in HHC services is a critical activity which the quality of the provided care depends on. From the admission of the patient, the service provider has to decide which operators will follow the patient during his stay as well as the detailed care delivery plan.

The resource assignment problem refers to the decision of which operators will take care of which patients. The operator routing problem specifies the sequence in which the patients are visited on a daily basis. To obtain the routes for operators, the assignment lists of operators and therefore the travel times between assigned patients should be known. The routing decision can either be held simultaneously with the assignment decision or it can be done just after the assignment procedure. In other words, the assignment and routing problems can be solved with two main approaches. The first one is solving them independently by a two-stage procedure where the output of the assignment problem is integrated as an input to the routing problem of each individual operator (Traveling Salesman Problem, TSP). The second approach aims at solving them simultaneously in a single model (Vehicle Routing Problem, VRP).

The literature available on the assignment and routing problems in HHC services has been enriched by recent works [6, 11]. Here we present some of the existing works. Akjiratikarl et al. [1] generate daily schedules by using the VRP with time windows. They focus on the determination of routes for each operator while minimizing the total distance traveled. Hertz and Lahrichi [5] propose two mixed integer programming models for assigning operators to patients. The objective is to balance the operators workloads. Trautsamwieser et al. [10] develop a model for the daily planning of the HHC services. The goal of the work is securing the HHC services in times of natural disasters. They develop the daily scheduling model as a VRP with state-dependent breaks. The objective of the model is minimizing the sum of travel times and waiting times, and also the dissatisfaction levels of the patients and health care operators. Lanzarone et al. [7] develop different assignment models to balance the operators' workloads considering several peculiarities of HHC services like the operators skills, the geographical areas of patients and operators, and the stochastic patient requests. Yalcindag et al. [12] propose a two–stage approach for assignment and routing decisions in HHC organizations. Their

main goal is to analyze the interaction between the assignment and routing processes where travel times between patients are estimated based on average values in the assignment phase.

Current literature mainly focuses on the simultaneous decisions of the assignment and routing problems. Although the simultaneous approach is theoretically the best alternative, it falls into the category of the NP-Hard problems. Due to this complexity, in the existing works either a heuristic solution method is adopted or very small instance sets are used to solve the developed models. Actually, heuristics are the only way to manage complexity in real applications where hundreds of patients receive the care service delivered by a single organisation.

The simultaneous approach is mainly based on the geographical locations of patients and aims to minimize the total traveling times of operators. However, in the HHC services there are other patient attributes that should be considered while trying to minimize travel times of operators such as patients' skill requirements, care profiles, special service requests etc. In these cases, the simultaneous approach may not be able to take into account such patient attributes or it could be computationally harder. Furthermore, each professional can construct the routing based on his (her) specific skills. These operator specific criteria can be hardly modeled in a mathematical programming model.

In order to cope with the problem complexity and special structure of the HHC services, this paper proposes a two-stage procedure for short-term planning of human resources. With this procedure, first the assignment problem needs to be solved and then, with the obtained patients lists and travel times between patients, the routes for each individual operator needs to be constructed. In this procedure, since the routing process is held independently and exact distances between patients are not available at assignment level, estimation of operator travel times is required to be able to solve the assignment problem. In the work of Yalcindag et al. [12], travel times are estimated based on average values. Although this is a intuitive approach, more accurate travel time estimation method is necessary to obtain results close to the ones the simultaneous approach provides. In particular, inaccurate estimations may create infeasibilities between the two stages (e.g., operator availability constraints in the routing problem) in addition to workload unbalancing and high travel times. Estimation is made on the basis of the operator travel times observed from previous periods in order to try capturing the specific operator behaviour. This paper partially addresses the stated problem by considering travel time minimization as the only criterion to construct the routes for the operators. The proposed estimator is assessed on a set of numerical cases.

The rest of the paper is organized as follows. The assignment and routing models are described in Sect. 4.2. In Sect. 4.3, the two-stage approach and travel time estimation methods are presented. In Sect. 4.4, the simultaneous approach is presented. Computational experiments are reported in Sect. 4.5. Finally, concluding remarks and future research directions are presented in Sect. 4.6.

4.2 Problem Definition

The assignment problem of the HHC services is used to determine which operators will provide the service to which patients, whereas the routing problem is used to decide the visiting sequence of patients for each operator. The problem can be defined on a complete directed network $G = (N,A)$, having n nodes where each node i corresponds to a patient.

In this work, we assume that the assignment and routing processes are held within a single category of operators (nurse or doctor) with same professional capabilities. In practice, operators are usually divided into several districts (as groups) based on their main skills and geographical areas to serve. A single district for a single planning period (e.g. day or week) is assumed.

Models are proposed under continuity of care where the newly admitted patient has to be assigned to only one principal operator in the set Ω of all operators. Each operator k, with $k \in \Omega = \{1,\ldots,K\}$, has one main skill that is used to handle a set of patients. The main skill refers to the patients for which the operator is best suited to care. For sake of simplicity, operators have no patient allocated from previous periods. Each operator k is assumed to have a deterministic capacity a_k, which is the maximum amount of time that the operator can accomplish according to his (her) working contract. In particular, it is also assumed that operators can handle excess load with respect to their capacities (i.e., overtime is allowed).

In the following sections, we present the details of the two-stage and simultaneous approaches to solve the assignment and routing problems.

4.3 Two-Stage Approach

In this section we provide details about the decomposed assignment and routing problems and also the travel time estimation methods.

4.3.1 Assignment Model

The considered assignment problem consists in matching operators with patients in a way that the utilization rates of operators (defined as the ratio between the actual workload of the operator and his (her) capacity) are balanced and the total traveling time of operators is minimized.

Each patient i (with $i = 1,\ldots,n$) has deterministic demand λ_i (expressed in amount of time) which denotes the total amount of the care volume needed by the patient. The demand of patient i is calculated as follows:

$$\lambda_i = (\tau_i + s_i) f_i \tag{4.1}$$

where s_i is the service time required by the patient, τ_i is the estimated travel time to reach the patient and f_i is the frequency of visits required by patient i.

The assignment problem is formulated as follows:

$$\min \quad h + \gamma \sum_{k=1}^{K} y_k \tag{4.2}$$

$$\text{s.t.} \sum_{k=1}^{K} x_{ik} = 1 \quad \forall i \tag{4.3}$$

$$y_k = \sum_{i=1}^{n} \tau_i x_{ik} \quad \forall k \tag{4.4}$$

$$w_k = \sum_{i=1}^{n} \lambda_i x_{ik} \quad \forall k \tag{4.5}$$

$$h \geq \frac{w_k}{a_k} \quad \forall k \tag{4.6}$$

$$x_{ik} \in \{0,1\} \quad \forall i,k \tag{4.7}$$

$$w_k \geq 0 \quad \forall k \tag{4.8}$$

$$y_k \geq 0 \quad \forall k \tag{4.9}$$

where variable x_{ik} takes the value 1 if the patient i is assigned to the operator k and 0 otherwise. The decision variable w_k is a continuous variable and is used to calculate the total workload of operator k. The decision variable y_k denotes the total travel time of operator k and γ is a parameter between 0 and 1. The auxiliary variable, h, is used to estimate the maximum utilization rate of the operators from above.

Equation (4.3) implies that all newly admitted patients must be assigned to only one operator. Equation (4.4) calculates the total travel time of each operator k. Equation (4.5) defines the total workload of each operator k. Inequality (4.6) expresses the maximum utilization rate h, which is minimized in the objective function (4.2) together with the penalized sum of travel times.

4.3.2 Travel Time Estimation Methods

Since in the two-stage approach the routing problem is solved after the assignment problem, at the time of the assignment decision the visiting sequences of patients are not known. In this section we provide details on how to build the travel time functions. We adopt a non parametric method to estimate travel times from real data observations. The reason is due to the distribution-free property of non parametric methods and the asymptotic convergence of some estimators. In particular, Kernel Regression (KR) is used to estimate the travel time functions. Remind that, in this

paper, we only consider the geographical locations of patients without taking into account their other attributes.

In the literature, only Average Values (AV) are used to estimate travel times. Thus, in the following section, in addition to the proposed method based on KR, we also describe the existing AV method.

4.3.2.1 Average Value Approach

The estimate of the travel time related to a patient is calculated as the weighted average travel time to reach his (her) home from all other patients, including also the common health care center. In such a case, the weights can be assumed to be proportional to the care volume required by each patients (frequency of required visits). Thus, the following estimator $\bar{\tau}_i$ is used:

$$\bar{\tau}_i = \frac{\sum_{j \neq i} w_j t_{ij}}{\sum_{j \neq i} w_j} \tag{4.10}$$

where t_{ij} denotes the traveling time from patient i to j, $(i, j) \in A$ and w_j is the weight related to the patient j.

Since average values are used to calculate the time to reach a patient, this can result in high travel times in comparison with the optimal travel times that are obtained with the simultaneous approach.

4.3.2.2 Kernel Regression Technique

KR is a non-parametric regression technique that does not require a predetermined (e.g. linear) form as the predictor is built with the information derived from the existing data [13]. KR exploits the correlations existing among the observations by assuming a radial basis function explaining the data. Since HHC patients have spacial relationship between each other (i.e., locations, skill requirements, etc.), KR can be adopted to estimate the travel time to visit a set of patients located in a geographical area.

KR technique estimates the expectation of the outcome variable Y (i.e., total travel time of operator) conditional on the random variable X (i.e., patient locations, care profiles), $E(Y|X)$. Than main reason for using KR is that it imposes few restrictions on the functional relationship between the covariates X and the outcome variable Y. This relationship can be shown with the following simple model:

$$Y = \tau(X) + \varepsilon \tag{4.11}$$

where τ is an unknown function and ε is the error term which is independent and identically distributed with $[0, \sigma^2(X)]$.

For our analysis we focus on the Multivariate Kernel Regression since our response variable Y depends on a vector of exogenous variables X. Thus, we try to estimate the following conditional expectation:

$$E(Y|X) = E(Y|x_1,\ldots,x_d) = \tau(X), \tag{4.12}$$

where $X = (x_1,\ldots,x_d)^T$ and d is the dimension of the covariate X.

To estimate the unknown function we use the Nadaraya-Watson estimator [13]:

$$\hat{\tau}(x) = \frac{\sum_{p=1}^{m} K\left(\frac{X_p - x}{h}\right) Y_p}{\sum_{p=1}^{m} K\left(\frac{X_p - x}{h}\right)}, \tag{4.13}$$

where $K(.)$ is a d dimensional kernel function and h is the bandwidth array. With this approach, the function τ is estimated with a locally weighted average by using the kernel as a weighting function. The selection of the bandwith value is relevant as it affects the smoothness of the predictor. Several methods are available in the literature to select an optimal value for h.

The kernel function, $K(.)$, is chosen as the widely applied Gaussian Kernel,

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-(1/2\theta)x^2}, \tag{4.14}$$

where θ represents the correlation coefficient.

In our context, $\hat{\tau}$ is indicating the estimation of the total travel time function of any operator k and in the remainder of this work it is denoted as $\hat{\tau}_k$. In particular, X_k is used to denote the attributes (in this study only geographical locations) of the patients assigned to the operator k. The outcome variable Y_k is used to express the total travel time of operator k to reach the assigned patients. X_k and Y_k values are used to estimate the total travel time function, $\hat{\tau}_k$.

To test the accuracy of the proposed KR technique, we first run an experiment to compare the predictor in Eq. (4.13) with the observed total travel times. In the experiment, we randomly generate five patients in a geographical area and the TSP model is used to calculate the optimal route to visit them accordingly to the travel time minimization criterion. This total travel time represents one (out of m) observation on which the predictor is constructed. To do this, we use different sizes of historical data (i.e, $m = 25, 35, 50, 100, \ldots$) to study the behavior of the predictor as the number of observations increases. At each generation, the five patients are randomly sampled with a triangular distribution between 0 and 100 and the mode equal to 40.

For each data set we calculate $\hat{\tau}$ on the basis of the m observations. Then the predictor is used to estimate the travel times for 100 new data sets randomly generated *out-of-sample*. For these new data sets the TSP model is used to obtain the optimal total travel times. These last are used as benchmark to study the accuracy of the estimator. The error between the estimated values and the optimal TSP values are

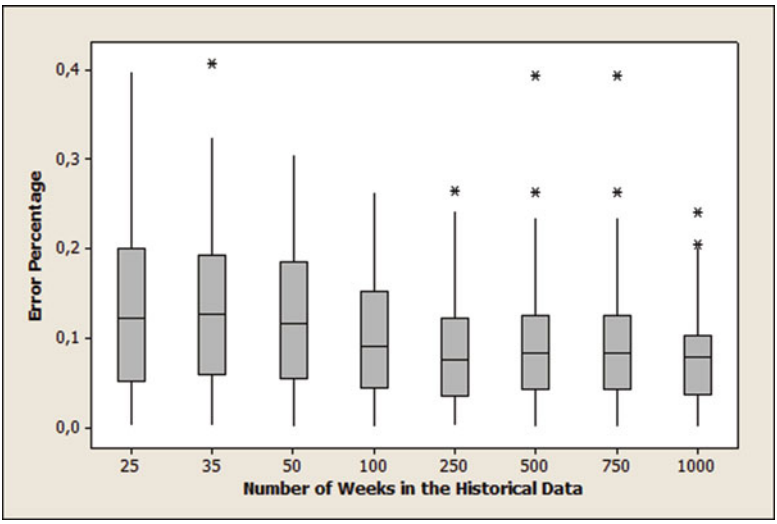


Fig. 4.1 Box-plot for the error between of the estimated and optimal travel times (100 samples)

shown on the box-plot in Fig. 4.1. As it can be seen, as the number of observations in the history increases, the predictor provides better estimates. Similar results were obtained by repeating the experiment with 6, 7 and 8 patients visited in the route. Obviously, the predictor performance deteriorates as the number of patients in the route increases.

The following section provides details on how the two travel time estimation alternatives can be considered in the assignment phase.

4.3.3 Use of Travel Time Estimators in the Assignment Problem

One of the most important point while solving the assignment problem is how to incorporate estimated travel times into the mathematical model. As far as the AV approach, it can be done simply by calculating the average travel times over all patients with Eq. (4.10) and plugging this value into Eq. (4.4) where the total travel time of each operator is calculated.

For the KR estimate, since the regression function is fitted to calculate directly the total travel time of an operator, it is more complex than the AV approach. To incorporate this into the current setting, two different approaches can be followed. A proper approach is to enumerate all possible assignment combinations for all operators and to estimate the related travel times using the KR functions. This can be done off-line, i.e., before the assignment problem is solved. Since the procedure may not be easy in practice, an heuristic approach can be applied as alternative to

solve the assignment problem. Indeed, it is very practical to embed the KR function in an heuristic approach such as genetic algorithm, tabu search, etc. In this way, first the heuristic selects (with some specific rules) assignments and then the KR function is executed to get the travel time estimate from which the objective function is calculated. This is repeated for several iterations until an exit condition is satisfied.

In this paper, a genetic algorithm is adopted to solve the assignment problem using KR for the estimation of travel times. The implemented heuristic solves the same assignment problem formulated in Eqs. (4.2)–(4.9).

4.3.4 Routing Model

At routing phase, a TSP model is used to create the routes for all operators in the considered planning period. With the patients lists obtained from the assignment phase, K independent TSP models are solved and the visiting sequences for all operators are determined. In other words, the output of the assignment phase is incorporated into the routing phase and the routes of all operators are obtained from the solution of the TSP models.

As the TSP model, we use the conventional formulation proposed by Dantzig et al. [4] with the objective of minimizing the total traveling time of each operator.

4.4 Simultaneous Approach

To be consistent with the modeled assignment problem, we need to formulate the VRP with the same objective function that balances the trade-off existing between workload balancing and total travel times. The problem has been formulated using the models proposed in [3, 9]. Two consecutive VRP models are solved to balance the total travel times of operators. In the first model, we find an upper bound on the maximum tour length. With the solution from this model, we solve a second VRP problem where the objective is minimizing the total travel times of all operators. As a result, the routes are constructed in a way that the total travel times of the operators are minimized according to the balancing purposes.

Balancing the total travel times of all operators can be considered as the balancing the total workloads of all operators when the service times required from each patient are assumed to be equal.

4.5 Computational Study

In this section we analyze and compare the proposed travel time estimation method with the AV approach. The travel time estimation alternatives are tested on three different instance groups, A, B, C. In the first instance group (A), locations of 15

Table 4.1 Results with instances from groups A and B (15 patients)

Group	Number	T(AV)	T(KR_2)	T(VRP)	$\%\triangle_{AV}$	$\%\triangle_{KR_2}$
A	1	639.14	604.14	601.73	6.2	0.4
	2	630.45	601.49	593.73	6.2	1.3
	3	662.00	619.06	614,58	7.7	0.7
	4	696.48	669.99	668.40	4.2	0.2
	5	666.86	637.02	609.88	9.3	4.5
B	1	882.73	715.62	703.97	25.4	1.7
	2	860.82	784.83	737.32	16.8	6.4
	3	847.01	769.86	718.74	17.6	7.1
	4	873.38	855.82	734.01	19.0	16.6
	5	876.83	792.34	739.71	18.6	7.1

Table 4.2 Results with instances from group C (56 patients)

Group	Number	T(AV)	T(KR_2)	T(KR_1)	$\%\triangle_{AV-KR_2}$
C	1[a]	139.04	105.58	128.69	31.7
	2	141.46	108.96	127.62	29.8
	3	139.47	94.49	129.65	47.6
	4	151.31	100.08	130.04	51.2
	5	147.90	105.14	127.77	40.7
	6	147.62	112.09	127.51	31.7
	7	135.81	109.18	127.55	24.4
	8	138.21	113.82	124.62	21.4

[a]The original real instance

patients are randomly sampled from a triangular distribution between 0 and 100 and the mode equal to 40. In the second instance group (B), patients are generated based on the grouping (clustering) structure. Here, 3 subsets of 15 patients (5 patient for each subset) are used. Each subset is located in a different geographical area and within each subset patients are located closely to each other. In the third group (C) we generate instances with 56 patients based on real data provided by an Italian HHC provider.

All of the presented results are obtained for a single planing period (e.g., day) for a single district. Small sized instances (A and B) are executed with three identical operators whereas the instances based on real setting (C) are solved with seven identical operators.

In all the experiments the historical data are randomly generated according to the instance type and the TSP is used to calculate the optimal travel times for building the KR predictor. In particular, bandwidth values, h, are used as the optimal values [2].

Small instances (A and B) are executed on both assignment and routing methods (two-stage and simultaneous (VRP) approach) whereas due to computational difficulties larger instances (C) are only solved for the two-stage approach. All of the results obtained with three instance groups are presented in Tables 4.1 and 4.2.

In the tables, the total travel time of all operators obtained in the two-stage approach with the AV and KR methods are shown with T(AV) and T(KR_2) notations, respectively. At the same way, the total travel time in the simultaneous approach is denoted with T(VRP). T(AV) and T(KR_2) values are obtained by solving several (as the number of operators) independent TSP models with the outputs obtained from the assignment stage and summing the results of each TSP models. Since we use a genetic algorithm for solving the assignment problem, T(KR_2) values are the average values resulting from five replications of the algorithm.

The percentage differences in Table 4.1 between the VRP approach and the two-stage approach with the two estimation methods are denoted with $\triangle_{AV}$ and $\triangle_{KR_2}$ calculated as follows:

$$\triangle_{\cdot} = \frac{|T(.) - T(VRP)|}{T(VRP)} \tag{4.15}$$

The results in Table 4.1 show that, if the patients are randomly scattered in the region (instance type A), the slight difference between AV and KR methods does not seem to justify the proposed approach. In particular, the two-stage approach has similar performance as the simultaneous approach. When the KR is used the differences with the VRP are quite small except for the instance 4-B.

In the other instance group (B) where patients are concentrated on specific locations, the two-stage approach with KR method is able to provide better solutions than the AV method. Thus, it seems from the reported numerical results that the KR technique performs better when the patients are located close to each other in some specific areas as it happens in the real case. To support this idea we test another instance set based on real data (C).

In the group C, one instance is directly generated from the real data with 56 patients distributed over 7 cities. By using the same patients, other seven instances are generated where, in each instance, the patients are randomly spread over the cities. The results of these instances are presented in Table 4.2.

The percentage differences on the total travel times between the two-stage approach with KR and AV methods are shown as $\triangle_{AV-KR_2}$ and calculated with Eq. (4.15) by replacing T(VRP) value with T(KR_2) value. As it can be seen from Table 4.2, the two-stage method with KR approach provides up to 51.2 % lower total travel times with respect to the AV approach.

The table also reports KR_1 representing the total travel times estimated by the KR method from the solution of the assignment problem. T(KR_1) values are used to test the accuracy of the proposed predictor with respect to the value obtained by the TSP approach, T(KR_2). Since we use the historical data with only 100 weeks to estimate the regression function, the observed differences between the KR_1 and KR_2 are slightly high. But, according to the considerations made in relation with Fig. 4.1, these differences can be reduced with the use of a larger number of historical data. Indeed, if one data corresponds to 1 day the KR can be successfully applied with hundreds of historical observations.

4.6 Conclusions

In this work, we propose a new travel time estimation method and we analyze the performance of this estimator with respect to the existing method. The results show how the proposed estimation method is used in a two-stage approach to decompose a complex problem.

We conclude that the proposed travel time estimator is performing good enough when patients are distributed in a special way (clustered). In particular, we also observe that even with a scattered distribution of patients our approach is providing lower total travel times in comparison to the existing approach.

We also compare the results of the two-stage approach with the simultaneous approach (VRP) and observe that, for the tested instances with few patients, the two-stage approach is able to provide very similar total travel times as the VRP approach provides.

The results reported in this paper suffer of a limited experimentation, thus they have to be confirmed on a larger design of experiments.

An on-going activity is to analyze the decomposition process of the assignment and routing problems in more details and try to compare the solutions with the VRP approach for larger instances according to the real framework. Another on-going activity is the improvement of the proposed travel time estimator to handle with more complex cases where more patient attributes are considered.

References

1. Akjiratikarl, C., Yenradee, P., Drake, P.R.: PSO-based algorithm for home care worker scheduling in the UK. Comput. Ind. Eng. **53**, 559–583 (2007)
2. Bowman, A.W., Azzalini, A.: Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations. Oxford University Press, Oxford (1997)
3. Cappanera, P., Gouveia, L., Scutella, M.G.: The skill vehicle routing problem, In: Pahl, J., Reiners, T., Voss, S. (eds.) Network Optimization: International Network Optimization Conference (INOC 2011), Hamburg. LNCS, vol. 6701, pp. 354–364 (2011)
4. Dantzig, G., Fulkerson, R., Johnson, S.: Solution of a large-scale travelling salesman problem. Oper. Res. **2**, 393–410 (1954)
5. Hertz, A., Lahrichi, N.: A patient assignment algorithm for home care services. J. Oper. Res. Soc. **60**, 481–495 (2009)
6. Hulshof, P.J.H., Kortbeek, N., Boucherie, R.J., Hans, E.W., Bakker, P.J.M.: Taxonomic classification of planning decisions in health care: a review of the state of the art in OR/MS. Health Syst. **1**, 129–175 (2012)
7. Lanzarone, E., Matta, A., Sahin, E.: Operations management applied to home care services: the problem of assigning human resources to patients. IEEE Trans. Syst. Man Cybern. A Syst. Hum. **42**, 1346–1363 (2012)
8. Matta, A., Chahed, S., Sahin, E., Dallery, Y.: Modelling home care organisations from an operations management perspective (2012). doi:10.1007/s10696-012-9157-0
9. Schwarze, S., Voss, S.: Improved load balancing and resource utilization for the skill vehicle routing problem. Optimization (2012). doi:10.1007/s11590-012-0524-2

10. Trautsamwieser, A., Gronalt, M., Hirsch, P.: Securing home health care in times of natural disasters. OR Spectr. **3**, 787–813 (2011)
11. Yalcindag, S., Matta, A., Sahin, E.: Human resource scheduling and routing problem in home health care context: a literature review. In: Proceedings of 37th Conference on Opisto Research Applied to Health Services, Cardiff, pp. 8–22 (2011)
12. Yalcindag, S., Matta, A., Sahin, E.: Operator assignment and routing problems in home health care services. In: Proceedings of CASE 2012, Seoul, pp. 325–330 (2012)
13. Wand, M.P., Jones, M.C.: Kernel Smoothing. Chapman and Hall, London (1995)

Chapter 5
Applying the Cardinality–Constrained Approach in Health Care Systems: The Home Care Example

Ettore Lanzarone and Giuliana Carello

Abstract Many approaches are applied to deal with uncertainty in health care optimization problems. However, a recently proposed technique, namely, the cardinality–constrained approach, is only marginally applied in health care. This approach accounts for a given degree of uncertainty with a reasonable computational effort, providing a trade-off between computational time and robustness. In this paper, we apply such approach to the nurse-to-patient assignment problem under continuity of care arising in home care services. A linear programming model is developed for solving the problem, and the robustness is included in the formulation according to the cardinality–constrained approach. The overall robust model is applied to a Home Care provider operating in Italy, in order to evaluate its capability of reducing the costs related to nurses' overtimes, and to compare the results both with the real practice of the analyzed provider and with previously developed approaches. Relevant benefits are achieved by applying the proposed model in the practice, and results suggest that such benefits could be also achieved in other optimization problems within the health care domain.

5.1 Introduction

Uncertainty is a key feature of many health care optimization problems, which cannot be neglected and may have a significant impact on the problem solution. In locating emergency vehicles, uncertainty is associated to the availability

E. Lanzarone (✉)
Istituto di Matematica Applicata e Tecnologie Informatiche (IMATI), Consiglio Nazionale delle Ricerche (CNR), Via Bassini 15, 20133 Milan, Italy
e-mail: ettore.lanzarone@cnr.it

G. Carello
Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, P.za Leonardo da Vinci 32, 20133 Milan, Italy
e-mail: giuliana.carello@polimi.it

A. Matta et al. (eds.), *Proceedings of the International Conference on Health Care Systems Engineering*, Springer Proceedings in Mathematics & Statistics 61,
DOI 10.1007/978-3-319-01848-5_5,

of ambulances [1], while in planning and scheduling operating room theaters uncertainty is mainly related to the duration of surgery [2]. Uncertainty also occurs in managing home care (HC) services, where sudden variations in the amount of service required by patients, which is in general highly variable, are the most critical and frequent random events.

Different approaches are usually applied to deal with uncertainty in optimization problems, such as probabilistic models, stochastic optimization approaches, and, more recently, the cardinality–constrained approach proposed in [3]. This accounts for a certain degree of uncertainty (which can be tuned) with a reasonable computational effort, providing a trade-off between needed computational time and robustness.

Although the approach seems to match many health care optimization problems, to the best of our knowledge it has not been often applied in this field so far (only four papers with keyword *health care* that cite [3] were found in March 2013 through a search on ISI web of knowledge and Scopus).

In this paper, we present an application of the cardinality–constrained approach to the nurse-to-patient assignment problem under continuity of care in HC. The approach can be easily applied to the problem and proved to produce good quality solutions with a reasonable computational effort. Therefore, it is also worthy of being tested on other health care optimization problems.

5.1.1 Home Care Service

HC consists of delivering medical, paramedical and social services to patients at their domicile rather than in hospital. This leads to a significant improvement in the quality of life for patients, as they continue to live at their home, and to considerable cost savings for the entire health care system, as hospitalization costs are avoided. Moreover, HC is a relevant and growing sector in western countries, due to the population aging, to the increasing of chronic pathologies, to the introduction of innovative technologies, and to the pressure of governments to contain health care costs.

Many resources are involved in delivering HC services, including nurses, other operators, support staff and material resources. In addition, the presence of peculiar constraints, such as the continuity of care and the operator risk of incurring burnout, makes the HC resource planning different from the planning problems arising in other health care systems.

Continuity of care means that a HC provider assigns only one nurse to each patient, called *reference nurse*, and the assignments are kept for a long period. This is an important quality indicator since patients are always cared for by the same nurse, instead of continuously developing new relationships, and potential loss of information among operators is avoided. However, continuity of care limits the flexibility of the service, and some providers do not adopt it to increase the

operational efficiency. In general, for a good balance between quality and flexibility, the continuity of care should be preserved at least for critical patients (e.g., palliative patients) or patients with particular needs.

5.1.2 Literature Review

The literature about HC management can be mainly divided into two groups: the first one deals with daily schedule of visits and routing of nurses, and the second one deals with staff planning and management from a mid-term and long-term perspective. The nurse-to-patient assignment is related to the mid-term management. Different features may be considered, such as the continuity of care and the uncertainty in patients' demands.

Nurse-to-patient assignment has been rarely studied as a stand alone problem (i.e., not considering the scheduling [4]) and, to the best of our knowledge, the assignment problem taking into account the continuity of care is only marginally addressed in the literature [5–7]. Besides, continuity of care is often considered as an objective rather than a strict requirement [8]. If continuity of care is not considered, the assignment problem turns out to be an assignment of operators to visits rather than to patients, in which the aim is to jointly optimize the operator-to-visit assignment and the scheduling and routing problem [9, 10]. In districts with a limited territorial extension (e.g., in Europe), the impact of travel times on scheduling and routing is not very significant; hence, assignment and scheduling problems are separately solved since the joint optimization requires a significant computational effort and, consequently, reduces the length of the considered time horizon.

As mentioned, uncertainty inherently arises in HC due to unpredictable changes in patients' needs. In [11] it is managed by representing the whole system as a Markov chain and developing admittance policies for patients.

The nurse-to-patients assignment problem, in which both continuity of care and demand uncertainty are considered, has been rarely addressed in the literature. The problem was tackled with stochastic programming [6] and with analytical policies [7]. However, both these approaches proved limited even if they improve the quality of the assignment with respect to those actually applied by the HC structures. The stochastic programming approach is based on scenario generation and, due to the high number of patients and the associated demand variability, requires to include a very high number of scenarios. Only a limited number of them can be consequently considered for a computationally acceptable solution. Therefore, a high expected value of perfect information (EVPI) and a low value of the stochastic solution (VSS) are obtained [6]. The analytical policies are related to strict assumptions regarding, e.g., the shape of workload probability density functions, the number of assignable patients, and the number of periods in the planning horizon [7].

With the cardinality–constrained approach, we aim at exploiting the potentialities of a linear programming model rather than an analytical approach, without the necessity of generating scenarios.

5.2 Robust Assignment Model

We consider the problem of assigning a set of patients P to a set of nurses I over a time horizon T divided into a set of time slots. Three continuity of care requirements are considered:

- *Hard continuity of care*: patients must be assigned to only one reference nurse for the entire time horizon. These patients are partitioned into two subsets P_c^a and P_c^n. Patients in P_c^a are already under treatment and assigned at the beginning of the time horizon, and they keep their assignment. Patients in P_c^n start their treatment at the beginning of the time horizon, and they are not yet assigned.
- *Partial continuity of care*: the reference nurse can be changed from time slot to time slot. However, each reassignment is penalized by a cost γ to keep the number of reassignments limited. As for the previous case, these patients are partitioned into two subsets P_{pc}^a and P_{pc}^n. Patients in P_{pc}^a are already under treatment at the beginning of the time horizon, while patients in P_{pc}^n start their treatment at the beginning of the time horizon.
- *No continuity of care*: patients can be assigned to more than one nurse even in the same time slot and the assignments can be changed from a time slot to another without penalties (set P_{nc}).

The division in districts is taken into account: a parameter m_{ij} is given for each nurse $i \in I$ and patient $j \in P$, which is equal to 1 if nurse i operates in the district of j, and 0 otherwise.

The amount of working time required by patient $j \in P$ in time slot $t \in T$ is an uncertain parameter r_{jt}, with expected value $\bar{r}_{jt}$ and maximum value $\bar{r}_{jt} + \hat{r}_{jt}$. Each nurse $i \in I$ has an amount of available working time per time slot v_i, and overtime must be paid if v_i is exceeded. The overtime cost depends on its amount. A set of overtime levels L_i are defined for each nurse $i \in I$, and two parameters are given for each level $l \in L_i$: a threshold Δ_i^l and a cost per time unit c_l for each overtime unit above $v_i + \sum_{k=1}^{l-1} \Delta_i^k$ and below $v_i + \sum_{k=1}^{l} \Delta_i^k$.

The problem consists of assigning all of the patients to the nurses, according to the required continuity of care, with the aim of minimizing the overtime costs and the number of reassignments for patients with partial continuity of care.

The problem is modeled as follows. A binary variable x_{ji} is defined for each patient $j \in P_c^a \cup P_c^n$ and nurse $i \in I$ ($x_{ji} = 1$ if j is assigned to i during the whole time horizon, and 0 otherwise). Similarly, a binary variable ξ_{ji}^t is defined for each patient $j \in P_{pc}^a \cup P_{pc}^n$, nurse $i \in I$ and time slot $t \in T$ ($\xi_{ji}^t = 1$ if nurse i is in charge of patient j during time slot t, and 0 otherwise). The assignments of patients to reference nurses

before the considered time horizon are described with parameters $\tilde{x}_{ji}$ ($\tilde{x}_{ji} = 1$ if $j \in P$ is initially assigned to $i \in I$, and 0 otherwise). Furthermore, a binary variable y_j^t is introduced for each patient $j \in P_{pc}^a \cup P_{pc}^n$ ($y_j^t = 1$ if the assignment of patient j is changed from time slot $t-1$ to time slot t, and 0 otherwise). Finally, the fraction of time needed by $j \in P_{nc}$ in time slot $t \in T$ provided by nurse $i \in I$ is represented by a continuous variable $\chi_{ji}^t \in [0,1]$. The overtime assigned to each nurse $i \in I$ in time slot $t \in T$ is described by a continuous variable w_{it}^l for each level $l \in L_i$, which represents the extra workload related to c_l.

The objective function aims at minimizing the overtime costs and the number of reassignments: these two parts are both relevant, as the first one reduces the burnout risk, while the second one guarantees a suitable quality of provided service.

$$\min \left\{ \sum_{i \in I} \sum_{t \in T} \sum_{l \in L_i} \left(c_l w_{it}^l \right) + \gamma \sum_{j \in P_{pc}^a \cup P_{pc}^n} \sum_{t \in T} y_j^t \right\} \tag{5.1}$$

subject to:

$$\sum_{i \in I} m_{ij} x_{ji} = 1, \qquad \forall j \in P_c^a \cup P_c^n \tag{5.2}$$

$$\sum_{i \in I} m_{ij} \xi_{ji}^t = 1, \qquad \forall j \in P_{pc}^a \cup P_{pc}^n, t \in T \tag{5.3}$$

$$\sum_{i \in I} m_{ij} \chi_{ji}^t = 1, \qquad \forall j \in P_{nc}, t \in T \tag{5.4}$$

$$\sum_{j \in P_c^a \cup P_c^n} r_{jt} x_{ji} + \sum_{j \in P_{pc}^a \cup P_{pc}^n} r_{jt} \xi_{ji}^t + \sum_{j \in P_{nc}} r_{jt} \chi_{ji}^t \leq v_i + \sum_{l \in L_i} w_{it}^l, \quad \forall i \in I, t \in T \tag{5.5}$$

$$0 \leq w_{it}^l \leq \Delta_i^l, \qquad \forall i \in I, t \in T, l \in L_i \tag{5.6}$$

$$x_{ji} \geq \tilde{x}_{ji}, \qquad \forall i \in I, j \in P_c^a \tag{5.7}$$

$$y_j^t \geq \xi_{ji}^t - \xi_{ji}^{t-1}, \qquad \forall t \in T \setminus \{t_1\}, j \in P_{pc}^a \cup P_{pc}^n, i \in I \tag{5.8}$$

$$y_j^{t_1} \geq \xi_{ji}^{t_1} - \tilde{x}_{ji}, \qquad \forall j \in P_{pc}^a, i \in I \tag{5.9}$$

Constraints (5.2)–(5.4) guarantee that each patient is assigned to a suitable nurse; constraints (5.5) compute nurse workloads and overtimes for each level; constraints (5.6) set the thresholds for the overtime workload; constraints (5.7) guarantee that patients in P_c^a do not change their assignment at the beginning of the time horizon; constraints (5.8) and (5.9) compute the number of reassignments.

To deal with uncertainty in constraints (5.5) we apply the cardinality–constrained robust model proposed in [3]. The basic idea of the approach is that only a subset of the uncertain parameters are likely to assume their maximum value simultaneously. The approach provides solutions which are feasible even if at most Γ uncertain parameters assume their worst possible value (i.e., the maximum value) rather than

their expected value. As the solution must be feasible for any choice of Γ parameters for each constraint, the subset which represents the worst possible case is selected. The impact is then computed exploiting duality properties, yielding to a linear formulation.

We apply the cardinality–constrained approach to the proposed formulation by including, for each nurse and time slot, three subsets S_c^{it}, S_{pc}^{it} and S_{nc}^{it} of patients assigned to i (with $S_c^{it} \subseteq P_c^a \cup P_c^n$, $S_{pc}^{it} \subseteq P_{pc}^a \cup P_{pc}^n$, and $S_{nc}^{it} \subseteq P_{nc}$), whose demand charged to nurse i in time slot t is equal to the maximum treatment time $\bar{r}_{jt} + \hat{r}_{jt}$. Cardinality is constrained as at most Γ_c^i, Γ_{pc}^i and Γ_{nc}^i patients (with Γ_c^i, Γ_{pc}^i and Γ_{nc}^i integer) are assumed to belong to these subsets, respectively. The charged demand of all other patients is the expected value $\bar{r}_{jt}$.

The robustness is taken into account considering the worst possible charge for each nurse i at each time slot t in constraints (5.5). As example, for patients requiring hard continuity of care, the term $\sum_{j \in P_c^a \cup P_c^n} r_{jt} x_{ji}$ is replaced with:

$$\sum_{j \in P_c^a \cup P_c^n} \bar{r}_{jt} x_{ji} + \max_{\substack{S_c^{it} | S_c^{it} \subseteq P_c^a \cup P_c^n, \\ |S_c^{it}| = \Gamma_c^i}} \left\{ \sum_{j \in S_c^{it}} \hat{r}_{jt} x_{ji} \right\}$$

Let us denote the maximum related to a given solution $\{x^*\}$ with $\beta_c^{it}(x^*, \Gamma_c^i, t)$:

$$\beta_c^{it}(x^*, \Gamma_c^i, t) = \max_{\substack{S_c^{it} | S_c^{it} \subseteq P_c^a \cup P_c^n, \\ |S_c^{it}| = \Gamma_c^i}} \left\{ \sum_{j \in S_c^{it}} \hat{r}_{jt} x_{ji}^* \right\}$$

This is computed for each nurse i and time slot t by solving the following linear programming problem:

$$(\mathscr{P}_c^{\beta it}) = \max \sum_{j \in P_c^a \cup P_c^n} \hat{r}_{jt} x_{ji}^* z_{ji}^t \tag{5.10}$$

$$\sum_{j \in P_c^a \cup P_c^n} z_{ji}^t \leq \Gamma_c^i \tag{5.11}$$

$$0 \leq z_{ji}^t \leq 1, \quad \forall j \in P_c^a \cup P_c^n \tag{5.12}$$

where $z_{ji}^t \in [0,1]$ are continuous variables which represent the choice of the elements in subset S_c^{it}. The associated dual problem is:

$$(\mathscr{D}_c^{\beta it}) = \min \sum_{j \in P_c^a \cup P_c^n} \pi_{jit}^c + \Gamma_c^i \zeta_{it}^c \tag{5.13}$$

$$\zeta_{it}^c + \pi_{jit}^c \geq \hat{r}_{jt} x_{ji}^*, \quad \forall j \in P_c^a \cup P_c^n \tag{5.14}$$

$$\pi_{jit}^c \geq 0, \quad \forall j \in P_c^a \cup P_c^n \tag{5.15}$$

$$\zeta_{it}^c \geq 0 \tag{5.16}$$

where ζ_{it}^{c} are the dual variables associated with (5.11), and π_{jit}^{c} the dual variables associated with $z_{ji}^{t} \leq 1$ (5.12).

Optimal values ($\mathscr{P}_c^{\beta it}$) and ($\mathscr{D}_c^{\beta it}$) coincide and, therefore, the maximum can be replaced by $\sum_{j \in P_c^a \cup P_c^n} \pi_{jit}^{c} + \Gamma_c^i \zeta_{it}^{c}$ adding the following variables and constraints to the model:

$$\zeta_{it}^{c} + \pi_{jit}^{c} \geq \hat{r}_{jt} x_{ji}, \quad \forall i \in I, j \in P_c^a \cup P_c^n, t \in T$$
$$\zeta_{it}^{c} \geq 0, \quad \forall i \in I, t \in T$$
$$\pi_{jit}^{c} \geq 0, \quad \forall i \in I, j \in P_c^a \cup P_c^n, t \in T$$

The same idea is applied to $\sum_{j \in P_{pc}^a \cup P_{pc}^n} r_{jt} \xi_{ji}^{t}$ and $\sum_{j \in P_{nc}} r_{jt} \chi_{ji}^{t}$, thus obtaining the robust cardinality–constrained version of the model.

In this way, each feasible solution remains feasible if any subset of at most Γ_c^i, Γ_{pc}^i and Γ_{nc}^i patients, respectively, require the highest number of visits.

5.3 Real Case Analysis

Computational tests are run in order to evaluate the applicability of the proposed approach to a real HC provider. The quality of the solutions and their impact when applied to realistic scenarios are taken into account.

The analysis is conducted on the same HC provider already studied in other papers dealing with assignment techniques under continuity of care [6, 7], so as to compare the outcomes of the proposed model with other approaches. Furthermore, a patient stochastic model to estimate the future patients' demands is available for this provider [12]. The considered HC provider operates in the north of Italy, covering a region of about 800 km^2, with about 1,000 patients assisted at the same time by about 50 nurses. The provider includes three independent divisions, and the analysis is carried out for the nurses of the largest one. The division consists of six districts and the analysis is carried out in four of them where more than one nurse is present (Table 5.1). The assignments are planned considering the districts as independent.

Table 5.1 Analyzed districts

Name of the district	Code of territory	Skill of the nurses	Number of nurses
NPA	A	Non-palliative	8
PA	A	Palliative	3
NPB	B	Non-palliative	4
NPC	C	Non-palliative	5

Table 5.2 Analyzed instances

Type of continuity for palliative	Γ values	Robust solution	Non-robust solution
C	1	Conf. A	Conf. E
	2	Conf. B	Conf. E
random 80% C	1	Conf. C	Conf. F
and 20 % PC	2	Conf. D	Conf. F

5.3.1 Experimental Setup

We consider data related to 26 weeks from April to September 2008 [6, 7]. The model is applied according to a rolling approach, and each time slot t is 1 week. An initial assignment of nurses is computed at the initial week (named week 0) considering all patients as newly admitted ones, while the successive assignments are provided on a rolling basis: at the beginning of each week, newly admitted patients are included in the mix and discharged patients are excluded. For each rolling week, the planning horizon includes the considered week and the next seven ones ($T = 8$). The assignments computed for the first horizon week are then kept, and the model is solved again for the next rolling week taking into account the information about patients assigned in the previous rolling weeks. This is consistent with the policy of the analyzed HC service provider, where assignments are mainly decided at the beginning of each week on a weekly basis. The initialization at week 0 is obtained neglecting the robustness (i.e., all patients require the expected demand $\bar{r}_{jt}$).

The reassignment penalty γ is assumed equal to 2.5, and 10 overtime levels are considered ($l = 1, \ldots, 10$), with $c_l = l \; \forall l$ and $\Delta_i^l = 0.1 v_i \; \forall i, l$.

The number of patients in charge at each week and their features are taken from the historical data of the provider (considering real arrivals of new patients and real discharges), while patients' demands are estimated with the stochastic model proposed in [12]. The expected demand $\bar{r}_{jt}$ and the maximum demand $\bar{r}_{jt} + \hat{r}_{jt}$ of each patient are taken from an empirical probability density function given by such stochastic model (maximum value $\bar{r}_{jt} + \hat{r}_{jt}$ is taken neglecting the right tail of the distribution with probability 0.1).

The continuity of care requirement for each patient is determined based on his/her characteristics. Patients belong to 15 different care profiles (CPs) [12] and the type of continuity of care required by each patient is once decided according to the CP when the patient is first considered. For non-palliative patients, low intensity CPs require no continuity of care, middle intensity CPs partial continuity of care, and high intensity CPs hard continuity of care. Two different configurations are taken into account for palliative patients: either they all require hard continuity of care, or they require hard or partial continuity according to a random choice: each palliative patient is randomly considered requiring hard continuity of care (with probability 0.8) or partial continuity (with probability 0.2).

Two levels of robustness are considered, either $\Gamma_c^i = \Gamma_{pc}^i = \Gamma_{nc}^i = 1, \forall i$ or $\Gamma_c^i = \Gamma_{pc}^i = \Gamma_{nc}^i = 2, \forall i$. Moreover, also the case in which the robustness is neglected is studied (Table 5.2).

5.4 Results

The model has been implemented with OPL 5.1 and solved with CPLEX; computational tests have been run on a PC equipped with CPU Intel Core i7 1.73 GHz and 6 GB of RAM. A stopping condition on the gap is set so as to limit the computational time (1 % for configurations A and C; 4 % for configurations B and D). No stopping condition is set for the non-robust configurations E and F.

Table 5.3 shows the computational time, the objective function and the number of reassignments for patients with partial continuity of care. Results are expressed in terms of minimum, maximum and average values among the weeks from 1 to 25; week 0 is excluded as it refers to the non-robust initialization.

Results show that, with the adopted gaps, computational times are reasonable for any configuration. The objective function increases with the values of Γ_c^i, Γ_{pc}^i and Γ_{nc}^i due to both the overtime costs and the number of reassignments, as the demand of the worst scenario increases and more robust solutions are selected. The overtime cost is significantly affected by the degree of robustness of the solution, as the maximum demands of patients belonging to S_c^{it}, S_{pc}^{it} and S_{nc}^{it} have an impact on the overall workload.

Then, the question arises on how a robust solution behaves if no patients require the maximum amount of care. For evaluating the behavior of the solutions with respect to the expected demands, the assignments are applied assuming that each patient is requiring the expected demand $\bar{r}_{jt}$. The obtained overtime costs are reported in Table 5.4 in terms of minimum, maximum and average values among the weeks from 1 to 25.

It can be seen that robustness determines an increase of overtime costs. However, it is worth noting that, when considering the expected demands, the robust assignment is not significantly penalized with respect to the optimal non-robust counterpart. Indeed, overtime expected costs are always lower than the double of the non-robust case.

Table 5.3 Computational time in seconds, objective function and number of reassignments

	Computation time			Objective function			Num. of reassignments		
Configuration	Min	Max	Average	Min	Max	Average	Min	Max	Average
A	5	115	31	59.3	157.1	98.4	0	4	0.8
B	4	7,339	505	177.7	484.5	299.6	0	8	1.2
C	5	987	83	101.6	265.6	168.2	0	4	0.8
D	4	7,580	644	181.7	579.9	349.0	0	6	1.2
E	1	3	2	3.9	27.8	13.6	0	1	0.1
F	1	2	2	3.9	27.8	13.4	0	1	0.0

Table 5.4 Overtime cost from the objective function and overtime cost recomputed with the expected demands

	Overtime cost			Overtime expected cost		
Configuration	Min	Max	Average	Min	Max	Average
A	59.3	149.6	96.4	7.9	35.6	16.5
B	175.2	482.0	296.6	6.3	43.4	20.5
C	99.1	265.6	166.2	7.5	42.1	20.3
D	179.2	564.9	346.0	8.3	41.6	20.9
E	3.9	27.8	13.4	3.9	27.8	13.4
F	3.9	27.8	13.3	3.9	27.8	13.3

Table 5.5 Executed mean overtime cost per nurse: minimum, maximum and average values among the weeks from 1 to 25

	Sample paths			Real execution		
Configuration	Min	Max	Average	Min	Max	Average
A	0.00	9.90	1.85	0.25	16.27	4.55
B	0.00	9.24	1.80	0.07	13.29	4.55
C	0.00	10.98	1.51	0.10	13.59	3.72
D	0.00	9.46	1.55	0.41	11.53	3.95
E	0.00	13.35	2.17	0.64	16.12	5.81
F	0.01	11.79	2.30	0.92	19.16	5.94

5.4.1 Execution of the Assignments

Each obtained solution is applied to 10 sample paths (generated with the same procedure of [6, 7]) and to the real historical patients' demands.

The quality of the solutions is analyzed in terms of the mean overtime cost per nurse. This is obtained at each week as the ratio between the total cost of the district (computed with the same levels $c_l = l$ and thresholds $\Delta_i^l = 0.1 v_i$) and the number of nurses in the district. This indicator is directly taken for the execution with the historical demands, while for the sample paths the analyzed indicator is the average at each week among the paths. Hence, for each configuration and district, the result is the list of average costs over the weeks in two cases: executed with the historical demands or averaged among the sample paths (Table 5.5). We remark that planned costs, reported in Table 5.4, refer to the entire planning horizon (8 weeks), while for the execution only the first week of the planning horizon is extracted from each rolling week.

Results show that robust solutions perform better than their non-robust counterparts, both for sample paths and real data; thus, robustness provides the desired cost savings. To give an idea of the obtained cost savings, we can assume that one unit of cost corresponds to about 15 euros. Considering that the 4 districts include 20 nurses

(see Table 5.1) and that the observed period refers to 25 weeks, each cost reduction of 1 unit corresponds to a global saving of 7,500 euros in the period. As example, comparing solution C with the corresponding non robust solution F, a global cost saving of 16,650 euros is observed for the real execution, and of 5,925 euros for the average among the paths.

Considering the detail of each district, the main benefits are obtained in districts NPA and PA both in terms of average and maximum values. A low benefit is observed in district NPC and hardly any benefit in district NPB. Then, it seems that larger benefits are obtained in the presence of critical patients with higher demands (i.e., palliative patients) or many nurses.

It must be stressed that non robust models are always solved to optimality, while an optimality gap is accepted for the robust counterparts. A robust, even if sub-optimal, solution computed in reasonable time is able to improve the solution upon its optimal non-robust counterpart on the considered case study.

Finally, if compared to other methodologies applied to this instance [6, 7], the cardinality–constrained approach is able to solve problem in a lower computational time (while including the stochasticity with the scenario generation of the stochastic programming approach requires huge computational times) with few assumptions on the demands (while the analytical approach based on stochastic ordering requires to introduce many assumptions on the shape of the density functions).

5.5 Discussions and Conclusions

In this paper, we apply the robust cardinality–constrained approach proposed in [3] in the health care area and, in particular, to the nurse-to-patient assignment in HC services under continuity of care. HC is chosen because of its novelty within the health care domain and the high randomness related to the workload amount, which is strongly higher than in other services. Thanks to this approach, the deterministic assignment model is easily modified to take into account the uncertainty in patients' demands, without the necessity of assuming probability density functions or deriving a relevant number of stochastic scenarios.

The proposed model has been tested on a set of generated instances and on historical data, and it provides good quality solutions in terms of overtime costs. The application of the cardinality–constrained approach to HC is then promising. Moreover, due to the general characteristics of HC within the health care domain, the obtained benefits could extend to other health care problems.

The main limit of the proposed approach is that patients are not allowed to have a demand for visits lower than the expected value $\bar{r}_{jt}$, while in the real practice some patients have a demand lower than the expected value. Such limit could be overcome by introducing different levels of demand for each patient rather than the two ones considered in this work; this will be the aim of our future work.

References

1. Brotcorne, L., Laporte, G., Semet, F.: Ambulance location and relocation models. Eur. J. Oper. Res. **147**, 451–463 (2003)
2. Denton, B.T., Miller, A.J.: Optimal allocation of surgery blocks to operating rooms under uncertainty. Oper. Res. **58**, 802–816 (2010)
3. Bertsimas, D., Sim, M.: The price of robustness. Oper. Res. **52**(1), 35–53 (2004)
4. Boldy, D., Howell, N.: The geographical allocation of community care resources – a case study. J. Oper. Res. Soc. **31**, 123–129 (1980)
5. Hertz, A., Lahrichi,N.: A patient assignment algorithm for home care services. J. Oper. Res. Soc. **60**(4), 481–495 (2009)
6. Lanzarone, E., Matta, A., Sahin, E.: Operations management applied to home care services: the problem of assigning human resources to patients. IEEE Trans. Syst. Man Cybern. A **42**(6), 1346–1363 (2012)
7. Lanzarone, E., Matta, A.: A cost assignment policy for home care patients. Flex. Serv. Manuf. J. **24**(4), 465–495 (2012)
8. Nickel, S., Schroder, M., Steeg, J.: Mid-term and short-term planning support for home health care services. Eur. J. Oper. Res. **219**, 574–587 (2012)
9. Trautsamwieser, A., Hirsch, P.: Optimization of daily scheduling for home health care services. J. Appl. Oper. Res. **3**(3), 124–136 (2011)
10. Rasmussen, M.S., Justesen, T., Dohn, A., Larsen, J.: The home care crew scheduling problem: preference–based visit clustering and temporal dependencies. Eur. J. Oper. Res. **219**, 598–610 (2012)
11. Koeleman, P.M., Bhulai, S., Van Meeresbergen, M.: Optimal patient and personnel scheduling policies for care-at-home service facilities. Eur. J. Oper. Res. **219**, 557–563 (2012)
12. Lanzarone, E., Matta, A., Scaccabarozzi, G.: A patient stochastic model to support human resource panning in home care. Prod. Plan. Control **21**(1), 3–25 (2010)

Chapter 6
Synchronization Between Human Resources in Home Health Care Context

Maria Di Mascolo, Marie-Laure Espinouse, and Can Erdem Ozkan

Abstract This paper deals with the scheduling and routing problem in a Home Health Care structure, when synchronization is needed between two types of human resources, and time windows are considered for patients and caregivers. Our objective is to minimize the total waiting time of caregivers between patients. We give a mathematical formulation of this problem as a mixed integer linear program. We present some experiments in order to analyze the execution time and test the capability of the MILP to solve the problem in real cases, within reasonable execution times, to measure the impact of the proportion of synchronized visits, and to analyze the average workload of an operator.

6.1 Introduction

Home Health Care (HHC) is defined as medical and paramedical services delivered to patients at home. It helps patients to maintain and improve their life conditions. HHC have seen a significant evolution in France, as well as in several other countries. Among the reasons of this development, we can cite economic factors, ageing of populations, increase in the number of people with chronic diseases, congestion of the hospitals, improvements in medical technologies, and choices of the patients.

Due to its numerous specificities (resources mobility, human resources with specific skills and constraints, importance of the quality of service, uncertainties, ...), HHC has become a particularly important application area for Industrial engineering. In this paper, we are interested in the scheduling and routing problem of HHC staff (i.e. deciding which human resource visits which patient at what time).

M. Di Mascolo (✉) • M.-L. Espinouse • C.E. Ozkan
Grenoble INP/UJF-Grenoble 1/CNRS, G-SCOP UMR, 5272 Grenoble, France
e-mail: Maria.Di-Mascolo@g-scop.grenoble-inp.fr;
marie-laure.espinouse@g-scop.grenoble-inp.fr; canerdemozkan@live.com

A. Matta et al. (eds.), *Proceedings of the International Conference on Health Care Systems Engineering*, Springer Proceedings in Mathematics & Statistics 61,
DOI 10.1007/978-3-319-01848-5_6,

We focus on the case when two human resources are required at the same time for some cares. This synchronization between resources occurs for example when a nurse and an auxiliary nurse are simultaneously required for a patient who needs help to get in or out of bed.

This paper is organized as follows: we define our problem in Sect. 6.2, and discuss of related work in Sect. 6.3. Section 6.4 proposes then a mathematical formulation for our problem as a mixed integer linear programming model, and some experiments are presented in Sect. 6.5.

6.2 Problem Description

As pointed out in [1–4], the synchronization, in HHC context, between the visits of the different stakeholder is a difficult and crucial problem.

Here a pair wise synchronization is studied between nurses and auxiliary nurses. The coordination of human resources corresponds to the pair wise synchronization constraints for multiple traveling salesman problem formulation. Synchronization implies temporal constraints.

Furthermore, hard constraints as working hours of human resources and time windows of patients must be taking into account. Time windows of patients represent either wishes of the patient or medical constraints.

In this study, short term planning is treated. Per time period considered, each patient must have exactly one visit. For this visit either a nurse is required or an auxiliary nurse is required or a nurse and an auxiliary nurse are required.

Our objective is to minimize the total waiting time of operators between patients. This objective is very important in practice. Indeed, the nurses and auxiliary nurses, besides the medical tasks also have preparatory tasks and administrative tasks to realize. It is thus important that the nurses and the auxiliary nurses have time at the beginning and at the end of a tour to realize these tasks and for that they should not waste too much time waiting between patients. Let us note that this waiting time is all the more important within the framework of this study as, on one hand, windows of patients are considered and, on the other hand, synchronization between the nurses and the auxiliary nurses are taken into account. As far as we know, in the literature, this objective is never used as an objective function.

We propose a Mixed Integer Linear Programming Model, and the data used for the tests is inspired by real data and is created in a random way.

6.3 Literature Review

In the literature, several issues are considered while dealing with resource planning of HHC, such as the resource dimensioning, partitioning of a territory into districts, allocation of resources to districts, assignments of care providers to patients, or the

visits and the resource scheduling and routing. The most frequently treated issue is the last one, routing and scheduling. Readers can refer to Yalçındağ et al. [5] for a review of papers addressing the scheduling and routing problem as a Travelling Salesman Problem (TSP) or Vehicle Routing Problem (VRP) in the HHC context.

We focus here on the papers addressing the problem of coordination between human resources.

There are some papers considering "shared visits" [6–8]. They all consider human resources with the same qualifications, who should sometimes be more than one for some visits.

Here, we are more especially interested with papers dealing with synchronization constraints between human resources who have different qualifications.

Bredström and Rönnqvist [1] develop a general branch and price algorithm for routing and scheduling problem with time windows. The problem is formulated as a set partitioning problem, considering synchronization constraints. LP relaxations are used in order to solve the problem. In [2], they go further by considering both synchronization and precedence constraints as temporal constraints. They use a multi criteria objective function, minimizing preferences, travelling time, and maximal workload difference, and propose a heuristic to solve their model.

Kergosien et al. [3] propose an integer linear programming model and propose some technical improvements to solve the routing problem in HHC context. They consider the problem under the constraint of synchronization, disjunction (some operators cannot work together), time windows for operators and patients, and continuity of care. They formulate the problem as a multiple traveling salesman problem with time windows with some additional constraints with the objective of minimizing total travelling distance. They test the model on randomly generated instances with Cplex solver. Results show that the proposed integer linear program is not able to deal with instances of real size.

More recently, Rasmunssen et al. [4] consider four temporal constraints. They formulate the problem as a set partitioning problem, and develop visit clustering schemes for home care personnel, in order to explore how much they decrease run times, and how much they compromise optimality. They propose LP-based branch and price framework. The algorithm is tested with real life problem instances. Results show that visit clustering schemes decrease the execution time significantly but cause a loss of quality for a few instances. Furthermore they outline that visit clustering schemes allow finding solutions that could not be solved to optimality.

Note that all these papers dealing with synchronization or shared visits use total travelling cost/distance as objective function. Some of them [1, 2, 4] consider a multi objective function, considering also visit time preference for [1], referential operator for [2, 4], number of uncovered visits for [4], and workload difference for [2]. None of them considers the same objective as ours, namely minimization of the total waiting time of operators between patients, although it is very important in practice, as explained above. As far as the constraints are concerned, these papers take into account most of the constraints that a HHC center has to deal with in practice, as we also do in our problem.

6.4 Model Description

6.4.1 Assumptions and Notations

- **Operators**: We consider two kinds of operators: nurses and auxiliary nurses and assume that there are N nurses and M auxiliary nurses. We denote by IN the set of nurses, and by IM the set of auxiliary nurses.

  We assume that nurses and auxiliary nurses might have different working hours.

  Nurses and auxiliary nurses can take care of a patient related to their qualification. All the nurses have the same qualification, and all the auxiliary nurses have the same qualification.

  Each operator has to start/finish his/her work at Home Health Care Center (HHCC), which is denoted by 0 when it is the starting point, and by d when it is the ending point.
- **Patients**: There are P patients, needing each exactly one visit per time period considered. Among them, there are PN patients who need a care by a nurse, PAN patients who need a care by an auxiliary nurse, and PS patients who need a simultaneous care by a nurse and an auxiliary nurse (synchronization of two different operators). We denote by IP_I the set of patients that need a care by a nurse: $IP_I = \{1, \ldots, PN\}$, by IP_{AN} the set of patients that need a care by an auxiliary nurse: $IP_{AN} = \{PN+1, \ldots, PN+PAN\}$, by IP_{sync} the set of patients that need a simultaneous care by a nurse and an auxiliary nurse: $IP_{sync} = \{PN+PAN+1, \ldots, PN+PAN+PS\}$ and by IP the set of all the patients ($IP = IP_{NI} \cup IP_{AN} \cup IP_{sync}$)

Each patient has a time window within which the operator(s) has to arrive at the patient's house. Duration of care for each patient and travelling time between each patient are fixed. Note that for these parameters, HHCC is considered as a patient.

6.4.2 Problem Formulation as Mixed Integer Linear Programming Model

6.4.2.1 Indexes

n: $1, \ldots, N$, for nurses
m: $1, \ldots, M$, for auxiliary nurses
i: $1, \ldots, P$ for patients

6.4.2.2 Parameters

$[a_n,b_n]$: Working hours of nurse $\boldsymbol{n}$, $n \in IN$
$[c_m,d_m]$: Working hours of auxiliary nurse $\boldsymbol{m}$, $m \in IM$
$[e_i,l_i]$: Time window of patient $\boldsymbol{i}$ for the arrival time, $i \in IP_{sync} \cup IP_N \cup IP_{AN} \cup \{0,d\}$
D_i: Duration of care for patient $\boldsymbol{i}$, $i \in IP_{sync} \cup IP_N \cup IP_{AN} \cup \{0\}$
T_{ij}: Travelling time between patient $\boldsymbol{i}$ and patient $\boldsymbol{j}$, $i \in IP_{sync} \cup IP_N \cup IP_{AN} \cup \{0\}$; $j \in IP_{sync} \cup IP_{AN} \cup IP_N \cup \{d\}$
A: Big number with $A \geq \max_{m,n} \{a_n, c_m\}$
B: Big number with $B \geq \max_{m,n} \{b_n, d_m\}$

6.4.2.3 Decision Variables

Nurses

$$X_{nij} = \begin{cases} 1 \; if \; nurse \; n \; takes \; care \; of \; patient \; j \; immediately \; after \; patient \; i \\ 0 \; otherwise \end{cases}$$

$$\forall i \in IP_{sync} \cup IP_N \cup \{0\}; \forall j \in IP_{sync} \cup IP_N \cup \{d\}; i \neq j; \forall n \in IN$$

t_{nj}: Arrival time of nurse $\boldsymbol{n}$ to the house of patient $\boldsymbol{j}$.

$$\forall j \in IP_{sync} \cup IP_N \cup \{0,d\}; \forall n \in IN$$

Auxiliary Nurses

$$Y_{mij} = \begin{cases} 1 \; if \; auxiliary \; nurse \; m \; takes \; care \; of \; patient \; j \; immediately \; after \; patient \; i \\ 0 \; otherwise \end{cases}$$

$$\forall i \in IP_{sync} \cup IP_{AN} \cup \{0\}; \forall j \in IP_{sync} \cup IP_{AN} \cup \{d\}; i \neq j; \forall m \in IM$$

s_{mj}: Arrival time of auxiliary nurse $\boldsymbol{m}$ to the house of patient $\boldsymbol{j}$.

$$\forall j \in IP_{sync} \cup IP_{AN} \cup \{0,d\}; \forall m \in IM$$

Waiting Time

Q_{nij}: Waiting time of nurse $\boldsymbol{n}$ between patient $\boldsymbol{i}$ and patient $\boldsymbol{j}$.

$$\forall i \in IP_N \cup IP_{sync} \cup \{0\}; \forall j \in IP_N \cup IP_{sync} \cup \{d\}; i \neq j; \forall n \in IN$$

S_{mij}: Waiting time of auxiliary nurse $\boldsymbol{m}$ between patient $\boldsymbol{i}$ and patient $\boldsymbol{j}$.

$$\forall i \in IP_{AN} \cup IP_{sync} \cup \{0\}; \forall j \in IP_{AN} \cup IP_{sync} \cup \{d\}; i \neq j; \forall m \in IM$$

6.4.2.4 Mathematical Formulation

$$\min \sum_n \sum_i \sum_j Q_{nij} + \sum_m \sum_i \sum_j S_{mij}$$

Subject to:

$$\sum_n \sum_i X_{nij} = 1 \qquad \forall n \in IN; \forall i \in IP_{sync} \cup IP_N \cup \{0\}; \ \forall j \in IP_{sync} \cup IP_N; \ i \neq j \tag{6.1}$$

$$\sum_m \sum_i Y_{mij} = 1 \qquad \forall m \in IM; \forall i \in IP_{sync} \cup IP_{AN} \cup \{0\}; \ \forall j \in IP_{sync} \cup IP_{AN}; i \neq j \tag{6.2}$$

$$\sum_j X_{n0j} = \sum_j X_{njd} = 1 \qquad \forall j \in IP_N \cup IP_{sync}; \quad \forall n \in IN \tag{6.3}$$

$$\sum_j Y_{m0j} = \sum_j Y_{mjd} = 1 \qquad \forall j \in IP_{AN} \cup IP_{sync}; \quad \forall m \in IM \tag{6.4}$$

$$X_{n0d} = 0 \qquad \forall n \in IN \tag{6.5}$$

$$Y_{m0d} = 0 \qquad \forall m \in IM \tag{6.6}$$

$$\sum_j X_{nij} = \sum_k X_{nki}. \qquad \forall i \in IP_{sync} \cup IP_N; \quad j \in IP_{sync} \cup IP_N \cup \{d\}; \quad k \in IP_{sync} \cup IP_N \cup \{0\}; \quad i \neq j; \quad i \neq k; \quad \forall n \in IN \tag{6.7}$$

$$\sum_j Y_{mij} = \sum_k Y_{mki} \qquad \forall i \in IP_{sync} \cup IP_{AS}; j \in IP_{sync} \cup IP_{AS} \cup \{d\}; \quad k \in IP_{sync} \cup IP_{AS} \cup \{0\}; i \neq j; i \neq k; \forall m \in IM \tag{6.8}$$

$$l_i * \sum_j X_{nij} \ \geq \ t_{ni} \geq e_i * \sum_j X_{nij} \qquad \forall i \in IP_N \cup IP_{sync}; \quad j \in IP_N \cup IP_{sync} \cup \{d\}; i \neq j; \quad \forall n \in IN \tag{6.9}$$

$$l_i * \sum_j Y_{mij} \ \geq \ s_{mi} \geq e_i * \sum_j Y_{nij} \qquad \forall i \in IP_{AN} \cup IP_{sync}; \quad j \in IP_{AN} \cup IP_{sync} \cup \{d\}; i \neq j; \quad \forall m \in IM \tag{6.10}$$

$$t_{nj} + (1 - X_{nij}) * l_i \geq \ t_{ni} + (D_i + T_{ij}) * X_{nij} \quad \forall i \in IP_{sync} \cup IP_N \cup \{0\}; \quad \forall j \in IP_{sync} \cup IP_N \cup \{d\}; i \neq j; \quad \forall n \in IN. \tag{6.11}$$

$$s_{mj} + (1 - Y_{mij}) * l_i \geq s_{mi} + (D_i + T_{ij}) * Y_{mij}$$
$$\forall i \in IP_{sync} \cup IP_{AN} \cup \{0\}; \quad \forall j \in IP_{sync} \cup IP_{AN} \cup \{d\}; i \neq j; \quad \forall m \in IM \tag{6.12}$$

$$t_{n0} \leq t_{nj} + A * \left(1 - \sum_i X_{nij}\right)$$
$$i \in IP_N \cup IP_{sync}; \quad \forall j \in IP_N \cup IP_{sync} \cup \{d\}; i \neq j; \quad \forall n \in IN \tag{6.13}$$

$$s_{m0} \leq s_{mj} + A * \left(1 - \sum_i Y_{mij}\right)$$
$$i \in IP_{AN} \cup IP_{sync}; \quad \forall j \in IP_{AN} \cup IP_{sync} \cup \{d\}; i \neq j; \quad \forall m \in IM \tag{6.14}$$

$$t_{nd} \geq t_{nj} \qquad \forall n \in IN; \quad \forall j \in IP_N \cup IP_{sync} \cup \{0\} \tag{6.15}$$

$$s_{md} \geq s_{mj} \qquad \forall m \in IM; \quad \forall j \in IP_{AN} \cup IP_{sync} \cup \{0\} \tag{6.16}$$

$$t_{n0} \geq a_n \qquad \forall n \in IN \tag{6.17}$$

$$s_{m0} \geq c_m \qquad \forall m \in \mathrm{IM} \tag{6.18}$$

$$t_{nd} \leq b_n \qquad \forall m \in IM \tag{6.19}$$

$$s_{md} \leq d_m \qquad \forall m \in IM \tag{6.20}$$

$$\sum_n t_{nj} - \sum_m s_{mj} = 0 \qquad \forall j \in IP_{sync} \tag{6.21}$$

$$t_{nj} - (t_{ni} + D_i + T_{ij}) \leq Q_{nij} + B * (1 - X_{nij})$$
$$\forall i \in IP_N \cup IP_{sync} \cup \{0\}; \quad \forall j \in IP_N \cup IP_{sync} \cup \{d\}; \quad i \neq j; \quad \forall n \in IN \tag{6.22}$$

$$s_{mj} - (s_{mi} + D_i + T_{ij}) \leq S_{mij} + B * (1 - Y_{mij})$$
$$\forall i \in IP_{AN} \cup IP_{sync} \cup \{0\}; \quad \forall j \in IP_{AN} \cup IP_{sync} \cup \{d\}; \quad i \neq j; \quad \forall m \in IM \tag{6.23}$$

$$X_{nij} \in \{0,1\} \qquad \forall i \in IP_{sync} \cup IP_N \cup \{0\}; \quad \forall j \in IP_{sync} \cup IP_N \cup \{d\}; i \neq j; \quad \forall n \in IN \tag{6.24}$$

$$Y_{mij} \in \{0,1\} \qquad \forall i \in IP_{sync} \cup IP_{AN} \cup \{0\}; \quad \forall j \in IP_{sync} \cup IP_{AN} \cup \{d\};$$
$$i \neq j; \quad \forall m \in IM \tag{6.25}$$

$$t_{nj} \in R_+ \cup \{0\} \qquad \forall j \in IP_{sync} \cup IP_N \cup \{0, d\}; \quad \forall n \in IN \tag{6.26}$$

$$s_{mj} \in R_+ \cup \{0\} \qquad \forall j \in IP_{sync} \cup IP_{AN} \cup \{0, d\}; \quad \forall m \in IM \tag{6.27}$$

$$Q_{nij} \in R_+ \cup \{0\} \quad \forall i \in IP_N \cup IP_{sync} \cup \{0\}; \quad \forall j \in IP_N \cup IP_{sync} \cup \{d\}; \ i \neq j; \ \forall n \in IN \tag{6.28}$$

$$S_{mij} \in R_+ \cup \{0\} \quad \forall i \in IP_{AN} \cup IP_{sync} \cup \{0\}; \quad \forall j \in IP_{AN} \cup IP_{sync} \cup \{d\}; \ i \neq j; \ \forall m \in IM \tag{6.29}$$

Our objective is to minimize the sum of waiting times of operators between patients. We must remember that waiting time of each resource at HHCC is not considered. This occurs in two different ways. First, an operator starts his/her work at HHCC and waits before visiting a patient. Second, an operator finishes his/her work at HHCC before ending working time. We do not deal with these two cases, because a resource can spend the waiting time at HHC with paper works.

Constraint sets (6.1) and (6.2) ensure that each patient $\boldsymbol{j}$ that needs a care from a nurse (resp. an auxiliary nurse) is visited by only one nurse (resp. an auxiliary nurse).

Constraint sets (6.3) and (6.4) ensure that each nurse (resp. an auxiliary nurse) has to leave HHCC and return to HHCC. Constraint sets (6.5) and (6.6) avoid that an operator visits only $\{0,d\}$ that corresponds to HHCC, making sure that each operator works.

Constraint sets (6.7) and (6.8) ensure that if a nurse (resp. an auxiliary nurse) enters a patient's house, he/she has to leave it.

Constraint sets (6.9) and (6.10) ensure that each operator can arrive at a patient's house respecting his/her time window.

Constraint sets (6.11) and (6.12) formulate arrival time to the patients. Here we deal with the duration of care and travelling time of an operator between two patients that he/she takes care of, considering that these two patients are visited one after the other one.

Constraint sets (6.13) and (6.14) make sure that if an operator visits patient $\boldsymbol{j}$, his/her arrival time will be greater or equal to the arrival time to HHCC. When an operator does not visit a patient, his/her arrival time to this patient will be zero. Constraint sets (6.15) and (6.16) force arrival time of operators to HHCC at the end of the day to be greater or equal than the arrival time to any patient $\boldsymbol{j}$. Constraint sets (6.13), (6.14), (6.15) and (6.16) also force each operator to start/finish at HHCC. We need these constraints for our objective.

Constraint sets (6.17) and (6.18) ensure that each operator can start his/her work after his/her beginning of working time. Constraint sets (6.19) and (6.20) ensure that each operator has to finish his/her work before his/her ending of working time.

Constraint sets (6.21) are for the patients who need a simultaneous care by a nurse and an auxiliary nurse (synchronization between two different operators). Constraint sets (6.21) force two different operators to arrive to the patients at the same time.

Constraint sets (6.22) and (6.23) formulate the waiting time of operators between two patients that he/she visits consecutively.

