Springer Proceedings in Mathematics & Statistics

Valérie Bélanger
Nadia Lahrichi
Ettore Lanzarone
Semih Yalçındağ *Editors*

Health Care Systems Engineering

HCSE, Montréal, Canada,
May 30–June 1, 2019

Springer

Springer Proceedings in Mathematics & Statistics

Volume 316

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of selected contributions from workshops and conferences in all areas of current research in mathematics and statistics, including operation research 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.

More information about this series at http://www.springer.com/series/10533

Valérie Bélanger · Nadia Lahrichi ·
Ettore Lanzarone · Semih Yalçındağ
Editors

Health Care Systems Engineering

HCSE, Montréal, Canada,
May 30–June 1, 2019

Springer

Editors
Valérie Bélanger
Logistics and Operations Management
HEC Montreal
Montreal, QC, Canada

Ettore Lanzarone
CNR-IMATI
Milan, Italy

Nadia Lahrichi
Mathematics and Industrial Engineering
Polytechnique Montreal
Montreal, QC, Canada

Semih Yalçındağ
Industrial and Systems Engineering
Yeditepe University
Istanbul, Turkey

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-030-39693-0 ISBN 978-3-030-39694-7 (eBook)
https://doi.org/10.1007/978-3-030-39694-7

Mathematics Subject Classification (2010): 37N40, 46N10, 65K05

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Patient Flow and Transportation

Non-emergency Patient Transfer Scheduling and Assignment 3
Travis Foster, Peter VanBerkel, Uday Venkatadri and Theresia van Essen

Optimizing Operator's and Users' Objectives in Non-emergency Patients Transportation 13
Jamal Abdul Nasir, Yong-Hong Kuo and Reynold Cheng

Modelling Hospital Medical Wards to Address Patient Complexity: A Case-Based Simulation-Optimization Approach 25
Paolo Landa, Micaela La Regina, Elena Tànfani, Francesco Orlandini, Mauro Campanini, Andrea Fontanella, Dario Manfellotto and Angela Testi

Benefits of a Broader View: Capturing the Hospital-Wide Impact of Surge Policies with Discrete-Event Simulation 41
Carolyn R. Busby and Michael W. Carter

Coxian Phase-Type Regression Models for Understanding the Relationship Between Patient Attributes, Overcrowding, and Length of Stay in Hospital Emergency Departments 53
Laura M. Boyle, Adele H. Marshall and Mark Mackay

Emergency

A Realistic Simulation Model of Montreal Emergency Medical Services 67
Gabriel Lavoie, Valérie Bélanger, Luc de Montigny and Nadia Lahrichi

A Two-Phase Approach to the Emergency Department Physician Rostering Problem 79
Paola Cappanera, Filippo Visintin and Roberta Rossi

Using a Slotted Queuing Model to Compare the Efficacy of Emergency Departments Operating with and Without a Physician in Rural Communities 93
Peter T. Vanberkel, Benjamin Wedge, Alix J. E. Carter and Ilze Ziedins

A Meta Algorithm for Reinforcement Learning: Emergency Medical Service Resource Prioritization Problem in an MCI as an Example 103
Kyohong Shin and Taesik Lee

Facing Implementation Barriers to Healthcare Simulation Studies 117
Clio Dosi, Manuel Iori, Arthur Kramer and Matteo Vignoli

Operating Room

Reallocating Operating Room Time: A Portuguese Case 133
Mariana Oliveira, Luísa Lubomirska and Inês Marques

Evaluating Replenishment Systems for Disposable Supplies at the Operating Theater: A Simulation Case Study 147
Karen Moons, Geert Waeyenbergh, Paul Timmermans, Dirk De Ridder and Liliane Pintelon

Stochastic Master Surgical Scheduling Under Ward Uncertainty 163
Asgeir Orn Sigurpalsson, Thomas Philip Runarsson and Rognvaldur J. Saemundsson

Home Health Care

Integration of User's Preferences into the Home Healthcare Routing and Scheduling Multi-objective Problem: A Hierarchical Approach with Pareto-Optimal Alternative Solutions 179
Laura Musaraganyi, Simon Germain, Nadia Lahrichi and Louis-Martin Rousseau

A Two-Phase Method for Robust Home Healthcare Problem: A Case Study 193
Mahdyeh Shiri, Fardin Ahmadizar, Houra Mahmoudzadeh and Mahdi Bashiri

Adverse Event Prediction by Telemonitoring and Deep Learning 205
Antoine Prouvost, Andrea Lodi, Louis-Martin Rousseau and Jonathan Vallee

Mass Casualty Events: A Decision Making Tool for Home Health Care to Discharge Conventional Hospitals 217
Alain Guinet and Eric Dubost

Radiology, Radiotherapy and Chemotherapy

Simultaneous Optimization of Appointment Grid and Technologist Scheduling in a Radiology Center 231
Dina Bentayeb, Nadia Lahrichi and Louis-Martin Rousseau

A Mathematical Programming Model for Radiotherapy Scheduling with Time Windows 241
Bruno Vieira, Derya Demirtas, Jeroen B. van de Kamer, Erwin W. Hans, Louis-Martin Rosseau, Nadia Lahrichi and Wim H. van Harten

Pattern-Based Online Algorithms for a General Patient-Centred Radiotherapy Scheduling Problem 251
Roberto Aringhieri, Davide Duma and Giuseppe Squillace

Multi-level Heuristic to Optimize the Chemotherapy Production and Delivery 263
Alexis Robbes, Yannick Kergosien and Jean-Charles Billaut

Primary Care

Acuity-Based Access Time Evaluation in Primary Care: A Case Study of an Ontario Clinic 277
Nazanin Aslani, Fariborz Fazileh, Donatus Mutasingwa and Daria Terekhov

Blood System

Uncertainty in the Blood Donation Appointment Scheduling: Key Factors and Research Perspectives 293
Ettore Lanzarone and Semih Yalçındağ

About the Editors

Valérie Bélanger is an Assistant Professor at the HEC Montréal. She holds an MS in Mechanical Engineering (Université Laval) and a PhD in Decision Sciences (HEC Montréal). Her research focuses on health care logistics and emergency service management, while her methodological expertise is operations research. She works with various organizations in the health care sector on projects related to patient and material transportation, home care, and network design.

Nadia Lahrichi holds a PhD in Operations Research from the Polytechnique Montréal, where she is now an Associate Professor at the Department of Mathematical and Industrial Engineering. Her research focuses in applying modeling and operational research tools in healthcare, and she has recently investigated patient flow optimization, scheduling and workforce planning problems.

Ettore Lanzarone is a Researcher at the Institute for Applied Mathematics and Information Technology (IMATI) of the National Research Council of Italy (CNR), and an Adjunct Professor at the University of Bergamo and the Politecnico di Milano. He holds a PhD in Bioengineering and an MS in Biomedical Engineering from the Politecnico di Milano. His research activities include operations research applied to health care, stochastic modeling, and bioengineering.

Semih Yalçındağ is an Assistant Professor at the Industrial and Systems Engineering Department of the Yeditepe University. He holds an MS in Industrial Engineering from Sabancı University, and a joint PhD in Industrial Engineering from Politecnico di Milano and École Centrale Paris. His research focuses on applied operations research, and he is currently woking on modeling and solving optimization problems in healthcare operations, scheduling and sequencing, transportation and sustainable operations.

Patient Flow and Transportation

Non-emergency Patient Transfer Scheduling and Assignment

Travis Foster, Peter VanBerkel, Uday Venkatadri and Theresia van Essen

Abstract Emergency Medical Services organizations are responsible for providing paramedic crews, vehicles and equipment to transfer patients from one location to another in emergency and non-emergency settings. They must solve difficult scheduling and assignment problems to ensure on-time arrival of patients and the efficient use of health care resources during non-emergency operations. Ambulances can serve both emergency and non-emergency requests but are rarely available to serve non-emergency requests. Therefore, non-emergency requests are the responsibility of Patient Transfer Units. The objective of this study is to develop a mathematical model that will assign Patient Transfer Units to non-emergency patient transfer requests, design a schedule that will minimize travel costs and balance workloads and apply it to a real-world case study. This paper also proposes a framework to utilize historical patient transfer data in the scheduling process. The mathematical model provides decision support for the non-emergency patient transfer scheduling process.

Keywords Healthcare · Emergency Medical Services · Vehicle Routing

T. Foster (✉) · P. VanBerkel · U. Venkatadri
Dalhousie University, 6299 South St, Halifax, NS B3H 4R2, Canada
e-mail: tfos@dal.ca

P. VanBerkel
e-mail: peter.vanberkel@dal.ca

U. Venkatadri
e-mail: uday.venkatadri@dal.ca

T. van Essen
Delft University of Technology, Van Mourik Broekmanweg 6,
2628, XE Delft, Netherlands
e-mail: j.t.vanessen@tudelft.nl

V. Bélanger et al. (eds.), *Health Care Systems Engineering*,
Springer Proceedings in Mathematics & Statistics 316,
https://doi.org/10.1007/978-3-030-39694-7_1

1 Introduction

This paper examines the scheduling and assignment of non-emergency patient transfer requests. Such transfers can be between any of the following locations: a special care facility, a hospital or a personal residence. Patient transfers are an important part of public safety systems as they improve patient care by allowing access to proper and continuing medical care to patients. In some jurisdictions of Canada, transfers are conducted by Emergency Medical Service (EMS) providers with paramedic crews in ambulances or similar vehicles. Increasing transfer volumes add pressure to EMS providers [10].

From a scheduling and assignment perspective, patient transfers present a unique challenge to EMS and Operations Research (OR) scientists. Patient transfers are non-emergency requests and are often scheduled according to:

- Arrival time of request (advance notice or same day).
- The requested time of pickup.
- The availability of transport vehicles in the region.
- Logistical issues surrounding the transfer such as equipment required, current vehicle location and future pickups.

Many health care organizations, including EMS providers, collect and store large amounts of historical data. This paper presents a framework for using this historical data to help a model for patient transfer scheduling and assignment at the offline operational level [6].

In Sect. 2 we describe the scheduling problem faced by Nova Scotia EMS providers. In Sect. 3 we review related literature and position our research. In Sect. 4 we present the scheduling model and the framework for integrating their data within the model. In Sect. 5 we review our results on a real-world case study. In Sect. 6 we present our conclusions and future work.

2 Problem Description

This research is motivated by Emergency Health Services (EHS). EHS is the organization that provides EMS to the province of Nova Scotia in Canada. EHS uses vehicles called Patient Transfer Units (PTUs) for a significant portion of non-emergency patient transfers. For simplicity, the term “patient transfer” will refer specifically to non-emergency patient transfers.

Patient transfer requests are phoned into EHS when the sending facility has determined a patient requires extra care for transport to a medical appointment. After a pickup time has been agreed upon between EHS and the requesting party, EHS schedules the transfer in their system. A preliminary schedule for the following day is created in the evening with all requests that were submitted in advance. Same

day requests are submitted to EHS the following day during the operation period. The schedule is adjusted to account for these same day requests. EHS also assigns a paramedic crew (who are operating a PTU) during the day-of operation. This assignment depends on several factors including the location of the patient, the location and status of the crews, future patient transfers and operator knowledge. Delaying or arriving late to a patient pickup can result in a cancelled or rescheduled appointment for the patient and a deadhead trip (a completed trip without a patient) for a PTU to the next patient pickup.

This paper focuses on scheduling advance notice patient transfer requests to minimize total travel time while using historical data to inform the model. We are also interested in introducing workload balancing features to the model. We do not examine the same day scheduling portion of the problem at this time nor do we address crew assignment; in this phase of the research, only vehicle assignment and scheduling are considered. This model is then applied to a real-world case study in Nova Scotia.

3 Related Research

Patient transfer systems are a common part of health care systems and as such, scheduling patient transfers has been studied extensively. Patient transfer systems can be modelled as dial-a-ride problems (DARP), a class of Vehicle Routing Problems (VRP) and part of the Travelling Salesman group of problems. The dial-a-ride model develops vehicle routes and schedules for n requests divided among k vehicles. We refer readers to Cordeau and Laporte [3] and Ho et al. [7] for additional information regarding the DARP.

Detti et al. [4] studied a real-world health care example of the DARP with several constraints and multiple vehicle depots. They analyze the effectiveness of their heuristics and a Mixed Integer Programming (MIP) model from real and randomly generated data. Guerriero et al. [5] solved a multi-objective DARP considering travel and patient waiting time and demonstrated computational results from their two step heuristic approach. Workload balancing features such as cost-related objective functions and constraints are explored by Matl et al. [9] in VRPs, including the DARP. They also review types of VRP workload balancing measures. Berg and Essen [1] examined scheduling vehicles for patient transfer coverage while minimizing the impact on emergency vehicles. Marković et al. [8] developed prediction models using statistical and machine learning algorithms for capacity requirements of a new dial-a-ride system. Yalçındağ et al. [11] used data driven methods to estimate travel times in home health care.

Our paper applies a DARP scheduling model including workload balancing, and data driven statistical models to estimate travel and service times from historical data to act as an efficient scheduling tool and applied to a real-world case study of a non-emergency patient transfer system.

4 Methods

In this section, we first review the approach and assumptions made when modelling the scheduling process with a DARP model. This approach is used to mimic the actual scheduling process where a preliminary schedule is created the previous day for all patient transfer requests submitted in advance. Second, we discuss the historical data and how it is used as an input to the model.

4.1 Advance Request Model

The model creates a set of routes for k PTUs and n requests while minimizing the travel time across all routes. The model assumes a single depot to act as the start and end point for every vehicle. We also assume that we have a homogeneous fleet and patients. In reality, patients do have different needs but this has little impact on the time required for a request.

Our DARP model is based off of the three-index formulation found in Cordeau [2]. However, our model includes workload balancing constraints and variable shift times for the PTU crews. We use the time windows to ensure patients arrive at their destination in a timely manner. It is formulated on a directed graph $G = (V, A)$. All vertices on the graph are represented by $i, j = (0, ..., 2n + 1)$. The pickup nodes are represented by $P = (1, ..., n)$ and the drop-off nodes are represented by $D = (n + 1, ..., 2n)$. The depot is represented by nodes 0 and $2n + 1$. These three indices make up the vertex set $V = (0, 1, ..., n, n + 1, ..., 2n, 2n + 1)$. Each request is treated as a pair $(i, n + i)$ that must be visited in order and by the same vehicle.

K represents the set of vehicles. Each vehicle $k \in K$ has a capacity of Q_k, a minimum shift start time of $Tmin_k$ and a maximum shift end time of $Tmax_k$. Each node $i \in V$ has a service time d_i and a load q_i such that $q_{n+i} = -q_i$. These values for the depot are such that $d_0 = d_{2n+1} = q_0 = q_{2n+1} = 0$. Each node has a time window $[e_i, l_i]$ where e_i and l_i are the earliest and latest times that service may begin at node i. Each arc (i, j) has an associated travel time t_{ij}. We use the parameters wb^+ and wb^- as the maximum and minimum workload, respectively, for each PTU.

The model has three types of decision variables:

- x_{ij}^k is a binary decision variable and is 1 if vehicle k will traverse the route from node i to node j and is 0 otherwise.
- u_i^k decides the service start time at node i by vehicle k.
- l_i^k decides the number of patients in vehicle k after visiting node i.

The MIP formulation is (1)–(19):

$$Min \sum_{k\in K}\sum_{i\in V}\sum_{j\in V} t_{ij}x_{ij}^k \tag{1}$$

$$\sum_{k\in K}\sum_{j\in V} x_{ij}^k = 1 \quad \forall\, (i \in P) \tag{2}$$

$$\sum_{i\in V} x_{0i}^k = \sum_{i\in V} x_{i,2n+1}^k = 1 \quad \forall\, (k \in K) \tag{3}$$

$$\sum_{j\in V} x_{ij}^k - \sum_{j\in V} x_{n+i,j}^k = 0 \quad \forall\, (i \in P, k \in K) \tag{4}$$

$$\sum_{j\in V} x_{ji}^k - \sum_{j\in V} x_{ij}^k = 0 \quad \forall\, (i \in P \cup D, k \in K) \tag{5}$$

Equation (1) is the objective function where we minimize the total travel time of the routes. Constraints (2) ensures each request is served once. Constraints (3) ensures that every vehicle route begins and ends at the depot. Constraints (4) ensures the pickup and delivery nodes of a request are served by the same vehicle. Constraints (5) certifies that the vehicle that enters a node will also depart from that node.

$$u_j^k \geq u_i^k + d_i + t_{ij} - M(1 - x_{ij}^k) \quad \forall\, (i, j \in V, k \in K) \tag{6}$$

$$u_{i+n}^k \geq u_i^k + d_i + t_{ij} - M(1 - x_{i,i+n}^k) \quad \forall\, (i \in P, k \in K) \tag{7}$$

$$u_{2n+1}^k \leq Tmax_k \quad \forall\, (k \in K) \tag{8}$$

$$u_0^k \geq Tmin_k \quad \forall\, (k \in K) \tag{9}$$

$$e_i \leq u_i^k \leq l_i \quad \forall\, (i \in V, k \in K) \tag{10}$$

Constraints (6) and (7) guarantee that service time is consistent among every node, and that the service time at node $i + n$ does not begin until after the service at node i is completed. Constraints (8) and (9) uphold vehicle shift start and end times. Constraints (10) ensures that a request is completed during it's time window.

$$l_j^k \geq l_i^k + q_j - M(1 - x_{ij}^k) \quad \forall\, (i, j \in V, k \in K) \tag{11}$$

$$l_j^k \leq l_i^k + q_j + M(1 - x_{ij}^k) \quad \forall\, (i, j \in V, k \in K) \tag{12}$$

$$l_i^k \leq Q_k \quad \forall\, (i \in V, k \in K) \tag{13}$$

$$l_i^k \leq Q_k + q_i \quad \forall\, (i \in V, k \in K) \tag{14}$$

$$l_i^k \geq 0 \qquad \forall\, (i \in V, k \in K) \tag{15}$$

$$l_i^k \geq q_i \qquad \forall\, (i \in V, k \in K) \tag{16}$$

$$\sum_{i \in K} \sum_{j \in V} (t_{ij} + d_i) x_{ij}^k \leq w b^+ \qquad \forall\, (k \in K) \tag{17}$$

$$\sum_{i \in K} \sum_{j \in V} (t_{ij} + d_i) x_{ij}^k \geq w b^- \qquad \forall\, (k \in K) \tag{18}$$

$$x_{ij}^k \in (0, 1) \qquad \forall\, (i \in V,\ j \in V, k \in K) \tag{19}$$

Constraints (11)–(16) ensures consistency for the passenger load in every vehicle. Inequalities (17) and (18) are the workload balancing constraints. We measure workload as the sum of the travel time and service time along a PTU's route.

4.2 Model and Data Framework

EHS captures data such as facility visited, length of time traveled, time spent with the patient and much more. We use historical patient transfer data to help estimate travel times, service times and generate samples to test the model. This process is illustrated in Fig. 1. The DARP model receives the service and travel time estimates from the patient transfer data in addition to the PTU shift schedules and q_i. The output from this model is the route set (x_{ij}^k), service time start times (u_i^k) and PTU load (l_i^k) for each node.

The patient transfer process can be broken into four stages as shown in Fig. 2. A PTU crew is sent to pick up a patient. They spend t_{ji} minutes travelling to pickup node i. They spend service time d_i at the pickup node. Once the patient is loaded into the PTU, the crew spends $t_{i,i+n}$ minutes travelling to the delivery node where they spend d_{i+n} minutes. When they have transferred care to the receiving facility, the patient transfer is complete. This is a special case where no additional patients are picked up once a patient transfer has begun, and $Q_k = 1$. We focus on this case as the majority of PTUs have a capacity of one. These model parameters are estimated from the patient transfer data. We first look at characteristics of patient transfer demand to determine how we will determine these parameters.

We check patient transfer requests by day of week and by time of day. The day of week analysis found the majority of patient transfer requests occur Monday to Friday. Focusing only on those days, Fig. 3 shows requests by time of day. Requests begin at 6 a.m. We see that requests taper off by 6 p.m.; therefore, that will be the end of the model's time horizon. We also note that advance notice requests are more likely to take place in the morning while same day requests take place in the afternoon.

With these patient transfer trips captured in the data, we use the historical average to estimate the expected time to travel from node i to node j. These estimates are used in the model as parameter t_{ij}.

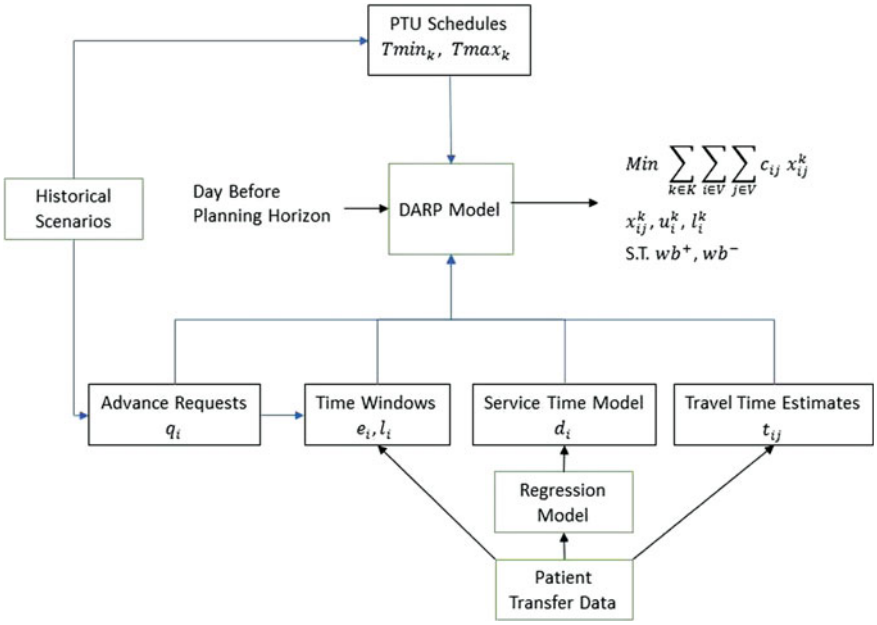


Fig. 1 Big data and OR framework

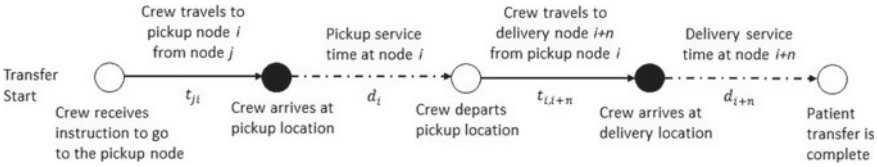


Fig. 2 Time stages in the patient transfer process

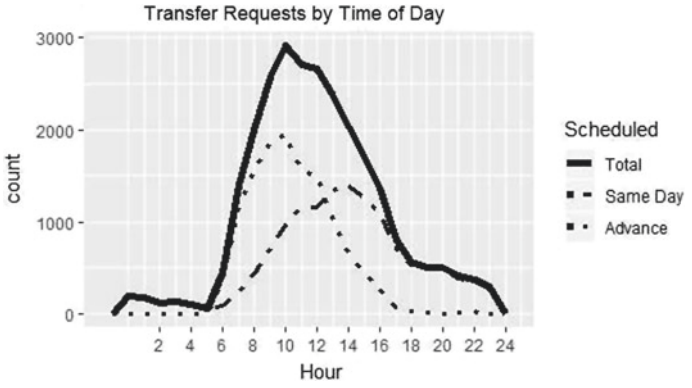


Fig. 3 Patient transfer requests by time of day

In this paper, we define two types of service time. The pickup service time is the time that paramedics spend with the patient at the pickup location before transit begins. Delivery service time is the time spent at the delivery location before completing the transfer. These are shown in Fig. 2 as d_i and d_{i+n} respectively. Service times are captured in the patient transfer data. However, numerous factors can influence the service time, including the facility type, equipment required and whether the patient is ready when the paramedics arrive. Using the patient transfer data, multi-linear regression models were created to estimate the expected pickup and delivery service times for requests. These models use common information that are included in every patient transfer record. The output of the regression model is used as input to the DARP model as d_i. We use day of week, time of day, facility, whether the patient is bariatric, whether special equipment is required and whether the request was advance notice or same day.

5 Results

EHS typically receives 15–45 advance requests per day in the city of Halifax. We test the model on randomly selected actual problems from the historical data. We can use the actual service times or generate predicted service times from the regression model. The computational times are plotted on a logarithmic scale in Fig. 4 shows that computation time (plotted on a logarithmic scale) increases with the number of requests. However, every problem tested solved in under one hour of run time using the Gurobi Optimizer. Since EHS would run the model overnight, we have shown that the model can be solved in a short time frame where the results will be of use to EHS. Table 1 shows a summary of our tests including the maximum, minimum and mean number of requests and time (measured in seconds) required to solve the DARP.

We compare our total travel time against the actual recorded travel time from the data for advance requests. However, we do not have the original schedules with only the advance requests. Instead, we have the final schedules with advance and same day requests. Therefore, we select the 9 days where same day requests make up the

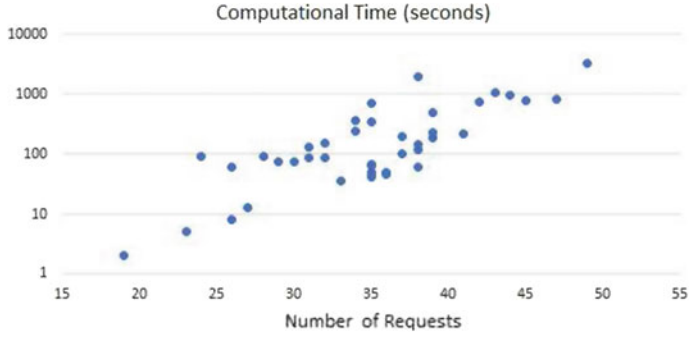


Fig. 4 DARP model computational time

Table 1 Computational results summary

	Requests	Computational time (s)
Max	49	3281
Min	19	2
Mean	33	319

Table 2 Advance requests study

Advance requests	Model travel time	Actual travel time	Actual completed requests	Travel time difference	Same day requests
24	645	747	22	102	8
26	771	833	26	62	8
32	847	1244	32	397	9
34	744	1048	34	304	11
35	768	993	35	225	10
37	890	933	36	43	8
39	920	1211	37	291	10
39	740	1191	38	451	11
43	1289	1271	41	−18	12

smallest percentage of all requests. While this is not a completely fair comparison, focusing on the days with the least impact from the same day requests is as close as we can get. The results are found in Table 2.