Constraint sets (6.24) and (6.25) force decision variables to take binary values. Constraint sets (6.26), (6.27), (6.28) and (6.29) ensure that decision variables take positive real number values.

6.5 Experiments

6.5.1 Data Generations

The data, namely working hours of operators, time windows of patients, duration of care and travelling time, is generated from a real case of the region Rhône-Alpes in France: Grenoble HHCC. This specific case is one of the biggest of this region. Our objective is to test the limits of MILP and show that our MILP is capable to solve a real and big case. The data is generated according to the answers collected by a survey done during the regional project "OSAD" [9] and, more particularly, the interviews done with Grenoble HHCC.

Operators: We need to determine beginning and end of working hours for each operator (nurses and auxiliary nurses). The data generation for nurses and auxiliary nurses is the same in nature. We define two cases in order to deal with the working hours of part time operators. Each generated operator belongs either to the first case or to the second case. The choice between these cases depends on a random number between 0 and 1. If the random variable belongs to [0, 0.5], the operator belongs to the first case, else it belongs to the second case.

Case 1: For each operator belonging to case 1, we determine the beginning of his/her working hours, while the end of his/her working hours is fixed to 300 min. The beginning of working hours of each operator is determined according to a random number which is between 0 and 1. To avoid short working hours, we give 40% chances to start at 0 min, 30% chances to start at 60 min, 20% chances to start at 120 min, 10% chances to start at 180 min.

Case 2: For each operator belonging to case 2, we determine the end of his/her working hours, while the beginning of his/her working hours is fixed to 0 min. The end of working hours is determined by a random variable which is between 0 and 1. As in case 1, in order to avoid short working hours, we give different chances for the end of working hours: Each operator has 10% chances to finish at 120 min, 20% chances to finish at 180 min, 30% chances to finish at 240 min, 40% chances to finish at 300 min.

Patients: We can group the patients within three categories, according to the required care giver: first, patients needing a care by a nurse; second, patients needing a care by an auxiliary nurse, and third, patients needing a simultaneous care by a nurse and an auxiliary nurse. The data generation of each patient is the same in nature. As data, we need the time window of the operator arrival time, and the duration of care for each patient.

Time window of arrival is composed of the following data: the earliest arrival time and the latest arrival time. They are determined as follows: the value of the earliest arrival time depends on a random number between 0 and 1. It can be 0, 30, 60, 90 or 120, and each one has an equal probability of 0.2.

The latest arrival time is determined in a different way: we first determine the length of the time window. This length is added to the earliest arrival time, in order to calculate the latest arrival time. The length of time window depends on a random variable between 0 and 1. It can be 60, 90, 120, 150 or 180 min and each one has an equal probability of 0.2.

Duration of care ranges between 20 and 180 min. The majority of cares take 45 min. We decide to determine the care duration as follows: we suppose that we have three different intervals for care duration. The first interval of care duration is between 20 and 35 min. 8% of the considered patients belong to this interval. The second one is between 35 and 55 min. 84% of the patients belong to this second interval. The third interval is between 55 and 180 min 8% of patients belong to this interval.

The care duration belonging to each of these intervals is determined by a uniform random number between the limits of the interval in question.

Travelling Time: We define an area for the locations of patient's house and HHCC. We consider a Cartesian coordinate system which is limited between 0 and 40 km for x axis and between 0 and 40 km for y axis. This area is inspired from Grenoble HHCC. We suppose that the area takes place in the positive side of the coordinate plane. Each patient (including HHCC) is defined as a point in the Cartesian coordinate system in two dimensions (x, y). We suppose that HHCC is located in the middle of the coordinate system, which is the point (20, 20). The location of each patient is determined randomly by generating one random number between 0 and 40 for each axis x and y.

Note that the travelling distance between each location corresponds to the travelling time. The travelling time between each patient, and between each patient and HHCC is calculated as the Euclidian distance between two points of the plane with Cartesian coordinates (x_1, y_1) and (x_2, y_2).

Instances Generation: In order to create the test problems, we create pools of operators and patients. We have two operator pools; a nurse pool containing 20 nurses and an auxiliary nurse pool containing 20 auxiliary nurses.

We have three patient pools, according to the required care giver. First pool is formed by 16 patients needing a care by a nurse; second pool is composed of 16 patients needing a simultaneous care by a nurse and an auxiliary nurse, and the third pool is formed by 16 patients needing a care by an auxiliary nurse.

Each operator and patient in each pool has a reference number. For each instance, a random number is generated in order to choose the operators and patients according to the reference number. Selection of operators (i.e. nurses and auxiliary nurses) and patients (three different categories according to the required care giver) is done separately.

For the experiments, we define the size of the problem as (number of nurses, number of auxiliary nurses, number of patients, and percentage of synchronized care). For example, instance (5, 5, 20, 30%) means that we randomly chose 5 nurses (resp 5 auxiliary nurses) among the 20 within the nurse pool (resp. the auxiliary nurse pool), and 20 patients among the 3*16 patients within the three patient pools. These 20 patients are chosen by considering the wished percentage of synchronized care, which means here that 30% of these 20 patients belong to the second pool (patients needing a simultaneous care by a nurse and an auxiliary nurse)

6.5.2 Results

We first analyze the execution time and test the capability of the MILP to solve the problem of the HHCC in the Rhône Alpes region, within reasonable execution times, second we measure the impact of the proportion of synchronized visits for the 40 patients' problem size. Next we analyze the average workload of an operator. The percentage of his/her time spent taking care of patients, the percentage of his/her time spent travelling and percentage of his/her time spent in HHCC.

These experiments are done by taking into account the following information: number of operators, working hours of operators, number of patients, time windows of patients, duration of care and travelling time.

ILOG on CLPEX 12.2 OPL STUDIO is used to solve the test problems. The resolution time is limited to 1 h for ILOG. The experiments have been conducted with CPU 3 GHz, 4 Go of RAM and Windows 7 (64 bits).

6.5.2.1 Analysis of Execution Times

We note that according to the generated data, an operator works on average 240 min. The time horizon is 300 min. That means that an operator works on average 80% of the time horizon. Table 6.1 shows the test problem sizes, the number of instances solved within 1 h, and minimum, maximum and average execution times of solved instances.

For the instances (5, 5, 20, 30%), 80% of instances are solved within 1 h. 70% of instances are solved within 5 min while 10% of them are solved in more than 15 min. We increase the size of the problem and for the instances (10, 10, 30, 33%). 40% of instances are solved in 10 min and 10% of instances are solved in more than 50 min. The instances are solved on average within 18 min. Lastly we test the

Table 6.1 Number of instances solved within 1 h for different problem sizes and execution times of instances in minutes

Problem size	Percentage of instances solved within 1 h	Execution time in minutes		
		Minimum	Maximum	Average
(5, 5, 20, 30%)	80	0.6	16.4	3.5
(10, 10, 30, 33%)	70	2.9	57.9	17.3
(20, 20, 40, 30%)	90	2.5	32.3	13.9

Table 6.2 Number of instances solved within 1 h for different proportion of synchronized visits

Problem size	Percentage of instances solved within 1 h
(10, 10, 30, 33%)	70
(10, 10, 40, 30%)	0
(10, 10, 40, 20%)	10
(10, 10, 40, 10%)	80
(15, 15, 40, 30%)	20
(15, 15, 40, 20%)	70
(15, 15, 40, 10%)	100

instances (20, 20, 40, 30%), and 90% of instances are solved in less than 1 h, 20% of instances are solved within less than 5 min while the rest of them are solved between 10 min and 33 min. The average execution time is 14 min.

6.5.2.2 Impact of the Proportion of Synchronized Visits on the Execution Times

We tested several examples and, as we could expect, we observed that the impact of this proportion is really important. For example, as shown in Table 6.2 when we consider the example (10, 10, 30, 33%) and we try to increase the number of patients to 40, we observe that the proportion of synchronized visits has to be reduced until 10% in order to solve successfully those instances. Finally for the instances with 15 nurses, 15 auxiliary nurses and 40 patients, 20% of synchronized visits can be solved successfully within 1 h.

Note that the proportion 10% of synchronized visits is the proportion that is usually used in the literature. So we can conclude that the proposed MILP is able to solve the problems with until 40 patients in the conditions of the literature.

6.5.2.3 Analysis of the Average Workload of Operators

In order to analyze the workload of operators, three indicators are determined:

- % duration of care on working hours
- % travelling time on working hours
- % time spent in HHCC on working hours

Table 6.3 Values of three indicators for different size of the problems

	(5, 5, 20, 30%)			(10, 10, 30, 33%)			(20, 20, 40, 30%)		
Instances	Max	Min	Average	Max	Min	Average	Max	Min	Average
% duration of care on working hours	63	44	49	49	36	44	29	26	28
% travelling time on working hours	26	21	23	20	18	19	29	26	28
% time spent in HHCC on working hours	35	11	27	46	31	36	63	57	58

As it can be seen in Table 6.3, the time spent at patients' house on average decreases while increasing the size of the problem. The time spent on average for travelling first decreases and then increases while increasing the size of the problem. The range of this indicator is really small for each size. The time spent in HHCC on average is significantly increased while number of operators and patients increase. This can be explained because of the significant increase in the number of operators.

6.6 Conclusion and Future Research

In this study, we proposed a mathematical formulation for the problem coordination of human resources in home health care context. We tested the limits of the Mixed Integer Linear Programming. As a result, the MILP is able to solve different sizes of problems within 1 h. But a heuristic is required for bigger sizes of the problem. We measured the impact of the proportion of synchronized visits. As a result, the proportion of synchronized visits impacts the number of instances solved within 1 h because it affects the number of visits as well. For the problems with 40 patients, number of instances solved within 1 h is increased while reducing the proportion of synchronized visits. The average workload of an operator is analyzed.

For future works, material resource planning can be added to the problem. In this work we dealt with short term planning. Our problem can be extended to the midterm planning so that care continuity is considered. Stochasticity can be included into data generation as demand of patients, duration of care or working hours of operators.

References

1. Bredström, D., Rönnqvist, M.: A branch and price algorithm for the combined vehicle routing and scheduling problem with synchronization constraints. Technical report. Norwegian School of Economics and Business Administration, Department of Finance and Management Science (2007)
2. Bredström, D., Rönnqvist, M.: Combined vehicle routing and scheduling with temporal precedence and synchronization constraints. Eur. J. Oper. Res. **191**(1), 19–31 (2008)

3. Kergosien, C., Billaut, J.C., Lenté, J.-C.: Home health care problem an extended multiple Traveling Salesman Problem. In: Multidisciplinary International Conference on Scheduling: Theory and Applications (MISTA 2009) (2009)
4. Rasmussen, M.S., Justesen, T., Dohn, A., Larsen, J.: The home care crew scheduling problem: preference-based visit clustering and temporal dependencies. Eur. J. Oper. Res. **219**, 598–610 (2012)
5. Yalçındağ, S., Matta, A., Şahin, E.: Human Resource Scheduling and Routing Problem in Home Health Care Context: A Literature Review. ORAHS/Cardiff, United Kingdom (2011)
6. Ben Bachouch, R., Fakhfakh, M., Guinet, A., Hajri-Gabouj S.: Planification de la tournée des infirmiers dans une structure de soins à domicile, Conférence Francophone en Gestion et Ingénierie des Systèmes Hospitaliers (GISEH), Switzerland (2008)
7. Eveborn, P., Flisberg, P., Ronnqvist, M.: LAPS CARE—an operational system for staff planning of home care. Eur. Oper. Res. **171**, 962–976 (2006)
8. Thomsen, K.: Optimization on home care. Master Thesis Report, Informatics and Mathematical Modeling, Technical University of Denmark (2006)
9. Di Mascolo, M., Kharraja, S., Guinet, A., Augusto, V., Biennier, F., Teil, A.: Synthèse des résultats de l'enquête auprès des structures de HAD de la région Rhône-Alpes –OSAD project (Organisation des Soins A Domicile) funded by cluster GOSPI from Rhône-Alpes region (2011)

Chapter 7
Simulation-Based Analysis of Patient Flow in Elective Surgery

Dario Antonelli, Giulia Bruno, and Teresa Taurino

Abstract The reduction of waiting lists and length of stay in hospitals, together with an efficient utilization of system capacity is the challenge facing healthcare systems today. In an elective surgery department, as operations can be scheduled in advance, this goal is generally achieved by maximizing the utilization index of the operation theatres. Nevertheless, operations are only one of the many activities performed during patient flow inside hospital and these activities interact with each other. The optimization of any single stage of the process is pointless without an efficient management of the entire routing from admission to dismissal. The paper presents a thorough analysis of the patient flow in an elective surgery ward using data gathered in a large hospital in Italy. Data, derived from log files and questionnaires, together with solutions proposed by healthcare managers, are considered. A model is then built and validated, its parameters are defined, and a variety of experiments are simulated in order to select the solution that improves the performance of the system. The solutions are discussed and refined in the light of corresponding production management approaches.

7.1 Introduction

Health-care resources are getting more and more expensive. The administrators of health-care facilities are constantly faced with the difficult task of balancing the achievement of quality standards of health with the appropriate allocation of resources [1]. Cutting the waiting lists and the length of stay in hospital is therefore an important managerial goal for modern healthcare systems because it increases the

D. Antonelli (✉) • G. Bruno • T. Taurino
Dipartimento di Ingegneria Gestionale e della Produzione, Politecnico di Torino, Corso Duca degli Abruzzi, 24, 10129 Torino, Italy
e-mail: dario.antonelli@polito.it; giulia.bruno@polito.it; teresa.taurino@polito.it

A. Matta et al. (eds.), *Proceedings of the International Conference on Health Care Systems Engineering*, Springer Proceedings in Mathematics & Statistics 61,
DOI 10.1007/978-3-319-01848-5_7,

perceived quality of care and frees resources [2]. In elective surgery departments, system administrators can maximize managerial performance parameters only partially, as not all arrivals can be scheduled. The external performance indexes are the waiting time and the waiting list which both impact on the perceived quality. The internal performance indexes are the throughput time (time from arrival to dismissal), bed occupancy, dismissal rate and resources utilization rate. Operations management techniques show the correlations among internal and external parameters [3]. However hospital manager often prefer to adopt a more intuitive approach, trying to get the full occupation of beds and the maximum utilization of every resource. To this aim, several different tactics have been adopted: use of priority levels in the discipline of the waiting list, scheduling of patient arrival, increasing the utilization of operation theatres by reducing idle times and the redesign of the procedures for patient accommodation on wards. It is worth noting that changing this procedure is yet another way to discipline the waiting list after the patient has been hospitalized.

In present study, Discrete Event Simulation (DES) is applied to simulate the effects of interventions on pre-hospitalization and on bed allocation for an elective surgery department in an Italian hospital. The main factors influencing patient flow are extracted and analyzed in order to find key solutions for the improvement of the system's performance. The results are discussed by using analogies with the PULL (demand driven) production processes.

7.2 Problem Description

The case study considered in this work regards a large hospital in Italy. Just one division is taken into account for the current analysis: an elective surgery department. In management terms, it is a process with scheduled arrivals.

In order to optimize system performances, strategies proposed by the hospital management were simulated through experiments and were compared, extending a method already applied in a former study [4].

Simulated experiments were conducted by following prescribed formal stages: system observation, data collection, model implementation, run and validation, output analysis.

Several practical issues arose during the experiment such as errors in the collected data, high variability of system parameters, and self-adjusting behavior of personnel.

It is important to bound the analyzed case study on the type of patient considered, the inpatient. An inpatient is "admitted" to the hospital and stays overnight or for an indeterminate time. An early selection of inpatients from the outpatients could considerably reduce the waiting time. Thus, as diagnostic is not an exact science, it is unavoidable that triage admits some outpatients, too.

An important performance index, directly perceived by every patient, is related to the length of the waiting time before hospitalization. In order to improve this issue, it is possible to adopt different tactics:

- Queue discipline based on priority rules (already adopted).
- Improve the scheduling of patient arrival.
- Increase the utilization of surgery rooms.
- Redesign the procedures for the accommodation in wards beds.

Also patient scheduling was adopted by many hospitals but not everywhere [5]. Scheduling is effective only when the scheduled system is deterministic or with low variability. This is not the case as recovery times display a variance equivalent to the mean times. Therefore the ward under analysis uses a flexible scheduling in which it is scheduled only the date from which the patient should be ready for hospitalization, with the results of the diagnostic exams. Starting from that date, the actual hospitalization will occur as soon as a bed is actually free.

Alternatively pre-hospitalization analysis are a way to hospitalize patients just in time for the operation, saving beds [6].

Another improvement would be to cluster beds in two groups: standard stay patients and long terms patients. These latter delay the admission of new patients to surgery. The relative size of the two groups can be reallocated based on the demand [7]. Experiments on the actual patients are not advisable therefore it was decided to have recourse to simulated experiments.

Several approaches could be used to model and optimize patient flow: Markov and semi-Markov models, queuing theory, solved analytically or by discrete event simulation [8, 9]. Queuing theory models are usually based on some simple assumptions such as exponential inter-arrival and service time. However, for complex real-world systems, DES models are more flexible and adaptable [10, 11]. The model of the patient flow takes the form of a queuing network with G/G/m servers, there are m workstations in the server, the queue, intended as the waiting list, is virtually unlimited and the inter-arrival times and the process times follow a general distribution.

7.3 Elective Surgery Department

7.3.1 A Description of the Case Study

In the considered Elective Surgery Department, data were collected from different sources: the recovery logs (made anonymous) in the year 2008, integrated by inormation gathered through a questionnaire filled by hospital personnel.

When a surgery date is scheduled, the patient may be required to undertake pre-operatory analysis, such as laboratorial samples, cardiovascular and respiratory tests. Regular patients have a priority discipline for their waiting in queue. A triage is performed to assign a priority order to each single patient, with descending priorities A, B, C and D.

It is also important to state that patients ranked as B, C and D sometimes receive this designation because they must still undertake prior examinations before hospitalization that are mandatory for the surgical procedure. This forces them to wait longer before admission.

There is another category of patients, named urgent patients. Urgent patients arrive from other Wards as they have to submit to a surgical operation as a consequence of other diseases that were treated non surgically. They obviously don't have to undergo examinations as they are already hospitalized.

After entering the hospital, all patients are treated equally, disregarding their queuing priority and the surgeries follow the rule of First In First Out.

Whenever entering the hospital, a patient is allocated to a bed occupying this resource until the end of its recovery. It is clear that a patient only enters after there is a free bed.

During hospitalization, visits and examinations can be executed on patients (especially on patients A since the others had time to undergo examinations during the waiting). As a consequence of the analysis some patients are treated without recurring to surgery. Some patients may undergo complications during the surgery requiring, then, a second intervention that is executed as soon as possible. This is the only case in which the FIFO rule for the access to the operation theatre is not respected.

The previously described system was represented by means of a Process Flow Diagram that is reported in Fig. 7.1.

The process flow follows the vertical line from the top to the bottom, circles represent operations or activities, arrows represent transfers and the delay symbols represent waiting times according to the ASME (American Society Mechanical Engineering) symbology [12].

7.3.2 Data Collection

In the ward there are, totally, 24 beds, equally divided between the two genders. One of these beds is usually reserved for urgent patients. In the ward there is a single surgery room (also called operation theatre), and surgeries are only performed on Monday, Wednesday and Friday. According to the managerial staff, the estimated amount of surgeries performed in a week is 15.

The data collected from the log files of the ward cover the months of January and March 2008 for a total of 112 patients (i.e., patients that enter the hospital in those 2 months). From the logs it's possible to gather the percentage of patient types that

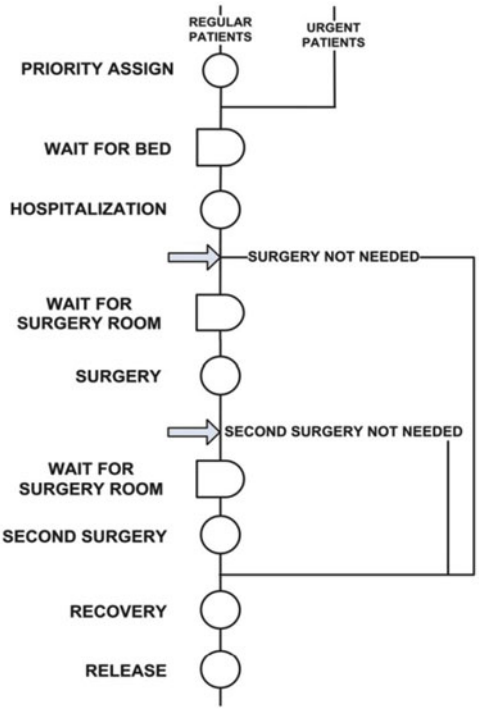


Fig. 7.1 Process Flow Diagram of the elective surgery department

Table 7.1 Basic statistics about the patients

Number of patients	Patient type	Percentage (%)
112	All	100
65	Regular A	58.0
32	Regular B	28.6
2	Regular C	1.79
1	Regular D	0.89
12	Urgent	10.7
112	First surgery	100
9	Second surgery	8.03
1	Third surgery	0.89
109	Released	97.3
2	Deceased	1.79
1	Transferred	0.89

arrived at the hospital, the number of operation they needed and the way patients got out the hospital. All these data are reported in Table 7.1. Since there are only three patients that belong to patient types C and D, and thus the number is not significant to model their distribution, they are considered as assigned to class B. The only patient that needed the third surgery was not modeled as a case a part from the others, but was included in the patients that needed two surgeries. An outlier

Table 7.2 Means and standard deviation of time spent by patients (in days)

Time	Mean	S.D.
WT1	26.2	35.8
WT2	4.33	4.86
WT3	0.46	1.93
RT	5.95	5.23
HT	10.7	8.46

Table 7.3 Means and standard deviation of time spent by patients (in days)

	Type A		Type B		Urgent	
Times	Mean	S.D.	Mean	S.D.	Mean	S.D.
WT1	18.2	23.8	49.2	46.9	–	–
WT2	4.14	3.64	3.6	3.49	7.82	10.9
WT3	0.78	2.48	–	–	–	–
RT	5.86	5.25	4.71	4.09	9.82	6.78
HT	10.9	7.00	8.31	6.78	17.6	15.8

patient that presents a waiting time before the second surgery of 89 days has been considered as an error of data entry and removed since it seems unfeasible that a patient remains for such a long time in the considered ward. The total number of considered patients is thus 111.

From the log data it's also possible to gather the times patients spent in the process. Particularly, the following times are analyzed: the waiting time before hospitalization (WT1), the waiting time from hospitalization to first surgery (WT2), the waiting time from first until second surgery (WT3), the recovery time (RT), and the total time spent inside the hospital (HT, equal to the sum of the three previous times). The derived mean and standard deviation of such times for the considered 111 patients are reported in Table 7.2.

A further analysis of these data is done in order to see differences in behaviors of patients based on their typology. The mean values and standard deviations of times for each type of patients are reported in Table 7.3.

Regarding WT1, there is a strong difference among the behavior of patients belonging to the three categories. As a matter of fact, urgent patients usually do not wait until entering the hospital, while patients A wait on average 18 days and patients B wait on average 49 days. Also the waiting for surgery (WT2) is quite different among the categories, and interestingly patients B usually wait less than the others (only 3.6 days on average), while urgent patients wait more than twice the time of patients B. This can be due to the fact that patients B have a long wait time before the hospitalization, which they can use to perform some examinations, thus saving time for when they are inside the hospital. On the contrary, urgent patients come directly to the hospital, and thus probably have to perform other kinds of analysis before being allowed to the operation.

For WT3 we do not have enough cases to differentiate among the three categories, having only 11 patients needing the second surgery, all of type A. The differences for RT are similar to the ones of WT2.

7.4 Simulation of the Process

7.4.1 Process Workflow

The process described in Fig. 7.1 has been modeled using the Rockwell Arena software [13] to perform the simulation. To assign the distribution of waiting times and patients' arrivals, data obtained by the hospitals were exploited to find the expression that best estimates data distributions. The Kolmogorov-Smirnov (K-S) test [14] is applied to select the best distribution.

The best probability distribution function that fits the arrival rate of patients is the Exponential distribution with mean equal to 0.53.

In the simulation, a priority level is randomly assigned to each patient in order to reflect proportions found in data (i.e., 58% of type A, 31% of type B and 11% of Urgent). Once the patient is assigned a type, it enters in a queue representing the waiting until there is an empty bed (WT1). The queue is of Lowest Attribute Value type, i.e., the precedence is given to patients with the lowest priority value, according to the real procedure in which patients of type A (i.e., priority level 2) have the precedence over patients of type B (i.e., priority level 3), and urgent patients (i.e., priority level 1) have the precedence over both of them.

Since all patients went through surgery, all of them spend some time waiting for the surgery room (WT2). The distributions that best fit the delays for surgery depending on patient type are reported in Table 7.4 (first column). Then, a decisional process sends some patients (7% of cases) to the second operation, represented by the delay process in which a patient waits for the second operation. The data distribution follows the expression $0.5 + 11$*BETA(0.802, 0.757). Finally, all patients perform a recovery step before leaving the hospital. From interviews to domain experts it appears that the distribution of recovery time is independent from the patient type. Therefore we put together all the values to provide an estimation of the distribution; the retrieved expression is $-0.5 + \text{GAMM}(3.11, 2.07)$. Table 7.4 reports the time distributions adopted in our simulation.

7.4.2 Simulation parameters

A simulation of the workflow of the process was executed with parameters' values reported in Table 7.5. The Warm up period was chosen using the Welch method.

Table 7.4 Process times distribution for each patient type

Patient type	WT2 distribution	WT3 distribution	RT distribution
Type A	$-0.5 + \text{WEIB}(5.02, 1.3)$	$0.5 + 11$*BETA (0.802, 0.757)	$-0.5 + \text{GAMM}$ (3.11, 2.07)
Type B	$-0.5 + \text{LOGN}(4.22, 4.5)$	"	"
Urgent	$-0.5 + \text{WEIB}(6.96, 0.76)$	"	"

Table 7.5 Simulation parameters

Parameter	Value
Number of replications	100
Warm-up period	730 days
Replication length	3,650 days

Table 7.6 Average values of obtained results for the standard case

Field	Real average value	Standard simulation average value 95% confidence interval
WT1	26.16	1.44 ± 0.11
WT2	4.33	4.41 ± 0.01
WT3	0.46	0.43 ± 0.005
RT	5.95	5.93 ± 0.01
HT	10.74	10.77 ± 0.13
Waiting patients	–	2.71 ± 0.22
Bed utilization rate	0.85	0.88 ± 0.001
Busy bed	20.4	20.28 ± 0.07

The obtained results in term of 95% confidence intervals for average values of WT1, WT2. WT3, RT, HT, number of waiting patients in the queue, bed utilization rate and number of busy beds are reported in Table 7.6 compared with real average values. The half-widths of the confidence intervals for average values suggest that an acceptable level of convergence is reached after 100 replications.

All of the values obtained in the simulation are coherent with the real data, except the waiting time WT1, that in the simulation is significantly lower. This is due to the fact that when a patient asks for a schedule, the admission date is not calculated by analyzing the current waiting list only, but adding a further 2 weeks to the date in order to allow the patient to perform some pre-operation exams. Thus, the length of this time depends not only from current resources or from organization of the ward, but also from the management rules of patients.

7.5 Proposal of Improvement

In the simulation of the ward, i.e., in the current state, the bed utilization rate is, on average, less than 90%; particularly, the utilization rate is in the range 0.88 ± 0.03. The objective of the ward's managers is to increase the utilization rate of beds to a value close to 0.95. Thus, we performed a simulation by changing the arrival rate to reach the desired utilization rate, in order to evaluate effects on waiting queues. As can be seen from Table 7.7, this change causes a sudden increase of waiting times.

The desired utilization rate is reached by decreasing the average time between arrivals from a value of 0.53 days (less than two patients a day) to a value of 0.49 days (more than two patients a day). Simulation results show that the average time spent to wait for a bed (WT1) strongly increases from 0.84 to 6.30 days with an average number of waiting patients of almost 13. In the last column of Table 7.7

Table 7.7 Average values of obtained results for the case with more patients

Field	Average value	Range average
WT1	6.30	(2.40, 16.31)
WT2	4.42	(4.27, 4.57)
WT3	0.43	(0.36, 0.49)
RT	5.94	(5.81, 6.05)
HT	10.78	
Waiting patients	12.87	(4.77, 34.30)
Bed utilization rate	0.96	(0.92, 0.99)
Busy bed	22.01	(21.24, 22.69)

range values for average obtained in replications are given. Thus, the problem becomes how to meet the manager objective without having such a worsening of performance.

The main point is that, if the ward is considered equivalent to a production line, buffers are not allowed (i.e., patients cannot be hospitalized without available beds). Therefore the ward corresponds to a pull system: the admittance of a new patient is based on the system status (availability of beds). Pull systems suffer from variability and unfortunately present case has high variability, as can be seen in Table 7.3. In industrial management, if a system displays an high variability it can be buffered by increasing the capacity, the WIP or the waiting time. Increasing the capacity (beds) has a direct cost. WIP increasing in this case is unfeasible because it corresponds to adopt an office-based surgery that has been excluded a priori for inpatients. The last way is by increasing the total cycle time that is the exact opposite to the objective of ward's managers.

To improve the system with no additional costs, another way exists: by addressing efforts directly to the reduction of variability on waiting times before surgery (WT2), for example by reducing the number of exams done during the hospitalization by increasing the pre-hospitalization activities. This operation involves a reorganization of the admission and recovery process and can be done by reinforcing a pre-hospitalization process. Infact, trying to anticipate some examinations before the admission to the ward can reduce the waiting time inside the hospital.

To simulate this scenario, the variance of waiting time before the first surgery (WT2) has been reduced by considering a process organization that admits exclusively Urgent patients and patients with a pre-hospitalization period (B patients) and by considering waiting time before the surgery equal to real average values. Table 7.8 reports results obtained by this simulation, which shows a consistent reduction of patients' waiting time and of queue length.

7.6 Conclusion

In this work an engineering approach is used to provide a process parameterization in order to reach managerial objectives of beds utilization. A simulation of the new process is done to test proposed parameters. Simulation results shows that small

Table 7.8 Average values of obtained results for the case after re-organization

Field	Average value	Range average
WT1	2.39	(1.05, 5.72)
WT2	4.06	(4.02, 4.09)
WT3	0.43	(0.36, 0.49)
RT	5.94	(5.82, 6.05)
HT	10.44	
Waiting patients	4.88	(2.09, 12.02)
Bed utilization rate	0.92	(0.89, 0.96)
Busy bed	21.26	(20.56, 22.01)

variation on the average value of inter-arrival times cause significant variations on waiting times. So a solution to find a compromise between bed utilization and waiting times is provided and simulated. The idea is to reduce variance on waiting times before surgery with a reorganization of the admittance process, placing more emphasis on the pre-hospitalization phase. On the basis of his/her objectives and requirements, the healthcare manager is provided with better guidance for an informed choice.

Acknowledgements The authors would like to thank Prof. Baudolino Mussa (University of Torino Medical School, Italy) for his support and fruitful hints.

References

1. Culyer, J.G., Cullis, J.G.: Some economics of hospital waiting lists in the NHS. J. Soc. Policy **5**(3), 239–64 (1976)
2. Antonelli, D., Bellomo, D., Bruno, G., Villa, A.: Evaluating collaboration effectiveness of patient-to-doctor interaction in a healthcare territorial network. In: PRO-VE 2012: Collaborative Networks in the Internet of Services, pp. 128–136. ISBN 9783642327742 (2012)
3. Young, T., Brailsford, S., Connell, C., Davies, R., Harper, P., Klein, J.H.: Using industrial processes to improve patient care. Br. Med. J. (BMJ) **328**, 162–164 (2004)
4. Antonelli, D., Taurino, T.: Application of a patient flow model to a surgery department. In: 2010 IEEE Workshop on Health Care Management, Venezia, 18–20 Febbraio 2010
5. Gupta, D., Denton, B.: Appointment scheduling in health care: challenges and opportunities. IIE Trans. **40**(9), 800–819 (2008)
6. Qi, E., Xu, G., Huo, Y., Xu, X.: Study of hospital management based on hospitalization process improvement. Proc. IEEE IEEM 74–78 (2006)
7. Akkerman, R., Knip, M.: Reallocation of beds to reduce waiting time for cardiac surgery. Health Care Manag. Sci. **7**, 119–126 (2004)
8. Vissers, J.: Patient flow-based allocation of inpatient resources: a case study. Eur. J. Oper. Res. **105**, 356–370 (1998). Elsevier Science
9. Xiong, H.H., Zhou, M.C., Manikopoulos, C.N.: Modeling and performance analysis of medical services systems Using Petri Nets. In: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, pp. 2339–2342 (1994)
10. Davies, R., Davies, H.T.O.: Modeling patient flows and resource provision in health systems. Omega Int. J. Manag. Sci. **22**, 123–131 (1994)

11. Koo, P.-H., Jang, J., Ielsen, K.B., Kolker, A.: Simulation-based patient flow analysis in an endoscopy unit, Health Care Management (WHCM). In: 2010 Institute of Electrical and Electronics Engineers (IEEE) Workshop on, pp.1–6, 18–20 Feb 2010
12. American Society of Mechanical Engineers: ASME Standard Operation and Flow Process Charts. New York (1947)
13. Kelton, W.D., et al.: Simulation with Arena. McGraw-Hill Professional, New York, NT, US (2006)
14. Smirnov, N.S.: Tables for estimating the goodness of fit of empirical distributions. Ann. Math. Stat. **19**, 279 (1948)

Chapter 8
Optimizing Efficiency and Operations at a Large California Safety-Net Endoscopy Center: A Modeling and Simulation Approach

Lukejohn W. Day, David Belson, Maged Dessouky, Caitlin Hawkins, and Michael Hogan

Abstract Improvements in endoscopy center efficiency, especially in safety net hospitals, are needed, but scant data are available. A time and motion study was performed and a discrete simulation model constructed to assess changes in scheduling, staffing models, and the pre- and post-procedure process and its impact on several performance measures in a safety net hospital endoscopy center. Decreasing the endoscopy appointment time from 60 to 45 min led to a 21% rise in the number of procedures performed per week, but unfortunately increased patient wait time by 42% while further reductions in appointment times led to even more significant queuing. However, increasing the number of pre-procedure nurses from 1.5 to 2 resulted in a 22% increase in the number of procedures performed per week and increased provider, nurse and procedure room utilization with minimal impact on patient wait time. Further increases in nurse staffing resulted in no significant changes to measured outcomes. Increasing the number of endoscopists by one each half day resulted in procedure volume rising, but there was a concomitant rise in patient wait time and nurse utilization exceeding capacity. A significant improvement in performance metrics was created by moving patient appointments from afternoon to morning appointments. In this simulation at 45 and 40 min appointments procedure volume rose by 23 and 34% respectively, all utilization metrics increased and patient time spent in the endoscopy center declined by 17 and 13%. Thus the combination of minor, cost-effective changes such as reducing

L.W. Day (✉)
Division of Gastroenterology, San Francisco General Hospital and Trauma Center, San Francisco, CA, USA
e-mail: lukejohn.day@ucsf.edu

D. Belson • M. Dessouky • C. Hawkins • M. Hogan
Daniel J. Epstein Department of Industrial and Systems Engineering, University of Southern California, Los Angeles, CA, USA
e-mail: belson@usc.edu; maged@usc.edu; cait.hawkins@gmail.com; mchogan@gmail.com

A. Matta et al. (eds.), *Proceedings of the International Conference on Health Care Systems Engineering*, Springer Proceedings in Mathematics & Statistics 61,
DOI 10.1007/978-3-319-01848-5_8,

appointment times, minimizing and standardizing recovery time, and making small increases in pre-procedure ancillary staff maximized endoscopy center efficiency across a number of performance metrics. The simulation made it possible to identify which changes were desirable and to what extent.

8.1 Introduction

There has been a dramatic rise in the request for gastrointestinal (GI) specialty care, and in particular endoscopic services, over the last decade. At the same time, access to GI care in the safety net healthcare system is limited. Such disparity highlights the need for creative and innovative ways to increase access to GI care for underserved patient populations. A method to address this inequality is to develop more efficient endoscopy centers that can provide increased endoscopic services while at the same time maximize patient and provider satisfaction.

There is a dearth of information on the study of efficiency in endoscopy centers. Of the scant literature available there are varying conclusions about how to improve endoscopy center efficiency with no clear consistent message. Some studies have focused on altering staffing specifically focusing on the endoscopist [1, 5, 14] and utilizing more staff in the pre-procedure process [5]. While such changes improve physician efficiency and utilization, it does so at impairing non-physician staff utilization, sub-optimizing facility utilization and increasing patient length of stay [14]. Using simulation modeling others have discovered that identifying bottlenecks in patient recovery [3, 13], reducing room turnover time [6], modifying the patient arrival schedule [1, 7, 11] or reengineering the scheduling of patients [4, 12] can improve efficiency and decrease patient time in the endoscopy center. However, there are a number of limitations to these studies; they are small, examine efficiency solely from a physician perspective, and all are set in either an ambulatory endoscopy center or tertiary referral service. Given these deficiencies and with changes to the U.S. healthcare system, with more underserved patients being cared for, it is imperative to better understand safety net endoscopy centers and to improve efficiency within them.

Our objective was to conduct a time and motion study of clinic work and use this data in simulation modeling to study changes in scheduling, staffing models, facility changes and changes in the pre- and post-procedure process in a safety net hospital endoscopy center. The simulation objective was to understand the bottlenecks limiting the endoscopy center's current operational performance and, in turn, to identify opportunities to improve patient throughput while balancing resource utilization and patient wait times.

8.2 Methods

8.2.1 Study Design, Setting and Population

We conducted a time and motion study of the San Francisco General Hospital and Trauma Center (SFGH) endoscopy center and performed discrete simulation modeling to assess proposed changes to the endoscopy center with respect to specific performance and efficiency metrics. The study was conducted between November 2011 and May 2012. The SFGH endoscopy center provides subspecialty care for the safety net healthcare system of the City and County of San Francisco.

The SFGH endoscopy center performs colonoscopies and upper endoscopies as well as other advanced procedures in an ethnically diverse patient population. The majority of the endoscopy center's time is devoted to performing colonoscopy and upper endoscopies (EGD) (89.0% of procedure volume) with a no show rate of 17.7%. SFGH is a teaching hospital for the University of California, San Francisco's medical school that has three GI fellows and one surgical resident rotating through the GI Division each month.

Prior to constructing a discrete event simulation model, multiple days of direct time observations and interviews were conducted to identify patient flow, key parameters and process attributes. Time was spent shadowing physicians, nurses, and support staff at the endoscopy center in order to develop an understanding of the work flow.

The SFGH endoscopy center has four distinct workflow processes: check-in, pre-procedure, procedure and recovery (Fig. 8.1). A patient's visit begins at check-in after which patients move to a waiting room where they remain until called to the pre-procedure room. Patients complete the pre-procedure process in a dedicated pre-procedure space (maximum of three beds). In situations where a procedure room is available and no prepared patients are waiting to begin a procedure, pre-procedure activities are conducted in the procedure room. From the pre-procedure process, a patient then moves to a procedure room. At the conclusion of the procedure, patients either recover in the recovery room (maximum three beds), or if a recovery bed is unavailable then patients are kept in the procedure room. Once in the recovery room, patients stay for at least 30 min as required by state regulations. Patients are held in the recovery room until a ride home arrives to sign them out.

Observation and timing of the processes was done to provide a statistically significant picture of operations. Arrival times were collected from the hospital's appointment scheduling system. The pre-procedure process was quantified using a paper form that nurses completed. Procedure data was collected from the time of endoscope insertion and removal, as documented in procedure and nursing notes. Recovery data was collected from time stamps present on discharge paperwork. Observation and use of the SFGH endoscopy center's electronic record keeping system provided 278 patient arrival times, 257 procedure times and 257 recovery times.

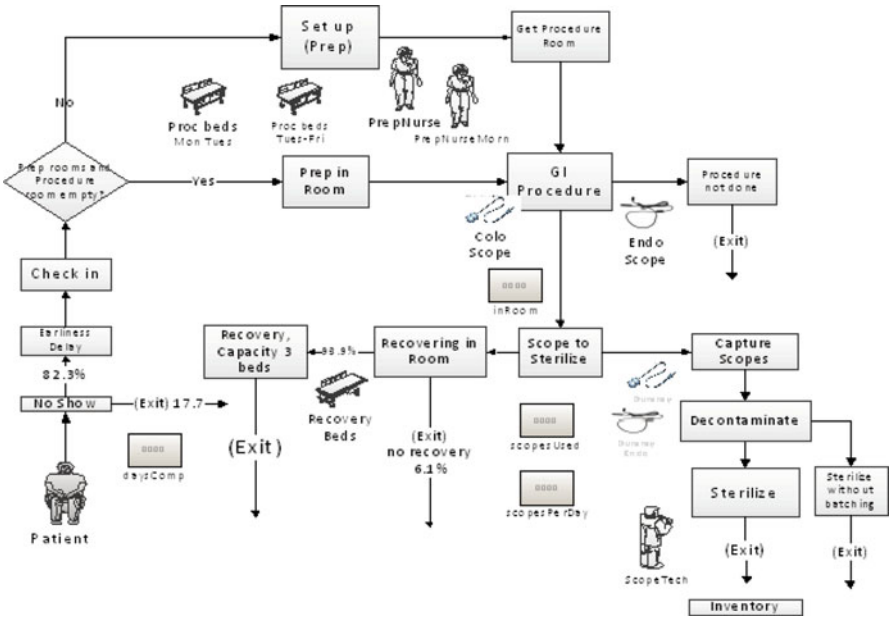


Fig. 8.1 Process model of the SFGH endoscopy center developed for discrete simulation modeling

8.2.2 Discrete Simulation Modeling

The discrete event simulation model was programmed using Process Simulator (Process Simulator is a Microsoft Visio add-on software from ProModel Corporation, 556 East Technology Ave., Orem) software based on the diagram shown in Fig. 8.1. The simulation included process times using probability distributions derived from clinic time observations. Patients often arrive before their appointment time and this earliness was also modeled as a probability distribution based on clinic observation. After arrival, the patient goes through a sequence of processes which, if busy, result in a queue of patients waiting. Each process is defined by a process time distribution that was determined through time measurements. The check-in, in room recovery, and recovery time process distributions were represented by a triangular distribution. The process time distributions for the pre-procedure and procedure processes followed a lognormal distribution and varied based on procedure type. The endoscope cleaning process required a discrete amount of time for decontamination and reprocessing. There was a certain probability that a patient did not show up to their scheduled appointment (i.e. no-show). There was also a probability that a patient did not require sedation and could therefore exit the system prior to being routed to the recovery room process step. All probability distributions were checked to assure a high confidence of fit between the modeled distributions and the distributions observed in the clinic.

Various scenarios were run and outcomes measured. The first outcome measured was overall time in the system spent by each patient. The percent of the patient's visit that was spent waiting was tracked to understand how much waiting was occurring due to bottlenecks. Total throughput was also tracked. Resource utilization rates including procedure room utilization, nurse utilization, and provider utilization were all calculated based on use and availability. Utilization was computed as the hours a resource was in use divided by the available hours for the resource.

Validation included the following to ensure the simulation was accurate:

- Parameters were verified by expert opinion.
- The workflow diagram's logic was verified by the providers.
- An assumptions document was developed and maintained for review by the providers during regular meetings to discuss updates to the model.
- The outcome results were verified with those experienced within the actual system.
- A separate analysis was conducted to ensure that the system was operating properly on Tuesdays, the most tightly scheduled day.
- Simulation animation was inspected by model developers and the providers at the endoscopy center to check that the patients were following the proper flow of events and queuing at various points.

All the parameters and distributions were based on historical data and Minitab was used to determine the distributions from the historical data.

After validation of the base case model was completed, several scenarios were studied. The primary scenarios included altering the patient appointment schedule from 60 to 45-, 40-, 35- and 30-min. appointment slots and assessing outcomes. For the shorter appointment slot schedules, limited resources were added to avoid extreme amounts of queuing. Also, changes to room availability through the adding of additional resources (i.e. endoscopists and nurses) were assessed.

8.2.3 Ethical Considerations

Our study was a quality improvement project and no personal health information was collected at any time. Thus formal institutional review was not required per the policy of the University of California San Francisco Committee on Human Research.

8.3 Results

8.3.1 Baseline Endoscopy Center Data

Utilizing data from the time and motion study, baseline arrival patterns as well as pre-procedure, procedure and recovery room times were determined. Patients with the latest scheduled appointment times arrived earliest; for example patients with

afternoon appointments (after 1 PM) arrived 179 min earlier for their appointment compared to patients scheduled at 8 AM who arrived 28 min earlier. Little variation was noted in the pre-procedure time regardless of the planned procedure, although EGD/colonoscopy required more time (31.2 min). Very little variation was noted with recovery room time and required 34.6 min if recovery occurred in the recovery room. Procedure time itself differed significantly depending on the type of procedure performed with EGD requiring 9.5 min, colonoscopy 28.5 min and combined procedures requiring 36.4 min. The mean number of procedures performed per week was 53.8. Patients spent 2.3 h at the endoscopy center with 22.3% of that time spent waiting.

In order to determine the optimal scenario(s) that would increase throughput, optimize utilization and minimize patient wait time a series of simulation models were run (Table 8.1). Scenarios included revising the endoscopy appointment times and weekly endoscopy schedule, increasing the number of nurses and providers, standardizing recovering room time and subsequently a combination of these scenarios.

8.3.2 Revision of Endoscopy Schedule

The first scenario examined a revised endoscopy schedule using shorter appointment times. When appointment time was decreased to 45 min (from 60 min) there was a 20.9% rise in the number of procedures performed/week with both patient time in the endoscopy center increasing to 3.2 h and percentage of time waiting rising to 41.2%. Additionally, there was a rise in overall utilization with the greatest rise noted in procedure room utilization. At shorter appointment times of 40 and 35 min the model was not sustainable without additional resources to serve patients; at these times there was queuing of patients in the pre-procedure area as the day progressed to the point where a significant number of patients would not have had their procedures performed by the end of the day. If the appointment was decreased further to 30 min the simulation was not feasible since the appointment time was nearly identical to the procedure times and a queue built up infinitely.

An additional change was to the overall weekly schedule. Given that patient's preferred earlier appointment times and the endoscopy center had been closed on Wednesdays – a half day of endoscopy appointments was moved from Friday afternoon to Wednesday morning. This change (compared to baseline) meant volume slightly increased to 55.7 procedures/week and procedure room utilization rose. Moreover, we found that when appointment times were shortened under this scenario to 45 and 40 min intervals there was a steady rise in procedures performed/week (an almost one-third increase) as well as improved procedure room and nursing utilization. However, these changes did so at a cost of increasing the number of hours a patient spent in the endoscopy center and increased patient wait time by 32.6 and 46.0% respectively when compared to baseline.

Table 8.1 Changes simulated with the endoscopy center model, mean and 95% confidence interval

	Procedures performed/ week	Patient time in endoscopy center (h)	Provider utilization (%)	Nurse utilization (%)	Procedure room utilization (%)	Wait time (%)
Baseline	53.8 (51.0–56.6)	2.3 (2.2–2.5)	24.0 (21.9–25.5)	40.0 (37.9–42.0)	49.8 (47.1–52.5)	22.3 (19.9–24.6)
Reducing endoscopy center appointment times						
45 min	68.0 (66.8–69.2)	3.2 (2.8–3.6)	28.9 (27.1–30.6)	48.0 (46.5–49.4)	55.8 (53.8–57.7)	37.9 (32.0–43.7)
Modifying endoscopy center weekly schedule[a]						
60 min	55.7 (53.8–57.6)	2.4 (2.3–2.5)	23.3 (21.3–25.4)	40.9 (39.4–42.4)	62.5 (58.4–66.5)	21.8 (18.7–24.8)
45 min	65.6 (64.5–66.7)	2.8 (2.6–2.9)	27.1 (24.8–29.4)	46.6 (45.3–47.9)	68.2 (65.9–70.5)	33.1 (29.1–37.2)
40 min	75.4 (73.6–77.3)	3.3 (3.1–3.6)	30.6 (28.2–32.9)	52.9 (50.4–55.4)	74.9 (71.7–78.2)	41.3 (39.1–43.5)
Expanding human resources						
Increase in pre-procedure nurses to 2[b]	68.4 (65.3–71.5)	2.4 (2.2–2.6)	30.3 (28.3–32.3)	65.6 (4.0–127.2)	61.7 (58.0–65.4)	22.4 (18.4–26.4)
Increase of 1 endoscopist to eachhalf-day of endoscopy						
60 min	70.0 (67.2–72.8)	2.3 (2.1–2.6)	24.9 (22.8–27.1)	169.9 (158.2–181.7)	56.2 (52.7–59.8)	21.0 (18.2–23.8)
45 min	87.6 (85.6–89.5)	3.0 (2.6–3.4)	30.2 (27.5–32.9)	211.9 (198.3–225.4)	65.5 (61.3–69.8)	34.9 (29.1–40.7)
40 min	93.9 (92.3–95.4)	3.9 (3.7–4.2)	32.8 (31.1–34.5)	230.7 (218.4–243.0)	70.8 (66.3–75.3)	49.2 (45.2–53.1)
Minimizing recovery room time						
No recovery in procedure room[b]	68.1 (65.8–70.3)	2.8 (2.4–3.2)	29.5 (26.6–32.3)	42.5 (39.9–45.0)	49.6 (46.5–52.7)	35.0 (29.0–41.0)
30 min Recovery[b]	67.6 (65.8–69.4)	2.5 (2.2–2.7)	28.7 (26.5–30.9)	38.9 (37.2–40.6)	45.3 (43.1–47.5)	35.4 (30.8–40.0)

[a] Shifting one afternoon half-day of endoscopy from Friday to a Wednesday morning session
[b] Appointment time of 45 min

8.3.3 Human Resources Expansion

The next area explored was to improve endoscopy center operational efficiency by adding human resources. The number of staff dedicated to the pre-procedure area was modeled to determine if such changes improved efficiency. The addition of 0.5 nurses to the pre-procedure area (from 1.5 to 2) resulted in no significant differences in outcomes when compared to baseline data. Yet, by increasing the number of pre-procedure nurses to 2 with an appointment time of 45 min resulted in a 21.3% increase in the number of procedures performed per week, rises in provider (20.8%), nurse (39.0%) and procedure room (19.3%) utilization, with minimal impact on patient wait time. There was no significant change in performance outcomes with more than two nurses in the pre-procedure area.

The number of providers that performed endoscopic procedures during the week was also varied; one additional endoscopist was added to each half-day of endoscopy. With appointment time held constant, procedure volume increased by 23.1% but it did so at a cost of increasing nursing utilization beyond capacity to over 100%. The results were similar if appointment times were lowered less than 60 min.

8.3.4 Minimizing Recovery Room Time

The next simulation examined minimizing patient time in the recovery room. Two simulations were tested: (1) limiting recovery room time to 30 min (minimum required by state regulations) and (2) not allowing patients to recover in a procedure room. In either simulation at 60 min appointment times there were no significant differences with respect to outcomes when compared to baseline data. But when limiting recovery room time and changing the appointment time to 45 min, procedure volume increased to 67.6 procedures/week, but wait time increased by 13.1%. Similar results occurred in the model when patients were only allowed to recover in the recovery room.

8.3.5 Simultaneous Changes Incorporated into Endoscopy Center Models

Using the insight learned from above, a number of scenarios were examined with multiple changes tested (Table 8.2). Simultaneous changes included reducing appointment time to 45 min, increasing the number of pre-procedure nurses, minimizing recovery room time and expanding the hours of the endoscope re-processor to increase equipment usage (in order to make shorter appointment times feasible). The first endoscopy center scenario (appointment time of 45 min, 2 pre-procedure nurses, recovery room time of 30 min, and extending the endoscope re-processor's

Table 8.2 Scenarios modeled and the resulting performance, mean and 95% confidence interval

Endoscopy center simulations	Procedures performed/week	Patient time in endoscopy center (h)	Provider utilization (%)	Nurse utilization (%)	Procedure room utilization (%)	Wait time (%)
Endoscopy center (baseline)	53.8 (51.0–56.6)	2.3 (2.2–2.5)	24.0 (21.9–25.5)	40.0 (37.9–42.0)	49.8 (47.1–52.5)	22.3 (19.9–24.6)
Endoscopy center 1[a]	68.6 (66.2–70.9)	1.9 (1.8–2.0)	30.1 (27.4–32.7)	69.0 (−8.5–146.5)	51.2 (47.1–55.2)	21.7 (17.6–25.9)
Endoscopy center 2[b]	78.0 (76.4–79.6)	2.1 (2.0–2.3)	34.5 (32.5–36.6)	47.4 (44.8–49.9)	56.4 (53.2–59.6)	25.2 (22.4–28.0)
Endoscopy center 3[c]	70.2 (68.7–71.8)	1.9 (1.8–2.0)	29.6 (27.1–32.2)	41.9 (39.3–44.6)	61.9 (58.8–65.1)	18.2 (15.6–20.8)
Endoscopy center 4[d]	81.8 (80.5–83.0)	2.0 (1.9–2.1)	35.1 (32.0–38.2)	49.6 (47.4–51.8)	71.0 (67.4–74.7)	22.7 (19.7–25.6)
Endoscopy center 5[e]	89.1 (87.6–90.7)	2.1 (1.9–2.2)	32.5 (29.6–35.4)	175.8 (166.4–185.2)	55.9 (52.7–59.1)	25.3 (22.7–27.9)
Endoscopy center 6[f]	101.0 (98.0–104.0)	2.5 (2.2–2.7)	36.4 (33.3–39.4)	201.9 (190.9–212.8)	62.1 (58.5–65.8)	33.1 (27.5–38.7)

[a] Endoscopy center 1: Appointment time of 45 min, 2 pre-procedure nurses, recovery of 30 min, recovery only in recovery room, and endoscope reprocessor hours extended to 4:30 PM

[b] Endoscopy center 2: Appointment time of 40 min, 2 pre-procedure nurses, recovery time of 30 min, recovery only in recovery room, and endoscope reprocessor hours extended to 5:00 PM

[c] Endoscopy center 3: Friday PM appointments moved to Wednesday AM, appointment time of 45 min, 2 pre-procedure nurses, recovery time of 30 min, recovery only in recovery room, and endoscope reprocessor hours extended to 4:30 PM

[d] Endoscopy center 4: Friday PM appointments moved to Wednesday AM, appointment time of 40 min, 2 pre-procedure nurses, recovery time of 30 min, recovery only in recovery room,and endoscope reprocessor hours extended to 4:30 PM

[e] Endoscopy center 5: One additional endoscopist added to each half-day of endoscopy, appointment time of 45 min, 2 pre-procedure nurses, recovery time of 30 min, recovery only in recovery room, and endoscope reprocessor hours extended to 4:30 PM

[f] Endoscopy center 6: One additional endoscopist added to each half-day of endoscopy, appointment time of 40 min, 2 pre-procedure nurses, recovery time of 30 min, recovery only in recovery room, and endoscope reprocessor hours extended to 4:30 PM

day by 30 min) resulted in a 21.6% increase in procedures performed per week, 17.4%, drop in the patient's time in the endoscopy center, and no significant change in patient wait time. A second endoscopy center scenario incorporated the same changes except appointment time was lowered to 40 min and endoscope reprocessing hours were extended by 1 h. There was a steep rise in procedure volume, further reduction of patient time in the center and wait times remained unchanged.

The above changes were then combined with a half-day of Friday afternoon endoscopy appointments moved to Wednesday morning. In this scenario the number of procedures performed rose significantly in conjunction with provider, nursing and procedure room utilization improving, and patient time in the endoscopy center decreasing. For example, in simulations at 45 and 40 min appointment times' procedure volume rose by 23.4 and 34.2%, and patient time spent in the endoscopy center declined by 17.4 and 13.0% respectively.

Finally, the same changes were also incorporated into a scenario whereby one additional provider was made available on each half day of endoscopy. Again, procedure volume markedly increased by 39.6 and 46.7% for simulations at 45 and 40 min appointment times with overall provider utilization increasing to its highest levels. However, nursing utilization exceeded capacity in both of these simulations. Furthermore, as appointment times were shortened patient wait time steadily increased to where patients spent nearly a third of their time in the endoscopy center waiting.

8.4 Discussion

Through observation and a time and motion study we found that a large, diverse safety net hospital endoscopy center has weekly operational patterns, although variable, that are consistent and predictable. Our simulation model provides insight into operational changes that are beneficial. We found that patient throughput as well as provider and nursing utilization are increased with only simple changes such as decreasing endoscopy appointment times (to a point), realigning the endoscopy schedule with patient preferences and minimizing the recovery room and pre-procedure processes. Additional improvements in throughput are possible but only with adding costly human resources, over utilizing nurses and having unacceptable wait times.

Our study is not the first to conduct a time and motion study or employ simulation modeling in the endoscopy center; however there is sparse and disparate literature on this topic. Some studies have used only a qualitative approach [15], conducted solely a time and motion study [5], incorporated only one endoscopic procedure in their models [1] or limited their simulations to just one component of the endoscopy center process [6]. In addition, these studies are limited by their setting in that all of them examined large tertiary hospitals or a private setting or included endoscopy centers outside of the U.S. Our study has strengths compared to the

available literature in that we examined multiple processes and procedures in the endoscopy center, utilized multiple scenarios that quantitatively studied their impact on a number of critical outcomes in an endoscopy center and we are the first to use such methods to examine efficiency and change in a large safety net hospital system. We involved all clinical staff in developing and testing changes while other studies generally utilized GI data and worked on it separately from providers and staff.

Similar to other studies, our study highlights the importance of two key areas in the endoscopy center: pre-procedure and recovery room processes. With respect to the pre-procedure process, no clear evidence exists on how to improve this process with only scant expert opinions available [8, 12]. A number of factors influence this process including obtaining vitals, placing intravenous catheters, completing paperwork, patient changing, and in some cases the use of interpreting services. The majority of these tasks center on nursing/medical assistant roles [5] and in most cases these tasks are fixed and difficult to streamline. Previous work in the operating room has realized this challenge and some work has demonstrated that parallel processing of tasks among staff members can lead to a dramatic reduction in operating room pre-procedure and room turnover time [2, 10]. In this same light, we modeled an increase in the pre-procedure personnel in order to utilize this strategy of parallel processing which to date has not been modeled in endoscopy centers. We noted an increase in procedure volume by 14.6 procedures/week (mean increase of 730 procedures/year) while at the same time significantly improving nursing, provider and room utilization and maintaining patient wait time constant. Other potential improvements in the pre-procedure process, but difficult to model, may focus on patient education for patient preparedness, prior communication with patients who do not speak English, and education programs aimed at improving the pre-procedure process for staff.

Another vital step in improving endoscopy center efficiency is the recovery room; specifically limiting recovery room time increases efficiency. Grossman modeled an ambulatory surgery center and demonstrated that recovery room time was the main bottleneck. In fact, a 50% reduction in recovery room time increased the number of patients per room per day and shortened the overall length of stay for patients. Similarly, in our study by limiting recovery to the recovery room (which reduces procedure room turnover time) and limiting recovery time to 30 min (a reduction of 13.3%) we observed an increase of 14.3 procedures/week with no harm to overall patient wait time. However, there is no clear method on how to address or improve this bottleneck. Aside from increasing the physical space of the recovery room (which is quite costly in a resource limited environment) the only specific intervention proposed to reduce this time has been sedation related. The use of Propofol or only using one sedating medication compared with two medications has been demonstrated to help not only reduce sedation time, but overall recovery time as well [9]. Further research on strategies aimed at improving the endoscopy recovery process is warranted.

Lastly, unlike previous work, our simulations/changes did not solely focus on maximizing the efficiency of endoscopists. Of the limited work on this topic,

all studies have focused on two key outcomes: increasing patient throughput and improving physician efficiency. However, only focusing on physician efficiency doesn't translate into overall efficiency for the endoscopy center. Rex et al clearly illustrated this concept by showing that increasing patients served and physician utilization did so at a cost of the endoscopy center being sub-optimized with increased patient length of stays and decreased non-physician staff utilization [14]. Our model echoed this point whereby in several scenarios adding an endoscopist did increase patient volume but did so at a detriment to overutilization of the nursing staff, increased patient time in the endoscopy center, and increased patient wait time (ranging from a third to almost half of a patient's visit). Also, adding additional endoscopists is a costly option (mean salary of $321,575/year) especially in resource limited areas such as public hospitals. On the other hand, personnel such as nursing, medical assistants, or extending endoscope re-processor's hours, which can impact processes before and after a procedure, are far less costly and in our simulations not only provided improvements in volume and provider efficiency but did so in a more balanced approach.

Our study setting occurred in a safety net hospital and may not be generalizable to other endoscopy centers. However, our model has much strength in that it demonstrates that with only small changes to resource assignments one can dramatically improve patient volume and other performance metrics and can be done so in a cost-effective manner. Also, by using time and motion studies and building a simulation model of an endoscopy center one can evaluate potential changes with a tool not currently being used in GI services. Lastly, we did not model other possible, but more complex changes, such as assessing the impact of same day bowel preparation which may increase the desirability of afternoon appointments, scheduling complex procedures at the end of the day as is done in surgery [4], scheduling a mix of procedures that vary by time throughout the day, or evaluating the impact of changes to arrival "earliness" as occurred in our patient population.

8.5 Conclusions

Through observation of the workflow and analysis of the results of a simulation model we illustrate that weekly patient flow patterns are predictable and simulation modeling provides insight into what changes are feasible and how they are beneficial to an endoscopy center. Relatively straight forward changes such as reducing appointment times, standardizing recovery room time and slightly increasing ancillary staff in the pre-procedure area significantly improves endoscopy center efficiency without substantially increasing costs nor changing procedure times. By balancing pre and post procedure capacity a continuous work flow is created and patient waiting is reduced. Thus, more patients can be seen – a critical need at safety net hospitals. More costly changes such as increasing the number of endoscopists can improve procedure volume but this may result in overutilization of other resources and increase waiting for patients. Overall, we discovered better understand patient flow responses to operational changes and to a simulation model

can be used to develop cost effective solutions. Thus, we recommend the use of this modeling tool to increase the capacity of GI patient services, particularly in a safety net setting.

Acknowledgements This work was supported by a Hearts Grant from the San Francisco General Hospital Foundation.

References

1. Berg, B., Denton, B., Nelson, H., et al.: A discrete event simulation model to evaluate operational performance of a colonoscopy suite. Med. Decis. Making. **30**(3), 380–7 (2007)
2. Friedman, D.M., Sokal, S.M., Chang, Y., Berger, D.L.: Increasing operating room efficiency through parallel processing. Ann. Surg. **243**(1), 10–14 (2006)
3. Grossman, P.LB. DR.: Where are ambulatory surgery center (ASC) bottlenecks? Use of computer simulation modeling to evaluate efficiency targets (abstract). Gastrointest. Endosc. **61**, AB 150 (2005)
4. Gul, S.: Optimization of Surgery Delivery Systems. Dissertation at University of Arizona 1–103 (2010). http://repository.asu.edu/attachments/56262/content/Gul_asu_0010E_10193.pdf. Accessed 1 Aug 2012
5. Harewood, G.C., Chrysostomou, K., Himy, N., Leong, W.L.: A "time-and-motion" study of endoscopic practice: strategies to enhance efficiency. Gastrointest. Endosc. **68**(6), 1043–50 (2008)
6. Joustra, P.E., de Wit, J., Struben, V.M., Overbeek, B.J., Fockens, P., Elkhuizen, S.G.: Reducing access times for an endoscopy department by an iterative combination of computer simulation and linear programming. Health Care Manag. Sci. **13**(1), 17–26 (2010)
7. Lang, L.: Study shows how to lower costs, waiting times for colonoscopies. Gastroenterology **137**, 1866–1866 (2009)
8. Marasco, J.A., Marasco, R.F.: Designing the ambulatory endoscopy center. Gastrointest. Endosc. Clin. N. Am. **12**(2), 185–204 (2002)
9. McQuaid, K.R., Laine, L.: A systematic review and meta-analysis of randomized, controlled trials of moderate sedation for routine endoscopic procedures. Gastrointest. Endosc. **67**(6), 910–23 (2008)
10. Olmstead, J., Coxon, P., Falcone, D., Ignas, L., Foss, P.: World-class OR turnaround times: secrets uncovered. Aorn. J. **85**(5), 942–5, 7–9 (2007)
11. Parks, J.K., Engblom, P., Hamrock, E., Satjapot, S., Levin, S.: Designed to fail: how computer simulation can detect fundamental flaws in clinic flow. J. Healthc. Manag. **56**(2), 135–44; discussion 45–6
12. Petersen, B.T.: Promoting efficiency in gastrointestinal endoscopy. Gastrointest. Endosc. Clin. N. Am. **16**(4), 671–85 (2006)
13. Pilgrim, H.TP, Chilcott, J., et al.: The costs and benefits of bowel cancer service developments using discrete event simulation. J. Oper. Res. Soc. **60**, 1305–1314 (2009)
14. Rex, D.K., Lahue, B., Dronzek, R., Lacey, M.J.: Impact of two procedure rooms per physician on productivity: computer simulation examines the impact of process change in the hospital gastroenterology department (abstract). Gastrointest. Endosc. (2005)
15. Zamir, S., Rex, D.K.: An initial investigation of efficiency in endoscopy delivery. Am. J. Gastroenterol. **97**(8), 1968–72 (2002)

Chapter 9
Analysis of Gastroenterology (GI) Clinic: A Systems Approach

Xiang Zhong, Jie Song, Jingshan Li, Susan M. Ertl, and Lauren Fiedler

Abstract This paper is devoted to the analysis of the gastroenterology (GI) clinic at the University of Wisconsin Medical Foundation (UWMF). First, the work flow at the GI clinic is studied. Then a Markov chain model is developed and then extended to non-Markovian case to evaluate patient length of stay and staff utilization. The model is validated by the data observed in the clinic. It is shown that the model can provide accurate estimation of system performance. Finally, using such a model, what-if analysis is carried out and different patient check-out processes are investigated.

9.1 Introduction

This study is conducted at a gastroenterology (GI) clinic owned and operated by University of Wisconsin Medical Foundation (UWMF) in Madison, Wisconsin. The goal of this work is to develop a quantitative model to analyze patient flow in the GI clinic, evaluate its performance, and propose recommendations for improvement. To accomplish this, a Markov chain model characterizing the work flow in the

X. Zhong • J. Li (✉)
Department of Industrial and Systems Engineering, University of Wisconsin – Madison, 1513 University Avenue, Madison, WI 53706, USA
e-mail: xzhong4@wisc.edu; jingshan@engr.wisc.edu

J. Song
Department of Industrial and Management Engineering, Peking University, Beijing 100871, China
e-mail: songjie@coe.pku.edu.cn

S.M. Ertl • L. Fiedler
University of Wisconsin Medical Foundation, Middleton, WI 53562, USA
e-mail: sue.ertl@uwmf.wisc.edu; lauren.fiedler@uwmf.wisc.edu

A. Matta et al. (eds.), *Proceedings of the International Conference on Health Care Systems Engineering*, Springer Proceedings in Mathematics & Statistics 61,
DOI 10.1007/978-3-319-01848-5_9,

existing GI clinic has been developed and validated by comparing with the results observed in the clinic. Using the justified model, what-if analysis has been carried out to evaluate the impacts of different configurations of the system.

Patient flow in hospitals and clinics has attracted substantial research effort. Most of the analytical studies use queueing theory models (see, for instance, reviews [3] and [7], and representative papers [1, 4–6]). However, many models oversimplify the flow process (such as representing all activities within a patient room by a single server) in order to make the analysis tractable. Many details may be ignored so that the sophisticated behavior of the system may not be characterized sufficiently. In addition, almost all the research on GI modeling addresses the clinic issues except that a case study at the Medical Center of University of California at San Diego is reported in [2], which introduces a simulation model of work flow in endoscopy testing procedures rather than GI clinic visits. Therefore, to study the patient flow in GI clinic, developing an efficient analytical method is of importance. Since the work flow in the GI clinic represents a typical clinic process, such a method will not only be useful for one particular division, but also can be applied to other departments or clinics. This paper is intended to contribute to this end.

The remainder of the paper is structured as follows: In Sect. 9.2, the operations in the GI clinic are described and an analytical model is formulated. Section 9.3 presents the analysis method for system performance. Section 9.4 is devoted to what-if analysis. Finally, conclusions are given in Sect. 9.5.

9.2 System Modeling

9.2.1 Work Flow Description

The current GI clinic has the following configuration: There are ten exam rooms in total and every two exam rooms are assigned to one care provider group, which consists of a clinician (physician (MD), physician assistant (PA), or nurse practitioner (NP) and one registered nurse (RN) or a medical assistant (MA)).

Within the GI clinic, patient visits primarily fall within two categories: office visit (OFV) and consult visit (CON). The OFV visit type is used for patients who have frequent visits to a GI specialist due to a chronic GI illness requiring frequent clinician care. The CON visit type is used for patients who are new to the GI service, often referred by other physicians (most frequently primary care). Consult visits are scheduled for a longer duration than office visits. The office visits and consult visits are distributed throughout the daily schedule based on demand, provider preference and office efficiency.

A typical visit to the GI clinic includes the following steps (see Fig. 9.1):

- Patient checks-in at the reception desk; the receptionist notifies the RN or MA, and the patient is seated in the waiting room.

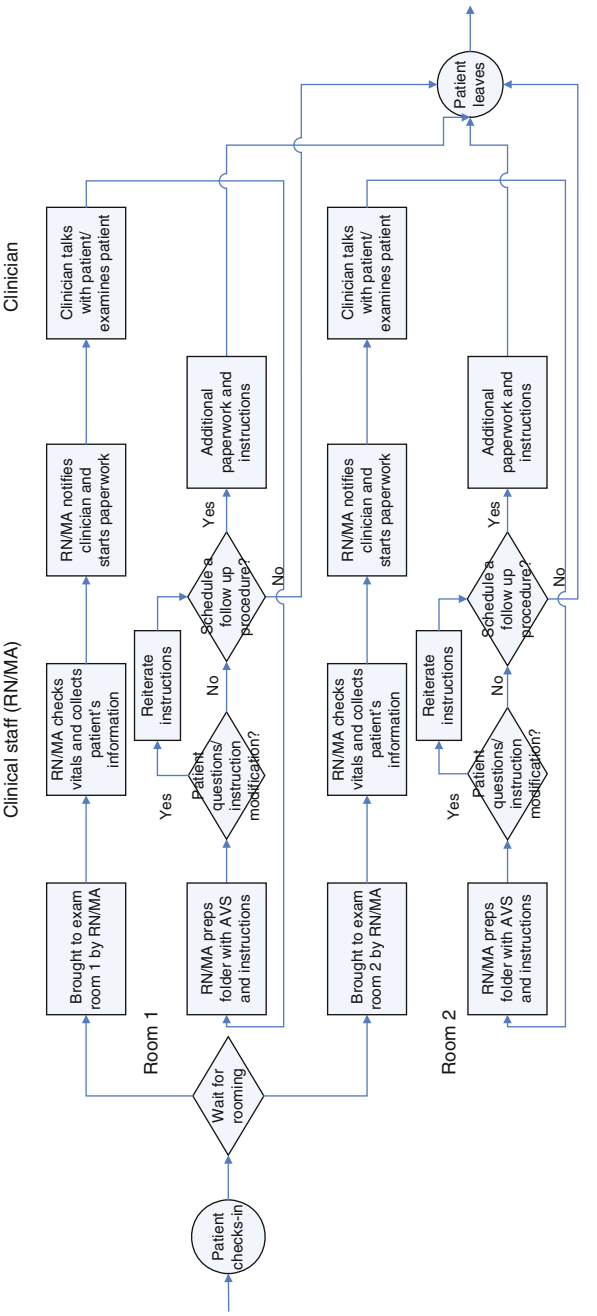


Fig. 9.1 Work flow in the GI clinic for one provider group

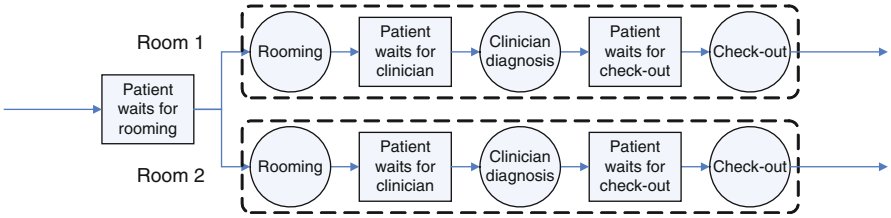


Fig. 9.2 Structural model of GI work flow

- The patient is escorted from the waiting room to the exam room by the RN or MA. The RN or MA collects basic information from the patient, obtains vitals, prepares paperwork, and records information into the electronic health record (EHR). This step is referred to as patient rooming.
- The clinician enters the exam room, assesses and diagnoses the patient condition, and develops a treatment plan.
- To discharge the patient, the RN or MA prepares the after visit summary (AVS) and follow-up instructions. The RN or MA then explains next steps and instructions to the patient and schedules any future clinic visits or procedures. The procedures typically include colonoscopy, endoscopy, MRI and CT Scan, etc. If such future procedure appointments are needed, additional documentation is required after the appointment.
- Finally, the patient leaves the clinic.

9.2.2 Structural Modeling

Since each provider group works independently, we focus our study on one provider group only. In this case, the work flow can be simplified into a serial process which includes patient waiting for rooming, patient rooming (information collection, vital check, paperwork, reporting, etc.), patient waiting for clinician, clinician examination and diagnosis, patient waiting for check-out, and check-out (including possible follow-up appointment scheduling, additional paperwork or instruction, etc.). Finally, the patient leaves the clinic. Such a work flow is illustrated in Fig. 9.2, where the circles represent the services, and the rectangles characterize patient waiting for the next service.

Within each exam room, only one patient is permitted at a time. Thus, simultaneous operations in the exam room are impossible. Moreover, since most of the patients arrive around their scheduled appointment time, the number of patients waiting for the exam rooms will be limited. Therefore, a finite capacity of waiting area can be assumed for patients waiting for rooming. Usually, assuming a capacity of 10 is more than enough.

Since the differences in service time among clinicians (such as MD, PA or NP) are small, the clinician types are not differentiated here. In this paper, MD or PA or NP are all referred to as clinicians, and RN or MA are referred to as clinical staff. To make the analysis tractable, we also group all the patient types into one, where the service time is calculated through weighted average. Such simplifications have been verified by simulation experiments, from which it is shown that the long run average length of stay based on these simplifications does not lead to much deviation.

9.2.3 Assumptions and Problem Formulation

To analyze such processes, the following assumptions are introduced to address the services, clinicians and clinical staffs (i.e., resources), and their interactions.

1. The following processes in the patient flow, from the start to the end, are labeled as 1–6, respectively: (1) patient waiting for rooming; (2) patient rooming/clinical staff visit; (3) patient waiting for clinician service; (4) clinician examination and diagnosis; (5) patient waiting for check-out; (6) clinical staff checking-out the patient.
2. The number of rooms assigned to one provider is M, where $M = 2$ in the current model. If a patient arrives while all the exam rooms are occupied, he/she needs to wait in the lobby. The maximum capacity of waiting lobby is Q. In this study, we select $Q = 10$.
3. The number of staffs is defined by $N = \{n_1, n_2\}$, representing the number of clinical staffs and clinicians, respectively. In the current setting, $n_1 = n_2 = 1$.
4. The patients arrive at the clinic based on their scheduled appointments, but with some variations. The inter-arrival time of the incoming patients follows exponential distribution with parameter λ.
5. There are three provider services in each exam room, RN/MA rooming, clinician visit, and check-out, denoted as services 1, 2, and 3, respectively. It is assumed that the corresponding services for each room are identical, and are exponentially distributed with cycle time τ_i, $i = 1,2,3$, i.e., the corresponding processing rates are $c_i = \frac{1}{\tau_i}$, $i = 1,2,3$.
6. The staff allocation for each process is denoted as θ_i, $i = 1,2,\ldots,6$. The current configuration is $\{\theta_1,\ldots,\theta_6\} = \{r_0, r_1, r_0, r_2, r_0, r_1\}$, where r_0 implies that no resource is needed, and r_1 and r_2 represent that the required resources are clinical staff and clinician, respectively.
7. Sometimes two services may require the same type of resource. In this case, priority is assigned to a later service, i.e., the check-out service has higher priority compared with the initial visit. If a patient needs to be discharged and another patient is waiting for rooming, the RN or MA will discharge the first patient and then room the next. There is no interruption of the ongoing service.

The problem to be addressed is: *Under assumptions (1)–(7), develop a method to evaluate the patient length of stay and staff utilization as functions of system parameters in the GI clinic and investigate the impacts of improvement strategies.*

9.3 Performance Analysis

9.3.1 State Space

To study this problem, a Markov chain model has been developed. Let $S = \{s_1, s_2, s_3, s_4, s_5, s_6\}$ denote the system state, where s_i represents the number of patients in stage i, $i = 1,\ldots,6$, i.e., $s_2 = m$ indicates that there are m patients in process 2 (rooming). The following constraints characterize the feasible states:

- $s_1 \leq Q$, queue length constrain,
- $s_2 + s_3 + s_4 + s_5 + s_6 \leq M$, room number constraint,
- $s_2 + s_6 \leq n_1$, clinical staff resource constraint,
- $s_4 \leq n_2$, clinician resource constraint.

In addition, for any feasible state, we have

- $s_3 > 0$ only when $s_4 = n_2$ (the clinician is busy);
- $s_5 > 0$ only when $s_2 + s_6 = n_1$ (the clinical staff is busy);
- $s_1 > 0$ only when $\sum_{i=2}^{6} s_i = M$ (all the rooms are occupied) or $s_2 + s_6 = n_1$ (the clinical staff is occupied).

Therefore, the number of feasible states, K, is reduced. In the current GI clinic, we have 79 feasible states. The steady state probability for a feasible state S_k, $k = 1,\ldots,K$, is then defined as

$$P_k = P(s_1^k, s_2^k, s_3^k, s_4^k, s_5^k, s_6^k), \quad k = 1,2,\ldots,K.$$

9.3.2 Transitions

For a feasible state S_j, there may exist a transition from another feasible state S_k to S_j triggered by one of the following events: (1) patient arrival; (2) patient rooming finishes; (3) clinician examination finishes; and (4) patient checks-out. Note that such events cannot occur simultaneously. Then, for a feasible state S_k, the rates going out of S_k and going into S_k can be written as μ_{out}^k and μ_{in}^k, respectively (the detailed derivation of them can be found in Zhong et al. (2012)). Then, the balance equations can be written as:

$$\mu_{in}^k = \mu_{out}^k, \quad k = 1,2,\ldots,K. \tag{9.1}$$

Next, introduce matrix Φ, where for $k = 1,\ldots,K$, $j = 1,\ldots,K$, $\Phi(k,j)$ defines the transition rate from state S_k to S_j, $k \neq j$. Note that $\Phi(k,j)$ can only take one of the values of c_1, c_2, c_3, or λ. Then,

$$\Phi(l,l) = -\sum_{j=1}^{K} \Phi(l,j), \quad l = 1,2,\ldots,K.$$

Thus, a transition matrix Φ with dimension $K \times K$ and rank $K-1$ is obtained. By taking the first $K-1$ columns of Φ and a normalization condition

$$\sum_{l=1}^{K} P_l = 1, \tag{9.2}$$

we construct a new matrix Γ, where

$$\begin{aligned} \Gamma(l,j) &= \Phi(l,j), \quad l = 1,\ldots,K, \quad j = 1,\ldots,K-1, \\ \Gamma(l,K) &= 1, \quad l = 1,\ldots,K. \end{aligned} \tag{9.3}$$

Then introduce vectors X and Y such that

$$\begin{aligned} X &= [P_1, P_2, \ldots, P_K], \\ Y &= [0,\ldots,0,1]. \end{aligned}$$

We obtain the balance equation as

$$X\Gamma = Y. \tag{9.4}$$

Therefore, the steady state probabilities can be obtained by solving the balance equation, i.e.,

$$X = Y\Gamma^{-1}. \tag{9.5}$$

Since we consider an irreducible Markov chain with finite number of states, there always exists a unique steady state solution.

9.3.3 Patient Length of Stay and Staff Utilization

Since the throughput and the average number of patients in the system can be evaluated by summing up all the states that the patient leaves and stays in the clinic, respectively. By Little's Law, the patient length of stay, T_s, can be obtained.

Theorem 1. *Under assumptions (1)–(7),*

$$T_s = \frac{c_3 \sum_{l=1}^{K} P_l s_6^l}{\sum_{l=1}^{K} \left(P_l \sum_{j=1}^{6} s_j^l \right)}. \tag{9.6}$$

In addition to patient length of stay, the staff utilizations ρ_i, $i = 1,2$, can be calculated as

Corollary 1. *Under assumptions (1)–(7),*

$$\rho_{clinical\ staff} = \sum_{l=1}^{K} P_l(s_2^l + s_6^l), \tag{9.7}$$

$$\rho_{clinician} = \sum_{l=1}^{K} P_l s_4^l. \tag{9.8}$$

9.3.4 Extensions to Non-exponential Scenarios

First, based on extensive numerical experiments using simulations, we verify that the patient length of stay is practically independent of the distribution type, but primarily depends on the mean and CVs of the inter-arrival time and service times, when such CVs are between 0 and 1. In practice, most of these CVs are less than 1. Next, if the scheduled inter-arrival time is long enough, and there is no variability in service time (i.e., $CV_i = 0$), then the patient length of stay can be calculated by the total service time. Thus, we define such a length of stay as

$$T_s^{fix} = \sum_{i=1}^{3} \frac{1}{c_i}. \tag{9.9}$$

Then the length of stay can be adjusted based on T_s^{fix} by the CVs of service times and inter-arrival time. Specifically, we define

$$CV_{eff} = \frac{\sum_{i=1}^{3} \frac{CV_i^2}{c_i}}{\sum_{i=1}^{3} \frac{1}{c_i}}, \tag{9.10}$$

and we hypothesize that there exists a linear relationship of lengths of stay between $CV = 0$ and 1 based on numerical investigations. In other words, we propose empirical formulas to calculate the patient length of stay in the system when both inter-arrival time and service times are non-exponential, $T_s^{non-exp}$, as follows:

$$T_s^{cv} = CV_{eff}(T_s^{exp} - T_s^{fix}) + T_s^{fix}, \tag{9.11}$$

$$T_s^{non-exp} = CV_{arrival}(T_s^{cv} - T_s^{fix}) + T_s^{fix}, \tag{9.12}$$

where $CV_{arrival}$ is the CV of patient inter-arrival time and the length of stay under exponential assumptions is denoted as T_s^{exp}.

Remark 1. Under different inter-arrival time and service time distributions, the staff utilization will be the same, since it depends on the number of patients and the average service time. Therefore, the staff utilization will not be affected by the distribution types and CVs.

Table 9.1 Model validation

LOS_{actual} (min)	$LOS_{simulation}$ (min)	LOS_{model} (min)	$\triangle_1$ (min)	$\triangle_2$ (min)	ϵ_1	ϵ_2
53.28	55.27	54.16	−0.88	1.11	−1.65 %	2.01 %

9.3.5 Model Validation

The model introduced above has been validated by comparing with the results observed in the current GI clinic. In addition, a discrete event simulation model has been developed to emulate the process. Let LOS_{actual}, $LOS_{simulation}$ and LOS_{model} denote the lengths of stay obtained by data collection, simulation, and analytical model, respectively. Introduce

$$\begin{aligned}
\triangle_1 &= LOS_{actual} - LOS_{model},\\
\triangle_2 &= LOS_{simulation} - LOS_{model},\\
\epsilon_1 &= \frac{LOS_{actual} - LOS_{model}}{LOS_{actual}} \cdot 100\,\%,\\
\epsilon_2 &= \frac{LOS_{simulation} - LOS_{model}}{LOS_{simulation}} \cdot 100\,\%,
\end{aligned}$$

to illustrate the differences of analytical result compared with the observed and simulated ones.

The results of such comparisons are shown in Table 9.1. As one can see, the differences between them are very small. Therefore, the model is validated and can be used for further analysis.

9.4 What-If Analysis

Using the validated model, what-if analysis has been carried out to investigate the impact of parameter changes. Table 9.2 summarizes all the scenarios in what-if analysis.

Note that the 50 % clinical staff availability is intended to model the scenario where one clinical staff is supporting two clinicians, so that roughly 50 % of the clinical staff's effort is devoted to each one. In this case, the model discussed above is still applicable with the modification that the rate of rooming and discharging patient should be decreased by half, i.e., $c_1' = \frac{c_1}{2}$, and $c_3' = \frac{c_3}{2}$. This implies that the patient may stay at the previous state after finishing it, due to the unavailability of the clinical staff. In addition, the last scenario is a combination of all parameter changes in scenarios 1–3. Below, the detailed results of these scenarios are introduced.

Table 9.2 What-if scenarios

Scenario	Category	Description
1	Staffing model	50 % clinical staff availability or two clinical staffs per clinician
2	Demand change	Increase demand by 10 or 30 %
3	Room configuration	One or three exam rooms per clinician
4	Service times	Change service times of clinical staff or clinician by 10 %
5	Combined scenarios	Add one room or one clinical staff and increase demand by 30 %

Table 9.3 Staffing model: clinical staff per clinician

	50 % availability			Two clinical staffs		
	From	To	Changes (%)	From	To	Changes (%)
LOS_{model} (min)	54.16	117.7	117.32	54.16	48.22	−10.9
$\rho_{clinical\ staff}$ (%)	43.95	81.14	84.61	43.95	21.99	−49.97
$\rho_{clinician}$ (%)	47.06	43.44	−7.69	47.06	47.09	0.06

9.4.1 Staffing Model

First, we investigate the impact of changes in the current staffing model. Instead of having one clinical staff to assist one clinician, we investigate the case of one clinical staff supporting two clinicians (i.e., 50 % clinical staff availability for each clinician), and the case of two clinical staffs for each clinician. The results are summarized in Table 9.3.

These results show that a clinical staff of 50 % availability is definitely not enough. However, the case of two clinical staffs for one clinician is also not necessary. Therefore, the current staffing model of one clinical staff for one clinician can well accommodate the current demand.

9.4.2 Demand Change

Next, we check the effects of patient demand change on system performance. The current inter-arrival time of 45 min is dictated by the clinic scheduling system. We test the system with the same structural model, but with decreased inter-arrival times (from 45 to 41 min, a 10 % increase in demand; and to 34.6 min, a 30 % increase in demand).

As shown in Table 9.4, if the demand is increased by 10 %, the increase in patient length of stay is 7.29 %, which is not favorable, but still can be accommodated. However, it can be found that the current GI Clinic does not have the capability to accommodate a 30 % demand surge, since the patient length of stay under

Table 9.4 Demand increase

	10 %			30 %		
	From	To	Changes (%)	From	To	Changes (%)
LOS_{model} (min)	54.16	58.11	7.29	54.16	70.32	29.85
$\rho_{clinical\ staff}$ (%)	43.95	48.14	9.53	43.95	56.06	27.56
$\rho_{clinician}$ (%)	47.06	51.54	9.52	47.06	60.02	27.54

Table 9.5 Room configuration

	One exam room			Three exam rooms		
	From	To	Changes (%)	From	To	Changes (%)
LOS_{model} (min)	54.16	81.92	51.26	54.16	51.22	−5.43
$\rho_{clinical\ staff}$ (%)	43.95	42.77	−2.68	43.95	43.99	0.09
$\rho_{clinician}$ (%)	47.06	45.79	−2.77	47.06	47.11	0.11

this setting increases substantially. In addition, both clinical staff and clinician utilizations are increased by about 30 %. Although more patients can be served, it results in excessive wait times for the patients and substantial over-time work for the providers. More capacity and resources are needed in this scenario.

9.4.3 Room Configuration

In the current clinic setting, each provider group is assigned to two rooms. Here we change the number of rooms to 1 and 3. The results are shown in Table 9.5.

It is shown that dropping one room increases patient length of stay by 51.26 %, which indicates that one room is not enough and causes a long wait for rooming. By adding one more room, the patient length of stay is decreased by 5.43 %, which is not significant. Therefore, the current setting of two rooms per provider group is reasonable.

9.4.4 Service Times

The change in service times of clinical staff and clinician are also investigated. Suppose the service times of the clinical staff and clinician can be decreased by 10 %. The resulting performance is shown in Table 9.6.

As one can see, decreasing the service time of either the clinical staff or the clinician would have similar impact on system performance, due to their similar workload in the current system setting. It is observed that the clinician and the clinical staff may ask the patient the same questions repeatedly during their visits. Therefore, if possible, improving coordination between the clinician and the clinical

Table 9.6 Service times decreased by 10 %

	Clinical staff			Clinician		
	From	To	Changes (%)	From	To	Changes (%)
LOS_{model} (min)	54.16	51.66	−4.62	54.16	51.17	−5.52
$\rho_{clinical\ staff}$ (%)	43.95	39.57	−9.97	43.95	43.98	0.07
$\rho_{clinician}$ (%)	47.06	47.08	0.04	47.06	42.38	−9.95

Table 9.7 Combination scenarios: increase demand by 30 %

	Add a room			Add a clinical staff		
	From	To	Changes (%)	From	To	Changes (%)
LOS_{model} (min)	54.16	61.265	13.12	54.16	54.02	−0.26
$\rho_{clinical\ staff}$ (%)	43.95	56.896	29.46	43.95	25.65	−41.64
$\rho_{clinician}$ (%)	47.06	60.858	29.32	47.06	61.04	29.71

staff to decrease duplicate work could help to reduce staff service time. In addition, some of the paperwork and information patients with frequent visits can be prepared prior to the visit. Thus, there exist some opportunities to reduce service time without sacrificing care quality and patient satisfaction.

9.4.5 Combined Scenarios

Finally, we study the scenario that multiple parameters may change. In this scenario, the demand is increased by 30 %, and at the same time, one more room is added, or one more clinical staff is added in the system.

When only demand is increased (see Table 9.4), a 30 % demand surge leads to a roughly 30 % increase in patient length of stay. However, such an increase will be shrunk to 13 % when an additional room is introduced (Table 9.7). Similarly, if an additional clinical staff is added, the length of stay will not increase, but decrease by 0.26 %. Therefore, additional clinical staff would be needed to accommodate the high volume of patients.

9.5 Conclusions

In this paper, an analytical model is developed to study the work flow in the gastroenterology (GI) clinic of University of Wisconsin Medical Foundation. The patient length of stay and utilization of the staff are evaluated. What-if analysis is carried out to investigate the impacts of different configurations of workforce

and resources. The results of this work could provide hospital/clinic professionals a quantitative tool to evaluate current system performance, investigate the effects of different configurations, and to predict care service efficiency for future plans.

Acknowledgements This work is supported in part by NSF Grant No. CMMI-1233807.

References

1. Bekker, R., de Bruin, A.M.: Time-dependent analysis for refused admissions in clinical wards. Ann. Oper. Res. **178**, 45–65 (2010)
2. Cono, M., Dawson, K.A.: Determining the size of the gastroenterology division expansion using simulation: a case study. In: Proceedings of the Annual HIMSS Conference, vol. 2, pp. 127–137 (1993)
3. Fomundam, S., Herrmann, J.W.: A survey of queuing theory applications in healthcare. Technical Report, University of Maryland, College Park (2007)
4. Green, L.V.: Queueing analysis in healthcare. In: Hall, R.W. (ed.) Patient Flow: Reducing Delay in Health Care Delivery. International Series in Operations Research & Management Science, vol. 91, pp. 281–307. Springer, New York (2006)
5. Green, L.V., Soares, J., Giglio, J.F., Green, R.A.: Using queueing theory to increase the effectiveness of emergency department provider staffing. Acad. Emerg. Med. **13**, 61–68 (2006)
6. Mayhew, L., Smith, D.: Using queuing theory to analyse the governments 4-h completion time target in accident and emergency departments. Health Care Manag. Sci. **11**, 11–21 (2008)
7. Wiler, J.L., Griffey, R.T., Olsen, T.: Review of modeling approaches for emergency department patient flow and crowding research. Acad. Emerg. Med. **18**, 1371–1379 (2011)

Chapter 10
Operating Room Joint Planning and Scheduling

Niccolò Bulgarini, David Di Lorenzo, Alessandro Lori, Daniela Matarrese, and Fabio Schoen

Abstract In this paper we suggest a mixed approach in which medium term planning for surgery is combined with short term scheduling of resources. Combining scheduling with planning has the beneficial effect of producing feasible schedules for the next week taking into account waiting lists. Experiments performed with real data from the Careggi Hospital in Florence support the evidence that a significant improvement of waiting list management can be obtained this way.

10.1 Introduction and Motivation

Surgery activities play a major role in hospital and clinic operations. This is due not only to the importance society gives to health care, but also to the reason that surgery generates a significant component of revenues, accounting for almost a half of hospital resource costs [21, 27]. Within surgery activities management, new technologies and advances in surgery techniques on one hand, and the aging of populations on the other makes operations handling an even more challenging topic.

Specifically, optimized surgical operations are needed to reduce costs, while ensuring high quality service to patients. This leads to the formulation and the study of new optimization problems that can be effectively solved by means of Operations Research tools and techniques. Within this context, two decision layers can be identified: (i) strategical planning must be conducted to handle relatively long-term

N. Bulgarini (✉) • D. Di Lorenzo • A. Lori • F. Schoen
Dipartimento di Ingegneria dell'Informazione, Università degli Studi di Firenze, 50139 Firenze, Italy
e-mail: niccolo.bulgarini@unifi.it

D. Matarrese
Azienda Ospedaliera Universitaria Careggi, Firenze, Italy

A. Matta et al. (eds.), *Proceedings of the International Conference on Health Care Systems Engineering*, Springer Proceedings in Mathematics & Statistics 61,
DOI 10.1007/978-3-319-01848-5_10,

resource optimization that meets deadlines and aggregated resource requirements; (ii) detailed scheduling is needed to effectively support daily operations and management decisions.

Traditionally, planning and scheduling problems have been solved using a cascading approach, due to their computational complexity. Moreover it often happens in many organizations that planning and scheduling fall under the responsibility of different managers. This separation leads to suboptimal solutions that may severely affect resources capacity. This paper investigates health care resource allocation problems using mathematical programming methods. Specifically, we propose a new joint operating room planning and scheduling formulation to show the benefits such approach gives both in terms of strategical long-term decisions and day by day resource allocations.

The proposed approach is capable of generating a suggested schedule which from one side is feasible w.r.t. operating room and surgeon availability and from the other considers the effect of current decisions on future waiting lists.

The paper is organized as follows: in Sect. 10.1.1 we review related approaches to planning and scheduling resource allocation problems. We describe the experimental settings of our work in Sect. 10.2. Section 10.3 reports the problem formulation and details the ratio behind a joint strategical and tactical approach. Section 10.4 reports computational results and a performance analysis.

10.1.1 Survey of Related Contributions

Detailed literature reviews on operating room planning and scheduling have been published by Cardoen et al. [6] and [12]. Lamiri et al. [17] solve planning problems with uncertainties by a column generation approach. Guinet et al. [13] develop a planning model and a heuristic which allocates patients to resources, taking also into account post-operation bed availabilities, to be later assigned by a scheduling procedure. Lamiri et al. [19] propose algorithms and models taking also into account emergency cases. Agnetis et al. [1] propose a long-term planning model which takes into consideration both the quality of solution and the hospital's management organizational issues. Also in [32] a similar planning model is solved by a three-phase heuristic algorithm.

Scheduling problems have been addressed in [5], where branch-and-price and column generation are used to schedule surgeries and nurse times, reducing the peak loads for the staff. In [16] a row/column generation algorithm is applied over a MILP model, and uncertainty is taken into consideration. Min et al. [24] consider the surgery scheduling problem across patients with different priorities. Santibanez et al. [30] propose a scheduling model which takes into account further real-world constraints. Some authors have developed scheduling models that also consider post-surgery constraints like bed availability (see for example Augusto et al. [2]). In [22] two MILP formulations are presented to address the problem of balancing patient queue lengths among different specialties. Ghazalbash et al. [11] propose

a scheduling model which handles many elements like equipments, post-surgery cleaning time, surgeons, nurses and specialists; since the scheduling is done day by day, the problem is simple enough to be solved exactly. Herring and Herrmann [15] present a scheduling model based on similar assumptions as our own, since they consider the more realistic situation where every day new surgeries are released; surgeries may be postponed, incurring into a penalty. Marques et al. [23] show a scheduling model that has been tested using real data from a Lisbon hospital. Similarly to our approach, the surgeries are divided in classes, some of which have hard constraints, like being scheduled in the first week.

Solving both the scheduling and planning problems together gives the best solution in terms of quality, but imposes drawbacks in terms of computational resources, so additional heuristics are usually needed. A review about integration of planning and scheduling in supply chain systems is available in [25]. Riise et al. [28] and Fei et al. [10] solve both problems using heuristics. In [29] an heuristic is used to plan and schedule surgeons, anesthetists and nurses. A genetic algorithm is proposed in [7] to solve the integrated scheduling and planning problem minimizing the "makespan". In [31] the analysis is extended to multisite, multiproduct and multipurpose batch plants using an augmented Lagrangian method. Such method is also used in [20] to solve a large scale integration problem.

In literature a distinction is usually made between *block-scheduling* and *open-scheduling* models: *block-scheduling* [1,4,5,8,22,30,33] refers to additional model constraints forcing the same specialty surgeries to be operated only in specific time blocks (e.g. cardiac surgeries take place every day from 8 am up to 2 pm, orthopedic surgeries from 2 to 8 pm, etc.). Of course this distinction is valid only if multiple specialty surgeries are taken into account.

Uncertainties in surgical times and in the set of surgeries to be performed encourage researchers to use robust optimization techniques to solve such problems. The model proposed by Denton et al. [9] also handles pre and post surgery operations under uncertainties. Lamiri et al. [18] propose a planning model and algorithm which handles emergency surgeries, which are unknown during the planning phase. In [3] and [4] a scheduling model that handles uncertainties in the surgery lengths is analyzed. In [14] and [8] several heuristics for robust scheduling are studied, and surgery durations are stochastically modeled. In a similar work [33] robustness is taken into consideration, and the portfolio effect has been considered to manage the uncertainties.

10.2 Experimental Setting

Our experiments were based on data collected during the period January 1st, 2010 to August 31st, 2010 on the surgical unit of the Neurology Department of Careggi Hospital in Florence (Italy). The ward is specialized in surgery of the skull and of the spine. The data we used was originally collected by a Master degree student [26] and successively revised by our team.

The choice of this Department was originally motivated by the small size of the unit and the interesting patient mix. The analysis was based exclusively on elective surgeries, and we adopted the Departmental policy of reserving part of one of the operating rooms' time for urgency and emergency. Two operating rooms are available in this department, with 12 h per day availability of OR 1 (from Monday to Friday) and 6.5 daily hours availability for OR 2.

We collected data on all surgeries performed in the observation period; we observed data pertaining a total of 523 surgical operations. For each of these, we recorded:

- The "release date", i.e. the time (expressed in weeks) the patient asked for surgery
- The effective date in which surgery took place (again, we recorded this data in weeks)
- The "class" of the patient. Patients in Tuscany are grouped into three classes: class A, requiring surgery not later than 30 days after arrival, class B, with a 60 days maximum allowed, and class C patients which need to be operated not later than 90 days after arrival
- The code of the main surgical operation performed
- The name of the main surgeon who operated the patient
- A Boolean field specifying whether a special purpose equipment (a brightness amplifier, in the situation analyzed) was required
- The operating room used by the patient

We choose to represent dates as weeks, in order to be prepared for the planning phase, which is based on weeks, and in order not to have too detailed data. As a consequence, we artificially associated to class A a period of 4 weeks, 8 for class B and 12 for class C. This way, we imposed stricter deadlines on the scheduling and planning phases; this was deliberate, not only in order to simplify the planning phase, but also in order to try to "squeeze" as much as possible operations in the shortest compatible period of time. During the observed period, a total 300 class A, 75 class B and 148 class C patients were operated.

The time required by each operation was obtained from the surgical documents; on average, each operation took 181.45′ (with a median of 175′, quartiles at 120′ and 240′) and a standard deviation of 84.66′. Splitting the data according to the class, a median of 180′ was observed for class A patients, while class B and C ones were characterized respectively by a median of 160′ and 175′ respectively.

In the experiments reported in this paper, we assumed that the surgeon was pre-assigned to the patient, even if the model allows for much more flexibility. We adopted a quite conservative assumption for what concerns the time availability of each surgeon: after having observed the list of surgeries performed by each of the 13 surgeons available, we decided to give each surgeon a weekly time availability equal to the effective time spent in surgery in that week. We did not impose any restriction on day by day availability, apart from that obtained by the OR availability and by the total weekly time allotted. This choice was a consequence of the real data available

for our simulations. Of course, if the day by day availability of every resource were known (as it happens here for the operating rooms), it would be very easy to include them in the proposed model, and surely the resulting schedules would turn out to be even more realistic.

For what concerns the OR's, we first listed all surgical codes and checked whether a surgery of a specific type had always been performed in OR 1 or in OR 2. From this analysis, for some types of surgery, we pre-assigned the operating room, while for others we left this choice to the scheduler/planner. This way, the OR was pre-assigned in 51 % cases, while a choice remained to be made in 49 % operations. In a future set of experiments, we will perform a similar generalization for the choice of the surgeon.

Finally, a total of 267 surgeries (51 % of the total) required a special equipment.

10.3 Joint Scheduling and Planning

In order to optimize planning and scheduling, we developed a mixed integer linear programming model which captured most of the characteristics of the situation. We choose to solve this model, which is quite of a large scale, by means of an exact algorithm (CPLEX), as the CPU time required was considered acceptable; we are currently developing and testing a fast heuristic method which will be used in place of the exact one for larger size problems, as those which we will study when exporting this first set of experiments to the whole hospital.

Our choice, in developing the model, was to build a joint model for planning and for scheduling, guided by our feeling that trying to base tactical and strategical decisions on a single model would lead to better choices, both in term of the objective function to be optimized and in terms of feasibility. In fact, from one side, scheduling without any planning consideration, which is a common practice in many organizations, has a natural tendency to "follow the emergency", overlooking the overall objective of giving appropriate treatment to all patients. A pure scheduling approach is in general too myopic and typically generates schedules which satisfy quite well the constraints for class A patients, but at the expense of very long delays on patients with less strict deadline. Of course if we could schedule for a long period, taking into account the whole current waiting list, this could lead to a very good service to patients and an optimal usage of resources; however it is well known that scheduling is a very hard computational task and when the number of "jobs" (patients) and potentially conflicting resources (operating rooms, surgeons, special equipments) becomes large, no exact method is viable and most heuristic approaches deliver a moderately good solution.

On the other hand, many methods are based on two phases: first a pure planning problem is solved, where resources are aggregated in families and conflicts are not taken into account. Secondly, given the plan for the first week, a schedule is built trying to allocate all patients according to the results of planning. This approach

has a high risk of generating either infeasible schedules, as it might be impossible to satisfy all arising resource conflicts, or it might generate a costly solution, with frequent recourse to overtime.

The approach we present in this paper is an attempt to obtain a feasible schedule for the short term, typically 1 week, and the plan of the following periods.

10.3.1 A Mathematical Programming Model

In our model we can thus distinguish two types of variables and constraints, those related to the scheduling phase and those concerning the planning one. Let $W = \{0, 1, \ldots, T\}$ be the set of planning periods (weeks) and assume that period 0 is composed of G days (usually between 5 and 7). We will use the subscript w to denote planning periods (weeks) and d to denote scheduling ones (days). A pure planning model could be represented as follows:

$$\min \quad \text{Planning obj function} \tag{10.1}$$

$$\sum_{w \in W, r \in \mathrm{O}(p)} \mathbf{x}_{p,w,r} = 1 \quad \forall p \in \mathrm{P} \tag{10.2}$$

$$\sum_{p \in \mathrm{P}: Needs(s,p)} \mathrm{Dur}_p \sum_{r \in \mathrm{O}(p)} \mathbf{x}_{p,w,r} \leq \mathrm{Avail}_{s,w} + \mathbf{sl}_{s,w} \quad \forall s \in \mathrm{S} \cup \mathrm{E}, w \in W \tag{10.3}$$

$$\sum_{p \in \mathrm{P}} \mathrm{Dur}_p \mathbf{x}_{p,w,r} \leq \mathrm{Avail}_{r,w} + \mathbf{sl}_{r,w} \quad \forall r \in \mathrm{O}(p), w \in W \tag{10.4}$$

$$\mathbf{x}_{p,w,r} \in \{0,1\} \quad p \in \mathrm{P}, w \in W, r \in \mathrm{O}(p) \tag{10.5}$$

$$\mathbf{sl}_{r,w} \geq 0 \quad r \in \mathrm{P} \cup \mathrm{E} \cup \mathrm{O}, w \in W \tag{10.6}$$

In the above model, $\mathrm{P}, \mathrm{S}, \mathrm{E}, \mathrm{O}$ are, respectively, the sets of patients, surgeons, special equipments, operating rooms; Dur_p is the surgery length for patient p and $\mathrm{Avail}_{s,w}$ is the time availability of resource s in week w. Variable $\mathbf{x}_{p,w,r}$ is a Boolean with value 1 if and only if patient p will undergo surgery in week w in the operating room r. Variable $\mathbf{sl}_{r,w}$ is a non negative real which contains the amount of extra time required in week w from resource r (overtime for surgeons, rooms, equipments). With $\mathrm{O}(p)$ we denoted the (sub-)set of OR's compatible with the surgical operation needed by patient p; the Boolean parameter $Needs(s, p)$ indicates whether patient p needs the resource (either surgeon or equipment) s.

Constraint (10.2) forces all patients to be scheduled for operations between now and the time horizon T and allocates one of the compatible OR's to the patient; constraint (10.3) limits the total resource consumption for all patients planned for a specific week which require a specific resource (either a surgeon or a special equipment). Analogously, constraint (10.4) limits the time allocated to all surgeries in a specific OR to the time available, during that week, in the specified OR.

The objective function can have many different forms, but typically includes penalties for late surgery and penalties for extra time required for some resources.

The scheduling model adds to the above schema, a set of variables and constraints needed to obtain a feasible schedule for the first week (number 0). In order to build the scheduling model, a set of additional variables are required. Let $\mathbf{y}_{p,d,r}$ be a binary variable with value 1 if and only if patient p is operated on day d (week 0) in room r; $\mathbf{S}_{p,d}$ is the start time of this operation (in minutes). δ_{p_1,p_2} is a binary variable with value 1 if patient p_1 is scheduled before patient p_2; $\mathbf{C}_{d,r}$ is a non negative variable which represents the last operating time in room r of day d, i.e., the finish time of the lat surgery performed in that day. Finally, **ScSl** is variable totally analogous to **sl**, but specific for the scheduling period. Let C be the set of potentially conflicting pairs of patients. Conflict might arise either because they share the same surgeon or the same equipment.

$$\min \quad \text{Scheduling obj function} \tag{10.7}$$

$$\sum_d \mathbf{y}_{p,d,r} = \mathbf{x}_{p,0,r} \quad \forall p \in \mathrm{P}, r \in \mathrm{O}(p) \tag{10.8}$$

$$\sum_{r \in \mathrm{O}(p_1)} \mathbf{y}_{p_1,d,r} + \sum_{r \in \mathrm{O}(p_2)} \mathbf{y}_{p_2,d,r} \leq 1 + \delta_{p_1,p_2} + \delta_{p_2,p_1} \quad \forall d, (p_1,p_2) \in C \tag{10.9}$$

$$\mathbf{y}_{p_1,d,r} + \mathbf{y}_{p_2,d,r} \leq \delta_{p_1,p_2} + \delta_{p_2,p_1} + 1 \quad \forall d, p_1, p_2 \in \mathrm{P}, r \in \mathrm{O}(p_1) \cap \mathrm{O}(p_2) \tag{10.10}$$

$$\mathbf{S}_{p_2,d} \geq \mathbf{S}_{p_1,d} + \mathrm{Dur}_p \delta_{p_1,p_2} - M(1 - \delta_{p_1,p_2}) \quad \forall d, (p_1,p_2) \in C \tag{10.11}$$

$$\mathbf{C}_{d,r} \geq \mathbf{S}_{p,d} + \mathrm{Dur}_p \mathbf{y}_{p,d,r} - M(1 - \mathbf{y}_{p,d,r}) \quad \forall d, p \in \mathrm{P}, r \in \mathrm{O}(p) \tag{10.12}$$

$$\mathbf{C}_{d,r} \leq \mathrm{Avail}_{r,d} + \mathbf{ScSl}_{d,r} \quad \forall d, r \in \mathrm{O}(p) \tag{10.13}$$

In this model, constraint (10.8) force the linkage between the planning variables **x** of week 0 with the scheduling variable **y** – in fact these constraints impose that, if and only if planning prescribes operation in week 0, one and only one of the schedule variables associated to the patient will assume value 1. Constraint (10.9) requires that if two patients are possibly in conflict one another, then either they undergo surgery in different days (or weeks) or one of the two must precede the other; constraint (10.10) is exactly the same as the previous one, for each operating room. Equation (10.11) is a conflict resolving constraint, stating that if patient p_1 precedes patient p_2, than the start time of p_2 should follow the finish time of p_1; M is a "sufficiently large" constant (e.g., the time availability on day d). The last two constraint respectively define the maximum completion time on each room for each day and the extra time required for each resource in each day.

The objective function we used in our model had the following form:

$$\sum_{w\in W, p\in \mathrm{P}, r\in \mathrm{O}} \mathrm{Penalty}(w-\mathrm{Release}_p, \mathrm{Class}_p)\mathbf{x}_{p,w,r} + \tag{10.14}$$

$$K \sum_{\substack{r\in \mathrm{S}\cup\mathrm{O}\cup\mathrm{E} \\ w\in\{1,\ldots,T\}}} \mathbf{sl}_{r,w} + K \sum_{d,r\in \mathrm{O}} \mathbf{ScSl}_{d,r} \tag{10.15}$$

where (10.14) is a penalty term depending on the delay between the time of arrival of the patient and the time the operation is scheduled, (10.15) are two penalties associated with extra time required during the planning and scheduling periods for all resources.

More specifically, for a patient of class Class_p arrived in week v and planned for operation in week w, whose class prescribed an operation within Δ weeks, the penalty is defined as

$$\mathrm{Penalty}(w-v, \mathrm{Class}_p) = \begin{cases} \alpha(w-v) & \text{if } w-v \le \Delta \\ \beta(w-(v+\Delta))^2 & \text{otherwise} \end{cases}$$

with $\alpha = 0.1$ and $\beta = 100$ for the three classes of patients. This way, a strong penalty is given if the maximum allowed time is exceeded; however, a slight penalty is imposed even for on-time schedules, in order to prefer earlier operation dates in any case. These constants can be changed in order to augment or to diminish these effects, resulting in possibly different schedules.

The remaining parts of the objective functions contain costs associated to the violation of constraints on resource time availability. In our experiments we tried different values of K in order to analyze the trade-off between scheduling/planning and resource consumption. In the experiments reported in this paper, we used $K = 500$.

Before concluding this section, we would like to recall a particularly important point: the aim of this model is to find a good schedule for the first week, taking into account the effect that this week's decisions might have on the following week plans. However, when applied in practice, during week 0 new patients will arrive which were not considered; thus the whole model should be repeated, in week 1, with a dataset consisting of all patients, except those scheduled in week 0 and including the new patients arrived during week 0. In our experiments we adopted this "rolling horizon" view and performed several optimization runs in order to obtain a complete schedule for a sufficiently long period of time.

10.4 Experimental Results

We performed several experiments with the available data in order to find a good combination of the parameters used in the model definition. We observed a significant robustness in the choice of the penalty parameters, as quite similar results are obtained even when significantly varying their orders of magnitude.

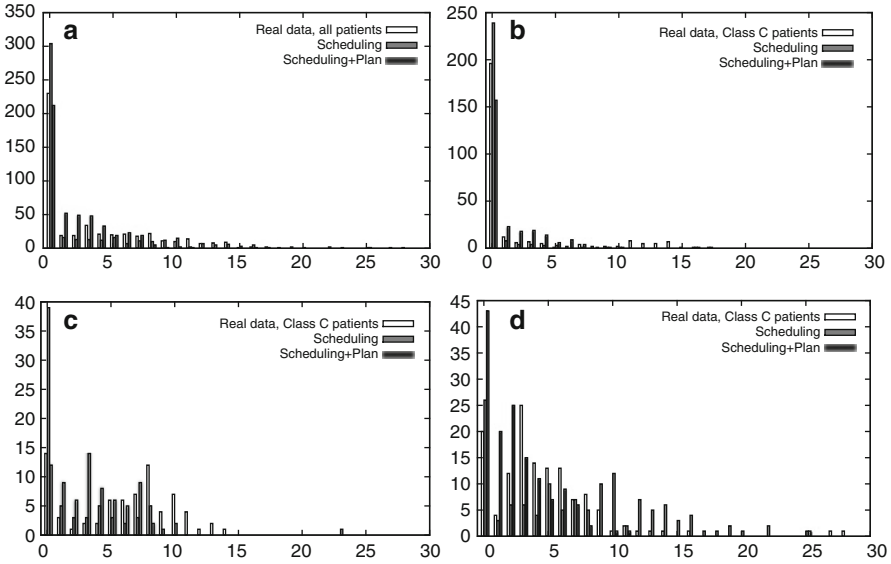


Fig. 10.1 Delay statistics – number of patients vs delay (in weeks). (**a**) Delay statistics for the whole dataset. (**b**) Delay statistics for class A patients. (**c**) Delay statistics for class B patients. (**d**) Delay statistics for class C patients

In Fig. 10.1a–d we report a comparison between the "as-is" situation and the results of the optimized schedules. In each of the figures on the horizontal axis we report the delay (in weeks) with respect to the planned due date; in the vertical axis we report the number of patients which, in the real situation and in the optimized ones, were scheduled for surgery with a specific delay. The first of the figures reports the histograms relative to the whole data set (with the exception of the last 4 weeks which were not included in the statistics in order to eliminate some tail effect), while the other three figures report similar plots for class A, B and C patients respectively.

In all figures we compare the results of the real schedules used in the hospital with those obtained with our joint planning and scheduling algorithm as well as the results obtained by means of a myopic scheduling strategy, obtained by simply scheduling in the best possible way only a single week at a time, without any consideration on the effect that this policy has on the planning of the following weeks. It is quite immediate to see that both optimized schedule are more beneficial than the actual system in use, as they strongly reduce the vast majority of delays.

Some delays reported (and some which are not reported, as they fall outside the displayed range) are still quite large, but they are the consequence of the starting state of the system, in which many patients, who already had accumulated very large delays, where in the waiting list.

When comparing the myopic scheduling policy with the optimized planning & scheduling one, the interpretation of the results require some interpretation. If we look the overall picture, we see that with the pure scheduling approach a large

number of additional patients can be scheduled for operation with no delay. The following table reports the average delays (in weeks) obtained during the simulation and compared with the real data. The results in the table are a consequence of the fact

Average delays	Class A	Class B	Class C	Total
Real	1.97	6.20	5.21	3.60
Pure scheduling	0.43	2.50	8.40	2.98
Planning+Scheduling	1.35	3.30	3.06	2.16

that, in giving weights to different delays in the objective function, we chose to use the same penalties for all patients. Having the freedom to change these weights gives the planner a lot of flexibility and, for example, augmenting the penalty for class A patients will produce schedules in which more class A patients will be scheduled on time, while not deteriorating the service provided to other patients in a significant way.

This data confirms the fact that following the most urgent cases has a beneficial effect on these ones at the expense of a severe worsening in the quality of service for other patients, while taking into account planning improves the overall performance while not deteriorating significantly the performance on class A patients.

The analysis of different choices of the weight on the resulting schedules is outside the scope of this paper and will be the subject of a future publication.

Concerning resource consumption, even with all the restrictions already described in the experimental settings, the results here obtained are based on quite small constraint violations (i.e., with very small recourse to extra time). In particular, the extra time observed in our simulation were only due to some extra time requested to a few surgeons in just 6 days (in the 30 week scheduling horizon), ranging from a 5 min extra time request to a maximum of 65 min. In the myopic scheduling simulation, no extra time was required in the 30 weeks.

10.5 Conclusions

These were the first experiments performed in order to check whether the idea of scheduling in the short time without disregarding the medium term consequences was a practical and useful one. Although the situation analyzed is relatively simple and the amount of scenarios analyzed quite limited, we are quite confident that the approach has been proven to be useful. The capacity of scheduling with a planning objective included into consideration helps in balancing the workload of the surgery, anticipating non urgent surgeries in order to have more flexibility on the future usage of precious resources. While the idea of “looking far” in the planning horizon is not new, our proposal for a rolling horizon approach which takes into account scheduling, planning as well as resource overuse is new in the surgical operation literature. The model we used was an exact one and, as the number of

OR's and patients increases, it is sure that the time required to solve it, even within a reasonable tolerance with respect to the optimum, will increase exponentially. Thus it will become necessary to switch to an heuristic approach in order to be able to solve real size scheduling and planning problems. However, the aim of this paper was that of providing confidence on the correctness of the approach. We think that this objective has been obtained and thus it is now reasonable to invest in the implementation of a fast heuristic to be applied to large scale instances.

References

1. Agnetis, A., Coppi, A., Corsini, M., Dellino, G., Meloni, C., Pranzo, M.: Long term evaluation of operating theater planning policies. Oper. Res. Health Care **1**, 95–104 (2012)
2. Augusto, V., Xie, X., Perdomo, V.: Operating theatre scheduling with patient recovery in both operating rooms and recovery beds. Comput. Ind. Eng. **58**, 231–238 (2010)
3. Batun, S., Denton, B.T., Fitts, E.P., Huschka, T.R.: The benefit of pooling operating rooms and parallel surgery processing under uncertainty. Technical Report, University of Pittsburgh, 2010
4. Beliën, J., Demeulemeester, E.: Building cyclic master surgery schedules with leveled resulting bed occupancy. Eur. J. Oper. Res. **176**, 1185–1204 (2007)
5. Beliën, J., Demeulemeester, E.: A branch-and-price approach for integrating nurse and surgery scheduling. Eur. J. Oper. Res. **189**, 652–668 (2008)
6. Cardoen, B., Demeulemeester, E., Beliën, J.: Operating room planning and scheduling: a literature review. Eur. J. Oper. Res. **201**, 921–932 (2010)
7. Choi, H.R.I.M., Park, B.J.O.O.: Integration of process planning and job shop scheduling using genetic algorithm. In: Proceedings of the 6th WSEAS International Conference on Simulation, Modelling and Optimization, Lisbon, Portugal, pp. 13–18 (2006)
8. Denton, B., Viapiano, J., Vogl, A.: Optimization of surgery sequencing and scheduling decisions under uncertainty. Health Care Manag. Sci. **10**, 13–24 (2007)
9. Denton, B.T., Miller, A.J., Balasubramanian, H.J., Huschka, T.R.: Optimal allocation of surgery blocks to operating rooms under uncertainty. Oper. Res. **58**, 802–816 (2010)
10. Fei, H., Meskens, N., Chu, C.: A planning and scheduling problem for an operating theatre using an open scheduling strategy. Comput. Ind. Eng. **58**, 221–230 (2010)
11. Ghazalbash, S., Sepehri, M.M., Shadpour, P., Atighehchian, A.: Operating room scheduling in teaching hospitals. Adv. Oper. Res. **2012**, 1–16 (2012)
12. Guerriero, F., Guido, R.: Operational research in the management of the operating theatre: a survey. Health Care Manag. Sci. **14**, 89–114 (2011)
13. Guinet, A., Chaabane, S.: Operating theatre planning. Int. J. Prod. Econ. **85**, 69–81 (2003)
14. Hans, E., Wullink, G., van Houdenhoven, M., Kazemier, G.: Robust surgery loading. Eur. J. Oper. Res. **185**, 1038–1050 (2008)
15. Herring, W., Herrmann, J.: The single-day surgery scheduling problem: sequential decision-making and threshold-based heuristics. OR Spectr. **34**, 429–459 (2012)
16. Holte, M., Mannino, C.: The implementor/adversary algorithm for the cyclic and robust scheduling problem in health-care. Eur. J. Oper. Res. (2012)
17. Lamiri, M., Dreo, J., Xie, X.: Operating room planning with random surgery times. In: IEEE International Conference on Automation Science and Engineering (CASE 2007), Scottsdale, pp. 521–526 (2007)
18. Lamiri, M., Xie, X., Dolgui, A., Grimaud, F.: A stochastic model for operating room planning with elective and emergency demand for surgery. Eur. J. Oper. Res. **185**, 1026–1037 (2008)
19. Lamiri, M., Grimaud, F., Xie, X.: Optimization methods for a stochastic surgery planning problem. Int. J. Prod. Econ. **120**, 400–410 (2009)

20. Li, Z., Ierapetritou, M.G.: Production planning and scheduling integration through augmented Lagrangian optimization. Comput. Chem. Eng. **34**, 996–1006 (2010)
21. Macario, A., Vitez, T., Dunn, B., McDonald, T.: Where are the costs in perioperative care? Analysis of hospital costs and charges for inpatient surgical care. Anesthesiology **83**, 1138–1144 (1995)
22. Mannino, C., Nilssen, E.J., Nordlander, T.E.: A pattern based, robust approach to cyclic master surgery scheduling. J. Sched. **15**, 553–563 (2012)
23. Marques, I., Captivo, M., Vaz Pato, M.: An integer programming approach to elective surgery scheduling. OR Spectr. **34**, 407–427 (2012)
24. Min, D., Yih, Y.: An elective surgery scheduling problem considering patient priority. Comput. Oper. Res. **37**, 1091–1099 (2010)
25. Nikolopoulou, A., Ierapetritou, M.G.: Integration of operation planning and scheduling in supply chain systems: Scheduling Problems and Solutions, Nova Science Pub Inc. 1–20 (2011)
26. Nofri, C.: Modelli di ottimizzazione per la pianificazione di interventi chirurgici. Master's thesis, Ingegneria Gestionale, Università degli Studi di Firenze, Facoltà di Ingegneria (2011)
27. Pham, D.N., Klinkert, A.: Surgical case scheduling as a generalized job shop scheduling problem. Eur. J. Oper. Res. **185**, 1011–1025 (2008)
28. Riise, A., Burke, E.K.: Local search for the surgery admission planning problem. J. Heuristics **17**, 389–414 (2011)
29. Roland, B., Martinelly, C.D., Riane, F., Pochet, Y.: Scheduling an operating theatre under human resource constraints. Comput. Ind. Eng. **58**, 212–220 (2010)
30. Santibáñez, P., Begen, M., Atkins, D.: Surgical block scheduling in a system of hospitals: an application to resource and wait list management in a british columbia health authority. Health Care Manag. Sci. **10**, 269–282 (2007)
31. Shah, N.K., Ierapetritou, M.G.: Integrated production planning and scheduling optimization of multisite, multiproduct process industry. Comput. Chem. Eng. **37**, 214–226 (2012)
32. Tànfani, E., Testi, A.: A pre-assignment heuristic algorithm for the master surgical schedule problem (MSSP). Ann. Oper. Res. **178**, 105–119 (2009)
33. Van Houdenhoven, M., van Oostrum, J.M., Hans, E.W., Wullink, G., Kazemier, G.: Improving operating room efficiency by applying bin-packing and portfolio techniques to surgical case scheduling. Anesth. Analg. **105**, 707–714 (2007)

Chapter 11
Risk-Aware Scheduling of Elective Surgeries

Gabriella Dellino, Carlo Meloni, and Marco Pranzo

Abstract This paper addresses Operating Room scheduling problems in elective surgery. In particular, we study a model for determining the surgical schedule when uncertainty on surgery duration is taken into account in order to consider and evaluate the risk of overtime and the possible waste of operating time. Surgical cases are selected from the waiting lists according to several parameters, including surgery duration, waiting time and priority class of the operations. We apply the proposed approach to the operating theatre of a public, medium-size hospital in Italy, using Mathematical Programming formulations and Monte Carlo simulations, assuring the scalability of the approach on larger hospitals.

11.1 Introduction

The operating theater (OT), consisting of several operating rooms (ORs), is one of the most critical resources in a hospital because it has a strong impact on the quality of health service and represents one of the main sources of costs (surgical teams, equipment etc.). Given the patients' waiting list and various information on OT characteristics and status, OT planning problems consist in deciding the schedule

G. Dellino (✉)
IMT Institute for Advanced Studies, Piazza San Ponziano, 6, 55100 Lucca, Italy
e-mail: g.dellino@imtlucca.it

C. Meloni
Politecnico di Bari, Via E. Orabona, 4, 70125 Bari, Italy
e-mail: meloni@poliba.it

M. Pranzo
Università degli Studi di Siena, Via Roma, 56, 53100 Siena, Italy
e-mail: pranzo@dii.unisi.it

A. Matta et al. (eds.), *Proceedings of the International Conference on Health Care Systems Engineering*, Springer Proceedings in Mathematics & Statistics 61,
DOI 10.1007/978-3-319-01848-5_11,

of surgeries in a given time horizon, with the aim of optimizing several performance measures such as OR utilization, throughput, surgeons' overtime, lateness etc. [2,7–9,12].

Surgical cases are usually carried out in OR sessions, i.e., uninterrupted time blocks (typically, half day or a full day). In the management policy usually referred to as *block scheduling*, each OR session is devoted to a specific surgical discipline. This organizational solution is often preferred, since performing the same discipline in a given room during a given time span typically simplifies the physical handling of equipment and/or materials. A more flexible solution is the *open scheduling* policy [3], in which no pre-specified session-to-discipline assignment exists, so two cases corresponding to different disciplines can be scheduled in the same OR session. This paper focuses on the block scheduling policy. Thus, surgical planning in operating theaters can be seen as involving three distinct decision steps:

(i) Deciding the surgical discipline that will be performed in each OR session;
(ii) Selecting elective surgeries to be performed in each OR session;
(iii) Sequencing surgeries within each OR session.

Problem (i) is often referred to as the Master Surgical Scheduling Problem (MSSP), and returns the Master Surgical Schedule (MSS). Problem (ii) determines the Surgical Case Assignment (SCA), and is therefore denoted as Surgical Case Assignment Problem (SCAP). Problem (iii) outputs the detailed calendar of elective surgeries for each session. Literature on all three above decision levels is wide and growing, and it has thoroughly been reviewed by several researchers [1,15,16].

The three above decision problems have been addressed by a multiplicity of approaches. Research focused on either all three levels concurrently, two of them, or even single problems. Some approaches have been designed to fit specific issues that may or may not be present in various real-life settings. In this paper we focus on SCAP considering uncertain durations of surgeries. Our assumptions are similar to those proposed by Agnetis et al. [1], who design a deterministic model for MSSP and SCAP, on the basis of the current state of the waiting list. Their approach consists in concurrently defining the MSS and the list of surgical cases to be performed during each OR session over the planning horizon, whereas we consider the MSS as given and focus our analysis on the SCAP.

The main contribution of this paper is to assess the risks of overtime and possible waste of operating time in each operating session associated with an OR plan obtained through a deterministic optimisation model. This analysis allows to evaluate the impact of uncertainties in the surgical times when the solution of the deterministic model is implemented. The role played by the variability in surgery times in creating delays, resources waste and non-compliance in health care is documented in the literature, but is often ignored in OR planning and scheduling systems. Completely ignoring this kind of variability, i.e., using the basic assumption of deterministic times, could be unrealistic and rather optimistic from the start, generating schedules that promise more than can be delivered to both customers and managers of the health care system [5,6].

The analysis conducted in this study yields two main advantages: it evaluates the risk associated with a specific OR plan, and gives sufficient information to take suitable decisions in the operational plan to limit risks and/or reduce costs; e.g., using overtime or processing additional case surgeries. Based on these considerations, the output of this analysis provides a more realistic view of a specific OR plan performance. Moreover, it could suggest and justify some improvements in the planning system to take into account the risks associated to the times variability in the assignment process.

The remaining sections are organized as follows: Sect. 11.2 introduces the addressed problem; Sect. 11.3 describes the setting adopted in our computational experiments and reports on the results obtained. Section 11.4 draws some conclusions, outlining perspectives for future research.

11.2 Problem Description

The aim of this work is to design a decision support tool for the risk assessment in elective surgical planning. We assume that the MSS is given. The OT management provided us with the MSS currently adopted by the hospital; therefore, only the SCAP has to be solved. Note that, once a MSS is determined, it identifies which OR sessions are assigned to each surgical discipline; so, a distinct SCAP can be solved for each discipline independently from the others.

Following [1], we associate a score K_{is} to each surgical case i of discipline s, defined as $K_{is} = P_{is}(W - R_{is})$, where P_{is} denotes the *nominal* surgery duration, W corresponds to the maximum allowed waiting time for the least urgent surgeries (as prescribed by regional regulations), and R_{is} is the *slack time*, i.e., days to the due date. To assign elective surgeries to OR sessions, we maximize the score associated to the selected surgeries, accounting for their priority class as well as for their duration. Let Q_s be the number of OR sessions assigned to surgical discipline s by the actual MSS, and T_{hs} the duration of the h-th OR session of discipline s, $h = 1, \ldots, Q_s$. We introduce the binary decision variables x_{ish} such that $x_{ish} = 1$ if the i-th surgery of discipline s is assigned to the h-th OR session of discipline s, otherwise $x_{ish} = 0$. Then, the optimization problem can be formulated as follows:

$$\max \sum_s \sum_h \sum_i K_{is} \cdot x_{ish} \tag{11.1}$$

$$\sum_h x_{ish} \leq 1 \qquad \forall i, s \tag{11.2}$$

$$\sum_i P_{is} \cdot x_{ish} \leq T_{hs} \qquad \forall s, h \tag{11.3}$$

$$x_{ish} \in \{0, 1\} \qquad \forall i, s, h \tag{11.4}$$

where constraint (11.2) guarantees that each surgery is performed at most once, while constraint (11.3) sets a maximum duration for the surgical cases assigned to the same OR session.

When addressing this problem, the surgery duration is commonly supposed to be deterministic and known in advance, based on estimates provided by surgeons: in fact, for each surgical case in the waiting list, a nominal (i.e., fixed) duration is specified (P_{is}). However, the whole surgical process is affected by uncertainty, so different surgical durations can be observed in practice. This may result in OR underutilization, if surgeries last less than expected, or overtime, if the actual surgery duration is higher than planned; both cases lead to inefficiencies in the OT management, and may significantly affect the patients' service level. Therefore, it is important to evaluate the risk associated to a deterministic planning. To this aim, we propose a statistical analysis on the surgical records for each discipline. More specifically, for each type of surgery appearing in the surgical records—coded according to the classification of surgical procedures International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM, [11])—we collect the actual duration of each surgical case over a period of 1 year; then, we fit a probability distribution on these data and estimate its parameters.

In this way, for each surgical case, we associate a probability distribution to its surgical duration. Then, we run Montecarlo simulations to derive several realizations of surgical durations for a given surgery. We now plug these alternative observations into the solution of the deterministic formulation (11.1)–(11.4) for SCAP, to evaluate the possible variation of the makespan related to each filled OR session.

Evaluating the solution obtained by the deterministic planner on a set of alternative scenarios enable us to assess the related risk of overtime/underutilization of each OR session. More specifically, once the SCA has been generated, this solution is implemented in a number replicates of the instance under study taking into account the possible surgery times variability. The results of this simulation enables to conduct a probabilistic analysis to estimate the distribution of the makespan of the OR sessions associated with the specific SCA solution. This analysis allows the decision maker to evaluate specific risk measures including the probability of meeting specific targets on the SCA performance.

11.3 Computational Experiments

This section describes the computational experiments that we ran for the OT of a medium-size Italian hospital in Tuscany. We first present our experimental setting (Sect. 11.3.1); then, we discuss the results obtained.

11.3.1 Experimental Setting

The hospital's OT performs elective surgeries for the following disciplines: general surgery, paediatric surgery, otolaryngology, urology and gynaecology. For the sake of brevity, our experiment focuses on a single surgical discipline; namely, general

surgery. Nevertheless, since the SCAP can be solved separately for each discipline, our risk assessment method is equally applicable to any surgical discipline, without affecting the results obtained for the other disciplines.

Model (11.1)–(11.4) is solved using CPLEX 12.4 on a 1.8 GHz Intel Core i7 with 4 GB of RAM. Based on a preliminary experimental campaign, we truncate the solver after 5 min of computation; the optimality gap w.r.t. the best solution found is on average 0.7 %. From a computational viewpoint, we work with 1-min temporal grain and, compared to [1], we do not discretize time in 15-min time units.

The hospital provided us with the surgical records associated to 1,470 cases from general surgery performed in the last months. Based on the MSS adopted by the hospital, $Q_s = 13$ OR sessions are assigned to general surgery ($s = \text{gs}$) in 1 week. In particular, there are 10 full-day sessions, each lasting 10 h (so $T_{hs} = 600$ min, for $h = 1, \ldots, 10$), 2 morning sessions, each lasting 6 h (i.e., $T_{hs} = 360$ min, for $h = 11, 12$), and one afternoon session, lasting 4 h (i.e., $T_{hs} = 240$ min, for $h = 13$). From each OR session capacity we leave a planned buffer time for possible delays and/or uncertainties affecting surgery duration: we considered two buffer time values; namely, $0.1T_{hs}$ and $0.2T_{hs}$ ($s = \text{gs}$, $h = 1, \ldots, 13$).

Several researches in the literature suggest that surgical duration usually follows a lognormal distribution [13, 14, 17]; alternatively, it may follow a Weibull distribution [4, 5, 10]. For this reason, we selected these two families of probability distributions in our tests: for each surgery type, we fitted both distributions to our data, and identified their parameters (mean and standard deviation for the lognormal distribution, scale and shape parameters for the Weibull distribution) through Maximum Likelihood Estimation (MLE). Further, we used our data from surgical records to estimate a truncation point cutting the right tail off: in fact, using a truncated probability distribution appears reasonable, since surgical duration higher than a given threshold will (almost) never occur. Validating both models supported the hypothesis that surgical duration follows a lognormal distribution for almost all surgical classes in general surgery.

We use information collected through our statistical fitting for quantifying risk associated to the deterministic SCAP on a number of test instances. In particular, we build $M = 10$ test instances by sampling (with replacement) $N = 300$ surgeries from the historical data provided by the hospital. Each instance represents a possible realization of the waiting list that can be given as input to solve the SCAP. Once a deterministic solution is obtained, it is evaluated on stochastic surgical durations sampled from the corresponding distribution; i.e., for each surgery included in the SCAP solution, we replace its nominal duration by a set of (say) $n_t = 1{,}000$ stochastic realizations extracted from the estimated probability distribution associated to that surgical case. We repeat this procedure for each of the M test instances.

Notice that, in general, surgical duration for each class can be characterized by a different distribution. This would prevent us to identify a closed-form expression of the distribution for the proposed risk assessment procedure.

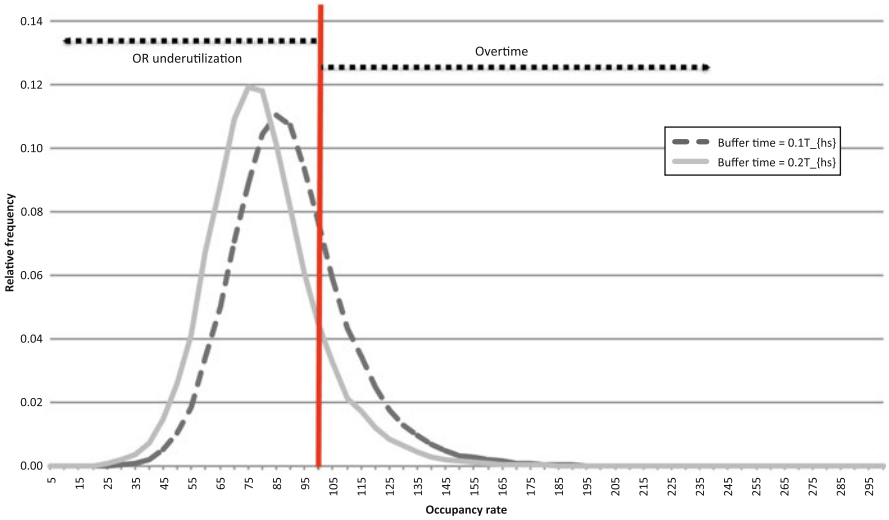


Fig. 11.1 Histogram of OR sessions occupancy rate for different buffer time values: $0.1T_{hs}$ (*dashed curve*) and $0.2T_{hs}$ (*solid curve*)

11.3.2 Results

As discussed in Sect. 11.2, we ran our Montecarlo simulations to derive n_t alternative scenarios of actual surgery duration for the SCAP deterministic solution. Based on the simulation outcomes, we measured the occupancy rate of each OR session in the M instances, w.r.t. the maximum capacity of each OR session. The results obtained are summarised in Fig. 11.1.

Figure 11.1 shows the histogram plot associated to the OR sessions occupancy rate, including both buffer time values. The vertical line at $x = 100$ corresponds to full occupancy rate; lower values imply underutilization of ORs, while higher values denote overtime. This plot provides an overview on OR utilization, including all the M instances and for all the OR sessions. This figure, reporting aggregated data, effectively describes the type of variability associated with the use of a deterministic planner when surgical times are affected by uncertainty. Note that the curve associated to a higher buffer time is shifted to the left w.r.t. the other curve, showing higher exposure to OR underutilization. On the other hand, comparing the two curves to the right of the vertical line, we note that overtime occurs more frequently when the buffer time is smaller, thus underlining that such a buffer is not big enough to deal with possible delays in surgery duration. When the risk of overtime is very high, and the expected delays are significant, the OT management may decide to postpone some surgical cases to the next working days. This would help to save additional costs to the hospital and further inconvenience to patients,

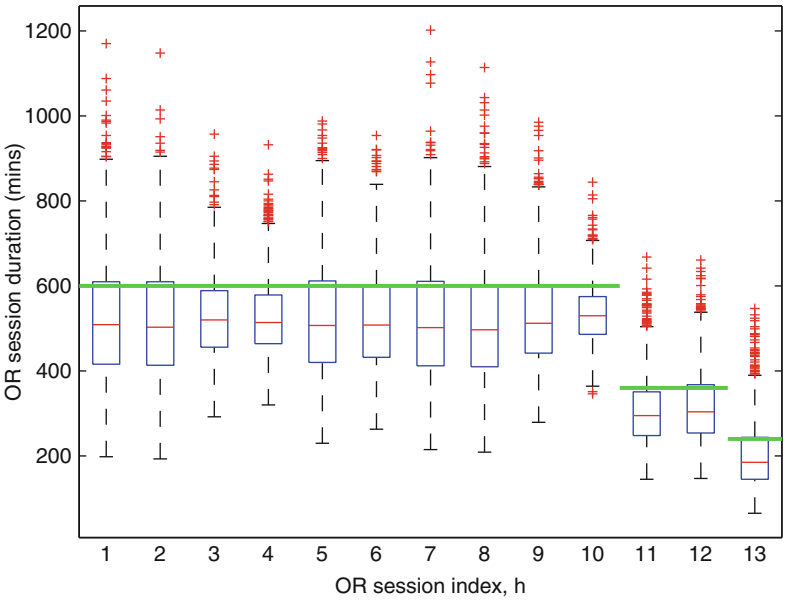


Fig. 11.2 Box plots on a sample test instance (# 1), with a buffer time of $0.1T_{hs}$

which would have resulted instead from a deterministic planning. The overall behavior described may suggest the decision maker to intervene in different ways: to adequately arrange overtime; to make ORs readily available when not in use, and to adopt techniques to reduce the variability of surgical times [5,6]. On the other hand, the scheduler may consider some changes to the models in use so that they can take into account the variability of the surgical times. However, these observations suggest a compromise between modeling improvements and the efforts dedicated to reduce that variability.

A deeper look into a single instance (say, instance #1) is provided by Figs. 11.2 and 11.3. These two figures provides box plots of surgical duration (expressed in minutes) for the Q_s OR sessions allocated to $s = \text{gs}$; the former is based on a buffer time of $0.1T_{hs}$, while the latter uses a buffer time of $0.2T_{hs}$. The tick horizontal lines denote the capacity T_{hs} of each OR session $h = 1, \ldots, 13$.

These two figures show the information that the decision maker can use to evaluate how to organize resources and activities within the surgical block. On the basis of his/her aversion to overtime, the decision maker will be willing to use a certain level of buffer time in the optimization model; this choice has direct consequences on the possibility of OR underutilization.

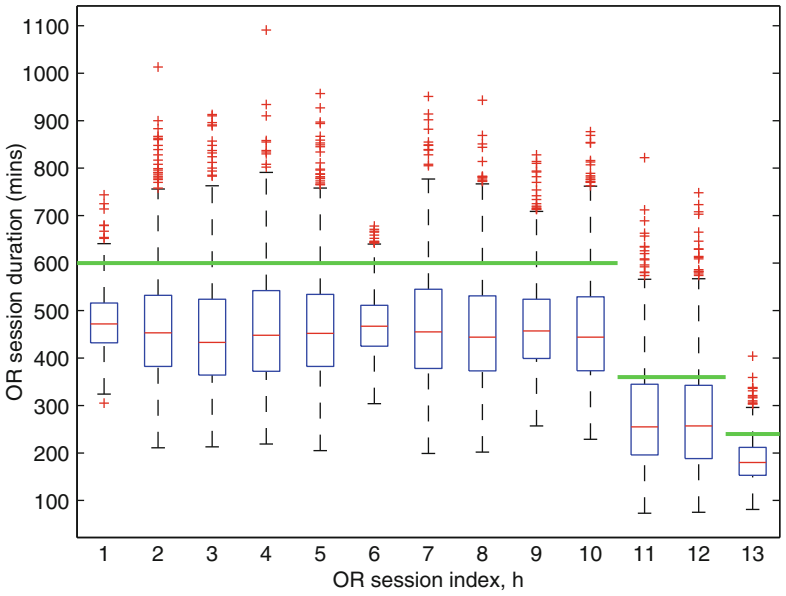


Fig. 11.3 Box plots on a sample test instance (# 1), with a buffer time of $0.2T_{hs}$

11.4 Conclusions

This paper is devoted to assess risks of overtime and possible waste of operating time associated with an OR plan obtained through a deterministic model for the SCAP. The proposed analysis allows to evaluate the impact of uncertainties in surgery times when implementing the solution of the deterministic model. A critical issue in applying our method relies on the availability of surgical records to provide accurate estimations of the statistical distributions for surgical times. Moreover, a careful preprocessing might be required, to manage possibly wrong or incomplete data records. This approach is equally scalable to larger hospitals, whose OT size may impact on the optimization methods adopted to solve the OR planning problem.

The output of the analysis suggests different actions to the decision maker, mainly related to the following issues: overtime administration, management of operating rooms become available throughout an OR session; methods to reduce surgical times variability. Further, the analysis can motivate some changes in the optimization models in order to exploit the variability of surgical times in the planning phase.

The results of this study highlight the significant impact uncertainty has on OR planning and scheduling, motivating the need for a decision support tool explicitly accounting for stochastic components affecting the planning activity. A set of easy-to-read indicators could be included to summarize the results in a compact and effective way, thus facilitating OT management decisions. Future research directions

will also cover risk assessment methods in OR planning, introducing adequate risk indicators and identifying a trade-off between modeling improvements and efforts dedicated to uncertainty management.

Acknowledgements This research is partially supported by the grant "Gestione delle risorse critiche in ambito ospedaliero" ("Critical resource management in hospitals") of the Regione Toscana—PAR FAS 2007–2013 1.1.a.3.- B51J10001140002.

References

1. Agnetis, A., Coppi, A., Corsini, M., Dellino, G., Meloni, C., Pranzo, M.: Long term evaluation of operating theater planning policies. Oper. Res. Health Care **1**(4), 95–104 (2012)
2. Cardoen, B., Demeulemeester, E., Beliën, J.: Operating room planning and scheduling: a literature review. Eur. J. Oper. Res. **201**, 921–932 (2010)
3. Chaabane, S., Meskens, N., Guinet, A., Laurent, M.: Comparison of two methods of operating theatre planning: application in Belgian hospital. J. Syst. Sci. Syst. Eng. **17**(2), 171–186 (2008)
4. Combes, C., Meskens, N., Rivat, C., Vandamme, J.-P.: Using a KDD process to forecast the duration of surgery. Int. J. Prod. Econ. **112**, 279–293 (2008)
5. Denton, B., Viapiano, J., Vogl, A.: Optimization of surgery sequencing and scheduling decisions under uncertainty. Health Care Manag. Sci. **10**(1), 13–24 (2007)
6. Dexter, F., Traub, R.D., Macario, A.: How to release allocated operating room time to increase efficiency: predicting which surgical service will have the most underutilized operating room time. Anesth. Analg. **96**(2), 507–512 (2003)
7. Guerriero, F., Guido, R.: Operational research in the management of the operating theatre: a survey. Health Care Manag. Sci. **14**(1), 89–114 (2011)
8. Guinet, A., Chaabane, S.: Operating theatre planning. Int. J. Prod. Econ. **85**, 69–81 (2003)
9. Jebali, A., Alouane, A.B.H., Ladet, P.: Operating rooms scheduling. Int. J. Prod. Econ. **99**(1–2), 52–62 (2006)
10. Li, Y., Zhang, S., Baugh, R.F., Huang, J.Z.: Predicting surgical case durations using ill-conditioned CPT code matrix. IIE Trans. **42**, 121–135 (2010)
11. National Center for Health Statistics: ICD-9-CM Addenda, conversion table, and guidelines (2011). Available on http://www.cdc.gov/nchs/icd/icd9cm_addenda_guidelines.htm
12. Sier, D., Tobin, P., McGurk, C.: Scheduling surgical procedures. J. Oper. Res. Soc. **48**, 884–891 (1997)
13. Spangler, W.E., Strum, D.P., Vargas, L.G., May, J.H.: Estimating procedure times for surgeries by determining location parameters for the lognormal model. Health Care Manag. Sci. **7**, 97–104 (2004)
14. Strum, D.P., May, J.H., Vargas, L.G.: Modeling the uncertainty of surgical procedure times: comparison of log-normal and normal models. Anesthesiology **92**(4), 1160–1167 (2000)
15. Testi, A., Tanfani, E., Torre, G.C.: A three phase approach for operating theatre schedules. Health Care Manag. Sci. **10**, 163–172 (2007)
16. Testi, A., Tanfani, E., Torre, G.C.: Tactical and operational decisions for operating room planning: efficiency and welfare implications. Health Care Manag. Sci. **12**, 363–373 (2008)
17. Wright, I.H., Kooperberg, C., Bonar, B.A., Bashein, G.: Statistical modeling to predict elective surgery time. Anesthesiology **85**(6), 1235–1245 (1996)

Chapter 12
Investigating the Relationship Between Resources Balancing and Robustness in Master Surgical Scheduling

Carlo Banditori, Paola Cappanera, and Filippo Visintin

Abstract In this paper: (i) we present a MIP model to address the Master Surgical Scheduling problem; (ii) we discuss the impact of different resources balancing strategies upon the schedule's efficiency and robustness. Each balancing strategy is associated with a different objective function. The resources whose utilization is balanced are the Operating Rooms and the post-surgical beds. The MIP model is solved considering deterministic values for the surgical times and length of stays. The schedule robustness against the variability of both these times is assessed via discrete event simulation.