For these DARP instances, we see a total reduction in travel time of 1857 min, or 206 min per day. This is a 19.6% improvement on the actual travel time spent on the advance requests. The PTUs also only completed 301 advance requests versus the 309 that were planned for as 8 requests were cancelled. Cancellations can happen for a number of reasons, but it is most likely that this happens when the patient is not ready for transport. The average travel time spent per request as per the model is 24.6 min versus 31.5 min in the data, an improvement of 21.7%.

6 Conclusions and Future Work

In this paper, we studied non-emergency patient transfers and applicable operations research models. We modelled the advance request system using a Dial-A-Ride Problem approach, and developed a model that will schedule and route PTUs for a set of known requests. We also develop statistical models for delivery service times from historical patient transfer data. The historical data is used to inform model parameters as described in Fig. 1. We present this framework as a method to

incorporate Big Data and Analytics into non-emergency patient transfer scheduling. We test our model on historical data and find the model is computationally feasible for problem sizes we are interested in. Finally, we compare our model's output to the actual travel times for days where impact from same day requests is minimal and find travel time improvements of approximately 20%.

Future work involving this problem could include incorporating the same day requests into a model to dynamically update the routes as new demand arrives in real time and investigating the stochastic nature of travel and service times.

Acknowledgements Thank you to NSCERC, NSHRF and MITACS for their funding. Thank you to the Delft University of Technology Optimization Department and Emergency Health Services for their support.

References

1. Berg, P., Essen, T.: Scheduling non-urgent patient transportation while maximizing emergency coverage. Transp. Sci., Accepted (2018)
2. Cordeau, J.F.: A branch-and-cut algorithm for the dial-a-ride problem. Oper. Res. **54**, 573–586 (2006)
3. Cordeau, J.F., Laporte, G.: The dial-a-ride problem: models and algorithms. Ann. Oper. Res. **153**, 29–46 (2007)
4. Detti, P., Papalini, F., Manrique de Lara, G.Z.: A multi-depot dial-a-ride problem with heterogeneous vehicles and compatibility constraints in healthcare. Omega **70**, 1–14 (2016)
5. Guerriero, F., Pezzella, F., Pisacane, O., et al.: Multi-objective optimization in dial-a-ride public transportation. Transp. Res. Procedia **3**, 299–308 (2014)
6. Hans, E., van Houdenhoven, M., Hulshof, P.J. H.: A framework for healthcare planning and control. In: Hall, R. (ed.) Handbook of Healthcare System Scheduling. International Series in Operations Research & Management Science, vol. 168. Springer, Boston (2012)
7. Ho, S., Szeto, W.Y., Kuo, Y., et al.: A survey of dial-a-ride problems: literature review and recent developments. Transp. Res. Part B **111**, 395–421 (2018)
8. Marković, N., Myungseob, K., Schonfeld, P.: Statistical and machine learning approach for planning dial-a-ride systems. Transp. Res. Part A **89**, 41–55 (2016)
9. Matl, P., Hartl, F., Vidal, T.: Workload equity in vehicle routing problems: a survey and analysis. Transp. Sci. (2017). https://doi.org/10.1287/trsc.2017.0744
10. Robinson, V., Goel, V., MacDonald, R.D., Manuel, D.: Inter-facility patient transfers in ontario: do you know what your local ambulance is being used for? Healthc. Policy (2009). https://doi.org/10.12927/hcpol.2009.20478
11. Yalçındağ, S., Matta, A., Şahin, E., Shanthikumar, J.G.: The patient assignment problem in home health care: using a data-driven method to estimate the travel times of care givers. Flex. Serv. Manuf. J. (2016). https://doi.org/10.1007/s10696-015-9222-6

Optimizing Operator's and Users' Objectives in Non-emergency Patients Transportation

Jamal Abdul Nasir, Yong-Hong Kuo and Reynold Cheng

Abstract In this work, we study a Non-emergency Patients Transportation (NEPT) problem in the context of delivering such services in Hong Kong. The purpose of our work is to examine the user inconvenience (waiting time) with the goals of optimizing the vehicle utilization and operating costs. We developed a Mixed Integer Linear Programming (MILP) formulation for the NEPT problem and solved the mathematical model by CPLEX to get optimal results. Using a weighted objective function-based sensitivity analysis, the behaviour of the MILP model is analyzed regarding multiple performance measures, namely operating cost, underutilization level, and user waiting time. This study provides decision makers with the insights into the impacts of objective weights on different performance measures. Our solutions not only reduce the operating costs but also the patient inconvenience. Moreover, our computational experiments based on a case study demonstrate the effective implementation of the model and show the practicality of our methodology.

Keywords Dial-a-ride problem · Health care system · Non-emergency Patients · Transportation · User inconvenience

1 Introduction

Transportation services to reach the hospitals/clinics are an important part of the health care system and essential for the well-being of elderly and disabled citizens.

J. A. Nasir · Y.-H. Kuo (✉)
Department of Industrial and Manufacturing Systems Engineering,
The University of Hong Kong, Pok Fu Lam Road, Hong Kong
e-mail: yhkuo@hku.hk

J. A. Nasir
e-mail: janasir@hku.hk

R. Cheng
Department of Computer Science, The University of Hong Kong, Pok Fu Lam Road, Hong Kong
e-mail: ckcheng@cs.hku.hk

V. Bélanger et al. (eds.), *Health Care Systems Engineering*,
Springer Proceedings in Mathematics & Statistics 316,
https://doi.org/10.1007/978-3-030-39694-7_2

Non-urgent patient transportation needs and associated circumstances in which patients are unable to reach or depart from the health care facilities on their own are defined by Bellamy et al. [1]. In order to transport non-urgent patients between homes and hospitals and among different care units, non-emergency ambulances are usually employed following the respective appointment specifications of the patients. The patients who require transportation submit requests with their specific pickup and delivery addresses and the times ready for pick up. The scheduling and routing plans for non-emergency ambulances to serve these patients are made considering several constraints, e.g., time windows, route length, and ride time. Such planning problems are usually modeled as dial-a-ride problem (DARP) [2] with the aim to minimize the travelling costs and patient inconvenience. The assignment and scheduling decisions for non-emergency patients have been also studied in the context of home health care [3, 4]. From the modeling perspective, DARP resembles Vehicle Routing Problem (VRP) with pickups and deliveries. However, a clear distinction can be made between the DARP and other route planning problems by taking into account the user inconvenience which must be balanced against the operating costs minimization [5].

The problem studied in this paper emerges from the Hong Kong public health care system as it investigates the Non-Emergency Ambulance Transfer Service (NEATS) dedicated to the transportation of non-urgent patients. NEATS works under the Hospital Authority and provides transportation services to three types of patients: (i) elderly patients attending medical appointments, (ii) patients in need of transfer between hospitals, (iii) the patients who seek transportation for home when being discharged from the hospital. Considering the non-emergency nature of these services, requests are usually made in advance of the scheduled appointments. Each ambulance can carry more than one patient, hence group transfers are preferred to avoid low utilization and extra operating costs. Consequently, instead of a first come, first served basis, the trip of a patient is scheduled with other patients heading towards the same locality. Although this strategy improves the ambulance utilization rate and in some case also helpful in reducing travelling costs, the patients' waiting time increases more than expectations. Hence there are increasing number of complaints about the waiting time which directly affects the quality of service. In this work, we examine the patient inconvenience in terms of their waiting time against the vehicle utilization rate and operating costs. An MILP formulation based on the DARP is developed to model the problem. In the perspective of economic efficiency and service quality of the transportation services, the conflicting objectives need to be considered. Hence the objective function of this mathematical model aims to minimize a weighted sum of three components that characterize different performance measures with respect to non-emergency patients transportation service: (1) the total distance traveled cost by all the vehicles and the associated vehicle assignment cost, (2) the total length of waiting time before the arrival of an ambulance to pick the patients and (3) the total underutilized capacity of ambulances with respect to the visited nodes.

The remainder of this article is organized as follows. In Sect. 2, we review the relevant literature. In Sect. 3, we describe the mathematical model for our problem.

Section 4 provides the computational experiments and results. Conclusion and future work are given in Sect. 5.

2 Literature Review

There have been studies on non-urgent patient transportation problem in the context of DARP. A NEPT problem with multi-trip ambulance routes and staff planning was investigated by Lim et al. [6]. Their model formulation contained a hierarchical objective function. A local search-based metaheuristic method was proposed to solve the problem. Zhang et al. [7] proposed a memetic algorithm to solve the NEPT problem. The authors developed a mathematical model that aims to minimize the travelling cost and the number of unserved requests. Van Den Berg and Van Essen [8] introduced a method to optimize the routes for NEPT by utilizing a part of the capacity while the remaining capacity was maximized to handle the emergency patients. Two approaches were suggested to solve their problem. In the first approach, an integer linear programming formulation was solved with a time limit, whereas the second approach presented an alternative formulation with discretized time to obtain efficient solutions. Similarly, Kergosien et al. [9] considered both emergency requests and NEPT requests with respect to dynamic arrivals throughout the day. The authors used a discrete event simulation based tool to analyze the performance of the proposed management approaches regarding the ambulance fleets. Three management approaches were investigated in their study: (i) independent management, (ii) reactive integrated fleet management, and (iii) proactive integrated fleet management. It was observed that the proactive integrated fleet management approach provided more promising results and was easy to implement in real-life situations.

Ritzinger et al. [10] studied a DARP in the context of transporting elderly and handicapped people where user inconvenience was taken into account. The authors used Dynamic Programming (DP) based exact and heuristic algorithms to provide solutions for their problem. Taking into account the stochastic information about the future return transports, Schilde et al. [11] investigated the DARP in the context of daily operations of the Austrian Red Cross. Four different types of metaheuristic solution approaches were used to find the solutions. It was observed that their look-ahead approach regarding the future patient transportation assists to improve the solution quality if the number of patient requests who require the return transport service are low compared to overall transportation requests. Parragh et al. [12] developed a two-phase heuristic solution procedure for a multi-objective DARP which aimed to find the efficient ambulance dispatch plans for the Austrian Red Cross. The proposed mathematical formulation included patient inconvenience and travelling cost in the objective function. Path relinking module was used to obtain Pareto optimal solutions, where initial solutions were obtained by a variable neighborhood search-based method. A DARP with heterogeneous vehicles and users was studied by Parragh [13]. The author considered different usage requirement modes (stretcher, wheel chair and seats) and proposed two mathematical formulations. The proposed

2-index formulation was found better than the 3-index formulation considering a branch and cut-based solution framework.

Distinguishing between the static and dynamic natures of transportation requests arrivals, Cordeau and Laporte [5] identified two distinct modes for DARPs. The static case only serves the transportation requests received ahead of the certain time limit, whereas dynamic case considers requests throughout the day and updates the scheduling plan accordingly. In order to serve the dynamic patient transportation requests between the care units, Kergosien et al. [14] proposed a Tabu Search (TS) based heuristic algorithm to obatian solutions. The capacity of each vehicle was limited to one patient only and subcontracting can be done to compensate the capacity shortage. Three different types of transportation requests were considered in their work. The suggested solution method stores the routes through adaptive memory and then the initial solutions are improved by iteratively running the TS algorithm. The main contribution of our current work is to analyze different performance measures in the context of NEPT by giving preference to these measures through a weighted objective function. This study conducts a sensitivity analysis to observe the potential tradeoff between the operating costs and waiting time considering the NEPT services in Hong Kong.

3 Problem Description

Each patient request is characterized by a pickup node and a corresponding delivery node. Let $P = \{1, ..., n\}$ be the set of pickup nodes and $D = \{n + 1, ..., 2n\}$ be the set of patient delivery nodes, where $n + j$ is the delivery node of pickup node j. The route of each non-emergency ambulance starts from the origin node 0 and terminates at the destination node $2n + 1$. Let $Nr = P \cup D$ be the set of all the locations to be visited by an ambulance after leaving the depot. Similarly, $N^o = \{0\} \cup Nr$ represent the set of origin node and all the pickup and delivery nodes, whereas $N^d = Nr \cup \{2n + 1\}$ denote the set of all the pickup and delivery nodes and the destination node. The NEPT problem is defined on a graph $G = (N, A)$, where $N = \{0\} \cup P \cup D \cup \{2n + 1\}$ is the set of all nodes and $A = \{(j, k) \mid j \in N, k \in N\}$ is the set of arcs. α_1, α_2 and α_3 are the assigned weights for respective performance measures. Table 1 gives a summary of related parameters and decision variables used in this mathematical model.

3.1 Mathematical Model

$$min\ \alpha_1 \left(b \sum_{i \in V} \sum_{(j,k) \in A} d_{jk}\, x_{ijk} + c \sum_{i \in V} z_i \right) + \alpha_2 \left(g \sum_{i \in V} \sum_{j \in Nr} u_{ij} \right) + \alpha_3 \left(f \sum_{i \in V} \sum_{j \in P} w_{ij} \right) \tag{1}$$

Table 1 Notations

Notation	Definition
Indices	
P	Set of patient pickup nodes
V	Set of non-emergency ambulances
N	Set of all nodes
Parameters	
Gc	Capacity of each ambulance $i \in V$
L	Maximum route length for each ambulance $i \in V$
M	A very large number
T	Service time duration for each patient pickup/delivery service
t_{jk}	Time required to travel from node j to k
d_{jk}	Travelling distance between two locations j and k
pi_j	The number of patients to be picked at pickup node $j \in P$
dr_j	The number of patients to be dropped at delivery node $j \in D$
Es_j	Time ready for pickup at pickup node $j \in P$
b	Traveling cost incurred per unit of distance
c	Cost for route assignment to an ambulance
f	Penalty for waiting per minute at pickup nodes
g	Penalty for underutilization of ambulance at each visited node
Decision variables	
x_{ijk}	1 if an ambulance $i \in V$ travels from node j to k, 0 otherwise
z_i	1 if an ambulance $i \in V$ has been assigned a route, 0 otherwise
q_{ij}	Service start time of ambulance $i \in V$ at node $j \in N$
VL_{ij}	Vehicle load for ambulance $i \in V$ at node $j \in N$
w_{ij}	Time waiting for ambulance $i \in V$ at pickup node $j \in P$
u_{ij}	Underutilization of ambulance $i \in V$ at node $j \in N$

Subject to:

$$\sum_{i \in V} \sum_{k \in Nr(k \neq j)} x_{ijk} = 1 \qquad \forall j \in P \tag{2}$$

$$\sum_{k \in N(k \neq j)} x_{ijk} - \sum_{k \in N(k \neq j)} x_{i,j+n,k} = 0 \qquad \forall j \in P, \quad \forall i \in V \tag{3}$$

$$\sum_{k \in N^d} x_{i0k} = z_i \qquad \forall i \in V \tag{4}$$

$$\sum_{j \in N^o} x_{ij(2n+1)} = z_i \qquad \forall i \in V \tag{5}$$

$$\sum_{j \in N^o (j \neq h)} x_{ijh} - \sum_{k \in N^d (k \neq h)} x_{ihk} = 0 \qquad \forall h \in Nr, \quad \forall i \in V \tag{6}$$

$$q_{ij} + \left(T + t_{jk}\right) x_{ijk} \leq q_{ik} + M \left(1 - x_{ijk}\right) \qquad \forall i \in V, \quad \forall (j,k) \in A \tag{7}$$

$$q_{ij} + \sum_{k \in N (k \neq j)} \left(T + t_{jk}\right) x_{ijk} \leq q_{i,j+n} \qquad \forall i \in V, \quad \forall j \in P \tag{8}$$

$$Es_j \sum_{(j,k) \in A} x_{ijk} \leq q_{ij} \qquad \forall j \in P, \quad \forall i \in V \tag{9}$$

$$Es_j \sum_{(j,k) \in A} x_{ijk} \geq q_{ij} - w_{ij} \qquad \forall j \in P, \quad \forall i \in V \tag{10}$$

$$VL_{ij} + (pi_k - dr_k) x_{ijk} \leq VL_{ik} + M \left(1 - x_{ijk}\right) \qquad \forall j,k \in N (j \neq k), \quad \forall i \in V \tag{11}$$

$$VL_{ij} + u_{ij} = Gc \sum_{k \in Nr (k \neq j)} x_{ijk} \qquad \forall j \in Nr, \quad \forall i \in V \tag{12}$$

$$\sum_{(j,k) \in A} d_{jk} \, x_{ijk} \leq L \qquad \forall i \in V \tag{13}$$

$$x_{ijk}, \; z_i \in \{0, 1\} \qquad \forall i \in V, \quad (j,k) \in A \tag{14}$$

$$q_{ij}, \; VL_{ij}, \; w_{ij}, \; u_{ij} \geq 0 \qquad \forall i \in V, \quad j \in N \tag{15}$$

The objective function (1) seeks to minimize a weighted sum of the operating cost (traveling cost and route assignment cost), underutilization penalty cost and penalty cost for patients waiting times. Constraints (2) ensure that all the requests are served. Constraints (3) guarantee that the pickup node and the corresponding delivery node are visited by the same ambulance for each specific patient request. Constraints (4) and (5) assure that the ambulance route will start at the origin node and terminate at the destination node respectively. Constraints (6) represent the flow conservation for ambulances with respect to nodes. Constraints (7) determine the ambulance arrival time at a pickup/delivery node. Constraints (8) implement the precedence and ensure that a delivery node is visited only after visiting the associated pickup node. Constraints (9) assure that the ambulance arrives at a pickup node after the start of ready-for-pickup time specified by a patient. Constraints (10) calculate the waiting time observed at each patient node. This waiting time is the difference between the ambulance arrival time and the ready-for-pickup time specified by the patient. Constraints (11) determine the vehicle load at each node visited by the

ambulance. Constraints (12) guarantee that the vehicle load will not exceed the ambulance capacity and they also define the underutilization level of an ambulance at each visited node before reaching the last delivery point. Constraints (13) ensure that the ambulances respect the route length limit. Constraints (14) and (15) show the binary and positive decision variables respectively.

4 Computational Experiments

We performed computational experiments to assess the performance measures by varying respective weightings in the objective function. The mathematical formulation was implemented in OPL and solved by CPLEX. The computational tests were performed on a computer with 3.40 GHz Intel Core i5 CPU.

The mathematical model developed in this study is applied to instances based on a case of Hong Kong. We consider sixty different locations as potential pickup and delivery points in the Central and Western district of Hong Kong. These locations include patient homes and public hospitals/clinics in this district. These locations are marked on the map in Fig. 1. The real distance (d_{jk}) and travel time (t_{jk}) between each pair of nodes is calculated through the Google Maps API. On the basis of this data, three different sets of problem instances are created which contain ten problem instances in total. CPLEX could only deliver the results for the instances with 10 pickups and 10 drop-off locations. Therefore, the maximum size of the test instances is limited to 10 pickup locations in this study. Each set contains problem instances with different sizes. The size of the problem instances remains same inside a set, whereas the size of the instances increases as we move from set A to set C. For instance, set A contains five problem instances and each problem instance in this set has same size (5 pickups and 5 drop-off locations), however, the geographical addresses (coordinates) for pickup and drop off locations are different for each instance. Table 2 presents the related characteristics and input parameters values used for all the instance sets in

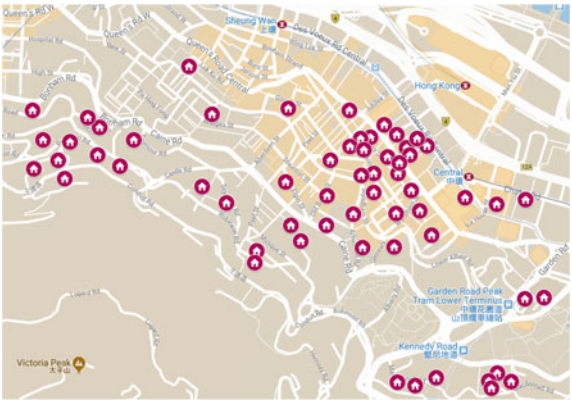


Fig. 1 Pickup and delivery locations for patients

Table 2 Instance characteristics

Inst.	# Pickup nodes	# Delivery nodes	# Ambulances	Capacity	Route length
A	5	5	3	7	50
B	7	7	3	7	50
C	10	10	4	7	50

terms of pickup nodes (P), delivery nodes (D), number of ambulances (V), capacity (Gc) and route length (L). The ready-for-pickup times are distributed randomly over a time period of seven hours. The cost values (b, c) are chosen based on the context of Hong Kong and these values are 4 Hong Kong Dollars (hkd) per unit of distance and 500 hkd per ambulance respectively. The penalty cost for underutilization (g) is four times less than the waiting penalty (f), as the underutilization will occur on almost every node and the respective values will always be large at the nodes earlier in the route. For the initial experiments equal weights are assigned to all the performance measures in the objective function ($\alpha_1 = \alpha_2 = \alpha_3 = 1$), though the varied weights are used in the sensitivity analysis to observe the tradeoff between these performance measures.

4.1 Results

Table 3 shows the results for all the instances. The run time is expressed in seconds and objective value is stated in terms of hkd. The results obtained by CPLEX show

Table 3 Results for the MILP model solution

Inst. type	CPU (s)	Total cost	Travelling cost	Ambulance route cost	Waiting time cost	Under utilization cost
A1	45	630.40	52.40	500	40	38
A2	63	584.00	10.00	500	36	38
A3	29	610.40	24.40	500	48	38
A4	80	636.40	26.40	500	72	38
A5	18	593.60	19.60	500	36	38
B1	2500	666.40	46.40	500	76	44
B2	2345	678.75	58.75	500	76	44
B3	2763	672.20	41.20	500	87	44
C1	5800	675.35	65.35	500	64	45
C2	4750	685.41	73.41	500	55	57

the successful implementation of the MILP model as the scheduling and routing decisions respect the associated constraints while minimizing the total cost. Clearly, the run time for each instance is highly dependent on the problem size and increases significantly with the increasing size of the problem as shown in the 2nd column. The total cost values increase from first to last set owing to the increasing size of the problem. The penalty cost values for user inconvenience are large in some cases (e.g., waiting time cost (B3) = 87). Such results are observed due to the equal preference given to all performance measures as far as respective weights are concerned ($\alpha_1 = \alpha_2 = \alpha_3 = 1$).

4.2 Sensitivity Analysis

Preference among the performance measures of the objective function is expressible by setting significantly different values for the respective weights of these measures. In order to observe the behavior of the MILP model against the variation in the values of assigned weights to performance measures, we performed a sensitivity analysis by varying the weights of first two measures (α_1, α_2) between 0 and 1, whereas α_3 is kept constant. For the purpose of this analysis, one instance from set A is selected and solved through CPLEX. This analysis aims to examine the impact of operating cost and underutilization level on the patients' waiting time. The weights for α_1 and α_2 are simultaneously changed in descending and ascending order respectively ($\alpha_1 = 1$ to 0.1, $\alpha_2 = 0.1$ to 1), whereas α_3 is kept at 1. It was noted that the waiting time shows considerable change when the value of α_1 is changed. The lowest value for waiting time is achieved when $\alpha_1 = 0.1$ and $\alpha_2 = 1.0$. The waiting time costs, operating costs

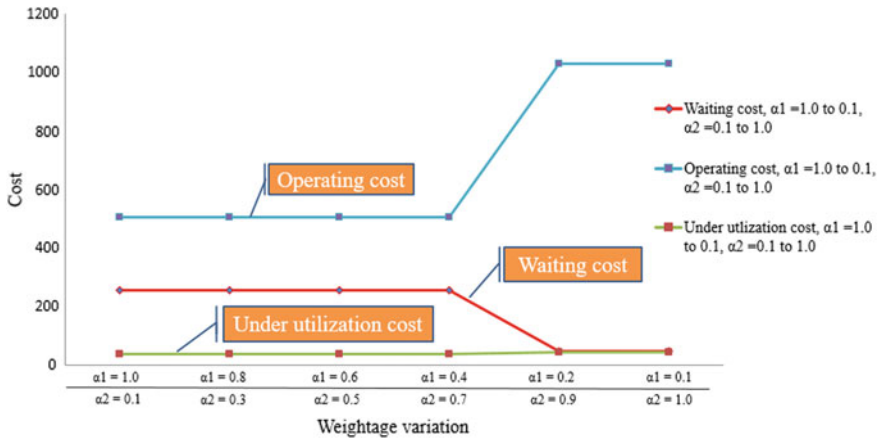


Fig. 2 Waiting costs, operating costs and underutilization costs against the considered scenario

and underutilization costs against different weights are shown in Fig. 2. Considering the results, clearly a trade-off exists between the operating cost and the waiting cost.

The results of sensitivity analysis indicate that the waiting time is more sensitive to the lower values of operating costs related to route assignment cost. It was observed that the waiting time changes significantly when very low weights are assigned to α_1 as shown in Fig. 2. Hence, the decision makers can reduce the user inconvenience by making a proper trade-off between operating costs and waiting time penalty.

5 Conclusion

This article studies the transportation services for non-emergency patients considering the NEPT services being provided in Hong Kong. This problem is modeled as a DARP with weighted objective function featuring three different performance measures. The problem takes into account the constraints related to the earliest start time, route length for ambulances and capacity limit. The results obtained by CPLEX indicate the successful implementation of the MILP model to perform the scheduling decisions. Moreover, a sensitivity analysis is performed to examine the behavior of the mathematical model with respect to the different performance measures. It was noted that the waiting time is reduced when a low preference is given to the operating costs. In the future work, these performance measures can be further studied by incorporating dynamic arrivals of patient requests and integration between emergency and non-emergency fleets of ambulances. Moreover, efficient heuristic methods can be developed to handle larger problem instances.

Acknowledgements This research is partially supported by Health and Medical Research Fund (HMRF) 14151771 from Food and Health Bureau, the Hong Kong SAR Government.