12.1 Introduction

The Operating Theatre (OT) is one of the most critical functional area in a hospital. In fact, it drives most of the hospital admissions and it is responsible for most of its costs [1]. Optimizing the OT operations, is therefore a primary concern for an increasing number of hospitals. One of the most challenging problem that hospitals need to face in this regard is the planning and scheduling of surgical activities. Such a problem is usually solved in cascade, addressing three intertwined sub-problems [2]: (i) the case mix planning, (ii) the master surgical scheduling (MSS) and (iii) the patients scheduling. In this paper we focus on the MSS problem. This problem consists, essentially, in: (i) determining the specialty (or specialties)

C. Banditori (✉) • F. Visintin
IBIS Lab, Dipartimento di Ingegneria Industriale, Università degli Studi di Firenze, Firenze, Italia
e-mail: carlo.banditori@unifi.it; filippo.visintin@unifi.it

P. Cappanera
IBIS Lab, Dipartimento di Ingegneria dell'Informazione, Università degli Studi di Firenze, Firenze, Italia
e-mail: paola.cappanera@unifi.it

A. Matta et al. (eds.), *Proceedings of the International Conference on Health Care Systems Engineering*, Springer Proceedings in Mathematics & Statistics 61,
DOI 10.1007/978-3-319-01848-5_12,

to assign to each operating room (OR) and session in each day of the planning cycle; and (ii) specifying the number and the typologies of surgeries that should be performed in each OR/session [3]. Solving a MSS problem is, indeed, remarkably complex. It requires, in fact, considering a wide set of issues (e.g. the expected surgical times and the expected length of stay (LoS)), constraints (e.g. resources availability, management of waiting lists), and the priorities, often conflicting, of the stakeholders (management, patients, surgeons, staff). Ideally, a MSS should be: efficient, robust and balanced.

Firstly, it should allow for obtaining a high patient throughput and high resource utilization thus to increase revenues, contain costs, and reduce waiting times (efficiency). Secondly, it should be easy to implement, i.e. the natural deviations of surgical time and LoS from their expected values should not cause schedule disruptions and, consequently, patients' dissatisfaction (robustness). Finally, the MSS should lead to a balanced distribution of the daily workload across the different ORs. A balanced solution, in fact, determines a fair distribution of workload among the OT staff (nurses, surgeons, etc.) and positively affects the employee satisfaction. These objectives, however, can be conflicting. A higher efficiency, for example, can lead to a lower robustness. When the resources utilization is very high, in fact, if a surgery lasts more than its expected duration, surgical teams might be requested to work overtime. Similarly, if a patient occupies a post-surgical bed (hereinafter bed) for a number of days exceeding the expected LoS, there might be no available beds to accommodate the patients scheduled for the following days, which could lead to surgery cancellations and/or postponement.

In this study, starting from a Mixed Integer Programming (MIP) model we developed in a previous work [4], we define new objective functions that allow to obtain efficient, robust and well-balanced MSSs. The MSS robustness against the surgical times and patients' LoS is assessed via simulation. Optimization and simulation models are then jointly used to investigate the relationships between efficiency, robustness and balancing.

The remainder of the paper is organized as follows: in Sect. 12.2 we provide a brief overview of the literature. In Sect. 12.3 we illustrate the main characteristics of the MSS problem and the specific aspects we have addressed. In Sect. 12.4 we describe both optimization and simulation models. In Sect. 12.5 we present the preliminary results of the study and finally, in Sect. 12.6, we draw out conclusions and outline the direction of our future research efforts.

12.2 Literature Review

The surgical scheduling problem has been the object of a relevant number of contributions in recent years [5–7]. Such a problem has been modeled using different mathematical techniques, mainly mathematical programming, and solved

through exact or heuristic approaches considering different objective functions [5, 6]. For example, Blake et al. [8] propose a MIP model to provide a MSS where the undersupply of OR time to the surgical specialties with respect to fixed target is minimized. Santibáñez et al. [9], instead, solve their MIP model using different objective functions, specifically the minimization of the deviation among scheduled and target throughput and the minimization of bed utilization (the latter with the extent to balance such an utilization). Several studies, however, are based on deterministic data. Other authors, instead, take into explicit consideration the impact that the variability either of surgical time or LoS may have on the MSS implementation. Specifically, Zhang et al. [10], Van Oostrum et al. [3] take into account the variability of surgical time whereas Belien et al. [11] consider the LoS variability. To the best of our knowledge, both sources of variability are considered only in our previous work [4] where, however, the relationships among efficiency, robustness and resource balancing are not investigated. With this paper we specifically address this literature gap.

12.3 Problem Addressed

The problem we address in this study is characterized by the following features.

First, we consider three resources: (i) ORs, whose opening time is divided into time slots; (ii) beds, which can be organized in different wards, each accommodating different types of patients; (iii) surgical teams, whose availability is defined in terms of time slots per week. Since scheduling a surgery requires the simultaneous availability of an adequate amount of these three resources, we have categorized the cases in the hospital waiting lists into surgery groups according with the required surgical team, surgical time and LoS. Each surgery group, thus, includes all the procedures requiring a surgical team of the same specialty and are characterized by similar surgical time and LoS.

Second, we assume that the mix of scheduled surgeries, in terms of short/long surgical time and/or LoS, i.e. the intensity care level, should reflect the one of the waiting lists. By doing so we avoid leaving an excessive amount of resource consuming and "complex" cases in the waiting list which would make the planning process more difficult in the following periods.

Finally, the decisions we address concern:

1. The assignment of surgical specialties to ORs and time slots
2. The determination of the amount of procedures in each surgery group to be scheduled in each time slot

with the aim to maximize the patient throughput, minimize the expected overtime and cancellations, and balancing the daily workload of the ORs.

12.4 Models Description

Here the optimization and the simulation models are described.

12.4.1 Optimization Model

In order to obtain efficient and robust solutions, the objective function of the optimization model includes, besides a term for the throughput maximization, other two terms that aim to level respectively the daily utilization of ORs and beds. The rationale of these terms is that if the daily utilization profiles of the ORs and the beds are nicely balanced there should be always enough idle resources to absorb the unexpected peaks caused by the variability of surgical time and LoS [11]. Two different balancing strategies are tested, giving rise to as many objective functions.

Let us define the following sets and parameters:

D	The set of days of the planning horizon, indexed by d
$\tilde{D}$	The set of days in D in which ORs are open
W	The set of weeks of the planning horizon, indexed by w
T	The set of time slots, indexed by t
O	The set of ORs, indexed by o
B	The set of bed types, indexed by b
S	The set of surgical specialties, indexed by s
K	The set of surgery groups, indexed by k
G	The set of intensity care levels, index by g
M	A suitably big constant
H_{odt}	The available time of OR o, on day d and time slot t
F_{bd}	The number of beds of type b available on day d
L_{sw}	The availability of surgical team s for week w, expressed in number of time slots
s_k	The specialty of surgery group k
f_k	The bed type required by surgery group k
g_k	The intensity care level to which the surgery group k belongs
γ_k	The average surgery duration of surgery group k
β_k, α_k	The average numbers of days of hospitalization, before and after surgery, required by surgery group k
$\overline{\overline{G}}_g, \underline{\underline{G}}_g$	The maximum and the minimum percentage of procedures with an intensity care level g that can be scheduled
$\overline{\overline{U}}, \underline{\underline{U}}$	The upper and the lower threshold on the total ORs utilization
W_1, W_2, W_3	The weights used in the objective functions.

Then let us define the following variables:

x_{sodt} Binary, 1 if specialty s is assigned to OR o on day d and time slot t, 0 otherwise

y_{kodt} The number of procedures of surgery group k assigned to OR o on day d in time slot t

Furthermore, let us define the following auxiliary variables:

z_{bd} The number of beds of type b occupied on day d

u_{odt} The utilization of OR o, on the day d and time slot t

v_{bd} The utilization of beds of type b, on day d.

Using these variables and parameters, we can state the feasibility set as follows:

$$\sum_{s \in S} x_{sodt} \leq 1 \qquad \forall o \in O, \forall d \in \tilde{D}, \forall t \in T \tag{12.1}$$

$$\sum_{o \in O} x_{sodt} \leq 1 \qquad \forall s \in S, \forall d \in \tilde{D}, \forall t \in T \tag{12.2}$$

$$\sum_{k \in K: s_k = s} y_{kodt} \leq M x_{sodt} \qquad \forall s \in S, \forall o \in O, \forall d \in \tilde{D}, \forall t \in T \tag{12.3}$$

$$\sum_{k \in K} \gamma_k y_{kodt} \leq H_{odt} \qquad \forall o \in O, \forall d \in \tilde{D}, \forall t \in T \tag{12.4}$$

$$\sum_{\substack{k \in K: f_k = b \\ o \in O, t \in T}} \sum_{d' = \max(1, d - \alpha_k)}^{\min(|D|, d + \beta_k)} y_{kod't} = z_{bd} \qquad \forall b \in B, \forall d \in D \tag{12.5}$$

$$z_{bd} \leq F_{db} \qquad \forall b \in B, \forall d \in D \tag{12.6}$$

$$\sum_{o \in O, t \in T} \sum_{d = 7w-6}^{7w} x_{sodt} \leq L_{sw} \qquad \forall s \in S, \ \forall w \in W \tag{12.7}$$

$$\underline{\underline{G}}_g \sum_{\substack{k \in K, o \in O \\ d \in D, t \in T}} y_{kodt} \leq \sum_{\substack{k \in K: g_k = g \\ o \in O, d \in D, t \in T}} y_{kodt} \leq \overline{\overline{G}}_g \sum_{\substack{k \in K, o \in O \\ d \in D, t \in T}} y_{kodt} \quad \forall g \in G \tag{12.8}$$

$$u_{odt} = \frac{\sum_{k \in K} \gamma_k y_{kodt}}{H_{odt}} \qquad \forall o \in O, \forall d \in \tilde{D}, \forall t \in T \tag{12.9}$$

$$v_{bd} = \frac{z_{bd}}{F_{db}} \qquad \forall b \in B, \forall d \in \tilde{D} \tag{12.10}$$

$$\underline{\underline{U}} \leq \frac{\sum_{\substack{o \in O, d \in \tilde{D}, \\ t \in T}} u_{odt}}{|O| \left|\tilde{D}\right| |T|} \leq \overline{\overline{U}} \tag{12.11}$$

$$x_{sodt} \in \{0,1\} \qquad \forall s \in S, \forall o \in O, \forall d \in D, \forall t \in T \tag{12.12}$$

$$y_{kodt} \in \mathrm{N} \qquad \forall k \in K, \forall o \in O, \forall d \in D, \forall t \in T \tag{12.13}$$

Two alternative objective functions are considered. For both of them specific variables and constraints are defined. The first one (12.14) minimizes the maximum ORs ($\overline{\overline{u}}$) and beds ($\overline{\overline{v}}$) daily utilizations, i.e.:

$$\min \quad W_1\overline{\overline{u}} + W_2\overline{\overline{v}} - W_3 \sum_{\substack{k \in K, o \in O \\ d \in D, t \in T}} y_{kodt} \tag{12.14}$$

$$u_{odt} \leq \overline{\overline{u}} \qquad \forall o \in O, \forall d \in \tilde{D}, \forall t \in T \tag{12.15}$$

$$v_{bd} \leq \overline{\overline{v}} \qquad \forall b \in B, \forall d \in \tilde{D} \tag{12.16}$$

The second objective function (12.17) minimizes the gaps between the maximum and the minimum values of ORs and beds daily utilizations: as before, $\overline{\overline{u}}$ and $\overline{\overline{v}}$ represent the maximum daily utilizations of ORs and beds, whereas $\underline{\underline{u}}$ and $\underline{\underline{v}}$ represent the minimum values of such utilizations.

$$\min \quad W_1\left(\overline{\overline{u}} - \underline{\underline{u}}\right) + W_2\left(\overline{\overline{v}} - \underline{\underline{v}}\right) - W_3 \sum_{\substack{k \in K, o \in O \\ d \in D, t \in T}} y_{kodt} \tag{12.17}$$

$$\underline{\underline{u}} \leq u_{odt} \leq \overline{\overline{u}} \qquad \forall o \in O, \forall d \in \tilde{D}, \forall t \in T \tag{12.18}$$

$$\underline{\underline{v}} \leq v_{bd} \leq \overline{\overline{v}} \qquad \forall b \in B, \forall d \in \tilde{D} \tag{12.19}$$

Both the objective functions are composed of three terms, whose relative importance can be set by means of weights. The first and the second term of both the objective functions are the balancing terms. The former acts on ORs utilizations while the latter on beds' ones. The third term of both the objective functions maximizes the number of scheduled cases.

In order to make a fair comparison between the strategies and to avoid trivial solutions (if $W_3 << W_1$ and $W_3 << W_2$), the average ORs utilization over the planning horizon is bound in a range (12.11).

A brief description of the constraints follows. Constraints (12.1) guarantee that at most one surgical specialty can be assigned to a given OR in each time slot of the planning horizon. Constraints (12.2) assure that in each time slot a given specialty cannot occupy more than one OR. Constraints (12.3) bind together assignment (x) and scheduling variables (y). Constraints (12.4) state that the total time consumed by all the procedures scheduled in a given OR, in each time slot, cannot exceed the

available OR time. Constraints (12.5) and (12.6) respectively compute the number of beds of each type occupied in each day and limit it to the bed availability. Constraints (12.7) control the surgical teams availability in each week. Constraints (12.8) are the mix constraints with respect to the intensity care level and control that for each level the number of scheduled procedures falls inside the pre-defined range. Constraints (12.9) and (12.10) compute the utilization factor respectively of each OR in each time slot and of each bed type in each day over the planning horizon. Constraints (12.12) and (12.13) define the bound on the variables. Constraints (12.15) and (12.16), that are considered when the objective function (12.14) is used, compute the maximum values of daily utilization of OR and beds. Finally Constraints (12.18) and (12.19), that are instead considered when the objective function (12.17) is used, compute both the maximum and the minimum values of daily utilization of OR and beds.

12.4.2 Simulation Model

As pointed out in Sect. 12.1, the effectiveness of the two strategies in terms of robustness is assessed via simulation. Specifically, we have used a discrete-event simulation model that works as follows: the model reads the schedule produced in the optimization phase, generates a number of entities equal to the number of surgeries planned for the planning horizon and links each entity with its surgery group. Hence, for each simulated day a number of entities equal to those planned for the day enter in the system. These entities seize the ORs and beds they have been assigned in the MSS, and release them after a time that is randomly sampled from the empirical distributions of surgical time and LoS associated with their surgical group. The model, thus, keeps track of the actual duration of the surgical sessions and of the number of occupied beds, as well as of the overtime and overbooking that may occur.

12.5 Computational Results

In this section we present the preliminary results of our study. The section is organized in three subsections. In the first one we give a brief description of the assumptions and the data considered in this experimental campaign. In the second and third subsection we show, respectively, the results of the optimization and of the simulation study. The optimization model has been coded in AMPL and solved through the IBM ILOG CPLEX solver (version 12.4). For all the analyzed

scenarios we have bounded the computational time to 30 min; solutions of very good quality are always obtained within such a time limit. The simulation model, instead, has been created with Rockwell Arena (version 13.9) and integrated with AMPL via VBA.

12.5.1 Input Data

The experimental campaign presented upon here is based on real data coming from one leading Italian hospital. In particular, we have considered:

- Elective surgery;
- A planning horizon of 14 days (2 weeks);
- Daily surgical sessions, i.e. one time slot in each day;
- 12 surgical specialties and 39 surgery groups;
- 4 interchangeable ORs and 47 beds;
- For each OR, a daily opening time equal to 690 min for 5 days per week;
- A percentage of low care intensity surgeries in the waiting list ranging from 30% to 40%;

Referring to this latter point, we have considered as low care intensity surgeries the so-called day surgeries, i.e. those surgeries for which the patients occupy a bed only for one day (the one they undergo a surgical procedure).

The weights in the objective functions are hierarchically set. Specifically, in order to obtain robust solutions the balancing terms are prioritized with respect to the throughput maximization term. Furthermore, since overbooking is more undesirable then overtime, balancing beds' daily utilization is prioritized with respect to balancing ORs daily utilization. Hence we have that $W_2 >> W_1 >> W_3$.

12.5.2 Optimization results

The performance of the two objective functions has been tested in correspondence with different OT workload. Specifically, through constraints (12.11), we have considered five different OR utilization ranges (70–75%, 75–80%, 80–85%, 85–90%, 90–95%) giving rise to $5 \times 2 = 10$ different scenarios. For each scenario we report (Table 12.1) the number of scheduled surgeries (N), the percentage of scheduled low care intensity surgeries (%LC) as well as the mean values of the surgical time (ST) and LoS (LOS) of the surgeries scheduled in the MSS. Finally we report the values of the mean (M), the standard deviation (Sd), the maximum (Max) and the range (Rng), calculated across the 10 working days, of both the ORs and beds daily utilizations.

Table 12.1 Optimization results

Objective function	Min(max)					Min(range)				
OT workload [%]	70–75	75–80	80–85	85–90	90–95	70–75	75–80	80–85	85–90	90–95
Scenario	1	2	3	4	5	6	7	8	9	10
N [patients]	215	226	246	268	295	278	288	319	333	333
%LC [%]	40	39.82	39.84	39.93	40	39.93	39.93	39.81	39.94	39.94
ST [min]	89.89	91.60	89.76	87.59	84.21	71.14	74.67	73.44	74.53	76.39
LOS [days]	1.67	1.72	1.73	1.69	1.72	1.66	1.66	1.70	1.64	1.65
M(ORs) [%]	70.02	75	80	85.05	90	71.66	77.91	84.88	89.92	92.17
Sd(ORs) [%]	30.84	30.01	27.15	23.61	16.69	1.41	1	1.43	1.86	1.09
Max(ORs) [%]	90.29	97.39	97.68	99.71	99.13	73.91	79.71	86.96	92.75	94.2
Rng(ORs) [%]	90.29	97.39	97.68	99.71	99.13	4.64	3.33	4.64	5.8	3.91
M(Beds) [%]	61.7	65.96	72.34	78.72	87.23	82.98	87.23	95.74	100	100
Sd(Beds) [%]	0	0	0	0	0	0	0	0	0	0
Max(Beds) [%]	61.7	65.96	72.34	78.72	87.23	82.98	87.23	95.74	100	100
Rng(Beds) [%]	0	0	0	0	0	0	0	0	0	0

As can be observed:

- For each scenario the daily beds utilizations are perfectly balanced (Sd(Beds) = 0); however, in correspondence with the same OT workload, the Max(Beds) values obtained with the min(max) strategy are lower than the ones obtained with the min(range) one;
- ORs are less balanced (Sd(ORs) > 0) than beds. In correspondence with each OT workload the min(range) strategy performs better than the min(max) one. Moreover the former strategy leads to lower maximum values of daily ORs utilization than the latter one. This fact can be justified as follows: the min(range) objective function assumes the minimum daily range value (Rng(Beds) = 0) in correspondence with different level of beds mean utilization. Therefore the solver is free to choose the solution that minimizes the daily range of ORs utilization. With the min(max) strategy, instead, the solution that minimizes the maximum daily bed utilization is clearly more penalized with respect to the ORs balancing, making substantially useless the effects of the second term;
- For each strategy, the higher the OT workload, the higher the number of scheduled surgeries. However, among the two strategies, the min(range) one seems to be the most efficient. This can be due to the fact that the min(max) strategy, in order to keep low the maximum daily bed utilization, chooses surgeries characterized by an higher surgical time (LOS is essentially bound by the constraint (12.8)), causing a lower efficiency.

12.5.3 Simulation Results

In this section we show the results of the simulation study. For each scenario we have performed |I| = 30 simulation runs and recorded the overtime and the overbooking. Overtime occurs when the difference between the duration of the surgical session in a OR and its available time is positive. Similarly, overbooking occurs when the number of hospitalized patients on a given day exceeds the number of available beds. For both overtime and overbooking we have calculated the mean values over the 30 replications as follows:

$$M(OVT) = \frac{\sum_{\substack{o \in O, d \in \tilde{D} \\ t \in T, i \in I}} \max\left(0; \sum_{k \in K} \widehat{\gamma}_{ki} y_{kodt} - H_{odt}\right)}{|O|\,|\tilde{D}|\,|T|\,|I|}$$

$$M(OVB) = \frac{\sum_{\substack{b \in B, d \in \tilde{D} \\ i \in I}} \max\left(0; \widehat{z}_{bdi} - F_{bd}\right)}{|B|\,|\tilde{D}|\,|I|}$$

with i indicating the i-th replication, $\widehat{\gamma}_{ki}$ the value assumed in the i-th replication by the surgical time of the case belonging to surgical group k and $\widehat{z}_{bdi}$ the value assumed in the i-th replication by the number of hospitalized patients in beds of type b, on day d. To assess the schedule robustness we have calculated the mean and the max values of both the overtime and the overbooking. The mean values (M(OVT), M(OVB)) allow us to understand the mean difference between the available resources and those actually needed to implement the MSS. The max values (Max(OVT), Max(OVB)), instead allow us to understand what happened in the worst case scenario. This latter information is very relevant as well. Extremely high overtime values, in fact, may be unacceptable for the OT staff and, in case of day surgery, make impossible to dismiss the last patients before the end of the day. In the same way extremely large overbooking would determine a number of cancellations which could seriously hamper patients satisfaction. In Table 12.2 we report the mean (M), the standard error of the mean (SEM), the third quartile (Q3) and maximum (Max) for OVT and OVB. Even if not explicitly reported in Table 12.2 we have performed several independent t-test to compare the mean values of the different indicators across scenarios. For the most relevant tests hereafter we show both the p-value (p) and the effect size (r).

Looking at Table 12.2, it is possible to notice that for each OT workload level, min(range) strategy has obtained a M(OVT) value that is significantly lower than the value achieved with the min(max) one ($p < 0.05$); in addition the r is always higher than 0.51, which indicates a fairly large effect [12]. Similar considerations can be made for the Max(OVT) value, except for the scenario 10 where Max(OVT) for min(range) is higher than for min(max). This fact however, is certainly due to an exceptionally large value of the surgical time characterizing the scenario 10 worst case. Indeed, if we compare Q3(OVT) of scenario 5 and 10 we can notice that the former is bigger than the latter. With regard to the overbooking, instead, min(range) strategy has obtained M(OVB) values that are significantly higher than the values achieved with the min(max) one ($p < 0.05$ and $r > 0.31$), in correspondence with each OT workload level; similar considerations can be made for Q3(OVB) and Max(OVB) values.

Instead, by comparing the results of optimization (Table 12.1) and simulation (Table 12.2), it is possible to observe that:

- The minimum values of Max(OVT), M(OVT) and Q3(OVT) are obtained in correspondence with the scenarios in which the Max(ORs) is minimum (scenarios 1, 6, 7, 8). In addition when Max(ORs) exceeds a threshold value (approximately 80%), M(OVT) becomes significantly higher than 0 ($p < 0.05$ and $r > 0.8$) and tends to grow with Max(ORs). However similar values of M(OVT) are obtained in scenarios characterized by different Max(ORs), Rng(ORs) and Sd(ORS), e.g. scenarios 3 and 9 ($p > 0.5$). This might be due to the fact that if all of the ORs are close to their maximum utilization (low Rng(ORs) and Sd(ORs)), the resulting overall overtime will be probably greater than the case in which some ORs are empty (high Rng(ORs) and Sd(ORs)), even if the maximum utilization is higher

Table 12.2 Simulation results

Objective function	Min(max)					Min(range)				
OT workload [%]	70–75	75–80	80–85	85–90	90–95	70–75	75–80	80–85	85–90	90–95
Scenario	1	2	3	4	5	6	7	8	9	10
M(OVT)	7.96	15.08	17	41.06	38.42	0	0.2	7.02	16.09	23.43
SEM(OVT)	1.25	1.81	2.06	3.32	2.66	0	0.11	1.21	1.91	2.18
Q3(OVT)	0	15	20	61.5	65	0	0	0	18	38
Max(OVT)	130	188	230	350	219	0	25	145	255	300
M(OVB)	0	0	0	0	0.14	0.01	0.09	0.84	2.34	2.12
SEM(OVB)	0	0	0	0	0.03	0.01	0.03	0.08	0.15	0.15
Q3(OVB)	0	0	0	0	0	0	0	1	4	4
Max(OVB)	0	0	0	0	5	1	4	7	11	11

than in the first case. Finally, it seems there is no clear relationship between the OVT and number of scheduled surgeries nor between the OVT and the bed utilization.

- There is a relationship between OVB and beds utilization. For each scenario, in fact, the higher the bed utilization (M(Beds) = Max(Beds)), the higher M(OVB) and Max(OVB). However, if M(Beds) is smaller than a threshold value (approximately 80%), then, there is not overbooking i.e., M(OVB) = 0 ($p < 0.05$ and $r > 0.31$) and Max(OVB) = 0. In addition, the higher the number of scheduled surgeries, the higher M(OVB) and Max(OVB). Nonetheless, if the number of scheduled surgeries is smaller than a threshold value (around 270) no overbooking occurs. On the contrary, it seems that there are no relationships between overbooking and ORs utilization.

12.6 Conclusions and Future Research

In this paper we have presented the preliminary results of a study aiming at investigating the relationships among efficiency, balancing and the robustness for the MSS problem.

In particular we have analyzed the MSSs produced by a MIP model in correspondence with two different objective functions: min(max), min(range). Each objective function incorporates a different strategy to balance the daily utilizations of two key resources: beds and ORs. In particular min(max), minimizes the maximum daily utilization value, while min(range), minimizes the difference between the maximum and the minimum daily utilization values. The obtained schedules, characterized by different efficiency and balancing levels, have been then simulated. The simulation allowed us to compute the overtime and overbooking associated with each schedule and to assess the relevant robustness. Specifically, the min(range) strategy seems to be the most efficient, i.e. for each OT workload level it schedules more surgeries than the min(max) one. In fact, although both strategies lead to a perfect beds daily utilization balancing, the mean bed utilization for min(range) strategy is higher than for min(max) one. Min(range) leads to a better ORs utilization balancing: its maximum values and ranges are lower than the one obtained with the min(max) strategy. Simulation has then revealed that min(max) strategy produces more robust schedules with respect to overbooking. On the contrary the min(range) strategy guarantees a more balanced workload between the different ORs and more robust solutions with respect to overtime. We can thus conclude that to obtain robust solutions it is better to focus on keeping low the maximum daily values of the resources utilizations rather than trying to reduce the gap between their maximum and minimum values. Instead, reducing the utilization range, is essential to obtain a balanced ORs workload. Unfortunately, in this study we have not obtained both these positive effects. For these reasons our future research efforts will be focused on testing objective functions (i) incorporating different balancing strategies for the two considered resources and (ii) investigating different weight (W_1, W_2, W_3) settings.

References

1. Denton, B., Viapiano, J., Vogl, A.: Optimization of surgery sequencing and scheduling decisions under uncertainty. Health Care Manag. Sci. **10**, 13–24 (2007)
2. Beliën, J., Demeulemeester, E.: Building cyclic master surgery schedules with leveled resulting bed occupancy. Eur. J. Oper. Res. **176**, 1185–1204 (2007)
3. Van Oostrum, J.M., Van Houdenhoven, M., Hurink, J., Hans, E., Wullink, G., Kazemier, G.: A master surgical scheduling approach for cyclic scheduling in operating room departments. OR Spectr. **30**, 355–374 (2008)
4. Banditori, C., Cappanera, P., Visintin, F.: A combined optimization–simulation approach to the master surgical scheduling problem. IMA J. Manag. Math. **24**(2), 155–187 (2013)
5. Cardoen, B., Demeulemeester, E., Beliën, J.: Operating room planning and scheduling: a literature review. Eur. J. Oper. Res. **201**, 921–932 (2010)
6. Guerriero, F., Guido, R.: Operational research in the management of the operating theatre: a survey. Health Care Manag. Sci. **14**, 89–114 (2011)
7. May, J.H., Spangler, W.E., Strum, D.P., Vargas, L.G.: The surgical scheduling problem: current research and future opportunities. Prod. Oper. Manag. **20**, 392–405 (2011)
8. Blake, J.T., Dexter, F., Donald, J.: Operating room managers' use of integer programming for assigning block time to surgical groups: a case study. Anesth. Analg. **94**, 143–148 (2002)
9. Santibáñez, P., Begen, M., Atkins, D.: Surgical block scheduling in a system of hospitals: an application to resource and wait list management in a British Columbia health authority. Health Care Manag. Sci. **10**, 269–282 (2007)
10. Zhang, B., Murali, P., Dessouki, M.M., Belson, D.: A mixed integer programming approach for allocating operating room capacity. J. Oper. Res. Soc. **60**, 663–673 (2008)
11. Beliën, J., Demeulemeester, E., Cardoen, B.: A decision support system for cyclic master surgery scheduling with multiple objectives. J. Sched. **12**, 147–161 (2009)
12. Field, A.: Discovering Statistics Using SPSS, 2nd edn. Sage, London (2005)

Chapter 13
Expert's Evaluation of Innovative Surgical Instrument and Operative Procedure Using Haptic Interface in Virtual Reality

G. Thomann, D.M. Pan Nguyen, and J. Tonetti

Abstract In the domain of designing innovative products in the medical field, investigations are often oriented towards communication between actors and needs comprehension. In the DESTIN (DEsign of Surgical/Technological INnovation) project, User Centered Design methodology with concrete experiments is applied. Researchers propose experimentation in operating room for innovative products and new adapted surgical procedures co-evaluation. In this paper, they intend to evaluate the usage of the product in a virtual environment using a 3D haptic feedback system. Researchers not only propose a better ergonomic situation of the physician in front of the operating screen, but also increase the performance of the simulator in order to allow the manipulation of the innovative surgical instrument developed. We used virtual reality environment and the manufactured prototype with the aim to validate the new surgical procedure and the innovative designed surgical instrument.

13.1 Research Context

The development of new technologies in medicine can significantly improve the effectiveness. On the contrary, the use of more complex systems tends to make the practice of medicine more difficult. In particular, this complexity reinforces the importance of preoperative planning and postoperative monitoring. New technologies in informatics and virtual reality allow physicians to better interpret the enormous amount of information that is provided by the imaging systems or therapy

G. Thomann (✉) • D.M.P. Nguyen
G-SCOP Laboratory, Grenoble Institute of Technology, Grenoble, France
e-mail: Guillaume.thomann@grenoble-inp.fr; duy_minh1988@yahoo.com

J. Tonetti
Orthopaedic and Traumatology Centre, Michallon Hospital, Grenoble, France
e-mail: jtonetti@chu-grenoble.fr

A. Matta et al. (eds.), *Proceedings of the International Conference on Health Care Systems Engineering*, Springer Proceedings in Mathematics & Statistics 61,
DOI 10.1007/978-3-319-01848-5_13,

systems [1]. Specifically, virtual reality allows better understanding, better planning and better work through visualization of three-dimensional images of anatomy and pathology. In addition, virtual reality can help the practitioner through the stages of diagnosis, therapy, and postoperative monitoring.

The main aim of DESTIN project (Design of Surgical-Technological INnovation) is to propose a new process approach focused on this specific context: How to create a new operative surgical procedure and coupled with an innovative surgical instrument when a new medical approach is imagined?

The specific surgical application we are working on addresses thoraco-lumbar fracture. The current "classical" procedure is carried out with the patient in the prone position under general anesthesia. The surgeon performs a posterior open approach through a 15 cm large incision. The posterior vertebral arch is exposed. Pedicle screw entry points are chosen by direct visual control and they are fixed to the vertebrae. Rods are placed to connect the pedicle screws together. Prone placement added with rod-screw connection provides reduction of the trauma deformity and durable stability Thus, vertebrae are preventing from moving while bone healing and graft fusion takes place.

The new surgical procedure proposed by the surgeon consists in inserted the rod inside the pedicular screws in MIS (Minimally-Invasive Surgery). Thus, new little incisions should allow the insertion of the rod in the three pedicle screws.

In this context, researchers, designers and the medical staff regularly work in the real operating room. This work was very effective but time consuming. It necessitates heavy organization and management (mainly in the hospital), creation of mannequins, manufacturing of many prototypes, etc.

To facilitate this organization by maintaining the essential experimental aspects, we create a CATIA CAD model of the virtual operating room. It integrates patient, medical equipment and surgical instruments. In this virtual environment, the surgeon has to manipulate the virtual surgical instrument on the virtual patient's spine (the spine has been modeled in a compatible format as the CATIA environment and integrated in a mannequin placed on the operative table). The goal of this exercise is to provide information to the designers for the validation of the innovative surgical instruments during the design process. At the same time, it also allows surgeons to perform the operative procedures with haptic feedback as in the real operative case.

The difficulties in this research concern the ability to sufficiently represent the virtual environment for the co-validation of the medical procedure and the innovative surgical instrument.

Supposing that surgeons can manipulate the virtual innovative surgical instrument using a 3D-Haption© haptic system in the virtual operating room, the research questions can be summarized as follow:

- How to modify the configuration of the virtual reality room and the physical interface for a better immersion of the physician in the virtual environment?
- Which are the optimal dimensions of the virtual surgical instruments for a better manipulation feedback using the 3D-Haption© haptic system?

To answer these questions, this research methodology is proposed:

- Research some ergonomic references in the surgery domain and compare them to our virtual reality room organization,
- Secondly design and link a new physical interface to the arm of the 3D-Haption© haptic system,
- Modify the virtual model and to compare the surgical intervention feedback with the real one.

In this article, we firstly present the User Centered Design methodology we use during our study. Then we focus on the virtual reality and ergonomic applications and research in the surgical field. This first step allows us to analyze the general situation in the world. From this work, we propose modifications and adaptations of our current virtual operating room and 3D-Haption© haptic system user interface.

Next, we present the first results of the manipulation in virtual environment and conclusions concerning its efficiency related to the situation in real situation.

13.2 User Centered Design

User Centred Design (UCD) is considered as one of the cornerstones theories about user involvement. UCD, as a design approach, was first time introduced in NF EN ISO 9241-210: Human-Centred Design Processes for Interactive Systems [2]. The main issue is how to involve, integrate and consider the end-user and its requirements throughout the product design process. This ISO 13407 model proposes technical points the project must encompass to be considered as *human centred: 1* – a certain knowledge of the end-users: their tasks and of their environment – 2 – an active participation of these end-users, the clear understanding of their needs and the requirements linked with the tasks – 3 – an appropriate distribution of the end-users/technological functions – 4 – an iterative design solution – 5 – the intervention of a multidisciplinary designing team. This is necessary to better interpret the end-user, its knowledge and how-know: human factors, information architecture, design, quality, marketing, etc.

The UCD cycle is decomposed into six main steps (Fig. 13.1). It is an iterative cycle (step 2–5) which ends when the system answers the end-user requirement (step 6).

To better understand this UCD design steps, Jokela et al. propose another interpretation of this NF EN ISO 9241-210 UCD Process. They explain more concretely how it can be applied on a project and suggest a new UCD process model [3]. Another important issue in UCD is how identifying and selecting relevant end-users in the development work. In practice it is commonly possible to involve only a limited number of users, and therefore it is very important to define criteria in order to select the most "representative users" to centre the design on their requirements and expectations.

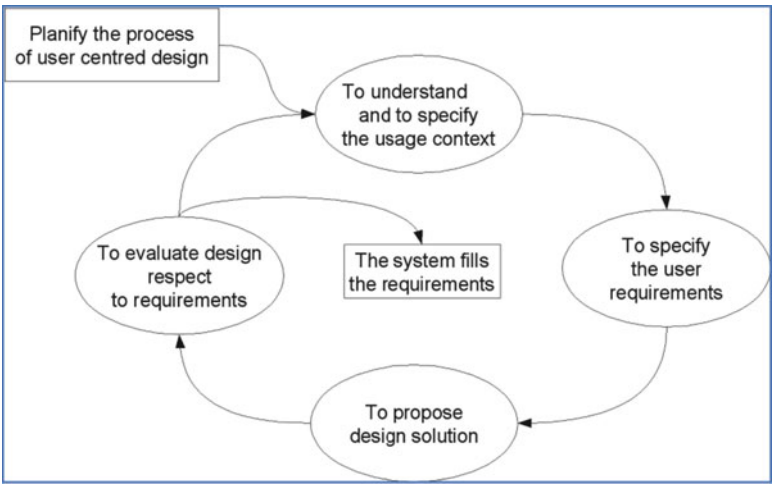


Fig. 13.1 The six steps of the UCD cycle

13.3 Virtual Reality, Ergonomics and Their Application in Surgical Field

13.3.1 Virtual Reality and Application to the Surgical Field

Virtual Reality (VR) is an interactive immersive data-processing simulation in real or imaginary environments. Currently, the technology of VR was applied in many different fields such as: formation by simulator (driving vehicles, aerospace), design of products, the simulation of surgery, meteorology . . .

In the surgical field, the laparoscopy is a procedure which requires surgeons to observe the surgical intervention on a monitor and requires acquisitions of new competences. This Minimally Invasive Surgery (MIS) differs from the open surgery by the fact that the surgeon operates through small incisions and uses specific instruments as scalpel, grips, nets, etc. [4]. In spite of its many advantages (faster recovery of the patients, less damage with healthy tissues and smaller scars, less pain and less need for drugs), the MIS requires a long time training eyes-hands coordination.

Researches developing the haptic control feedback device can be found in [5–7]. To follow the user intentional movements, by interaction between hand and device, high powerful haptic devices must be able to produce force feedback. Consequently, it is essential to closely examine the human touched and the constraints of application during the construction of these devices. A haptic interface with 4 degrees of freedom of freedom was designed by Guiatni et al. to compare it with devices commercially available [4]. This device has the capacity to offer force feedback in all the degrees of freedom available during the MIS procedure.

In our case, researchers, designers and physicians work together on the development of a virtual environment to simulate a MIS operation on the spinal column. The goal is to create a complete virtual surgical environment integrated surgical instruments, haptic feedback, the operating room and the necessary parts of the anatomy.

13.3.2 Ergonomics in the Surgical Field

Ergonomics is based on design models of machines and tools that optimize the performance of users. In our case, the aim of ergonomics consists on improving the simulation conditions in virtual reality surgery environment: creating a better immersion for surgeons by finding the factors that influenced its comfort during operations in the operating room. For example the optimum ergonomic position of the monitor was defined according to various sources in the literature [8–11]. The monitor was at a distance of 0.6 m apart from the subjects' eyes. The monitor height (from the middle of the screen to the ground) was between the operating surface and eyelevel height, and the monitor was inclined (to a maximum of 15°) as by the subjects. Moreover the optimal operating surface height was 80% of the elbow height and the table was positioned in 20° tilt. [12].

In Gurvinder Kaur [13], researcher conducted a test to find the height of the ergonomics table in the minimally invasive surgery. In this study, the height of the table has an effect on the upper joint movements of the shoulders, arms and wrist during laparoscopy. Table height should vary from 65 to 90 cm from the floor. The surgeon should be able to adjust the table corresponding to his/her height in order to bring upper joint movements to the minimum position with the resultant less discomfort in the shoulder, back elbow and the wrist. After analyzing the ratio between the surgeon's height with the height of the operating table, it was assumed that the height of the operating table should be calculated as follows:

$$\text{Table Height} = \text{Surgeon's Height} \times 0.49$$

The ideal posture for the MIS is supposed in the literature [14] and [15]. The arms are slightly removed, retroversion, and turned inward at the level of the shoulder (abduction $<30°$). The elbows are bent at about 90–120° of flexion. This position leads to the maximum force to be applied for a maximum duration. The head is slightly bent with an angle of between 15 and 45°.

Through this study literature, we find that the virtual reality technology plays an important role in many areas. In particular, the applications of VR technique in surgical simulation have been developed to provide better and better ergonomic solutions which satisfy users. Through these studies, we can better consider the virtual reality room and design the components that give a better immersion for the surgeons. Thus, we can improve the ergonomics in surgical simulation by changing haptic interface, the position of the surgeon and his posture.

13.4 Related Works

13.4.1 The Human Machine Interface

The practitioner manipulated the haptic arm using the 3D-prining machine handle (Fig. 13.2). The position of the physician was not comparable to the real operating room environment and the conditions of experimentation not ideal:

– The surgeon was not in front of the screen and the posture position not comfortable.
– The 3D-printing handle material was different than the final product's one.

Moreover, the idea is to use the same surgical instrument on mannequin in operating room and during the simulation.

Through ergonomics studies, we can better consider the virtual reality room and design the components that can make a better immersion of the practitioner.

13.4.2 The Surgeon Posture and Position

To perform the simulation with haptic sensation as in real surgical environment, we adapted and modified the haptic interfaces as well as the position of surgeon:

– Setting the table height corresponding to the surgeon's height. We chose table height is equal to 0.49 of surgeon's height [14]. Notice that it is now possible to adjust at real time the position of table?

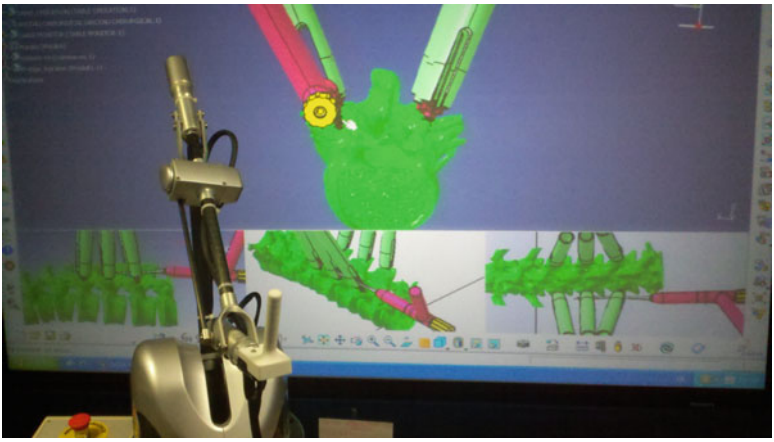


Fig. 13.2 Picture of the 3D-prining machine handle at the end of the haptic arm

Fig. 13.3 The instrument tightened by the piece is tied to the haptic arm

– Adjusting the distance between the screen and the surgeon's position. Normally, this distance is 0.60 m, but with the giant screen in Virtual Reality room at our laboratory; we chose the distance of 1.5 m.

Changing position as well as the posture of the surgeon: the surgery is always in front of the screen. We improve ergonomics in surgery simulation by changing haptic interfaces handle. In order that the surgeon can use the haptic arm in the operating simulation as in reality, we thought to create an intermediate mechanical piece to hang the surgical instrument prototype (Protige) at the end of the haptic arm. The objectives of this adaptation are to give the surgeon a real sensation when holding the real instrument Protige and then to carry it in a direction parallel to the spine's main axis.

Before the mechanic piece was fabricated, we carried out a numerical simulation to ensure the strength, deformation and constraints of the piece to work properly when it tightened the surgical instrument .We divided the simulation into two cases:

– Test the strength of the piece under the Protige's effort when the simulator is running maximum the instrument.
– Test the tightness of the piece under the load of the screws so it could tight well Protige.

The intermediate piece of aluminum has been made at our workshop by Numerical Control of Machine Tools (Fig. 13.3). We observe the 90° modification orientation compared to the previous 3D-prining machine handle (Fig. 13.2).

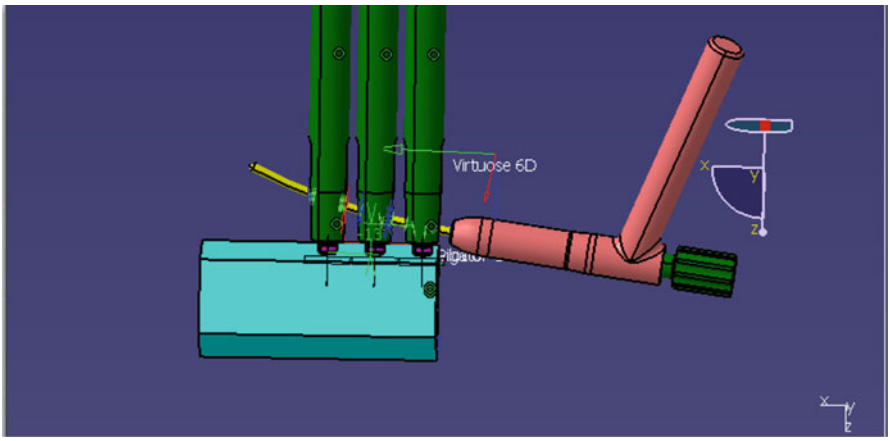


Fig. 13.4 The translation of Protige parallel with the spine position

13.4.3 Testing the Haptic Sensation During the Insertion of the Rod Inside Screws' Holes

Concerning the simulations, the previous virtual surgical instruments dimensions never allows the insertion of the rod inside the pedicle screws head. The main objective of this activity is to find optimal dimensions of the virtual surgical instruments and verify the friction sensation when inserting the rod into the holes of the pedicle screws head. Using a simplified virtual model, we test different manipulation situations (Fig. 13.4): changing screws holes diameters and rod diameters. We modified the diameter of the pedicle screw hole from 6 to 9 mm. We also used different rod diameters: from 4 to 6 mm.

To simplify the simulation with many different cases, we use a simple model of the spine. Of course, the positional parameters between screws and spine are similar than in MIS procedure.

We ran several simulations with different views ("multiple views" on IFC CATIA software). So we set up a kind of viewpoints allowing the surgeon to use more isometric views, each vignette characterizing a different spatial view.

User moves the virtual instrument over the screws and we set the test duration up to 1 min to validate the feasibility of the procedure phase. Five trials are conducted for each case. The experimental duration for each case were taken in order to determine the levels of difficulty in inserting the rod into the holes. A test was considered successful if the positioning time of the rod through the holes did not exceed 1 min. We got the test results with different cases (see an extraction of the results in Table 13.1). Simulation is recorded by video to analyze the results as well as confirmation of results.

We have tested a maximum of the possible experimental conditions. Depending of the trauma cases, the physician has to use two or more screws inside the body.

Table 13.1 Results of the trials (in duration time)

Rod diameter (mm)	Duration of the trials (s) Screws' hole: 8 mm					Mean (s)	Duration of the trials (s) Screws' hole: 9 mm					Mean (s)
4	8	7	6	4	10	7	12	8	7	4	7	7,6
4,5	60	50	120	90	60	76	5	7	8	8	9	7,4
5	Not possible						5	8	9	8	7	7,4
5,5	Not possible						10	8	8	11	9	9,2

Table 13.2 Conclusion for one specific configuration: extraction from the complete table. Hole of the screws: 8 mm/diameter of the rod 4 and 4.5 mm

	Number of screws			Duration of the trials (s)					
Rod diameter (mm)	1	2	3	1	2	3	4	5	Mean (5)
4	OK	OK	OK	8	7	6	4	10	7
4.5	diff.	imp.	imp.	60	90	120	50	60	76

We asked the user to test the virtual insertion of the rod in 1, 2 and 3 screws. An extraction of the complete results is presented in Table 13.2.

"OK" means that the corresponding experiment is working well. For example, inserting a 4 mm diameter rod through 3 pedicle screws' holes of 8 mm takes less than 10 s. Inserting a 4.5 mm diameter rod through one pedicle screw hole of 8 mm takes more than 1 min. We qualified this situation as difficult (diff.). Finally, it is impossible (imp.) for the user to insert the 4.5 mm diameter rod through two or more pedicle screw holes of 8 mm.

The complete experiment shows that the 9 mm pedicle screws' holes always allow the insertion of the rod from 4 to 5.5 mm. For the 8 mm pedicle screws' holes, they are compatible only with the 4 mm rod diameter (insertion through 3 screws) and 4.5 mm rod diameter (insertion through 1 screw). One of the reasons that prevent this insertion is the precision of the collision detection between parts using the IFC CATIA software coupled with the haptic device. It doesn't allow the relative movements between rode and holes even if the rod's diameter is smaller than screws' holes. Moreover, the durations of the trials depend of the user's experience.

13.5 Conclusion

In this study, we not only propose a better ergonomic situation of the physician in front of the operating screen, but also increase the calibration of the simulator in order to allow the manipulation of the real innovative surgical instrument developed.

We used virtual reality environment and the manufactured prototype with the aim to validate the new surgical procedure and the innovative designed surgical instrument. For that, an adaptation piece has been designed, manufactured and manipulated. This adaptation has really increased the real sensation of the user in front of the virtual reality screen.

Moreover, the disposition of the experimental room and the user has evolved. The modification of the model and the different trials with different users allow researchers to find parameters which influence the quality of physical sensation. This activity will allow

– Designers to propose tools and models more realistic for effective simulations during the design process. In consequences, design choices can be more precise.
– Physicians to quickly evaluate and validate an adapted operative procedure.

These experiments with users and researchers give us some qualitative results. The next step will be the evaluation of the complete virtual environment (with different dimensional models) with numerous expert surgeons to:

– Validate the design of the surgical instrument.
– Quantify the sensations of the experts.

The surgical instruments developed are generally composed of multiple mobile parts. One of the future objectives will be to work on the possibility to manipulate all the parts of the product in virtual reality. This objective imposes the integration of multiple cameras and markers in the experimental room.

References

1. Melton, G.B.: Biomedical and health informatics for surgery. Adv. Surg. **44**, 117–130 (2010)
2. NF EN ISO 9241-210: Human-Centred Design Processes for Interactive Systems. International Organization for Standardization, Genève, January (2011)
3. Jokela, T.: Making user-centred design common sense: striving for an unambiguous and communicative UCD process model. In: Proceedings of the Second Nordic Conference on Human Computer Interaction. ACM Press, Aarhus (2002)
4. Guiatni, M., Riboulet, V., Kheddar A.: Design and evaluation of a haptic interface for interactive simulation of minimally-invasive surgeries. In: IEEE/ASME international conference on advanced intelligent mechatronics, Suntec Convention and Exhibition Center, Singapore, 14–17 July 2009
5. Saupin, G., Duriez, C., Cotin, S.: Contact model for haptic medical simulations. In: Proceedings of the 4th International Symposium on Biomedical Simulation. Lecture Notes In Computer Science, vol. 5104, pp. 157-165. London (2008)
6. Zarrad, W., Poignet, P., Cortesão, R., Company, O.: Stability and transparency analysis of a haptic feedback controller for medical applications. In: Proceedings of the 46th IEEE, Conference on Decision and Control, New Orleans, 12–14 Dec 2007
7. Mizokamit, R., Abet, N., Kinoshitatt, Y., He, S.: Simulation of ICSI procedure using virtual haptic feedback model. In: 2007 IEEE/ICME International Conference on Complex Medical Engineering, (2007)
8. Matern, U., Waller, P., Giebmeyer, C., Rückauer, K.D., Farthmann, E.H.: Ergonomics: requirements for adjusting the height of laparoscopic operating tables. J. Soc. Laparoendosc. Surg. **5**, 7–12 (2001)
9. Burgess-Limerick, R., Mon-Williams, M., Coppard, V.L.: Visual display height. Human factors. J. Hum. Factors Ergon. Soc. **42**(1), 140–150 (2000)
10. Jaschinski, W., Heuer, H., Kylian, H.: Preferred position of visual displays relative to the eyes: a field study of visual strain and individual differences. Ergonomics **41**(7), 1034–1049 (1998)

11. Turville, K.L., Psihogios, J.P., Ulmer, T.R., Mirka, G.A.: The effects of video display terminal height on the operator: a comparison of the 15°and 40° recommendations. Appl. Ergon. **29**(4), 239–246 (1998)
12. Van Veelen, M.A., Kazemier, G., Koopman, J., Goossens, R.H., Miejer, D.W.: Assessment of the ergonomically optimal operating surface height for laparoscopic surgery. J. Laparoendosc. Adv. Surg. Tech. **12**(1), 47–52 (2002)
13. Gurvinder, K.: Role of OT table height on the task performance of minimal access surgery. World J. Laparosc. Surg. **1**(1), 49–55 (2008)
14. Van Veelen, M.A.: Human-Product Interaction in Minimally Invasive Surgery: A Design Vision for Innovative Products. Delft University of Technology, Delft, pp. 92e97 (2003)
15. Matern, U., Waller, P.: Instruments for minimally invasive surgery. Surg. Endosc. **13**(2), 174–182 (1999)

Chapter 14
A Robust Optimization Approach for the Operating Room Planning Problem with Uncertain Surgery Duration

Bernardetta Addis, Giuliana Carello, and Elena Tànfani

Abstract This paper deals with the Surgical Case Assignment Problem (SCAP) taking into account the variability pertaining patient surgery duration. In particular, given a surgery waiting list, a set of Operating Room (OR) blocks and a planning horizon, the decision herein addressed is to determine the subset of patients to be scheduled in the considered time horizon and their assignment to the available OR block times. The aim is to minimize a penalty associated to waiting time, urgency and tardiness of patients. We propose a robust optimization approach for the SCAP with uncertain surgery duration, which allows to exploit the potentialities of a mathematical programming model without the necessity of generating scenarios. Tests on a set of real-based instances are carried on in order to evaluate the solutions obtained solving different versions of the problem. Besides the value of the penalty objective function, the solution quality is also evaluated with regards to the number of patients operated and their tardiness. Furthermore, assuming lognormal distribution for the surgery times, we use a set of randomly generated scenarios in order to assess the performance of the proposed solutions in terms of OR utilization rate and number of cancelled patients.

B. Addis (✉)
LORIA, Université de Lorraine, CNRS, INRIA, 615 Rue du Jardin Botanique, Vandœuvre-lès-Nancy, France
e-mail: bernardetta.addis@loria.fr

G. Carello
Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, via Ponzio 34, Milano, Italy
e-mail: carello@elet.polimi.it

E. Tànfani
Dipartimento di Economia, Università di Genova, via Vivaldi 5, Genova, Italy
e-mail: etanfani@economia.unige.it

A. Matta et al. (eds.), *Proceedings of the International Conference on Health Care Systems Engineering*, Springer Proceedings in Mathematics & Statistics 61,
DOI 10.1007/978-3-319-01848-5_14,

14.1 Introduction and Related Work

In the last decades the increase of hospital costs have led health care managers to improve hospital organization, by optimizing resources and increasing operational efficiency. The crucial role that surgery departments play within hospitals has been raising an increasing number of research studies aimed at planning Operating Room (OR) activities. This is due both to the significant costs of development and management of surgical facilities and to the impact that surgical activities have on the demand for hospital services and on waiting times [14]. Exhaustive literature reviews on operating room planning and scheduling are reported in [2] and [6], where the authors analyze in detail different topics related to the problem settings and summarize significant trends in research and possible areas for future research.

In this paper we deal with the OR planning problem assuming a block scheduling, also known as closed block planning approach. This means that, in a given planning period, each specialty receives a number of OR blocks (usually half-day or full day length), in which it schedules its surgical cases [17]. Note that the OR planning and scheduling problem, within the framework of a block scheduling approach, can be viewed as made up of three phases/sub-problems related to three different levels of decisions [14]. In the first phase, the number, type and opening hours of the available ORs, as well as the OR capacity assignment among surgical groups or specialties are determined at a strategic level. Then, a cyclic timetable, denoted "Master Surgical Schedule" (MSS), is constructed on a medium term stand point to define the tactical assignment of specialties to days and ORs. The MSS must then be updated whenever the total amount of OR time assigned to each specialty changes. The last phase, which may be called operational "surgery process scheduling", is composed by two sub-problems referred as "advance scheduling" and "allocation scheduling" problem, respectively [8]. The first sub-problem (1 week to 1 month), called Surgical Case Assignment Problem (SCAP) solves a planning phase by assigning a surgery date and OR to each patient scheduled to be operated over the planning horizon. The second sub-problem, called Surgical Case Scheduling Problem (SCSP) solves a scheduling phase which determines the sequence of surgeries in each OR and day.

Efficient OR planning and scheduling is further complicated by the inherent variability of the duration of the surgical cases, which usually decreases the OR utilization level [15]. Accurate modeling of operating time and procedure time distributions has been recognized as an important factor in effective planning and scheduling systems [11].

In the following, we set our analysis at an operational level and we focus our attention on the problem of determining the assignment of patients to OR blocks, i.e. the SCAP, assuming that patient operating times are random variables that follow lognormal distributions as usually recommended to represent operating time variability, see [5, 9].

Recently, [3] deals with the surgery process scheduling and manages the uncertainty in operating times using a two-stage stochastic model with recourse,

including in the objective function the patient waiting times and the OR idle time and overtime. They compare different heuristics and also analyze the influence of patient sequencing inside the OR blocks. [18] builds a mathematical program considering probabilistic constraints to represent the uncertain duration of surgery procedures. The proposed model shows how to optimize OR utilization without increasing overtime and cancellations. [7] proposes different heuristics for the robust surgery loading problem, aimed at maximizing the utilization of operating theater and minimizing the overtime risk by introducing planned slack times. [10] develops a stochastic programming model with recourse and a sample average approximation method to obtain an optimal surgery schedule with the aim of minimizing patient costs and OR overtime costs. [4] develops a two-stage stochastic model with binary decision variables and simple recourse to deal with both block scheduling and open scheduling strategies. The model determines the surgeries assignment to ORs by minimizing the maximum cost associated with uncertain surgery duration. In [13] a two-level framework is proposed. In the first level, a MIP model finds a deterministic solution for the OR planning problem. In the second level, the variability of surgery duration is taken into account by means of individual chance constraints for each OR block and a robust solution is achieved by iteratively adding safety slacks to the first level deterministic model solutions.

In this paper, we propose a robust optimization approach to solve the SCAP with uncertain surgery duration, with the aim of minimizing a penalty function associated to waiting time, urgency and tardiness of patients. The robustness of solutions is achieved by applying the approach proposed in [1], which allows to exploit the potentialities of a linear programming model without the necessity of generating scenarios. The formulations of different versions of the problem are proposed and computational tests over a set of real life based instances are presented and compared in order to evaluate the different versions in terms of computational effort and solution quality. Besides the behaviour of the penalty objective function, the solution quality is also evaluated with regards to the number of patients operated and their tardiness. Moreover, assuming lognormal distributions for the surgery times, a set of randomly generated scenarios is used in order to compare the proposed solutions in terms of OR utilization rate and number of rescheduled patients.

The remainder of this paper is organized as follows: in Sect. 14.2 we introduce the problem under investigation and we give the different versions formulations. In Sect. 14.3, the results on a set of real-based randomly generated instances are reported and compared. Finally, in Sect. 14.4 conclusions and future research directions are given.

14.2 Problem Description and Models

In the following we concentrate our analysis on the so called Surgical Case assignment Problem (SCAP). The problem consists in determining the assignment of a set of elective patients I to a set J of OR blocks in a considered planning

horizon D. We assume a block scheduling approach and focus on a single surgical specialty. Note that the solution approach herein presented could be easily adapted for considering more than one specialty. The set J of OR blocks assigned to the specialty and their distribution during the planning horizon are given: each block is described by an operating room and a day. More precisely, a weekly based pattern describes blocks availability: according to such pattern, a set of days $D_j \subset D$ is given for each block j, which represents the set of day indexes in which block j is available. The available total time of each time block j, i.e. the OR block length, is denoted as γ_j. We consider also the case where overtime is allowed. In particular, δ represents the amount of overtime allowed for a given block, while Δ is the maximum number of blocks which can have overtime in the considered planning horizon. For each patient i, let denote with w_i the number of days which the patient has spent in the waiting list, i.e. the waiting time at the beginning of the planning horizon. Moreover, a maximum waiting time l_i and a corresponding urgency parameter u_i are set for each patient i. If the patient has spent w_i days in the waiting list, he/she must have surgery before day $dd_i = l_i - w_i$, otherwise he/she is considered tardy. The surgery time for each patient i is a random variable $\tilde{t}_i$ that follows a lognormal probability distribution $F(\tilde{t}_i)$. The mean and standard deviation parameters are equal to t_i and $\hat{t}_i$, respectively. The problem consists in selecting a subset of patients to be scheduled in the considered planning horizon and to assign them to OR blocks, while guaranteeing that the capacity of each block is not exceeded. The objective function aims at minimizing the overall penalty due to delay in serving the patients. As proposed in [12] it takes into account both the urgency and waiting time of scheduled and not scheduled patients. Moreover, the novelty of the objective function herein used is to consider also possible due date violation and patients tardiness.

The problem can be formulated using the following sets of variables:

- $x_{ij}^d \in \{0,1\} = 1$ if patient i is assigned to block j in day $d \in D_j$
- $v_j^d \in \{0,1\} = 1$ if overtime is assigned to block j in day d
- $o_j^d \geq 0$ amount of overtime in block j of day $d \in D_j$

The objective function is formulated as follows:

$$\min \sum_{i \in I} \sum_{j \in J} \sum_{d \in D_j} ([du_i] + [(w_i + d - l_i)^+]u_i) x_{ij}^d + \sum_{i \in I} ([w_i + |D| + 1]u_i + [(w_i + |D| + 1 - l_i)^+]u_i)(1 - \sum_{j \in J} \sum_{d \in D_j} x_{ij}^d), \tag{14.1}$$

where $(w_i + d - l_i)^+ = \max\{w_i + d - l_i, 0\}$ is the patient tardiness. The first term represents the penalty for the scheduled patients. For each patient i the penalty depends on the day of the planning horizon when the surgery is executed. Note that the number of waiting days is weighted by the patient urgency parameter u_i, in order to schedule first the more urgent patients. Besides, a penalty is given if the patient due date is violated, i.e. if $(w_i + d - l_i)^+ > 0$. The second term is associated

with the penalty of the unscheduled patients given by the sum of the total number of waiting days and the patient tardiness. Also for the unscheduled patients the waiting time and the tardiness are weighted by the urgency parameter u_i.

The set of constraints is the following:

$$\sum_{j\in J}\sum_{d\in D_j} x_{ij}^d \le 1 \qquad \forall i \in I \tag{14.2}$$

$$\sum_{i\in I} \tilde{t}_i x_{ij}^d \le \gamma_j + o_j^d \qquad \forall j \in J, \quad \forall d \in D_j \tag{14.3}$$

$$o_j^d \le \delta v_j^d \qquad \forall j \in J, d \in D_j \tag{14.4}$$

$$\sum_{j\in J}\sum_{d\in D_j} v_j^d \le \Delta \tag{14.5}$$

Constraints (14.2) ensure that each patient is operated at most once. Constraints (14.3) are the stochastic capacity constraints for each block forcing either the total time in block j of day d to be lesser than or equal to the maximum available time γ_j or variable o_j^d to be strictly positive. Constraints (14.4) and (14.5) limit, respectively, the amount of overtime for each block j and day d, and the resulting number of overtime blocks to be less than the a priori fixed values δ and Δ.

The deterministic version (DM) of model (1)–(5) is obtained using for each patient i a deterministic surgery time. In particular the mean parameter t_i of the distribution $\mathrm{F}(\tilde{t}_i)$ of the surgery duration random variables is used as expected surgery time, then constraints (14.3) are replaced by

$$\sum_{i\in I} t_i x_{ij}^d \le \gamma_j + o_j^d \qquad \forall j \in J, \quad \forall d \in D_j \tag{14.6}$$

In order to deal with uncertainty in the model (1)–(5) we apply a robust optimization approach [1] which allows exploiting the potentialities of a mathematical programming model without the necessity of generating scenarios.

According to the approach proposed in [1], assuming that random variables (in our case surgery times) may vary in a given interval $[a-\hat{a}, a+\hat{a}]$, uncertainty is dealt with in such a way to guarantee than any solution is feasible if, for each constraint (OR block capacity), at most Γ variables assume their maximum value and all the others assume the central value of the uncertainty interval. To apply in our case, we firstly assume that the "maximum value" we want to protect from is equal to $t_i + \hat{t}_i$, where $\hat{t}_i$ is the standard devition parameter of the $\mathrm{F}(\tilde{t}_i)$ distribution, and that the central value is t_i. Therefore, for each block, a subset S of patients, who require their maximum surgery time, such that $|S| = \Gamma$, is chosen among the patients assigned to the block. Among all the possible subsets, the one having the worst impact on the capacity constraint is chosen, and the solution is forced to be feasible for this subset:

$$\sum_{i\in I} t_i x_{ij}^d + \max_{S\subset I:|S|=\Gamma} \left\{ \sum_{i\in S} \hat{t}_i x_{ij}^d \right\} \le \gamma_j + o_j^d \qquad \forall j \in J, \quad \forall d \in D_j \tag{14.7}$$

The value $\max_{S\subset I:|S|=\Gamma}\left\{\sum_{i\in S}\hat{t}_i x_{ij}^d\right\}$ can be computed for each block j and each day d solving the following Linear Programming model (β^{jd}):

$$(\beta^{jd}) = \max\left(\sum_{i\in S}\hat{t}_i x_{ij}^d\right) z_i \tag{14.8}$$

$$\sum_{i\in I} z_i \le \Gamma \tag{14.9}$$

$$0 \le z_i \le 1 \qquad \forall i \in I \tag{14.10}$$

Let denote with ζ^{jd} the dual variables associated to constraints (14.9) and with π_i^{jd} the dual variables associated to the right hand side of constraints (14.10). The dual of (β^{jd}) problem can be formulated as follows:

$$\min \Gamma\zeta^{jd} + \sum_{i\in I}\pi_i^{jd} \tag{14.11}$$

$$\zeta^{jd} + \pi_i^{jd} \ge \hat{t}_i x_{ij}^d \qquad \forall i \in I \tag{14.12}$$

$$\zeta^{jd}, \pi_i^{jd} \ge 0 \tag{14.13}$$

The optimal values of objective functions (14.8) and (14.11) coincide. Thus, constraints (14.3) can be linearized, by replacing them with (14.16), (14.19) and (14.20), thus obtaining the following complete robust model (RM) formulation:

$$\min \sum_{i\in I}\sum_{j\in J}\sum_{d\in D_j}([du_i] + [(w_i + d - l_i)^+]u_i)x_{ij}^d +$$
$$\sum_{i\in I}([w_i + |D| + 1]u_i + [(w_i + |D| + 1 - l_i)^+]u_i)(1 - \sum_{j\in J}\sum_{d\in D_j} x_{ij}^d) \tag{14.14}$$

$$\sum_{j\in J}\sum_{d\in D_j} x_{ij}^d \le 1 \qquad \forall i \in I \tag{14.15}$$

$$\sum_{i\in I} t_i x_{ij}^d + \Gamma\zeta^{jd} + \sum_{i\in I}\pi_i^{jd} \le \gamma_j + o_j^d \qquad \forall j \in J, \quad \forall d \in D_j \tag{14.16}$$

$$o_j^d \le \delta v_j^d \qquad \forall j \in J, \quad d \in D_j \tag{14.17}$$

$$\sum_{j\in J}\sum_{d\in D_j} v_j^d \le \Delta \tag{14.18}$$

$$\zeta^{jd} + \pi_i^{jd} \ge \hat{t}_i x_{ij}^d \qquad \forall j \in J, \quad \forall i \in I \tag{14.19}$$

$$\zeta^{jd}, \pi_i^{jd} \ge 0 \qquad \forall j \in J, \forall d \in D_j, \quad \forall i \in I \tag{14.20}$$

$$x_{ij}^d \in \{0,1\} \qquad \forall j \in J, \forall d \in D_j, \quad \forall i \in I \tag{14.21}$$

Both for the deterministic (DM) and robust (RM) formulation of model (14.1)–(14.5), a version in which overtime is not allowed, i.e. $\delta = \Delta = 0$, is formulated and denoted in the following as DM-no and RM-no, respectively. In particular, DM-no model is obtained by replacing in model DM constraints (14.6) with $\sum_{i \in I} t_i x_{ij}^d \leq \gamma_j$, while RM-no model is obtained by replacing constraints (14.16) in RM model with $\sum_{i \in I} t_i x_{ij}^d + \Gamma \zeta^{jd} + \sum_{i \in I} \pi_i^{jd} \leq \gamma_j$.

14.3 Experimental Tests

The four formulations above introduced are tested and compared in order to evaluate the applicability of the proposed approach both in terms of computational effort and quality of the obtained solutions. First we tested our models on instances in which the average and standard deviation parameters of the surgery time distributions are taken from real life data. The obtained optimal solutions are compared with respect the number of operated patients, with the aim of evaluating the impact of allowing overtime and of different values of Γ, and thus different levels of required robustness. After this first analysis, the obtained assignments of patients to OR blocks are evaluated on a set of 100 randomly generated scenarios. This second series of computational results is aimed at studying the behavior of the proposed solutions in terms of utilization rate and number of cancelled patients. The instances are generated from two real data based waiting lists partially derived from [13]. Each waiting list is a different collection of patients who wait for surgery and should be scheduled (leading to a set I). The first waiting list is composed by 20 patients ($|I| = 20$), while the second by 40 ($|I| = 40$). For each patient i the urgency class and the elapsed waiting time (w_i) are based on real life data. In the following we refer to an already validated prioritisation system based on five urgency classes [16]. Each patient urgency class is associated with a maximum waiting time l_i expressed in days, that is the maximum number of days that a patients can wait before surgery without deteriorating his/her clinical conditions. The maximum waiting time of each patient i contributes in defining the urgency coefficient (u_i) which represents the speed at which the clinical need is assumed to increase along with the passing of time. In particular, for each patient i the urgency coefficient is stated by the ratio between the maximum waiting time of the least urgent class and his/her maximum waiting time.

According to the data herein used five urgency classes are defined with maximum waiting time l set at 8, 30, 60, 180 and 360 days, respectively, and corresponding urgency coefficients u equal to 45, 12, 6, 2, 1. For each patient the due date parameter dd_i can be derived as $dd_i = l_i - w_i$, by combining urgency and waiting time.

For each waiting list, we generated eight instances by assigning different surgery times to patients. Real data surgery time were derived from [7]. In particular, we selected eight different specialties, for each specialty different types of surgery are given, and for each of these types average surgery time, standard deviation and

percentage of this type over the total number of surgeries are given. Using these percentages we randomly assigned the types of surgery to the set of patients I, obtaining as result an average surgery time (t_i) and a standard deviation ($\hat{t}_i$) for each patient i. Each instance represents the combination of a waiting list and a surgery specialty. Each instance is named n_s, where n is the number of patients ($|I|$), and s is the specialty index used for the surgery times generation.

For the instances with 20 patients we considers an availability of two OR blocks per week, scheduled on Monday and Wednesday, while for the instances with 40 patients we assume to have three blocks per week, on Monday, Wednesday and Friday. Each block j has a capacity 6 h ($\gamma_j = 360$). The maximum allowed overtime per block is equal to 2 h ($\delta = 120$), while during the planning horizon at most $\Delta = \lceil \frac{1}{3}|J| \rceil$ overtime blocks are allowed to use overtime. We consider a 7 days time horizon, corresponding to 1 week.

For each instance we generated 100 different random realizations. In each realization, for each patient i the surgery time $\tilde{t}_i$ is randomly generated using a lognormal distribution F($\tilde{t}_i$) with average surgery time t_i and standard deviation $\hat{t}_i$. In particular, to avoid too short surgery times, we truncate the lognormal distribution at a minimum value equal to $\max(t_i - \hat{t}_i, 30)$. If r_i is the random generated number following the lognormal distribution, the surgery time assigned to patient i will be: $\tilde{t}_i = \max(r_i, t_i - \hat{t}_i, 30)$.

The deterministic and robust models, both with (DM and RM) and without (DM-no and RM-no) overtime availability are tested on the set of instances described above. The models have been implemented with AMPL and solved with CPLEX 12.2.0.0 on a Intel Xeon CPU E5335 (2 quad core cpus at 2 GHz). We set a 2 h time limit. All the considered instances have been solved to optimality within the time limit, while many of them have been solved in shorter computational time. Solving the (DM) model requires few seconds, while the computational time may significantly vary for the robust version. However, the required CPU time is never above 1 h and a half. The objective function increases with the increasing value of Γ, as a more robust solution is required and therefore a larger subset of patients requires the maximum surgery time. However the objective function is constant after a certain value, that can differ for different instances. As the results tend to stabilize for $\Gamma \geq \Gamma^*$, or at least the variations are meaningless, we report the values only for $\Gamma \leq \Gamma^*$. The stabilized values are denoted with '-'.

Allowing overtime reduces the overall penalty for both models (DM and RM).

In Table 14.1 the behavior of the optimal solutions in terms of operated and not operated patients is given. For each instance in the first column (max) the upper bound of the number of patients who can be operated is also reported. Such value is computed by maximizing the number of operated patients without considering patient penalities, while guaranteeing that the available time is not exceed. In the following columns three values are reported for each value of Γ: the number of operated patients (y), the number of patients operated after their due date (y>) and the number of non operated patients whose due date has been exceeded (n>).