References

1. Bellamy, G., Stone, K., Richardson, S.K., Goldsteen, R.: Getting from here to there: Evaluating west Virginia's rural nonemergency medical transportation program. J. Rural Health (Off. J. Am. Rural Health Assoc. Nat. Rural Health Care Assoc. **19**, 397–406 (2003)
2. Ho, S.C., Szeto, W.Y., Kuo, Y.-H., Leung, J.M.Y., Petering, M., Tou, T.W.H.: Literature review and recent developments: a survey of dial-a-ride problems. Transp. Res. Part B Methodol. **111**, 395–421 (2018)
3. 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. Part A: Syst. Hum. **42**, 1346–1363 (2012)
4. Yalçındağ, S., Matta, A., Şahin, E., George Shanthikumar, J.: A two-stage approach for solving assignment and routing problems in home health care services. In: Proceedings of the International Conference on Health Care Systems Engineering, pp. 47–59. Springer International Publishing (2014)

5. Cordeau, J.-F., Laporte, G.: The dial-a-ride problem: models and algorithms. Ann. Oper. Res. **153**(1), 29–46 (2007)
6. Lim, A., Zhang, Z., Qin, H.: Pickup and delivery service with manpower planning in hong kong public hospitals. Transp. Sci. **51**, 688–705 (2017)
7. Zhang, Z., Liu, M., Lim, A.: A memetic algorithm for the patient transportation problem. Omega **54**, 60–71 (2015)
8. Van Den Berg, P.L., Van Essen, J.T.: Scheduling non-urgent patient transportation while maximizing emergency coverage. Transp. Sci., 1–18 (2019)
9. Kergosien, Y., Gendreau, M., Ruiz, A., Soriano, P.: Managing a fleet of ambulances to respond to emergency and transfer patient transportation demands. In: Proceedings of the International Conference on Health Care Systems Engineering, pp. 303–315 (2014)
10. Ritzinger, U., Puchinger, J., Hartl, R.F.: Dynamic programming based metaheuristics for the dial-a-ride problem. Ann. Oper. Res. **236**, 341–358 (2016)
11. Schilde, M., Doerner, K.F., Hartl, R.F.: Metaheuristics for the dynamic stochastic dial-a-ride problem with expected return transports. Comput. Oper. Res. **38**(12), 1719–1730 (2011)
12. Parragh, S.N., Doerner, K.F., Hartl, R.F., Gandibleux, X.: A heuristic two-phase solution approach for the multi-objective dial-a-ride problem. Networks **54**(4), 227–242 (2009)
13. Parragh, S.N.: Introducing heterogeneous users and vehicles into models and algorithms for the dial-a-ride problem. Transp. Res. Part C Emerg. Technol. **19**(5), 912–930 (2011)
14. Kergosien, Y., Lent, Ch., Piton, D., Billaut, J.-C.: A tabu search heuristic for the dynamic transportation of patients between care units. Eur. J. Oper. Res. **214**(2), 442–452 (2011)

Modelling Hospital Medical Wards to Address Patient Complexity: A Case-Based Simulation-Optimization Approach

Paolo Landa, Micaela La Regina, Elena Tànfani, Francesco Orlandini, Mauro Campanini, Andrea Fontanella, Dario Manfellotto and Angela Testi

Abstract In this paper we focus on patient flows inside Internal Medicine Departments, with the aim of supporting new organizational models taking into account the patient relevant characteristics such as complexity and frailty. The main contribution of this paper is to develop a Discrete Event Simulation model to describe in detail the pathways of complex patients through medical hospital wards. The model has been applied to reproduce a case study of an Italian middle size hospital. The objective is quantifying the impact on resource use and outcome of introducing a new organizational model for medical departments. The re-organization is mainly focused on changing the available beds assignment among the wards to better address the complexity of care of patients with comorbidities. Following a patient-centered

P. Landa (✉)
Department of Economics, Università Degli Studi Di Genova, Via Vivaldi 5, 16126 Genova, Italy
e-mail: paolo.landa@unige.it

M. La Regina
S.S. Risk Management, ASL 5 Spezzino, Via Fazio 30, 19121 La Spezia, Italy
e-mail: micaela.laregina@asl5.liguria.it

E. Tànfani · A. Testi
Department of Economics, University of Genova, Via Vivaldi 5, 16126 Genova, Italy
e-mail: etanfani@economia.unige.it

A. Testi
e-mail: testi@economia.unige.it

F. Orlandini
Healthcare Director, ASL 4 Chiavarese, Via Gio Batta Ghio 9, 16034 Chiavari, Italy
e-mail: francesco.orlandini@asl4.liguria.it

M. Campanini
Department of Internal Medicine, Maggiore della Carità, Novara, Italy
e-mail: mauro.campanini@maggioreosp.novara.i

A. Fontanella
Department of Internal Medicine, Ospedale del Buonconsiglio—Fatebenefratelli, Napoli, Italy

D. Manfellotto
Department of Internal Medicine, Ospedale Fatebenefratelli Isola Tiberina, Roma, Italy
e-mail: dario.manfellotto@afar.it

V. Bélanger et al. (eds.), *Health Care Systems Engineering*,
Springer Proceedings in Mathematics & Statistics 316,
https://doi.org/10.1007/978-3-030-39694-7_3

approach, patients are segmented considering the clinical characteristics (i.e. the pathology, proxy of Diagnoses Related Groups classification) and sub-grouped considering other characteristics, such as comorbidities and ward of admission. Then, an optimization component embedded into the model chooses the best pooling strategy to reorganize medical wards, determining the corresponding number of beds able to improve process indicators, such as length of stay. The simulation model is presented, and preliminary results are analyzed and discussed.

Keyword Simulation-optimization · Internal Medicine Ward organization · Clinical pathway · Hospitalist-based model · Data segmentation

1 Introduction and Problem Addressed

In the last few years with the fast progress of medical knowledge, the education of doctors has evolved towards greater specialization. Within the medical area, many sub-specializations, such as cardiology, pulmonology, gastroenterology, geriatrics, etc., gemmated from Internal Medicine [2]. The need to investigate each medical condition has led, from an organizational point of view to the birth of different medical wards, each corresponding to a specific specialization [12]. Consequently, patients are today admitted to different wards depending on the prevalent clinical problem that led to the need for the hospital admission.

The problem arises from the fact that, to the greater specialization of medical knowledge, an evolution of the patient's conditions in the opposite sense is observed. The presence of multiple-pathologies and social frailty represent the epidemic of the third millennium, and they are mining the sustainability of national and worldwide health systems [17]. This problem affects mostly patients admitted in hospital that have an age over 65 year old, with an average of 2.7 chronic diseases, requiring medical care for an acute transient condition, i.e. an infection, that triggers a decompensation of chronic condition or acute decompensated heart failure, and/or a complication such as diabetes onset [8]. The clinical complexity is increased by functional and cognitive decline, adverse events given by the use of multiple drugs, socioeconomic deprivation and poor familiar support. These patients, often called frail, require urgent organizational changes to address their health needs appropriately [5].

The first change to be addressed concerns the professional education of medical specialists who should regain their main characteristics, being doctors of complexity capable of treating the patient following a holistic approach. The appropriate professional figure, already introduced 20 years ago in the US, could be the "Hospitalist", a medical specialist, more often a specialist in Internal Medicine hat should have the clinical, organizational and relational skills needed to the integrated care of complex patients with multi-pathologies [18]. The introduction of this new figure in a specialty-based hospital, however, is not sufficient to meet patient requirements, even if it seems to produce performance gain as literature proves [16, 19] but, in our opinion, it is not enough.

A second change is essential to take full advantage of this new professional figure, i.e. the reorganization of medical wards from specialty-based care to a patient-centered one. This change requires a cultural shift and a complete re-thinking of medical Departments, or even the whole hospital, where the divisions among sub-specialties should disappear. This does not mean, of course, that specialized cardiologists, pulmonologists, geriatricians and other specialized clinicians should disappear, but that they should not be assigned a specific ward. Instead, they should work in multidisciplinary teams coordinated by the global approach of the hospitalist. Some specialized units should remain for particularly severe intensive care such as the ICU for cardiac disease. This reconfiguration is the only one able to face the needs of new patients in the most appropriate clinical way as recent studies show it is a reconfiguration based on the patient and not on the hospital supply [4, 11]. However, before the introduction of organizational innovation, an evaluation of the expected impact should be carried on.

Whether this patient-centered reconfiguration also brings some advantages in terms of resource use and outcome to the traditional specialty-based one, is the specific aim of this work. The resource use is a proxy of the number of ward beds needed and costs for laboratory and diagnostics, while the outcome is assessed by means of the average length of stay. The main contribution of this paper is to develop a Discrete Event Simulation (DES) model to firstly reproduce the traditional (specialty-based) organization of a real case study and to evaluate the impact on resource utilization (beds and costs) and outcome (average length of stay) of re-organizing the stay areas using a patient-centered model. In the patient-centered model the specialist wards are merged into a unique Internal Medicine Ward (IMW) to better address the complexity of care of patients. Besides, the optimization component embedded in the DES model is used to determine the optimal (minimum) number of beds necessary to manage the overall cohort of patients flowing in the hospital IMWs following the patient-centered organizational model.

The paper is organized as follows: In Sect. 2, the study motivation is presented together with a brief description of the organizational models to be tested. Section 3 reports the case study, data collection, and analysis. In Sect. 4 the simulation model development is introduced, and some details of the methodology and assumptions are reported. The results given by the simulation-optimization for the case study are analyzed in Sect. 5. Finally, in Sect. 6, conclusions and future direction of the research are reported.

2 Study Motivation

This study began from a collaboration with a group of internists involved in an advanced master level course titled “Hospitalist: managing complexity in Internal Medicine inpatients”. The aim of the course was forming these internists as Hospitalist, for the Italian hospital sector. As reported in Sect. 1, Literature shows that the introduction of hospitalists in IMWs could result in reduced costs, shortened

lengths of stay, preserved or even enhanced the quality of care and patient satisfaction, in essence improving the "value of care" [16, 19]. However, the introduction of this figure poses additional issues on how healthcare services should be organized around acute multi-pathology patients. At least, to the authors' knowledge and experience, no studies are dealing with the evaluation of the re-organization of the stay area connected to the introduction of this new figure.

The organizational models herein compared are referred to as the specialty-based and the patient-centred model, respectively. The first reproduces the current practice where patients are admitted in a ward following the main acute clinical problem. Specialty-based hospitals cannot assure global and efficient care for multi-pathology and frailty of patients [15]. Their hospital stay will likely be fragmented in more, isolated episodes of care with transfers from the emergency department to other wards (e.g. infectious disease, cardiologic and metabolic wards). Movements among wards are uncomfortable and risky for patients. Transitions of care are invariably associated with loss of clinical information, duplication of tests, unintentional pharmacological discrepancies and much more. In the re-organization that follows the patient-centered model, the patient is admitted in a unique IMW where the hospitalist organizes and takes in charge the patient hospital stay managing a multidisciplinary medical team and assuring a holistic vision of the care.

Thanks to the collaboration of the clinicians involved in this study, we had the opportunity of collecting a large amount of clinical historical data of patients admitted in hospital with a diagnosis among the most prevalent in the Internal Medicine area. The inclusion criteria and the resulting cohort of patients analyzed are reported in Sect. 3. The clinical pathways of all patients with the same health problem, age, comorbidity conditions, severity of illness are analyzed with a focus on the differences in terms of resource use and outcome depending only on the organizational model: specialty-based or patient-centered. Starting with the data collected, a discrete event simulation model evaluates the benefits of introducing a patient-centered reconfiguration of the stay area in terms of resource use and outcome.

3 Case Study: Data Collection and Analysis

The case study herein reported refers to a Ligurian Local Health Authority (ASL5) sited in La Spezia province (Italy). ASL5 is one of the Local Health Authorities of Liguria Region. It provides, directly or through accredited public and private subjects, the following services: (i) services provided on the Essential Health Care Levels (LEA) in the form of district assistance and hospital care health services, (ii) high social and health integrated assistance, and (iii) emergency health services. It provides health services to 217,507 inhabitants (which 27.4% is over 65 years old). About 8500 inhabitants are frail and at risk of disability, while 8300 have a disability.

Administrative data coming from the Hospital Discharge Episodes Database (HDED) and medical data coming from Electronic Patient Record (EPR) collected from January 2016 to December 2016 were analyzed. The Hospital Discharge Report

Table 1 Number of patients admitted for each DRG and ward (year 2016)

	DRG					
Ward	087	089	090	127	576	*Total*
Cardiology	71	1	2	140	5	*219*
Geriatrics	64	83	14	155	61	*377*
Infectious diseases	–	21	15	1	97	*134*
General Medicine 1	199	52	36	129	109	*525*
General Medicine 2	566	37	16	163	351	*1133*
Respiratory Medicine	266	40	27	3	3	*339*
Total	1166	234	110	591	626	*2727*

includes administrative data, as well as the date of admission and discharge, the transfers of the patient between wards, the diagnosis, and the DRG assigned. Data from EPR include all the tests and consultations (blood transfusion, specialist visits, diagnostic tests, laboratory tests, and other tests) performed to the patient during the hospital stay. The cost of these specialist and diagnostics services were provided by the Italian National Health System official tariff list. Other data were collected by the Hospital management accounting service.

The analysis is focused on the six medical wards reported in Table1, two of them (General Medicine 1 and General Medicine 2) are generic, and the other four are specialist wards. With reference to the pathologies to be included, as suggested by the hospital physicians involved in our study, the analysis focused on five Diagnosis Related Groups (DRGs) covering on average 70% of the total cases (DRG 087: Pulmonary edema and respiratory failure, DRG 089: Pneumonia and pleuritis with complications, DRG 090: Pneumonia and pleuritis >17 year old, DRG 127: Heart failure and shock, DRG 576: Sepsis without medical ventilation).

All DRGs are treated within each of the six wards. The total number of patients admitted by each ward depends on the different ward capacity in terms of resource, but they are not distributed exclusively following the prevalent condition. For instance, specialist wards, as cardiology and infectious diseases, admit patients with heart failure and sepsis, respectively, but also with respiratory problems.

As a consequence, patients with heart failure are almost equally distributed among cardiology, geriatrics and general medicine wards, while patients with pulmonary edema and respiratory failure are mostly managed by respiratory medicine and general medicine wards. This situation however, engenders different organizational processes leading to a different length of stay and an average cost of treatment for each patient at the parity of DRG, as pointed in Tables 2 and 3. For instance, the same condition Heart failure has a LOS ranging from 5.9 in the specialist ward Cardiology to 9.4 in General Medicine 2. There is a large variability also across wards about all wards: sometimes the difference is due to the specific treatment—this seems to be the case for the Infectious disease ward. However, in other cases, differences appear

Table 2 Average length of stay (in days) for each DRG and ward of admission

	DRG					
Ward	087	089	090	127	576	*Average*
Cardiology	7.3	5.0	5.5	5.9	16.0	*6.6*
Geriatrics	8.7	9.5	8.2	8.9	9.5	*9.1*
Infection and Immunology	–	11.6	6.9	7.0	15.7	*14.0*
General Medicine 1	10.2	11.3	8.2	9.4	12.9	*10.5*
General Medicine 2	7.4	6.7	5.6	6.5	7.8	*7.3*
Respiratory Medicine	9.6	7.9	6.3	6.7	2.7	*9.0*
Average	*8.4*	*9.3*	*7.1*	*7.6*	*10.1*	*8.7*

Table 3 Average cost per patient (in Euro, €) for each DRG and ward of admission

	DRG					
Ward	087	089	090	127	576	*Average*
Cardiology	2793.40	1948.60	2131.50	2273.30	6013.40	*2524.60*
Geriatrics	2816.70	3087.60	2651.10	2894.20	3085.40	*2945.50*
Infection and Immunology	–	5805.60	3515.80	3679.60	7824.30	*6994.70*
General Medicine 1	3047.80	3348.10	2426.70	2850.50	3819.00	*3146.60*
General Medicine 2	2278.40	2053.20	1727.20	2030.30	2388.90	*2261.80*
Respiratory Medicine	4593.20	3785.50	3052.30	3212.40	1368.30	*4334.40*
Average	*2998.70*	*3340.30*	*2650.20*	*2502.30*	*3572.10*	*3038.00*

to be unjustified: for instance, General Medicine 1 has a larger average LOS for all the DRGs, while General Medicine 2 has on average three days less.

Large variability is also observed with regards to the average cost for each DRG (see Table 3). The average cost for each ward is given by the sum of different items: average utilization of diagnostics and laboratory and the average daily cost times the number of days.

The variability of the average cost depends, of course, by the clinical pathway (DRGs) requiring different bundle of services (diagnostics and so on), for instance in the case of Respiratory Medicine and Infectious Diseases. However, in other cases, as between General Medicine 1 and General Medicine 2, for the same DRG, the detected lower LOS seems to be justified by a different organizational model able to achieve larger productivity of the given beds and resources.

The comparison between the different organizational models for the same DRG, however, is correct only if patient complexity for each DRG is similar among the different wards. The analysis of the demographic and clinical data summarized in Tables 4 and 5 show large variability among the complexity of patients addressing different wards. Complexity is assessed by three characteristics drawn from administrative data (HDED): (i) demographic characteristics (age, sex); (ii) comorbidity

Table 4 Demographic and clinical characteristics: number of patients for age class, sex and Charlson Comorbidity Index

	Age			Sex		Charlson Comorbidity Index			
Ward	≤65	>65 and ≤80	>80	Male	Female	0	1–2	3–4	>5
Cardiology	22	65	132	109	110	161	48	9	1
Geriatrics	–	47	330	113	264	123	192	51	11
Infection and Immunology	50	59	25	72	62	90	27	11	6
General Medicine 1	52	151	322	267	258	222	249	35	19
General Medicine 2	69	258	806	513	620	417	592	104	20
Respiratory Medicine	88	147	104	188	151	127	195	8	9
Total	*281*	*727*	*1719*	*1262*	*1465*	*1140*	*1303*	*218*	*66*

Table 5 Severity conditions and mortality risk: number of patients for each ward and APR code

	APR severity class				APR mortality risk				
Ward	1	2	3	4	1	2	3	4	*Total*
Cardiology	99	115	5	0	42	146	31	0	*219*
Geriatrics	63	212	94	0	51	186	120	20	*377*
Infection and Immunology	33	52	42	8	54	34	30	16	*134*
General Medicine 1	114	350	55	7	115	247	147	16	*525*
General Medicine 2	101	642	357	6	103	396	536	98	*1133*
Respiratory Medicine	33	260	46	33	103	158	78	0	*339*
Total	*443*	*1631*	*599*	*54*	*468*	*1167*	*942*	*150*	*2727*

status (measured by the Charlson Comorbidity Index; (iii) severity condition and mortality risk (APR-DRG classes).

The Charlson Comorbidity Index (CCI) was computed following the specific criteria reported in Deyo et al. [7]. The CCI is a method of categorizing comorbidities of patients based on the International Classification of Diseases (ICD) diagnosis codes reported in administrative data, such as electronic patient records. Seventeen comorbidity categories are included with associated weight (from 1 to 6), based on the adjusted risk of mortality or resource use, and the sum of all the weights provides a final comorbidity score for the patient. A score of zero indicates no comorbidities. The higher is the score, the more likely the predicted outcome will result in mortality or higher resource use. In this study, we use four classes of comorbidity with score values of 0, 1–2, 3–4 and more than 5 respectively.

The Patient Refined Diagnosis Related Group (APR-DRG), is an inpatient classification system that assigns a Diagnostic Related Group value, a Risk of Mortality subclass and a Severity of Illness subclass ranging from 1 to 4 in ascendant order of risk and severity [13]. Regarding the Clinical pathways, we mean the main disease condition causing hospitalization (proxy of DRG, coded using ICD9-CM v.24). In Table 4 for each ward are reported the demographic characteristics and the comorbidity status while in Table 5 the severity conditions, i.e. severity class and mortality risk.

General Medicine wards have the largest quantity of patients covering about 70% of the overall sample, while the smallest units in terms of patient treated are Cardiology and Infection diseases wards. More than half (63%) of overall patients are older than 80 years old, while the patients between 65 and 80 years old and the patients with less than 65 years old represent 27% and 10% of the cohort, respectively. Most of the patients have a CCI of 1–2 (48%) and 0 (42%). Patients with a CCI of 3–4 and larger than 5, are 8% and 2%, respectively A larger quantity of CCI 3–4 is present in the Geriatric unit (14%).

The most frequent APR severity class is 2 (60%), where in General Medicine 1 and Respiratory Medicine has a maximum of 67% and 77%, respectively. The 16% and 22% of patients have a severity class of 1 and 3, respectively, while only the 2% has a severity class of 4. The most frequent APR mortality risk is 2 (43%), where in Cardiology and Geriatric units has a maximum of 67% and 49%, respectively. The 17% and 35% of patients have a mortality risk of 1 and 3, respectively, while only the 5% has a mortality risk of 4. Infection and Immunology ward treat patients with higher APR values (both severity class and mortality risk). Geriatrics, Immunology and General Medicine 1 and 2 have at least the 30% of patients with a high risk of mortality (3 or 4). Different combinations of complexity characteristics for each Clinical Pathway (represented by the DRG) define groups of patients that should be homogeneous with respect to the resource use and cost. After this adjustment, the residual variability among wards is due only to different organizational models.

In the next section a simulation model is developed to evaluate the impact on resource utilization (beds and costs) and outcome (average length of stay) of merging all IMW into a unique ward following the best model.

4 Simulation Model Development

The adoption of simulation modeling in the healthcare context derives from the need to reproduce the system reality and to provide to the decision maker a good or optimal solution for health policies. Since the 1970s were published several scientific articles where simulation techniques were applied to analyze healthcare services [3, 9, 10]. DES is a simulation technique that was used widely in health care to provide evidence of "what-if" and scenario analysis before implementation in reality [20]. DES is an effective modeling technique to represent the care pathways structures, it can include inside its structure resource constraints and health outcomes. "What-if" scenarios analyses and determines the effect of implementing changes and process re-organization in the whole system performance [6, 14]. The adoption of solutions provided by "what if" analysis through simulation models, enables to understand the system behavior and the implication of a process re-organization before their implementation [1]. In this paper, a DES model has been developed and implemented using the simulation software WITNESS to assess the impact of introducing a patient-centered reconfiguration of the medical wards stay area. The schematic flow chart of the resulting DES model is reported in Fig. 1.

Following a patient-centered perspective, new patients enter the system belonging to a Pathology-related Clinical Pathway, represented by the DRG. Note that, all patients arrive as urgent and are directly admitted from the Emergency Department. The number and time of arrivals of patients for each DRG are taken from the data collection as well as the main characteristics associated. To consider the current occupation of beds at the beginning of the planning horizon, the number and LOS of patients already in the hospital are generated using retrospective data and pushed into the stay area. Note that, using the real data to feed the system with the patients already present at the beginning of the simulation run, we do not need to perform a warm-up to reach steady-state simulation. In fact, in our analysis, we want to simulate the flows of the cohort of patients as collected by real data verifying the impact of different organizational settings.

During the simulation run, new patient arrivals are managed using an arrival profile input data. Patients arriving in the system are segmented using demographic and clinical characteristics, as reported in Tables 4 and 5, DRG and ward of admission. Different combinations of these characteristics define groups of patients homogenous with respect to the resource use and cost. Each identified group is then associated with a LOS and cost distribution function. After hospital admission, patients flow in the system depending on the clinical pathway and organizational model of the stay area used.

As introduced in Sect. 2, the organizational model refers to how the stay areas are organized, i.e. specialty-based versus patient-centered hospital organization. The first reproduces the current practice where patients are admitted in the ward collected by real data. Instead, in the re-organization that follows the patient-centered model, all patients are accepted into a generic ward, where the multidisciplinary team organizes and takes in charge the patient hospital stay and providing a holistic vision of the

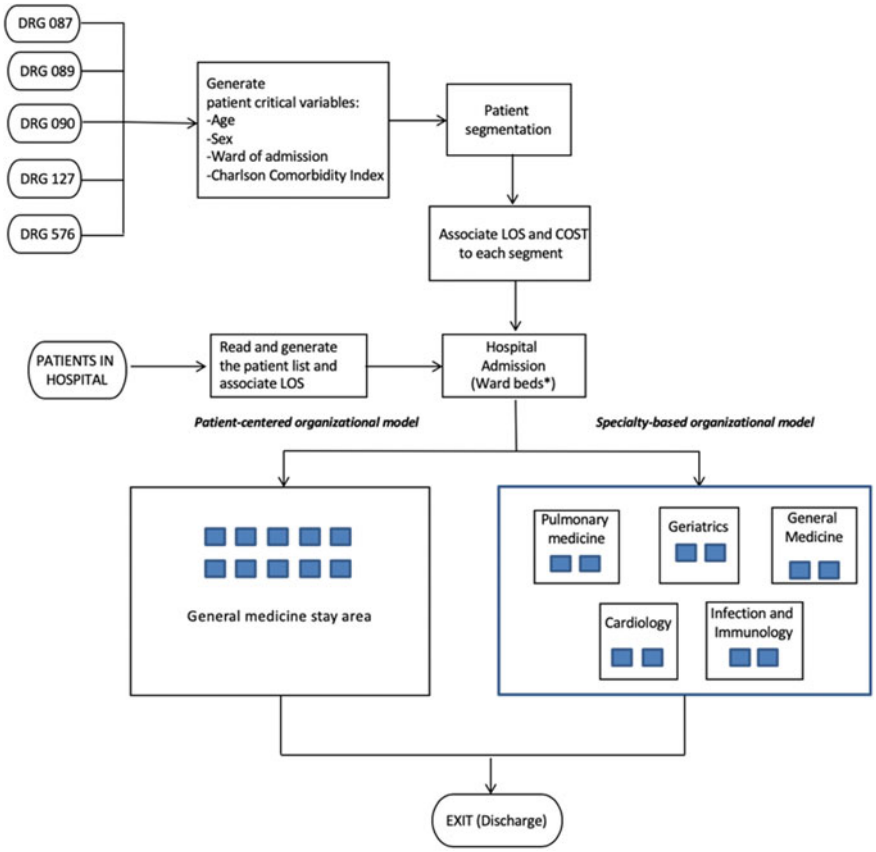


Fig. 1 Schematic flow chart of the system under study

care process. Dealing with multi-pathologies patients recovered in medical wards, the main resources in the care process are beds and clinical staff. Assuming that the number of clinicians and nurses are fixed, the main question herein addressed is determining how many beds are needed for each ward to treat the considered cohort of patients in both scenarios. To answer this question, we used the optimization module integrated into the simulation environment. We use as constraints the overall capacity in terms of the number of beds for each ward, as collected from real data. The objective function aims at defining the optimal number of beds to avoid cancellations and delayed admissions. Obviously, in the best scenario the objective function must reach the null value, guaranteeing that all patients arriving in the system in the exact timing of the real data (real arrival profile) are admitted.

5 Preliminary Results

The data-driven simulation model has been used to exactly reproduce the cohort of patients under study with their characteristics and their flow rules validated with the clinicians involved in our study to ensure its ability to represent the real system under investigation. Two scenarios are tested to evaluate the effect of re-organizing the "traditional", specialty-based, stay area (each medical ward has its available beds) into a new patient-centered organization (beds are shared among all medical wards and patients are all treated as they are in an IMW).

In Table 6, the number of beds needed and the average length of stay in the three scenarios are reported. Note that, concerning the patient-centered model, two configurations are tested using for each patient group the LOS distributions and costs of the data collected in General Medicine 1 and General Medicine 2, respectively.

In both patient-centered scenarios, a reduction of the total number of beds needed is shown passing from 119 beds, in the current scenario (Specialty-based), to 115 and 98, respectively, in Patient-centered configuration (1) and (2), with a percentage reduction of beds of 3.4% and 17.6%. The outcome, measured by the average length of stay, shows improvement only in the Patient-centered model (2), where it reduces from 9.4 days to 8.3 days on average with a percentage reduction of 11.7%.

In Table 7 the average cost for each DRG and the total cost of the cohort is reported for the two scenarios. For both configurations of the patient-centered model a cost reduction is observed for all DRGs. Note that the average cost herein reported is weighted for the number of the patient in the segment and reflects the differences among the number and types of tests performed to the patients belonging to the segment analysed. Shifting from a specialty-based model to a patient-centered one, a total average reduction of 3% and 28% is obtained in configuration (1) and (2) respectively.

The better results of configuration (2) can be explained by the different skills of the clinicians of the two wards that affect the clinical pathways and outcomes of patients treated. In particular, in General Medicine 2 ward, the skills and abilities of the physicians are similar to the hospitalist, as described by literature: they perform ultrasounds on their own, as well as most invasive procedures such as positioning of central venous catheters, they plan the controls themselves or some changes in therapies such as insulin or laxative, helping to anticipate the controls, identify early or prevent complications, and thus shorten the stay and reduce the costs accordingly.

6 Conclusions and Future Works

This study focuses on the analysis of the impact of the adoption of a new organizational model for medical wards (Patient-centered model) with respect to the standard organization currently in use (Specialty-based model), considering both resources

Table 6 Specialty-based versus Patient-centered model (number of beds and LOS)

	Specialty-based model		Patient-centred model (1)		Patient-centred model (2)	
# beds	# of beds	Average LOS	# of beds	Average LOS	# of beds	Average LOS
Cardiology	11	6.6	–	–	–	–
Geriatrics	16	9.1	–	–	–	–
Infection and Immunology	8	14.0	–	–	–	–
General Medicine 1	24	9.0	115	9.5	–	–
General Medicine 2	40	7.3	–	–	98	8.3
Respiratory Medicine	20	10.5	–	–	–	–
Total	*119*	*9.4*	*115*	*9.5*	*98*	*8.3*

Table 7 Specialty-based versus Patient-centered model (average cost for DRG and total cost in Euro, €)

	Specialty-based model	Patient-centred model (1)	Patient-centred model (2)
DRG	Average cost		
087: Pulmonary edema and respiratory failure	2998.68	2832.19	2269.35
089: Pneumonia and pleuritis with complications	3340.27	2818.04	1823.34
090: Pneumonia and pleuritis >17 years old	2650.19	2286.95	1684.89
127: Heart failure and shock	2502.26	2667.70	1975.23
576: Sepsis without medical ventilation	3572.08	3560.18	2455.67
Average total cost	*3037.98*	*2940.44*	*2186.53*

use and outcomes. The flow of patients within the hospital wards was modeled including patient-relevant characteristics such as severity, comorbidities, age, and sex. A Discrete Event Simulation model was developed to represent the pathways of complex patients through medical hospital wards. The model evaluates the length of stay of patients and the resource use (consultations, blood transfusions and diagnostic, cardiology, imaging and laboratory tests), using two organizational models. A real case study based on a medium hospital setting was analyzed. The results show that the patient-centered model provides an improvement in terms of beds needed and length of stay reduction of about 17% and 12%, respectively. The reduction of costs provided by the patient-centered models of 3% and 28%, respectively.

This study presents two main limitations: the first consists in the limited use of outcome indicators, where other outcomes should be included such as 90-days patient readmission and in-hospital mortality; the second derives from the hospital data which the model is based, a sensitivity analysis should be provided in order to verify the robustness of the results. Future work will be directed to test the model on a larger dataset, made up of three years of hospital data records also distinguishing in detail the results with respect to different DRGs. Indeed, we will use Machine unsupervised learning techniques, such as K-means clustering to identify the main characteristics able to create representative clusters of patients, with similar characteristics in terms of the intensity level of care and corresponding costs.

References

1. Abuhay, T.M., Krikunov, A.V., Bolgova, E.V., Ratova, L.G., Kovalchuka. S.V.: Simulation of patient flow and load of departments in a Specialized Medical Center. Procedia Comput. Sci. **101**, 143–151 (2016)
2. Bauer, W., Schumm-Draeger, P.M., Koebberling, J., Gjoerup, T., Garcia Alegria, J.J., Ferreira, F., Higgens, C., Kramer, M., Licata, G., Mittelman, M., O'hare, J., Unal, S.: Political issues in internal medicine in Europe. A position paper. Eur. J. Intern. Med. **16**(3), 214–217 (2005)
3. Brailsford, S.C., Harper, P.R., Patel, B., Pitt, M.: An analysis of the academic literature on simulation and modelling in health care. J. Simul. **3**(3), 130–140 (2009)
4. Bruzzi, S., Landa, P., Tànfani, E., Testi, A.: Conceptual modelling of the flow of frail elderly through acute-care hospitals: an evidence-based management approach. Manag. Decis. **56**(10), 2101–2124 (2018)
5. Clegg, A., Young, J., Iliffe, S., Rikkert, M.O., Rockwood, K.: Frailty in elderly people. The Lancet **381**, 752–762 (2013)
6. Demir, E., Southern, D., Rashid, S., Lebcir, R.: A discrete event simulation model to evaluate the treatment pathways of patients with cataract in the United Kingdom. BMC Health Serv. Res. **18**, 933 (2018)
7. Deyo, R.A., Cherkin, D.C., Ciol, M.A.: Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. J. Clin. Epidemiol. **45**(6), 613–619 (1992)
8. Duckitt, R., Palsson, R., Bosanska, L., Dagna, L., Durusu, T.M., Vardi M.: CDIME group. Common diagnoses in internal medicine in Europe 2009: a pan-European, multi-centre survey. Eur. J. Intern. Med. **21**(5), 449–452 (2010)
9. 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)
10. Katsaliaki, K., Mustafee, N.: Applications of simulation within the healthcare context. J. Oper. Res. Soc. **62**(8), 1431–1451 (2011)
11. La Regina, M., Guarneri, F., Romano, E., Orlandini, F., Nardi, R., Mazzone, A., Fontanella, A., Campanini, M., Manfellotto, D., Bellandi, T., Gussoni, G., Tartaglia, R., Squizzato, A.: What quality and safety of care for patients admitted to clinically inappropriate wards: a systematic review. J. Gen. Int. Med. (2019)
12. Malone, T.W., Laubacher, R., Johns, T.: The big idea: the age of hyprspecialization. Harvard Business Review 2011; available at https://hbr.org/2011/07/the-big-idea-the-age-of-hyperspecialization, last consultation on 18th march 2019
13. McCormick, P.J., Lin, H., Deiner, S.G., Levin, M.A.: Validation of the all patient refined diagnosis related group (APR-DRG) risk of mortality and severity of illness modifiers as a measure of perioperative risk. J. Med. Syst. **42**, 81 (2018)
14. Ozcan, Y.A., Tànfani, E., Testi, A.: Improving the performance of surgery-based clinical pathways: a simulation-optimization approach. Health Care Manag. Sci. **20**(1), 1–15 (2017)
15. Pietrantonio, F., Orlandini, F., Moriconi, L., La Regina, M.: Acute Complex Care Model: an organizational approach for the medical care of hospitalized acute complex patients. Eur. J. Intern. Med. **26**(10), 759–765 (2015)
16. Rachoin, J.S., Skaf, J., Cerceo, E., Fitzpatrick, E., Milcarek, B., Kupersmith, E., Scheurer, D.B.: The impact of hospitalists on length of stay and costs: systematic review and meta-analysis. Am. J. Manag. Care **18**(1), e23–e30 (2012)
17. Salive, M.E.: Multimorbidity in older adults. Epidemiol. Rev. **35**(1), 75–83 (2013)
18. Wachter, R.M., Goldman, L.: The emerging role of "hospitalists" in the American health care system. New England J. Med. **335**, 514–517 (1996)
19. Wachter, R.M., Goldman, L.: Zero to 50,000—the 20th Anniversary of the Hospitalist. New England J. Med. **375**, 1009–1011 (2016)
20. Zhang, X: Application of discrete event simulation in health care: a systematic review. BMC Health Serv. Res. **18**(1), 687 (2018)

Benefits of a Broader View: Capturing the Hospital-Wide Impact of Surge Policies with Discrete-Event Simulation

Carolyn R. Busby and Michael W. Carter

Abstract There are many examples in the literature of techniques aimed at optimizing bed capacity and allocation, reducing ED waiting time and maximizing operating room utilization; however, most draw narrow boundaries around a specific department or service (specialty) or tackle each problem in isolation. These approaches help diagnose department specific issues but miss the wider picture and fail to capture the upstream and downstream impacts. This is particularly important in congested hospitals where operational issues in one department will likely have a ripple effect throughout the hospital. This inter-department dependence in congested hospitals is reflected in the surge policies used when congestion reaches a threshold that is deemed to require a whole hospital response. A hospital-wide generic discrete-event simulation (DES) model, that factors in the effects of congestion, has been built to tackle these problems. Insights on the importance of wide model boundaries and surge protocol modelling, gained through application of the model at three hospitals in Ontario, Canada, are shared. Example scenarios used to illustrate this include early discharge planning, and surgical throughput improvement.

Keywords Discrete event simulation · Hospital · Patient flow · Surge policy

1 Introduction

In a congested hospital, the interactions between departments become pronounced as patients compete for resources across the hospital. In addition, hospitals routinely facing congestion use "surge protocols" to help alleviate the pressure. In order to accurately capture patient flow in these hospitals, a discrete-event hospital-wide model, that includes surge protocols, was constructed. The model can be used to

C. R. Busby (✉)
ORCHID Analytics, 1262 Baldwin Dr., Oakville, ON L6J 2W5, Canada
e-mail: carolyn.busby@orchidanalytics.ca

M. W. Carter
University of Toronto, 5 King's College Rd., Toronto, ON M5S 3G8, Canada
e-mail: mike.carter@utoronto.ca

V. Bélanger et al. (eds.), *Health Care Systems Engineering*,
Springer Proceedings in Mathematics & Statistics 316,
https://doi.org/10.1007/978-3-030-39694-7_4

aid decisions on inpatient capacity and allocation, operating room scheduling, and hospital flow policies. Output includes ED boarding time (time in ED waiting for an inpatient bed), operating room throughput and cancellations, inpatient occupancy and off-servicing rates (patients put in bed assigned to different service (specialty), alternately referred to as "misallocation" or "outlier"). The model was constructed as a generic model such that it could be used at multiple hospitals. In this paper we discuss the validation and application of this model at three hospitals, highlighting insights gained on the value of hospital-wide modelling and surge protocol inclusion.

Many healthcare and simulation reviews [6, 7, 10, 12, 13, 26] conclude that very few hospital-wide models have been reported in the literature. Modelling multiple departments allows interactions between departments to be captured and ensures that problems are addressed where they originate. If focus is on just one department, effort may be spent within that department to compensate for an issue originating in another department, when it would be more appropriately addressed at the source [26]. While, there is evidence that research is trending towards more integrated models with multi-facility, multi-stage patient flows [2], hospital-wide discrete event models are still rare. In addition, most whole-hospital models are high-level, lacking detailed patient flow and/or coverage [4, 15–17, 19]. On the other hand, Norouzzadeh et al. [21] focus in on the details of the transition between the ED and the inpatient wards, modelling the staff communication process at a community hospital. Our model does not explicitly capture staffing as beds are the main constraint in congested hospitals, and wait times are generally measured in hours rather than minutes.

In congested hospitals, problems in one area quickly spread to other areas of the hospital. Likewise, improvements in one area can alleviate congestion in another. Surge protocols employed routinely in congested hospitals provide an example of how congestion is addressed using a hospital wide approach. Surge protocols vary by hospital but are typically triggered by insufficient inpatient bed availability or crowding in the Emergency Department. Typical surge responses include: increased effort to discharge eligible patients ("reverse triage"); opening extra beds; delaying transfers from other hospitals; transferring patients to other hospitals; canceling surgery; moving patients to lower level of care to make room for more acute patients; and increasing off-servicing options. The importance of Bed Management and potential solutions, including surge protocols, are thoroughly discussed in Proudlove et al. [23]. The need to include surge policies is illustrated by Wolstenholme et al. [27]. In order to validate a nationally developed System Dynamics model at the local level, they found that they needed to incorporate local coping strategies used in times of congestion. Finally, Landa et al. [18] use DES and multi-objective optimization to determine the optimal bed management decision rules and bed allocation to achieve minimum ED boarding time, number of ED boarding patients, number of off-serviced patients, and cancelled elective admissions. The simulation model captures arrival streams from both the ED and elective patients into inpatient beds differentiated by service (ED patients could also flow through a short-observation unit before admission to an inpatient bed). Surgery was not explicitly modeled. The bed management options (analogous to surge protocol responses) considered included off-servicing and cancellation of elective admissions. The decision rule variables that could be

changed included triggers for when a bed management response should occur (the maximum time an individual ED patient can wait for an on-service inpatient bed and the maximum number of ED patients that can be waiting for an inpatient bed) and the duration over which elective patients are delayed. In our model, the later was fixed at one day, but could be triggered over multiple consecutive days.

Capturing surge protocols requires a whole-hospital approach. The clear drawback of models with wider boundaries is complexity and model size. Creating a detailed generic simulation model amortizes the time invested by only coding the model once and then applying it at multiple hospitals. Our model is what Fletcher and Worthington [6] classify as a "setting-specific generic model designed for multiple local use" (level 3B generic model). There are three notable examples of widely applied department-level 3B generic hospital simulation models: An ED model applied in ten hospitals [5]; a perioperative model applied in fourteen hospitals [25] (of which this model is an expansion); and a multi-department model [8, 9, 11] that does not include detailed OR flow, but does include more ED detail and clinic detail than this model.

2 Model Overview

In this section, an overview of the elements included in the model are described. A previous publication [3] offers additional details on the model development and techniques used.

The model is composed of a Simul8 simulation model that captures common elements of hospital flow while allowing hospital-specific resources, policies and patient characteristics to be uploaded via an Exel input file. Upon importing the data to the simulation model, code is triggered to configure the model according to the imported hospital data. The hospital-specific simulation is then ready to be run and hospital-specific output data collected. As shown in Fig. 1, the main areas covered include the emergency department, operating rooms and inpatient beds. Auxiliary services such as diagnostics are not included in this model, nor are outpatient clinics.

Patient files containing arrival time, priority level, services required and associated lengths of stay, are pulled in their entirety from historical hospital records. This technique maintains correlation between each element of the patient stay. These correlations are shown by Luangkesorn et al. [20] to affect the distribution of inpatient occupancy and so are important to maintain. In addition, this technique facilitates data updates and continued use of the model by client hospitals without the need to fit complex dependent distributions.

The daily volume of entities arriving via the ED is based on a Poisson distribution that is day-of-week dependent. The Poisson distribution is an ideal choice for a generic model. It generically representation random arrivals and allows the lay user to update the distribution simply by calculating the average arrival rate rather than having to fit data to a distribution. In addition, to capturing the variation in arrival rates across the week, intra-day arrival time variation is captured by basing the arrival

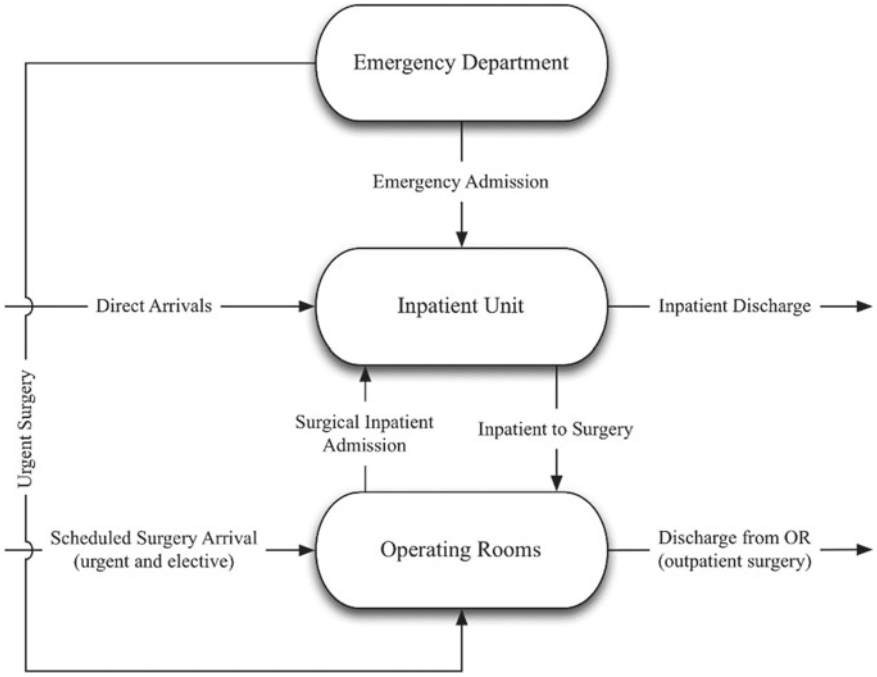


Fig. 1 Model coverage and flow

time of each patient on the patient file pulled. This is accomplished by having all patients for the day arrive to a queue at the start of each day. The arriving entity then chooses a patient file and enters the emergency room at the time specified on the patient file. This also ensures that any correlation between patient type and arrival time of day is captured. For example, car accident victims may be more likely to arrive during the day than at night and night admissions may be more likely to be emergent.

Upon admission to the hospital a check is done for an inpatient bed. If unavailable, the patient waits in the ED until one becomes available. If the patient requires urgent surgery, they are placed on an urgent surgical list. These patients may continue to wait in the ED or are moved to a surgical unit. Direct Admission patients, which typically includes maternity patients and transfers from other facilities or clinics, also arrive to the model according to a Poisson distribution by day-of-week. Patients then wait to be assigned an inpatient bed. Elective surgical patients arrive randomly to surgeon specific elective waiting lists where they wait to be scheduled for surgery.

At the start of each day, the model schedules patients for surgery based on an inputted master surgical schedule and surgeon assignments. This model is designed for use with a block schedule where hospital administrators assign full-day or half-day blocks in each OR to a particular surgical specialty. Each specialty then assigns specific surgeons to each allocated OR block. Schedules can be of any length but are

typically on a 1–4 week cycle. In the model, scheduled patients enter the OR and remain for the time specified in their patient file. Completed patients move to the post anaesthesia care unit (PACU), an inpatient bed (Critical Care or Ward), or leave the hospital. If a patient is meant to go to an inpatient bed that is not yet ready, the patient will wait in the PACU. If the PACU is full, the patient will remain in the OR blocking the next surgery. Checks are then done to see if there is time for the next scheduled surgery and if there will be a bed available (if required). If not, the surgery is cancelled.

Urgent surgeries must be completed within target times based on acuity level and availability of ORs. Urgent patients can be accommodated by scheduling them into an elective or reserved urgent block ahead of time, using time remaining at the end of an elective block, bumping a scheduled elective patient during operating hours, or completing them outside of regular operating hours. The policies governing these decision are hospital specific and therefore captured in the input file.

When a patient requires a bed, the model first looks for available on-service beds. If none are found, a check for a suitable off-service bed is done. Based on rules selected in the input file the patient is either sent to the off-service bed immediately or waits for an on-service bed that will open later in the day. Inpatient units available for off-servicing may increase when in surge. If no bed is available, patients wait until an appropriate bed opens up. When a bed becomes available a check is done to see if any patients are waiting for the bed. If multiple appropriate patients require the bed, it is assigned based on priorities specified in the input file. For each patient unit in the model an ordered priority is given to search for a waiting patient categorized by location waiting (e.g. ICU, ED) and match to bed (e.g. off-service, on-service). Patients also move between inpatient units as service requirements change.

There are three types of inpatient beds included in the units: funded, flex, and unfunded. Funded beds have planned, consistent nursing levels and set nurse-to-bed ratios. Flex beds are extra physical beds that can be temporarily used without adding nursing staff. Unfunded beds are physical beds that can be opened when necessary by bringing in extra nursing staff. These flex and unfunded beds are used in times of surge.

Surge policies are modelled by capturing the triggers and responses associated with each surge level and affected area of the hospital. Levels of surge correspond to the degree of congestion and therefore have different trigger thresholds and degrees of response. Triggers included in the model are: number of admitted patients in the ED; number of patients exceeding total anticipated free beds; previous days in surge; number of scheduled surgeries with no reserved bed; number of flex or unfunded beds currently open. Responses included in the model are: expediting discharge (“reverse triage”); expediting move to lower level of care; allowing increased off-servicing; opening flex or unfunded beds; delaying outside transfers into the hospital; canceling of elective surgery; and transferring ED patients to another hospital. Hospital-specific surge policies are defined by the combination of triggers and responses used at each level of surge. This allows surge protocols used by the hospital to be replicated. As an example, a level 1 surge may occur because the number of patients waiting in the ED for an admission to an inpatient bed may exceed the specified threshold. This

may then result in the administration instructing physicians to check patients in the ward to see if accommodations can be made to allow patients to be discharged home more quickly (expedited discharge). A level 4 surge may be called if the hospital has been in surge for several days, beds are full and there is a back-up of patients in the ED. At this stage hospital administrators may open any remaining unfunded beds, prevent patient transfers in from other hospitals, and contact downstream services, such as home care, to help support expedited patient discharge.

The surgical block schedule, surgeon assignments, surgical priority levels and deadlines, turnover time, number of inpatient units, number and type of beds in each unit, as well as rules for bed prioritization, surgical scheduling, overtime, and surge triggers, responses, affected areas and levels are all defined by the user in the input file.

3 Results and Validation

The model was tested at three hospitals with widely varying characteristics to ensure that one generic model could adequately represent each hospital, using the standard input file to generate hospital-specific models. The characteristics of each hospital are described in Table 1.

Model results, compared to historical data, are shown for the three hospitals in Figs. 2, 3 and 4. Number of runs varied between the three hospitals, ranging from 16 to 60 and were set to obtain 95% confidence intervals that are within a reasonable level of error for midnight census and ED boarding time. These two outputs were chosen because they are affected by all other model outputs and therefore have the most variability. Reasonable errors were determined to be a midnight census within 2% overall and within 2% or 1.5 beds in each inpatient unit and an overall average ED wait time and medical ED wait time within 10% or 1 h.

Validation was completed on these and other indicators, including surgical throughput, by comparing actual historical values to model outputs. All results were

Table 1 Pilot hospitals' characteristics

Hospital	Location	Type	Operating rooms	Inpatient beds	ED visits/year
A	Downtown Urban	Trauma centre Teaching hospital	22	~535	~75,000
B	Suburban	Community hospital	9	~270	~47,000
C	Small town	Small community hospital	2	~60	~21,000

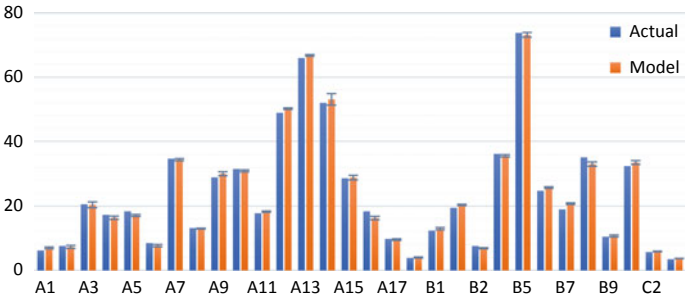


Fig. 2 Actual and predicted average occupancy by inpatient unit at the three pilot hospitals (A = Hospital A, 1 = unit 1 etc.)

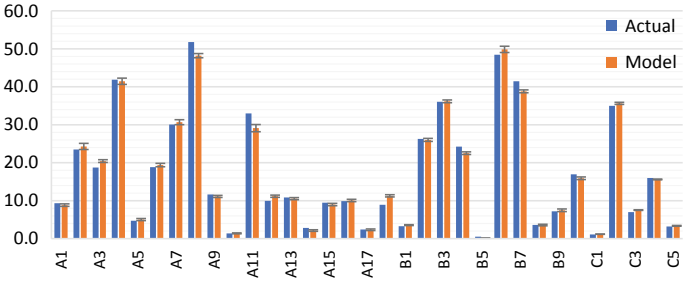


Fig. 3 Actual and predicted average surgical throughput by specialty service (cases/week) at the three pilot hospitals (A = hospital A etc., 1 = service 1 etc.)

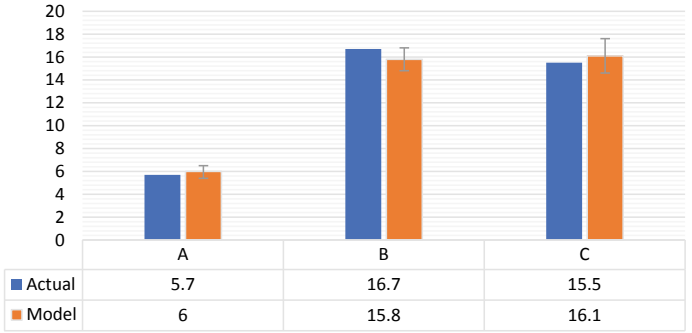


Fig. 4 Actual versus model predicted average ED boarding time (hours) at the three hospitals (A = Hospital A etc.)

also discussed with hospital personnel to determine if the model was sufficiently representative of their process.

4 Impact of Hospital-Wide Boundaries and Surge

With the assurance that the model was sufficiently representative in each case, several popular hospital management theories including increased AM discharge, 85% bed capacity, LOS changes and the impact of surge policies when full (from Proudlove et al. [23]). In addition, hospital-requested scenarios such as OR schedule optimization and bed reallocation were tested. In this section, two scenarios are reported that underscore how hospital-wide modelling was able to capture essential interactions leading to conclusions that would not be found in a department focused model. The direct impact of surge policy modelling is also highlighted in one of the examples.

4.1 Operating Room Throughput and Utilization

Operating room utilization improvement scenarios at one of the hospitals showed that an increase in surgical throughput was restricted by bed availability. Further, it was found that while there were sufficient surgical beds to accommodate increased throughput of surgical patients, the current level of patient off-servicing from the medical units prevented their use by surgical patients. Therefore, only surgical blocks dedicated to outpatient procedures could be added to the existing block schedule unless changes were made to reduce the off-servicing of medical patients into surgical beds.