Results show that if robustness is not required (DM and DM-no) the number of operated patients is quite close to the maximum possible. The number of operated

Table 14.1 Operated patients

					$\Gamma = 1$			$\Gamma = 2$			$\Gamma = 3$			$\Gamma = 4$			$\Gamma = 5$		
		y	y	n	y	y	n	y	y	n	y	y	n	y	y	n	y	y	n
			>	>		>	>		>	>		>	>		>	>		>	>
Inst	max	(DM-no)			(RM-no)														
20_1	8	8	5	5	6	4	6	6	3	7	4	4	7	4	4	7	–	–	–
20_2	9	8	5	4	8	5	5	7	4	5	6	5	5	6	5	5	–	–	–
20_3	7	7	5	5	6	4	6	5	4	6	5	4	6	5	4	6	–	–	–
20_5	10	9	7	2	8	7	3	8	6	4	6	5	5	6	5	5	–	–	–
20_7	12	11	7	2	10	7	2	9	7	2	10	6	3	9	6	3	–	–	–
20_8	8	7	5	5	7	5	5	6	4	5	6	4	6	5	4	6	–	–	–
20_9	8	7	5	5	7	5	6	7	5	6	6	4	6	6	5	6	–	–	–
20_10	8	6	5	5	7	5	6	6	5	6	5	5	6	5	5	6	–	–	–
40_1	14	12	8	5	9	7	6	7	6	7	5	5	8	5	5	8	–	–	–
40_2	15	13	11	2	11	10	3	12	8	4	10	9	4	10	9	4	–	–	–
40_3	12	10	10	3	9	8	5	8	8	5	7	7	6	7	7	6	–	–	–
40_5	15	14	11	0	13	12	1	12	12	1	10	10	3	9	9	4	–	–	–
40_7	20	18	11	1	17	12	1	15	11	1	15	11	1	13	10	1	–	–	–
40_8	14	10	9	4	9	8	5	10	7	5	9	7	5	8	7	6	–	–	–
40_9	13	10	9	4	11	8	5	8	8	5	8	7	6	8	7	6	–	–	–
40_10	13	10	8	5	9	8	5	9	7	6	7	7	6	7	7	6	–	–	–

(continued)

Table 14.1 (continued)

					$\Gamma = 1$			$\Gamma = 2$			$\Gamma = 3$			$\Gamma = 4$			$\Gamma = 5$		
		y	y	n	y	y	n	y	y	n	y	y	n	y	y	n	y	y	n
			>	>		>	>		>	>		>	>		>	>		>	>
Inst	max	(DM)			(RM)														
20_1	9	8	4	5	8	4	6	6	4	6	5	5	6	5	5	6	5	5	6
20_2	10	9	6	3	9	6	4	7	5	4	8	4	5	8	4	5	7	5	5
20_3	8	8	6	4	7	5	5	6	5	5	6	5	5	6	4	6	6	4	6
20_5	11	11	9	0	10	7	1	9	8	2	8	7	3	8	5	4	7	6	4
20_7	13	12	8	1	12	7	2	12	7	2	11	7	2	10	7	2	10	7	2
20_8	9	8	4	4	7	5	5	6	4	5	6	5	5	7	4	5	7	4	5
20_9	9	8	5	4	8	4	5	6	5	5	8	5	6	7	4	5	7	4	5
20_10	9	8	5	5	8	4	5	8	5	6	7	4	6	7	5	6	6	5	6
40_1	15	13	8	4	9	8	5	9	7	6	6	6	7	6	6	7	6	6	7
40_2	17	14	11	1	12	11	2	12	10	3	11	8	4	10	10	3	10	10	3
40_3	13	10	10	3	10	9	4	9	9	4	9	8	5	8	8	5	8	8	5
40_5	17	16	11	0	15	12	0	14	11	1	12	11	2	11	10	2	11	10	3
40_7	21	20	10	1	19	11	1	18	11	1	17	11	1	16	10	1	15	10	1
40_8	15	12	9	3	11	8	4	10	8	5	9	9	4	9	8	5	9	8	5
40_9	15	11	10	3	11	8	5	10	8	5	9	8	5	9	8	5	9	8	5
40_10	15	10	9	4	10	9	4	9	8	5	9	7	6	8	8	5	8	7	6

patients usually decreases with the increasing values of Γ, while it increases if overtime is allowed. Note that the set of operated patients is different for different values of Γ, and, in general, the set of operated patients for $\Gamma = n$ is not a subset of those operated for $\Gamma = n - 1$: the subsets may be completely different. Overtime always improves the solutions: it allows either to increase the number of operated patients or to reduce the number of patients operated after their due date.

The behavior of the proposed solutions on a set of 100 randomly generated scenarios is described in Table 14.2 and in Table 14.3. In particular in Table 14.2 the operating room utilization rate is given, while in Table 14.3 the average minimum number of cancelled patients for each block is reported. Concerning the case in which no overtime is allowed, results show that the operating rooms are well exploited in case no robustness is required: the utilization rate is about 100% for DM-no case, while the rate decreases when the value of Γ increases, as longer surgery time are considered for at least a subset of patients. The rate may fall to about 70% for most of the instances, but it is always above 50%. Allowing overtime increases significantly the utilization rate. On the other hand, with small values of Γ the number of cancelled patients is significant, while it decreases if the value of Γ increases. The selected assignment is almost completely respected for $\Gamma \geq 4$. Note that introducing overtime does not have a strong impact on the number of cancelled patients, although this value decreases a little for all instances (Table 14.3).

By properly tuning the value of Γ a tradeoff between the utilization rate and the number of cancelled patients can be obtained. In fact, from the hospital management point of view, smaller values of Γ are preferable, as they guarantee a higher utilization rate. However, such values impact on the solution robustness, as it is shown by the higher number of cancelled patients. From the perceived quality of service point of view, instead, higher values of Γ are better as they guarantee that the OR schedule is respected and no patients must be delayed from the plan and rescheduled. Besides, it is worth noting that an utilization rate below 100% means that there is some operating room capacity not utilized. Such available OR time, rather than being a loss for the system, could allow to manage emergency cases and/or reschedule cancelled patients, without changing the planned OR schedule.

14.4 Conclusions and Further Developments

We presented an approach based on robust optimization to deal with the Operating Room Planning problem in which surgery times are uncertain parameters. Waiting time, urgency and due date of patients are considered. The goal of the problem is to minimize the penalty associated with waiting time and tardiness of patients. The possibility of allowing overtime is considered as well, and its impact is evaluated. The proposed models have been tested on a set of real life based instances. The impact of different levels of required robustness is compared. Besides, we tested the obtained solutions on a set of randomly generated realistic scenarios assuming lognormal distributions for surgery duration. Results show that the proposed robust

Table 14.2 Utilization rate in percentage

		(RM-no)							(RM)					
Inst	(DM-no)	1	2	3	4	5	6	(DM)	1	2	3	4	5	6
20_1	101.72	69.77	62.06	54.54	–	–	–	118.62	90.62	68.71	61.81	–	–	–
20_2	99.17	83.53	77.03	69.73	69.73	69.73	70.08	120.31	102.64	92.31	81.83	82.64	82.09	–
20_3	97.74	83.56	72.16	–	.	–	–	119.97	103.03	88.04	83.04	83.30	–	–
20_5	94.30	84.93	80.11	63.02	–	–	–	120.12	106.79	94.00	83.03	78.23	–	–
20_7	95.81	89.20	82.94	82.82	76.25	74.36	70.28	116.47	103.44	102.27	94.52	87.54	88.05	88.15
20_8	99.86	87.62	78.86	77.04	68.91	–	–	117.64	98.73	96.89	88.73	86.53	–	–
20_9	92.53	79.84	79.84	68.51	69.28	69.28	69.44	115.37	101.53	82.78	90.52	88.44	–	–
20_10	91.42	80.82	70.95	61.30	61.30	.	–	115.40	105.28	91.47	78.03	79.20	70.95	–
40_1	102.11	73.82	64.80	49.10	–	–	–	119.42	83.51	76.02	57.78	58.30	58.30	57.67
40_2	101.11	84.85	80.19	70.75	70.62	70.67	70.67	114.10	97.49	90.46	83.65	79.85	79.28	79.28
40_3	97.83	81.41	73.53	69.86	69.86	69.86	69.86	112.55	95.16	84.82	80.47	76.64	77.04	77.04
40_5	96.34	87.38	81.47	67.87	62.17	62.10	61.92	113.29	103.29	94.70	79.28	73.23	73.50	73.23
40_7	99.94	92.77	85.18	83.18	75.70	74.59	70.13	114.21	106.03	98.84	93.43	87.64	83.08	83.29
40_8	94.71	86.47	81.04	75.56	76.36	75.92	75.92	114.06	101.87	92.24	87.51	85.74	83.29	83.25
40_9	91.39	86.28	69.25	62.69	62.74	–	–	106.27	87.03	81.10	75.00	–	–	–
40_10	92.53	79.84	79.84	68.51	69.28	69.28	69.44	109.93	98.18	83.82	75.91	74.72	70.32	70.44

Table 14.3 Average number of cancelled patients per block

		(RM-no)								(RM)						
Inst	(DM-no)	1	2	3	4	5	6	7	(DM)	1	2	3	4	5	6	7
20_1	0.47	0.13	0.13	0.03	–	–	–	–	0.37	0.25	0.11	0.06	–	–	–	–
20_2	0.32	0.19	0.07	0.06	–	–	–	–	0.32	0.19	0.05	0.07	0.03	–	–	–
20_3	0.19	0.02	–	–	–	–	–	–	0.17	0.01	0.01	0.05	0.00	–	–	–
20_5	0.25	0.07	0.08	0.00	0.00	0.01	0.00	–	0.28	0.16	0.06	0.02	0.02	0.00	–	–
20_7	0.21	0.12	0.05	0.03	0.03	0.02	0.00	0.02	0.27	0.05	0.07	0.02	0.03	0.03	0.03	0.02
20_8	0.15	0.11	0.05	0.02	0.00	–	–	–	0.20	0.07	0.02	–	–	–	–	–
20_9	0.25	0.05	0.05	0.02	0.00	–	–	–	0.26	0.10	0.05	0.03	0.01	–	–	–
20_10	0.29	0.10	0.04	0.01	–	–	–	–	0.23	0.15	0.07	0.07	0.03	0.01	–	–
40_1	0.26	0.07	0.08	0.01	0.00	0.00	0.01	–	0.20	0.13	0.05	0.07	0.01	0.00	0.00	0.02
40_2	0.41	0.16	0.08	0.04	0.05	0.05	0.05	0.06	0.36	0.14	0.08	0.04	0.02	0.04	0.05	0.05
40_3	0.17	0.05	0.03	0.01	–	–	–	–	0.15	0.05	0.04	0.06	0.01	0.00	0.00	0.01
40_5	0.41	0.26	0.12	0.04	0.00	–	–	–	0.34	0.25	0.12	0.05	0.02	0.01	0.02	0.01
40_7	0.43	0.16	0.08	0.06	0.00	0.00	0.00	0.00	0.30	0.15	0.09	0.05	0.04	0.02	0.02	0.03
40_8	0.26	0.08	0.03	0.00	0.02	–	–	–	0.27	0.10	0.05	0.04	0.03	0.03	0.01	0.02
40_9	0.25	0.07	0.03	0.00	0.01	–	–	–	0.24	0.09	0.05	0.03	–	–	–	–
40_10	0.65	0.31	0.12	0.06	–	–	–	–	0.59	0.24	0.19	0.07	0.07	0.07	0.08	0.07

models can be used, as the required computational time is compatible with the weekly schedule. Furthermore, the obtained robust solutions behave well when tested on the scenarios: in fact they reduce the number of cancelled patients w.r.t. the deterministic and non robust case. Although the robust solutions may produce an utilization rate below the 100%, nevertheless by properly tuning the value of Γ, which represents the level of robustness required, or the degree of risk accepted, a good tradeoff between hospital productivity and quality of service provided to patients can be achieved. Future works will be devoted to perform a more extensive computational analysis by varying the number of patients in the waiting list to be operated on and considering different values of overtime. As future development, the impact of different objective functions has to be studied, as well as an online procedure which re-assigns the patients to be rescheduled and deals with the emergency cases.

References

1. Bertsimas, D., Sim, M.: The price of robustness. Oper. Resour. **52**(1), 35–53 (2004)
2. Cardoen, B., Demeulemeester, E., Beliën, J.: Operating room planning and scheduling: a literature review. Eur. J. Oper. Resour. **201**, 921–932 (2010)
3. Denton, B., Viapiano, J., Vogl, A.: Optimization of surgery sequencing and scheduling decisions under uncertainty. Health Care Manag. Sci. **10**, 13–24 (2006)
4. Denton, B., Miller, J., Balasubramanian, H., Huschka, T.: Optimal allocation of surgery blocks to operating rooms under uncertainty. Oper. Resour. **58**, 802–816 (2010)
5. Dexter, F., Ledolter, J.: Bayesian prediction bounds and comparisons of operating room times even for procedures with few or no historic data. Anesthesiol. **103**, 1259–1267 (2005)
6. Guerriero, F., Guido, R.: Operational research in the management of the operating theatre: a survey. Health Care Manag. Sci. (2010). doi:10.1007/s10729-010-9143-6
7. Hans, E., Wullink, G., van Houdenhoven, M., Kamezier, G.: Robust surgery loading. Eur. J. Oper. Resour. **185**, 1038–1050 (2008)
8. Magerlein, J., Martin, J.: Surgical demand scheduling: a review. Health Serv. Resour. **13**, 418–433 (1978)
9. May, J., Strum, D., Vargas, L.: Fitting the lognormal distribution to surgical procedure times. Decis. Sci. **31**(1), 129–148 (2000)
10. Min, D., Yih, Y.: Scheduling elective surgery under uncertainty and downstream capacity constraints. Eur. J. Oper. Resour. **206**, 642–652 (2010)
11. Spangler, W., Strum, D., Vargas, L., Jerrold, H.: Estimating procedure times for surgeries by determining location parameters for the lognormal model. Health Care Manag. Sci. **7**, 97–104 (2004)
12. Tànfani, E., and Testi, A.: A pre-assignment heuristic algorithm for the master surgical schedule problem (mssp). Ann. Oper. Resour. **178**(1), 105–119 (2010)
13. Tànfani, E., Testi, A., Alvarez, R.: Operating room planning considering stochastic surgery durations. Int. J. Health Manag. Inf. **1**(2), 167–183 (2010)
14. Testi, A., Tànfani, E., Torre, G.: A three-phase approach for operating theatre schedules. Health Care Manag. Sci. **10**, 163–172 (2007)
15. Tyler, D., Pasquariello, C., Chen, C.: Determining optimum operating room utilization. Anesth. Analg. **96**(4), 1114–1121 (2003)

16. Valente, R., Testi, A., Tànfani, E., Fato, M., Porro, I., Santo, M., Santori, G., Torre, G. Ansaldo, G.: A model to prioritize access to elective surgery on the base of clinical urgency and waiting time. BMC Health Serv. Resour. **9**(1), 1 (2009)
17. van Oostrum, J., Bredenhoff, E., Hans, E.: Suitability and managerial implications of a master surgical scheduling approach. Ann. Oper. Resour. **178**(1), 91–104 (2010)
18. van Oostrum, J., van Houdenhoven, M., Hurink, J., Hans, E., Wullink, G., Kazemier, G.: A master surgical scheduling approach for cyclic scheduling in operating room departments. OR Spectr. **30**, 355–374 (2008)

Chapter 15
The Methodological Approach to Process Analysis for Robotic Surgical Procedures: The Experience of SAFROS and I-SUR Projects

Riccardo Dodi, Elettra Oleari, and Alberto Sanna

Abstract This work presents two methodological approaches followed for two distinct contexts of surgical robotics. The IRIS Unit of Fondazione Centro San Raffaele (Milano, Italy) is involved in two research projects related to robotic surgery, dealing with different purposes, procedures and technologies, thus requiring different approaches to knowledge formalization and process analysis. The first project, SAFROS – Patient Safety in Robotic Surgery, aims to improve patient safety for robotic surgery and a *systemic* approach has been adopted, in order to take into account several aspects related to the surgical procedure, from the device to the process itself and the whole environmental organization. The second project, I-SUR – Intelligent Surgical Robotics, aims at automatizing three basic surgical gestures with specific procedural constraints and targets. A *goal-based* approach has been chosen to analyze the process, extrapolate the operational workflow and setting the requirements for the underlying technological system. Both in SAFROS and I-SUR, safety for the patient and acceptability for the surgeons, considered as "final users" of such innovations, have taken a key role for the approach to process improvement.

15.1 Introduction

Nowadays, hospitals represent structures that are ever increasingly complex and are expected to provide a diversified range of services. Hospital functions are wide ranging and must evolve simultaneously to a world that develops always faster from the technological point of view [1].

R. Dodi (✉) • E. Oleari • A. Sanna
Fondazione Centro San Raffaele, via Olgettina 60, Milano, Italy
e-mail: dodi.riccardo@hsr.it; oleari.elettra@hsr.it; sanna.alberto@hsr.it

A. Matta et al. (eds.), *Proceedings of the International Conference on Health Care Systems Engineering*, Springer Proceedings in Mathematics & Statistics 61,
DOI 10.1007/978-3-319-01848-5_15,

The present work will describe the experience gained in surgical process analysis by the IRIS Unit – eServices for Life and Health (Fondazione Centro San Raffaele, Milano) within two different EU co-funded research project. During last years, IRIS Unit has worked on several projects involving IT for wellbeing and healthcare. The aim is to develop technological services to provide and manage real-time and personalized health data. The ultimate goal is to disseminate awareness toward the personal health status and to develop a proper healthcare culture to improve it and improve hospital, clinical and surgical processes.

Activities belonging to the robotic surgery area hold a large spectrum of field of research. For instance, new ICT methods have been developed to continuously detect the operating room activities, for the realistic simulation of surgical operations on anatomical models, monitor the surgical robot performance, update the organ position and shape and to identify potential safety risks with comparisons between planned and real situations. Another important activity is to assess the applicability of the current training methods for surgery to robotic surgery and develop new specific training methods for this last class of interventions.

As previously said, both the considered projects are focused on one of the fastest growing fields of interest in surgery: robotic assisted procedures. The project for "Patient SAFety in RObotic Surgery" (SAFROS) addresses the design of innovative tools and definition of methods capable to improve patient safety, not only focusing on the technological aspect but also embracing the entire surgical workflow. The project for "Intelligent SURgical Robotics" (I-SUR), instead, develops general methods for cognitive surgical robots that combine sensing, dexterity and cognitive features in order to carry out *autonomously* simple surgical actions, namely puncturing, cutting and suturing.

The key role of IRIS Unit in both projects emprises both the analysis of the surgical procedures taken into account and the collection of the requirements for the innovative technology chosen by the consortium to improve the current state-of-the-art. Deciding the methodological approach to face with the issues introduced by the scopes of the project is a crucial step. In fact, it is mandatory to have a clear idea on how to set up the work and optimize the extrapolation of the results. In this context, a relevant role is taken by the translation of the surgical knowledge in a logic/mathematic formalization in order to be understood by non-medical-educated professionals and then to be easily re-translated as input for a robotic system.

The Sect. 15.2 of this work presents the objectives of the two projects, which approaches and results are better detailed separately in the following Sects. 15.3 and 15.4. At the end, conclusions are drawn and possible future perspective discussed in Sect. 15.5.

15.2 Objectives

The requirements analysis plays a key role in both projects: thanks to this step the most significant goals can be outlined and are identified the proper addressing

features of project's solutions. Because of this, exploiting a modeling approach is especially helpful during this first phase of analysis, in order to outline a consistent and well-structured methodological framework.

Concerning SAFROS (European Union Seventh Framework Program FP7/2007–2013 under grant agreement n. 248960), the modeling aspect reflects the re-engineering process of a surgical procedure which is studied in risks terms. Thanks to such an approach, the risks embedded in the operational workflow are identified and ranked. This lays the basis for the implementation of project's solutions capable to guarantee an improvement in safety care of the patient, addressing proper safety criteria through the introduction of new technologies and a methodological safety culture [2].

For I-SUR (European Union Seventh Framework Program FP7/2007–2013 under grant agreement n. 270396), the approach is slightly different because the automation of the three surgical actions required the identification of the development of a new technological and methodological system acting from the preoperative phase to the completion of the task and the possible critical aspects of such innovation.

15.3 SAFROS: Modeling Robotic Surgical Procedures Through the Risk Analysis Approach

The primary goal of SAFROS is to develop new methods and tools to improve patient safety. Researches carried out during the project aimed at identifying proper metrics in order to assess the safety level achieved with project solutions and pointing out methods capable to correctly address these requirements. The methodology applied to support this analysis is divided into three different levels: *product* safety analysis, *process* safety analysis and *organizational* safety analysis. This framework is an effective tool to achieve the project objective considering the developed technologies first singularly (product safety), then widening the scope of the research towards their integration into a surgical workflow (process safety) and finally studying their impact onto an organizational level (organizational safety) [2]. What follows is the description of the modeling applied in the context of the process safety analysis of a reference robotic surgical procedure: Robotic-assisted laparoscopic radical prostatectomy (RALP) [3, 4].

The process safety methodological level consists of a systemic approach which evaluates the effects of the interaction of SAFROS products and their impact on the various phases of a robotic surgical intervention, also including the operating room environment (i.e.: the surgical staff) and patient related information as influencing factors. In fact, with the increasing deployment and sophistication of equipment within robotic surgery, technological failures are more likely to occur and are added to the inherent risks of the procedure. The scope of the analysis is to identify and prioritize the entirety of process-related criticalities, to provide a basis to discuss the ways in which solutions could directly impact on the surgical process

Table 15.1 Schema of the simplified Failure Mode and Effect Analysis applied

Activities	Related risk [RR]	Cause	Effects	Criticality index [CI]		
				O Occurrence	S Severity	O × S

Occurrence \ Severity:	No damage (1)	Minor damage (2)	Medium damage (3)	Serious damage (4)	Really serious damage (5)
Remote (1)	1	2	3	4	5
Occasional (2)	2	4	6	8	10
Probable (3)	3	6	9	12	15
Frequent (4)	4	8	12	16	20

= Accettable Risk = Medium Risk
= Low Risk = High Risk

Fig. 15.1 Risk matrix (Adapted from [2])

and mitigate the found risks. Among the different existing techniques for risk management in healthcare, a simplified version of the Failure Mode and Effect Analysis[1] (FMEA) has been chosen. The analysis was carried out by a team composed of a facilitator and various process owners, in order to benefit from added values derived from different professionals profiles and experiences. When needed, the work was supported by site visits during real interventions and interviews with OR personnel (e.g., expert robotic surgeons, anesthetists, nurses). The starting point was a detailed analysis of the pre-operative and intra-operative phases of the RALP procedure and of all the most important Related Risks (RR). Then, the cause and effects linked to each critical step were listed (see Table 15.1).

After this, a Criticality Index (CI) of each outlined risk was obtained by multiplying the estimated frequency of occurrence (O, 4 point rating scale) by the expected severity of the damage towards the patient (S, 5 point rating scale) and **framed** in a Risk Matrix as exemplified in Fig. 15.1 [2].

The analysis led to an accurate schematization of the robotic procedure workflow (see Fig. 15.2) and the definition of the risks related to its pre-operative and intra-operative phases with particular attention given to causes and effects of each RR.

In the following Table 15.2 is summarized an extract of the results of the described research: for the most critical surgical steps the embedded risks are described and evaluated according to their resulting criticality index (CI) []. The results derived from the procedural risks selection and ranking allowed to identify a group of safety indicators strictly linked to the surgical scenario and widely

[1]FMEA is a proactive risk assessment tool used to identify potential vulnerabilities in complex, high-risk processes and to generate remedial actions before the process results in adverse events [5].

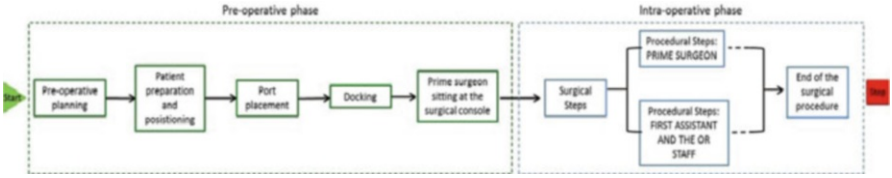


Fig. 15.2 The RALP workflow

applicable to different robot-assisted procedures. To name a few: accuracy of task execution, overall operative time, adequacy of the internal workspace, the need of coordination skills, ...(for more details see [2]). Combining these medical indicators with the merely technical specifications of the project innovative tools, allowed on the one hand to mirror the surgical safety requirements in a safety driven technological design of the SAFROS solutions. On the other, in such a way was moved the first step towards a comprehensive quantitative and qualitative assessment of the reachable improvement of patient safety.

15.4 I-SUR: A Goal-Based Approach Targeted to Automated Surgery

If designing and controlling a robotic tool for surgery is a very highly demanding task, the automation of a surgical gesture introduces even more challenging issues concerning the required imaging, sensing and control system to be applied. The first step to approach this problem regards the collection of all the information about the surgical knowledge of the target gestures (puncturing, suturing and cutting) in order to formalize the environment and the tasks themselves in mathematical models and develop a reasoning mechanism for the identification of pre and post conditions of the actions and their execution [6]. The procedural methodology is displayed in Fig. 15.3. The Puncturing task has been selected as the most promising one (in terms of feasible mid-term results and efficacy) and thus better analysed and modelled. As reference case study, the consortium has chosen the Percutaneous Cryoablation of Kidney Tumours, which consists of freezing the neoplastic tissue with gas injected by a specific probe [7].

The *Goal Model* [8] has been chosen as mathematical formalization of the selected procedures because it facilitates the translation from surgical to technical requirements in terms of state diagram, operative sequence, identification and managing of critical steps, design of sensing and monitoring activities. Interviews with worldwide renowned professional surgeons were conducted in Milano (San Raffaele Hospital) and Verona (General Hospital G. B. Rossi), to collect medical information about the three tasks. Then the Goal Model has been created and validated for each surgical action. Starting from all the elements of this model

Table 15.2 The most critical surgical steps the embedded risks for RALP

Step	Related risks	Causes	Effects	CI
PRE-OPERATIVE				
Pre-operative planning	Not adequate evaluation of the surgical strategy	Incomplete or inaccurate evaluation of the pre-operative tests	Prolonged surgery time Deviation from the surgical plan Not prompt reaction of the surgeon to complications or unforeseen events	6.60
Patient preparation and positioning	Uneven distribution of pressure points Limbs extra-rotation Slipping during table movements	Inadequate mattress or gel pads Not correct positioning of the patient or of the relative supports Inadequate securing of the patient through straps	Pressure induced sores Patient discomfort after surgery Lesions due to nerves compressions Cardiovascular or respiratory problems	9.65
Port placement	Access-related injuries	Not proper trocar positioning Not transillumination of the zone during the trocar insertion Surgeon's inattention	Bleeding Prolonged surgery time	5.77
Docking	Collision between robotic arms/external interferences between robot and human limbs or OR staff	Not proper trocar positioning and not sufficient space between robotic arms	Mechanical failure and breakage	5.24
		Incorrect position of the robot/patient	Prolonged surgery time	

INTRA-OPERATIVE				
	Lack of communication between the first surgeon and the assistant, especially in non-standardized procedure	Improper or missed training on human factors themes in robotic surgery	Difficulties or misunderstandings during surgery	5.77
Procedure-related steps: First surgeon	Inadequate Exposure of operative field (Structure outside instrument reach)	Incorrect positioning of the trocar and patient	Difficulties during surgery (e.g. not clear working field, unrecognised damage)	8.32
		Pneumoperitoneum not achieved or difficult to maintain		
	Surgical errors (procedure-specific)	Surgeon's skills (training and experience)	Damage to anatomical structure	6.21
			Prolonged surgery	
Procedure-related steps: first assistant and the OR staff (nurses, anaesthetists, technicians)	Improper handling of the organs	First assistant's skill in laparoscopy or insufficient confidence with robotic devices	Damage to anatomical structure	8.96
	Improper changing of the instruments	Miss communication between OR staff	Prolonged surgery time	3.63

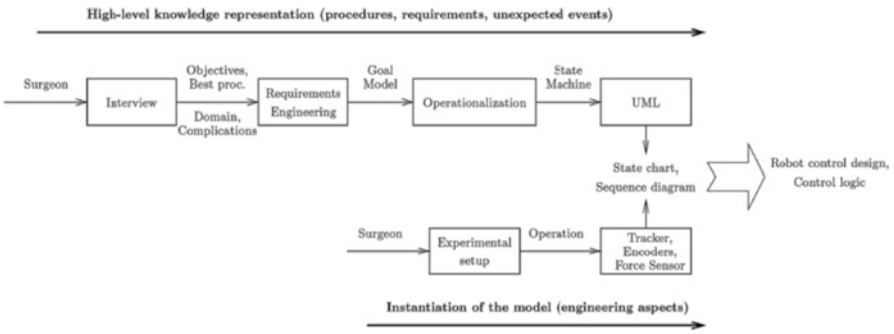


Fig. 15.3 The I-SUR methodology chosen for the requirement collection

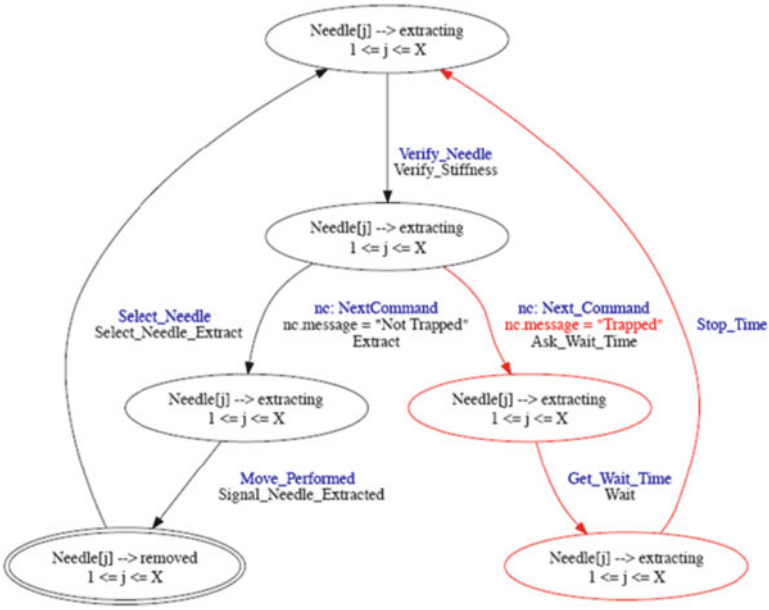


Fig. 15.4 State diagram for the "Needle extraction" macro-phase of the Puncturing task

[9, 10], i.e. *Domain* (entities and events), *Goals* (the main objectives of the surgical actions), *Operations* and *Adaptations*, a State Diagram has been inferred to synthesizes all possible ways through which the operations can be performed to satisfy the defined goals and sub-goals. Four sequential macro-phases can be identified, namely *Initialization* (planning phase of surgical action and tool trajectories starting from patient-specific characteristics), *Insertion* (heading of the needle to the skin, penetration and heading to the target point in the tumor), *Cryoablation* (standard cycles of freezing-thawing) and *Extraction* (getting the needle out and reach the initial position); a schematization of the last phase is represented in Fig. 15.4.

By analyzing the state diagram, it has been possible to gain the medical knowledge embedded in the task in order to obtain the mechanisms of surgical reasoning. These are the bases of the planning action and allow the surgeon to decide how to execute the operation, being aware of possible problems or errors which occur during the intra-operative execution and taking the respective adaptations to face them.

The a priori medical knowledge and the analysis of the state diagram allow identifying some constraints and parameters about surgical action execution to be monitored and controlled. These data will be acquired through the sensing system during the pre- and intra-operative phases: the sensing devices have to cooperate to provide the robotic system with accurate information in order to build a feasible, optimal plan of the required operation and autonomously execute it. In Table 15.3, any information to be controlled by the sensing system during the robotic execution of Percutaneous Kidney Cryoablation is collected. For each event to detect, or condition to monitor, the needed quantitative entities has been identified and a real or Boolean variable can be derived by combining them. Tolerances and thresholds for such variables have been collected during structured interviews and virtual/practical experiments involving expert surgeons and radiologists. These data will feed the reasoning mechanism, controlling the state transition of the system, branches and adaptation paths.

Moreover, analyzing the state diagram some branches can be found, i.e. deviations from the nominal behavior, which represent dangerous events of the procedure and need appropriate adaptation paths to return back to the nominal one. For each branch, an event regulating the transition to a state or another one can be identified. Such event generation depends on a specific condition or evolution of the system, i.e. a specific configuration of a definite set of variables.

The definition of the thresholds for the whole procedure represents the linchpin of the reasoning for preoperative assessment. The interviews with surgical staff were useful to extrapolate these values, which can be exploited as constraints for the intra-operative reasoning module. Since most of procedures require standard thresholds, sometimes the clinician has to adapt the therapy because of the peculiar anatomic configuration of the patient. For instance, the standard accuracy required for needle placement is around 3–4 mm, but in case of small lesions (e.g. ∅ 10 mm) this value should be inferior.

The strong potentiality of this approach is represented by the development of a reliable mathematic formalization starting from the analysis of the key points of the procedure, i.e. what are the objective and how to reach them, in terms of efficient surgical outcomes and patient safety. Then, a feasible operational model has been obtained. Thanks to its modularity, simply adapting the therapy-specific macro-phase (*Cryoablation* module) it can be potentially extended and applied to several procedures similar to the chosen case study, e.g. biopsies, nephrostomies, treatments of renal cysts and so on; in general, to any puncture, injection, extraction or drainage performed with a rigid needle that requires high accuracy, leading to a reduced operative time (e.g. emergency) and bleeding risk through a minimally invasive approach. Finally, a patient-specific procedure can be obtained through the instantiation of the operational model by analysing the pre-operative data. Surgeons,

Table 15.3 Procedural constraints and parameters identified for Kidney Cryoablation

Event	Entity (attribute)	Variable	Bound
Target missed	Tumor center	Distance_to_target	3–4 mm
	Needle tip		
Forbidden region touched	Needle tip	Distance_to_fr	10 mm
	FR surface		
Tumor not covered	Tumor volume	Coverage_percentage	0%
	Iceball volume (math model)		
Needle trapped	Needle stress	Needle_trapped	(Yes/no)
	Needle temperature		
	Time		
Move robot from "rest" to "init" and vice versa	Needle tip	Distance_to_patient	50 mm
	Skin surface		
Minor adjustments after insertion	Needle tip	Distance_to_skin	10 mm (tool-dependent)
	Skin surface		
Insertion area	Skin surface	Insertion_error_pos	5 mm
Insertion angle	Needle orientation	Insertion_error_deg	5°
	Longitudinal body axis		

radiologists and clinicians play a key role in the whole process and the development of the technological device needs to run in parallel with a strong, continuous interaction based on input-validation feedbacks.

15.5 Discussions and Future Perspectives

In the previous sections, two different methodological approaches to face with innovative technologies for robotic surgery have been analysed, reporting the experience gained in two different projects. The modelling approach has been a key element to complement the analysis both from a global point of view, such in the case of the entire surgical process, and from a more detailed one as for the automation of a surgical gesture.

Regarding SAFROS, the methodology developed and the applied risk analysis have allowed a systemic sectioning of the entire process of study. In this way, a holistic view of the objectives of the project has been guaranteed and it has been possible to develop well addressing solutions, capable to overcome the current

limits and inherent risks of both technology and procedures. As yet anticipated in [2], the extendibility of the proposed modeling approach to other surgical or medical specialties is one of the most promising prospects resulting from this research project. This hypothesis is also supported by the satisfactory results of the present work and by the flexibility inherent of modeling through analysis of risk.

For what concerns I-SUR, instead, a goal-based approach has been a useful tool to analyze the as-is process and obtain a preliminary version of the state diagram which the robotic system will be asked to follow. This approach led to the collection of a set of requirements, constraints and other parameters to instantiate such model and obtain a safe, feasible and reliable automatic execution. Up to now (Project Month n. 24), the consortium used this information to design system components, i.e. robotic tool, control algorithms, sensing system, surgical interface and artificial organs; once integrated together, an assessment and validation of this new technology will be possible.

Patient safety is one of the main issues in the development of technological solutions in the surgical world, and in robotics for surgery as well. A correct methodology and modeling approach allows integrating safety constraints and requirements from the beginning of the research process, making them an integral part of the objectives to be pursued and the corresponding solutions to be implemented.

Acknowledgments Researches leading to the described results have been funded by the European Union Seventh Framework Program FP7/2007–2013 under grant agreement n. 248960 (SAFROS) and n. 270396 (I-SUR). Authors wish also to acknowledge the surgical staffs of San Raffaele Hospital (Milan) and General Hospital G. B. Rossi (Verona) for their support and valuable cooperation to the projects' research activities.

References

1. Kumar, S.: Modeling hospital surgical delivery process design using system simulation: optimizing patient flow and bed capacity as an illustration. Technol. Health Care **19**, 1–20 (2011)
2. Morandi, A., Verga, M., Oleari, E., Gasperotti, L., Fiorini, P.: A methodological framework for the definition of patient safety measures in robotic surgery: the experience of SAFROS project. Intelligent autonomous systems 12. Adv. Intell. Syst. Comput. **194**, 155–164 (2013)
3. Australian Institute of Health and Welfare (AIHW) and Australasian Association of Cancer Registries (AACR): Cancer in Australia: an overview (2008). Cancer series no. 46. Cat. no. CAN 42. AIHW, Canberra (2008)
4. Murphy et al.: Robotic technology in surgery: current status in 2008 (2008)
5. Joint Commission on Accreditation of Healthcare Organizations –JCAHO – Standard LD 5.2
6. Bonfè, M., Boriero, F., Dodi, R., Fiorini, P., Morandi, A., Muradore, R., Pasquale, L., Sanna, A., Secchi, C.: Towards automated surgical robotics: a equirements engineering approach. In: 4th IEEE International Conference on Biomedical Robotics and Biomechatronics (2012)
7. Permpongkosol, S., Nielsen, M., Solomon, S.: Percutaneous renal cryoablation. Urology **68**(1 Suppl.), 19–25 (2006)

8. Baresi, L., Pasquale, L., Spoletini, P.: Fuzzy goals for requirements-driven adaptation. In: Proceedings of the 18th International Requirements Engineering Conference. IEEE Computer Society, pp. 125–134 (2010)
9. Baresi, L., Pasquale, L.: Live goals for adaptive service compositions. In: Conference SEMS'10. Cape Town, 2–8 May 2010
10. Baresi, L., Pasquale, L.: Adaptation goals for adaptive service-oriented architectures

Chapter 16
A Whole-System Approach to Identify the Sources of Variation in Patient Flow

Nasim Arbabzadeh, Mohsen A. Jafari, and Kian Seyed

Abstract The main objective of this paper is to develop a quantitative framework to identify the main sources of variation in patient flow. Since 1983, under Health Care Financing Administration (HCFA)'s system, generally referred to as the Prospective Payment System (PPS), each hospital inpatient is classified into one of around 500 Diagnosis-Related Groups (DRGs), and the hospital is paid the amount that HCFA has assigned to each DRG. In other words, irrespective of what the hospital charges for, it will be paid only a fixed price for each DRG through major reimbursement plans. Therefore, it is logical to expect that by reducing the within DRG discrepancies, hospitals can cut cost and improve patient safety and satisfaction. In order to reach this goal the first step is to identify the main sources of variations. In this paper, we apply classical quality/process control tools and well known data mining methods to determine significant factors affecting the patient sequence among tens or hundreds of potential factors.

16.1 Introduction

During their hospital stay, patients may experience redundant steps and procedures that may lead to unnecessary excessive expenses, lower Quality of Care (QoC) and customer dissatisfaction. The excessive costs are often covered by hospitals or paid by individual patients since insurance companies have standard payment plans ranging from the infamous charge master or fee-for-service (FFS) price list to bundled payment systems such as diagnosis-related groups (DRGs), with various

N. Arbabzadeh • M.A. Jafari (✉) • K. Seyed
Industrial and Systems Engineering Department, Rutgers, The State University of New Jersey, 96 Frelinghuysen Rd., CoRE Building, Room 201, Piscataway, NJ, USA
e-mail: majafari@gmail.com

A. Matta et al. (eds.), *Proceedings of the International Conference on Health Care Systems Engineering*, Springer Proceedings in Mathematics & Statistics 61,
DOI 10.1007/978-3-319-01848-5_16,

forms of “discounts off charges” and “per diems” somewhere in between [1, 2]. Regardless of who pays for these excessive and unnecessary expenses, the adverse societal impacts and negative business consequences are immense. In this paper, we focus on the patient flow process in a hospital with DRG based payment system for its inpatient claims.

Renewed focus on quality measurement and improvement and on medical-error reduction has heightened interest in paying for performance, rather than just reimbursing providers for services rendered. Private Pay for Performance (P4P) programs for hospitals usually pays bonuses as an incentive above the agreed-upon reimbursement rate. A more rational reimbursement system, which rewards quality of care rather than simply doing more to patients, is the short-term goal of paying for performance. The longer-term goal is also to make the health care system more efficient. It has become clear that under existing reimbursement structures, current market forces are insufficient to ensure either higher-quality or more-cost-effective care [2]. P4P programs can be seen as additional incentives for hospitals to seek to improve their patient flow processes which can be attained through our variation reduction framework.

Since 1983, under Health Care Financing Administration (HCFA)’s system, generally referred to as the Prospective Payment System (PPS), each hospital inpatient is classified into one of around 500 Diagnosis-Related Groups (DRGs), and the hospital is paid the amount that HCFA has assigned to each DRG. Thus all hospitals treating all patients who fall into a particular DRG may charge whatever they charge based upon their patients’ courses of treatment, but each will be paid the same. One limitation to this methodology is that individual DRG categories often combine subgroups of patients with predictably different expected resource costs. HCFA has repeatedly improved the DRG definitions since 1984; in fact a new DRG system, called Medicare Severity DRGs (MS-DRGs), was adopted in October 1, 2007 which replaced 538 DRG system with 745 new MS-DRGs [1]. This enhancement, while necessary, does not fully account for differences in illness severity associated with substantial disparities in providers’ costs.

The fact is only a part of these disparities is attributed to the patient profile including his/her demographics, medical history, medication, physical exams, and so on; these are uncontrollable factors in patient flow. There are also controllable factors that influence patient’s experience from hospital admission to discharge. These include, but not be limited to, the order of treatments patient receives, medical procedures, current medications, received resources including physicians, nurses, technicians, transporters, and administrative work. These sources of variability could severely impact patient safety, QoC, professional satisfaction, and hospital revenue. The potential reduction in costs and increase in QoC and patient safety and satisfaction will be too rewarding to ignore. All these tools become handier especially when the regular normal operation of hospital is affected by an external incident varying from highway crashes to earthquakes and terrorist attacks. It’s in such situations that having a managed patient flow can be of great help to the hospital management to increase patient care and lower the number of fatalities.

This article is organized as follows. Section 16.2 presents the literature survey. In Sect. 16.3 we present the formulation of our problem. The data to test our procedure and the results of applying our methodology are discussed in Sect. 16.4. Conclusions are presented in the final section.

16.2 Literature Survey

A number of researchers have used queueing models to study various aspects of the patient flow process. McClean et al. (2005) use phase-type distributions to carry out model-based clustering of patients using the time spent by the patients in hospital. They cluster patients into classes on the basis of the number of phases involved. Cadez et al. (2003) presented a new methodology for exploring and analyzing navigation patterns on a web site [3]. They partition site users into clusters such that users with similar navigation paths through the site are placed into the same cluster. Their proposed method clusters users by learning a mixture of first-order Markov models using the Expectation-Maximization algorithm. In this paper, we have used their proposed model to cluster patient sequences in the hospital.

16.3 Technical Approach

Patient flow is not a single datum but a pattern or a sequence of steps. Unlike classical statistics where singular or array of data is used, we need to work with flow patterns and ordered data. In this paper, we use a mixture of first-order Markov models to model patient flow. Each patient is admitted to an inpatient floor with an initial diagnosis determined by the admitting physician. After patient is discharged, her chart is reviewed by coders and a DRG is assigned based on the primary (definitive final diagnosis) and other diagnoses together with treatments, resources and procedures utilized towards treating patient's condition during her stay. For each DRG certain level of resources (treatments, diagnostic tests, procedures, etc.) are assigned and required. From admission to discharge, a patient goes through a sequence of steps both in terms of her condition and the utilized resources, treatments and procedures. Throughout this paper we will refer to this sequence of steps as *patient flow* vector and denote it by $\overrightarrow{S}_i$ which is defined as follows:

$$\overrightarrow{S}_i = [S_{i1}, S_{i2}, \cdots, S_{ij}, \cdots, S_{in}]', i = 1, 2, \cdots, m \text{ and } j = 1, 2, \cdots, n \qquad (16.1)$$

where $\overrightarrow{S}_i$ is a $n \times 1$ ordered vector with jth element, S_{ij}, as the state of patient i at step j ($j = 1, 2, \ldots, n$). S_{ij} takes on values (s_{ij}) from among N ($n = 1, 2, \ldots, N$) possible patient states. Therefore the sequence $[S_{i1}, S_{i2}, \cdots, S_{ij}, \cdots, S_{in}]'$ indicates that patient i first was at state s_{i1}, then s_{i2}, and so on. In our model, the last state

is always x_n, which is "discharged" state. The nature and definition of these states can be different according to the level of granularity of the problem, i.e. the level of detail at which patient flow is observed. They can be as aggregated as generic states that any patient may go through during a hospital stay (like admission, inpatient floor stay, and discharge), or they can be very detailed including all the steps in each of the above mentioned high level states.

As we mentioned earlier there can be several sources of variability that are intrinsic to all healthcare delivery systems. We have categorized these sources into three groups:

(i) Unique characteristics of each patient (patient profile), including demographics, medical history and other health conditions upon admission. $\overrightarrow{X}_i$ defines these characteristics:

$$\overrightarrow{X}_i = [X_{i1}, X_{i2}, \cdots, X_{ik}, \cdots, X_{ip}]', i = 1, 2, \cdots, m \text{ and } k = 1, 2, \cdots, p \quad (16.2)$$

where $\overrightarrow{X}_i$ is a $p \times 1$ vector whose kth element, X_{ik}, represents the kth explanatory variable quantifying the kth characteristic of patient i.

(ii) Hospital resources, including medical and non-medical (overhead) staff {direct (nurse, tech, doctor) and indirect (unit secretary, housekeeping) labor and overhead labor}, major equipment, units and their functionalities (hospital factor). We denote by $\overrightarrow{Z}_i$ these characteristics:

$$\overrightarrow{Z}_i = [Z_{i1}, Z_{i2}, \cdots, Z_{i1}, \cdots, Z_{iq}]', i = 1, 2, \cdots, m \text{ and } l = 1, 2, \cdots, q \quad (16.3)$$

where $\overrightarrow{Z}_i$ is a $q \times 1$ vector whose lth element, Z_{il}, represents the lth explanatory variable quantifying the lth hospital resource on patient i. Depending on the attribute which they quantify, X_i and Z_i can be of both types of explanatory variables: continuous or categorical.

(iii) Random noise denoted by ε_i which are assumed to be i.i.d. random variables with mean zero and standard deviation $_i$. There are always un-assignable causes, which are usually grouped under random noise. Since random noise is statistically un-controllable, it is imperative to reduce its effect as much as possible. Any significant reduction in un-controllable variations will increase "process capability" and improve the process, which will in turn lead to significant cost reductions.

Furthermore, we assume that reentry of patient i to the hospital is a new admission with an updated $\overrightarrow{X}_i$ vector due to the new set of treatments that he received during his most recent stay. Then a historical data set of size *m*, containing *m* vectors of $\overrightarrow{S}$, $\overrightarrow{X}_i$, and $\overrightarrow{Z}$ defines patient paths, patient characteristics and hospital resources of *m* observed patients categorized under a specific DRG during a given time interval.

We intend to determine the number of clusters defined on the basis of sampled data collected on $\overrightarrow{S}$. We also intend to link $\overrightarrow{X}$, and $\overrightarrow{Z}$ to $\overrightarrow{S}$ in order to determine significant factors that lead to clusters within a DRG. Finally by controlling the

important attributes and reducing their variation we expect to see a reduction in the variations inherent in the patient flow process. Sections 16.3.1, 16.3.2, 16.3.3, 16.3.4, and 16.3.5 explain the steps of our algorithms in details.

16.3.1 Data Collection

With the current practices and adoption of EMR technology it is safe to assume that there are sufficient medical and personal data on patients, which can be mined and inferences can be made from. For example, CPT (Current Procedural Terminology) and HCPCS (Healthcare Common Procedure Coding System) codes are numbers assigned to every task and service a medical practitioner may provide to a patient including medical, surgical and diagnostic services. In principle, the data supporting CPT codes exist in hospitals (either collected real time using RFID or other RTLS technologies, or with some time lags entered by medical staff). Only in rare cases, the above data categories are all in a single database and is easily accessible; in majority of hospitals they are scattered in different databases, and data transfer and data fusion will be necessary. The data accessibility problem, however, is outside of the scope of this article. We will assume that this data exists and can be accessed for patient samples at different times.

16.3.2 Brainstorming

This step requires expert opinion to extract, filter, and transform data into meaningful quantifiable variables that we can further feed into our statistical engine. For this purpose, we should build multidisciplinary teams whose members will bring different perspectives and knowledge about the problem [4]. It is important to ensure that the core team and extended members include individuals that have direct contact with the process. The team should be brought together to hold brainstorming sessions for two important tasks:

1. Defining the state space of patient flow vector ($\overrightarrow{S}$): Medical judgment should be used to construct states, which both exhibit the necessary independence and make sense in terms of the delivery of care. A state space must be constructed in a manner that results in state definitions, which are mutually exclusive and collectively exhaustive [5]. This is essential to ensuring that Markov modeling of patient flow is valid.
2. Quantifying vectors of patient profile and hospital resources ($\overrightarrow{X}$, and $\overrightarrow{Z}$): To perform this task, one must try to identify as many potential variables as possible. One of the well-known tools to identify the potential causes of an event is the fishbone diagram also known as Ishikawa diagram or cause-and-effect diagram [6]. In this diagram, causes are usually grouped into major categories to identify these sources of variation.

Finally, we need to translate these potential causes into quantifiable random variables of either continuous or categorical type. The easiest case is when there are only two classes, such as variable gender with classes of "male" or "female". Examples for categorical type are gender, severity of illness, and nurse's level of expertise.

16.3.3 Sequence Clustering

At this step, we apply a mixture of fist-order Markov models to model patient flow sequences. We assume that the flow of each patient in the data set, $\overrightarrow{S}_i$, is generated independently (the traditional i.i.d. assumption). Statisticians refer to such a model as a mixture model with R components (R is the number of clusters). We apply Expected Maximization (EM) method to train our Once the model is trained, we can use it to assign each patient to a cluster or fractionally to the set of clusters. A mixture model for $\overrightarrow{S}$ with R components has the form:

$$p\left(\overrightarrow{S}\middle|\theta\right) = \sum_{r=1}^{R} p\left(c_r\middle|\theta\right).p_r\left(\overrightarrow{S}\middle|c_r,\theta\right) \tag{16.4}$$

where c_r is the cluster assignment for a given patient, $p(c_r|\theta)$ is the marginal probability of the rth cluster ($\sum_r p(c_r|\theta) = 1$) and $p_r\left(\overrightarrow{S}\middle|c_r,\theta\right)$ is the statistical model describing the distribution for the variables for patients in the rth cluster, and denotes the parameters of the model. We further assume that each model component is a first-order Markov model capturing the sequence of steps taken by a patient to some degree. Then, the EM method is used to train the parameters of the mixture model with known number of components R, given training data $d_{train} = \left\{\overrightarrow{S}_1, \overrightarrow{S}_2, \cdots, \overrightarrow{S}_M\right\}$ such that the following equation holds:

$$\theta^{ML} = \arg\max_\theta\, p\left(d_{train}\middle|\theta\right) = \arg\max_\theta \prod_{i=1}^{M} p\left(\overrightarrow{S}_i\middle|\theta\right) \tag{16.5}$$

θ^{ML} are the maximum likelihood or ML estimates of the model parameters.

In this paper, we have used Microsoft Sequence Clustering algorithm (SQL Server Analysis Services or SSAS) to carry out the sequence analysis. Microsoft SQL Server provides us with the membership assignment of each patient. Therefore, having a training data set of size M, we can run the sequence clustering algorithm and obtain the *vector of class memberships*, denoted by $\overrightarrow{Y}$, as follows:

$$\overrightarrow{Y} = [Y_1, Y_2, \cdots, Y_M]' \tag{16.6}$$

where Y_i is the class membership of patient i, and can accept values of 1, 2, …, R. Later, we will feed this vector into the Variable Selection module.

16.3.4 Variable Selection

In this step, we will use a well-known classifier, namely random forest, to identify the most important variables which significantly affect the patient flow sequences. Random forest (or random forests) is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees [7]. The data-set used for training comes in records of the form $(\overrightarrow{Q}, \overrightarrow{Y})$ for each data-point, where $\overrightarrow{Q}$ denotes a vector of observed characteristics (also referred as features or factors) and $\overrightarrow{Y}$ denotes a group label (also called target variable). In our application, $\overrightarrow{Q}$ is a $(p+q) \times 1$ vector of $\begin{bmatrix} \overrightarrow{X}_{p\times 1} \\ \overrightarrow{Z}_{q\times 1} \end{bmatrix}$ which contains the information of patient profile and hospital resources, i.e. the explanatory variables, and $\overrightarrow{Y}$ is the *vector of class memberships*, i.e., the output of the sequence clustering algorithm.

In order to perform the classification task we will use the *randomForest* package available in R software [8]. The input to the software will be feature vector $\overrightarrow{Q}_{(p+q)\times 1} = \begin{bmatrix} \overrightarrow{X}_{p\times 1} \\ \overrightarrow{Z}_{q\times 1} \end{bmatrix}$ and vector of *class memberships* $\overrightarrow{Y}$.

Random forests can be used to rank the importance of variables. There are two criteria based on which the Breiman's random forest calculates the importance of variables: *Gini importance* which calculates the mean Gini gain produced by Q_i over all trees, and *permutation accuracy importance* which is the mean decrease in classification accuracy after permuting Q_i over all trees. The variable importance plot gives a relative ranking of significant features, and absolute values of the importance scores should not be interpreted or compared over different studies. We consider, the first B variables as the most important variables where $B < p+q$. We will refer to the vector of important variables as $\overrightarrow{Q}'_{B\times 1} = \begin{bmatrix} \overrightarrow{X}'_{p'\times 1} \\ \overrightarrow{Z}'_{q'\times 1} \end{bmatrix}$, and define $\overrightarrow{X}'_i$, and $\overrightarrow{Z}'_i$ as follows:

$$\overrightarrow{X}'_i = \left[X_{i1}, X_{i2}, \cdots, X_{ik}, \cdots, X_{ip'}\right]', i = 1,2,\cdots, m \text{ and } k = 1,2,\cdots,p' \quad (16.7)$$

$$\overrightarrow{Z}'_i = \left[Z_{i1}, Z_{i2}, \cdots, Z_{i1}, \cdots, Z_{iq'}\right]', i = 1,2,\cdots, m \text{ and } l = 1,2,\cdots,q' \quad (16.8)$$

where $p \leq p'$, and $q \leq q'$.

16.3.5 Monitoring and Controlling Important Variables

Monitoring and controlling of important variables is the last step in our model. In the previous steps we established a relationship between patient flow sequences and process attributes, and identified those attributes that affect the patient flow process

significantly. In this step we investigate how and why these attributes affected patient flow. For this purpose, questions must be asked to find the assignable causes of variations and then a proper corrective action must be taken to eliminate them. To maintain the gained improvement and be able to detect future assignable variations, advanced statistical tools such as single-variable or multivariate control charts can be used. Using control charts is an ongoing activity over time to bring continuous improvements to the process.

16.4 Numerical Experimentation

In this section, we illustrate the performance of our algorithm using a simulated data set. For confidentiality reasons and also for the lack of sufficient real data at this time, we will demonstrate our model using simulated data. But the data generation will closely mimic the true real life process. We assume that patient flow sequences of cases under DRG type xxx can at most have six steps (N = 6). Seven factors have been identified as the potential causes of variation two of which are patient profile-related attributes (p = 2), and five are hospital resources (q = 5). The definitions of these variables can be found in Appendix.

To simulate expert opinion correctly, we assume to have a priori knowledge that Z_1, Z_2, and Z_5 are the significant variables and the rest of the attributes may not affect patient sequences significantly. Furthermore, we assume that we are given expert opinion on particular relationship between process attributes (Z_1, Z_2, and Z_5) and patient flow ($\overrightarrow{S}$). According to this prior knowledge, we know that exhaustively there exist 13 distinct patient sequences. It means that, assuming the patient flow process is a stable and stationary process without any chaotic behaviors, the expected path of a given patient falls into the set of 13 sequences. We model this relationship with a multinomial logit function regressing the transition probabilities on the value of significant attributes. The model is given by:

$$P\left(S_{ij} = W \middle| S_{i(j-1)} = V\right) = f\left(Z_1, Z_2, Z_3\right); W, V = 1, 2, \cdots, N \tag{16.9}$$

where $P(S_{ij} = W | S_{i(j-1)} = V)$ is the probability that patient i is in state W at step j, given that he was in state V at step j-1. This definition comes from our assumption that the patient transfer between states follows a Markov model. f is a multinomial logit function, and is defined as follows:

$$P\left(S_{ij} = W \middle| S_{i(j-1)} = V\right) = e^{\beta_0^W + \beta_1^W z_1 + \beta_0^W z_2 + \beta_0^W z_5} / 1 + \sum_{w=1}^{N} e^{\beta_0^w + \beta_1^w z_1 + \beta_0^w z_2 + \beta_0^w z_5}$$

$$W = 1, 2, \cdots, N-1 \tag{16.10}$$

$$P\left(S_{ij} = W \middle| S_{i(j-1)} = V\right) = 1/1 + \sum_{w=1}^{N} e^{\beta_0^w + \beta_1^w z_1 + \beta_0^w z_2 + \beta_0^w z_5}, W = N \tag{16.11}$$

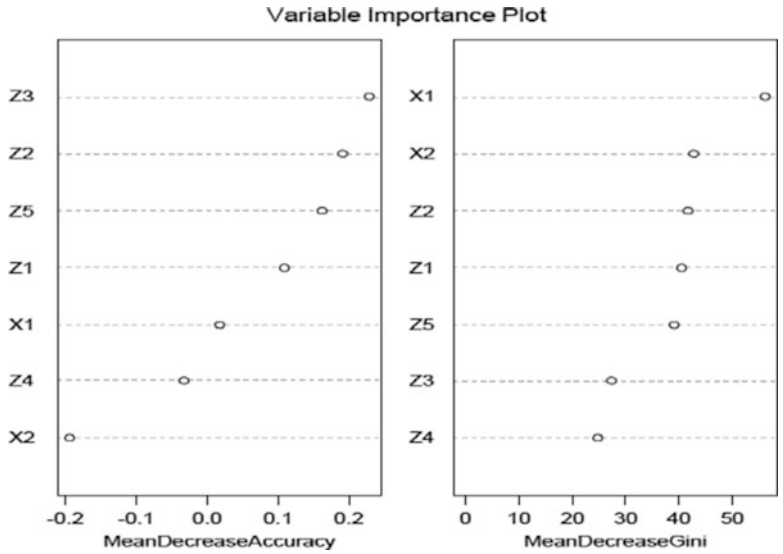


Fig. 16.1 Variable importance plot, case I- incomplete data set

We estimate the parameters of the logit model ($\overrightarrow{\beta}^{w} = [\beta_0^w, \beta_1^w, \beta_2^w, \beta_3^w]$). To generate a simulated data set, we start with an initial set, $d^{(initial)} = \left\{ \overrightarrow{S}, \overrightarrow{X}, \overrightarrow{Z} \right\}$, and use the logit model to estimate the transition probabilities matrix as a function of significant attributes. In our example, the initial set included only 3 distinct sequences out of the set of 13 original sequences. Then, we keep fine-tuning the parameters by adding new sequences until no further improvement is gained in our estimations and the built model completely captures the original relationship between $\overrightarrow{S}$ and $\overrightarrow{X}, \overrightarrow{Z}$. The variable importance plot can verify this gradual improvement.

Following the above approach, at each iteration a training data set of 1,000 cases was generated and fed into the statistical engine. Figures 16.1 and 16.2 show the variable importance plots for two cases: case I- an incomplete data set including 7 distinct sequences, case II- a complete data set including all the13 distinct sequences.

As it can be seen in Fig. 16.1, if the incomplete data set is fed into the engine, we would mistakenly be led into the conclusion that either the set of variables Z_2, Z_3, and Z_5, according to the permutation accuracy importance, or the variables Z_2, X_1, and X_2, according to Gini importance, were the significant factors affecting the patient flow sequences. While if the complete data set is fed into the algorithm according to both measures of variable importance, Z_1, Z_2, and Z_5 are correctly identified as the most important variables. The values of importance measures for X_1, X_2, Z_3, and Z_4 are close to zero meaning that relatively speaking their effects on patient flow are trivial.

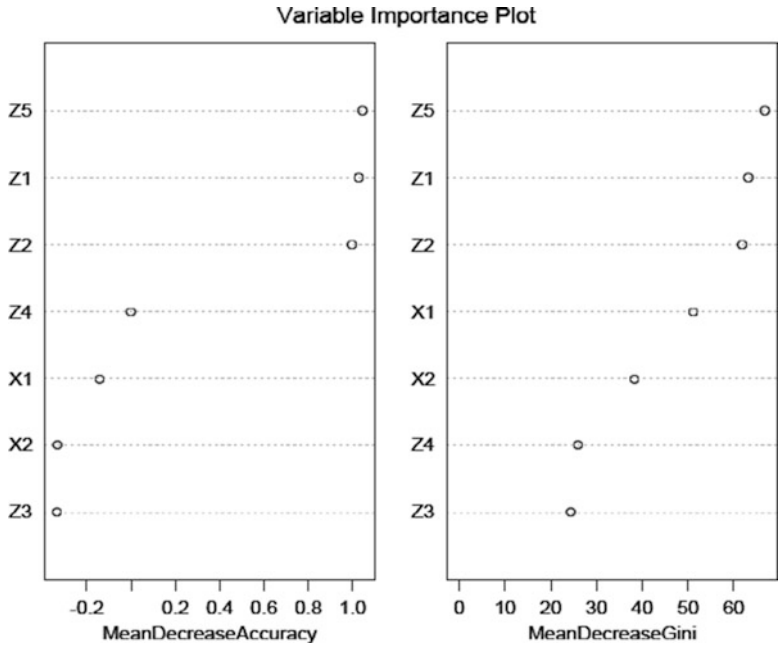


Fig. 16.2 Variable importance plot, case II- complete data set

In summary, by using the random forest classifier we have been able to identify the significant factors that truly impact patient flow. With this valuable information, the hospital management should focus his efforts and money to improve these attributes, which can consequently improve and facilitate patient flow in the hospital. Finally, to maintain the acquired improvements, the use of multinomial or multiattribute control charts is suggested to constantly monitor and control the important attributes and be alerted if a disturbance occurs in the patient flow process [9].

Note that in our example all the important variables are hospital resource-related attributes. In case a patient profile attribute is identified as a significant variable one should use other alternative solutions to control the process. One solution would be the use of robust optimization methods to control such a process since we cannot control or change statistical distributions of patient profile attributes into our favor [10].

16.5 Conclusions

In this paper, we have proposed a novel framework to identify the sources of variations in the patient flow process. The main idea is that by reducing the variations of these single processes we will be able to reduce the variation of patient

sequences. Our simulated results show that having a statistically large historical data set, the classifier can correctly determine the important variables, which truly had relationships with patient sequences. We further suggest the use of statistical control charts to maintain the gained improvements. The hospital management can use this valuable information to improve the quality of its patient flow process which consequently improve patient and staff satisfaction and results in a better cost management.

A.1 Appendix

Patient profile variables are as follows:

$$\text{Patient's age, } X_1 = \begin{cases} 1, & 0 \le age < 15 \\ 2, & 15 \le age < 30 \\ 3, & 30 \le age < 50 \\ 4, & 50 \le age \end{cases}$$

$$\text{Patient's gender, } X_2 = \begin{cases} 1, & Female, \\ 2, & Male, \\ 3, & Intersex. \end{cases}$$

Hospital resource-related variables are:

$$\text{Care taker's level of expertise: } Z1 = \begin{cases} 1, & Licensed\ Practical\ Nurse \\ 2, & Registered\ Nurse \\ 3, & Advanced\ Practice\ Nurse \end{cases}$$

$$\text{Work shift of the nurse, } Z_2 = \begin{cases} 1, & Morning\ shift, \\ 2, & Evening\ shift, \\ 3, & Night\ shift. \end{cases}$$

$$\text{The doctor's ranking: } Z_3 = \begin{cases} 1, & Physician, \\ 0, & Physician\ assistant. \end{cases}$$

$$\text{Availability of bedside monitoring tools: } Z_4 = \begin{cases} 1, & Available, \\ 0, & Otherwise. \end{cases}$$

$$\text{X\_Ray machine types: } Z1 = \begin{cases} 1, & Type\ M1\ maintained\ monthly, \\ 2, & Type\ M2\ maintained\ semiyearly, \\ 3, & Type\ M3\ maintained\ yearly. \end{cases}$$

References

1. Lynk, W.J.: One DRG, one price? The effects of patient condition on price variation within DRGs and across hospitals. Int. J. Health Care Finance Econ. **1**(2), 111–137 (2001)
2. Nichols, L.M., O'Malley, A.S.: Hospital payment systems: will payers like the future better than the past? Health Aff. **25**(1), 81–93 (2006)

3. Cadez, I., Heckerman, D., Meek, C., Smyth, P., White, S.: Model-based clustering and visualization of navigation. Data Min. Knowl. Discov. 399–424 (2003)
4. McHugh, M., Dyke, K. V., McClelland, M., Moss, D.: Improving Patient Flow and Reducing Emergency Department Crowding: A Guide for Hospitals. AHRQ Publication. 11(12)-0094 (2011)
5. Weiss, E.N., Cohen, M.A., Hershey, J.C.: An iterative estimation and validation procedure for specification of semi-Markov models with application to hospital patient flow. Oper. Res. **30**(6), 1082–1104 (1982)
6. Ishikawa, K. (Translator: J. H. Loftus): Introduction to Quality Control, 448 p (1990). ISBN 4-906224-61-X OCLC 61341428
7. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
8. Liaw, A., Wiener, M.: Classification and regression by randomForest. R News **2**(3), 18–22 (2002)
9. Topalidou, E., Psarakis, S.: Review of multinomial and multiattribute quality control charts. Qual. Reliab. Eng. Int. **25**, 773–804 (2009)
10. Wu, C.F.J., Hamada, M.S.: Experiments: Planning, Analysis and Optimization. Wiley (2000)
11. McClean, S., Faddy, M., Millard, P.: Markov model-based clustering for efficient patient care. In: Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems (CBMS'05) (2005).

Chapter 17
A Broader View on Health Care System Design and Modelling

Catherine Decouttere and Nico Vandaele

Abstract Many rigorous models have been developed to support health care system design. However, embedding these models in a broader stakeholder based framework, will substantially enhance the societal and human impact. Moreover, the acceptance of the proposed health care system (re)design suggestions will be more evident. Building on the model of an NMR scanning department, we propose an integrated health care design approach to support the modelling, the stakeholder analysis, the generation of alternative scenario's and the final design choice.

17.1 Introduction

The initial goal of a health care system, is not only to address the medical needs of individuals but also involves other factors affecting their well-being. An important underlying factor is patient satisfaction, which has been measured indirectly by capturing the patient experience [1]. The three main goals of a health system, as considered by the WHO, [2, 3] are: health improvement, responsiveness and fairness in financial contribution. The responsiveness of a health system has been put forward as a general concept designating the way individuals are treated and the environment in which they are treated. This responsiveness has been characterized by a common set of measurable domains. A first component, which involves ethical considerations, is called "respect for persons". It is built on the following pillars: respect for dignity, respect for individual autonomy and respect for confidentiality. A second component is called "client orientation". It includes prompt attention to health needs, basic amenities, access to family and community support, clarity of communication and choice of institution and individual providing care.

C. Decouttere (✉) • N. Vandaele
Katholieke Universiteit Leuven, Naamsestraat 69, 3000 Leuven, Belgium
e-mail: Catherine.decouttere@kuleuven.be; Nico.vandaele@kuleuven.be

A. Matta et al. (eds.), *Proceedings of the International Conference on Health Care Systems Engineering*, Springer Proceedings in Mathematics & Statistics 61,
DOI 10.1007/978-3-319-01848-5_17,

From a modelling point of view, health care systems have typically been approached by a limited selection of tangible performance dimensions such as technical capacity, waiting times, cost of care [4].

Starting from the performance measures of a national health system [3], the goals for subsystems and individual organizations, e.g. hospitals, can be derived. This results in a set of performance indicators which encompasses the more diverse aspects of patient experience, health improvement and fairness of financial contribution. It is clear that some inherently conflicting goals need to be brought into balance, which is why an integrated approach for design and modelling of the health care system is proposed. The challenges faced by today's health care systems are two-fold, either emerging from the demand side or the supply side. On the one side (the demand/receiving side) there are broader, more dynamic and advanced patient aims: the comfort of the patient, the exchange of information, the possibility of quick response and waiting times, the experience of patient rooms, the delivery and availability of drugs, the relation with the caregivers (nurses, physicians,) and the role of information towards patient and relatives, and many more. On the other side, the use of health care resources (the system/supply side) experiences also more and more the pressure of efficiency: government budgets, scarce skilled resources, logistics expenses, increased regulations, extremely expensive equipment, considerable environmental impact and production of hazardous outputs, etc. Also it became clear that the adoption of a newly designed health care system depends on the support it gets from the key stakeholders involved, e.g. the medical staff. All of these contribute on top of the issues raised from the demand side, among others, to a very complex design problem [5].

Typically the modelling of a health care system has been directed in a bottom up way: adding more and more incremental improvements to the operational models under study [6, 7]. In this paper we look at the health care design problem in a more top down way: from a design point of view and not from a modelling point of view as a starting point for our analysis. Health care systems need to excel on both technical, economic, and a vast amount of human and social aspects. Due to the multitude of stakeholders involved, it is a challenge to identify improvements for an existing health care system or to design radically new health care systems leading to an overall better societal, economic and technical performance. A patient-centred design approach, instead of a disease-centred one, is expected to deliver such radical steps forward [8]. Since there are a large amount of stakeholders in the health care system, and significant budgets involved, radical changes cannot be realised by one entity alone. A group-decision will preceed the adoption of a new system.

The starting point is a good understanding of the different stakeholders involved: understanding what their needs are, the strength of these needs, how well they are being served with the actual solutions, and what they will be needing more in the future. Through stakeholder analysis and user research, as part of a human-centred design approach, these insights can be gathered [9]. For complex systems with a diversity of stakeholders such as most health care systems, we additionally apply elements from the customer value chain approach [10]. It allows us to identify the stakeholders and to map the impact of each stakeholder on the design elements

of the health care system, which will eventually boil down to a bundle of product and service issues. Our approach builds further on the insights more widely known as value analysis, which also integrates the customer point of view in regular commercial relationships as for instance in [11].

Narrowly defined, on the operational and incremental level an innovative health care system can be designed or improved from the observed functional patient need for care and other stakeholder's requirements, which will be a combination of both technical opportunities and economic feasibility. We suggest a broader but complementary approach which is essential for more breakthrough innovations of the health care system. The seminal point here is both the mission of the supplying organization and the stakeholder's impact on functional, emotional, financial and decision making level as defined earlier. The supplying organisation is, e.g. a hospital, where the specific health care subsystem, for instance an emergency unit, is present. The mission statement and the values, define the position the organization wants to claim in the balancing exercise between economic, technical and human relevance. The translation of the mission fits in the multidimensional goals for the health care system design. It results in system innovations based on stakeholders' needs and the strengths and potential of the organization. Therefore, the evaluation of a product/service innovation as part of a new health care system requires a multidimensional approach, which is able to reflect this strategic alignment. In the next section we introduce the example of an NMR scanning unit as to illustrate our approach. Section 17.3 describes the integrated stakeholder approach built up from both the demand and supply side of the health care system. Section 17.4 concludes the paper.

17.2 Case Study: Redesign of NMR Scanning Unit

To illustrate the general idea of Sect. 17.1, we build upon an example from our own past research, which deals with an effort to improve patient waiting times for a NMR scanning department. In the sequel we review the case shortly. More in depth details can be found in [12].

17.2.1 Initial Flow Model

The objective was to improve patient waiting times (backlog times) as the hospital envisioned the patient lead time as primary key performance indicator for a sustainable customer service and thus preserving the long term success of the hospital. This objective was put forward without questioning about the impact of various stakeholders. Like in many (re)design projects, it turned out that the suggestions for operational improvements were valid but hardly implemented and suffered from resistance of particular stakeholders.

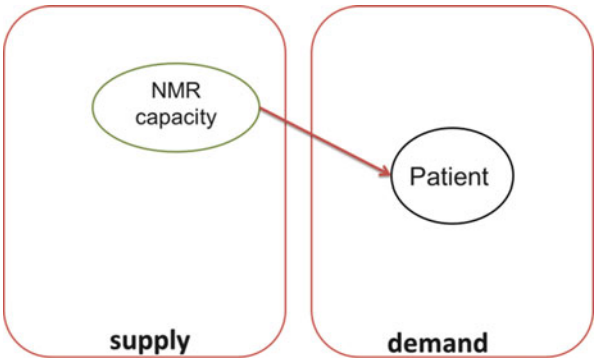


Fig. 17.1 The flow model approach

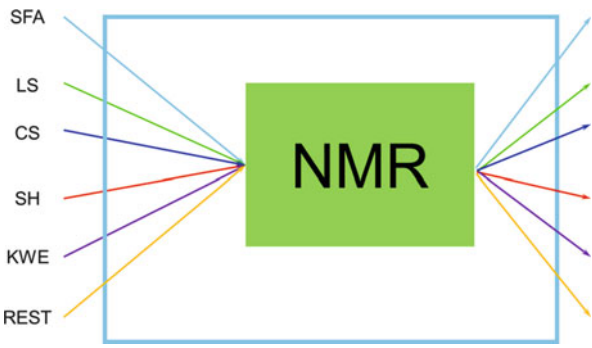


Fig. 17.2 The flow model of the NMR department

Both the supply side and the demand side have been modelled by use of a queueing model as can be seen in Fig. 17.1. From Fig. 17.2 it can be understood that the demand side consists of a patient flow represented by six classes. The flow originates from both recommendations from physicians as well as from patient's own initiatives. Each class represents a family of scans with similar technological characteristics:

1. Skull/foot/ankle SFA
2. Lumbal spine LS
3. Cervical spine CS
4. Shoulder/hip SH
5. Knee/wrist/elbow KWE
6. Rest (neck, breast, etc.) REST

Related to the supply side, per class, several scans are grouped to form a batch for which a general setup is performed after which each patient undergoes his scan.