Upon closer examination of the model, it was also observed that medical patients were flowing into surgical beds because “Alternate Level of Care” (ALC) patients were occupying medical beds. ALC patients are those that no longer need an acute bed in the hospital but are not sufficiently well to be discharged home. They are designated as requiring an alternate form of care such as home care, rehab or long-term care. However, if those services are not immediately available, they remain in the hospital until the required services become available.

By using a hospital-wide model we were able to better understand how ALC patients remaining in the hospital and taking up beds in the medical ward impact OR utilization and throughput. We were able to show that adding surgical blocks that serviced patients requiring and inpatient stay would lead to increased ED and inpatient congestion as well as cancelled surgeries. An OR-focused model would not have flagged the need to limit additional blocks to outpatient procedures, as the model would show that there were sufficient beds to accommodate an increase in inpatients. The hospital-wide model can better identify the true impact of the change and isolate the barriers to OR throughput expansion so that barriers can be addressed in advance.

4.2 Morning Discharge Policy to Reduce ED Boarding Time

There has been a great deal of attention in practice and in the literature, in recent years on the policy of discharging patients as early as possible in the day to free up beds for incoming patients and reduce crowding in the ED. Most literature indicates that if discharges are moved earlier in the day, there is a positive effect on crowding, wait time for an inpatient bed, and peak and average hospital occupancy [1, 14, 22, 28]. However, Shi et al. [24] found mixed results in both empirical observations, and simulated scenarios.

The effect on ED boarding time of dismissing patients by 9 am was tested at each hospital with the generic simulation model. Discharge times after 9 am were adjusted to between 7 am and 9 am. This represents an aggressive version of the policy that would be difficult to achieve in practice, so gives a "best case" outcome. The policy was effective at two of the three hospitals, reducing ED boarding time by 2–3 h on average. However, at the third hospital the ED boarding time remains stable.

The surge trigger at Hospital B was based on the number of admitted patients waiting in the ED. So, the early movement of patients reduced the spike of admitted patients waiting, and therefore the chance of surge being called; the percentage of time in surge was reduced by 43%. The result was fewer extra beds opened, but unchanged ED boarding time. If the hospital wished to shift the balance to reduce boarding time, they could test changing surge protocol triggers as part of the policy change. Models focused on a narrower process or that don't include surge policies would have failed to capture these complex relationships. If simply piloted live, management could have concluded that the pilot was a failure based on measured outcomes in the ED alone.

5 Conclusion

The generic discrete event simulation model presented demonstrates the importance of using sufficiently wide model boundaries. The examples discussed show how interventions can have unexpected impacts due to hospital wide flow that will be missed if model boundaries are drawn too tightly. In addition, the critical role of surge protocols in congested hospitals is demonstrated. These critical protocols cannot be accurately captured in narrowly focused models. The drawback of widening model boundaries, without loss of detail, is clearly complexity, leading to increased cost and effort. This is mitigated in this case by creating a generic model that can be amortized across several hospitals. This wide-boundary model can complement smaller, faster models by testing findings to see if the conclusions hold when placed in a wider setting.

Acknowledgements This work has been funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Ontario government.

References

1. Ben-Tovim, D., Filar, J., Hakendorf, P., Qin, S., Thompson, C., Ward, D.: Hospital event simulation model: arrivals to discharge-design, development and application. Sim. Model. Pract. Th. **68**, 80–94 (2016)
2. Bhattacharjee, P., Ray, P.K.: Patient flow modelling and performance analysis of healthcare delivery processes in hospitals: a review and reflections. Comput. Ind. Eng. **78**, 299–312 (2014)
3. Busby, C.R., Carter M.W.: Data-driven generic discrete event simulation model of hospital patient flow considering surge. In: Chan, W.K.V., D'Ambrogio, A., Zacharewicz, G., Mustafee, N., Wainer, G., Page, E. (eds.) Proceedings of the 2017 Winter Simulation Conference, pp. 3006–3017. Institute of Electrical and Electronics Engineers, Inc., Piscataway, New Jersey (2017)
4. Demir, E., Gunal, M., Southern, D.: Demand and capacity modelling for acute services using discrete event simulation. Health Syst. **5**, 1–8 (2016)
5. Fletcher, A., Halsall, D., Huxham, S., Worthington, D.: The DH accident and emergency department model: a national generic model used locally. J. Oper. Res. Soc. **58**, 1554–1562 (2007)
6. Fletcher, A., Worthington, D.: What is a 'Generic' hospital model? A comparison of 'Generic' and 'Specific' hospital models of emergency patient flows. Health Care Manag. Sci. **12**, 374–391 (2009)
7. Fone, D., Hollinghurst, S., Temple, M., Round, A., Lester, N., Weightman, A., Roberts, K., Coyle, E., Bevan, G., Palmer, S.: Systematic review of the use and value of computer simulation modelling in population health and health care delivery. J. Public Health (Bangkok) **25**, 325–335 (2003)
8. Gunal, M.M., Pidd, M.: Interconnected DES models of emergency, outpatient, and inpatient departments of a hospital. In: Henderson, S.G., Biller, B., Hsieh, M.-H., Shortle, J., Tew, J.D., Barton, R.R. (eds.) Proceedings of the 2007 Winter Simulation Conference, pp. 1440–1445. Institute of Electrical and Electronics Engineers, Inc., Piscataway, New Jersey (2007)
9. Gunal, M. M., Pidd, M.: DGHPSIM: Supporting Smart Thinking to Improve Hospital Performance. In: Mason, S.J., Hill, R.R., Mönch, L., Rose, O., Jefferson, T., Fowler, J.W. (eds.) Proceedings of the 2008 Winter Simulation Conference, pp. 1484–1489. Institute of Electrical and Electronics Engineers, Inc., Piscataway, New Jersey (2008)
10. Günal, M.M., Pidd, M.: Discrete event simulation for performance modelling in health care: a review of the literature. J. Simul. **4**, 42–51 (2010)
11. Günal, M.M., Pidd, M.: DGHPSIM: generic simulation of hospital performance. ACM T. Model. Comput. S. **21**, 1–22 (2011)
12. 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 structured review of the state of the art in OR/MS. Health Syst. **1**, 129–175 (2012)
13. 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**, 109–123 (1999)
14. Khanna, S., Sier, D., Boyle, J., Zeitz, K.: Discharge timeliness and its impact on hospital crowding and emergency department flow performance. Emerg. Med. Australas. **28**, 164–170 (2016)
15. Kortbeek, N., Braaksma, A., Smeenk, F.H.F., Bakker, P.J.M., Boucherie, R.J.: Integral resource capacity planning for inpatient care services based on bed census predictions by hour. J. Oper. Res. Soc. **66**, 1061–1076 (2015)
16. Landa, P., Sonnessa, M., Tànfani, E., Testi, A.: A discrete event simulation model to support bed management. In: Proceedings of the 4th International Conference on Simulation and Modeling Methodologies, Technologies and Applications, pp. 901–912. IEEE (2014)
17. Landa, P., Sonnessa, M., Tànfani, E., Testi, A.: Managing emergent patient flow to inpatient wards: a discrete event simulation approach. In: Obaidat, M.S., Ören, T., Kacprzyk, J., Filipe, J. (eds.) Simulation and Modeling Methodologies, Technologies and Applications. Advances in Intelligent Systems and Computing, vol. 402, pp. 333–350. Springer, Cham (2015)

18. Landa, et al.: Multiobjective bed management considering emergency and elective patient flows. Int. Trans. Oper. Res. **25**, 91–110 (2018)
19. Lin, D., Patrick, J., Labeau, F.: Estimating the waiting time of multi-priority emergency patients with downstream blocking. Health Care Manag. Sci. **17**, 88–99 (2014)
20. Luangkesorn, K.L., Bountourelis, T., Schaefer, A., Nabors, S., Clermont, G.: The case against utilization: deceptive performance measures in in-patient care capacity models. In: Laroque, C., Himmelspach, J., Pasupathy, R. (eds.) Proceedings of the 2012 Winter Simulation Conference, pp. 847–856. Institute of Electrical and Electronics Engineers, Inc., Piscataway, New Jersey (2012)
21. Norouzzadeh, S., Garber, J., Longacre, M., Akbar, S., Riebling, N., Clark, R.: A modular simulation study to improve patient flow to inpatient units in the emergency department. J. Hosp. Adm. **3**(6), 205–215 (2014)
22. Powell, E.S., Khare, R.K., Venkatesh, A.K., Van Roo, B.D., Adams, J.G., Reinhardt, G.: The relationship between inpatient discharge timing and emergency department boarding. J. Emerg. Med. **42**, 186–196 (2012)
23. Proudlove, N.C., Gordon, K., Boaden, R.: Can good bed management solve the overcrowding in accident and emergency departments? Emerg. Med. J. **20**, 149–155 (2003)
24. Shi, P., Chou, M.C., Dai, J.G., Ding, D., Sim, J.: Models and insights for hospital inpatient operations: time-dependent ED boarding time. Manag. Sci. **62**, 1–28 (2016)
25. Sniekers, D., Busby, C.R., Carter, M.W.: Implementation of Generic Perioperative Simulation at Multiple Hospitals. Working Paper, Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, Ontario, Canada (2018)
26. VanBerkel, P.T., Boucherie, R.J., Hans, E.W., Hurink, J.L., Litvak, N.: A survey of health care models that encompass multiple departments. Int. J. Health Manag. Info. **1**, 37–69 (2010)
27. Wolstenholme, E., Monk, D., McKelvie, D., Arnold, S.: Coping but not coping in health and social care: masking the reality of running organisations beyond safe design capacity. Sys. Dyn. Rev. **23**, 371–389 (2007)
28. Zhu, Z., Hen, B.H., Teow, K.L.: Estimating ICU bed capacity using discrete event simulation. Int. J. Health Care Qual. Assur. **25**, 134–144 (2012)

Coxian Phase-Type Regression Models for Understanding the Relationship Between Patient Attributes, Overcrowding, and Length of Stay in Hospital Emergency Departments

Laura M. Boyle, Adele H. Marshall and Mark Mackay

Abstract Hospital emergency departments (EDs) operate under significant pressure worldwide. Overcrowding is a frequent occurrence, caused by a combination of high presentation numbers, and long delays in the admission of patients due to a lack of availability of hospital beds. Understanding the factors which influence length of stay (LoS) in the ED is a vital aspect of any strategy to improve patient flow. The determinants of ED patient flow are complex and varied, due to the diverse population of patients competing for limited resources. This research uses Coxian phase-type distributions to cluster patients into groups by their LoS, using a unique diagram for improved communication of patient flow issues. A novel application of survival analysis is presented to simultaneously evaluate the effect of patient attributes, system factors, and overcrowding on ED LoS. The approach is demonstrated with an application to data from a hospital in South Australia.

Keywords Emergency department overcrowding · Hospital length of stay · Coxian phase-type distributions · Patient flow

1 Introduction

Hospital emergency departments (EDs) are frequently 'overcrowded' and operating under pressurised conditions, due to an excess of patient demand for the available

L. M. Boyle (✉)
School of Mathematical Sciences, ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS), University of Adelaide, Adelaide, Australia
e-mail: laura.boyle@adelaide.edu.au

A. H. Marshall
Mathematical Sciences Research Centre, Queen's University Belfast, Northern Ireland, UK
e-mail: a.h.marshall@qub.ac.uk

M. Mackay
School of Nursing and Midwifery, University of South Australia, Adelaide, Australia
e-mail: mark.mackay@unisa.edu.au

V. Bélanger et al. (eds.), *Health Care Systems Engineering*,
Springer Proceedings in Mathematics & Statistics 316,
https://doi.org/10.1007/978-3-030-39694-7_5

hospital resources [8]. Until recently, Australian EDs were subject to the National Emergency Access Target (NEAT) [3], which specified that pre-determined proportions of patients should be admitted, discharged, or transferred within four hours of their presentation to hospital. Since the end of 2017, NEAT is no longer officially in use on a national level within Australia, however individual states continue to use iterations of the four-hour rule. Statistics from the Australian Institute of Health and Welfare for 2016–2017 report that the proportion of patients seen 'on time' (ie. within four hours) in South Australian EDs was 64% [2]. Failure to meet performance targets is attributed to a number of factors, including high attendance rates. The number of presentations to EDs in Australia continues to increase by 3.7% annually [2], placing a growing strain on EDs to treat an increased volume of patients within a capacity-constrained physical space. Further problems are caused by 'boarding' patients, who remain in ED for lengthy periods awaiting availability of inpatient hospital beds [15]. Both the high attendance and boarding patient problems pose a considerable challenge, as these factors lie outside the control of the ED.

1.1 Review of Relevant Literature

Understanding the factors which influence length of stay (LoS) in ED is a vital aspect of any strategy to improve patient flow and increase system performance. The determinants of ED patient flow are complex and varied, due to the diverse population of patients competing for limited resources. Regression models are commonly used, however they require a log-transformed or discretised outcome variable to handle the skewness typically found in LoS data [9]. Survival analysis techniques are more appropriate for modelling positively skewed data [18]. Cox proportional-hazards (PH) regression has been utilised [16], however difficulties arise when handling the assumption of proportional hazards with a sizeable number of variables, and obtaining the distribution of baseline hazard for prediction in the presence of tied event times [5], which are often present in large ED datasets. These limitations can be avoided by using accelerated failure time (AFT) models to predict ED LoS [4], where interpretation of the parameter effects as acceleration factors is more intuitive and clinically meaningful than through hazard ratios. Coxian phase-type distributions are a type of Markov model which have been used in patient flow modelling to represent LoS as a series of latent stages prior to departure from the system [7]. These models offer advantages over other survival distributions, both in flexibility, and opportunity to gain additional information regarding the underlying rates of flow through phases of the survival process prior to the event of interest [1]. Coxian phase-type regression models have been used to evaluate the effect of explanatory variables on ED LoS for a cohort of respiratory patients [19]. The authors noted an improvement in fit of their data over standard AFT models.

The models discussed thus far did not consider the effect of overcrowding on ED LoS. Despite a sizeable amount of literature on quantifying ED overcrowding

[8], there are relatively few studies which investigate it's effect on ED LoS [12]. McCarthy et al. [12] considered the relationship between ED occupancy and LoS using a discrete time survival model.

1.2 Overview and Contributions

The work presented includes the following contributions;

1. Coxian phase-type distributions are used to cluster patients into groups by their length of stay [11]. A unique diagram approach is developed to visualise the differences between patient attributes in each group. This type of analysis addresses a recently reported gap in the ED LoS literature [9] which indicated a lack of clustering methods to identify groups of patients with protracted LoS;
2. To the best of the authors' knowledge, this is the first study to simultaneously evaluate the effect of patient attributes, system factors, and overcrowding on ED LoS;
3. The work of McCarthy et al. [12] on overcrowding is extended from discrete-time to continuous-time survival analysis.

The remainder of this paper is laid out as follows. Section 2 outlines the methodology of Coxian phase-type regression models, Sect. 3 demonstrates an application of the methods to data from an Australian ED. The paper concludes with a discussion in Sect. 4.

2 Methodology

Phase-type distributions are a type of Markov process, representing time until absorption of a finite Markov chain in continuous time, where there is a single absorbing state and the process begins in a transient state [14]. Phase-type distributions can be generalised to approximate almost all continuous distributions [6]. The main limitation of phase-type distributions is that they are over-parameterised, typically requiring $(k^2 + k)$ parameters to represent a distribution with k phases. Coxian phase-type distributions are a special subclass with $(2k - 1)$ parameters, meaning they are more computationally efficient [6].

Coxian phase-type distributions consist of a Markov process where entities move sequentially through ordered transient stages until absorption occurs [14]. Figure 1 displays the k-phase Coxian representation. The parameters λ_k represent the instantaneous risk of transitioning between transient phases i and $(i + 1)$, and parameters μ_k represent the instantaneous risk of transitioning between transient phase i and the absorbing phase $(k + 1)$. This process flow corresponds to the movement of patients through treatment stages in ED departing from the system.

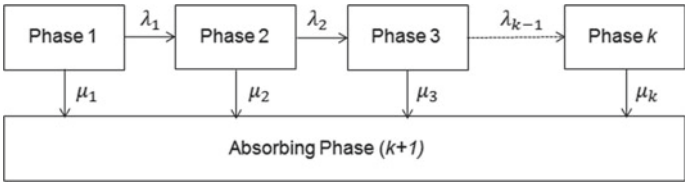


Fig. 1 Transition diagram for a k-phase Coxian distribution, with arrows indicating all possible transitions and directions of the Markov process

More formally, Coxian phase-type distributions may be defined by a latent Markov chain in continuous time, as $\{X(t);\, t \geq 0\}$ with state space $\{1, 2, \ldots, k, k+1\}$, where the process always begins in phase one of the model (initial state $X(0) = 1$). Then the probability of a transition occurring in a small time interval, say δt between transient states $i = 1, 2, \ldots, k$ is:

$$\text{prob}\{X(t+\delta t) = i+1 | X(t) = i\} = \lambda_i \delta t + \text{o}(\delta t)\ , \tag{1}$$

where λ_i represents the probability of transitioning between transient phases of the model. Furthermore, the probability of transitioning from each of the transient phases $i = 1, 2, \ldots, k$ into the absorbing phase $k+1$ may be written as:

$$\text{prob}\{X(t+\delta t) = k+1 | X(t) = i\} = \mu_i \delta t + \text{o}(\delta t)\ , \tag{2}$$

where μ_i represents the probability of absorption from transient phases of the model, and Eqs. (1) and (2) are correct to terms of order $\text{o}(\delta t)$ [10, 14].

The probability density function of the Coxian phase-type distribution can be expressed in the following matrix notation:

$$f(t) = \mathbf{p} \exp\{\mathbf{Q}t\}\mathbf{q}, \tag{3}$$

where $\mathbf{Q}$ is the $(k \times k)$ generator matrix, containing the transition intensities between the transient phases of the Markov model:

$$\mathbf{Q} = \begin{pmatrix} -(\lambda_1+\mu_1) & \lambda_1 & 0 & \ldots & 0 & 0 \\ 0 & -(\lambda_2+\mu_2) & \lambda_2 & \ldots & 0 & 0 \\ . & . & . & & . & . \\ . & . & . & & . & . \\ . & . & . & & . & . \\ 0 & 0 & 0 & \ldots & -(\lambda_{k-1}+\mu_{k-1}) & \lambda_{k-1} \\ 0 & 0 & 0 & \ldots & 0 & -\mu_k \end{pmatrix}, \tag{4}$$

$\mathbf{p}$ is a $(1 \times k)$ vector of probabilites for the process beginning in each state:

$$\mathbf{p} = (1, 0, 0, \ldots, 0, 0), \tag{5}$$

and $\mathbf{q}$ is a $(k \times 1)$ vector of transition intensities into the absorbing phase:

$$\mathbf{q} = -\mathbf{Qe} = (\mu_1, \mu_2, \ldots, \mu_k)^T. \tag{6}$$

Marshall and McClean [11] outlined a method for calculating the probability of absorption from each of the transient phases. Let π_i represent the probability that an individual departs from the ith phase of the distribution, then $\pi_1, \pi_2, \ldots, \pi_k$ can be calculated as follows:

$$\pi_k = \frac{\mu_k}{\lambda_k + \mu_k} \left\{ \prod_{i=1}^{k-1} \frac{\lambda_i}{\lambda_i + \mu_i} \right\}. \tag{7}$$

These probabilities represent the proportion of patients in the data which are absorbed from each phase. Patients can be separated into similar length of stay clusters grouped in the ratio $\pi_1 : \pi_2 : \ldots : \pi_k$. By using this technique, the length of stay groups can be analysed to determine whether patients in the same cluster possess similar attributes which might help to explain the pattern of their LoS.

The Coxian phase-type distribution has been identified as particularly suitable for representing LoS data [18]. Tang et al. [17] proposed the Coxian phase-type regression model as a method of assessing the influence of covariates upon rates of flow through the model. The regression coefficients are estimated through direct incorporation of parameters into the probability density function as follows:

$$f(t) = \mathbf{p} \exp[\exp\{-\mathbf{x_i}^T \boldsymbol{\beta}\}\mathbf{Q}t][\exp\{-\mathbf{x_i}^T \boldsymbol{\beta}\}\mathbf{q}], \tag{8}$$

where $\boldsymbol{x}_i$ is a vector of covariates for individual i and $\boldsymbol{\beta}$ is the corresponding vector of regression parameters.

3 Application

The data used in this study was obtained from the patient database of an ED within a large teaching hospital located in Adelaide, South Australia. The ED dataset consists of 119,306 patients who presented to the ED over a period of 20 months between 2012 and 2013 and contains clinical information (such as triage category and arrival method) and patient details (for example age and gender). Time-stamps indicating the progression of each patient through ED are also recorded, for triage, seen by doctor (first occasion), admission, and outcome for each patient. The total LoS in ED is calculated as the sum of three distinct stages, as outlined in Fig. 2.

The methodology is illustrated using the data for 'boarding' [15] stage 3 of ED LoS, which represents the delay experienced by patients who are waiting for an inpatient bed to become available for hospital admission. Boarding patients are a major

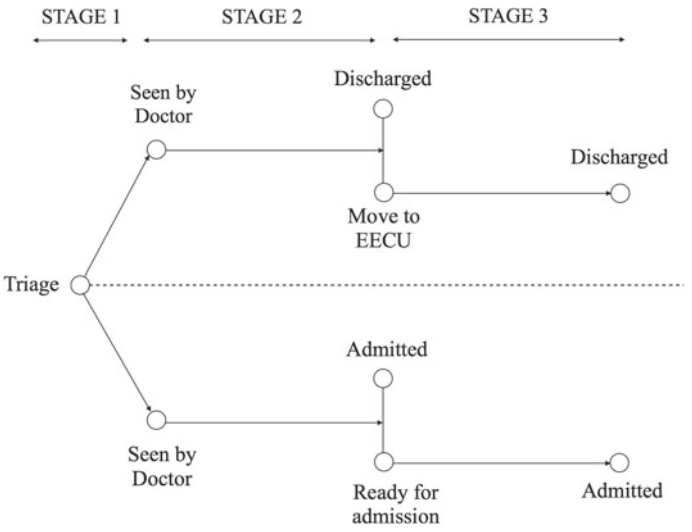


Fig. 2 Pathways through the ED as defined by the date and time stamps. Stage 1: waiting time between triage and initial consultation with a doctor. Stage 2: treatment time for each patient in the ED. Stage 3: any additional time spent in the ED. For admitted patients, this refers to 'boarding' [15] i.e. the time spent waiting for an inpatient bed to become available. For discharged patients this is time which in an extended emergency care unit (EECU)

contributor to overcrowding, as they continue to utilise resources before transfer to inpatient wards, thereby reducing the ability of the system to cater for new arrivals.

3.1 Analysing Length of Stay Groups from the Coxian Phase-Type Distribution

The LoS for boarding patients was fitted using Coxian phase-type distributions. The optimal order of Coxian phase-type distribution for representing a dataset is determined using a sequential fitting approach i.e. a one-phase distribution is initially fitted, then additional phases are added until little or no difference is observed in the model fit. Bayesian Information Criterion (BIC) is used for model selection, as it provides a good balance between goodness-of-fit and model parsimony. The fminsearch procedure in MATLAB was used to find the maximum likelihood value and parameter estimates, by employing the Nelder-Mead algorithm [13]. Table 1 displays the results of the optimal 6-phase model, which had the lowest BIC value Fig. 3 demonstrates that the 6-phase Coxian captures the shape of the empirical LoS data. The order of optimal Coxian distribution corresponds to the number of LoS clusters, where the probability of departure from each phase was calculated using Eq. (7) then utilised to group patients, as shown in Table 2.

Table 1 Coxian phase-type distributions fitted to emergency department LoS (admitted)

No. of phases	Fitted parameters		Log-likelihood	BIC
1	$\hat{\mu_1} = 0.0474$		−155,792.9	311,596.3
2	$\hat{\mu_1} = 2.99 \times 10^{-14}$	$\hat{\lambda_1} = 0.1597$	−151,168.5	302,368.7
	$\hat{\mu_2} = 0.0673$			
3	$\hat{\mu_1} = 0.0052$	$\hat{\lambda_1} = 0.3018$	−150,534.9	301,122.6
	$\hat{\mu_2} = 7.52 \times 10^{-212}$	$\hat{\lambda_2} = 0.0680$		
	$\hat{\mu_3} = 0.2896$			
4	$\hat{\mu_1} = 0.0063$	$\hat{\lambda_1} = 0.0616$	−149,625.1	299,324.1
	$\hat{\mu_2} = 3.61 \times 10^{-63}$	$\hat{\lambda_2} = 0.4249$		
	$\hat{\mu_3} = 0.0002$	$\hat{\lambda_3} = 0.4260$		
	$\hat{\mu_4} = 0.4262$			
5	$\hat{\mu_1} = 0.0076$	$\hat{\lambda_1} = 0.3350$	−149,462.4	299,019.8
	$\hat{\mu_2} = 0.0049$	$\hat{\lambda_2} = 0.0985$		
	$\hat{\mu_3} = 0.0001$	$\hat{\lambda_3} = 0.3408$		
	$\hat{\mu_4} = 0.3006$	$\hat{\lambda_4} = 0.0331$		
	$\hat{\mu_5} = 0.0287$			
6	$\hat{\mu_1} = 0.0059$	$\hat{\lambda_1} = 0.3636$	−149,427.0	**298,970.2**
	$\hat{\mu_2} = 0.0148$	$\hat{\lambda_2} = 0.4364$		
	$\hat{\mu_3} = 3.48 \times 10^{-55}$	$\hat{\lambda_3} = 0.0957$		
	$\hat{\mu_4} = 0.0182$	$\hat{\lambda_4} = 0.3449$		
	$\hat{\mu_5} = 0.9633$	$\hat{\lambda_5} = 0.0823$		
	$\hat{\mu_6} = 0.0260$			

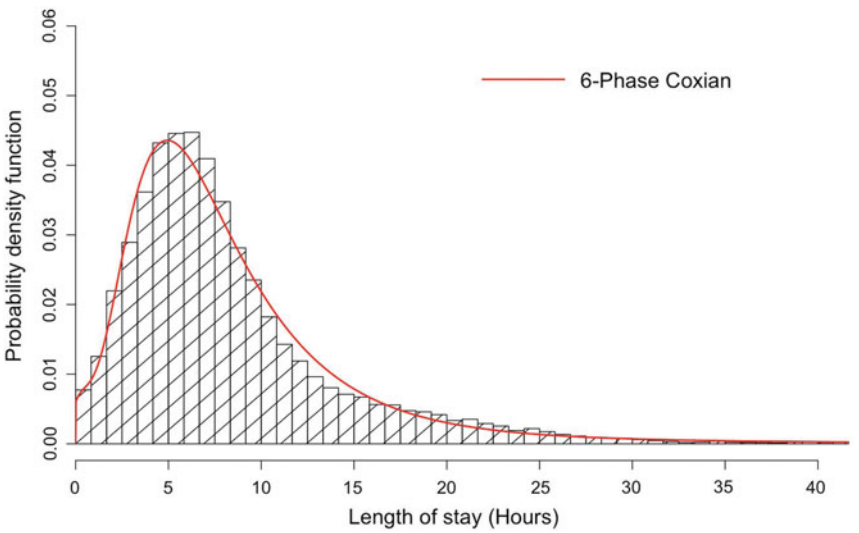


Fig. 3 Six-phase Coxian distribution plotted over empirical LoS data for admitted patients

Table 2 Probability of departure from each stage of the fitted six-phase Coxian distribution for admitted patients

Phase	S_1	S_2	S_3	S_4	S_5	S_6
Probability of departure	0.01607	0.03238	3.46×10^{-54}	0.04774	0.83265	0.07117

Most of the patients departed from the fifth phase of the model. The percentage of patients departing from each LoS group S_1 to S_6 were calculated by the variable 'primary complaint category', of which there are 98 illness categories, to identify differences between the extreme phases of the distribution. Patients departing from the first and last phase of the model have unique behaviour, and are consequently of primary interest in this study. Figures 4 and 5 display four complaint categories which had a particularly high proportion of patients depart from phases 1 and 6 of

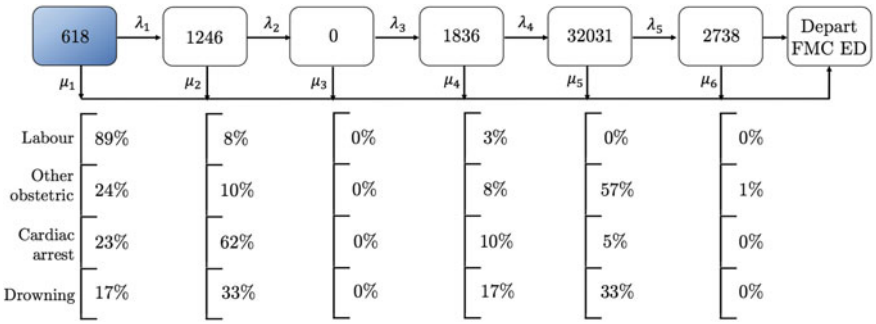


Fig. 4 Four complaint categories containing a relatively high proportion of patients that departed from phase 1 of the fitted Coxian distribution. The highest values were the 'labour', 'other obstetric', 'cardiac arrest', and 'drowning' categories

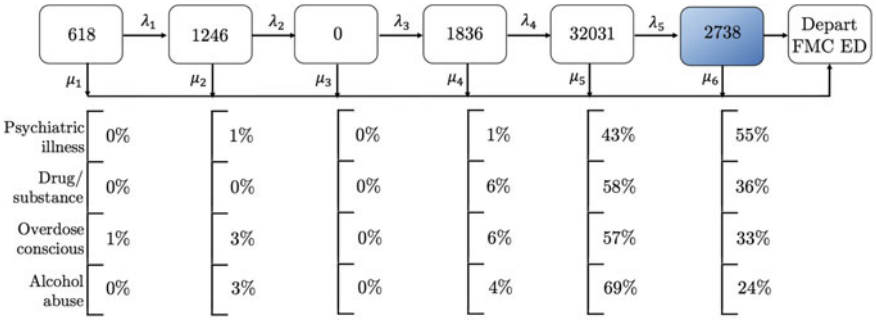


Fig. 5 Four complaint categories containing a relatively high proportion of patients that departed from phase 6 of the fitted Coxian distribution. The highest proportions were the 'psychiatric illness', 'drug/substance abuse', 'overdose unconscious', and 'alcohol abuse' complaints

the model respectively. These diagrams were utilised to communicate patient flow patterns with hospital management. It was identified that patients presenting in the psychiatric illness and substance abuse (PISA) categories had a protracted ED LoS.

3.2 Assessing the Effect of Patient and System Factors on LoS

Coxian phase-type regression models were fitted to the LoS stage 3 (boarding) in the ED data. The patients in each of five triage categories were considered separately, to identify the unique characteristics of each group. Table 3 displays the optimal Coxian phase-type regression model for each triage category.

A range of variables were identified as influential on the boarding time of patients using stepwise selection. Patients who presented to ED between 3 pm and 7 am had an increased boarding time over patients who presented between 7 am and 3 pm. Ambulance arrival was associated with an increased LoS for patients in triage category 2 only. PISA patients in triage categories 1–4 had an increased boarding time over other patient types.The inpatient division destination for patients was found to be influential on boarding time for triage categories 1–4, but not 5. Patients who were admitted to the PICU (mental health) division generally had longer boarding times than those in the reference group (general medicine). By contrast, patients who were admitted to the Women and Children's division generally had shorter boarding times than the reference group. Patients who were admitted to the Surgery division had a mixture of effects in comparison to the reference group. ED occupancy was included as a covariate in the regression model, as a proxy measure for overcrowding. Occupancy was significantly influential on ED LoS, with the exception of patients in triage category 5. The regression parameter estimates of 0.03 indicated that each additional boarding resulted in an increase of approximately 3% in the length of boarding time.

4 Discussion

This research has presented an application of the Coxian phase-type clustering method to group patients by LoS, and thereby identify characteristics of subjects who departed from the extreme distribution phases of the distribution. The techniques presented in this paper provide a new method to understand ED patient flow, by using both Coxian patient-flow diagrams and regression models to quantify the effect of patient and system factors on LoS, and in particular on ED boarding times. Such information can be used by health service managers to facilitate the development of improvement strategies that can be targeted towards the patients who are most likely to breach LoS targets, and to address system issues that may reduce time spent in ED. Additionally, the models can rapidly be re-fitted using new data, en-

Table 3 Coxian phase-type regression models fitted to stage 3 (boarding) of ED LoS (Inpatient division destination is abbreviated to DIV, time of day to TDAY, psychiatric illness and substance abuse to PISA, ambulance to AMB, and overcrowding to O)

Triage cat. (no obs)	No. of phases	Fitted estimates		
		Rates	Covariates	Log-likelihood
1 (1001)	3	$\hat{\mu}_1 = 0.0559$, $\hat{\mu}_2 = 0.2670$, $\hat{\mu}_3 = 0.0591$, $\hat{\lambda}_1 = 0.1288$, $\hat{\lambda}_2 = 0.0851$	$\hat{\beta}_{TDAY1} = 0.1960$, $\hat{\beta}_{TDAY2} = 0.3713$, $\hat{\beta}_{DIV1} = 1.6091$, $\hat{\beta}_{DIV2}$ =-0.1095, $\hat{\beta}_{DIV3}$ =-1.0685, $\hat{\beta}_O = 0.0263$	−3850.5
2 (7558)	3	$\hat{\mu}_1 = 0.0086$, $\hat{\mu}_2 = 2.99 \times 10^{-28}$, $\hat{\mu}_3 = 0.0793$, $\hat{\lambda}_1 = 0.8069$, $\hat{\lambda}_2 = 0.9302$	$\hat{\beta}_{AMB} = 0.0624$, $\hat{\beta}_{TDAY1} = 0.3053$, $\hat{\beta}_{TDAY2} = 0.2337$, $\hat{\beta}_{DIV1} = 1.2349$, $\hat{\beta}_{DIV2} = 0.0163$, $\hat{\beta}_{DIV3}$ =-0.8286, $\hat{\beta}_{PICU} = 0.3771$, $\hat{\beta}_O = 0.0286$	−29,640.5
3 (15346)	3	$\hat{\mu}_1 = 1.46 \times 10^{-08}$, $\hat{\mu}_2 = 0.1391$, $\hat{\mu}_3 = 0.0404$, $\hat{\lambda}_1 = 0.1785$, $\hat{\lambda}_2 = 0.0394$	$\hat{\beta}_{TDAY1} = 0.1925$, $\hat{\beta}_{TDAY2} = 0.2097$, $\hat{\beta}_{DIV1} = 1.2734$, $\hat{\beta}_{DIV2} = -0.0196$, $\hat{\beta}_{DIV3}$ =-0.6833, $\hat{\beta}_{PICU} = 0.3179$ $\hat{\beta}_O = 0.0283$	−59,952.0
4 (6450)	3	$\hat{\mu}_1 = 1.40 \times 10^{-09}$, $\hat{\mu}_2 = 0.1792$, $\hat{\mu}_3 = 0.0430$, $\hat{\lambda}_1 = 0.2248$, $\hat{\lambda}_2 = 0.0456$	$\hat{\beta}_{TDAY0} = -0.1281$, $\hat{\beta}_{TDAY2} = 0.1088$, $\hat{\beta}_{DIV1} = 1.3248$, $\hat{\beta}_{DIV2} = -0.1295$, $\hat{\beta}_{DIV3}$ =-0.7031, $\hat{\beta}_{PICU} = 0.3501$, $\hat{\beta}_O = 0.0262$	−24,459.1
5 (314)	4	$\hat{\mu}_1 = 0.0037$, $\hat{\mu}_2 = 0.0341$, $\hat{\mu}_3 = 0.2361$, $\hat{\mu}_4 = 0.0251$, $\hat{\lambda}_1 = 0.2787$, $\hat{\lambda}_2 = 0.2482$, $\hat{\lambda}_3 = 0.0463$		−1020.3

suring up-to-date analysis, and the opportunity for comparison of ED performance between multiple hospitals.

To ensure that health services benefit from this method a range of strategies are required to improve translation. Engagement between modellers and health service management staff, together with demonstration of uses and benefits of the approach is required. The ideal solution to ensure ongoing uptake of the method is to provide the health services management team with ongoing regular access to a modeller who can provide the necessary technical expertise as required. Future work will consider use of the models for obtaining accurate LoS predictions across different EDs and how the translation of the method into practice may be best facilitated.

Acknowledgements The authors wish to thank the Northern Ireland Department for the Economy for funding this research.

References

1. Aalen, O.: Phase type distributions in survival analysis. Scand. J. Stat. **22**, 447–463 (1995)
2. Australian Institute of Health and Welfare.: Emergency department care 2016-2017: Australian hospital statistics. In: Health services series no. 780 (2017). https://www.aihw.gov.au/reports/hospitals/emergency-department-care-ahs-2016-17. Cited 9th March 2018
3. Baothers, C., Owler, B., Grigg, M., et al.: Expert panel review of elective surgery and emergency access targets under the National Partnership Agreement on improving public hospital services. Report to the Council of Australian Governments (2011)
4. Chaou, C.-H., Chen, H.-H., Chang, S.-H., et al.: Predicting length of stay among patients discharged from the emergency department—using an accelerated failure time model. PloS one **12**(1) (2017)
5. Collett, D.: Modelling Survival Data in Medical Research. Chapman & Hall/CRC (2003)
6. Faddy, M.J.: Examples of fitting structured phase-type distributions. Appl. Stochast. Models Bus. Ind. **10**, 247–255 (1994)
7. Faddy, M., McClean, S.: Analysing data on lengths of stay of hospital patients using phase-type distributions. Appl. Stochchast. Models Bus. Ind. **15**(4), 311–317 (1999)
8. Hwang, U., McCarthy, M.L., Aronsky, D., et al.: Measures of crowding in the emergency department: a systematic review. Acad. Emerg. Med. **18**(5), 527–538 (2011)
9. Kreindler, S.A., Cui, Y., Metge, C.J., Raynard, M.: Patient characteristics associated with longer emergency department stay: a rapid review. Emerg. Med. J. **33**, 194–199 (2016)
10. Latouche, G., Ramaswami, V.: Introduction to matrix analytic methods in stochastic modeling. Society for Industrial and Applied Mathematics (1999)
11. Marshall, A.H., McClean, S.: Using Coxian phase-type distributions to identify patient characteristics for duration of stay in hospital. Health Care Manag. Sci. **7**, 285–289 (2004)
12. McCarthy, M.L., Ding, R., Pines, J.M., Zeger, S.L.: Comparison of methods for measuring crowding and its effects on length of stay in the emergency department. Acad. Emerg. Med. **18**(12), 1269–1277 (2011)
13. Nelder, J.A., Mead, R.: A simplex method for function minimization. Comput. J. **7**(4), 308–313 (1965)
14. Neuts, F.: Structured stochastic matrices of M/G/1 type and their applications. Probability: Pure and Applied. Taylor & Francis, New York (1989)
15. Richardson, D.B., Mountain, D., et al.: Myths versus facts in emergency department overcrowding and hospital access block. Med. J. Aus. **190**(7), 369 (2009)

16. Rose, L., Gray, S., Burns, K., et al.: Emergency department length of stay for patients requiring mechanical ventilation: a prospective observational study. Scand. J. Trauma Resuscitation Emerg. Med. **20**(1), 30 (2012)
17. Tang, X., Luo, Z., Gardiner, J.: Modeling hospital length of stay by Coxian phase- type regression with heterogeneity. Stat. Med. **31**, 1502–1516 (2012)
18. Vasilakis, C., Marshall, A.H.: Modelling nationwide hospital length of stay: opening the black box. J. Oper. Res. Soc. **56**, 862–869 (2005)
19. Zhu, T., Luo, L., Zhang, X., Shen, W.: Modeling the length of stay of respiratory patients in emergency department using Coxian phase-type distributions with covariates. IEEE J. Biomed. Health Inf. **22**(3), 955–965 (2018)

Emergency

A Realistic Simulation Model of Montreal Emergency Medical Services

Gabriel Lavoie, Valérie Bélanger, Luc de Montigny and Nadia Lahrichi

Abstract Emergency Medical Services (EMS) provide pre-hospital care and transportation to hospitals following an emergency call. Some EMS will also offer inter-hospital transportation services for patients. This paper presents a simulation model based on Urgences-santé, an EMS covering a population of 2.4 million people in Quebec (Canada). The goal of the simulation tool is to produce a highly realistic model that focuses on some of the lesser studied aspects of EMS management. These aspects include relocation, reroute-enabled dispatching, hospital selection, break management, complex priority system, ambulance specialization, and the integration of inter-hospital transport. The simulation tool allows us to measure the impact of changes to the rerouting and relocation policies, as well as to the fleet composition, on the system's performance using what-if analysis.

Keywords Discrete-event simulation · Emergency medical services · Inter-hospital transfers

G. Lavoie · N. Lahrichi
Department of Mathematical and Industrial Engineering,
Polytechnique Montreal, Montreal, Canada
e-mail: gabriel-2.lavoie@polymtl.ca

N. Lahrichi
e-mail: nadia.lahrichi@polymtl.ca

V. Bélanger (✉)
Department of Logistics and Operations Management, HEC Montreal, Montreal, Canada
e-mail: valerie.3.belanger@hec.ca

L. de Montigny
Urgences-santé Corporation, Montreal, Canada
e-mail: luc.demontigny@urgences-sante.qc.ca

Epidemiology, Biostatistics and Occupational Health, Faculty of Medicine,
McGill University, Montreal, Canada

V. Bélanger et al. (eds.), *Health Care Systems Engineering*,
Springer Proceedings in Mathematics & Statistics 316,
https://doi.org/10.1007/978-3-030-39694-7_6

1 Introduction

For many patients, Emergency Medical Services (EMS) are the point of entry into the health care system. From the moment they call for emergency services until they are successfully taken in charge at the hospital, their quality of care depends on their EMS provider. Among the performance measures used for the quality of care of EMS, Response Time (RT), which is defined as the time elapsed after a call has been assigned a priority up until an ambulance arrives on the scene, is one of the more commonly found measures cited in the literature [1–3]. It is notably used because of its important impact on mortality rates [4, 9].

The management of an EMS is a complex task. It requires the operation of an ambulance fleet, with its corresponding paramedics, and a dispatch center. In this center, a medical dispatcher evaluates the caller's condition and manages the fleet.

Many decisions must be taken into account in order to effectively manage the fleet. The most recent literature reviews separate these decisions into three categories based on their time horizon [1, 3, 8].

Strategic decisions usually impact the EMS for several years. These include choosing both the number of bases of operation, as well as their location. Deciding on the number of ambulances and paramedics is also considered a strategic decision. And finally, choosing performance objectives could also be included in this category [1].

At the tactical level, the two main decisions are the scheduling problem for ambulances and the location problem. The scheduling decision consists of choosing the start time and duration of the shift to ensure that the number of ambulances on the road is adequate for the frequency of calls. The location decision is about deciding the position of the standby site where the ambulance is waiting for calls. The position of these sites is important since adequate positioning ensures that the ambulance can promptly reach any part of the territory to answer calls. Location problems have been the main focus in the literature so far.

Operational decisions are those taken in the day-to-day management of the EMS. The relocation decision is closely linked to the location decision since it implies choosing where to send idle ambulances among the standby sites and, as such, has also been studied often. The dispatching decisions imply choosing available ambulances to send to each call. Most papers use the simplest approach of sending the closest available ambulance, however more sophisticated rules for dispatching have also been considered, although it is not a focus in the literature. For example, already assigned ambulances can be sent to another call thus ending their current assignation. This is referred to as a reroute-enabled-dispatch or simply "reroute" [7]. Finally, another operational decision is the hospital selection. Outside of the specific case of disaster response, most studies consider that patients are sent to the closest hospital. To the best of our knowledge, the only study to test other approaches outside of disaster response has been Lee [6].

In addition to serving emergency calls, some EMS also offer inter-hospital transportation services for patients [8]. Since most EMS do not include this service, it has been less studied. Among researches on the subject, Kergosien et al. [5] compared the

performance gained by using a single fleet of ambulances for both types of requests versus using dedicated resources.

By its nature, the request for emergency care is highly uncertain. This has led to an important part of the research done on EMS to use discrete-event simulation to account for this variability. Aboueljinane et al. [1] presents a detailed review of simulation models applied to EMS operations. Simulation models for EMS tend to focus on one or two decisions that are modeled in great detail, while keeping the other aspects of the EMS simple. While this paper does focus on operational level decisions, we aimed to minimize the number of simplifications done in order to make our model both very realistic and flexible.

The simulation model presented in this paper is based on Urgences-santé, a state-owned EMS system that covers a population of 2.4 million people in Quebec (Canada). Its territory covers the city of Montreal and its suburb, Laval. Their territories include several general and specialized hospitals. Some of these hospitals offer their services primarily in English while most do so in French. Urgences-santé's main mission is to respond to emergency calls made in its territory, but they will also transfer patients between health facilities when required. These transports represent around 10% of their activities.

The model incorporates many aspects of the Urgences-santé system. Notably, it includes Urgences-santé's dispatching rules for both emergency calls and inter-hospital transfers, including the possibility to reroute an ambulance. It also incorporates complex policies for relocation, hospital selection, paramedics' lunch break, and end of shift policies. Among these, re-routing and hospital selection have been studied in very few cases; while the impact of a lunch break and end of shift policies has never been considered, to our knowledge. Furthermore, it's one of the few models to consider ambulance specialization, in our case with three different types of ambulances.

The remainder is organized as follows. Next section describes the process at Urgences-Santé. The simulation model is detailed in Sect. 3 and results in Sect. 4. A discussion concludes this paper.

2 Process

In order to accurately model Urgences-santé's process, we consulted their internal documentation and observed their dispatch center operate for several hours. The resulting process maps were subsequently validated and enriched by a medical dispatch instructor working for Urgences-santé. Finally, the process maps were presented and validated with several of their executives. The five main components to understand the process are the priority system and ambulance fleet, the emergency calls (how they are handled), the inter-hospitals transfers, the relocation policies and the paramedic work shifts.

Table 1 Priority system summary

Priority	Type	911	Hybrid	Inter	Urgent	Description
0	Calls	Yes	Yes	Yes	Yes	High risk of cardial arrest
1	Calls	Yes	Yes	No	Yes	Immediate mortality risk
2	Transfer	Yes	Yes	Yes	Yes	High risk of morbidity or mortality
3	Calls	Yes	Yes	No	Yes	Risk of morbidity
4	Calls	Yes	Yes	No	No	Risk of morbidity in the next hours
5	Transfer	Yes	Yes	Yes	No	Immediate transfer
6	Transfer	Yes	Yes	Yes	No	Transfer with appointment
7	Calls	Yes	Yes	No	No	No identified risk of morbidity
8	Transfer	Yes	Yes	Yes	No	Non-emergency transfer

2.1 Priority System and Ambulance Fleet

Urgences-santé's priority system goes from 0 to 8, 0 being the most urgent. Table 1 summarizes the priority system. Their ambulance fleet is divided into three types of vehicles: 911 vehicles, which mostly respond to emergency calls; Inter vehicles, which mainly perform transfers; and Hybrid vehicles, which will do both types of requests with a focus on transfers. The priority of the request has several impacts on vehicle dispatching. First, it represents the relative priority of the requests. It also impacts which kind of vehicle can respond to each request. Priorities of 3 and under are considered urgent. For these priorities, paramedics use emergency driving protocols, which include the use of lights and sirens. Furthermore, a vehicle already assigned can be sent to a new urgent call if it's the closest available vehicle and if the current assignation is not as urgent. Finally, non-urgent requests cannot be dispatched using 911 ambulances if there are less than seven 911 ambulances available; this is explained in further detail in Sect. 2.4.

2.2 Emergency Calls

Emergency calls come from calls addressed to the city's 911 service which transfers them to Urgences-santé after an initial assessment. A medical dispatcher then talks

to the caller and assigns a priority to the call. If there's a vehicle available, one will be dispatched to the call. The vehicle will move to the call location and will provide pre-hospital emergency care to the patient. Unless the patient refuses transport or has left the location before the ambulance's arrival, the paramedics will communicate with a medical dispatcher who will then select an appropriate hospital. While patients can refuse transportation, the paramedics cannot, for legal reasons, advise them to refuse. We estimate that about 15% of interventions will not lead to a need for transport. Following this step, the patient is transported to the selected hospital where he is subsequently transferred. The ambulance becomes available to take another call once the transfer is finished.

Once the transport is confirmed, the hospital selection takes places. It takes into consideration multiple variables including the patient's condition, age, spoken language, and medical history. These factors ensure that each patient is sent to a hospital equipped for his needs. In cases where more than one hospital could be suitable, the patient is sent to the closest one. In some cases, patient preferences will also be considered. Furthermore, Urgences-santé has an allocation agreement with the hospitals in its territory, to an agreed upon number, under which they agree to limit both the percentage of patients sent to each hospital and the number of patients sent in a single hour to the same hospital. For some of the lower-urgency cases, Urgences-santé will select the hospital in order to respect these numbers.

2.3 Inter-hospital Transfers

Inter-hospital transfers cover cases where the patient is already in a medical facility and needs to be transferred to a new one. Requests for transfers are sent by the hospital to Urgences-santé. Some requests will be made by appointment before the actual request, while others will be made immediately before the transfer is needed. Requests for immediate transfer are considered urgent if the patient's condition is at risk of deteriorating. This category of requests also includes transfer to and from an air ambulance. In all cases, transfers are treated in a similar manner. An hour before the time of an appointment or when an immediate request is made, the dispatch center will try to assign an ambulance to the request. If a dispatch is possible, the ambulance will go to the hospital of origin where the patient will be transferred to the ambulance. The patient will then be transported to the hospital of destination where he will be dropped off. Afterward, the ambulance will immediately become available for another transport request or call. Contrary to emergency call response, the probability of cancellation is very low.

2.4 Relocation Policies

Urgences-santé's territory includes 34 waiting stations where idle 911 ambulances can be sent. These stations are divided into 3 groups, depending on their importance. Each station usually contains no more than one ambulance. Medical dispatchers prioritize the most important stations over the others. Ambulances are not usually moved from station to station. If at least 7 ambulances are stationed, the territory is considered to be covered by Urgences-santé, regardless of which locations are occupied in its territory. This coverage means that there should be enough idle ambulances in the territory to ensure a fast response to new high-urgency calls. If this condition is not met, 911 ambulances cannot be used to respond to low-urgency requests and will instead be sent to one of the empty waiting stations. Hybrid and Inter ambulances are not affected by the coverage condition and will not be sent to a waiting station. If they are idle, they will, instead, stay close to one of the major hospitals.

2.5 Paramedic Work Shifts

Each Urgences-santé ambulance is manned by two paramedics. At the start of their work shift, they leave one of Urgences-santé's three operational centers and become available to respond to emergency calls and transfer requests. The length of the paramedic shift is generally between 8 and 12 h. Outside of their normal work, two events happen during the shift: their lunch break and the end of shift period.

The lunch break varies from 30 to 60 min. It is not taken at an exact time. Instead, after a set delay, each paramedic team becomes eligible for their lunch break. At this point, they will be given their break if there is no unassigned request of priority 0–3. Before the actual start of the break, there is a 15-min pre-break period. During this period, the paramedic can freely move their ambulance but are still considered available for priority 0 calls and in some cases priority 1 and 2 as well. It is used by paramedics to reach the desired location of their lunch break. If a paramedic team has not taken their lunch break past the middle of their shift, the condition for their break will change to only require no unassigned requests of priority 0–2. This ensures that paramedics can take their breaks when the system is overloaded.

The end of shift period starts 45 min before the actual end of shift. During this period, the ambulance stops responding to low priority requests and will try to get closer to its starting operational center. As time passes, the range of priorities not covered increases. At the actual end of shift, the ambulance stops responding to all priorities and the paramedics exit the vehicle if they've reached the operational center. This approach reduces the occurrence of overtime by making paramedics less likely to be dispatched close to the end of their shift. Despite these policies, overtime is a common occurrence, with an average value per shift of around 20 min.

3 Simulation Model

The simulation model is implemented using Arena Simulation. The two main entities of the models are ambulances and patients. For now, parameters for both entity types, as well as the other distributions in the model, are based on estimates and do not follow Urgences-santé's data rigorously.

Patient arrivals follow a Poisson distribution whose parameter changes by the hour of the day. The Poisson distribution is the most popular distribution for patient arrival found in the literature [1]. In total, around 820 patients are created each day. Patients are divided into two mains categories: emergency call patients and transfer patients. The attributes linked to each patient will differ based on these categories. For emergency calls, patient attributes include their priority, location, condition, and whether they accept to be transported. For transfer requests, attributes include priority, hospital of origin, and hospital of destination. Refusal of transport for transfer is not taken into account since it is uncommon. The priorities used in the model are the same as those used by Urgences-santé.