The overall objective was to minimize the aggregate weighted lead time over all classes as a function of the six group sizes. The arrival process was modelled by use of generally distributed inter-arrival times wherefore field data were collected to obtain first and second moments. The same accounts for the setup and process times. The modelling details can be seen in Fig. 17.3.

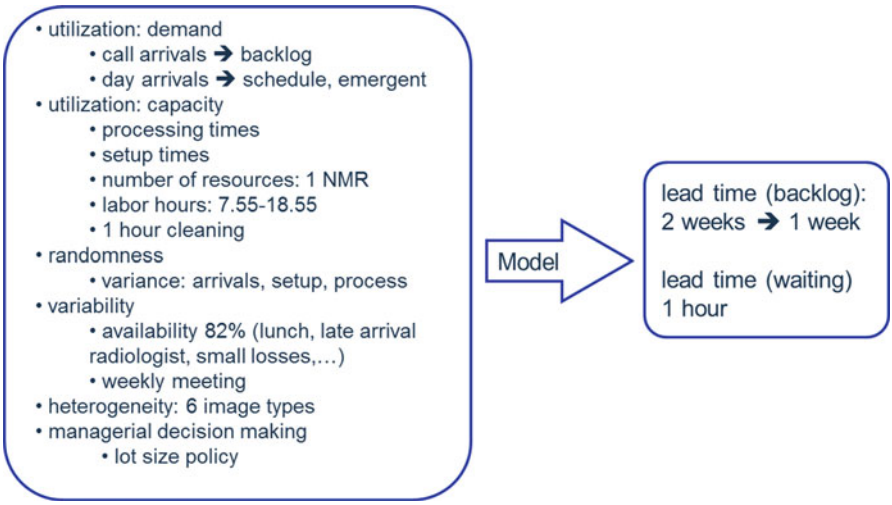


Fig. 17.3 The flow model details of the NMR department

The outcome was that the by taking proper measurements and appropriate managerial decision making in terms of batch sizing, the patient backlog could be reduced from two weeks to one week and that the waiting times on the day of scanning was about 1 h. The suggested improvements were not implemented, the question remains why? Clearly, this queueing model did not take into account the various stakeholder issues which made a proper implementation difficult. This will be discussed in the next subsections. It should be clear that our queueing approach can be replaced by any predicting (forward) flow modelling approach. In the literature, simulation and systems dynamics are popular alternatives [4].

17.2.2 From Stakeholders Needs to the Design Problem Setting

The stakeholders for the NMR scanning process were identified based on a stakeholder diagram which has been proven to be well understood in the health care environment [13], it can therefore be applied for stakeholder identification and participatory (re)design involving the medical staff.

Subsequently the stakeholders were mapped according to their role in the service, either on the supply side (care giving) or on the demand side (care receiving) as shown in Fig. 17.4. The service which is delivered by the NMR system is providing information from radiologist to the doctor or GP and to the patient and his or her close relatives. For each of the stakeholders, the nature of interaction during the NMR scanning process is determined following a customer value chain analysis [10]: functional, financial or decision making as shown in Fig. 17.5.

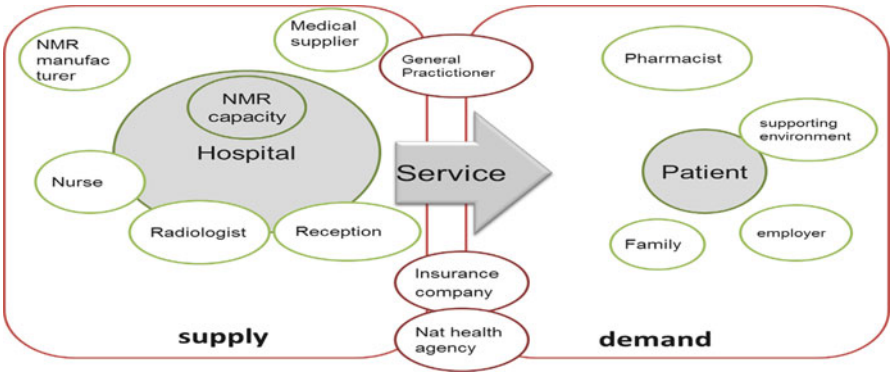


Fig. 17.4 Stakeholder mapping derived from stakeholder diagram

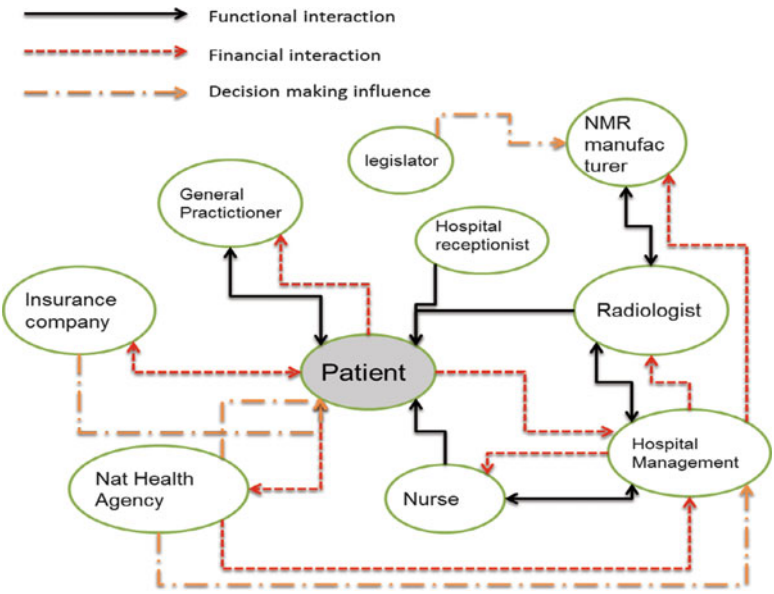


Fig. 17.5 The stakeholder interaction diagram for the NMR design problem

It becomes clear that some stakeholders will be crucial for the effective implementation and proper *functioning* of the health care system, as they will interact directly with the system and will experience its performance during the delivered service. These stakeholders involve the patients, doctors and nurses. Other stakeholders will have a *financial* interaction with the system and do not necessarily take part in direct interaction. They will set restrictions to the system design in terms of investments and cost of operation e.g. the insurance companies or social security bodies in general. A third kind of stakeholder are those with a strong *decision power* on the system design in terms of approving or rejecting a certain solution

by the use of laws, rules or standards. They act as gatekeepers for the system design and their aim can be the controlled acceptance of new technology with respect to ethics or safety. The national health security organization is an example of this. Stakeholders can interact on several levels at the same time, e.g. a patient interacts on the functional and the financial level, but usually not on the decision level. From this point on, the user research can be carried out. It will involve a mix of qualitative research methods such as patient observations and interviews, in order to gain empathy with the patients and other stakeholders involved. The resulting insights deliver additional information to the technical data from the process.

In product design, a human-centred approach usually starts from the user needs, usually the end-user, and takes a selection of other stakeholders into account. The user is preferably actively involved from the earliest stages in the design process, the idea generation phase throughout the concept definition and product development, for concept testing and prototype validation [14]. When designing a complex product/service system such as a health care system, a lot of different stakeholders' needs must to be taken into account simultaneously and conflicting requirements need to be solved in the new service design [15].

At this point we notice that the patient's decision power is sometimes quite low compared to that of other stakeholders: medical staff, procurement managers, nurses, family members of the patient, hospital board, governmental regulators among others. As it is reflected in some of the WHO's common set of domains [2], such as "choice of care provider" and "respect for autonomy", the patient's decision power is expected to become more important in future health care systems. However, the impact of the patient on the functioning level of the health care system is very high: most of the KPI's for improvement of health and responsiveness of the system are directly related to the patient.

The result of the user research will reveal the patient needs and stakeholder insights to induce a holistic system design. The design process should actively involve the stakeholders from the demand side and the supply side, and is based on design thinking and participatory design [16, 17].

17.2.3 From Mission to Design goals

As declared in the hospital's mission statement and institutional values,[1] the hospital under investigation strives for the highest quality treatment and care. Its central values are patient kindness and safety. The engagement towards the patient comes down to enhancing the patient's quality of life on both physical and psychological level, taking into account the uniqueness of each patient. Furthermore, the hospital values the interaction with the supporting environment of the patient, i.e. his family

[1] http://www.jessazh.be/over-jessa/algemeen/mission-statement

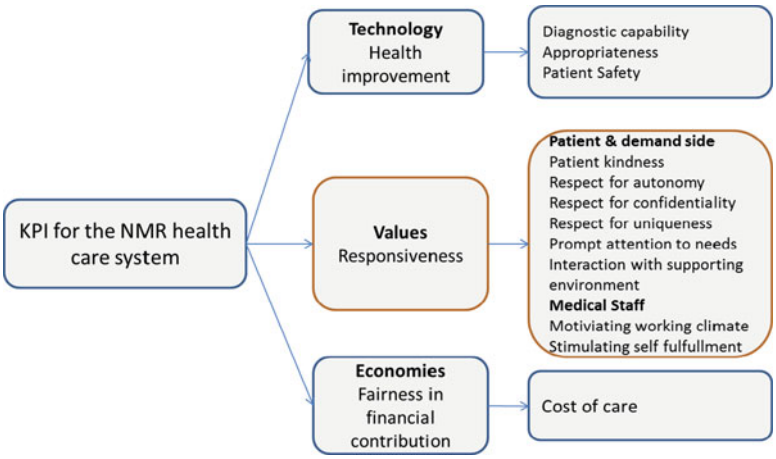


Fig. 17.6 The multi-dimensional KPI's of the NMR design problem

and other caregivers. Logically, these elements are fully in line with the WHO's common health care system goals, but furthermore, they inspire the organization in its daily operations.

The engagement from the hospital towards its employees, including doctors focuses on creating a motivating working climate with room for self fulfillment.

The analysis of these means and beliefs as stated by the hospital, leads to a set of high level aggregate KPI's for system innovation, such as the redesign of the NMR process, as depicted in Fig. 17.6. This figure is an adaptation of a methodology developed for R&D portfolio management [18].

17.3 Health Care System Design via an Integrated Stakeholder Approach

When we go down the road to the modelling effort behind the (re)design of the system, we end up with rigorous modelling from Fig. 17.3 fully embedded in the multi-dimensional design approach as depicted in Fig. 17.6. The KPI's for the design problem will be related to stakeholder's needs and will generate ideas for improved or new NMR systems. The integration is visualized in Fig. 17.7.

On the three main dimensions of KPI's, each system concept is represented by a set of inputs, limited resources, and outputs, desired outcomes. The flow model calculates the relation between a subset of inputs and outputs from the technology-pillar. The other inputs and output variables are the result of design activities. Many of the human-related outputs will be measured by qualitative techniques from user research and concept testing.

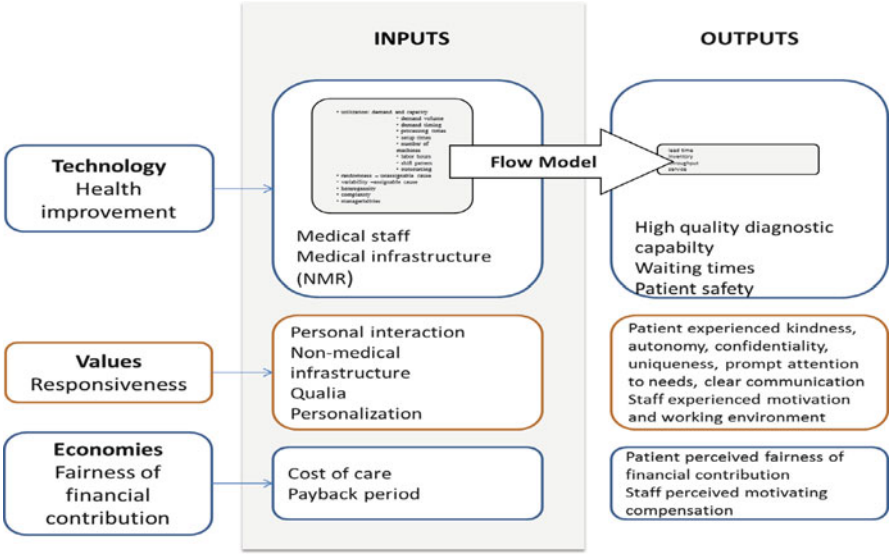


Fig. 17.7 The integrated NMR design problem

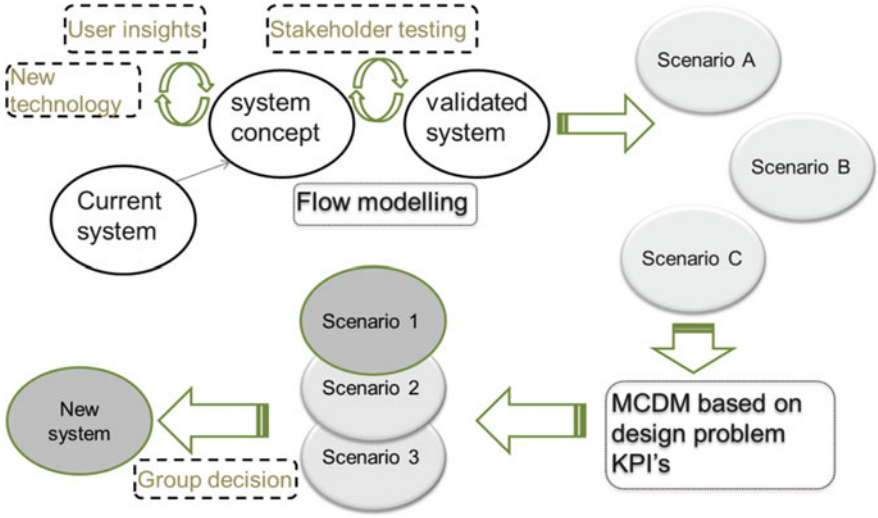


Fig. 17.8 The integrated health care design problem

Note that even the original flow model's objective of minimizing aggregate lead time is only one aspect, part of the technical main goal of improving health. Also the NMR equipment described by capacity, utilization and availability, is a myopic and limited view of the NMR supply side and part of the health improvement main goal.

At this point we can expose an overview of the broader health care system design and modelling process we propose. All steps are visualized in Fig. 17.8. The start

is the current health care system were user and stakeholders insights and possibly new technology, offer opportunities for improvement along the diverse KPI's. The decision making stakeholders put the limits to the solution space. The development of new system concepts is followed by testing with the stakeholders, in short iterative cycles, each time improving the concept. The technological modelling is used as partial knowledge to model input and output characteristics of the health care system concepts. The iterative design/model process step continues until a satisfying validated system is reached, called a scenario. Each scenario is characterized by its set of input and output variables. A number of scenarios are constructed, possibly very diverse in the solution they offer to the design problem. We argue that a scenario building methodologies can be useful here (see for instance [19] for a nice example).

As a single design will unlikely be championing on all dimensions, we expect a couple of designs to be top of class and thus candidates for implementation. At this point, a multi-criteria ranking method can be of great value to give insight into the multiple dimensions of the decision problem. With the additional help of effective infographics techniques, the final choice is usually made on the basis of a group decision which can even take additional elements into account to make the ultimate choice.

17.4 Conclusions

In this paper we revisited the (re)design of an NMR service system. The experience showed a very weak willingness to implement the model based suggestions for improvement. A major reason was the ignorant exclusion of major stakeholders in the design process. Therefore a broader approach is put forward based on stakeholder analysis and user-centred design. Based on this analysis, more mind expanding design propositions can be put forward which will then be dealt with by a multi-criteria decision method in order to select the best design. Additionally, the early involvement of key stakeholders in the design process can lead to better fitting designs and a higher willingness to implement the new service system. In this way we believe that health care system design will have a much higher probability of reaching the full-fledged implementation benefits for all stakeholders involved.

Future research enholds more formalization of the proposed approach and application of the methodology to other health care system design problems of which we have two in mind. One on the laboratory operations where samples have to be analysed and turned into information back to the primary process from which the samples were drawn (examples: clinical labs in hospitals and labs in pharmaceutical companies). Here the challenge is to avoid a sole focus on lean lab operations as to serve a predetermined customer service level. Another application deals with the staffing decisions in an emergency department, where obviously multiple stakeholders have their stake.

References

1. Bleich, S.N., Özaltin, E., Murray, C.J.L.: How does satisfaction with the health-care system relate to patient experience? Bull. World Health Organ. **87**(4), 271–278 (2009)
2. Murray, C.J.L., Frenk, J.: A framework for assessing the performance of health systems. Bull. World Health Organ. **78**(6), 717–731 (2000)
3. World Health Organization et al.: The world health report, 2000. Health systems: improving performance. World Health Organization, Geneva (2000). http://www.who.int/whr/2000/en/index.html. 2005
4. Brailsford, S.C., Lattimer, V.A., Tarnaras, P., Turnbull, J.C.: Emergency and on-demand health care: modelling a large complex system. J. Oper. Res. Soc. **55**(1), 34–42 (2004)
5. Beyan, O.D., Baykal, N.: A knowledge based search tool for performance measures in health care systems. J. Med. Syst. **36**(1), 201–221 (2012)
6. Rechel, B., Wright, S., Barlow, J., McKee, M.: Hospital capacity planning: from measuring stocks to modelling flows. Bull. World Health Organ. **88**(8), 632–636 (2010)
7. Taboada, M., Cabrera, E., Iglesias, M.L., Epelde, F., Luque, E.: An agent-based decision support system for hospitals emergency departments. Procedia Comput. Sci. **4**, 1870–1879 (2011)
8. Barry, M.J., Edgman-Levitan, S.: Shared decision making: the pinnacle of patient-centered care. N. Engl. J. Med. **366**(9), 780–781 (2012)
9. Human Centered Design Toolkit: Ideo. Retrieved on 24th November, 2008
10. Donaldson, K.M., Ishii, K., Sheppard, S.D.: Customer value chain analysis. Res. Eng. Des. **16**(4), 174–183 (2006)
11. Keen, P.G.W.: Value analysis: justifying decision support systems. MIS Q. **5**, 1–15 (1981)
12. Vandaele, N., Van Nieuwenhuyse, I., Cupers, S.: Optimal grouping for a nuclear magnetic resonance scanner by means of an open queueing model. Eur. J. Oper. Res. **151**(1), 181–192 (2003)
13. Jun, G.T., Ward, J., Morris, Z., Clarkson, J.: Health care process modelling: which method when? Int. J. Qual. Health Care **21**(3), 214–224 (2009)
14. Martin, J., Barnett, J.: Integrating the results of user research into medical device development: insights from a case study. BMC Med. Inform. Decis. Mak. **12**(1), 74 (2012)
15. Clarkson, P.J., Buckle, P., Coleman, R., Stubbs, D., Ward, J., Jarrett, J., Lane, R., Bound, J.: Design for patient safety: a review of the effectiveness of design in the UK health service. J. Eng. Des. **15**(2), 123–140 (2004)
16. Brown, T., et al.: Design thinking. Harvard Bus. Rev. **86**(6), 84 (2008)
17. De Rouck, S., Jacobs, A., Leys, M.: A methodology for shifting the focus of e-health support design onto user needs: A case in the homecare field. Int. J. Med. Inform. **77**(9), 589–601 (2008)
18. Vandaele, N.J., Decouttere, C.J.: Sustainable R&D portfolio assessment. Decis. Support Syst. **54**(4), 1521–1532 (2012)
19. Brailsford, S.C.: System dynamics: what's in it for healthcare simulation modelers. In: Simulation Conference, 2008. WSC 2008. Winter, pp. 1478–1483. IEEE (2008)

Chapter 18
Epidemic State Estimation with Syndromic Surveillance and ILI Data Using Particle Filter

Taesik Lee and Hayong Shin

Abstract Designing effective mitigation strategies against an influenza outbreak requires an accurate prediction of a disease's future course of spreading. Real time information such as syndromic surveillance data and reporting of influenza cases by clinicians can be used to generate an estimate of the current state of spreading of a disease. Syndromic surveillance data is immediately available compared to clinical reports that require data collection and processing. On the other hand, syndromic data is less credible than the clinically confirmed case reports. In this paper, we present a method to combine immediately-available-but-highly-uncertain syndromic surveillance data with credible-but-time-delayed clinical case report data. This problem is formulated as a non-linear stochastic filtering problem and solved by a particle filtering method. Our experimental results on a hypothetical pandemic scenario show that state estimation is improved by utilizing both data sets than when using only one of them, but the amount of improvement depends on relative credibility and length of delay of clinical case report data. This result is explained with a preliminary analysis for a linear, Gaussian case.

18.1 Introduction

Designing optimal containment and mitigation strategies upon an epidemic outbreak is of critical interest. The first step to designing a containment strategy is to detect an outbreak of an epidemic and to predict the future course of spreading of the disease. The magnitude and speed of the spread of an epidemic need to be accurately estimated for public health authorities to develop mitigation strategies. The capability to accurately predict the future course of an epidemic influenza is a key to effective real-time decision making.

T. Lee (✉) • H. Shin
KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon 305-701, Korea
e-mail: taesik.lee@kaist.edu; hyshin@kaist.ac.kr

A. Matta et al. (eds.), *Proceedings of the International Conference on Health Care Systems Engineering*, Springer Proceedings in Mathematics & Statistics 61,
DOI 10.1007/978-3-319-01848-5_18,

Prediction of the future course of disease spread requires a high-fidelity disease spread model that, given a current state of the spread of an epidemic, produces quantitative estimates on how quick and severe it will be. Largely, there are two approaches to model the spreading of an epidemic disease: equation-based models and simulation-based models. Once we have a high-fidelity model for disease spreading, we can build a variety of containment strategies into the model and assess their effectiveness.

Quality of the prediction of an epidemic spread depends also on what we know about the current state of the system. Information on the current state – i.e., how many people have been exposed and infected as of today – is fed into an epidemic model to identify the most likely scenario for the disease progress. If an estimate on the current state is wrong, the prediction by even the most accurate epidemic model will be wrong, leading to suboptimal responses. This paper addresses the problem of state estimation for the spread of an epidemic influenza using a nonlinear stochastic filtering technique.

A traditional source of information for the current epidemic state is Influenza-Like-Illness(ILI) data from a government health agency such as Center for Disease Control and Prevention (CDC) [3]. ILI refers to a medical diagnosis of a possible influenza case. ILI data are gathered from healthcare providers as patients with relevant symptoms visit hospitals and clinics. This data is used as an indicator for the number of people infected with an epidemic flu. Although ILI data have some uncertainty [6], it is generated from reports by physicians based on their medical diagnosis, and thus can be considered a reasonably reliable indicator for the spread of a flu. That said, there is one important shortcoming when using ILI data for estimating a current state of disease spread. Generally, it takes 1–2 weeks to gather and process data from a large surveillance network [1, 4, 6]. This 1–2 week's lag makes ILI data outdated by the time it is released. It does not provide a real-time information on the current state of an epidemic.

Another source of disease spreading information is syndromic surveillance data, which is recently getting significant attentions from research community [1, 2, 4, 7, 12]. Examples of syndromic surveillance data includes school or work absenteeism, over-the-counter drug sales, search engine queries as well as clinical data [5]. With proper tools and systems, syndromic surveillance data can be made available in almost real-time, which offers an advantage for making timely state estimation. However, this data is based on the "syndromes," including population's behavioral responses, and thus it has lower credibility than ILI data.

This paper proposes a method that can combine reliable-but-time-delayed ILI data with less-reliable-but-real-time syndromic surveillance data to improve the accuracy of estimation of an epidemic state. As the two sets of data compliment each other, it is expected that combining the two will enhance estimation outcomes. One of the tools commonly used for epidemic state estimation and prediction is a recursive Bayesian state estimation technique [6,8,11,12]. Bayesian state estimation assumes some knowledge on an underlying dynamic model, and recursively updates the degree of belief in system states by using sequentially available observation

data. For example, [12] uses a modified stochastic SIR model as a system model along with an observation model based on [4], and formulates a Bayesian filtering problem. Our method also uses a recursive Bayesian state estimation. Specifically, we use particle filtering, a stochastic nonlinear filtering technique, to estimate the size of infected population in a community when both ILI data and syndromic data are available.

18.2 Epidemic Model

We consider a hypothetical outbreak of an epidemic, where two streams of observation information – syndromic surveillance data and ILI data – arrive sequentially. Syndromic surveillance data is assumed to be immediately available but with high uncertainty, while more reliable ILI data are delayed by a certain time lag.

We use a stochastic version of an equation-based epidemic model to describe the spreading dynamics of an epidemic. Equation-based epidemic models typically divide the population into a few compartments, and express the rate of increase of each compartment by a set of non-linear ordinary differential equations. One of the simplest models is the S-I-R model, where S, I, and R represents susceptible, infectious, and recovered compartments. [12] presents a stochastic version of the basic S-I-R model, which incorporates stochastic fluctuations:

$$\frac{ds}{dt} = -\beta i s^{\nu} + \sigma_q \xi \; ; \; \frac{di}{dt} = \beta i s^{\nu} - \gamma i - \sigma_q \xi + \sigma_\gamma \zeta \; ; \; r = 1 - s - i \tag{18.1}$$

where s, i, and r are the sizes of each compartment normalized by the total population. β represents the rate of infectious contacts, and γ is the recovery rate, which is an inverse of an average infectious period for the disease. ν, $\sigma_q \xi$ and $\sigma_\gamma \zeta$ are introduced to account for heterogeneity and stochasticity. ξ and ζ are uncorrelated, white Gaussian noises with zero mean and a unit variance. For the purpose of discussions in this paper, we simplify (18.1) by assuming that the parameters in the model – $\beta, \nu, \sigma_q, \gamma$, and σ_γ – are all known constants.

By defining a state vector $\mathbf{x} = [s, i]^T$, we have a discretized, state-space model for (18.1):

$$\left\{ \begin{matrix} s_{k+1} \\ i_{k+1} \end{matrix} \right\} = \left[\begin{matrix} -\beta i_k s_k^{\nu} \\ (\beta i_k s_k^{\nu} - \gamma i_k) \end{matrix} \right] \Delta t + \left[\begin{matrix} \sigma_q \xi \\ (-\sigma_q \xi + \sigma_\gamma \zeta) \end{matrix} \right] \Delta t \tag{18.2}$$

Parameters in (18.2) assume constant values: $\beta = 0.3, \gamma = 0.1, \nu = 1.0$, and $\sigma_q = \sigma_\gamma = 0.001$.[1]

[1]We also tested cases with $\sigma_q = \sigma_\gamma = 0.005$. Results from those cases are similar to when $\sigma_q = \sigma_\gamma = 0.001$, and not included due to the limited space.

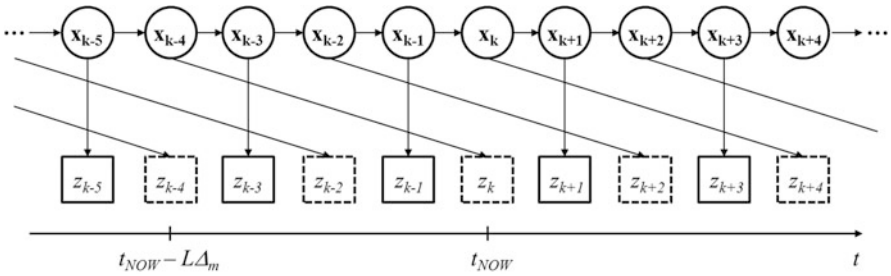


Fig. 18.1 Relationship between syndromic surveillance data (*solid box*) and ILI data (*dotted box*)

For a measurement model, we follow [12] to assume:

$$z = b_j i^{\varsigma_j} + \sigma_{w,j}\eta_j \tag{18.3}$$

where j is an index to indicate different types of measurement data. For simplicity, we also assume that b_j and ς_j are known constants, η_j is independent Gaussian noise with a unit variance, and $\sigma_{w,j}$ is a known constant.

For measurement model (18.3), index $j = 1$ denotes syndromic surveillance data and $j = 2$ for ILI data. We assume $b_j = 1$ and $\varsigma = 1.0$ for $j = 1, 2$. $\sigma_{w,1}$ and $\sigma_{w,2}$ are used to represent reliability of the two data sets. $\sigma_{w,1}$, for syndromic surveillance data, is set to 0.05, while $\sigma_{w,2}$ is varied in the range of 0.0005–0.1 to depict the relative difference in reliability between the 2.

We assume that measurement data arrive with a fixed interval Δ_m, with each type of data having an interval of $2\Delta_m$, arriving in an alternating sequence. An arrival pattern for measurement data is shown in Fig. 18.1. ILI data have a fixed time-delay of $L\Delta_m$. Suppose that, at $t = t_{NOW}$, z_k is a measurement from ILI data reported at the k^{th} sequence. z_k corresponds to the system state $[s, i]$ at time t_k, which is $(t_{NOW} - L\Delta_m)$.

Equations (18.1) and (18.3) serve as a system model and a measurement model for particle filter formulation of our problem.

18.3 Problem Formulation and Particle Filter Implementation

In this section, we briefly discuss the basics of standard particle filter technique, and present its modification to handle out-of-sequence measurement(OOSM) data. For more details on particle filter and OOSM particle filter, see [9, 10].

Particle filter is a recursive Bayesian filter that is particularly useful for non-linear, non-Gaussian problems. This technique is commonly used for estimating

the state of a dynamic system where the state variables are not directly observable. A series of measurement data are combined with a known system model to update belief on the true state of the system.

To formally describe the technique, consider a system whose dynamics is described by the following system model: $\mathbf{x}_k = f_{k-1}(\mathbf{x}_{k-1}) + \mathbf{v}_{k-1}$. $\mathbf{x}_k$ denotes a state vector at time index k, f_{k-1} is a possibly non-linear and time-varying function, and $\mathbf{v}_{k-1}$ represents process noise. Suppose for this system a set of observable variables, $\mathbf{z}_k$ are measured, and $\mathbf{z}_k$ is related to $\mathbf{x}_k$ by the following observation model: $\mathbf{z}_k = h_k(\mathbf{x}_k) + \mathbf{w}_k$, where h_k is a possibly non-linear and time-varying function, and $\mathbf{w}_k$ denotes measurement noise.

Let $\mathbf{z}_{0:k}$ denote a series of measurement data up to k, then a recursive Bayesian filtering seeks to construct a posterior pdf $p(\mathbf{x}_k|\mathbf{z}_{1:k})$ as an estimate for the true state of the system. This is done in two stages – *prediction* and *update*. Suppose we know a posterior pdf at $k-1$, i.e. we know $p(\mathbf{x}_{k-1}|\mathbf{z}_{1:k-1})$. Prediction stage computes the distribution of system state at k given $\mathbf{z}_{1:k-1}$ based on our knowledge of the system model. In the update stage, a posterior density $p(\mathbf{x}_k|\mathbf{z}_{1:k})$ is computed as a product of $p(\mathbf{x}_k|\mathbf{z}_{1:k})$ and $p(z_k|\mathbf{x}_k)$.

Particle filter carries out these steps by using a sample representation for a posterior density $p(\mathbf{x}_k|\mathbf{z}_{1:k})$ as follows:

$$p(\mathbf{x}_k|\mathbf{z}_{1:k}) \simeq \sum_{i=1}^{N_s} \omega_k^i \delta(\mathbf{x}_k - \mathbf{x}_k^i) \tag{18.4}$$

where $\delta(\mathbf{x} - \mathbf{x}^i)$ is 1 for $\mathbf{x} = \mathbf{x}^i$ and 0 otherwise. ω_k^i is a weight assigned to sample x_k^i, and N_s is the number of samples (i.e., *particles*). Following the principle of importance sampling, ω_k^i can be written in a recursive form:

$$\omega_k^i \propto \omega_{k-1}^i \frac{p(\mathbf{z}_k|\mathbf{x}_k^i)p(\mathbf{x}_k^i|\mathbf{x}_{k-1}^i)}{q(\mathbf{x}_k^i|\mathbf{x}_{0:k-1}^i, \mathbf{z}_{1:k})} \tag{18.5}$$

where $q(\cdot)$ is an importance density, from which the samples $\mathbf{x}_k^i$ are drawn.

Out-of-sequence measurements (OOSMs) refer to measurement data that arrive with delay such that they represent the system state at some point in the past. Let t_k denote actual time instant for the kth measurement $\mathbf{z}_k$. When $t_k > t_{k-1}$, $\mathbf{z}_k$ is in sequence, and when $t_k < t_{k-1}$, then it is out of sequence. OOSM particle filter provides a means to update importance weight ω_k^i upon an arrival of an out-of-sequence measurement, $\mathbf{z}_k$.

Suppose we have a series of in-sequence measurement data, $\mathbf{z}_{1:k-1}$, and at k, an out-of-sequence data $\mathbf{z}_k$ arrives. Let a and b denote the time indices right before and after $\mathbf{z}_k$. That is, $t_b < t_k < t_a$ and $a = b + 1$. [9] shows that the joint posterior density $p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k})$ can be written in the following recursive form:

$$p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k}) = p(\mathbf{x}_{0:k-1}|\mathbf{z}_{1:k-1}) \times \frac{p(\mathbf{x}_a|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{x}_b)p(\mathbf{z}_k|\mathbf{x}_k)}{p(\mathbf{x}_a|\mathbf{x}_b)p(\mathbf{z}_{1:k})} \tag{18.6}$$

With (18.6), we now have a weight update equation similar to (18.5).

$$\omega_k^i \propto \omega_{k-1}^i \frac{p(\mathbf{z}_k|\mathbf{x}_k^i)p(\mathbf{x}_k^i|\mathbf{x}_a^i,\mathbf{x}_b^i)}{q(\mathbf{x}_k^i|\mathbf{x}_{0:k-1}^i,\mathbf{z}_{1:k})} \tag{18.7}$$

It was shown in [9] that the optimal importance density function is $p(\mathbf{x}_k^i|\mathbf{x}_a^i,\mathbf{x}_b^i,\mathbf{z}_k^i)$. However, sampling form this optimal importance function is quite difficult, and [9] suggests $p(\mathbf{x}_k^i|\mathbf{x}_a^i,\mathbf{x}_b^i)$ as a tractable approximation. This choice of importance function reduces (18.7) to $\omega_k^i \propto \omega_{k-1}^i p(\mathbf{z}_k|\mathbf{x}_k^i)$.

It turns out that for our problem, we have a more straightforward implementation for the above OOSM particle filter framework. Here, we depart from the OOSM particle filter algorithm of [9] and propose a modified version for two reasons. First, applying [9] requires sampling from $p(\mathbf{x}_k^i|\mathbf{x}_a^i,\mathbf{x}_b^i,\mathbf{z}_k)$ or from its approximation $p(\mathbf{x}_k^i|\mathbf{x}_a^i,\mathbf{x}_b^i)$. Neither is straightforward in our case due to the nonlinearity present in the system model (18.1). Secondly, unlike general OOSM cases discussed in [9], we assume a known and fixed amount of lag $L\Delta_m$ for the OOSMs. This eliminates concerns for having to store the entire history of particles throughout filtering time horizon. With the assumption of fixed lag, we only need to store particle history from t_{NOW} to $(t_{NOW} - L\Delta_m)$.

A basic idea behind our OOSM particle filter approach is the following: when we obtain a measurement for a past state, the best thing to do is to go back and re-compute from the past point as if a set of in-sequence measurement data are arriving. We call this approach a *roll-back-and-update* scheme. Our roll-back-and-update scheme is described using Fig. 18.1. Suppose we have a posterior density $p(\mathbf{x}_{k-1}|\mathbf{z}_{1:k-1})$ as a set of weighted particles $\{\mathbf{x}_{k-1}^i, \omega_{k-1}^i\}$. We also have a past history of particles back to $\mathbf{x}_{k-5}^i$. That is, at $k-1$ $(t = t_{k-1})$, we have updated particles $\{\mathbf{x}_{k-5}^i, \omega_{k-5}^i\}$, $\{\mathbf{x}_{k-3}^i, \omega_{k-3}^i\}$ and $\{\mathbf{x}_{k-1}^i, \omega_{k-1}^i\}$ using z_{k-5}, z_{k-3} and z_{k-1}. Without the presence of OOSM data, this is exactly what we would get with a standard particle filter. Now at k $(t = t_{NOW})$, we obtain an OOSM data z_k. If we ignore the previously computed $\mathbf{x}_{k-3}^i$ and $\mathbf{x}_{k-1}^i$ and set us back to $t_{NOW} - L\Delta_m$, we simply have another instance of standard particle filtering: we have $\{\mathbf{x}_{k-5}^i, \omega_{k-5}^i\}$ and a series of in-sequence measurements $\{z_k, z_{k-3}, z_{k-1}\}$. $\mathbf{x}_{k-5}^i$ is propagated to find $\mathbf{x}_{k-4}^i$, its weight ω_{k-4}^i is updated using z_k, and the process continues to $k-3$. Thus, we can re-compute for each particle to update $\{\mathbf{x}_{k-1}^i, \omega_{k-1}^i\}$. This process continues, repeatedly updating a portion of history of the particles from $k-4$ to $k-1$.

A procedure of the algorithm used in this paper is summarized below:

- $\{\mathbf{x}_{k-1}^i, \omega_{k-1}^i\}_{i=1}^{N_s}$ is given, and z_k arrives
- If z_k is a syndromic surveillance data (i.e., an in-sequence measurement)
  - Use a standard particle filter to compute $\{\mathbf{x}_k^i, \omega_k^i\}_{i=1}^{N_s}$

(continued)

(continued)

- Else if z_k is a clinical case report data (i.e., an out-of-sequence measurement)
  - From the stored particle history, retrieve $\{\mathbf{x}^i_{k-L-1}, \omega^i_{k-L-1}\}$
  - Given an in-sequence measurement $\{z_k, z_{k-L+1}, z_{k-L+3}, \ldots, z_{k-1}\}$, execute a standard particle filter to update $\{\mathbf{x}^i_k, \omega^i_k\}$, $\{\mathbf{x}^i_{k-L+1}, \omega^i_{k-L+1}\}$, $\{\mathbf{x}^i_{k-L+3}, \omega^i_{k-L+3}\}, \ldots, \{\mathbf{x}^i_{k-1}, \omega^i_{k-1}\}$
  - Sample $\mathbf{x}^i_k$ using (18.2), and set $\omega^i_k = \omega^i_{k-1}$ to obtain $\{\mathbf{x}^i_k, \omega^i_k\}_{i=1}^{N_s}$

18.4 Experimental Results and Discussion

Our experiments have been motivated by a few simple questions: given two sets of observation data – immediately-available-but-highly-uncertain data and credible-but-time-delayed data, which one would yield a better state estimation? Will the answer depend on the amount of delay? Would it be always better to use both sets of data than using only one, and if so, how much?

We first generate a *true* state sequence, $\mathbf{x}_{0:T}$ using (18.2) with the initial state $[s^*_0, i^*_0] = [0.99, 0.01]$. Measurement data is then generated according to (18.3). Syndromic surveillance data and ILI data arrive each with a measurement interval $\Delta_m = 10\Delta_t$. Figure 18.2 shows an example of true system state $i(t)$ and the two measurement data. In this example, measurement noise for syndromic surveillance $\sigma_{w,1} = 0.05$ and for ILI, $\sigma_{w,2} = 0.005$. (For the rest of this paper, we use σ_{synd} and σ_{ILI} instead of $\sigma_{w,1}, \sigma_{w,2}$.) ILI data, denoted by '$*$', has a lag of 50 time units.

For an initial prior, we use a uniform distribution such that $i_0 \sim U(0, 2i^*_0)$ and $s_0 = 1 - i_0$. The number of particles N_s is 300, and particles are resampled at all steps. A typical example of particle filtering results is shown in Fig. 18.3. On the left, we use syndromic surveillance data ('$\times$') only, and in the middle, only ILI data ($*$) with lag$=$ 50 is used. Shown on the right is the estimation result when both syndromic surveillance data and ILI data are used together.

Comparing the left and middle plots, we can see that the relative uncertainty of syndromic surveillance data ($\sigma_{synd} = 0.05 > \sigma_{ILI} = 0.005$) manifests as a wider range of particle distribution, i.e., larger variance of a posterior density, even when ILI data has a non-trivial lag. When comparing the first two plots and the rightmost plot, it is not readily visible whether such improvement exists when both sets of data are used. To make these comparisons quantitative, we measure RMSE value taken over a trajectory:

$$RMSE = \sqrt{mean_k\{(i^{true}_k - \hat{i}_k)^2\}} \tag{18.8}$$

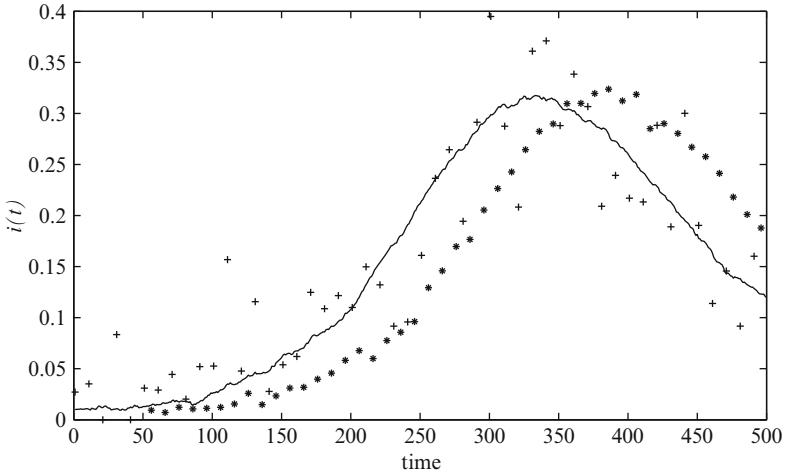


Fig. 18.2 An example of a trajectory of true state $i(t)$ (*solid line*) and the two measurement data ('+' for syndromic surveillance data and '$*$' for ILI data). ILI data arrives with a lag of 50 time units

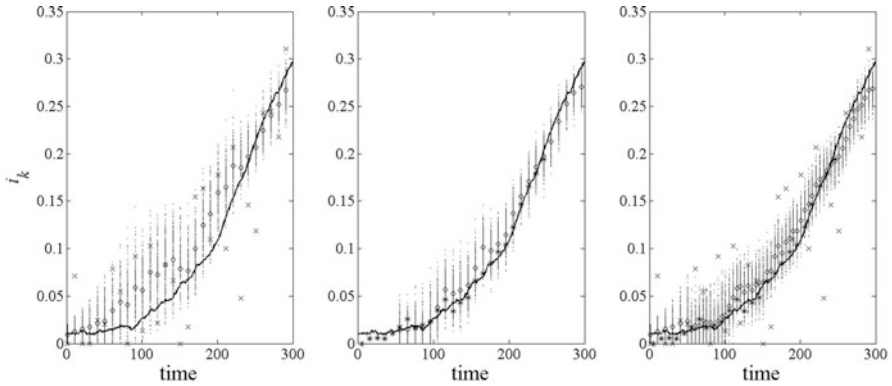


Fig. 18.3 Particle filter estimates a posterior density of true state $p(\mathbf{x}_k|z_{1:k})$ as an approximate density represented by a set of particles. At each k, i_k of a set of resampled particles (equal weights) are plotted along the vertical direction, and their mean value, $\hat{i}_k$ is denoted by a *circle*. *Solid line* shows a true trajectory of $i(t)$. Filtering results are shown for: (*Left*) syndromic surveillance data only, $\sigma_{synd} = 0.05$; (*Middle*) ILI data with lag $= 50$, $\sigma_{ILI} = 0.005$ (note that data points have been shifted to indicate their actual measuring point); (*Right*) both syndromic surveillance and ILI data

We vary the amount of lag $\{0, 10, \ldots, 140\}$, and test 6 levels of $\sigma_{ILI} = \{0.001, 0.002, 0.005, 0.01, 0.02, 0.05\}$ while fixing σ_{synd} at 0.05. We run 20 replications for each set of parameter values to obtain average *RMSE* over the replications. For each case, we evaluate average *RMSE* under (1) using syndromic surveillance data only, (2) using ILI data only, and (3) using both sets of data. Results are shown in Fig. 18.4.

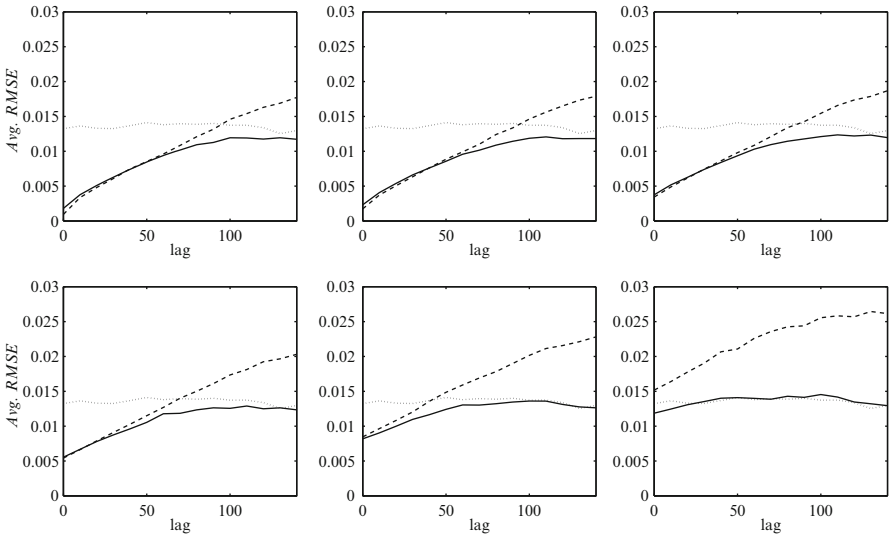


Fig. 18.4 Average *RMSE* as a function of lag in the ILI data: $\sigma_s = \sigma_i = 0.001; \sigma_{synd} = 0.05; \sigma_{ILI} = 0.001, 0.002, 0.005$ (*top left to right*), 0.01, 0.02, 0.05 (*bottom left to right*); *dotted line* for a case where only syndromic surveillance is used, *dashed line* for ILI only, and *solid line* for both sets of data

Across all levels of σ_{ILI}, the average *RMSE* displays a consistent pattern, and we can make the following few observations.[2] First, when using only ILI data (dashed line), average *RMSE* monotonically increases. An intuitive explanation for this monotonic increase is that for a given level of uncertainty, *value* of such measurement information decreases as their acquisition is more delayed.

Second, the monotonic increase observed in the average *RMSE* seems to approach a certain limit. This is particularly visible in the last subplot of Fig. 18.4.[3] Again, we may offer an explanation based on intuition. A very large measurement lag would make measurement information obsolete, and at an extreme, it will be equivalent to having no (useful) measurement data at all. In this case, we will be left with a system model only, and our estimation of system states will be as good as the system model (its process noise). Thus, the average *RMSE* would approach to a limit, which depends on the underlying process noise.

Third, the average *RMSE* when using the both measurement data stays *below* the *RMSE* curves for single data case. While this is also rather expected – state estimation using both data is better or at least as good as that of using only

[2]Dotted lines in six subplots of each figure are identical since it is not a function of σ_{ILI}. They remain to be more or less constant along the x-axis (lag) since it is not a function of lag either. Slight variations visible in Fig. 18.4 are due to non-systematic causes.

[3]This asymptotic behavior is more visible in the cases of $\sigma_q = \sigma_\gamma = 0.005$, which was not presented due to space limitation.

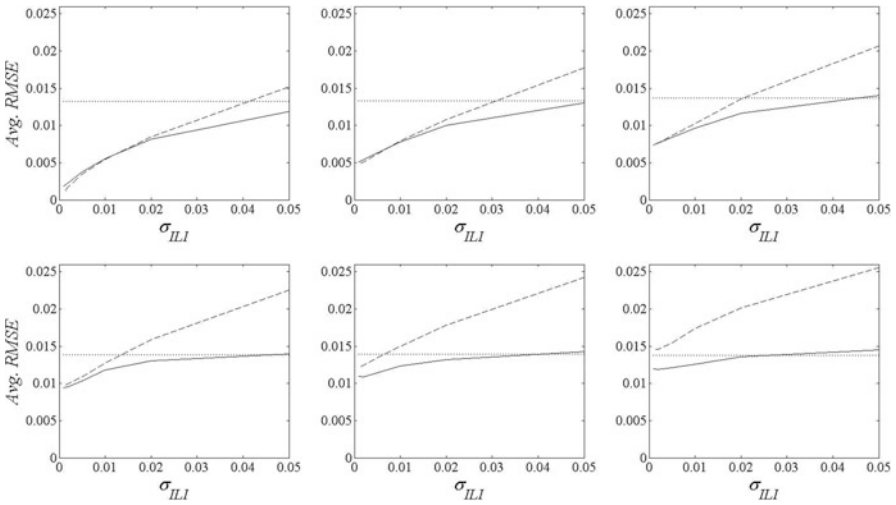


Fig. 18.5 Average *RMSE* as a function of σ_{ILI} for a fixed lag: $\sigma_s = \sigma_i = 0.001; \sigma_{synd} = 0.05;$ lag $= 0, 20, 40$ (*top left to right*), 60, 80, 100 (*bottom left to right*); *dotted line* for a case where only syndromic surveillance is used, *dashed line* for ILI data only, and *solid line* for both sets of data

one set of data –, a closer examination suggests a more interesting behavior. It approaches to the *RMSE* curve of ILI-only when the lag goes to zero and to the syndromic-surveillance-only curve when the lag becomes very large. We also note that the difference between $min\{RMSE_{synd}, RMSE_{ILI}\}$ and $RMSE_{synd+ILI}$ seems to be maximized where the two *RMSE* curves of individual-data intersect. A following hypothesis for this observation is possible: when the value of one of the two sets of measurements dominates the other, benefit of using both sets of measurement diminish and its state estimation is no better than when using the superior measurement data. Using both measurement is most rewarded when the two individual measurements have *comparable* value.

Figure 18.5 presents the same experimental results along σ_{ILI} axis for a fixed lag. It displays the same pattern observed in Fig. 18.4. As σ_{ILI} gets small, the average *RMSE* using both data sets (solid line) approaches ILI-only curve, and vice versa. The benefit of using both sets of data appears to be maximized when the two *RMSE* curves of individual data case intersect.

Recall that the first motivating questions for our work was "given immediately-available-but-highly-uncertain data and credible-but-time-delayed data, which one would yield better state estimates?" The experimental results suggest that it depends on relative uncertainty and an amount of delay, which is reasonable to expect. The second question was "would it be always better to use both sets of data than using only one?" The answer seems to be not always so. It is advantageous to use both sets of data when they have *comparable values*. Otherwise, it is only as good as using measurement data of a higher value.

All these results seem to indicate that there is a systematic mechanism behind the patterns observed in the above figures. It certainly warrants further investigations using analytic models to support the observations and conjectures mentioned above. Below, we present a brief sketch on our preliminary analysis. While it needs further extension and refinement, it shows a promising pathway to a precise analysis for future work.

Consider the following simple model where we attempt to compute a posterior density for a state variable x, $p(x|z_1, z_2)$, given two types of measurements, z_1 and z_2. For now, we assume a normal distribution for its prior, $p(x)$, and likelihood, $p(z_1|x)$ and $p(z_2|x)$. That is, $p(x) = N(\mu_0, \sigma_0^2), p(z_1|x) = N(x, s_1^2)$, and $p(z_2|x) = N(x, s_2^2)$. Let β_0 denote a precision of $N(\mu_0, \sigma_0^2)$, i.e., $\beta_0 = 1/\sigma_0^2$. Likewise, $b_1 = 1/s_1^2$ and $b_2 = 1/s_2^2$.

A posterior density of x given a measurement z_1 is written as $p(x|z_1) \propto p(z_1|x)p(x)$. Since we assume a normal prior and normal likelihood, we know the posterior is also a normal density, $N(\mu_1, 1/\beta_1)$, where $\beta_1 = b_1 + \beta_0$ and $\mu_1 = (b_1 z_1 + \beta_0 \mu_0)/\beta_1$. Precision of a posterior is improved by adding the precision of a measurement, and its mean is an average of the prior mean and measurement weighted by relative precision of each. Similarly, $p(x|z_2) \propto N(\mu_2, 1/\beta_2)$ where $\beta_2 = b_2 + \beta_0$ and $\mu_2 = (b_2 z_2 + \beta_0 \mu_0)/\beta_2$.

When both measurement data are given, a posterior density can be written as a factorized form: $p(x|z_1, z_2) \propto p(z_1|x, z_2)p(z_2|x)p(x) = p(z_1|x)(z_2|x)p(x)$. Note the conditional independence between z_1 and z_2 (i.e., $z_1 \perp z_2|x$) is used in the second step. $p(x|z_1, z_2)$ is a product of three normal densities, and it is straightforward to see that it is a normal density $N(\mu_{12}, 1/\beta_{12})$ with $\beta_{12} = b_1 + b_2 + \beta_0$. Thus, we have the following results for the simple model:

$$\sigma_1^2 = 1/\beta_1 = \frac{1}{b_1 + \beta_0} = \frac{1}{1/s_1^2 + 1/\sigma_0^2}; \quad \sigma_2^2 = 1/\beta_2 = \frac{1}{b_2 + \beta_0} = \frac{1}{1/s_2^2 + 1/\sigma_0^2}$$

$$\sigma_{12}^2 = 1/\beta_{12} = \frac{1}{b_1 + b_2 + \beta_0} = \frac{1}{1/s_1^2 + 1/s_2^2 + 1/\sigma_0^2}$$

For a fixed value of s_1 and σ_0, we can compute $\sigma_{12}^2, \sigma_2^2, \sigma_1^2$ by varying s_2. Figure 18.6 shows a plot of $\sigma_{12}, \sigma_2, \sigma_1$ as a function of s_2.

Now, let index 1 and 2 represent the syndromic surveillance data and ILI data, respectively. σ_1^2 is then a posterior variance given syndromic surveillance data only, and σ_2^2 for ILI data only. σ_{12}^2 is a posterior variance when both sets of data are given. Then, Fig. 18.6 is analogous to Fig. 18.5, and we see that the behavior observed in the experimental results is almost exactly reproduced in Fig. 18.6. Using the formula for $\sigma_1^2, \sigma_2^2, \sigma_{12}^2$, it can be further shown that the benefit of using both sets of data is maximized when $\sigma_1 = \sigma_2$, which is consistently observed in Figs. 18.4–18.5. Hence, with some reservations on the assumptions made in the simplified model, we see that the behavior exhibited by $RMSE_{synd+ILI}$ curve is a logical consequence of combining two sets of data within a Bayesian framework.

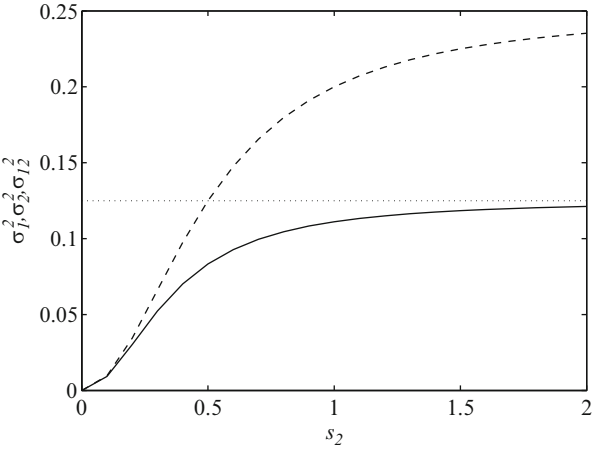


Fig. 18.6 σ_{12}(*solid*), σ_2(*dashed*), σ_1(*dotted*) as a function of s_2

18.5 Conclusion

We study a problem of estimating current epidemic state by combining syndromic surveillance data and ILI data through particle filtering. The two sets of data compliment each other: syndromic surveillance data is immediately available but contains large noise while more reliable ILI data is delayed by some lag in its reporting process. Our experimental results show that using both sets of data is advantageous only when informative value of the two data sets is comparable. Preliminary analysis on a linear, Gaussian case suggests that this behavior is a logical consequence of using a Bayesian stochastic filtering framework. Considering there is a possible trade-off between timeliness and credibility of clinically validated surveillance data, appropriate design of surveillance data collection and processing is a valid optimization problem. We believe that understandings and insights as well as the state estimation method presented in this paper will aid in such decision makings for public health authorities.

Acknowledgements This research was supported by the Public Welfare&Safety Research Program (No.2011-0029881, No.2011-0029883), and by Basic Research Program (No.2010-0025224) through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology.

References

1. Chen, L., Achrekar, H., Liu, B., Lazarus, R.: Vision: towards real time epidemic vigilance through online social networks. In: ACM Workshop on Mobile Cloud Computing and Services, San Francisco (2010)
2. Chew, C., Eysenbach, G.: Pandemics in the age of twitter: content analysis of tweets during the 2009 H1N1 outbreak. PLoS ONE **5**(11) (2010). doi:10.1371/journal.pone.0014118

3. FluView: A weekly influenza surveillance report. Centers for Disease Control and Prevention. http://www.cdc.gov/flu/weekly/. Accessed 5 Feb 2013
4. Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L.: Detecting influenze epidemics using search engine query data. Nature **457**, 1012–1014 (2009)
5. Henning, K.J.: Overview of syndromic surveillance: what is syndromic surveillance? Morbidity and mortality weekly report. Cent. Dis. Control Prev. **53**(Supplement), 5–11 (2004) http://www.cdc.gov/mmwr/preview/mmwrhtml/su5301a3.htm.
6. Jegat, C., Carrat, F., Lajaunie, C., Wackernagel, H.: Early detection and assessment of epidemics by particle filtering. In: Soares, A., Pereira, M., Dimitrakopoulos, R. (eds.) geoENV VI – Geostatistics for Environmental Applications, pp. 23–35. Springer, Netherlands (2008)
7. Lampos, V., Bie, T., Cristianini, N.: Flu detector – tracking epidemics on twitter. In: Balcazar, J., Bonchi, F., Gionis, A., Sebag, M.(eds.) Machine Learning and Knowledge Discovery in Databases, pp. 599–602. Springer, Heidelberg (2010)
8. Ong, J.B.S., Chen, M.I-C., Cook, A.R., Lee, H.C., Lee, V.J.: Real-time epidemic monitoring and forecasting of H1N1-2009 using Influenza-Like Illness from general practice and family doctor clinics in Singapore. PLoS ONE **5**(4) (2010). doi:10.1371/journal.pone.0010036
9. Orton, M., Marrs, A.: Particle filters for tracking with out-of-sequence measurements. IEEE Trans. Aerosp. Electron. Syst. **41**(2), 693–702 (2005)
10. Ristic, B., Arulampalam, S., Gordon, N.: Beyond the Kalman filter: particle filters for tracking applications. Artech House, Boston (2004)
11. Shaman, J., Karspeck, A.: Forecasting seasonal outbreaks of influenza. Proc. Natl. Acad. Sci. (2012). doi:10.1073/pnas.1208772109
12. Skvortsov, A., Ristic, B.: Monitoring and prediction of an epidemic outbreak using syndromic observations. Math. Biosci. **240**, 12–19 (2012)

Chapter 19
A Decision-Making Approach Supporting Hospital Drug Logistics

Anna Corinna Cagliano, Sabrina Grimaldi, and Carlo Rafele

Abstract The use of innovative technologies to improve the hospital drug management process, and in particular its part concerned with logistics, needs to be supported by appropriate decision-making models basing their assessments on both technological issues and the characteristics of healthcare organisations. The frameworks existing in literature are sometimes too complicated to guide towards clear solutions, especially when scarce and/or qualitative information is available. This work presents a new decision-making approach in order to provide hospital managers with a simple but complete tool assisting in the selection of adequate technologies to make drug logistics more efficient. Its first application to some of the most diffused technologies is discussed.

19.1 Introduction

The today's hospital mission of providing patients with an increased level of service is faced with a very limited amount of resources, especially economic ones. In such a context, a careful organisation of material, equipment, and information flows is crucial to improve the healthcare performance while reducing costs [1]. In particular, the logistics of drugs plays an important role because the associated purchasing and managing expenditures have a high impact on budgets.

However, the hospital drug management process is currently affected by significant limitations such as insufficient warehouse spaces, high inventory values, expired products, ward demand that greatly exceeds the real needs, and reduced time spent on clinical activities by healthcare operators [2, 3]. The awareness that this process is still far from being optimised has induced national and international

A.C. Cagliano (✉) • S. Grimaldi • C. Rafele
Department of Management and Production Engineering, Politecnico di Torino, corso Duca degli Abruzzi 24, 10129 Torino, Italy
e-mail: anna.cagliano@polito.it; sabrina.grimaldi@polito.it; carlo.rafele@polito.it

A. Matta et al. (eds.), *Proceedings of the International Conference on Health Care Systems Engineering*, Springer Proceedings in Mathematics & Statistics 61,
DOI 10.1007/978-3-319-01848-5_19,

healthcare organisations to find and test alternative ways to improve drug logistics [4]. In particular, a great variety of technologies, here intended as devices and procedures provided within the healthcare system as it delivers health services [5], have been introduced, many of them inspired by Just In Time (JIT) principles that are already well-known in the manufacturing industry.

The use and dissemination of innovative healthcare technologies need to be supported by appropriate evaluation models relying on an in-depth understanding of their advantages and disadvantages and on a comparison with the peculiar characteristics of the organisations at issue. Many conceptual frameworks existing in literature can be useful to this purpose but they may be either too complicated to guide towards clear solutions or not enough based on empirical evidence.

This work proposes a novel decision-making framework in order to provide hospital managers with a simple approach assisting in a preliminary selection of adequate technologies to make drug logistics more operationally and economically efficient. The paper also discusses its application to a context that does not implement any of the advanced solutions available for stocking, distributing, and administering drugs. Then, benefits and limitations as well as future research directions are highlighted.

19.2 The Hospital Drug Management Process

Drug management in hospitals is a cross-functional process that includes all the activities from prescription to administration [6]. It starts with the physician's prescription in the patient record and its subsequent transcription in the medication record. Every day nurses check the availability of the required drugs on the medication cart and pick the missing ones from the ward inventory. After patient identification, nurses read the associated treatment in the medication record, find the necessary drugs on the cart, and administer them. In parallel to the clinical part of the drug management process, nurses and administrative staff periodically check the level of ward inventory, request products to the hospital pharmacy, receive and stock them, and manage expiration dates. The hospital pharmacy in turn prepares and delivers the products requested by wards, monitors its level of inventory, and orders and receives drugs from suppliers.

This hospital drug management process suffers from several criticalities. Drugs are often stocked without a precise codification and a same product may be placed in multiple inventory positions, thus determining an inefficient space utilisation and making it difficult to apply a First-In-First-Out (FIFO) rule. A poor control over actual consumption, together with the unpredictability affecting healthcare demand, causes high inventory levels both at wards and in hospital pharmacies, with considerable holding costs, obsolescence risk, and an ultimate negative effect on the hospital service level. In addition, drug requesting to the hospital pharmacy is usually performed independently from prescription and administration and quantities are determined based on personal experience and knowledge and not on formalised

inventory management models. Furthermore, nurses tend to be overly involved in logistics duties reducing the time available for clinical activities [7]. Finally, a number of errors related to prescription, transcription, interpretation of treatments, distribution from the hospital pharmacy to wards, and administration may originate from a not optimised drug management process [8]. Adverse drug events (ADEs) are connected to incorrect drug handling in the whole drug management process, being the 70–80% of clinical risk related to therapy prescription, transcription, and administration and the remaining 20% influenced by the preparation and distribution of drugs.

Therefore, a process re-engineering aimed at rationalising the flow of materials and replacing a high level of inventory with an enhanced availability of accurate information can benefit hospitals from both the organisational and the medical perspectives.

19.3 Main Technologies for Improving Hospital Drug Logistics

Literature proves that the JIT philosophy is particularly suitable to overcome the weaknesses of the logistics part of the hospital drug management process [9]. Table 19.1 presents the most diffused technologies to ensure the so-called "five

Table 19.1 Main technologies for hospital drug logistics

Technology	Definition	Reference
Kanban Systems	Mobile cabinets that are periodically replenished to their maximum level of stock based on actual consumption	[7]
Automated Dispensing Cabinets and Carts	Cabinets whose drawers are opened through operators' electronic identification. Carts equipped with laptops and barcode or Radio-frequency identification (RFID) readers	[10]
Computerised Physician Order Entry Systems	Physicians' prescriptions are recorded on laptops or palmtops and may be associated with patients by scanning their identification wristbands	[6]
Unit Dose	Drugs are packed in single doses, each of them identified by a barcode	[11]
Personalised Dose	The pharmacy combines the single doses of drugs in packages containing the therapy for each patient according to medical prescriptions	[3]
Automated Patient Identification	Wristbands equipped with barcodes or RFID tags worn by patients and nurses to allow their identification	[12]

rights", that is the right drug, to the right patient, at the right time, in the right dose, and in the right way. They will be used to illustrate the developed decision-making framework.

19.4 A Decision-Making Framework to Assess Technologies for Drug Logistics

In order to assist in a first selection of technologies supporting the hospital drug management process the authors, in collaboration with a panel of clinical (physicians and nurses) and engineering (logistics and information systems) experts from a number of Local Healthcare Agencies in Northern Italy, developed an approach taking into account not only pure technological features but also specific elements concerning the hospital environment. The following sections describe the proposed framework as well as the results of its first application.

19.4.1 The Proposed Framework

The framework to assess technologies is founded on comparison aspects and weights expressing the importance of each aspect. To be more precise, four different aspects are considered: each of them is further detailed in sub-aspects (for a total of 23 sub-aspects) that can be more easily estimated by people working in healthcare organisations. A review of the existing literature suggested taking into account technological and organisational requirements, economic factors, and the impacts of the reduction of clinical errors. Following the list of aspects and sub-aspects. The latter were determined based on the knowledge and experience of the panel of experts.

- *Information technology*: adaptability to existing technology systems (I1), traceability of information (I2), rapidity of implementation (I3), independence from suppliers' technical support (I4) [12].
- *Organisational*: easiness of implementation (O1), physician time reduction (O2), nursing time reduction (O3), space reduction and order (O4), reversibility of technological solutions (O5) [13].
- *Economic*: limited investment (E1), rapid return of the investment (E2), independence from suppliers' financial solidity (E3), reduction in drug consumption (E4), reduction in ward inventory (E5), reduction in expired products (E6), reduction in operational staff (E7), significant reduction in premium (E8) [13].
- *Risk management*: prescription (R1), understanding, and transcription of therapies (R2), management of ward cabinets (R3), therapy preparation (R4), patient identification (R5), therapy administration (R6) [8, 13].

Table 19.2 Rating scale for the semi-quantitative analysis

Rating scale	No impact	Low negative impact	Medium negative impact	High negative impact
	\	−	−−	−−−
	No impact	Low positive impact	Medium positive impact	High positive impact
	\	+	++	+++

Table 19.3 Weights of aspects

	Information technology	Organisational	Economic	Risk management
Scenario 1	0.25	0.25	0.25	0.25
Scenario 2	0.1	0.20	0.30	0.4
Scenario 3	0.1	0.1	0.3	0.5

Since no specific and complete data about experiments on this issue are currently available from Italian healthcare institutions, the comparison between candidate technologies for the drug management process is performed by using both a qualitative and a semi-quantitative approach.

The aim of the first analysis is making a qualitative assessment of the effects technologies may have on the considered comparison aspects. This allows a preliminary differentiation among technologies and their potential applications. To be more precise, a "Yes" assessment means a positive impact of a technology on a specific sub-aspect, a "No" assessment means a negative impact of a technology on a sub-aspect, and an "Indifferent" assessment indicates that a technological solution does not impact a given sub-aspect (see the application of the framework, first part of Tables 19.4 and 19.5).

Once the qualitative assessment is complete, the analysis is detailed by associating a semi-quantitative evaluation of the impact to each technological solution and sub-aspect (see the application of the framework, second part of Tables 19.4 and 19.5) according to the symbols in the rating scale presented in Table 19.2.

Thus, the more positive the impact of a technology on the aspects included in the proposed decision-making approach, the higher its score. Similarly, the more negative the impact, the lower its score. An indifferent assessment expresses that the introduction of a new technology does not change the existing situation.

Then, in order to allow a robustness analysis of the results, weights are assigned to the four aspects in the framework. As an example, Table 19.3 shows three possible scenarios. In the first scenario all the weights are set equal to 25%, while in the other two ones more importance is given to the Economic and Risk management aspects due to the attention these issues are recently receiving in healthcare environments [13].

In order to keep the method as simple as possible, the score of a technology in each comparison aspect is calculated by adding up positive and negative symbols in the related sub-aspects. Each symbol is associated with a value ranging from +1

Table 19.4 Qualitative and semi-quantitative assessments of technologies

Information technology aspect						
Technology	I1	I2	I3	I4		Total score
Kanban Systems	(Yes)++	(Ind)\	(Yes)++	(Yes)+++		7
Automated Dispensing Cabinets and Carts	(Yes)+	(Yes)+	(No)−	(No)-		0
Automated Dispensing Cabinets with Drawers	(Yes)+	(Yes)++	(No)−	(No)-		1
Computerised Physician Order Entry Systems	(Yes)+	(Yes)++	(No)−	(Yes)+		3
Unit Dose	(No)−−	(Yes)++	(No)−−	(No)−−		−4
Personalised Dose	(No)−−	(Yes)+++	(No)−−	(No)−−		−3
Automated Patient Identification	(Yes)+	(Yes)++	(No)−	(Yes)+		3
Organisational aspect						
Technology	O1	O2	O3	O4	O5	Total score
Kanban Systems	(Yes)++	(Ind)\	(Yes)+	(Yes)+	(Yes)++	6
Automated Dispensing Cabinets and Carts	(No)−	(Ind)\	(No)−	(Yes)++	(Yes)+	1
Automated Dispensing Cabinets with Drawers	(No)−	(Ind)\	(No)−	(Yes)++	(Yes)+	1
Computerised Physician Order Entry Systems	(No)−	(No)−	(Yes)+++	(Ind)\	(Yes)+	2
Unit Dose	(No)−−	(Ind)\	(Yes)++	(Yes)++	(No)−	1
Personalised Dose	(No)−−	(Ind)\	(Yes)+++	(Yes)+++	(No)−−	2
Automated Patient Identification	(No)−	(Ind)\	(No)−−	(Ind)\	(Yes)+	−2

Table 19.5 Qualitative and semi-quantitative assessments of technologies

Economic aspect									
Technology	E1	E2	E3	E4	E5	E6	E7	E8	Total score
Kanban Systems	(Yes)+++	(Yes)++	(Yes)+++	(Yes)++	(Yes)++	(Yes)++	(Ind)\	(Ind)\	14
Automated Dispensing Cabinets and Carts	(No)−	(No)−	(No)−−	(Yes)++	(Yes)++	(Yes)++	(Ind)\	(Yes)++	4
Automated Dispensing Cabinets with Drawers	(No)−	(No)−	(Yes)+	(Yes)++	(Yes)++	(Yes)++	(Ind)\	(Yes)++	4
Computerised Physician Order Entry Systems	(Yes)+	(No)−	(Yes)+	(Yes)+	(Yes)+	(Ind)\	(Ind)\	(Yes)+++	4
Unit Dose	(No)−−	(No)−−−	(No)−−	(Yes)+++	(Yes)+++	(Yes)+++	(Yes)+	(Yes)+++	7
Personalised Dose	(No)−−	(No)−−	(No)−−	(Yes)+++	(Yes)+++	(Yes)+++	(Yes)++	(Yes)+++	8
Automated Patient Identification	(Yes)+	(No)−	(Yes)+	(Yes)+	(Ind)\	(Ind)\	(Ind)\	(Yes)+++	10
Risk management aspect									
Technology	R1	R2	R3	R4	R5	R6	Total score		
Kanban systems	(Ind)\	(Ind)\	(Yes)+	(Ind)\	(Ind)\	(Ind)\	1		
Automated Dispensing Cabinets and Carts	(Ind)\	(Ind)\	(Yes)++	(Yes)+	(Ind)\	(Ind)\	3		
Automated Dispensing Cabinets with Drawers	(Ind)\	(Ind)\	(Yes)+++	(Yes)++	(Yes)+	(Yes)+	7		
Computerised Physician Order Entry Systems	(Yes)+++	(Yes)+++	(Ind)\	(Ind)\	(Yes)++	(Yes)++	10		
Unit Dose	(Ind)\	(Ind)\	(Yes)+++	(Yes)++	(Yes)++	(Yes)+++	10		
Personalised Dose	(Ind)\	(Ind)\	(Yes)+++	(Yes)+++	(Yes)+++	(Yes)+++	12		
Automated Patient Identification	(Ind)\	(Ind)\	(Ind)\	(Ind)\	(Yes)+++	(Yes)+++	6		

to +3 according to the number of plus signs, in case of a positive impact, and with a value from −1 to −3 according to number of minus signs, in case of a negative impact. An indifferent assessment is associated with the value 0. The score thus obtained is multiplied by the weight corresponding to the aspect. Finally, the total score of a technology is obtained by adding up its weighted scores for all the aspects of the framework.

Of course, although the structure of the decision-making approach is general, the values assumed by the qualitative and semi-quantitative assessments and by the weights are strictly related to the peculiar characteristics of the drug logistics process where it is applied as well as to the process phases supported by the technologies under investigation.

19.4.2 First Application of the Framework

As a first test, the developed decision-making approach was applied to a hospital in Northern Italy where none of the technologies for improving hospital drug logistics introduced in Sect. 19.3 were in use. Thus, the primary aim of such hospital was having a preliminary understanding of what of these solutions could be potentially suitable.

Tables 19.4 and 19.5 present the qualitative and semi-quantitative assessments. According to the data and information collected and processed, the Kanban technology turns out to be the dominant choice in each of the scenarios presented in Table 19.3 as far as the Information Technology, Organisational, and Economic aspects are concerned. Unit Dose, Personalised Dose, and Automated Patient Identification report negative values in the Information Technology and Organisational aspects. This means that the implementation of such systems may initially generate disadvantages/problems and/or inefficiencies associated with their use. As regards the Risk Management aspect, the scores obtained in Table 19.5 are more homogeneous: Personalised Dose reveals to be the dominant technology in each scenario followed by Computerised Physician Order Entry Systems and Unit Dose. Unlike in the previous aspects, the Kanban technology turns out to be the worst solution compared to the other drug management systems.

Table 19.6 highlights the total scores of the different technologies in the three scenarios. Scenario 1 shows that the Kanban technology is the dominant one: this is due to the weights that are equal in each aspect. Instead, Personalised Dose, Computerised Physician Order Entry System, and Automated Patient Identification total nearly the same values, therefore they can be regarded as comparable. Automated Dispensing Cabinets and Carts as well as Automated Dispensing Cabinets with Drawers appear to be outclassed by the other technologies. Looking at scenario 2 and scenario 3, the Kanban technology is no more the dominant solution but Personalised Dose reports the highest score. Unit Dose and Computerised Physician Order Entry Systems can be considered as equivalent. The situation of Automated

Table 19.6 Summary of the comparisons among single technologies

Technology	Scenario 1	Scenario 2	Scenario 3
Kanban Systems	7	6.5	6
Automated Dispensing Cabinets and Carts	2	2.6	2.8
Automated Dispensing Cabinets with Drawers	3.25	4.3	4.9
Computerised Physician Order Entry Systems	4.75	5.9	6.7
Unit Dose	3.5	5.9	6.8
Personalised Dose	4.75	7.3	8.3
Automated Patient Identification	4.25	5.3	6.1

Table 19.7 Summary of the comparisons among technology mixes

Technology mix	Scenario 1	Scenario 2	Scenario 3
Computerised Physician Order Entry Systems + Kanban + Automated Patient Identification	16	17.7	18.8
Automated Dispensing Cabinets and Carts + Computerised Physician Order Entry Systems + Automated Patient Identification	11	13.8	15.6
Automated Dispensing Cabinets with Drawers + Computerised Physician Order Entry Systems + Automated Patient Identification	12.25	15.5	17.7
Computerised Physician Order Entry Systems + Unit Dose + Automated Patient Identification	12.5	17.1	19.6
Computerised Physician Order Entry Systems + Personalised Dose + Automated Patient Identification	13.75	18.5	21.1

Dispensing Cabinets and Carts and Automated Dispensing Cabinets with Drawers does not change compared to scenario 1: they are again dominated by the other technologies.

The assessment of individual technologies helps understand and compare their strengths and weaknesses but the assessment of combinations of them is even more important when choosing among possible solutions because each technology covers just a partial phase of the drug management process and a complete control over it implies adopting more than one technology.

For this reason, the possibility of combining together single technologies was explored (Table 19.7). Again to keep the approach simple, semi-quantitative scores for combinations of technologies were obtained by summing up the scores of the single technologies in the mix. According to the collected data, the mix of the Computerised Physician Order Entry Systems, Kanban, and Automated Patient Identification technologies emerges to be the best option for the Information Technology, Organisational, and Economic aspects. This emphasises the fact that such mix is quite interesting from different points of view. The other four technology mixes report very similar scores in these aspects and reveal comparable

characteristics. The main reason why the mix of the Computerised Physician Order Entry Systems, Kanban, and Automated Patient Identification technologies is the dominant solution is the fact that the Kanban technology does not need too many pieces of information and economical resources and has a low impact on the organisation of hospital activities.

As far as the Risk Management aspect is concerned, the mix of Computerised Physician Order Entry Systems, Personalised Dose, and Automated Patient Identification technologies is the solution that totals the highest score.

Table 19.7 highlights that in scenario 1 the combination of Computerised Physician Order Entry Systems, Kanban, and Automated Patient Identification technologies outclasses the other solutions which on the contrary report quite similar scores. In scenarios 2 and 3 the mix of Computerised Physician Order Entry Systems, Personalised Dose, and Automated Patient Identification technologies reveals to be the dominant solution due to the greater importance given to the Economic and Risk Management aspects. The mix of Computerised Physician Order Entry Systems, Automated Dispensing Cabinets with Drawers, and Automated Patient Identification technologies is a dominated solution in scenarios 1 and 2, whereas in scenario 3, where the weight assigned to the Risk Management aspect becomes considerable, the difference between its score and the ones of the other solutions is much reduced. Finally, the mix of Automated Dispensing Cabinets and Carts, Computerised Physician Order Entry Systems, and Automated Patient Identification technologies results to be the least interesting option in any scenario.

19.4.3 Analysis of Results

Looking at single technologies, the analysed solutions can be adopted in different phases of the drug management process. Computerised Physician Order Entry Systems and Automated Patient Identification can be placed at the two ends of the process: physician's prescription and administration by nurses. These two activities generate the information flow along the process, allow its control, and have a key role in increasing patient safety. Moreover, Computerised Physician Order Entry Systems influence inventory and consumption management because they enable the knowledge of the actual drug demand, which can be used by the hospital pharmacy to supply wards with an appropriate quantity of materials. However, such technologies do not ensure the synchronisation between drug demand and consumption having no impact on the monitoring of stocks from the hospital pharmacy as far as patients. Computerised Physician Order Entry Systems and Automated Patient Identification can be introduced with limited financial efforts but they largely influence physicians' and nurses' activities. For this reason, they need a strong involvement and commitment of personnel in order to avoid resistance to change that could prevent their successful applications. Finally, these two technologies are very versatile and can be integrated with other solutions.

The remaining technologies that have been investigated may be considered as partially alternative systems for a correct inventory management, with a minor impact on the reduction of clinical risk.

Two of the studied mixes of technologies can be mentioned: Computerised Physician Order Entry Systems, Kanban System, and Automated Patient Identification and Computerised Physician Order Entry Systems, Unit/Personalised Dose, and Automated Patient Identification. The first mix is relatively simple but is able to cover all the phases of the drug management process. It supports the management of the information flow, clinical risk reduction, and the monitoring of ward inventory by the pharmacy thanks to the Kanban system. Although not directly impacting on patients, the Kanban approach is a storing alternative to traditional cabinets that allows a first monitoring of the information flow related to drugs with a very limited investment. Conversely, the second mix is the most sophisticated one from a technological point of view but it ensures the best performance in terms of risk reduction and control of the drug flow from the hospital pharmacy to patients.

The application of the proposed decision-making approach provided the case hospital with an understanding of the benefits and drawbacks of possible technologies to improve drug logistics. Its outcomes will support more sophisticated assessments to choose the specific solutions that should be implemented.

19.5 Discussion and Conclusions

The proposed framework aims to support decision-making about the different solutions improving hospital drug logistics with respect to multiple comparison aspects. Being developed by experts working in the healthcare sector and tested in a hospital, it is particularly appropriate to be applied to real environments. Also, the approach enables a simple but complete analysis of technologies and is straightforward and intuitive and, thus, suitable for those situations where scarce information is available and the use of more complicated approaches would not be possible. The two phases of the methodology, qualitative and semi-quantitative analysis, can be applied to extremely different informational contexts, assuring a great flexibility. The structure of the framework, made up of heterogeneous aspects and sub-aspects, allows taking into account the main perspectives on the issue and prevents from neglecting some of them. Finally, assigning weights to comparison aspects makes possible to give each of them the correct importance, according to the goals of specific healthcare organisations.

However, the developed approach suffers from some limitations. First, it provides a semi-quantitative assessment that heavily relies on subjective knowledge and experience. Second, it just gives a preliminary knowledge about potential technologies that could be adopted and requires to be used together with more specific and objective methods to make a final decision. Finally, although its first application yielded consistent results, it needs a systematic test in multiple situations in order to identify the necessary refinements. Therefore, future research efforts will be directed

towards applying the approach to a wide range of hospital cases, including situations where some technologies to improve drug logistics have been already implemented. This could bring changes in either the number or the nature of the aspects and sub-aspects considered by the framework that could ultimately contribute to a more effective decision-making.

References

1. Landry, S., Philippe, R.: How logistics can service healthcare. Supply Chain Forum Int. J. **5**(2), 24–30 (2004)
2. Nathan, J., Trinkaus, J.: Improving health care means spending more time with patients and less time with inventory. Hosp. Mater. Manag. Q. **18**(2), 66–68 (1996)
3. Rafele, C., Grimaldi, S.,: Aspetti organizzativi nella gestione del farmaco in ambito ospedaliero. In: Agenzia Regionale Socio Sanitaria del Veneto (ed.): Governo del farmaco: elementi organizzativi e tecnologie. Esperienze a confronto, Il Pensiero Scientifico Editore, Rome (2009)
4. Mustaffa, N., Potter, A.: Healthcare supply chain management in Malaysia: a case study. Supply Chain Manag. Int. J. **14**(3), 234–243 (2009)
5. Velasco Garrido, M., et al.: Developing health technology assessment to address health care system needs. Health Policy **94**(3), 196–202 (2010)
6. Wentzer, H.S., et al.: Unintended transformations of clinical relations with a computerized physician order entry system. Int. J. Med. Inform. **76S**, S456–S461 (2007)
7. Persona, A., Battini, D., Rafele, C.: Hospital efficiency management: the just-in-time and Kanban technique. Int. J. Healthcare Technol. Manag. **9**(4), 373–391 (2008)
8. Cunningham, T., et al.: Impact of electronic prescribing in a hospital setting: a process-focused evaluation. Int. J. Med. Inform. **77**, 546–554 (2008)
9. Baum, N.H.: 'Just in time' means more dimes in your pocket: stocking only what your practice needs takes careful planning, but offers big savings. (The Bottom Line). Urology Times **34**(1), 28 (2006)
10. Paparella, S.: Automated medication dispensing systems: not error free. J. Emerg. Nurs. **32**(1), 71–74 (2006)
11. Fontan, J.-E., et al.: Medication errors in hospitals: computerized unit dose drug dispensing system versus ward stock distribution system. Pharm. World Sci. **25**(3), 112–117 (2003)
12. Nanji, K.C., et al.: Overcoming barriers to the Implementation of a pharmacy bar code scanning system for medication dispensing: a case study. J. Am. Med. Inform. Assoc. **16**, 645–650 (2009)
13. Lampe, K., et al.: The HTA core model: A novel method for producing and reporting health technology assessments. Int. J. Technol. Assess. Health Care **25**(2), 9–20 (2009)

Chapter 20
Analyzing the Impact of Lean Approach in Pharmaceutical Supply Chain

Alberto Portioli Staudacher and Alice Bush

Abstract Pharmaceutical industry is experiencing a time of change because of several reasons. This time of change has generated new needs as efficiency and effectiveness. Pharmaceutical Supply Chains need to compete in their industry and the attention paid to those issues is continuously increasing. Some studies have been conducted for this research line, but many of them are non-mathematical and non-numerical studies. The aim of this paper is to present a modeled Supply Chain to understand how Lean Approach can impact on the Pharmaceutical Supply Chain. In fact we implemented some Lean practice along the Supply Chain and we measured the obtained performances.

20.1 Introduction

The Pharmaceutical industry has been defined as the complex of processes, operations and organizations involved in the discovery, development and manufacture of drugs and medications [1].

Pharmaceutical Industry today is experiencing a time of change because of several reasons: reduction of healthcare drug budgets, increased costs to put new medicines on the market, harder global competition and increased cost along the Supply Chain [2]. Thus the Pharmaceutical industry is under tremendous pressure to improve all the related business.

This research field is becoming more and more relevant for all the players involved in this industry, because the whole Chain needs to be efficient to stay competitive in the Pharmaceutical industry.

A.P. Staudacher (✉) • A. Bush
Department of Management, Economics and Industrial Engineering,
Politecnico di Milano, Milano, Italy
e-mail: alberto.portioli@polimi.it; alice.bush@mail.polimi.it

A. Matta et al. (eds.), *Proceedings of the International Conference on Health Care Systems Engineering*, Springer Proceedings in Mathematics & Statistics 61,
DOI 10.1007/978-3-319-01848-5_20,

On the other hand, today to improve performances in terms of effectiveness and efficiency in a specific Company, working on its internal operations is not enough. An open issue is to improve performance along all the stages of the Supply Chain.

As a contribution of the new challenges of Pharmaceutical Industry, in this paper we will present a modeled Pharmaceutical Supply Chain and we will propose the Lean Approach to answer to the new Industry's needs.

20.2 Literature Review

In 2012 Narayana, Pati and Vrat published a literature review of Pharmaceutical Industry showing the major issues of this contest. The authors collected articles published from 1999 to 2009 and they genereted a classification of these studies in two general categories : non-behavioural studies (64% of the collected literature, 304 studies) and behavioural studies (36%, of the collected literature, 304 studies). Considering only the non-behavioural studies, Narayana, Pati and Vrat highlight that the most addressed issue are The Pricing & Medical Expenditures contributing with 21.7%, R&D and Supply Chain Management, with 10.1% and 9.2% respectively.

The number of articles about SCM in Pharmaceutical Industry is 34 and it's interesting to notice that half of them have been published in the last 2 years of the considered period, indicating an increasing interest on this topic.

Narayana, Pati and Vrat also show that the case study is the most frequently used methodology in studying Supply Chain managerial issues in the Pharmaceutical industry. The second one is mathematical modelling and data analysis. The interest in the mathematical methodology is increasing in the last years in this particular sector; this is due to the new need to finding some principles to improve the whole system [3].

In the United States a Securities and Exchange Commission (SEC) intervened in the US Pharmaceutical SCs, catalysing the adoption of Information Sharing Approach along the Chains. This caused a significant inventory reduction along the Supply Chains for Pharmaceutical products [4]. The distributors used investment buying to gain higher margins and then the performances along the whole Supply Chains were un-efficient. The investment buying was replaced with a fee for service model [5] with inventory management agreements. Indeed Pharmaceutical distributors receive fees from manufacturers for the distribution services that the distributors provide.

Other authors investigated the impact of information sharing Approach on Supply Chain performances (see [6, 7]) showing a significant impact on the inventory level reduction.

But information sharing is not always applicable; therefore it is interesting to investigate the impact of alternative approaches.

Lean management represents one of the most effective practices to improve systems, in general, and to reduce inventory, in particular.

A study carried out by Robert E. Spector [8] represents the Lean implementation in the Pharmaceutical industry. The performance index measured has been the inventory turns as it indicates how the company is improving its processes.

Spector concluded that there wasn't a significant overall improvement. Spector claimed that the poor results were due to the fact that Lean was implemented in only one stage of the Supply Chain, rather than at all stages.

In this study we want to deepen the knowledge on the possible impact of adopting Lean Approach in a Pharmaceutical Supply Chain, and to find a possible explanation why Lean is not widely implemented at the Supply Chain level.

In particular, we want to understand the impact of adopting Lean at one stage only, in the Supply Chain, and the impact of adopting Lean at all stages. In order to do so, we built a simulation model of a typical Pharmaceutical Supply Chain, and we simulated the impact of two Lean Practices: reducing order and production batch sizes and focusing on the flow, by adopting a FIFO rule at all stages, rather than a rule focusing on the single stage efficiency, as, for example, minimum setup.

20.3 The Model

20.3.1 The Modeled Supply Chain

A common Pharmaceutical Supply Chain is composed by Primary Manufacturers, Secondary Manufacturers, Distributors, Retailers and/or Hospitals [1].

The Primary Manufacturer is in charged to produce the active ingredient (API or AI). The characterizing processes are chemical synthesis and separation stage. The secondary Manufacturer add to the Primary Manufacturer's output the "excipient", we can summarize the principals processes as granulation, compression, coating, quality control and packaging [1] (Fig. 20.1).

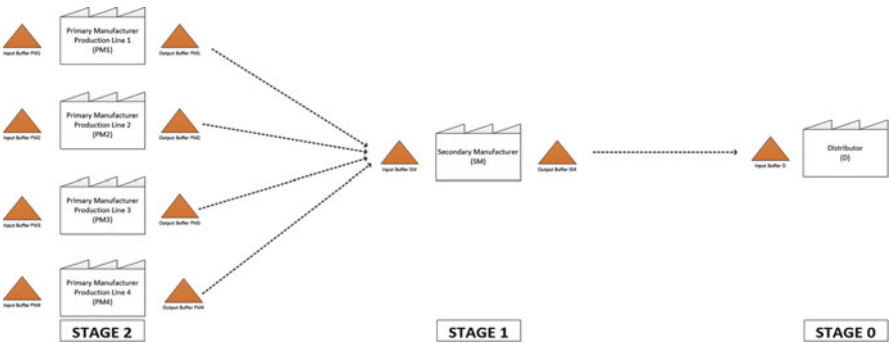


Fig. 20.1 The modeled Supply Chain

We focused on the upstream section of this Supply Chain. There are four different Primary Manufacturers; each one handles a different product line. Every product line is made by six products and differs, from the others, for the input demand.

Each Primary Manufacturer supplies not only this Supply Chain but also others; therefore there is an interaction with other products feeding others Supply Chains. There is no constraint in the availability of raw material. Downstream of the Primary Manufacturers there is a Secondary Manufacturer processing all products and its capacity is fully dedicated to this Supply Chain. Downstream of the Secondary Manufacturer there is a Distributor facing end retailers' and hospitals' demand. Each stage of the Supply Chain is decoupled by an input buffer and an output buffer.

The production stages have set ups to change from a product to another and set ups are shorter if the two products appertain to the same product line, longer if they appertain to different product lines. The orders' processing times are deterministic; they're different for each product and for each stage. When orders are queued to a production stage they're dispatched with a minimum set-up rule, both for the Primaries and the Secondary manufacturers. The value of the processing times is set in order to ensure a saturation index of 89% in stage 1 and 85% in the stage 2.

For the sake of simplicity Production processes have no downtimes due to failures, lack of information or other causes. The deliveries between the different stages are carried out by trucks with limited capacity and the transportation times are deterministic. In the modeled system it's assumed that the trucks cannot move unless a minimum quantity of materials to transport is reached, in order to limit the transportation cost.

As in Chen et al. [9] and Lee et al. [7] we have used the following formula to set the average final customer demand faced by the Distributor:

$$D_t = k + \rho * D_{(t-1)} + \varepsilon * \gamma$$

Where:

k Is a nonnegative constant?
D_t Is the demand in period t (t = 1, 2000);
ρ Is the correlation parameter; in this study $\rho = 0, 7$ (See [7, 9])
$\varepsilon^* \gamma$ Variability factor
ε Parameter normally distributed with mean 0 and variance σ^2.
γ Experiment parameter; in this study $1 \leq \gamma \leq 2$

20.3.2 The Supply Chain Planning Model

The logic that governs each stage is simple: each stage receives an order from the stage downstream and satisfies it from stock. Each stage decides individually when to place an order to the stage upstream and the quantity of each order.

In this model all the stages use an EOQ policy as in Gavirneni et al. [6], Lee et al. [7], Chen et al. [9], to define when to order and how much to order. The logic used considers the delivery times and the production lead times together with the demand faced by the different stages in order to calculate the order batches.

When a customer order arrives to the Distributor, if he has enough availability of stock fulfills completely the incoming order, otherwise the order is backlogged and is fulfilled only when there will be enough stock. If the order is satisfied from stock, a control on the inventory position of this buffer is carried out. If the inventory position is below the order level, the Distributor places an order to the Secondary manufacturer. The order is satisfied from the Secondary Manufacturer output buffer and the inventory position is checked. If it is below the order level, then a production order is generated and queued at the Secondary Manufacturer stage. Next order to produce is selected according to the minimum setup rule (setups are sequence dependent). The production batch size is set equal to 1 week of average demand.

When a production order is manufactured the required material is taken from the input buffer, and the inventory level checked. If the level is below the order level, an order is placed to the stage upstream: the output buffer of the Primary Manufacturer of that product line.

The same mechanism is applied by the upstream stages PM1 (Primary Manufacturer 1), PM2, PM3, and PM4. The production batch size for the Primary Manufacturers is fixed as 2 weeks of average demand. The Primary Manufacturers' input buffers are fed up by a Supplier who has an infinite capacity.

For each buffer the inventory position is calculated as follows:

$$\begin{aligned} &\textit{Inventory Position} = \textit{Stock in the buffer} + \textit{Orders placed to the stage upstream} \\ &\textit{but not yet arrived} - \textit{Order received by the downstream stage, but not delivered} \\ &\textit{yet } (\textit{Backlogged orders}) \end{aligned}$$

20.3.3 Design of Experiment

In this chapter we will present the model's parameters and the design of the experiments set up to analyze the impact of Lean Approach on the Supply Chain. In particular we decided to investigate the impact of set-ups and batch sizes reduction and the impact of reducing the Lead Time variability.

We adopted a discrete-event simulation study because it allows a detailed replication of the behavior of the Supply Chain even under complex configurations and scenarios.

To determine the simulation run length we used the procedure described by Law and Carson [10]. The initial warm-up period was calculated via Welch procedure [11]. The simulation run time as been set to 2,000 days with a warm-up period of 500 days those are not considered during the collection of the statistics. For every tested experiment we made 10 runs with the same parameters but different stochastic numbers, in order to increase the confidence of the results.

Table 20.1 Parameters of simulated model

	PM1	PM2	PM3	PM4	SM
Order processing time (min)	0.662	0.662	0.662	0.662	0.931
Mean set-up time (min)	20.17	20.17	20.17	20.17	25.17
Production capacity dedicated to the Supply Chain	35%	35%	35%	35%	100%
Daily available time for production (hours)	8	8	8	8	8
Transportation time to downstream stage (hours)	16	16	16	16	8
Minimum number of units to start the truck	150	150	150	150	50

To allow an easier comparison of different experiments, we defined a desired service level of 96% and for every experiment we set the order level of each buffer as the minimum quantity in the system to reach this performance target. The most important output of the simulations is the inventory level as in Gavirneni et al. [6] and Lee et al. [10].

The following table present the values of the parameters in the Supply Chain (Table 20.1).

20.3.3.1 Set-Ups and Batch Sizes Reduction

To investigate the impact of making the system more flexible we decreased the set up times and we reduced the production order batch sizes by the same percentage so the overall capacity saturation remains the same, but the system becomes more flexible.

This reduction of set-ups is one of the most important point of the Lean Approach as depicted by Silva et al. [12]. We have tested set-ups reduction ranging between 20% and 45% to both Primary and Secondary Manufacturers.

We tested the impact of reducing set-ups and batch sizes only at the Primary Manufacturers, then only at the Secondary Manufacturer and finally to both stages simultaneously. Finally we investigated the impact of these changes under three different demand variability condition: low, medium and high.

Plan of the first set of the experiments is depicted in the following table (Table 20.2).

20.3.3.2 The Impact of Production Order Sequence Practice: FIFO Versus Minimum Set-Up

The second aspect of Lean Approach we tested is to move the attention from the single stages to the flow. Minimum set-up practice allows an increasing performances to the single stage, while FIFO practice allows minimizing Lead Time variability (see [13, 14]), so the system would be more reliable and more predictable.

We want to highlight that when there are faster and more predictable time reactions, then Safety Stocks decrease and also the Bullwhip effect decreases.

Table 20.2 First set of experiments

	Variability	Percentage of reduction PMs	Percentage of reduction SM
Set-up reduction at PMs only	$\gamma = 1$ $\gamma = 1.5$ $\gamma = 2$	20%/30%/45%	0%
Set-up reduction at SM only	$\gamma = 1$ $\gamma = 1.5$ $\gamma = 2$	0%	20%/30%/45%
Set-up reduction at PMs and SM coordinated	$\gamma = 1$ $\gamma = 1.5$ $\gamma = 2$	20%/30%/45%	20%/30%/45%

Table 20.3 Second set of experiments

	Variability	Percentage of reduction PMs	Percentage of reduction SM	Production order priority rule
Set-up reduction at PMs and SM coordinated	$\gamma = 1$	20%/30%/45%/ 55%/70%/85%	20%/30%/45%/ 55%/70%/85%	MINIMUM SET-UP
Set-up reduction at PMs and SM coordinated	$\gamma = 1$	20%/30%/45%/ 55%/70%/85%	20%/30%/45%/ 55%/70%/85%	FIFO

The minimum setup rule is set as follow: when the production stage completes an order, the queued order that causes the shortest setup is processed next. If a production order has queued for 3 days or more, then it has the priority to be processed, even if the set-up time isn't the shortest one.

The plan of the second set of experiment is depicted in the following table (Table 20.3).

20.4 Results

20.4.1 Set-Ups and Batch Sizes Reduction

First we decreased (−20%/–30%/–45%) the set-ups time and the batch sizes only for the Primary Manufacturers, leaving the Secondary Manufacturer parameters unchanged. Then we returned at the base case for the Primary Manufacturers, and we decreased the set-ups time and batch sizes of Secondary Manufacturer only.

Last we decreased the parameters (set-ups time and batch sizes) simultaneously both for the Primary Manufacturers and Secondary Manufacturer.

The inventories reduction is referred to the base case.

In the figure below we can see the impact of set-ups and batch sizes reduction to Supply Chain inventories in a medium variability contest.

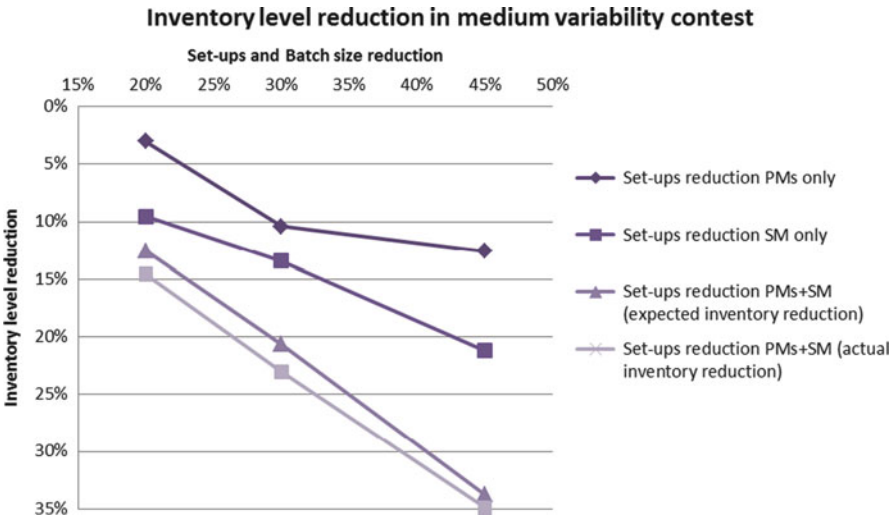


Fig. 20.2 Inventory level reduction in medium variability contest

As we can see in the Fig. 20.2 if we decrease the set-ups time and batch sizes of the Primary Manufacturers, the Supply Chain inventory level can be reduced up to 13% (for set-ups and batch sizes reduction of 45%). On the other hand if we reduce the set-ups time and batch sizes of the Secondary Manufacturer, the Supply Chain inventory level can be reduced until 21% (for set-ups time and batch sizes reduction of 45%).

Reducing only the Secondary Manufacturer gives a greater benefit, in terms of total inventory reduction, because it affects the downstream and upstream stages, while reducing only the Primary Manufacturers influences only the downstream stages (the upstream supplier has an infinite capacity and a fixed lead time).

It is now interesting to add-up the inventory reduction achieved by PMs only and the reduction achieved when reducing SM batch sizes and set-ups time, and compare the result with the actual reduction achieved by reducing set-ups time and batch sizes of both, PMs and SM.

If the actual reduction is the same, it means that it is possible to sum the effects of the two actions, if the actual inventory reduction is larger than the sum of the two single ones, it means that there is a synergetic effect between the two actions.

If the actual result is lower, it means that there is a saturation effect.

The simulation runs show that there is a synergetic effect.

Finally In the figure below we present the impact of the demand variability on the inventory reduction effect:

Figure 20.3 presents the percentage of the difference between the sum of the inventory reduction achieved reducing set-ups time and batch sizes of PMs and SM only, and the inventory reduction achieved by reducing set-ups and batch sizes of PMs and SM simultaneously.

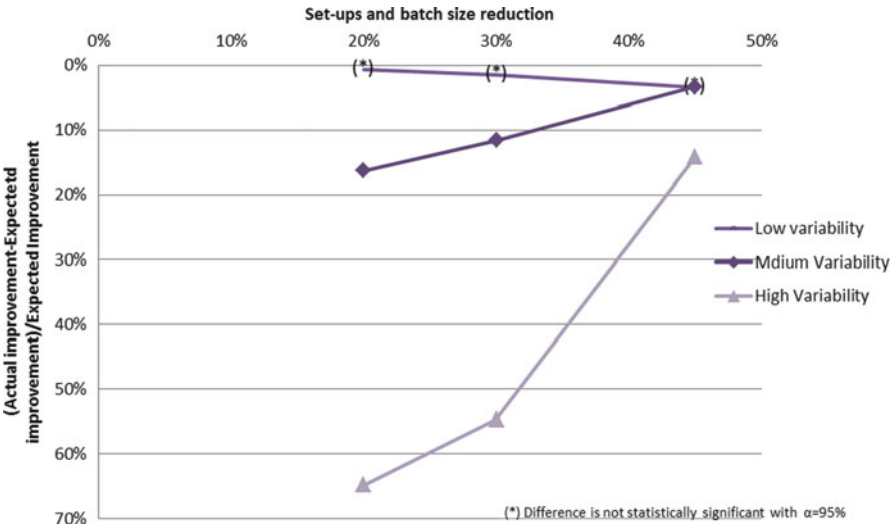


Fig. 20.3 Comparison of expected improvement and actual improvement in three variability contests

Results show that the synergetic effect is higher when demand variability increases, moving from 2% with low variability to 65% with high variability, when set-ups time and batch sizes reduction is set at 20%.

When set-ups time and batch sizes reduction increases, the synergetic effect decreases.

20.4.2 The Impact of Production Order Sequence: FIFO Versus Minimum Set-Up

In the figure below we can see the impact of Production Order sequence practice:

We noticed that for small set-ups time reduction FIFO rule gives to the whole system a disadvantage. In-fact the saturation of the system increases (97%) due to the larger amount of time spent making set-up and the total inventory is higher.

There is a level, decreasing set-ups and batch sizes further, where FIFO is no longer unfavourable. In fact, as we can see in Fig. 20.4, differences of inventory reduction between FIFO and Minimum setup aren't large in the zone between 50% and 70% of set-ups and batches reduction but even inventory level, in FIFO scenario, mainly decrease.

This means that reducing set-ups and batch sizes, not only gives a greater performance to the whole Supply Chain, but it allow to take advantage from the lead times variability reduction.

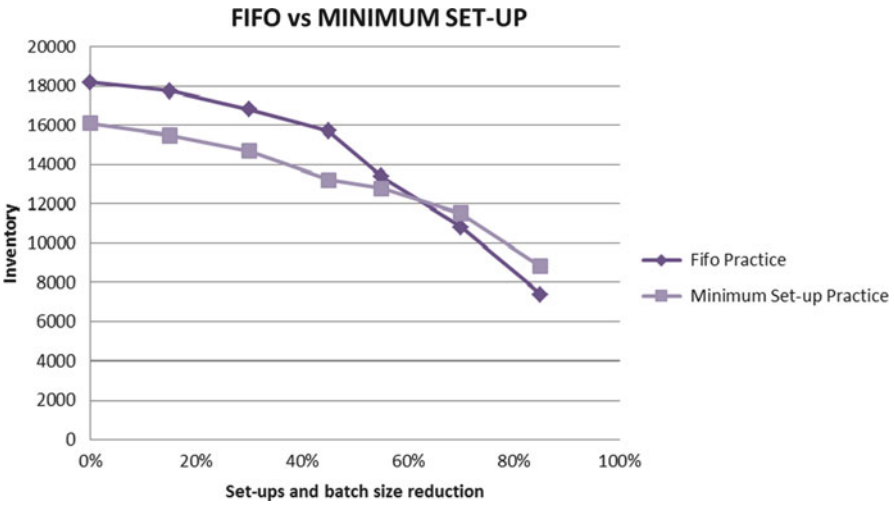


Fig. 20.4 Comparison of two different practice of Production Order sequence practice

20.5 Conclusion

Pharmaceutical Companies are looking for greater efficiency and improving practice along the Supply Chain is a great opportunity to achieve this.

Many authors investigated the impact of adopting an Information Sharing approach (see [4]) and showed that it gives interesting advantages, but other intervenes are possible.

Lean Approach has shown to reduce inventories and lead times, to improve service level, quality, and, in general, the performance of companies, but, there is very little research on quantitative analysis of the impact of adopting Lean Approach, in particular in Supply Chain management, and in Pharmaceutical industry.

The research work presented in this paper, investigates the impact of adopting Lean Approach along a Pharmaceutical Supply Chain, through a simulation model, and a campaign of experiments aimed at measuring and at understanding the impact of adopting typical actions of Lean: reducing set-ups time and batch sizes, and looking at smoothing the flow, by reducing lead time variability.

An important element emerged from the analysis is the synergetic effect of the Lean Approach when it is applied to more than one stage along the Supply Chain.

Reducing set-ups time and batch sizes, not only decreases the inventory level needed to achieve the desired service level, but it also reduces the absolute effect of sequence dependent set-ups time.

This allows to move from set-ups orientated sequencing rules to FIFO rule that gives to the system a strong decrease in lead time variability.

Different players of the Supply Chain have to operate simultaneously and coordinately. They also need to use the same practice and the same strategies. In fact introducing Fifo without bringing set-ups time at the right level make the performances worse as well as reducing batch sizes only in one stage gives to the system a very limited benefit.

References

1. Shah, N.: Pharmaceutical Supply Chains: key issues and strategies for optimization. Comput. Chem. Eng. **28**, 929–941 (2004)
2. Huw, T.: Transforming the pharma industry: lean thinking applied to Pharmaceutical manufacturing
3. Narayana, S.A., Pati, R.K., Vrat, P.: Research on management issues in the Pharmaceutical industry: a literature review. Int. J. Pharm. Healthcare Mark. **6**, 351–375 (2012)
4. Schwarz, L., Zhao, H.: The unexpected impact of Information sharing on US Pharmaceutical supply chains. Interfaces **41**(4), 354–364 (2011)
5. Zhao, H., et al.: Fee-for-service contracts in Pharmaceutical distribution supply chains: design, analysis, and management. Manuf. Serv. Oper. Manag. **14**, 685–699 (2012)
6. Gavirneni, S., et al.: Value of information in capacitated supply chain. Manag. Sci. **45**(1), 17–24 (1999)
7. Lee, H., et al.: The value of information sharing in a two level supply chain. Manag. Sci. **46**(5), 626–643 (2000)
8. Spector, R.: How Lean is Pharma?: A 10-Year Progress Report. http://www.pharmamanufacturing.com/articles/2010/109.html (2010)
9. Chen, F., et al.: Quantifyng the bullwhip effect in a simple supply chain: the impact of forecasting, lead times and information. Manag. Sci. **46**(3), 436–444 (2000)
10. Law, A., Carson, J.S.: A sequential procedure for determining the length of a steady state simulation. Oper. Res. **27**, 1011–1025 (1979)
11. Welch, P.D.: Computer performance modeling handbook. In: The Statistical Analysis of Simulation Results. Academic, New York (1983)
12. Silva, C., Salviano, K., Tantardini, M., Portioli Staudacher, A.: Lean production implementation: a survey based comparison between Italian and Portuguese companies. In: Pre-prints of the 16th International Working Seminar on Production Economics, pp. 481–492 (2010)
13. Portioli Staudacher, A., Tantardini, M.: A Lean-based ORR system for non-repetitive manufacturing. Int. J. Prod. Res. **50**(12), 3257–3273 (2012)
14. Portioli Staudacher, A., Tantardini, M.: Lean implementation in non-repetitive companies: a survey and analysis. Int. J. Serv. Oper. Manag. **11**(4), 385–406 (2012)

Chapter 21
Portable Optokinetic Stimulator for Vestibular Rehabilitation

Cândida Malça, Fernando Moita, and Inês Araújo

Abstract Based on optokinetic simulation technique, a lightweight, compactness, portable and low cost device for vestibular rehabilitation is designed and constructed. All functions are controlled remotely through a multiplatform connectivity especially developed for this application. This multiplatform not only allows for interconnection with different operating systems, but also makes it possible to control the overall clinical parameters and data acquisition and storage of the trial on each patient, thus enabling continuous monitoring of their evolution. These and other advantages are presented and demonstrated throughout this paper.

21.1 Introduction

The optokinetic stimulation is recognized as an effective treatment in vestibular rehabilitation of patients with vestibular etiology or balance center disorders, e.g., unilateral vestibular deficits, vestibular neuritis, Meniere's disease, labyrinthitis, labyrinthine fistula, vertebrobasilar insufficiency, presbivertigo, vestibular paroxysmia, cholesteatoma, acoustic neuroma, bilateral vestibular deficits, elderly

C. Malça (✉)
Mechanical Engineering Department, Instituto Superior de Engenharia de Coimbra, Rua Pedro Nunes, Quinta da Nora, 3030-199 Coimbra, Portugal
e-mail: candida@isec.pt

F. Moita
Electrical Engineering Department, Instituto Superior de Engenharia de Coimbra, Rua Pedro Nunes, Quinta da Nora, 3030-199 Coimbra, Portugal
e-mail: moita@isec.pt

I. Araújo
Audiology Department, Escola Superior de Tecnologia da Saúde de Coimbra, Rua 5 de Outubro, 3046-854 Coimbra, Portugal
e-mail: ines@estescoimbra.pt

A. Matta et al. (eds.), *Proceedings of the International Conference on Health Care Systems Engineering*, Springer Proceedings in Mathematics & Statistics 61,
DOI 10.1007/978-3-319-01848-5_21,

multi-sensory deficits, central vestibular deficits and visual dependence, among others [1–5]. The main purpose of this treatment is to induce a conflict neuro-sensory. The optokinetic nystagmus, induced by a moving optical stimulation, promotes the retinal slip of the target forcing the oculomotor system to a slow motion chase followed by a saccadic movement. The sensory stimulation is therefore promoted to one of the hallways leading to greater efficiency and an increase in the vestibulo-ocular reflex (VOR) [2, 4–6].

Most of optokinetic stimulators commercially available, e.g. [7–9], present some drawbacks namely in what concerns: (i) the number of axes of rotation since some devices only allow two rotational directions, (ii) the impossibility of the patient's ocular fixation as the current equipment doesn't have fixation light spots, which means that the decreases or inhibition of nystagmus spontaneous present in many vestibular disorders can't be worked on and the consequent of increase vestibular reflex gain reached, (iii) the portability and compactness due to their dimensions and weight already devices are not portable; and (iv) the acquisition and storage of the patient's data, as well as the recording of trial conditions are not available and existing equipment does not integrate an interface enabling the establishment of the communication with any operating system and a suitable software to do so.

To overcome these shortcomings a new concept of optokinetic device is developed and a prototype is built. This low cost, portable and compact system enables: (i) the independent rotational motion under the three main axes; (ii) the use of one, two or maximum three adjustable fixation light spots; (iii) that all system features are controlled remotely; and (iv) the patient's report database. These advantages are only possible due to the customized multiplatform interface developed. This multiplatform interface allows for the interconnection of equipment with different operating systems from computers to any mobile device available on the market that has Bluetooth technology. The additional developed software enables the management and control of overall clinical parameters as well as data reception and storage of the patient's trial allowing the continuous monitoring of their evolution. Each one of these characteristics will be described in the following section and the device capacities will be demonstrated through the running work underway. Finally, these results will be presented in the final paper.

21.2 Equipment Description

Figures 21.1 and 21.2 represent the sketch and physical model of the optokinetic stimulator developed, will be now described the components and functionalities that integrate the system. The mirror ball (1) sheds light illuminating the whole area of the room where the trial will take place through the LED coupled to the articulated arm (3),and its orientation is manually adjustable but the intensity remotely controlled by the customized multiplatform interface developed and shown in Fig. 21.3. The light spot fixation is achieved by the laser located at

Fig. 21.1 Sketch model of the optokinetic simulator developed

Fig. 21.2 Prototype of the optokinetic simulator developed

the end of articulated arm (4) whose orientation and intensity are also adjustable. The articulated arms (3) and (4) are fixed to the cover/box (5) through a support suitably designed to incorporate two more articulated arms. The mirror ball is supported by three spheres of equal size (2), which in turn are inferiorly and

Fig. 21.3 Customized multiplatform interface developed prototype

supported laterally and from the bottom by smaller ball bearings. These spheres receive the rotation motion produced by the electric motors and transmit it to the mirror ball (1) according to the three possible axes of rotation.

The angular velocity and direction of rotational motion produced by each motor are controlled individually by each one of the three power drivers, which in turn are controlled by the microcontroller ATmega328. The microcontroller board communicates via wireless with the interface management multiplatform specifically developed for this application. As aforementioned this tool is illustrated in Fig. 21.3 and allows the interconnection between the optokinetic equipment with different operating systems from computers to any mobile device available on the market with Bluetooth interface. Furthermore, this interface allows for the control of overall clinical parameters as well as for the receipt and storage of data from each patient's clinical trial, thus enabling continuous monitoring of their evolution.

21.3 Kinematics of the Optokinetic Equipment

The Optokinetic Stimulator is compact and portable, allowing for the free rotation around the three axis. The mirror ball is removable and supported by three rubber balls. This is only possible due to the special geometry and mechanics of the motorized platforms. Figures 21.4 and 21.5 show lateral and top views of the optokinetic system.

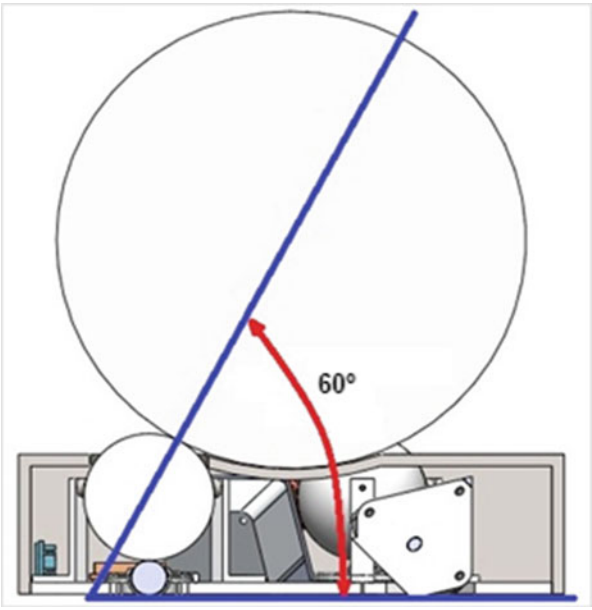


Fig. 21.4 Lateral view of the optokinetic mechanics geometry

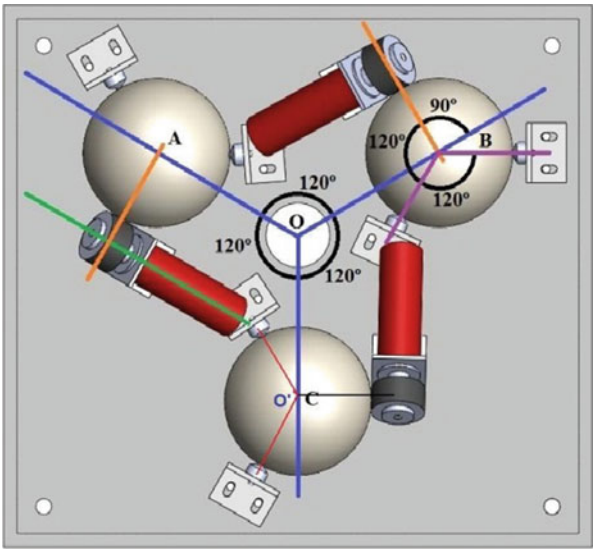


Fig. 21.5 Top view of the optokinetic mechanics geometry

Considering the optokinetics mechanics a rigid body described as a subset of $\mathbb{R}^3$, the individual point movements in the body can be deduced using the general robotic kinematics mathematics [10].

To rotate the mirror ball with a specific direction and velocity, we need to compute the contribution from each independent motors and the associated rotational speed, normal to the motor axis, and defined by vector $W_m = (w_{m1}, w_{m2}, w_{m3})^T$.

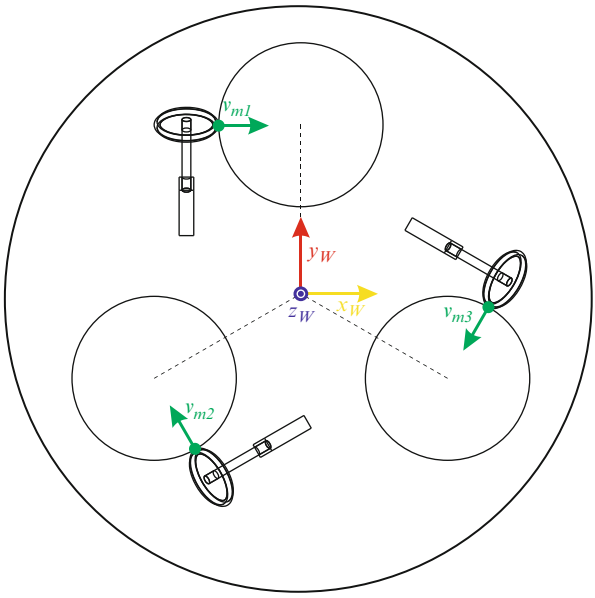


Fig. 21.6 Kinematic diagram of the omnidirectional mirror ball transmission (*top view*)

In order to achieve this, we consider the interfaces among the motor flanges and the rubber balls sliding frictionless normal to the motor shaft axis. Also, we consider the coupling among rubber balls and the main mirror working without slippage.

The axis and speed of rotation for the main ball can be specified by a 3D angular velocity vector $W = (w_x, w_y, w_z)^T$, where:

$$\text{Angular Speed} = \frac{\partial angle}{\partial t} = |W(t)| = \sqrt{W_x^2 + W_y^2 + W_z^2} \tag{21.1}$$

$$\text{Normalized axis} = (w_x, w_y, w_z) \,/\, |W(t)| \tag{21.2}$$

W is expressed in the world coordinate system, fixed in inertial space and denoted by $(x_W,\ y_W, z_W)$ (see Fig. 21.6).

As we are interested in deriving the angular velocities vector W_m transferred by motors flanges to the rubber balls, we need to consider the $(x_B,\ y_B, z_B)$ and $(x_b,\ y_b,\ z_b)$ body coordinate system, fixed in the inertial space and located at the geometrical center of each ball as represented in Fig. 21.7.

Figure 21.7 represents a lateral view of the body coordinate systems. To obtain the tangent linear velocity vector v_t at point p we start converting the angular velocity vector W expressed in the world coordinated system to body fixed coordinates. According to the optokinetic arrangement geometry (Fig. 21.6) two coordinate rotations are necessary: first, a z_W axis rotation, α, is around the geometric center of the mirror ball and a second rotation, β, is always a 60° clockwise rotation about

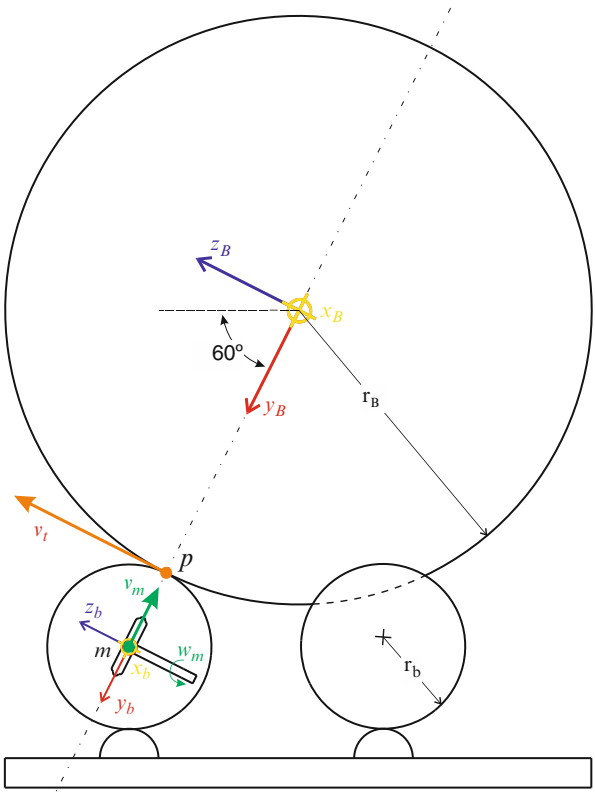


Fig. 21.7 Kinematic diagram of the omnidirectional mirror ball transmission (*lateral view*)

the x_W axis. The new mirror ball angular velocity W_B related to (x_B, y_B, z_B) body coordinate system will be:

$$W_B = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\beta) & \sin(\beta) \\ 0 & -\sin(\beta) & \cos(\beta) \end{bmatrix} \begin{bmatrix} \cos(\alpha) & \sin(\alpha) & 0 \\ -\sin(\alpha) & \cos(\alpha) & 0 \\ 0 & 0 & 1 \end{bmatrix} W \tag{21.3}$$

The point p linear velocity is the cross product of the mirror ball angular velocity and its distance from the center.

$$P_B = (0, r_B, 0) \tag{21.4}$$

$$v_t = W_B \times p_B \tag{21.5}$$

Because the two body coordinate systems are parallel, the point p linear velocity referred to rubber ball coordinate system (x_b, y_b, z_b) should be the same. That way, given W_b, the rubber ball angular velocity vector, and point p distance from its geometrical center, v_t can also be given by:

$$p_b = (0, -r_b, 0) \tag{21.6}$$

$$v_t = W_b \times p_b \tag{21.7}$$

Taking (21.5) and (21.7) we can find the solution for W_b:

$$W_b = \begin{bmatrix} -\frac{r_B}{r_b}(w_x \cos\alpha + w_y \sin\alpha) \\ 0 \\ -\frac{r_B}{r_b}(w_x \sin\alpha \sin\beta - w_y \cos\alpha \sin\beta + w_z \cos\alpha) \end{bmatrix} \tag{21.8}$$

Using (21.8) we can calculate v_m tangent linear velocity at m point that makes the contact among motor flanges and rubber balls.

$$m = (-\mathrm{r_b}, 0, 0) \tag{21.9}$$

$$v_m = W_b \times \mathrm{m} \tag{21.10}$$

$$v_m = \begin{bmatrix} 0 \\ \mathrm{r_B}(w_x \sin\alpha \sin\beta - w_y \cos\alpha \sin\beta + w_z \cos\alpha) \\ 0 \end{bmatrix} \tag{21.11}$$

As expected, v_m has only one component because it is always a tangent vector parallel to y_m axis. Finally, we need to know the angular velocity contribution for each motor. Taking in to account all the system arrangement geometry, the angular contribution for each motor will be defined by:

$$W_m = \begin{bmatrix} 0 & -0.69 & 0.40 \\ -0.60 & 0.34 & 0.4 \\ 0.6 & 0.34 & 0.4 \end{bmatrix} \begin{bmatrix} w_{m1} \\ w_{m2} \\ w_{m3} \end{bmatrix} \left(\frac{cycles}{sec}\right) \tag{21.12}$$

Figure 21.8 illustrates the contribution for each individual motor when the mirror ball is rotating at constant velocity, $\frac{1}{4}\frac{\mathrm{cycle}}{\mathrm{sec}}$, around an axis in the plane xz. For example, when rotating horizontally, rotating axis at 90º or 270º, every motor is providing equal contribution as expected.

21.4 Mechanisms of Vestibular Recovery

The vestibular rehabilitation through the use of new technologies is a continuously emerging field with promising advances to treatment. Adaptation of specific vestibular parameters has been noted after exposure to optokinetic stimulation. This includes changes in the gain of the vestibulo-ocular reflex in primates,

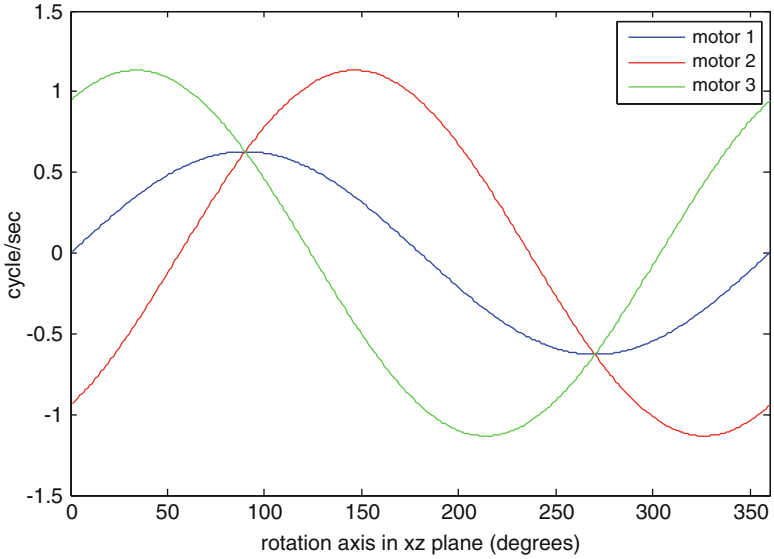


Fig. 21.8 Individual motor contribution for mirror ball rotating with constant velocity, $\frac{1}{4}\frac{cycle}{sec}$, around a axis in the plane xz

healthy individuals and individuals with chronic peripheral vestibular and central disorder. During small field optokinetic stimulation, activation in cortical areas related to visual motion processing and control of eye movement are noted, along with deactivation of parieto-insular vestibular cortex. Neuronal substrates in the cerebellum and brainstem are also involved in the process of horizontal and vertical optokinetic stimulation [5, 11–13].

21.4.1 The Portable Optokinetic Stimulator

The Portable Optokinetic Stimulator light is projected all over the visual field (wall, ceiling and floor), creating effects such as "planetarium" in complete obscurity (those surfaces must be completely homogeneous, without any reference point). The light source should have an adjustable and flexible intensity in order to allow precise focusing on the walls for variable size of the room, and thus a variable size of the light spots. This equipment allows ocular fixation, reducing or inhibiting spontaneous nystagmus, which stemmed mainly from peripheral vestibulopathy.

The patient is standing facing the wall at a distance of 2 m and Optokinetic Portable Stimulator is behind the patient's head level. The patient is instructed to look for the bright spots that pass in front of the wall letting the eyes move freely without moving the head and trying to keep a balance.

The success of vestibular rehabilitation is to find the direction and speed that causes greater instability to the patient, and this made by increasing the difficulty from session to session.

21.5 Work Underway

This study aims to verify the effects of optokinetic exercise. We are carrying out a program of optokinetic stimulation in patients with balanced disorders. These patients underwent a total of ten sessions of rehabilitation through optokinetic exercises, with three weekly sessions lasting 15 min each. To evaluate the success of vestibular rehabilitation (through the use of a new portable optokinetic stimulator) a computerized dynamic posturography (CDP) was performed during the exercises to register the values of deviation from the body center gravity and stability limits. Dizziness Handicap Inventory (DHI) scale is also being used to assess the degree of disability that causes imbalance in the sample and it is applied to the sample before and after the completion of the treatment.

The body posturography device using inclinometer technique and vision is also being developed in our lab in order to be integrated with the Optokinetic Stimulator. This new device detects body sway motion using trunk, head and the eye movement allowing instantaneous relation with optokinetic stimulation [14, 15] (see Fig. 21.9).

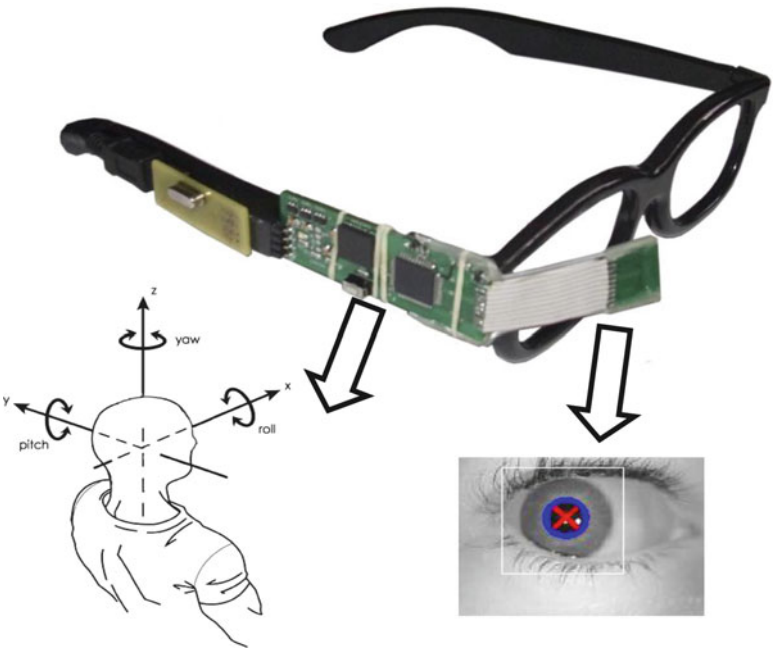


Fig. 21.9 The body posturography device using inclinometer technique and vision to measure eye gaze

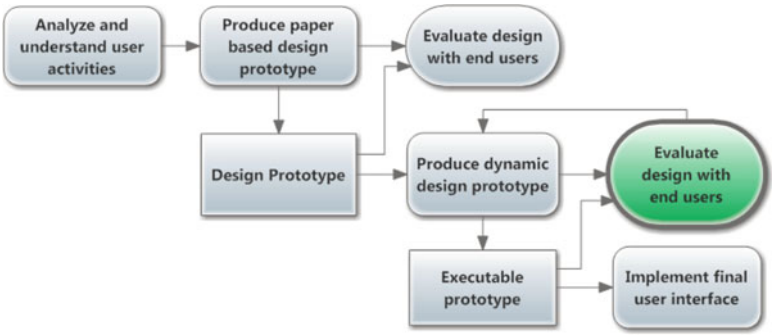


Fig. 21.10 Stages of a Human-Computer-Interface project

With this new equipment it's possible to measure, the deviation from body gravity centre and the eye movement (optokinetic nystagmus), during the optokinetic rehabilitation session, allowing to draw conclusions about the improvements and necessary changes in the rehabilitation strategy.

In the development and design of a medical device interfaces, we must perform the following steps: analysis and understanding of user needs; preparation of schematic development and implementation of the prototype and final implementation; reviewed by users (see Fig. 21.10). At current stage, while testing the rehabilitation strategy, we are evaluating the design with end users in order to improve the usability of the device and also collecting data in realistic situations, to identify all problems that exist in the user interface.

21.6 Conclusion

A new portable, low cost, lightweight and compact optokinetic apparatus for vestibular rehabilitation is presented here. When compared with current optokinetic stimulators, this device is characterized for: (i) integrating a LED light and laser sources with adjustable direction and intensity; (ii) allowing the rotational movement under the three possible axes; and (iii) including a customized multiplatform interface. The angular velocity and direction of rotation of each of the motors as well as the intensity of the LED light and laser are controlled remotely by a Bluetooth interface. The wireless interface also allows the interconnection of equipment with different operating systems by also enabling control over the general clinical parameters as well as the receipt and storage of data from each patient's clinical trial, thus enabling a continuous monitoring of their evolution. These characteristics are an added value from both a technical and economical (financial, monetary, business) perspective.

References

1. Baloh, R., Halmagyi, G.: Disorders of the Vestibular System. Oxford University Press, Inc., New York, (1996)
2. Barros, C., Bittar, R., Bottino, M.: Restoration of the Corporal Balance in Bilateral Vestibular Loss with a Man–machine Interface (MMI): Preliminary Study. Arq. Int. Otorrionolaringol. **11**, 271–277 (2007)
3. Bower, J., Parsons, L.: Rethinking the lesser brain. Sci. Am. **289**, 50–57 (2003)
4. Highstein, S., Fay, R., Popper, A.: The Vestibular System, Springer Handbook of Auditory Research. Springer-Verlag, New York, Inc., **19,** (2004)
5. Huygen, P., Verhagen, W.: Optokinetic response in patients with vestibular areflexia. J. Vestib. Res. **21**(4), 219–225 (2011)
6. Semont, A., Vitte, E., Berthoz, A., Freyss, G.: Repeated Optokinetic Stimulation in Conditions of Active Standing Facilitates the Recovery from Vestibular Deficits. Exp. Brain Res. **102**, 141–148 (1994)
7. http://www.framiral.fr/fr/stimulopt.php, Cited Dec 2012
8. http://www.difra.be/eng/Balance-systems, Cited Dec 2012
9. http://www.technoconcept.fr, Cited Dec 2012
10. Murray, R., Li, Z., Sastry, S.: A Mathematical Introduction to Robotic Manipulation, CRC Press, Taylor & Francis Group, USA (1994)
11. Lopez, C., Blanke, O.: The Thalamocortical Vestibular System in Animals and Humans. Brain Res. Rev. **67**, 119–146 (2011)
12. Pavlou, M, Quinn, C., Murray, K., Spyridakou, C.: The effect of repeated visual motion stimuli on visual dependence and postural control in normal subjects. Gait Posture. **33**(1), 113–118 (2011)
13. Schubert, M., Minor, L.: Vestibulo-Ocular Physiology Underlying Vestibular Hypofunction. Phys. Ther. **84**, 373–385 (2004)
14. Moita, F., Oliveira, R., Santos, V., Silva, M.: EyeSEC Project, Development of interfaces for impaired users. In: 10th Portuguese-Spanish Congress in Electrical Engineering-XCLEEE, Portugal, July 2011
15. Moita, F., Oliveira, R., Santos, V., Silva, M.: Development of interfaces for impaired users. PRZEGLD ELEKTROTECHNICZNY (Electrical Review), ISSN 0033–2097, R. 88 NR 1a/2012

Chapter 22
Modeling and Simulation of a French Extended White Plan: A Hospital Evacuation Before a Forecasted Flood

Wanying Chen, Alain Guinet, and Angel Ruiz

Abstract As high level emergencies can have serious consequences on hospital activities, an emergency management plan to face a crisis situation must be specified and assessed. Even though more and more research is devoted to this area, most studies are based on academic assumptions and the proposed improvement methods are difficult to be applied in the real world. This paper addresses the French Extended White Plan i.e. a hospital evacuation plan facing a flood based on a real life scenario. First, a global model is built, using linear programming to roughly estimate the resources needed for an evacuation and to get a lower bound of the evacuation time. Second, a detail model is proposed in two steps, using the software, ARIS and SIMIO. In the first step, a frame model, which considers the processes of the vested interest actors and the information flow among them, is established to get the sequence of events based on activities from the ARIS diagrams. In the second step, a simulation model, which integrates the information transmission and the different activities, is proposed based on the frame model in the SIMIO programme. The correctness of the detail model has been checked, using the linear model results and the rationality of the simulation model is verified by various experiments. Through experiments, the best way to assign the resource has been found and two organizational improvements have been proposed. With such improvements, 1 h and 18 min is saved in evacuating all the patients and the improvement rate is as high as 26.2%. Thus, our work can provide some guidelines for managers who work in hospitals to improve their evacuation management plan.

W. Chen (✉) • A. Guinet
DISP (laboratoire de Décision et d'Information des Systèmes de Production), INSA de Lyon, Bât. Jules Verne, 19 av. Jean Capelle, 69621 Villeurbanne, France
e-mail: wanying.chen@insa-lyon.fr; alain.guiet@insa-lyon.fr

A. Ruiz
Faculté des sciences de l'administration et CIRRELT, Université LAVAL, 2325 rue de la Terrasse, Québec (Québec) G1V 0A6, Canada
e-mail: angel.ruiz@fsa.ulaval.ca

A. Matta et al. (eds.), *Proceedings of the International Conference on Health Care Systems Engineering*, Springer Proceedings in Mathematics & Statistics 61,
DOI 10.1007/978-3-319-01848-5_22, © Springer International Publishing Switzerland 2014

22.1 Introduction

Hydrological disasters, triggered by flood or wet mass movement (mudslides), bring about heavy loss of life and property damage. For instance, in March 2011, Japan's tsunami and earthquake, accounted for the biggest amount of money distributed for disasters in Japan ($210 billion), as well as the most casualties (15,500 deaths with 7,300 still counted as missing). After Japan, the costliest disaster was the flood from December 2010 to January 2011, in Australia, which caused a $7.3 billion loss [1]. In the Asian continent in 2010, hydrological disasters were responsible for 92.9% of disaster victims, the highest number since the 1980s. Extensive floods and landslides, following heavy monsoonal rains in Southern China affected 134 million people. Floods and flash floods in Pakistan brought about another 20.4 million victims. Hydrological disasters are among the most serious disasters worldwide [2]. Moreover, the impact of those catastrophic floods in Pakistan in July 2010 showed how disaster-risk and poverty are closely interlinked [2]. An effective way to tackle the serious hydrological disaster consequences is to establish a prudent and comprehensive emergency management plan.

The emergency management plan is made up of four phases: mitigation, preparedness, response, and recovery [3]. Mitigation efforts are intended to keep dangers from escalating into disasters and to reduce the impact of disasters, thus reducing loss of life and damage to property if a disaster happens. Preparedness means that we take measures beforehand, to reduce the impact of disasters on human beings. The response phase seeks to minimize the consequence of a disaster. The aim of the recovery phase is how to restore the affected area to its normal state. To minimize human suffering and deaths is the most important criteria in the emergency logistic management amongst the numerous objectives, such as minimizing the economical cost and reducing the impact to the environment [4]. Realizing the importance of the emergency management plan, many countries have made different emergency management plans, based on the situations in their own countries. The French White Plan (Plan Blanc) is the emergency plan for the sudden increase of activity in a hospital. Moreover, as this activity could affect several hospitals, it is called An Extended White Plan. Its aim is to organize the rescue resources to cope with the concentrated number of casualties.

Our work is devoted to studying the part of the French Extended White Plan dedicated to a hospital evacuation. A hospital evacuation occurs when a threat is posed to the hospital itself or when the patients in one hospital must be transported to other hospitals, in the case of an emergency. Our research focus is the hospital evacuation in which the patients have to be moved to other hospitals before a forecasted flood. As far as we know, there are very few investigations of the French Extended White Plan. Our objective is to improve the whole evacuation system from two aspects: optimize the resource dimensioning and minimize the needed evacuation time.

Our paper has made the following scientific contributions:

- The model and the simulation we have done are based on a real scenario. Therefore the model is well suited to the hospital's practices and the improvement methods that we propose can be applied on the ground easily.
- All the cooperation among different departments during the evacuation is taken into consideration, which is often ignored by other pieces of research.
- The quantitative approach we used guarantees the reliability of our result and the feasibility of the whole process.
- Our attention is focused on the hospital evacuation, the research of which is limited, but could inspire other applications.

The remainder of this paper is organized as follows. Section 22.2, briefly, reviews the related literature. Section 22.3 describes the evacuation problem from the context of the vested interest actors and activities. A global model and a detailed model have been formulated separately in Sects. 22.4 and 22.5. In Sect. 22.6, several experiments have been done to analyze the evacuation process and to identify where improvements could be made. The conclusion and perspectives can be found in Sect. 22.7.

22.2 Literature Review

This part is a brief review of the related papers in terms of the research content and the approach. Optimization model and computer simulation seem to be the two most popular research approaches to study evacuation problems. As early as 1982, Sheffi et al. [5] used computer simulation to build a network emergency evacuation model based on a simulator capable of estimating traffic patterns and evacuation time on road networks surrounding nuclear power plants. Filippoupolitis [6] presented a distributed decision support system which consisted of a number of decision nodes helping evacuees to find the best available outlet in a disaster. A multi-agent simulation platform for building evacuation was developed to evaluate the proposed system in various emergency scenarios. Su et al. [7] built a discrete-event computer simulation model for assessing evacuation programs and provided a comprehensive idea of evacuation plans for hospital buildings in the event of a possible bomb threat. Wu et al. [8] proposed a dynamic discrete disaster decision simulation system, which combined the ARENA simulation model with a geographic information system and an SQL Server database to simulate evacuation process and resource deployment. Russo and Vitetta [9] implemented a formulation of the general evacuation problem in the standard simulation context of a 'what if' approach with the consideration of the transportation system in ordinary conditions.

The researchers studying the evacuation model with optimization tools always combine the evacuation problem with the 'facility location problem' or 'relief distribution problem'. All location-evacuation models built for large-scale emergency situations seek to minimize total evacuation time. Kongsomsaksakul et al. [10]

proposed a bi-level program based on the Stackelberg game to find the optimal shelter locations for flood evacuation planning. The upper level problem is a location model which solves the shelter location. The lower level problem is a combined distribution and assignment model that deals with the evacuee route choices. Sherali et al. [11] established a location-allocation model in the event of hurricanes. This model selects the candidate shelters among a given set of admissible alternatives and prescribes an evacuation plan which minimizes the total congestion-related evacuation time. As the evacuees can be regarded as a vehicle or a commodity, it is understandable that some researchers study the relief distribution and the evacuation together using the vehicle routing. This methodology simplifies the model and the solution, but it is not practical in the real world. Different evacuees have different situations and disruptions are likely to happen to patients during an evacuation trip. Adding a probability of disruption parameter may be a solution. Odamar [12] described a hierarchical cluster and route procedure for coordinating vehicle routing in large-scale post-disaster distribution and evacuation activities. He used a multi-level clustering algorithm that groups demand nodes into smaller clusters at each planning level, enabling the optimal solution of cluster routing problems. Wei [13] proposed a mixed integer multi-commodity network flow model that coordinates logistics support and evacuation activities while maintaining equilibrium among service rates of medical facilities. Both wounded people and commodities are categorized into a priority hierarchy, where different types of vehicles are utilized to serve priority transportation needs. The model is based on a network flow formulation. Song et al. [14] formulated a location-routing model with uncertain demands. This model identifies the optimal serving areas and transit vehicle routings to move evacuees from the affected zone to safe destinations.

To sum up, from a view of the content, it can be found that even though the papers studying emergency evacuation are abundant, the papers focusing on hospital evacuation are few. From a view of the approach, the optimization model and the computer simulation are the more suitable in this field. However, none of the research on hospital evacuation was based on a real scenario, which led to two problems: one is to ignore the detail, and the other is to fail to respect the hospital's rules and its laws. These two problems make the existing plans and the improvement methods difficult to put into practice. These gaps have been our main drive: to study the Extended White Plan and to use the computer simulation and the optimization model together based on a real scenario.

22.3 Problem Description

The flood situation we intend to study here is that of a dam, located in Commune Cernon (Jura, France), which, if damaged could rapidly generate a disaster. The dam is about 103 m tall and 36,000 m long. Its water volume is about 600 million m^3. The speed of the flood is estimated at 20 km/h and the height of the water is between 1 and 8 m. The Hospital Saint-Joseph/Saint-Luc is located in the

passage of the potential flood, which, in the event of a breach, would arrive in 5 or 6 h in Lyon and in 8 or 9 h at the Hospital Saint-Joseph/Saint-Luc. Water in the hospital would rise to between 6 and 9 m. The whole situation would last approximately 24 h and its overall impact is difficult to estimate. The flood would affect buildings, infrastructure, electrical and telecommunication networks as well as water networks. Therefore, the evacuation of all the patients in the Hospital Saint-Joseph/Saint-Luc, to other unaffected hospitals is necessary. The patients can be classified into non-autonomous patients, who must be transported with ambulances, and autonomous patients who can be evacuated with public transportation. This latter group will not be considered here. According to the requirement of the Hospital Saint-Joseph/Saint-Luc, the benchmark evacuation time is 5 h (with 6 nurses, 2 coordinators, 12 ambulance teams and 10 stretchers) and the number of non-autonomous patients evacuated is estimated to be 120.

22.3.1 The Vested Interest Actors

The hospital Saint-Joseph/Saint-Luc is located beside the river Rhone. This hospital has 1,207 employees and a capacity of 344 beds. In 2011, the number of people who are treated in the emergency department was 35,767. During the holiday, at night, or on Saturday and Sunday, 19 physicians are on duty. These hospital employees take care of autonomous and non-autonomous patients.

Hospices Civils de Lyon (HCL) is a network of hospitals providing expertise in all disciplines – both medical and surgical. The whole annual budget of HCL is 1.5 billion Euros. It has 23,000 professionals and a capacity of more than 5,400 beds. SAMU is a health care coordinator in France standing for 'Service d'Aide Médicale Urgente'. SAMU controls the response vehicles and ambulances from SMUR (Service Mobile d'Urgence et Reanimation) which is a 'mobile intensive care unit' (MICU). The tasks of SAMU are as follows: to evaluate the patient's needs according to the calls; to find the best care solution for the patient's requirements; to dispatch the most appropriate mobile care resource (MICU, Ambulance...) to move the patient to hospital... In this project, SAMU decides the destination of evacuated patients and the assignment of medical vehicles.

22.3.2 The Activities

The emergency committee is the leader of the evacuation and will control the whole situation. The non-autonomous patient evacuation can be divided into three main processes: preparing the patients, using stretchers to move the patients in the Hospital Saint-Joseph/Saint-Luc, and assigning the ambulances to transport the patients from Hospital Saint-Joseph/Saint Luc to Hospices Civils de Lyon. Every process consists of several activities. The patient preparation includes the

preparation of: the list of the patients to be evacuated, the list of places to which the patients will be moved, the medication that the patients will need during the evacuation and the clinical history of each patient. Ambulance utilization is similar to stretcher utilization respectively both outside and inside the hospital. These two equipment utilization processes involve the demand of the ambulance/stretcher, the assignment of the ambulance/stretcher and the transportation of the patient needing the ambulance/stretcher.

22.4 Global Model

In this part, an optimization model is proposed firstly to model the evacuation globally, mainly for resource dimensioning reasons and also in order to measure the patient evacuation time.

22.4.1 Optimization Model

As linear programming technique is an exact method, we adopted this optimization technique to find our resource dimensioning. A linear programme has been built according to the three aforementioned processes.