Ambulance arrivals follow a fixed frequency for each hour of the day. In total, around 160 ambulances are created each day. Ambulance attributes include their type, the length of their shift, the length of their lunch break and an identifier for their operational base.

The simulation model uses the same decision rules as those used by Urgences-santé and described in Sect. 2. More specifically, the model uses the full set of dispatching rules of Urgences-santé, including rerouting rules. The processes for emergency calls, including hospital selection and inter-hospital transport are as described in Sects. 2.2 and 2.3 respectively. Finally, the paramedics' lunch break and end of shift period are modeled.

The transportation model used to calculate travel time of ambulances in the simulation divides Urgences-santé's territory into 950 zones of 775 m by 1110 m. This shape corresponds to an area of a hundredth of a degree latitude by a hundredth of a degree longitude for the Montreal area. Since in future work we plan to use Urgences-santé's data, and they have agreed to give us call coordinates whose position will be rounded down to 1/100 of a degree to ensure patients anonymity, we have decided to use this shape.

For each zone, a node has been located on either its main road or major intersection. A total of 3582 edges link these nodes to form a graph. Edges link neighboring nodes if there is no geographical obstacle between them. The weight of these edges was measured using Google's Distance Matrix API. Transportation time and shortest path between each pair of nodes were pre-calculated and integrated into the model. At the moment, the shortest path is static and does not change based on time. This approach allows tracking of an ambulance's position while it travels from one zone to another. This is necessary to correctly apply Urgences-santé's rerouting policy. It also offers good modeling of geographical obstacles, which is important since Montreal and Laval are separated by a river. The transportation model did not account for the use of lights and sirens.

4 Experimentation and Results

With the simulation model, we tested several scenarios where we modified some of Urgences-santé's policies. The scenario goals are to reduce Response Time (RT) for low-urgency calls and transfers, both of which are very high, without compromising high-urgency RT. Currently, the system has 4 main mechanisms in place to ensure short RT for high-urgency: (1) The priority system ensures that the higher the urgency, the faster an ambulance is sent; (2) most ambulances are type 911 and thus focus on emergency calls, most of which are high-urgency; (3) the coverage system prevents dispatching for low-urgency requests if not enough ambulances are available; and (4) high-urgency can reroute ambulances sent to low-urgency requests. The scenarios serve to test if some of these mechanisms can be modified without increasing high-urgency RT significantly. Since high-urgency requests have more risks of mortality and morbidity, no significant increase of RT is justifiable even with the decrease to low-urgency RT.

Besides a baseline scenario, we tested 8 scenarios. The **Baseline scenario** uses the same set of rules as Urgences-santé, as described in Sect. 3. It serves as a point of comparison for the other scenarios. **Scenario 1** increases the percentage of the fleet made of hybrid vehicles. **Scenario 2** uses the same approach to increase the number of Inter ambulances. **Scenario 3** combines both previous scenarios. **Scenario 4** modifies the re-dispatching rules by limiting their application to priority 0–2 instead of 0 to 3. **Scenario 5 to 8** test alternative requirements for coverage, ranging from 3 to 6 stationed 911 ambulances.

Table 2 presents the scenario results. Response Times are expressed in minutes. The confidence intervals use a percentage value of 95%. RT were divided into four categories: high-urgency covering requests of priority 0–3; low-urgency range from priority 4 to 8; emergency calls covering priorities 0, 1, 3, 4 and 7; and transfer requests including priorities 2, 5, 6 and 8. For transfer, RT includes two elements: the time difference between the appointment and the arrival time of the ambulance for transfer requests made by appointment; and the RT for immediate transfer requests. Reported results are based on 20 replications of 30 days. Each replication had a warm-up period of 5 days. Total computation time per scenario is between 8 to 10 min.

Scenarios 1–3 focus on inter-hospital since they add additional Hybrid and Inter ambulances to the fleet, both of which are specialized in patient transfers. Scenarios 2 and 3 reduce Inter RT significantly at the cost of an increase to high-urgency RT, while also positively impacting low-priority RT.

Scenario 4 positively impacts the RT of every category except for the high-urgency, which increases by about 50 s. In its current state, we found that ambulances assigned to priority 4 and higher have difficulty reaching their destination without being rerouted to a higher urgency request. While re-dispatching help reduces RT for high-urgency, it is also a cause of unproductivity since the time spent by ambulances while they travel to a low-urgency call is wasted if they are re-dispatched to another request. Scenario 4's main benefit is to reduce this loss of time. The increase to high-urgency RT for this scenario should be carefully considered since the rule

Table 2 Scenarios features and impact on responses times

Name	Scenario features					Responses times by categories			
	Fleet composition			Coverage criteria	Rerouting for P3	High urgency	Low urgency	Calls	Transfer
	911 (%)	Hybrid (%)	Inter (%)						
Baseline	92	4	4	7	Yes	12.0±0.0	165.9±27.6	23.0±1.2	180.2±36.0
#1	88	8	4	7	Yes	12.3±0.0	144.3±25.2	22.5±1.2	151.8±33.6
#2	88	4	8	7	Yes	12.8±0.0	89.1±6.0	22.0±0.6	74.6±7.2
#3	84	8	4	7	Yes	13.1±0.0	102.9±7.8	23.6±0.6	89.0±10.2
#4	92	4	4	7	No	12.8±0.0	75.9±7.8	18.9±0.6	72.8±9.0
#5	92	4	4	6	Yes	12.2±0.0	155.4±30.0	22.8±1.2	167.7±40.2
#6	92	4	4	5	Yes	12.2±0.0	152.4±25.8	22.4±1.2	165.4±34.2
#7	92	4	4	4	Yes	12.3±0.0	133.1±15.0	22.2±0.6	138.2±19.8
#8	92	4	4	3	Yes	12.5±0.0	143.6±21.6	22.6±1.2	151.4±28.8

change only affects priority 3, the least urgent priority among its group, and one for which no risk of mortality is identified.

Scenarios 5–8 reduce the number of 911 ambulances that need to be available before ambulances can be dispatched to low-urgency demands. Each reduction of the coverage criteria from 7 to 4 has a positive impact on every category except high-urgency, which steadily increases for each reduction. A further reduction to the criteria, to a value lower than 4, leads to an increase in response time in every category. This can be explained by a longer travel time as fewer ambulances cover the territory when the coverage criteria is reduced. The increase in travel time when reducing the coverage to 3 seems to exceed the benefit of a lower criteria for low-urgency requests, resulting in a general deterioration in the system's performance.

Overall, none of the scenarios manage to decrease the low-urgency or transfer RT without increasing high-urgency RT. While changing the relocation policies or the fleet distribution leads to major changes on the system's performance, modifying the coverage criteria has a more moderate impact.

5 Discussion

In this paper, we present a simulation model that aims to replicate, in detail, the inner-working of an Emergency Medical Service (EMS) which provides both assistance and transport for emergency calls and inter-hospital transport services. Before building the model, we consulted the EMS internal documentation and spent time observing their activities to ensure appropriate modeling. The resulting model includes many aspects of the EMS that are often left out of simulation studies.

The scenarios tested aim to assess potential improvement in the performance of the EMS by making changes to some of its current policies. The first 3 scenarios show that changing the ratio between specialized ambulances can have a vast impact on overall performance. In one scenario, the proposed change leads to a reduction in the average RT for inter-hospital transfer by more than an hour and a half, at the cost of an increase of 1.1 min to high-urgency RT. While the increase to high-urgency RT is not justifiable, the scenarios show the impact of these ratios on the system. Another scenario modifies the rerouting policies by making them more restrictive. This results in an improved RT for lower-urgency requests at the cost of an increase in high-urgency RT. This increase was, however, solely caused by an increase in RT for priority 3, the lowest priority of its group. The last four scenarios show that modifying the redeployment policies can improve mean RT for low-urgency, emergency call, and transfers up to a certain point while increasing the high-priority RT. While the improvement of RT for low-urgency and transfer offered by these scenarios seems promising, they do not necessarily justify the increase in high-urgency RT regardless of how small they might seem.

These results show that modeling the details of an EMS can help us understand the relation between the system rules and the resulting response times for different categories of requests. It also allows estimation of the trade-off between the RT of different categories when trying to improve one of them. The modifications tested in these scenarios would all be relatively simple to implement for Urgences-santé and could be implemented for no cost other than some training for the emergency dispatchers. Applying a similar approach to other EMS might provide solutions to improve other EMS' specific situations without requiring any major changes.

Furthermore, we find that the work required to accurately model an EMS system helps to build a better understanding of its operations and can provide insight into some lesser-studied aspects of EMS management. For example, shift management proved to be an important part of Urgences-santé day-to-day operations. Together, the pre-break period and end of shift period reduce the availability of vehicles for an hour per shift. Optimizations in those areas could possibly improve overall system performance, but—to the best of our knowledge—has never been studied.

In future work, we plan to improve the modeling of the request generation using historical data and add out-of-territory transports to the model. This will allow for direct validation of the model by comparing it to Urgences-santé's actual metrics. Once validated, the proposed scenarios could be measured more accurately in regard to their impacts on the Urgences-santé system. Additional scenarios must be considered since the hospital selection policy of Urgences-santé is complex and offers opportunities for improvement. The impact of changes to the break policies and end of shift policies should also be considered. It would also be possible to evaluate, and potentially improve, the location of Urgences-santé's waiting stations. However, this problem would require the use of different tools, in addition to simulation. Finally, additional performance indicators might give a better understanding of the impact of the scenarios.

References

1. Aboueljinane, L., Sahin, E., Jemai, Z.: A review on simulation models applied to emergency medical service operations. Comput. Ind. Eng. **66**(4), 734–750 (2013)
2. Aringhieri, R., Bruni, M.E., Khodaparasti, S., Van Essen, J.T.: Emergency medical services and beyond: addressing new challenges through a wide literature review. Comput. Oper. Res. **78**, 349–368 (2017)
3. Bélanger, V., Ruiz, A., Soriano, P.: Recent optimization models and trends in location, relocation, and dispatching of emergency medical vehicles. Eur. J. Oper. Res. **272**(1), 1–23 (2019)
4. Burger, A., et al.: The effect of ambulance response time on survival following out-of-hospital cardiac arrest: an analysis from the German resuscitation registry. Deutsches ärzteblatt Int. **115**(33–34), 541 (2018)
5. Kergosien, Y., Bélanger, V., Soriano, P., Gendreau, M., Ruiz, A.: A generic and flexible simulation-based analysis tool for EMS management. Int. J. Prod. Res. **53**(24), 7299–7316 (2015)
6. Lee, S.: The role of hospital selection in ambulance logistics. IIE Trans. Healthc. Syst. Eng. **4**(2), 105–117 (2014)

7. Lim, C.S., Mamat, R., Braunl, T.: Impact of ambulance dispatch policies on performance of emergency medical services. IEEE Trans. Intell. Transp. Syst. **12**(2), 624–632 (2011)
8. Reuter-Oppermann, M., van den Berg, P.L., Vile, J.L.: Logistics for emergency medical service systems. Health Syst. **6**(3), 187–208 (2017)
9. Wilde, E.T.: Do emergency medical system response times matter for health outcomes? Health Econ. **22**(7), 790–806 (2013)

A Two-Phase Approach to the Emergency Department Physician Rostering Problem

Paola Cappanera, Filippo Visintin and Roberta Rossi

Abstract In this study, we address the physician rostering problem occurring in an Emergency Department of an Italian pediatric hospital. Motivated by the paramount importance that workload balance has in this setting, we propose a tailored two-phase approach and we present two optimization models on which the proposed approach is based. In the first phase, we assign all the weekend (and holidays) shifts to physicians in a medium-term planning horizon pursuing a fair distribution of weekend and night shifts among the physicians, whereas in the second phase, we assign all the weekday shifts to physicians in short-term planning horizons so that each physician works almost the same number of morning and afternoon shifts. We present preliminary results of an ongoing research whose ultimate goal is to develop a decision support system to facilitate the creation of physicians' rosters.

Keywords Physician rostering · Workload balance · Emergency department · Optimization

1 Introduction

For every organization, satisfying customers' needs requires having the right staff on duty at the right time. In Emergency Department (ED), where the service being delivered saves human life, relying on a well-designed physician roster is of paramount importance.

In general, solving a staff rostering problem involves building a work schedule that: (i) allows meeting a time dependent demand for service, (ii) complies with

P. Cappanera · R. Rossi
Dipartimento di Ingegneria dell'Informazione, University of Florence, Florence, Italy
e-mail: paola.cappanera@unifi.it

R. Rossi
e-mail: rorossi@unifi.it

F. Visintin (✉)
Dipartimento di Ingegneria Industriale, University of Florence, Florence, Italy
e-mail: filippo.visintin@unifi.it

V. Bélanger et al. (eds.), *Health Care Systems Engineering*,
Springer Proceedings in Mathematics & Statistics 316,
https://doi.org/10.1007/978-3-030-39694-7_7

regulatory constraints and work-place agreements, (iii) attempts to satisfy individual staff constraints and preferences [7].

Effectively addressing personnel management is very complex. Consequently, the problem is typically split in two sub-problems [7]: the first one, referred to as *staff dimensioning*, involves the determination of the number and type of staff needed to meet the demand for each time-shift. The second one, referred to as *rostering*, instead, involves the assignment of individual staff members to each shift to meet demand.

In this study, we consider the latter sub-problem within the context of an emergency department. This will be thus referred to as Emergency Department Physician Rostering Problem (EDPRP). In particular, we assume that the number of physicians to assign to each work-shift in a medium-term planning horizon (6 month) is known and we determine the assignment of physicians to shifts in a way that maximizes personnel's satisfaction and perceived equity.

In emergency departments—including the one inspiring this study—the number of physicians covering each shift is typically determined once a year. Such a decision takes into account the seasonality of the demand (which is quite significant in pediatric ED, see [15]), involves a negotiation between hospital management, trade unions and local government, and usually drives hiring policies. Once the staff dimensioning is determined, the hospital management iteratively solves the EDPRP to determine the physician schedules for the incoming weeks. This is a very difficult task. It requires, in fact, not only to comply with articulated regulations (defining, for example, the maximum and/or minimum number of daily, weekly and monthly hours that a physician can work, the minimum number of days between two consecutive night shifts and so on) but, also, to deal with the fact that, given the seasonality of the ED arrival rates and the fact that ED are open 24/7, certain shifts (e.g. weekend and night shifts) are less desirable than others. Moreover, when building the schedule, it is necessary to continuously deal with physicians' requests that can be strict and thus managed as hard constraints (e.g. when the physician is unavailable because s/he has already assigned to other hospital duties) or expressed as preferences and thus managed as soft constraints (e.g. when s/he asks for a day-off for personal motivations). In the hospital under study, this task is performed by two seasoned physicians, with enough authority to deal with colleagues' pressures.

In this paper we propose a two-phase approach to the EDPRP and present two Integer Linear Programming (ILP) models that allow implementing the proposed approach. In the first phase, we assign all the weekend (and holidays) shifts to physicians in a medium-term planning horizon (e.g. 6 months). In the second one, instead, we assign all the weekday shifts to physicians in short-term planning horizons (e.g. one month). In this phase, we consider the weekend shifts as already assigned, and we assign the remaining ones. The optimal solution of the first phase, thus, is given in input to the model used in the second phase. Such an approach ensures that: (i) physicians know their weekend shifts well in advance and (ii) physicians' requests, if communicated reasonably in advance (e.g. one month ahead), can be accommodated without incurring in schedule disruptions.

The proposed approach, thus, aims at increasing personnel's satisfaction while at the same time simplifying the rostering efforts. The objective functions of our

models are defined in a way that maximizes the perceived shift equity. Ensuring shift equity, in fact, is proven to influence staff morale and motivation [2, 4]. Specifically, the objective function used in the first phase allows to evenly distribute weekend and night shifts among physicians. The one used in the second phase, instead, allows to evenly distribute morning and afternoon shifts. As pointed out earlier, the shifts assigned to each physician in the first phase are considered as preassigned shifts in the second one.

This paper reports the preliminary results of an ongoing research whose ultimate goal is to develop a decision support system to help ED physicians developing staffing rosters.

The manuscript is organized as follows. Section 2 briefly reviews the literature with a specific focus on those contributions related to physician rostering problems in which workload balance plays a pivotal role. Then, in Sect. 3 the two optimization models proposed to formulate the rostering problem are presented emphasizing their common structure. Section 4 reports the preliminary results obtained when facing the rostering problem at the ED of an Italian pediatric hospital, while Sect. 5 concludes the paper.

2 Literature Review

Personnel scheduling is a wide spread problem which has been extensively studied in the literature [8, 13]. One of the most challenging setting to address personnel scheduling problems is the healthcare one. Here, in fact, the problem is complicated by the fact that personnel, especially physicians, are usually understaffed, burnout exposed, specialized and thus difficult to replace [6].

Moreover, developing personalized schedules (rosters) taking into consideration physician preferences and constraints (most of which are related with the different services that physicians provide both within and outside the hospital) is of paramount importance, as it drives staff satisfaction and service quality. In addition, physicians are resources usually shared by different wards and their rosters in the ED need to consider the activities already assigned to them in other wards as preassigned activities [14]. Thus, the management of preferences and preassigned activities, and the plurality of activities performed by physicians make the physician rostering problem different from other rostering problems addressed in the healthcare setting. The physician rostering problem, however, has received limited research attention to date [1]. In the literature, there is consensus [1, 11, 17] that when dealing with physician rostering problems, it is necessary to ensuring shift equity by balancing personnel workload. Workload balance is usually pursued while considering other patient-related constraints. As an example, Adams et al. [1] deal with workload balance while taking into consideration the continuity of care that patients experience.

Several papers (e.g., [9]) propose cyclic roster solutions to ensure workload balance. However, approaches based on cyclic roster make it difficult to satisfy individual preferences [10]. Indeed, in cases in which individual requests are manifold and diversified among physicians—such as in our case study—cyclic approaches are not suitable.

In addition, the management of preferences makes the problem computationally difficult to solve. This computational complexity has often led to adopt two-phase approaches in which shifts are organized in two groups (for example working shifts and rest shifts) assigned separately in each phase. Typically, rest days are assigned in the first phase, whereas working shifts are assigned in the second phase assuming the rest shift as preassigned activities [12, 17]. In other studies, the complexity introduced by constraints ruling workload equalization is managed by relaxing the complicating constraints. Then, either they are penalized as typically done in metaheuristic approaches [16] or introduced dynamically in the relaxed problem as done in branch-and-cut algorithms [5].

In this study, we address the physician rostering problem and we design a two-phase approach to ensure workload balance. Indeed, in our setting, equity has two dimensions and physicians ask that (i) weekend and night shifts are fairly assigned in a medium-term planning horizon, (ii) each physician works almost the same number of morning and afternoon shifts in a medium-term planning horizon. The proposed decomposition algorithm reflects in each phase the peculiarities of a specific type of balancing and it is not intended to reduce computational complexity. In addition, the workload balancing is explicitly incorporated in the objective function. We are not aware of similar approaches in the literature.

3 The Optimization Models

In modeling the two rostering problems characterizing the two phases above described, we exploit their common structure. Indeed, each of the two problems has its peculiar constraints, whereas they both share the same network structure. Thus, both of them are modeled as multicommodity flow problems. This approach has been already investigated in the literature for other kinds of rostering problems [3]. Specifically, a layered network is used in each phase. In each of the two phases, the layered network is characterized by a level for each day in the planning horizon, and the nodes in each level correspond to the shifts that must be covered in that day. Each physician corresponds to a commodity and the sequence of activities assigned to a given physician identifies a path in the layered network from a source node to a destination node. The layered network is characterized by three sets of arcs: (i) the first set of arcs connect the source node to all the nodes in the level corresponding to the first day of the planning horizon; (ii) the second set of arcs connect two intermediate and consecutive layers in the network, and (iii) the third set of arcs connect the nodes in the day corresponding to the last day of the planning horizon to the destination node. The resulting network is acyclic and for each physician exactly one shift—either a work shift or a rest shift, has to be assigned on each day of the planning horizon (there are specific situation in which a physician can work more than one shift per day, if that is the case, the graph is modified accordingly by adding arcs linking shift nodes within the same day). Thus, for a given physician, a path from the source to the destination node visits exactly one node in each layer of the

network and the node visited in a specific day represents the shift assigned to him on that day. The multicommodity structure allows to consider a personalized network for each physician and to consider implicitly compatibility constraints between two consecutive shifts: as an example, if two shifts cannot be assigned consecutively to a physician, in the corresponding network the arc between the two shifts is not inserted for any couple of consecutive days in the planning horizon. There are two additional issues which are common to both the models: (i) the management of desiderata expressed by physicians, and (ii) the management of undesirable sequences of shifts. Concerning desiderata, a physician may ask that a certain activity is assigned on a certain day or not assigned, and that his request is strict (the request has to be satisfied) or granted only if it does not deteriorate the optimal solution. In the former case the desiderata are treated as hard constraints, whereas in the latter case they are managed as soft constraints, i.e., they are penalized when not granted and the total number of not satisfied desiderata is minimized in the objective function. On the other hand, the assignment of two activities one after the other is discouraged in the objective function for any sequence belonging to the set of undesirable sequences.

Let us denote with:

H	set of physicians
$\overline{H} \subseteq H$	subset of temporary staff
S	set of shifts
$S_m \subseteq S$	subset of morning shifts
$S_a \subseteq S$	subset of afternoon shifts
$S_n \subseteq S$	subset of night shifts
$S_r \subseteq S$	subset of rest shifts
$S_s \subseteq S$	subset of night-call duty shifts
$U = \{(i, j) \text{ s.t. } i, j \in S\}$	set of undesirable couples of consecutive shifts
$D = \{1, \ldots, \lvert D\rvert\}$	set of days to be considered—extended planning horizon consisting of $\lvert D\rvert$ days
$L = D \cup \{0, \lvert D\rvert + 1\}$	set of levels
$\overline{D} \subseteq D$	subset of days corresponding to the planning horizon (possibly different from D due to extension on the left and on the right)
$M \subseteq D$	set of days corresponding to Mondays
W	set of weekends in the planning horizon—each $w \in W$ is a subset of D
$G^h = (N^h, A^h)$	graph relative to physician h
$o^h \in N^h$	origin node for physician h; by default o^h belongs to level 0
$d^h \in N^h$	destination node for physician h; by default d^h belongs to level $\lvert D\rvert + 1$
$\Delta^h_{v,t} = \{(i, l) \text{ s.t. } i \in S, l \in D\}$	with $v \in \{0, 1\}, t \in \{P, F\}$ set of desiderata for physician h, expressed as couples of activity-day which have to be avoided ($v = 0$) or done ($v = 1$) in a soft ($t = P$) or hard way ($t = F$)
$\overline{n}^h$	maximum number of night shifts physician h can work
$\overline{w}^h$	maximum weekly workload for physician h—expressed in hours
$\overline{m}^h$	maximum monthly workload for physician h—expressed in hours
$\overline{s}^h$	monthly number of night-call duties for physician h
$\overline{b}^h$	monthly number of working hours for temporary staff
$\underline{d}$	minimum number of days that must elapse between two night shifts
c_{il}	number of physicians required in day l on duty i
w_i	workload of shift $i \in S$
$\overline{M}$	big-M value
α_c	weight used in the objective functions to discriminate criterion c.

Then, let us define the following families of variables in order to model the decisions:

$$x^h_{iljl+1} = \begin{cases} 1 \text{if shift } i \text{ in level } l \text{ and shift } j \text{ in level } l+1 \text{are assigned consecutively to physician } h \\ 0 \text{ otherwise} \end{cases}$$

$h \in H, l \in L, (i_l, j_{l+1}) \in A^h,$

$$\varepsilon^h_{il} = \begin{cases} 1 \text{ if the preference expressed by physician } h \text{ for shift } i \text{ on day } l \text{ is not satisfied} \\ 0 \text{ otherwise} \end{cases}$$

$h \in H, l \in L, i \in S$

In the following, we first describe, in terms of variables and constraints the common features of the two rostering problems and then their distinguishing features. Using the variables and notation above, the constraints common to both the models are the following:

$$\sum_{j \in S} x^h_{o^h 0 j 1} = 1 \qquad \forall h \in H \tag{1}$$

$$\sum_{j \in S} x^h_{j|D|d^h|D|+1} = 1 \qquad \forall h \in H \tag{2}$$

$$\sum_{j \in S \cup o^h} x^h_{jl-1il} - \sum_{j \in S \cup d^h} x^h_{iljl+1} = 0 \qquad \forall h \in H, \forall l \in D, \forall i \in S \tag{3}$$

$$\sum_{h \in H} \sum_{j \in S \cup o^h} x^h_{jl-1il} \geq c_{il} \qquad \forall l \in \overline{D}, \forall i \in S \tag{4}$$

$$\sum_{j \in S \cup o^h} x^h_{jl-1il} = 0 \qquad \forall h \in H, \forall (i,l) \in \Delta^h_{0F} \tag{5}$$

$$\sum_{j \in S \cup o^h} x^h_{jl-1il} = 0 + \varepsilon^h_{il} \qquad \forall h \in H, \forall (i,l) \in \Delta^h_{0P} \tag{6}$$

$$\sum_{j \in S \cup o^h} x^h_{jl-1il} = 1 \qquad \forall h \in H, \forall (i,l) \in \Delta^h_{1F} \tag{7}$$

$$\sum_{j \in S \cup o^h} x^h_{jl-1il} = 1 - \varepsilon^h_{il} \qquad \forall h \in H, \forall (i,l) \in \Delta^h_{1P} \tag{8}$$

$$x^h_{iljl+1} \in \{0,1\} \qquad \forall h \in H, \forall l \in L, \forall (i_l, j_{l+1}) \in A^h \tag{9}$$

$$\varepsilon^h_{il} \in \{0,1\} \qquad \forall h \in H, \forall v \in \{0,1\} \forall (i,l) \in \Delta^h_{vP} \tag{10}$$

The x^h_{iljl+1} variables allow to design the schedule for each physician h: for each physician, flow conservation constraints (1)–(3) impose that a path is determined from the origin o^h to the destination d^h visiting exactly one node (one shift) in each layer (each day). Constraints (4) guarantee demand coverage, while constraints (5)–(8) manage desiderata as hard constraints—(5) and (7), or as soft constraints—(6) and (8). Finally, the rest of the constraints impose the integrality of the variables.