The data of this model is:

- T: the number of periods t (in hours)
- M: the number of activities
- Nbpat: the number of non-autonomous patients to evacuate
- Dur (j): the duration of the activity j (by minutes)
- Suc (j): the set of successors of the activity j
- Cap (j,t): the capacity of the resource associated to the activity j (by minutes)

The variables are:

- X (j,t): the number of patients who benefit from the activity j during period t
- S (j,t): the number of patients who wait for the next activity during period t

Model:

$$Min \sum_{t=1}^{T} X(M,t).t \tag{22.1}$$

Subject to

$$S(j,t-1)+X(j,t)-S(j,t) \geq X(k,t) \quad \forall k \in Suc(j) \quad \forall j = 1,\ldots,M-1 \quad \forall t = 1,\ldots,T \tag{22.2}$$

$$X(j,t).Dur(j) \leq Cap(j,t) \quad \forall j = 1,\ldots,M \quad \forall t = 1,\ldots,T \tag{22.3}$$

$$\sum_{t=1}^{T} X(M,t) = Nbpat \tag{22.4}$$

The objective function endeavors to minimize the sum of two series where each term of same rank in both series is multiplied. To be minimized, the series X(M,t) has to be ranked in decreasing order as the series t is in increasing order. So activities are realized at the earliest. We try to use this dynamic model to find the suitable resource dimensioning. The constraint (22.2) controls the patient flow. Constraint (22.3) respects the resource capabilities. Constraint (22.4) defines the number of patients to be evacuated. Using the software Cplex, we got the needed resources are: 6 nurses, 12 ambulance teams, 10 stretchers and 2 coordinators. This result indicates that, according to the policies of the hospital, it will cost at least 5 h to evacuate 120 non-autonomous patients.

22.4.2 Problem Complexity

Even though the three main processes can be modeled via our linear programming, several problem complexities lie in our study and make our linear programming not perfect enough to properly reflect the real world. First, since the evacuation situations are plagued with uncertainties [15], it is better to use stochastic data, which can reflect the situation better. With the stochastic data, however, the solution of the optimization model is difficult to find. Second, the optimization model is an analytical method which can only represent the as-is system in an aggregated way and lot of details are omitted. Third, the optimization model cannot achieve the organization improvement objective effectively in practice. So, it is necessary to find another suitable approach. The results calculated by the linear program, however, can be used to check the result obtained by the next approach i.e. to define a lower bound.

22.5 Detailed Model

Judging from our literature review, the most suitable way to solve problems related to real-world complex systems is a computer simulation approach [16]. First, a computer simulation can capture the stochastic data by a user-friendly interface [17]. Second, unanticipated problems and the sequential or parallel events can be easily exposed by simulation. Third, the combination of 3-D technology makes simulation powerful at presenting detailed information, hence helping people to find the weak points [18]. Fourth, simulation can act as a 'what if' tool and will be useful to support training exercises, performance and impact evaluation [19].

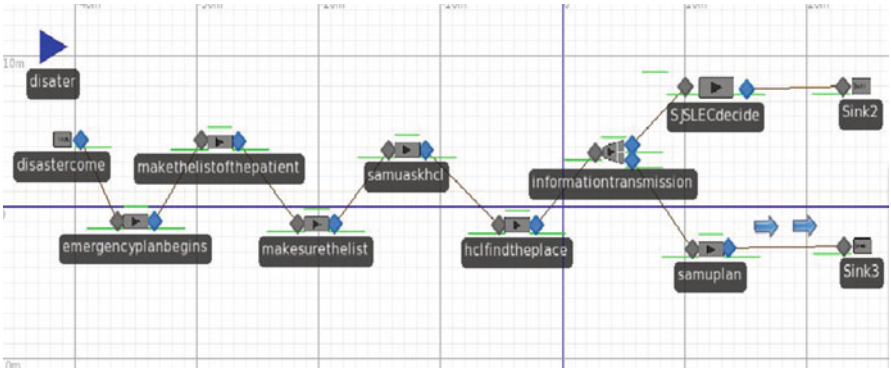


Fig. 22.1 The process of sending the information in SIMIO

According to our objective and the way proposed by Bradley et al. [20] to choose a Business process re-engineering (BPR) software, we decided to adopt two software tools in our detail model. At the beginning, we used ARIS to build a frame model which can detail the evacuation activities because ARIS is a powerful communication tool. However, the simulation ability is a weakness in ARIS. SIMIO is impressing in the previous model, with its simulation ability. So it is adopted to implement the later simulation work. ARIS, as one of successful products of IDS Scheer, is widely used for the purpose of business process design and management. It is powerful in process modeling and provides reliable information flowcharts. ARIS was chosen to build a frame model which clearly shows the co-operation and information sharing among the different hospital departments. Our frame model [21] is made up of 17 processes and 108 activities. Once the frame model is finished, we intend to use SIMIO to implement the simulation. SIMIO is designed to support the object modeling paradigm and supports both discrete and continuous systems, along with large scale applications based on agent-based modeling. These advantages help us to build a simulation model based on the frame model and display logically and graphically the characteristic of discrete events in our model. The strong 3-D simulation effect is useful tool for presenting the result to the practitioners in the hospital.

Our simulation model consists of two levels. Level One is to show the information communication among the vested interest actors. Level Two is to simulate the evacuation activities. Different trigger events are used here to synchronize these two levels. Figure 22.1 presents Level One. At the beginning of Level One, six different activities will happen in chronological sequence. After these six activities, two activities take place in parallel. A model entity is used to represent the disaster. Among the six chronological sequence activities, five servers are used to represent the first five activities separately. The first one is triggering Emergency Management Plan and assembling the Saint-Joseph/Saint-Luc Emergency Committee (SJSLEC). The second one is the process of making the list of the patients to be evacuated. The third one is assessing the needs of those patients and passing it to a SAMU. The 4th

one is that the SAMU requires the HCL to find available beds for the patients to be evacuated. The 5th one is the process of making a list of the available places in HCL. The 6th activity shows the information process from HCL to SAMU and Saint-Joseph/Saint-Luc. After the information process has been completed, the SAMU will direct the ambulances and the SJSLEC will plan the utilization of the available beds in HCL.

22.6 System Improvement

After the correctness of the detail model was checked, several experiments were launched and from these two proposals for improvements were made.

22.6.1 Experiments

In order to evaluate the performance of the as-is system and find a way to improve the existing evacuation process, a good design of the experiment is necessary. The common and normal steps for the experiments are the following [22, 23]: define the goal of the project, identify and classify the experiment variables and choose an experiment design. Because our goal is clear enough, we begin with the second step. A cause-effect diagram can be applied to look for the two kinds of experiment variables, the dependent variable and the independent variable. The dependent variables are the total evacuation time and the time used to evacuate the first patient. The independent experiment variables are the number of nurses, coordinators, ambulance teams and stretchers. For the third step, a factorial design experiment has been retained. This experiment allows studying the effect of each independent variable on the dependent variables, as well as the effects of interactions between the independent variables and the dependent variables.

Four experiments have been conducted. Every combination of the experiments ran 20 times and the confidence level has been set to 95%. After calculation, it is found that the effect of every independent variable is similar. However, the effects of interactions between the independent variables and the dependent variables are significantly different. According to the benchmark evacuation time, the minimal and reasonable resources that we must assign are 6 nurses, 2 coordinators 12 ambulance teams and 10 stretchers. These correspond to the linear model result.

22.6.2 The Improvement Scenarios

After the experiments, two improvement methods have been proposed. As the material resource and the human resource are limited, we have modified the way

to use the resource, and thus a first modification was proposed. It was found that the time to evacuate the first patient appeared too long, and, therefore, we have sought to solve the problem and proposed the second change.

To further enhance the organization, a change is needed to the practice of stretcher use. In the basic model, the worker who pushes the stretcher will transport the patient to the car park and wait for the ambulance until it comes and then he will go back to the care unit with the stretcher. In the improved model, the worker will move the stretcher from the care unit to the car park, and then go back to the care unit to continue the transportation of patients, using a free stretcher from the car park. As a consequence, the number of stretchers must be increased. The patient will wait in the car park and be cared for by an employee until he can be transferred to the ambulance. When the patient is on the ambulance, the stretcher is released and is left in the car park, for further use.

From the result of the simulation we can find that the time to evacuate the first patient is too long. If we assign 6 nurses, 2 coordinators, 12 ambulances and 10 stretchers, the time used to evacuate the first patient will be more than 1.4 h. Our experiment found that not a lot of time can be saved even if more resources are available. After the analysis, we found that the main reason is that the information preparation tasks consume a lot of time. To be exact, it takes too much time to make two lists, the patient list and the available place list. In the as-is model, humans make the patient list and the placement list. In fact, an appropriate information system can do this task better and faster than a human. So the second way to improve the system is to add an information system. The information system can create the patient list and the placement list automatically and quickly. The Regional Health Agency in France is beginning to build a web site where each private or public hospital must specify the number of their available beds twice a day. In the as-is system, it would take about 10 min to assemble the patient list and 15 min to build the placement list. With the web information system, it takes just 1 min to create the patient list and just 5 min to produce the placement list. To evacuate 120 patients, the original system takes 4.94807 h, our first modification can reduce the time to 4.68333 h and the second modification further reduces the time to 3.98220 h. When these two improvements are combined, an overall improvement of 1.29622 h will have been achieved, with it taking only 3.65185 h to evacuate 120 patients.

22.7 The Conclusion and the Future

This is one of the first papers studying the French Extended White Plan and a prudent and comprehensive detail model has been made based on a real scenario in the context where the hospital suffers from a threatening flood and all the patients must be evacuated. In order to evaluate the needed resources and to ensure the correctness of the evacuation model, we first used a global linear model and second a detailed simulation model. The best way to assign the resources and improvements based on new IT information has been proposed. The results of the experiments

confirm the idea behind our suggested modifications: that the evacuation time could be largely reduced. One of the limitations of our research is the reliability and availability of the data we obtained. All the data we obtained is solely based on the experience of the people in the hospital. The possibility of an interruption was not taken into account during the evacuation such as transportation congestion, for example. In that case, helicopters should be used. In such a case, fuzzy theory will probably be adopted to get more valuable data and build an extended simulation model. This work was sponsored by the Rhone-Alps Region.

References

1. Miguel Llanos: 2011 already costliest year for natural disasters. http://www.msnbc.msn.com/id/43727793/ns/world_news-world_environment/t/already-costliest-year-natural-disasters/ (2011). Accessed on 21 Mar 2013
2. Guha-Sapir D, Vos F, Below R, Ponserre S.: Annual Disaster Statistical Review 2010: The Numbers and Trends. CRED, Brussels (2011)
3. Altay, N., Green III, W.G.: OR/MS research in disaster operations management. Eur. J. Oper. Res. **175**(1), 475–493 (2006)
4. Chen, A., Peña-Mora, F., Ouyang, Y.: A collaborative GIS framework to support equipment distribution for civil engineering disaster response operations. Automat. Constr. **20**, 637–648 (2011)
5. Sheffi, Y., Mahmassani, H., Powell, W.B.: A transportation network evacuation model. Transp. Res. Part A **163**, 209–218 (1982)
6. Filippoupolitis, A., Gelenbe, E.: A decision support system for disaster management in buildings. In: Proceedings of the 2009 Summer Computer Simulation Conference, San Diego, pp. 141–147 (2009)
7. Su, S., Tsai, S.T., Shih, C.L., Kuo, R.J., Wang, H.C., Chen, J.C.: Use of computer simulation in the evacuation system for hospitals. In: Proceedings of the Summer Computer Simulation Conference, Edinburgh, pp. 205–212 (2008)
8. Wu, S., Shuman, L., Bidanda, B., Kelly, M., Sochats, K., Balahan, C.: Embedding GIS in disaster simulation. In: Environmental Systems Research Institute International User Conference Proceedings, San Diego (2007)
9. Russo, F., Vitetta, A.: A methodology for evacuation design for urban areas: theoretical aspects and experimentation. In: Proceedings of the Summer Computer Simulation Conference, Edinburgh, pp. 1–14 (2008)
10. Kongsomsaksakul, S., Yang, C., Chen, A.: Shelter location-allocation model for flood evacuation planning. J. East. Asia Soc. Transp. Stud. **6**, 4237–4252 (2005)
11. Sherali, H., Carter, T., Hobeika, A.: A location-allocation model and algorithm for evacuation planning under hurricane/flood conditions. Transp. Res. Part B Methodol. **25**(6), 439–452 (1991)
12. Özdamar, L., Demi, O.: A hierarchical clustering and routing procedure for large scale disaster relief logistics planning. Transp. Res. Part E Logist. Transp. Rev. **48**(3), 591–602 (2012)
13. Yi, W., Özdamar, L.: A dynamic logistics coordination model for evacuation and support in disaster response activities. Eur. J. Oper. Res. **179**, 1177–1193 (2007)
14. Song, R., He, S., Zhang, L.: Optimum transit operations during the emergency evacuations. J. Transp. Syst. Eng. Info. Technol. **9**(6), 154–160 (2009)
15. Aakil, M., Nie, X., Shaligram, P.: Optimization models in emergency logistics: a literature review. Socioecon. Plann. Sci. **46**, 4–13 (2012)

16. Longo, F.: Emergency simulation: state of the art and future research guidelines. SCS M&S Mag. **4**, 1–8 (2010)
17. Benneyan, J.C.: An introduction to using computer simulation in healthcare: patient wait study. J. Soc. Health Syst. **5**(3), 1–15 (1997)
18. Pidgeon, N.: O'Leary M: Man-made disasters: why technology and organizations (sometimes) fail. Saf. Sci. **34**, 15–30 (2000)
19. Stewart, R.: Simulation – The Practice of Model Development and Use. Wiley, Chichester (2004)
20. Bradley, P., Browne, J., Jackson, S., Jagdev, H.: Business process re-engineering (BPR) – a study of the software tools currently available. Comput. Ind. **25**, 309–330 (1995)
21. Wang, T., Guinet, A., Belaidi, A., Besombes, B.: Modelling and simulation of emergency services with ARIS and Arena. Case study: the emergency department of Saint Joseph and Saint LucHospital. Prod. Plann. Control **20**(6), 484–495 (2009)
22. Barton, R.R.: Designing simulation experiments. In: Proceedings of the 2002 Winter Simulation Conference, New Jersey (2002)
23. Barton, R.R.: Experimental design for simulation. In: Proceedings of the 2003 Winter Simulation Conference, New Orleans (2003)

Chapter 23
Using Simulation to Analyze Patient Flows in a Hospital Emergency Department in Hong Kong

Omar Rado, Benedetta Lupia, Janny M.Y. Leung, Yong-Hong Kuo, and Colin A. Graham

Abstract This paper presents a case-study of applying simulation to analyze patient flows in a hospital emergency department in Hong Kong. The purpose of our work is to analyze the impact of the enhancements made to the system after the relocation of the emergency department. We developed a simulation model (using ARENA) to capture all the key relevant processes of the department. Using the simulation model, we evaluated the impact of possible changes to the system by running different scenarios. This provides a tool for the operations manager in the emergency department to "foresee" the impact on the daily operations when making possible changes (such as, adjusting staffing levels or shift times), and consequently make much better decisions.

23.1 Introduction

The Prince of Wales Hospital (PWH) is one of the largest public general hospitals in Hong Kong and the teaching hospital for the Medical Faculty of the Chinese University of Hong Kong. It provides 1,360 hospital beds, employs around 4,000

O. Rado • B. Lupia
Università degli Studi di Padova, Padova, Italy
e-mail: omar.rado@gmail.com; benedetta.lupia@libero.it

J.M.Y. Leung
Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong
e-mail: janny@se.cuhk.edu.hk

Y.-H. Kuo (✉) • C.A. Graham
Faculty of Management and Administration, Macau University
of Science and Technology, Avenida Wai Long, Taipa, Macau

Accident and Emergency Medicine Academic Unit, The Chinese University
of Hong Kong, Shatin, New Territories, Hong Kong
e-mail: yhkuo@must.edu.mo; cagraham@cuhk.edu.hk

A. Matta et al. (eds.), *Proceedings of the International Conference on Health Care Systems Engineering*, Springer Proceedings in Mathematics & Statistics 61,
DOI 10.1007/978-3-319-01848-5_23,

people and operates as the regional hospital of the New Territories East (serving more than 1.5 million people). In order to provide a good quality of service, PWH has to best-utilize its resources because of the large number of patients served and its limited budget due to tight government financial support. One of the departments facing this challenge head-on is the Emergency Department (ED) which provides 24-h Accident and Emergency (A&E) services. In preparation for the growing (and aging) population in Hong Kong, the ED was relocated in October 2010 to accommodate the increasing patients' demands. Since then, the hospital management has been trying to enhance the daily operations in the new department.

The ED handles 420 cases a day on average. In the daytime, the department is internally divided into two independent divisions: the *Walking* division and the *Non-walking* division respectively treating mobile patients (who can walk) and patients on a trolley or a wheel chair (thus non-walking). After 23:00, the Walking division is closed and the walking patients are diverted to the Non-walking division until 07:00 (i.e., walking patients and non-walking patients are merged to have the same treatment procedure.).

Critical patients arriving by ambulance are rushed to the resuscitation rooms and treated immediately. Otherwise, after registration, patients are assessed by a triage nurse and classified by category (level of urgency) so as to assign priorities for receiving treatments. There are five categories of patients: 1(critical), 2(emergency), 3(urgent), 4(standard) and 5(non-urgent). In our work and the rest of this paper, we put category 5 patients into category 4 as they have the same flow and priority in real practice and there are only small proportion of category 5 patients. *Critically-ill patients* (categories 1 and 2 patients), *less urgent walking patients* (categories 3 and 4 walking patients) and *less urgent non-walking patients* (categories 3 and 4 non-walking patients) follow different procedures of receiving treatments. Critically-ill patients have the highest priority and category 3 patients have a higher priority over category 4 patients. Within the same category, patients are seen on a first-come, first-served (FCFS) basis.

To provide 24-h A&E services, the ED employs different shifts (8 h a shift including a meal break of an hour and a short break of 20 min) of doctors to cover the patients' demands over a whole day. Basically, there are three shifts: morning (08:00–16:00), evening (16:00 to midnight) and mid-night shifts (00:00–08:00). In addition, an off-duty doctor is on-call.

As the ED has to handle a large number of patients a day, it must operate at a very high level of efficiency and quality. Ineffective operations can lead to serious consequences such as delay of treatments or even death of critical patients. To guarantee good quality of services, the ED aims to achieve the following service targets, as recommended by the Hospital Authority of Hong Kong.

1. Critical and emergency patients have to be given immediate care after they are admitted to the ED.
2. 90% of urgent patients (category 3 patients) have to be seen by a doctor within 30 min after registration.
3. Most patients are expected to be seen within 2.5 h after registration.

As mentioned, there are a large number of patient visits but the manpower is insufficient, the ED has the very difficult task of trying to offer a good quality service (minimizing patients' waiting times whilst not compromising the required attention for each patient), and making sure that valuable resources (e.g., doctors' and nurses' time and treatment equipment) are well-utilized. Our project team was asked by the ED to analyze and improve patient flows so as to enhance the quality of services provided. We adopt a simulation approach to provide the operations manager in the ED with a set of measurements (e.g., patients' waiting times and doctors' utilization) to assess the department performance and evaluate the impacts on the daily operations with different policies. Our previous investigations were reported in [15].

This paper is organized as follows. In Sect. 23.2, we give a literature review on related work. In Sect. 23.3, we compare the original and the current layouts of the ED. In Sects. 23.4 and 23.5, we describe our simulation model and present the results of the tests with different scenarios. Section 23.6 summarizes our work.

23.2 Literature Review

The applications of operations management techniques in the health-care industry are vast. We refer the reader to [18] for a recent survey. Here we focus on the applications to operations enhancement, in particular patient flows in emergency rooms.

In recent years, researchers have successfully built queueing models for analyzing and improving patient flows, and proposing decision strategies and policies in EDs. Green et al. [9] used a Lag stationary independent period by period queueing analysis to allocate staff to reduce the number of patients who leave without being seen. Cochran and Roche [5] presented a spreadsheet implementation of a queuing network model with split patient flows (accounting for patient categories of different acuity and arrival patterns and volume), to help reduce patient "walk-aways" and improve service provision of the ED. Huang et al. [12] considered the control of patient flow, in which physicians have to choose between seeing patients right after triage (facing deadline constraints on their time-till-first-service) and those who are in process but possibly need to return to physicians several times during their ED sojourn (resulting in feedbacks to the queueing system). They also proposed and analyzed scheduling policies with two types of costs: queueing costs incurred per individual doctor visit and congestion costs accumulate over all visits during patient sojourn-times. Saghafian et al. [20, 21] proposed patient streaming (based on their likelihood of being admitted to the hospital) and complexity-based triage (an up-front estimate of patient complexity) for improving operations in EDs. In both papers, they used a combination of analytic and simulation models to show the effectivenesses of the policies. While there has been much work on deriving analytical models for helping operations enhancements in EDs, we adopt a simulation approach for improving patient flows in the ED of PWH as

it is easier to examine many "what-if" scenarios with the complex system of the ED (such as time and category-dependent arrival rates of patients, different service-time distributions and time-varying staffing levels). And more importantly, a simulation approach is much convenient for real implementations as practitioners, who are not necessarily equipped with advanced mathematical and programming knowledge, can easily understand and make changes in the system within a user-friendly graphical interface to "foresee" the outcomes, which are basic statistical performance measures such as maximum and average waiting time of patients and utilization of staff.

The applications of simulation in the area of health-care management have been studied for more than half of a century, e.g. [7]; and the academic literature on simulation in health-care is immense. We refer the reader to [10, 14] for an overview. In EDs, reported successful cases of applying simulation models were mainly to improve the efficiencies of daily operations. A major proportion of work with the use of simulation is staff scheduling. The approaches are mainly to evaluate process performance with different staff shift schedules, e.g. [8, 19]. Some papers integrated optimization techniques with simulation. Ahmed and Alkhamis [1] presented a simulation optimization approach to determine the optimal number of doctors, lab technicians and nurses required to maximize patient throughput and to reduce patient time in the system subject to a set of constraints imposed on budgets, patient waiting time and number of servers. Centeno et al. [4] integrated simulation (for establishing the staffing requirements for each period) and integer linear programming to help ED managers optimize staff schedules so as to maximize utilization within given budgets. Yeh and Lin [22] utilized simulation and a genetic algorithm to obtain a near-optimal nurse schedule based on minimizing the patients' queue time. There has also been work on examining queueing priorities by running simulation experiments. Connelly and Bair [6] developed a simulation model for system-level investigation of ED operations and to compare a fast-track triage approach with an acuity ratio triage approach. Other related applications include policy/decision making. Hoot et al. [11] used simulation of patient flow to predict near-future ED operational measures and to forecast with several measures of ED crowding. Lane et al. [16] used simulation to analyze the functioning of the ED system and the policies with different bed capacity and demand pattern scenarios. Baseler et al. [2] developed a simulation model to estimate the function of patients' time in system and the maximum level of patients' demand that the system can absorb.

Although, according to [3], the number of articles related to health-care simulation or modelling is currently expanding at the rate of about 30 articles a day, Jahangirian et al. [13] found out that only 8% of the related papers actually applied simulation to a real problem where real data was used. This proportion is substantially smaller than the corresponding percentages in the areas of commerce (49.1%) and defense (39.4%) [13]. This highlights the fact that real implementations of simulation models in practice in the health-care sector are still rare and we still need to put more effort on promoting the use of simulation for advancing health-care management. In this paper, we present a real case of analyzing and improving patient flows in an ED in Hong Kong with the use of simulation.

23.3 Comparisons Between the Original and Current Settings

In October 2010, the ED of PWH was relocated to a new building with a new layout. Several changes were also made in the new system to accommodate the growing patients' demands. In this section, we analyze two major changes in the operations and compare the efficiencies of the original and current systems. To make fair comparisons, we present the data of the month of December 2009 (when operating in the old location) and December 2010 (after relocation). There were 12,945 patient visits in December 2009, and 13,287 visits in 2010, which translate to around 418 and 429 cases per day respectively. (The reason why we did not choose the first month after the relocation to make comparisons is that a "warm-up" period was needed since initially most of the staff needed time to get used to the new layout, system and settings.) Below, we describe two key changes in layout and operations and their impacts.

23.3.1 A Closer Sub-waiting Area for Consultation in the Walking Division

After the relocation of the ED, the waiting area for doctor's consultation in the Walking division was moved from the main waiting area to a new sub-waiting area, which is closer to the consultation rooms than before. This aims to shorten the walking time of patients. More importantly, this enables the nurses to more easily notify the patients that they will soon be seen by a doctor, so that they would not leave the waiting area (e.g. for a meal) while waiting. Consequently, this reduces the inactivity times of doctors waiting for "missing" patients, and hence reduces the waiting times of subsequent patients seen.

We compare the net times from triage to consultation for category 4 patients, who are mostly walking patients, before and after the relocation.

Comparing the data of 2009 and 2010, although there was an increase of 2.64% in the total number of patient visits, the average net time from triage to consultation for category 4 patients decreased from 112.91 to 107.77 min (a 4.55% decrease). This shows that walking patients benefit from the change of the layout of the waiting area in the Walking division. From Fig. 23.1, we observe the patterns of the net times from triage to consultation in the 2 months are similar but the one in December 2009 has a heavier tail. The percentage of category 4 patients who had net time from triage to consultation more than 3 h decreased from 21.57 to 16.01% (a 25.78% decrease). This indicates the increase in walking time of patients could amplify the waiting times of patients.

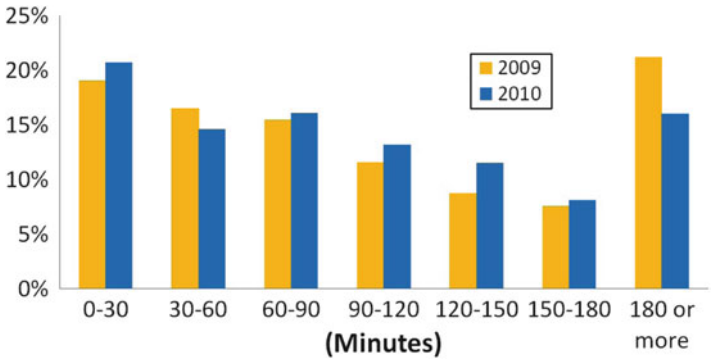


Fig. 23.1 Net time from triage to consultation for category 4 patients

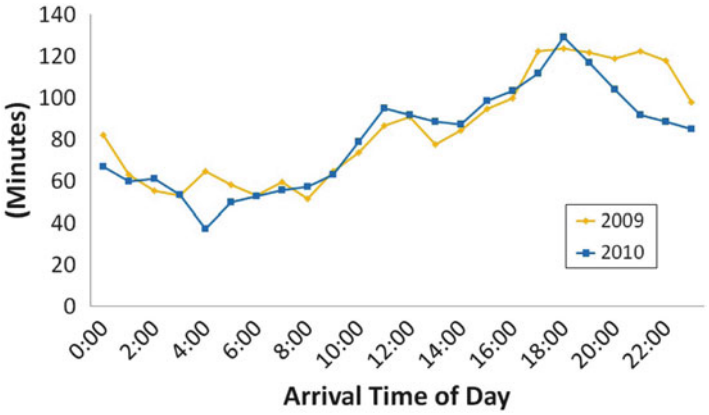


Fig. 23.2 Average net time from triage to consultation for less urgent patients by arrival time of day

23.3.2 Consolidation of the Walking and Non-walking Divisions in Nighttime

Before the relocation, the Walking and Non-walking divisions operated independently, each with its own staff and resources. After the relocation, the ED started to implement the policy that during nighttime (from 23:00 to 07:00) the Walking division is closed and the walking patients would join the system of the Non-walking division. It aims to better-utilize the reduced workforce (about half of the workforce of daytime) due to the low arrival rates of patients in nighttime.

Figure 23.2 depicts the average net time from triage to consultation for less urgent patients by arrival time of day in 2009 and 2010. From 07:00 to 20:00, the net times were similar in the 2 months. From 20:00 to 07:00, a significant improvement was observed. An interesting finding is that patients arriving after 20:00 but before 23:00

also benefited from the consolidation of the divisions. We believe it is due to the fact that some of these patients might need to wait for consultation for more than 3 h so that they might start consultation after 23:00 and hence benefited from the change.

23.4 Simulation Model

We developed a more detailed model of the new ED to analyze patient flows. As reported by other researchers, it is very difficult to build analytical models for the ED as there are many complicating factors in reality (such as time and category-dependent arrival rates of patients, multiple shift-times of doctors and re-entrant flows to the many "service stations"of the system). For this reason, we adopt a simulation approach which facilitates examination of many "what-if" scenarios, and provide valuable indications as to where the major bottlenecks of the system might be. It also provides a way to explore possible changes without jeopardizing patient care. Figure 23.3 depicts the screen-view of our simulation model, built using the software ARENA.

Our simulation model captures: all relevant treatment processes (triage, consultation, lab tests, etc.), the complexities of intertwining and re-entrant patient-flows, complicated arrival rates that vary by time and patient category and adjustable staff deployment (shift, breaks, doctors on reserve). The necessary input parameters/data are arrival rates, probability distributions of service times, available resources and schedules of doctors and nurses. To model the non-stationary time-varying arrival

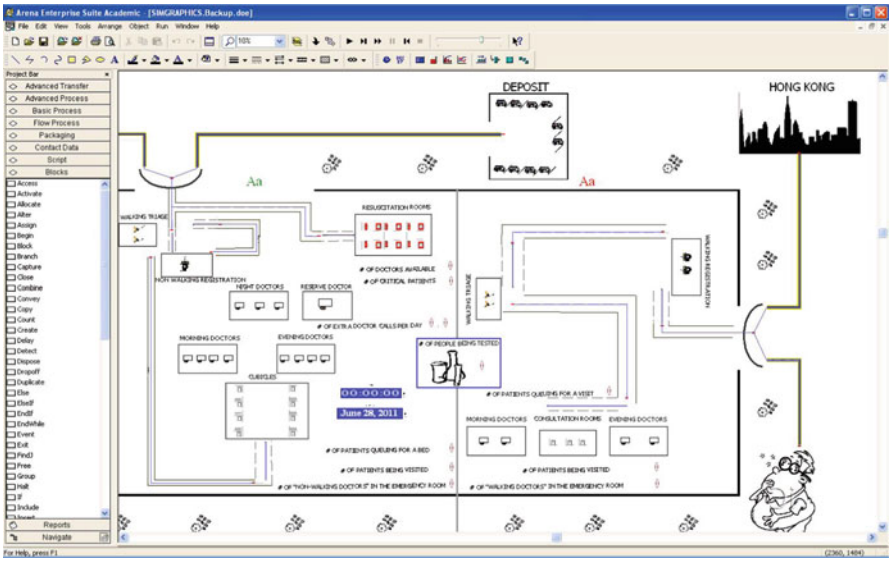


Fig. 23.3 The screen-view of our simulation model in ARENA

Table 23.1 Actual and simulated net times between services for less urgent patients

	Actual (min)	Simulated (min)	% error
Registration to triage	7.06	7.02	−0.57
Triage to consultation	85.27	81.76	−4.12

rates of patients, for each patient, his/her arrival time is the arrival time of the previous patient from the same category plus an interarrival time, which follows an exponential distribution with arrival rate in the time period of the previous arrival. We also tackled a challenge that the service-time distributions were not directly obtainable from the historical data. To resolve the problem, we assume that each service activity follows a Weibull distribution and develop two meta-heuristics, Descent Method and Simulated Annealing, to search for a good estimate of the parameters of the distributions, by considering the available time points in the data provided. For detailed descriptions of the challenge and parameter estimation procedure, we refer the reader to [15]. The outputs of the model are the key performance measures such as patients' waiting times, queue lengths, utilizations of doctors, which help us study and understand the performance of the ED.

To validate our simulation model, we presented the model and the simulated key performance indicators such as queue lengths and waiting times of patients to a consultant in the ED. He believed the model was sufficient to capture all the key activities in the ED and those values agreed with his estimations. The simulation model was also validated by comparing several key statistics as generated by the model to actual observations. As an illustration, Table 23.1 shows the comparison between the actual and simulated net times between services for less urgent patients. More details of the validation of the simulation model are discussed in [15].

23.5 Simulation Results

By running simulations, we have a way to obtain performance measures for the ED under different scenarios and, thus, to evaluate possible policies and changes in the system. We used the current arrival rates and the actual staff schedule as the input parameters for our base case. We did a series of simulation runs to evaluate different possible scenarios. In this section, we present some key findings.

23.5.1 10% Growth in Patient Arrivals

The population in Hong Kong keeps increasing (with an annual growth of around 1% according to the statistics of 2012 [17]), mainly due to an influx of immigrants. Moreover, more and more non-immigrant visitors from Mainland China also come

Table 23.2 Average waiting times of the patients based on the current arrivals and the simulated scenario (10% growth in patient arrivals)

	Current situation (min)	10% growth (min)	% change
Triage(walking)	10.20	12.66	+24.11
Triage(non-walking)	2.64	3.26	+23.48
Consultation(category 3 walking)	31.11	31.55	+1.41
Consultation(category 4 walking)	236.91	274.81	+15.99
Consultation(category 3 non-walking)	10.71	15.39	+43.69
Consultation(category 4 non-walking)	92.34	149.12	+61.49

to Hong Kong for a better quality of medical treatments. Thus, the demand for medical services in Hong Kong is expected to have a significant growth in the coming future. This is of particular concern for the EDs, which are often viewed as inexpensive clinics by the non-critical patients who visit them.

To study how the growth of patient visits impact on the daily operations in the ED, we increased the arrival rates of all patient categories by 10% (which is equivalent to the percentage increase in 3–4 years using the estimated annual increase of 2.64%) and keep all the capacities and resources at the current levels. We ran simulations and recorded the waiting times of patients. (In this section, waiting time for consultation is defined as the total waiting time for the first consultation and the "follow-up" consultation after extra tests, if needed, for the same patient visit.)

From Table 23.2, we observe that a 10% growth of patient arrivals leads to a big increase in the waiting times of patients for triage and for consultation. The waiting times of patients increase mostly more than 20%. As expected, a larger increase in waiting times is observed for categories 4 patients since a lower priority is given to them. The 10% growth in patient arrivals leads to an increase in doctors' utilization, from 88.44 to 94.3%. Some doctors are overloaded with utilization more than 100% (i.e. working time is extended). Moreover, it is important to point out that, based on the above results, the targets of services set by the ED probably cannot be met after 10% growth of patients arrivals.

23.5.2 Adding an Extra Doctor

In order to reduce waiting times of patients, we evaluate the performance of the ED if an extra doctor is hired, based on the current situation. This activity is useful to determine the optimal trade-off between the cost of additional workforce and the services provided.

Before adding an extra doctor to the simulation model, we calculated the utilization of every doctor in order to assess which doctors are overloaded. We observe a significant overuse of the doctors working the afternoon shift in the Walking division and those for the mid-night shift. Therefore, we simulated the two scenarios when an extra doctor is added to each of the shift.

Table 23.3 Average waiting times of the patients for consultation based on the current situation and the simulated scenario (an extra doctor added to the afternoon shift in the Walking division)

	Current situation (min)	Doctor added (min)	% change
Consultation(category 3 walking)	31.11	28.02	−9.93
Consultation(category 4 walking)	236.91	188.50	−20.43

Table 23.4 Average waiting times of the patients for consultation based on the current situation and the simulated scenario (an extra doctor added to the mid-night shift)

	Current situation	Doctor added	% change
Consultation(category 3 walking)	31.11	29.79	−4.24
Consultation(category 4 walking)	236.91	236.06	−0.36
Consultation(category 3 non-walking)	10.71	8.47	−20.92
Consultation(category 4 non-walking)	92.34	49.80	−46.07

Table 23.3 lists the average waiting times of walking patients if we add an extra doctor to the afternoon shift in the Walking division. Not surprisingly, on average, the relative time reduction for category 3 patients waiting for consultation is smaller than category 4 patients' as category 3 patients have a higher priority. A significant reduction in average waiting time of categories 4 patients for consultation is observed.

Alternatively, if we add an extra doctor to the mid-night shift, we observe a significant reduction in the average waiting times for the patients in the Non-walking division (see Table 23.4). Surprisingly, although the walking patients are directed to the Non-walking division for consultation in nighttime, we cannot make any significant reduction in the waiting times for walking patients after adding an extra doctor to the mid-night shift. We believe the surprising result is due to the fact that the consultation is still not fast enough to clear the patients of the lowest priority, who are category 4 walking patients. Another possible reason is that patients usually experience longer waiting times during afternoon but not nighttime, which is shown in Fig. 23.2. It seems that adding an extra doctor to the mid-night shift cannot benefit the walking patients. Finally, we note that having an extra doctor can contribute to a decrease in doctors' utilization from 88.44 to 81.64%.

23.5.3 Shift Planning

Although adding more resources to the ED is the best way to improve the patient flows, the financial issue is one of the major concerns of the hospital management. Given limited budgets, one way to improve the patient flows is to best-utilize the current resources. Therefore, we would like evaluate how the schedules of the doctors, who are the most valuable resources in the ED, might be changed to improve the efficiency of the ED. By measuring the utilizations of doctors in

Table 23.5 Average waiting times of the patients for consultation based on the current situation and the simulated scenario (reallocation of doctor)

	Current situation (min)	Reallocation (min)	% change
Consultation (category 3 walking)	31.11	27.85	−10.48
Consultation (category 4 walking)	236.91	185.37	−21.76
Consultation (category 3 non-walking)	10.71	20.19	+88.52
Consultation (category 4 non-walking)	92.34	102.03	+10.49

the current scenario, we can find out the doctors with the heaviest and lightest workloads. They are the doctors in the Walking division and Non-walking division, respectively, in the afternoon. An interesting scenario would be to assign the doctor who has the lightest workload to the shift of heaviest workload. (i.e. In the afternoon shift, extract a doctor in the Non-walking division and assign him/her to the Walking division). The results are shown in Table 23.5.

As expected, the walking patients benefited from this reallocation. A significant time reduction in the average waiting times for consultation is observed for walking patients. However, the average waiting times for consultation for non-walking patients increase as a doctor is removed from the Non-walking division. The reallocation of doctor, of course, benefits the majority, but at the same time, hurts the more urgent minority. To decide whether we should employ this schedule, we have to determine the optimal trade-off. Simulation is a tool for decision makers to "predict" how good or how bad a change impacts on the system in order to make the right balance. We would like to point out that, although there is a large percentage increase in the waiting time for consultation for category 3 non-walking patients after this reallocation, the absolute increase (9.48 min) is still small enough to be within range of the target waiting time set by the Hospital Authority for patients of this category. Moreover, this increase is comparably much smaller than the absolute reduction for the category 4 walking patients (51.45 min). As the majority of patients are category 4 walking patients, a reduction in total average waiting time is expected after the reallocation. Nonetheless, this balance between benefits to the majority and urgency of service to those in need is a difficult decision for the hospital management.

Although we just presented some of the issues examined, the simulation model could be used by the operations manager in the ED to evaluate many other possible changes in the system, such as layout, capacities and resources.

23.6 Conclusions

This paper presents a case-study of analyzing patient flows in a hospital emergency department in Hong Kong. We analyzed the enhancements of the system changes after the relocation of the ED in October 2010. We also developed a simulation tool

for the ED to evaluate the impacts on patient flows with different scenarios. The simulation tool can also throw some light on key issues of decision making for the operations manager.

Finally, it is important to remark that, in general, it is very difficult (or nearly impossible) to build a simulation model for an ED to capture all the activities and events in the system, particularly when key parameters cannot be estimated directly. However, the inclusion of the major activities and events, as captured by our simulation model, was already sufficient to let operations managers in EDs "foresee" the impacts on the daily operations due to possible changes, and consequently enable them to make much better decisions.

Acknowledgements The authors would like to thank Mr. Stones Wong, Operations Manager of the Emergency Department of the Prince of Wales Hospital, for his assistance in data collection.

References

1. Ahmed, M.A., Alkhamis, T.M.: Simulation optimization for an emergency department healthcare unit in Kuwait. Eur. J. Oper. Res. **198**(3), 936–942 (2009)
2. Baesler, F.F., Jahnsen, H.E., DaCosta, M.: The use of simulation and design of experiments for estimating maximum capacity in an emergency room. In: Proceedings of the 2003 Winter Simulation Conference, New Orleans, pp. 1903–1906 (2003)
3. Brailsford, S.C., Harper, P.R., Patel, B., Pitt, M.: Analysis of the academic literature on simulation and modelling in health care. J. Simul. **3**, 130–140 (2009)
4. Centeno, M.A., Giachetti, R., Linn, R., Ismail, A.M.: A simulation-ilp based tool for scheduling ER staff. In: Proceedings of the 2003 Winter Simulation Conference, New Orleans, pp. 1930–1938 (2003)
5. Cochran, J.K., Roche, K.T.: A multi-class queuing network analysis methodology for improving hospital emergency department performance. Comput. Oper. Res. **36**(5), 1497–1512 (2009)
6. Connelly, L.G., Bair, A.E.: Discrete event simulation of emergency department activity: a platform for system-level operations research. Acad. Emerg. Med. **11**(11), 1177–1185 (2004)
7. Fetter, R.B., Thompson, J.D.: The simulation of hospital systems. Oper. Res. **13**(5), 689–711 (1965)
8. Evans, G.W., Gor, T.B., Unger, E.: A simulation model for evaluating personnel schedules in a hospital emergency department. In: Proceedings of the 1996 Winter Simulation Conference, Coronado, pp. 1205–1209 (1996)
9. Green, L.V., Soares, J., Giglio, J.F., Green, R.A.: Using queuing theory to increase the effectiveness of emergency department provider staffing. Acad. Emerg. Med. **13**(1), 61–68 (2006)
10. Günal, M.M., Pidd, M.: Discrete event simulation for performance modelling in health care: a review of the literature. J. Simul. **4**(1), 42–51 (2010)
11. Hoot, N.R., LeBlanc, L.J., Jones, I., Levin, S.R., Zhou, C., Gadd, C.S., Aronsky, D.: Forecasting emergency department crowding: a discrete event simulation. Ann. Emerg. Med. **52**(5), 116–125 (2008)
12. Huang, J., Carmeli, B., Mandelbaum, A.: Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. Working Paper, National University of Singapore (2012)
13. Jahangirian, M., Naseer, A., Stergioulas, L., Young, T., Eldabi, T., Brailsford, S., Patel, B., Harper, P.: Simulation in health-care: lessons from other sectors. Oper. Res. **12**, 45–55 (2012)

14. Jun, J.B., Jacobson, S.H., Swisher, J.R.: Application of discrete-event simulation in health care clinics: a survey. J. Oper. Res. Soc. **50**(2), 109–123 (1999)
15. Kuo, Y.H., Leung, J.M.Y., Graham, C.A.: Simulation with data scarcity: developing a simulation model of a hospital emergency department. In: Proceedings of the 2012 Winter Simulation Conference, Berlin (2012)
16. Lane, D.C., Monefeldt, C., Rosenhead, J.V.: Looking in the wrong place for healthcare improvements: a system dynamics study of an accident and emergency department. J. Oper. Res. Soc. **51**(5), 518–531 (2000)
17. Mid-year Population for 2012, Census and Statistics Department, the Government of the Hong Kong Special Administrative Region. http://www.censtatd.gov.hk/press_release/pressReleaseDetail.jsp?charsetID=1&pressRID=2994. 23 Jan 2013
18. Rais, A., Viana., A.: Operations Research in Healthcare: a survey. Int. Trans. Oper. Res. **18**(1), 1–31 (2011)
19. Rossetti, M.D., Trzcinski, G.F., Syverud, S.A. : Emergency department simulation and determination of optimal attending physician staffing schedules. In: Proceedings of the 1999 Winter Simulation Conference, Phoenix, pp. 1532–1240 (1999)
20. Saghafian, S., Hopp, W.J., Van Oyen, M.P., Desmond, J.S., Kronick, S.L.: Patient streaming as a mechanism for improving responsiveness in emergency departments. Oper. Res. **60**(5), 1080–1097 (2012)
21. Saghafian, S., Hopp, W.J., Van Oyen, M.P., Desmond, J.S., Kronick, S.L.: Complexity-based triage: a tool for improving patient safety and operational efficiency. Working Paper, Arizona State University (2012)
22. Yeh, J.Y., Lin, W.S.: Using simulation technique and genetic algorithm to improve the quality care of a hospital emergency department. Expert Syst. Appl. **32**(4), 1073–1083 (2007)

Chapter 24
Managing a Fleet of Ambulances to Respond to Emergency and Transfer Patient Transportation Demands

Y. Kergosien, M. Gendreau, A. Ruiz, and P. Soriano

Abstract Organizations of pre-hospital emergency medical services have as first mission to provide medical assistance to patients including transport to medical centers if necessary, and a second one that concerns the transfers of patients from one medical facility to another one. Most organizations in Canada and in North-America use two independent fleets to perform these missions. Although operating two separated fleets seems easier to do, it appears to be less efficient than an integrated fleet management approach able to deal with both types of demands. Taking this into consideration, this paper aims to design and evaluate the performance of three management approaches. This is a very challenging problem, since it involves solving simultaneously two difficult vehicle routing problems: an ambulance relocation problem and a dial a ride problem.

Y. Kergosien (✉)
Université François Rabelais Tours, Laboratoire d'Informatique (EA 6300), Equipe Ordonnancement et Conduite (ERL CNRS 6305), 64 avenue Jean Portalis, 37200 Tours, France
e-mail: yannick.kergosien@univ-tours.fr

M. Gendreau
École Polytechnique de Montréal, 2900 boul. Édouard-Montpetit, Montréal, QC H3T 1J4, Canada
e-mail: michel.gendreau@cirrelt.ca

A. Ruiz
Département opérations et systèmes de décision, Faculté des sciences de l'administration, Université Laval, QC G1K 7P4, Canada
e-mail: angel.ruiz@fsa.ulaval.ca

P. Soriano
Service de l'enseignement des méthodes quantitatives de gestion, HEC Montréal, 3000 chemin de la Côte Sainte-Catherine, Montréal, QC H3T 2A7, Canada
e-mail: patrick.soriano@cirrelt.net

A. Matta et al. (eds.), *Proceedings of the International Conference on Health Care Systems Engineering*, Springer Proceedings in Mathematics & Statistics 61,
DOI 10.1007/978-3-319-01848-5_24,

24.1 Introduction

The primary mission of pre-hospital emergency medical services organizations (EMS) is to provide the first medical assistance and, if required, the transportation of patients to a medical center. Transportation is performed by paramedic teams using ambulances that are generally deployed at strategic places to respond as soon as possible to emergency demands. The territory deserved is divided into areas. An area is recognized as covered if at least one ambulance can reach any demand in this area within a prefixed time. When an urgent demand appears, the nearest available ambulance is sent to the demand's location. It then may happen that idle ambulances need to be *redeployed* to new waiting locations in order to compensate the void created by the departure of the dispatched ambulance. In addition, most of EMS organizations carry out another mission, which is to respond to transportation demands between medical facilities, and sometimes, to or from the patients' homes. These transfer demands arrive in real time, but fairly in advance with respect to the desired patient departure. This slack allows to schedule them to form ambulances' routes.

Both types of demands, "emergency" and "transfer", can be performed by the same teams and ambulances. Nevertheless, Urgences-Santé currently manages these two types of demands independently by dividing their ambulances into two fleets and managing them separately. This solution was chosen to simplify the management of the vehicles, but it appears to be less efficient than an integrated fleet management approach dealing with both types of demands. Therefore, the aim of this paper is to assess whether or not an integrated management approach can lead to better service quality and to a reduction of costs. In this paper, we first present a management approach to manage efficiently two independent fleets. Then, we proposed two management strategies considering a single integrated fleet. The performances of these three models are compared by several simulation experiments.

This paper is structured as follows. The next section presents a short literature review. Section 24.3 introduces the problem considered here: the management of an integrated fleet of ambulances to respond to both emergency and transfer patient transportation demands. Section 24.4 describes the three management strategies proposed. Section 24.5 focuses on the simulation experiments and results' analysis. Finally, conclusions and further research avenues are provided.

24.2 Literature Review

In the literature, there are two well known families of problems which are closely related to the dynamic management of the two transportation requests here described. The first one is named "Dial A Ride Problem" [13]. In this problem, a set of customer's transportation demands has to be performed by a set of vehicles

under various constraints like maximum ride time, vehicle capacity, maximum ride time constraints, time windows, etc. The problem consists in finding the routes to be done by each vehicle (sequence of requests to perform) minimizing one or several criteria like the total distance traveled or mean user ride time. This problem has been studied both in a static context by Cordeau and Laporte [12] and Parragh et al. [25] as well as in a dynamic context by Attanasio et al. [3] and Xiang et al. [32]. In our case, however, ambulances can transport only one patient at a time. The dial a ride problem (DARP) has been studied in a medical context in [6]. A DARP with heterogeneous users and vehicles is studied in [24] where different modes (seated, stretcher and wheelchair) and types of vehicles are considered. The problem studied in [21] was inspired by a real case of transportation of patients in a hospital complex in France, where transportations are subject to particular constraints. Among them, its worth to mention priority of urgent demands, disinfection of a vehicle after the transportation of a contagion, respect of the type of vehicle needed and the opportunity to outsource demands to a private company.

The second family of problems concerns the ambulance relocation problem. For each emergency demand, the choice of an ambulance to be dispatched must be made. In some cases, a redeployment of ambulances is applied after the dispatching of an ambulance. Redeployment consists in assigning ambulances to potential sites to provide adequate coverage and in order to respond as quickly as possible to a new emergency demand. Several literature reviews have been published over the past years focusing on both ambulance location and relocation problems [9], and recently [7]. The problem was first studied in its static and determinist versions [31] and [11] with one coverage and with multiple coverage have been developed, like in [14, 16] and [15]. Stochastic approaches have also been proposed in [1, 5, 20] and [8]. In a dynamic version, two approaches can be found in the literature. The first one, multi-period, consists in decomposing the day into several periods and apply a redeployment at the beginning of each period [28] or [4]. Several studies using simulation have been conducted to test different deployment plans for each period, either without taking into account the relocation costs between two successive period [27] or taking them into account [10], or even considering a variation in the size of the ambulance fleet between periods [26]. The second approach consists in applying a redeployment according to the evolution of the system's state, [18] and [2]. In [23], it was proposed a dynamic programming approach combined with a Monte Carlo Simulation to determine where the next available ambulance should be redeployed in order to increase the number of demands reached within a given lapse of time. Recently, in [29], an approximate dynamic programming formulation is proposed to solve a dynamic version of the ambulance dispatching and relocation problem taking into account time-dependent information like variations in the demand volumes and travel times throughout the course of the day.

To the best of our knowledge, these two families of problems have been always studied independently in the literature except in one paper [22], where the authors deal with a problem that has a common feature with our case: the simultaneous management of emergency transportations and of transfer transportations between

hospitals. The authors tested different strategies to manage ambulances based on the selection of waiting points, which can only be the hospitals. However they did not solve the associated location problem.

The contributions of this is paper to improve the knowledge on the EMS fleet management problem are twofold. Firstly it proposes three strategies for managing both separated and integrated fleets. Secondly, it proposes efficient solving approaches to tackle the two problems underlying the EMS fleet management: the dial a ride problem and the ambulance relocation problem in a dynamic context.

24.3 Problem Description

A fleet of identical ambulances has to perform two types of transports during a given planning period. A team composed of two technicians having their own work time schedule is associated to each ambulance. At the beginning of their working shift, the team picks an ambulance at a given depot and must return it to the same depot at the end of their shift. Managing the fleet consists basically in assigning transportation requests to ambulances, and locating idle ambulances to standby points in such a way that they will respond as quick as possible to emergency demands. The whole problem can be divided into two subproblems (a dial a ride problem and an ambulance relocation problem) where the pool of ambulances, is shared. To evaluate the performance of a management strategy, we consider three types of objectives. The first one concerns how the fleet is able to perform all the transfer demands respecting their required time windows. This objective is measured by computing the sum of all the transportation delays. A second objective aims at maximize the total urgent demands covered. The last objective seeks at minimizing the operating costs, measured by the sum of empty running of ambulances as well as the technicians' overtime.

24.3.1 Transfer Patient Transportation Problem

At every moment, the set of transfer transportation demands is known. This set can change over time due to the arrival of new demands or canceling of existing demands. Each transfer transportation demand is characterized by:

- An origin point and a destination which are usually a hospital but may also be a patient's home,
- A time window that modelizes the earliest time at which the patient is ready to be transported and the latest time accepted for the beginning of the transportation. After such a latest time, a delay is incurred.
- A time needed to transfer the patient outside of the ambulance (administrative tasks, stretcher transportation, etc.) that has to be taken into account for both the destination and the origin of the transport.

24.3.2 Ambulance Relocation Problem

When the coverage offered by the available ambulances is not acceptable, a new relocation plan has to be computed. The problem consists in finding new standby locations for the ambulances. The potential standby points are known in advance and are usually strategic locations in the region that should be covered. Also, a maximum number of ambulances can be assigned to each standby point. The desired coverage is the same as that defined in [17]: two types of covering constraints in agreement with the United States EMS Act of 1973. These constraints specify that all emergency demands must be satisfied by an ambulance within S' minutes and a proportion α of the total demand is also satisfied within S minutes ($S' > S$). Thereby each zone of the region is characterized by:

- Two sets of potential sites. One represents the potential sites from which an ambulance can reach all points of the zone within S, and within S' for the other set.
- A density of population.
- A probability vector. The probability vector of a zone A gives, for each period, the probability that the next demand occurs in A. Since this probability can fluctuate according to the time of day, a day is decomposed into several periods (2 h long by default).

When an emergency demand occurs, an ambulance should be selected to respond to this demand. An emergency demand is defined by:

- An intervention time at the scene,
- Whether a transport to a hospital is required or not,
- The hospital destination,
- And a time needed at the hospital to transfer the patient to the hospital staff.

Finally, a minimum time has to be respected between two consecutives redeployments of a given ambulance in order to avoid to redeploy an ambulance too many times in a short period.

24.3.3 Performance Evaluation

To evaluate the performance of a given management strategy, we consider the following criteria related to service quality and costs:

- The elapse time between the arrival of an emergency call and the arrival of the ambulance at the scene for emergency demands,
- The delays for transfer demands,
- The workload of each team and overtime if any,
- The number of times that ambulances were redeployed or diverted,
- The number of deployments or redeployments,
- And the total distance traveled by empty ambulances.

24.4 Management Approaches

In order to assess whether or not an integrated management approach may improve service quality and lead to reduction of costs, this section proposes three management strategies. The first one, named Independent management, corresponds to the dominant strategy that consists in dividing the ambulances into separated fleets, which will deal with transfer and urgent requests, respectively. The second and third strategies consider a single fleet to respond to both types of demand. However, the second strategy, named Reactive integrated fleet management, adopts a pure reactive approach (i.e. demands are considered at their execution time) whereas the third one, named Proactive integrated fleet management, uses a proactive scheduling that consists in deciding the execution date of each transfer demand in order to minimize the number of ambulances that will be busy simultaneously. The three strategies as well as the related tools proposed to solve the ambulances relocation and assignment decisions are described in the next subsections.

24.4.1 Independent Management

This management is based on two fleets: the first one responds to emergency demands only and the second one responds to transfer demands. Both fleets are managed independently and the number of ambulances assigned to them is kept constant during the planning horizon. Strategies to manage each fleet are now described.

24.4.1.1 Management of the Emergency Requests Fleet

In order to have an adequate coverage at any time, an ambulance redeployment is computed and applied if and only if not all zones are covered within S and the last redeployment is not too recent (e.g. more than 15 min ago). Therefore, after each event like an ambulance becomes available (e.g. a team starts its shift or an ambulance finishes serving a demand) or becomes unavailable (e.g. a team finishes its shift or an ambulance is assigned to a demand), a redeployments can occur. If an ambulance becomes available and no redeployment is needed, two cases are possible. In the first case at least one zone is not covered within S'; then the ambulance is sent to the site that maximizes the number of zones that are covered. Otherwise: the ambulance is sent to the site that maximizes the sum over all doubly covered zones within S of the probability that the next demand will appear in that zone. When a new demand occurs, the nearest available ambulance is dispatched.

Redeployment decisions are done by solving an integer linear program based on [16] and solved by CPLEX. In [16], the ambulance redeployment problem is modeled with the same types of coverage. The objective function is to maximize the

sum of the zones that are covered twice within S minutes weighted by the probability of a new demand occurring in that zone. The covering constraints and the constraints on the minimum time before redeployment for each ambulance are relaxed to avoid getting infeasible solutions. However, their violation is strongly penalized in the objective function. The solution produced indicates the new number of ambulances that should located at each standby point. Then, the specific instructions for each ambulance are decided by minimizing the total travel distance. This is done after solving a min-cost max-flow problem.

24.4.1.2 Management of the Transfer Requests Fleet

To manage the ambulance fleet and the transfer demands, each ambulance is associated to a route (e.g. a sequence of transfer requests to be performed). Routes are recomputed by using a fast and efficient tabu search algorithm each time a new event happens. An event consists in a new request arrival or the cancellation of an existing request. The initial solution for each execution of the tabu search is the best solution found at the previous event, with some updates like the new position of each ambulance, the demands completed since them, etc. When a new demand occurs, it is included in the route of one ambulance before executing the tabu search in such a way that the sum of delays is minimized. The tabu search uses a lexicographic objective function: it first minimizes the sum of transportation delays and crew overtimes, then the sum of traveled distances.

The tabu search algorithm has a structure as in [19]. A solution is the set of routes for each ambulance. The stopping criterion is the maximum number of iterations without improving the best solution found so far. The neighborhood is built by the CROSS exchange operator, which is particularly well suitable for vehicle routing problems with time windows [30]. In CROSS, the neighborhood of a solution is obtained by exchanging all the sub-segments (parts of a route) of all routes. As in [30], the tabu list contains the objective function values. This way of managing the tabu condition helps reducing computation time as well as the risk of cycling between visited solutions, since the likelihood of having two different solutions with the same objective is very low.

24.4.2 Reactive Integrated Fleet Management

This approach considers a single ambulance fleet responding to both types of demands. The main idea of this strategy is to consider all the demands, transfer or urgent, as emergency demands. The "planning" part of the problem (scheduling of the transfer requests) disappears and only the location and relocation plans need to be solved.

Whenever a transfer request arrives to the system, it is transformed into a *dummy* urgent request that will happen at the patient earliest transportation date. Dummy emergency demands are then managed like the real emergency demands but, unlike

them, they can be postponed in some cases that will be described later. To take into account dummy demands in the ambulance relocation problem, the probability vector $P_t(a)$ of a zone a, that define the probability that the next demand will arrive in that zone for each period, is changed according to the Eq. 24.1. For a period t, this probability is a weighted average of the probability that a new real emergency demand happens and an average of the probability of a transfer demand occurring in that zone. The former probability is an average between the probability that a transfer demand occurs in one of the hospitals belonging to that zone and the proportion of the transfer demands known in that zone.

$$P_t(a) = (1-\beta)Pe_t(a) + 0.5\beta\left(\sum_{h\in a} Ps_t(h) + \frac{U(a)}{\#dmd}\right) \tag{24.1}$$

- β: Proportion of transfer demands (a parameter computed from historical data).
- $Pe_t(a)$: Probability of a new emergency demand occurring in zone a.
- $Ps_t(h)$: Probability of a transfer demand occurring at hospital h.
- $U(a)$: Number of transfer demands known from hospitals belonging to zone a.
- $\#dmd$: Total number of transfer demands known.

The fleet management strategy corresponds to the one described in Sect. 24.4.1.1 section. However, in order to avoid situations where too many ambulances will be occupied by dummy demands, the number of ambulances devoted to these requests is limited. Beyond this limit, the demands are added to a list of *postponed* demands. Once an ambulance becomes available, and if the number of dummy requests being performing is below the mentioned limit, then an ambulance may assigned to the first request in the postponed list.

24.4.3 Proactive Integrated Fleet Management

The main idea of this management approach is to improve the previous ones by anticipating the best dates to perform the transfer demands. Basically, the proactive management reproduces the strategy in Sect. 24.4.2 without the constraint which limits the number of ambulances that are responding to transfer demands simultaneously. The execution dates for transfer requests are calculated in order to minimize the number of ambulances that will be busy at the same time, and the sum of transfer transportation delays. These execution dates represent the dates at which the dummy emergency demands will occur. The new problem consists then in schedules the transfer demands, and can be modeled as parallel machines tasks scheduling problem where the tasks are the requests and the machines are the ambulances. To solve this problem we propose a method that can be seen as a proactive scheduler for the transfer demands. Processing times of tasks are

an estimate of the time required to perform the transfer demands obtained by computing an estimate of the travel time to move to the departure hospital, plus the travel time, plus the patient transfer time to hospitals staff. The machines have some periods of unavailability according to ambulance activities and their work schedules. Once the problem is solved, the solution indicates the execution dates to perform the transfer demands by sending an ambulance, but the specific ambulances/tasks assignments are not used to keep additional flexibility. The ambulance that will be sent will not be the nearest but the one minimizing the deterioration of the coverages.

We used a tabu search algorithm to solve efficiently this problem. An indirect encoding of the solutions, based on a sequence of tasks, was used in order to simplify the method. To build and evaluate an actual solution, each task is placed iteratively in its best place in the order of the encoding sequence. A partial schedule to the problem is evaluated by the sum of assignment costs where the cost of assigning a task i at a date x is given by the Eq. 24.2. The neighborhood is built by a swap operator and the tabu list contains the objective function values.

$$c_i^x = (1-\lambda) \sum_{y=x}^{y=x+p_i} W_y n_y + \lambda \; delay_i(x) \tag{24.2}$$

- p_i: processing time of i.
- W_y: probability that a new emergency demand occurs at time y.
- n_y: number of ambulances used at time y.
- $delay_i(x)$: delay of i.
- λ: weight applied to two criteria ("coverage degradation" and "transport delay").

24.5 Numerical Experiments and Preliminary Results

To assess the performance of the described strategies, we designed several simulation experiments. To this end, we developed a generic discrete-event simulation model. In order to increase the flexibility of the model, we separated the decision logic from the routines simulating the physical processes (ambulances movements).

To generate realistic instances, we inspired by the real case of the Urgences-Santé, the EMS organisation in Montreal, Canada. Based on previous published works, we proposed parameters and probability distributions of random variables. Some of the main characteristics of this case are now reported. Montreal is divided into 595 zones. Across the city, there are also the 39 potential sites that can hold up to 4 ambulances, 2 depots and the 15 hospitals. A total of 153 paramedical teams were considered. An exponential distribution is used to model the inter-arrival times between two emergency demands. Depending on the period of the day, the mean of the exponential distribution varies between 110 and 278 s. The time at call location

and the time needed at the hospital to take care the patient to the end are generated using the gamma distribution Gamma(k;θ) with θ the scale parameter and k the shape parameter. We assumed that, if no transport to hospital is needed, the time at call location in minutes is generated with Gamma(28.5;19.5) with a probability of 0.85 and Gamma(6.5;4) otherwise. If a transport to hospital is needed, the time at call location is generated with Gamma(16.5;7.2) and the time needed at the hospital with Gamma(42;15). The destination hospital is randomly chosen such as the nearest hospitals have a strong probability of selection.

We also assumed that the number of transfer demands and their arrival times follow a combination of two normal distributions, one during the morning and second one during the afternoon. The probability of selection of the morning's normal distribution for the arrival times were considered slightly larger (0.55). If we note $N(\mu;\sigma^2)$ a normal distribution with a mean μ and a variance σ^2, the morning's normal distribution and the afternoon's normal distribution were set respectively to $N(10\,\text{h};2\,\text{h})$ and $N(17\,\text{h};2\,\text{h})$. We also assumed that approximately 5 % of the demands are canceled. The time at which the cancelation is known is randomly generated at a moment between half an hour after the time where the demand is known and the latest date for the demand execution. Each transfer demand can start during a time window. The earliest date is generated between half an hour and 5 h after the time where the demand is known, using a normal distribution $N(2\,\text{h}30;1\,\text{h})$. The size of the time windows is randomly generated between half an hour and 4 h. The probability that the departure or the destination is a hospital is equal to 0.85, but the departure and the destination cannot both be a patient home. The times need at the places of the departure and the destination are generated using a normal distribution $N(20;7)$ (minutes).

Based on these assumptions, we generated 20 instances of 7 days covering 5 months. However, our analyses and results are based only on the 5 middle-days to remove the transitory states of the first and last day. We tested the three management strategies on all instances. To reduce as much as possible the variance in the results, the random events were kept exactly the same for all the three management strategies for a given instance. The results are summarized in Table 24.1. For each management strategy we report the average of the following indicators over the 20 instances: the number of emergency demands without and with transports, the average response times, the coverages according to S and S', the number of transfer demands, the sum of transport delays and the number of late transfer demands. We also noted some criteria related to efficiency of the strategy: the number of diversions, the number of redeployments, the number of ambulances redeployed, the percentage of ambulances empty travels and the overtime of paramedics.

We can conclude that, regarding the quality of services for emergency and transfer demands, the results of Independent fleet management are clearly worse than the ones produced by the other management strategies. However, the fleet efficiency indicators (number of diversions, average ambulances empty travels and the overtime of paramedics) are better when using Independent fleet management. Even if the Reactive fleet management seems a rather simple strategy, it produces

Table 24.1 Results

	Management type		
	Independent	Without planning	Robust
Emergency demands			
Number of demands without transport	589.1	589.1	589.1
Number of demands with transport	1,762.2	1,762.2	1,762.2
Average response time (R.T.) in sec.	469.5	454.6	446.1
Percentage of demands such that R.T.$\leq S$	68.8	71.0	73.4
Percentage of demands such that R.T.$\leq S'$	85.1	86.9	88.3
Percentage of demands such that R.T.$> S'$	14.9	13.1	11.7
Transfer demands			
Number of demands	452.2	452.2	452.2
Sum of transport delays in sec.	2,397.7	473.1	661.0
Number of late demands	46.3	3.2	16.0
Ambulances			
Number of diversions	2,407.9	2,640.3	2,584.4
Number of redeployments	202.6	182.7	182.8
Number of ambulances redeployed	611.2	585.7	594.7
Percentage of average ambulances empty travels	22.8	24.6	24.5
Average overtime of paramedics	1,521.1	1,593.8	1,709.2

a good balance between the service quality of emergency demands and transfer demands: both the delays and the coverages are strongly improved with respect to the Independent fleet management case. Unfortunately, if we decrease the number of ambulances limited to perform the transfer transports, the coverage is slightly improved but the transfer transport delays are considerably increased. Finally, the Proactive fleet management strategy improves the service quality of the emergency demands but it deteriorates the delays of transfer demands.

24.6 Conclusion

This paper proposes new fleet management strategies to deal with two types of demands, transfer and emergency demands, in an integrated manner. We showed that these strategies can improve the quality of the service without increasing the number of ambulance in the fleet. The proposed Proactive fleet management strategy is very promising, and it is also simple to implement in a real setting. One of our future research questions concerns the consideration of the breaks in paramedics schedules. We also intend to formulate a single model for the two problems (DARP and relocation problem) as well as integrated solving approach to improve the results.

References

1. Alsalloum, O.I., Rand, G.K.: Extensions to emergency vehicle location model. Comput. Oper. Res. **33**, 2725–2743 (2006)
2. Andersson, T., Petersson, S., Varband, P.: Decision support for efficient ambulance logistics. ITN research report LiTH-ITN-R-2005-1, Linkopings Universiteit (2007)
3. Attanasio, A., Cordeau, J.F., Ghiani, G., Laporte, G.: Parallel tabu search heuristics for the dynamic multi-vehicle dial-a-ride problem. Parallel Comput. **30**, 377–387 (2004)
4. Başar, A., Catay, B., Unluyurt, T.: A multi-period double coverage approach for locating the emergency medical service stations in Istanbul. J. Oper. Res. Soc. **62**, 627–637 (2011)
5. Batta, R., Dolan, J.M., Krishnamurty, N.N.: The maximal expected covering location problem: revisited. Transp. Sci. **23**, 277–287 (1989)
6. Beaudry, A., Laporte, G., Melo, T., Nickel, S.: Dynamic transportation of patients in hospitals. OR Spectr. **32**, 77–107 (2009)
7. Bélanger, V., Ruiz, A., Soriano, P.: Déploiement et redéploiement des véhicules ambulanciers dans la gestion des services préhospitaliers d'urgence. INFOR. **50**, 1–30 (2012)
8. Beraldi, P., Bruni, M.E., Krishnamurty, N.N.: A probabilistic model applied to emergency service vehicle location. EJOR **196**, 323–331 (2009)
9. Brotcorne, L., Laporte, G., Semet, F.: Ambulance location and relocation models. EJOR **147**, 451–463 (2003)
10. Carpentier, G.: La conception et la gestion d'un réseau de service ambulancier. Mémoire de maîtrise, Université Laval (2006)
11. Church, R.L., ReVelle, C.S.: The maximal covering llocation problem. Pap. Reg. Sci. Assoc. **32**, 101–118 (1974)
12. Cordeau, J.F., Laporte, G.: A tabu search heuristic for the static multi-vehicle dial-a-ride problem. Transp. Res. B Methodol. **37**, 579–594 (2003)
13. Cordeau, J.F., Laporte, G.: The dial-a-ride problem: models and algorithms. Ann. Oper. Res. **153**, 29–46 (2007)
14. Daskin, M.S., ReVelle, C.S.: A hierarchical objective set covering model for emergency medical service vehicle deployment. Transp. Sci. **15**, 137–152 (1981)
15. Doerner, K.F., Gutjahr, W.J., Hartl, R.F., Karall, M., Reimann, M.: Heuristic solution of an extended double-coverage ambulance location problem for austria. EJOR **13**, 325–340 (2005)
16. Gendreau, M., Laporte, G., Semet, F.: Solving an ambulance location model by tabu search. Locat. Sci. **5**, 75–88 (1997)
17. Gendreau, M., Laporte, G., Semet, F.: A dynamic model and parallel tabu search heuristic for real-time ambulance relocation. Locat. Sci. **27**, 1641–1653 (2001)
18. Gendreau, M., Laporte, G., Semet, F.: The maximal expected relocation problem for emergency vehicles. J. Oper. Res. Soc. **57**, 22–28 (2006)
19. Glovert, F.: Future paths for integer programming and links to artificial intelligence. Comput. Oper. Res. **13**, 533–549 (1986)
20. Ingolfsson, A., Budge, S., Erkut, E.: Optimal ambulance location with random delays and travel times. Health Care Manag. Sci. **11**, 262–274 (2008)
21. Kergosien, Y., Lenté, C., Piton, D., Billaut, J.-C.: A tabu search heuristic for the dynamic transportation of patients between care units. EJOR **214**, 442–452 (2011)
22. Kiechle, G., Doerner, K.F., Gendreau, M., Hartl, R.F.: Waiting strategies for regular and emergency patient transportation. In: Operations Research Proceedings 2008, Augsburg, vol. 6, pp. 271–276 (2008)
23. Maxwell, M.S., Restepo, M., Henderson, S.G., Topaloglu, H.: Approximate dynamic programming for ambulance redeployment. INFORMS J. Comput. **22**, 266–281 (2010)
24. Parragh, S.N.: Introducing heterogeneous users and vehicles into models and algorithms for the dial-a-ride problem. Res. C Emerg. Technol. **19**, 912–930 (2011)
25. Parragh, S.N., Doerner, K.F., Hartl, R.F.: Variable neighborhood search for the dial-a-ride problem. Comput. Oper. Res. **37**, 1129–1138 (2010)

26. Rajagopalan, H.K., Saydam, C., Xiao, J.: A multiperiod set covering location model for dynamic redeployment of ambulances. Comput. Oper. Res. **35**, 814–826 (2008)
27. Repede, J.F., Bernardo, J.J.: Developping and validating a decision support system for location emergency medical vehicles in Louisville. EJOR **75**, 567–581 (1994)
28. Schmid, V.: Solving the dynamic ambulance relocation and dispatching problem using approximate dynamic programming. EJOR **219**, 611–621 (2012)
29. Schmid, V., Doerner, K.F.: Ambulance location and relocation problems with time-dependant travel times. EJOR **207**, 1293–1303 (2010)
30. Taillard, E., Badeau, P., Gendreau, M., Guertin, F., Potvin, J.Y.: A tabu search heuristic for the vehicle routing problem with soft time windows. Transp. Sci. **31**, 170–186 (1997)
31. Toregas, C., Swain, R., ReVelle, C.S., Bergman, L.: The location of emergency service facilities. Oper. Res. **19**, 1363–1373 (1971)
32. Xiang, Z., Chu, C., Chen, H.: The study of a dynamic dial a ride problem under time dependant and stochastic environments. EJOR **185**, 53–551 (2008)

Erratum to:

Chapter 23
Using Simulation to Analyze Patient Flows in a Hospital Emergency Department in Hong Kong

Omar Rado, Benedetta Lupia, Janny M.Y. Leung, Yong-Hong Kuo, and Colin A. Graham

A. Matta et al. (eds.), *Proceedings of the International Conference on Health Care Systems Engineering*, Springer Proceedings in Mathematics & Statistics 61,
DOI 10.1007/978-3-319-01848-5_23, © Springer International Publishing Switzerland 2014

DOI 10.1007/978-3-319-01848-5_25

The affiliation of the author names Y.-H. Kuo and C.A. Graham were incorrect. The correct information is given below:

Y.-H. Kuo
Faculty of Management and Administration
Macau University of Science and Technology
Avenida, Taipa, Macau
e-mail: yhkuo@must.edu.mo

C.A. Graham
Accident and Emergency Medicine Academic Unit
The Chinese University of Hong Kong
Shatin, New Territories, Hong Kong
e-mail: cagraham@cuhk.edu.hk

The original online version of this chapter can be found at
http://dx.doi.org/10.1007/978-3-319-01848-5_23

A. Matta et al. (eds.), *Proceedings of the International Conference on Health Care Systems Engineering*, Springer Proceedings in Mathematics & Statistics 61,
DOI 10.1007/978-3-319-01848-5_25, © Springer International Publishing Switzerland 2014