The weekend shift management problem makes use of the following peculiar variables:

Y the maximum number of weekends on duty assigned to a physician
V the maximum number of night shifts assigned to a physician
$y_w^h = \begin{cases} 1 \text{ if physician } h \text{ is on duty in weekend } w \\ 0 \text{ otherwise} \end{cases} \qquad h \in H, w \in W.$

The peculiar set of constraints in the first phase are the following:

$$\min \quad \alpha_y Y + \alpha_v V + \alpha_x \sum_{l=1}^{|D|-1} \sum_{(i,j)\in U} x_{iljl+1}^h + \alpha_\varepsilon \sum_{h\in H} \sum_{v\in\{0,1\}} \sum_{(i,l)\in \Delta_{vP}^h} \varepsilon_{il}^h \tag{11}$$

$$\sum_{l\in\overline{D}} \sum_{i\in S_n} \sum_{j\in S\cup d^h} x_{iljl+1}^h \leq V \qquad \forall h \in H \tag{12}$$

$$\sum_{l\in w} \sum_{i\in S\setminus S_r} \sum_{j\in S\cup d^h} x_{iljl+1}^h \leq \overline{M} y_w^h \qquad \forall h \in H, \forall w \in W \tag{13}$$

$$\sum_{w\in W} y_w^h \leq Y \qquad \forall h \in H \tag{14}$$

$$y_{w_i}^h + y_{w_{i+1}}^h \leq 1 \qquad \forall h \in H, \forall i = 1, \ldots, |W| - 1 \tag{15}$$

$$y_w^h \in \{0, 1\} \qquad \forall h \in H, \forall w \in W \tag{16}$$

The first phase is guided by balancing criteria which hierarchically minimize the maximum number Y of weekends on duty and the maximum number V of night shifts among the physicians. The other two terms in the objective function discourage respectively undesirable sequences of activities and unsatisfied desiderata. Variable y_w^h takes value one (see 13), if h is assigned a work shift in any day of weekend w, i.e., if h is on duty in weekend w. Constraints (12) and (14) guarantee the correctness of the value assumed by variables Y and V, whereas constraints (15) prevent that a physician is on duty for two consecutive weekends. Constraints (16) define the y_w^h's domain.

The weekday shift management problem makes use of the following peculiar variable:

Z the maximum difference (in absolute value) of the number of morning and afternoon shifts among the physicians.

The peculiar set of constraints in the second phase are the following:

$$\min \quad \alpha_z Z + \alpha_x \sum_{l=1}^{|D|-1} \sum_{(i,j)\in U} x^h_{iljl+1} + \alpha_\varepsilon \sum_{h\in H} \sum_{v\in\{0,1\}} \sum_{(i,l)\in\Delta^h_{vP}} \varepsilon^h_{il} \tag{17}$$

$$\sum_{l\in\overline{D}} \sum_{i\in S} \sum_{j\in S\cup d^h} w_i x^h_{iljl+1} \le \overline{m}^h \qquad \forall h \in H \tag{18}$$

$$\sum_{l=d}^{\min(d+6,|D|)} \sum_{i\in S} \sum_{j\in S\cup d^h} w_i x^h_{iljl+1} \le \overline{w}^h \qquad \forall h \in H, \forall d \in M \tag{19}$$

$$\sum_{l\in\overline{D}} \sum_{i\in S_s} \sum_{j\in S\cup d^h} x^h_{jl-1il} \le \overline{s}^h \qquad \forall h \in H \tag{20}$$

$$\sum_{l\in\overline{D}} \sum_{i\in S} \sum_{j\in S\cup d^h} w_i x^h_{iljl+1} = \overline{b}^h \qquad \forall h \in \overline{H} \tag{21}$$

$$\sum_{l\in\overline{D}} \sum_{i\in S_n} \sum_{j\in S\cup d^h} x^h_{iljl+1} \le \overline{n}^h \qquad \forall h \in H \tag{22}$$

$$\sum_{l=d}^{l=d+\underline{d}} \sum_{i\in S_n} \sum_{j\in S\cup d^h} x^h_{iljl+1} \le 1 \qquad \forall h \in H, \forall d \in D \tag{23}$$

$$\sum_{l\in\overline{D}} \sum_{i\in S_a} \sum_{j\in S\cup d^h} x^h_{iljl+1} - \sum_{l\in\overline{D}} \sum_{i\in S_m} \sum_{j\in S\cup d^h} x^h_{iljl+1} \le Z \qquad \forall h \in H \tag{24}$$

$$-\sum_{l\in\overline{D}} \sum_{i\in S_a} \sum_{j\in S\cup d^h} x^h_{iljl+1} + \sum_{l\in\overline{D}} \sum_{i\in S_m} \sum_{j\in S\cup d^h} x^h_{iljl+1} \le Z \qquad \forall h \in H \tag{25}$$

As the first phase, also the second phase is guided by a balancing criterion which attempts at assigning to each physician almost the same number of morning and afternoon shifts. Specifically, for each physician the (absolute) difference between morning and afternoon shits assigned in the planning horizon is computed (see 24 and 25) and the objective minimizes the maximum of these differences among the physicians. The rest of the constraints guarantee, for each physician, respectively a correct value for the monthly and weekly workloads (see 18 and 19), for the night-call duty time (20), for the working time of temporary staff (21), for the number of night shifts (22), and the correct interchange between night shifts and other shifts.

4 Numerical Results

In this section we present an example of the model's output. Due to space constraints the example will report the output relevant to the shifts from 04/02/2019 to 17/02/2019 (2 weeks). As showed in Table 1, the ED under study has to cover 10 types of shift per day. Some shifts refer to the main ED (MO, AF, ENI, NI), whereas

Table 1 Shift duration and coverage

Timeframe	Shift code	Shift name	Duration (h)	Coverage weekends	Coverage weekdays
Morning	MO	Morning	6	2	3
	MO_O	Morning observational unit	6	1	1
	MO_T	Morning trauma	6	1	1
Afternoons	AF	Afternoon	6	2	4
	AF_O	Afternoon observational unit	6	1	1
	AF_T	Afternoon trauma	6	1	1
Early night	ENI	Early night	4	1	1
Night	NI	Night	12	2	2
	NC	Night call duty	0	1	1
	NC_T	Night call duty trauma	0	1	1

others refer to different areas (observational unit and trauma unit) which are covered by dedicated resources during the day (while at night they share the same resources with the main ED). Not all the physicians can cover all the shifts (e.g. only a few of them cover the trauma shift), not all the physicians are expected to work the same amount of hours per week. Each shift is characterized by its length and coverage. For the same shift, the coverage changes between weekends and weekdays. The weekend shifts include the Friday afternoon shift. Night-call duty shifts (NC and NC_T) are assigned with duration equal to 0 as physicians have to guarantee their availability but they will work only in case of emergency.

Table 2 shows the results of the first model/iteration where we assign weekend shifts. As we can notice, physicians working one weekend do not work the following one (and indeed the number of weekends worked by each physician is well balanced in the planning horizon of 6 months).

Table 3 shows the results of the second model/iteration. Here, weekend shifts are already assigned, and we assign the remaining ones. In this case, the objective is to balance afternoon and morning shifts, while considering the shifts that have already been assigned in the previous iteration as fixed.

The resulting schedule complies with the weekend schedule and balances the number of morning and afternoon shifts worked by each physician in a planning horizon of 1 month. Each phase of the two-phase approach can be solved to optimality within 10 min. However, the decomposition of the problem in two-phases is not

Table 2 Weekend shifts (from 04/02/2019 to 17/02/2019)

	MO	MO	MO	MO_O	MO_T	AF	AF	AF	AF	AF_O	AF_T	ENI	NI	NI	NC	NC_T
04/02/2019																
05/02/2019																
06/02/2019																
07/02/2019																
08/02/2019						Phy17	Phy26			Phy23	Phy23	Phy17	Phy1	Phy6	Phy9	Phy1
09/02/2019	Phy20	Phy12		Phy26	Phy3	Phy12	Phy14			Phy25	Phy3	Phy22	Phy17	Phy23	Phy9	Phy1
10/02/2019	Phy14	Phy12		Phy5	Phy27	Phy6	Phy14			Phy27	Phy25	Phy6	Phy3	Phy26	Phy9	Phy1
11/02/2019																
12/02/2019																
13/02/2019																
14/02/2019																
15/02/2019						Phy24	Phy4			Phy13	Phy24	Phy4	Phy7	Phy16	Phy21	Phy16
16/02/2019	Phy18	Phy2		Phy15	Phy24	Phy2	Phy13			Phy18	Phy24	Phy11	Phy4	Phy8	Phy21	Phy16
17/02/2019	Phy13	Phy2		Phy18	Phy7	Phy2	Phy11			Phy18	Phy7	Phy11	Phy15	Phy19	Phy21	Phy16

Table 3 All shifts (from 04/02/2019 to 17/02/2019)

	MO	MO	MO	MO_O	MO_T	AF	AF	AF	AF	AF_O	AF_T	ENI	NI	NI	NC	NC_T
04/02/2019	Phy10	Phy16	Phy11	Phy26	Phy23	Phy1	Phy24	Phy8	Phy26	Phy16	Phy23	Phy2	Phy15	Phy22	Phy19	Phy7
05/02/2019	Phy9	Phy4	Phy6	Phy17	Phy16	Phy10	Phy2	Phy8	Phy11	Phy24	Phy7	Phy12	Phy13	Phy21	Phy26	Phy1
06/02/2019	Phy10	Phy6	Phy7	Phy8	Phy23	Phy24	Phy11	Phy9	Phy15	Phy2	Phy3	Phy22	Phy16	Phy19	Phy3	Phy16
07/02/2019	Phy11	Phy9	Phy26	Phy17	Phy7	Phy7	Phy21	Phy6	Phy13	Phy4	Phy16	Phy4	Phy8	Phy2	Phy24	Phy1
08/02/2019	Phy16	Phy11		Phy21	Phy3	Phy17	Phy26			Phy23	Phy23	Phy17	Phy6	Phy1	Phy9	Phy1
09/02/2019	Phy12	Phy20		Phy26	Phy3	Phy12	Phy14			Phy25	Phy3	Phy22	Phy23	Phy17	Phy9	Phy1
10/02/2019	Phy12	Phy14		Phy5	Phy27	Phy14	Phy6			Phy27	Phy25	Phy6	Phy3	Phy26	Phy9	Phy1
11/02/2019	Phy13	Phy17	Phy12	Phy15	Phy24	Phy10	Phy8	Phy22	Phy7	Phy2	Phy16	Phy17	Phy21	Phy9	Phy19	Phy24
12/02/2019	Phy6	Phy23	Phy1	Phy11	Phy16	Phy4	Phy10	Phy1	Phy8	Phy2	Phy24	Phy11	Phy22	Phy15	Phy13	Phy1
13/02/2019	Phy8	Phy7	Phy2	Phy17	Phy16	Phy23	Phy7	Phy10	Phy11	Phy26	Phy3	Phy4	Phy13	Phy12	Phy8	Phy7
14/02/2019	Phy9	Phy17	Phy3	Phy16	Phy23	Phy1	Phy15	Phy22	Phy23	Phy4	Phy3	Phy4	Phy6	Phy10	Phy19	Phy1
15/02/2019	Phy17	Phy8		Phy26	Phy3	Phy4	Phy24			Phy13	Phy24	Phy4	Phy7	Phy16	Phy21	Phy16
16/02/2019	Phy18	Phy2		Phy15	Phy24	Phy2	Phy13			Phy18	Phy24	Phy11	Phy8	Phy4	Phy21	Phy16
17/02/2019	Phy13	Phy2		Phy18	Phy7	Phy11	Phy2			Phy18	Phy7	Phy11	Phy15	Phy19	Phy21	Phy16

motivated by efficiency reasons, but by the fact that the two problems usually involve planning horizon of different lengths and have peculiar constraints and objective functions which differentiate the one from the other.

5 Conclusion

In this paper we presented a novel two-phase approach to the EDPRP. The approach is based on two optimization models sharing the same network structure. The first model supports medium-term (6 months) planning decision concerning the assignment of weekend shifts, and allows balancing the weekend and night shifts assigned to each physician. The second one, instead, allows assigning the weekday shifts in the short term (1 month) in a way that (i) is compatible with the weekend shifts already assigned and (ii) ensures that morning and afternoon shifts are equally distributed across physicians. The models have been successfully tested using data from an Italian pediatric hospital.

Our future research effort will be aimed at embedding the presented models in a decision support system facilitating the physician rostering process in the hospital under study.

References

1. Adams, T., O'Sullivan, M., Walker, C.: Physician rostering for workload balance. Oper. Res. Health Care. **20**, 1–10 (2019)
2. Burke, E., Cowling, P., De Causmaecker, P., Vanden Berghe, G.: A memetic approach to the nurse rostering problem. Appl. Intell. **15**(3), 199–214 (2001)
3. Cappanera, P., Gallo, G.: A multi-commodity flow approach to the crew rostering problem. Oper. Res. **52**(4), 583–596 (2004)
4. Cappanera, P., Scutellà, M.G.: Color-coding algorithms to the balanced path problem: computational issues. INFORMS J. Comput. **23**(3), 446–459 (2011)
5. Damcı-Kurt, P., Zhang, M., Marentay, B., Govind, N.: Improving physician schedules by leveraging equalization: cases from hospitals in US. Omega (2018)
6. Erhard, M., Schoenfelder, J., Fügener, A., Brunner, J.O.: State of the art in physician scheduling. Eur. J. Oper. Res. **265**(1), 1–18 (2018)
7. Ernst, A.T., Jiang, H., Krishnamoorthy, M., Owens, B., Sier, D.: An annotated bibliography of personnel scheduling and rostering. Ann. Oper. Res. **127**(1–4), 21–144 (2004)
8. Ernst, A.T., Jiang, H., Krishnamoorthy, M., Sier, D.: Staff scheduling and rostering: a review of applications, methods and models. Eur. J. Oper. Res. **153**(1), 3–27 (2004)
9. Ferrand, Y., Magazine, M., Rao, U.S., Glass, T.F.: Building cyclic schedules for emergency department physicians. Interfaces **41**(6), 521–533 (2011)
10. Knust, F., Xie, L.: Simulated annealing approach to nurse rostering benchmark and real-world instances. Ann. Oper. Res. **272**(1–2), 187–216 (2019)
11. Stolletz, R., Brunner, J.O.: Fair optimization of fortnightly physician schedules with flexible shifts. Eur. J. Oper. Res. **219**(3), 622–629 (2012)
12. Valouxis, C., Gogos, C., Goulas, G., Alefragis, P., Housos, E.: A systematic two phase approach for the nurse rostering problem. Eur. J. Oper. Res. **219**(2), 425–433 (2012)

13. Van den Bergh, J., Beliën, J., De Bruecker, P., Demeulemeester, E., De Boeck, L.: Personnel scheduling: a literature review. Eur. J. Oper. Res. **226**(3), 367–385 (2013)
14. Visintin, F., Cappanera, P., Banditori, C.: Evaluating the impact of flexible practices on the master surgical scheduling process: an empirical analysis. Flex. Serv. Manuf. J. **28**(1–2), 182–205 (2016)
15. Visintin, F., Caprara, C., Puggelli, F.: Experimental design and simulation applied to a paediatric emergency department: a case study. Comput. Ind. Eng. **128**, 755–781 (2019)
16. Wong, T.C., Xu, M., Chin, K.S.: A two-stage heuristic approach for nurse scheduling problem: a case study in an emergency department. Comput. Oper. Res. **51**, 99–110 (2014)
17. Zhong, X., Zhang, J., Zhang, X.: A two-stage heuristic algorithm for the nurse scheduling problem with fairness objective on weekend workload under different shift designs. IISE Trans. Healthc. Syst. Eng. **7**(4), 224–235 (2017)

Using a Slotted Queuing Model to Compare the Efficacy of Emergency Departments Operating with and Without a Physician in Rural Communities

Peter T. Vanberkel, Benjamin Wedge, Alix J. E. Carter and Ilze Ziedins

Abstract Nova Scotia has developed a novel way to manage Emergency Department (ED) patients in rural communities. Staffed by a paramedic and a registered nurse, and overseen by physician via telephone, Collaborative Emergency Centres (CECs) have replaced traditional physician-led EDs overnight. We modeled the performance of CECs using a slotted queuing model to determine how well they perform in larger communities. It is shown that a CEC's success is related to the proportion of demand for primary care appointments compared with the supply of primary care appointments. Furthermore, we show that larger communities employing CECs will experience diminishing returns.

Keywords Slotted queueing model · Emergency medicine · Lindley's recursion · Rural emergency departments

1 Introduction

Delivery of after hours health care in rural Nova Scotia (NS) has undergone major changes in recent years to improve access to primary care and decrease overnight Emergency Department (ED) closures. It is common in rural NS for physicians to both have primary care practices and also provide ED coverage. When there is a physician shortage, access to primary care suffers and overnight ED closures become more common [1]. The shortage of daytime primary care appointments can be exacerbated when physicians shorten or cancel their clinic before or after working an overnight shift in the ED. This combination of overnight ED closures and primary

P. T. Vanberkel (✉) · B. Wedge
Department of Industrial Engineering, Dalhousie University, Halifax, Canada
e-mail: peter.vanberkel@dal.ca

A. J. E. Carter
Division of Emergency Medical Services, Dalhousie University, Halifax, Canada

I. Ziedins
Department of Statistics, University of Auckland, Auckland, New Zealand

V. Bélanger et al. (eds.), *Health Care Systems Engineering*,
Springer Proceedings in Mathematics & Statistics 316,
https://doi.org/10.1007/978-3-030-39694-7_8

care appointment shortages, coupled with low overnight demand [2], motivated the NS government to rethink how ED services are provided in rural NS. Following an extensive review of emergency health care services, [2] the province began to implement Collaborative Emergency Centres (CECs) in rural areas. CECs are staffed by a nurse and paramedic team overnight who treat patients with the assistance of a Medical Oversight Physician (MOP) who oversees multiple CECs and provides advice over the phone. During daytime hours, the CEC is staffed by a physician who sees his/her regularly scheduled primary care patients, walk-in patients, and those referred by the night time CEC team.

The implementation of CECs began in the summer of 2011. As of June 2017, there are eight CECs in NS and all operate in small communities. The largest CEC community has approximately 7600 residents in the catchment area, whereas the smallest CEC community has approximately 4100 residents. An assessment of the performance of these sites was commissioned by the NS Department of Health and Wellness. The assessment, reported in [1], found that the program has been effective in reducing overnight closures of EDs, has increased the availability of primary care, and has saved money. The provincial government has indicated they will continue to use the model and assess additional (and larger) sites. Six years after the first CEC opened, this paper seeks to determine if similar gains can be expected in larger population centres.

The most comprehensive collection of evidence related to the CEC model of care found was a Rapid Knowledge Synthesis project by Hayden et al. [3]. This project summarized, among other things, the typical structure of CECs, the common challenges for CEC implementation, and the locations across Canada which have CEC-type centres. The authors conclude that "there is limited scientific literature on the concept of CEC-type models as a health care delivery model." We refer readers to their comprehensive report for further details on the care model. We instead present a high-level description of how patients typically flow through the system and the resources with which they interact.

The patient flow in the traditional ED environment involves physician(s) and nurse(s) in an ED at night treating patients. Patient care is complete at the end of the ED visit with the patient discharged either home or admitted as inpatients for further care. In rural areas these are the same physicians who provide primary care the following day, and as such, they may see fewer patients preceding or following an ED shift than on other days [1]. The patient flow in the CEC model involves a nurse and paramedic in the ED at night treating patients with the support of a physician available via telephone. In terms of patient flow, the primary difference compared to a traditional ED is that these patients may be asked to come back the following day to complete or reassess their treatment. Furthermore, because there is not a physician overnight, physicians are not cancelling day time primary care appointments the following day. Some of the following day's primary care appointments may be used by patients referred by the overnight CEC team. The patient flow diagram for CECs is displayed in Fig. 1 where all disposition options are detailed. There is some variation in the CEC staffing complement and processes as is reviewed by Wedge [4].

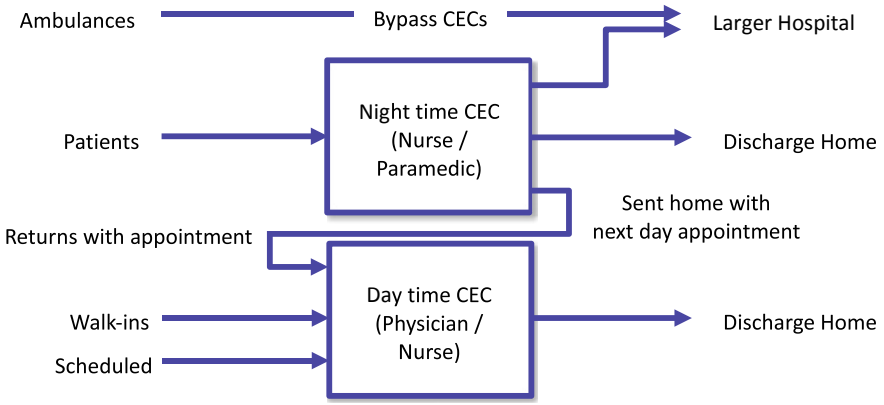


Fig. 1 CEC patient flow diagram

2 Methodology

To measure the efficacy of the CEC model in other (larger) communities we model the operations and performance of the traditional ED and then compare that to the operations and performance of the CEC. Although there are considerable clinical differences between these two care models (as described in [3]), the flow of patients and capacity to treat patients only changes slightly. By changing model parameters, the model described in this section is used to evaluation both the traditional ED and the CEC.

2.1 Model Description

Patient flow is modelled as a discrete time slotted queuing model. A discrete time queueing approach is used because it can represent the patient flow in a straightforward manner and allows for more generic and repeatable models suitable for multiple settings [5]. The model has two slots, one representing the daytime and one representing the night time. A slotted queueing model aggregates periods of time and considers arrivals and services as batches during this time period. The queue is computed at the end of the slot. As an example, consider a single slot representing a single day. If there were no patients waiting at the beginning of the day and 12 patients arrive and 10 patients are served, then the queue would be 2 at the end of the slot. For further details on slotted queueing models see [6].

A single day is divided into two slots, one representing all daytime primary care (including the daytime CEC) and one representing the night time CEC or ED depending on the scenario. Patients arriving to the daytime slot will be served if there is sufficient capacity. Patients who are served leave the system. Patients who are

not served from the daytime slot will either try again the next day or renege to the night time slot. Patients arriving to the night time slot will have either reneged or be new demand. Patients served at the night time slot will either be discharged to the Regional Hospital, home, or to the daytime slot the next day.

Patient arrivals at the daytime slot, denoted by A_t^P where t is the present day and P denotes primary care. The daytime slot has a daily capacity to see S_t^P patients on day t independent of the patient's acuity. Note that patient arrivals denote appointment requests i.e. calls to the clinic for an appointment, not patients physically arriving to the clinic. The shorter label "patient arrivals" is used for simplicity and clarity.

Those patients who are not seen on the day they requested an appointment are denoted by $L_{t,r}$, where r indexes the number of days that have passed since requesting an appointment. These patients will have priority for appointments over new appointment requests for the next daytime slot, however they renege and go to the night time slot with probability p_r. We denote these reneging patients by $A_{t,r}^E$. All other night time slot arrivals are denoted by A_t^E. It follows that the aggregate arrival rate to the nighttime slot is $A_t^C = A_t^E + \sum_{r=0}^{\infty} A_{t,r}^E$. The night time slot has a capacity to see S_t^E patients per night.

After service in the night time slot, patients are discharged to another hospital with probability p_1', patients are sent back to the daytime slot with probability p_2', and patients are discharged home with probability p_3'. The number of patients discharged in each manner is: $D_{1,t} = p_1' A_t^C$, $D_{2,t} = p_2' A_t^C$, $D_{3,t} = p_3' A_t^C$. Patients sent back to the daytime slots ($D_{2,t}$) receive first priority for appointments. For the ED, $D_{2,t} = 0$. That completes the feedback loop demonstrating how workload from the daytime slots overflow to the night time slots and vice versa.

The state of the system is described by Lindley's Recursion [7]. Let $L_{t,r}$ be the number of patients at the beginning of day t that have been waiting for service for r days. $L_{t,0}$ is increased by arrivals (A_t^P, $D_{2,t}$), and $L_{t,r}$, $r \geq 0$ is decreased by patients served by the daytime slot and patient reneging ($A_{t,r}^E$). Formally,

$$L_{t+1,r+1} = \{L_{t,r} - \{S_t^P - D_{2,t} - \sum_{j=r+1}^{\infty} L_{t,j}\}^+\}^+ - A_{t,r}^E \tag{1}$$

where $L_{t,0} = A_t^P$ and $x^+ = max\{0, x\}$.

Consider that the number of patients waiting $r + 1$ days tomorrow is the number of patients waiting r days today ($L_{t,r}$) minus those that received an appointment today ($\{S_t^P - D_{2,t} - \sum_{j=r+1}^{\infty} L_{t,j}\}^+$), and then if any are not seen today, minus those that reneged to the night time CEC ($A_{t,r}^E$). To determine how many waiting patients receive an appointment today we must consider how many daytime clinic appointments were available today (S_t^P) and how many appointments were consumed by patients of higher priority. Patients of higher priority include those referred from the night time slot ($D_{2,t}$) and those who have waited more days ($\sum_{j=r+1}^{\infty} L_{t,j}$).

2.2 Modelling the Random Processes

The two arrival processes, A_t^P and A_t^E, are assumed to be Poisson processes with mean λ_P and λ_E respectively. Poisson distributions have been shown to effectively model non-scheduled arrivals in healthcare settings [8]. Reneging ($A_{t,r}^E$) is modelled using a binomial distribution. The probability that $A_{t,r}^E$ patients waiting r days renege is equal to,

$$\mathbb{P}(A_{t,r}^E) = \frac{(L_{t,r}!)}{(L_{t,r} - A_{t,r}^E)! A_{t,r}^E!} p_r^{A_{t,r}^E} (1 - p_r)^{L_{t,r} - A_{t,r}^E} \tag{2}$$

This assumes that patients renege independently and with equal probability p_r given the number of days they have waited for an appointment. The assumption is what would be expected from a waiting list where patients do not interact, as in this case. For example, the decision by any patient to renege and go to a CEC does not influence any other patient's decision. The number of patients being sent back to the daytime clinics ($D_{2,t}$) is also modelled using a binomial distribution and conforms to the requirements of the binomial distribution.

We analyzed our model using simulation programmed in Visual Basic. This approach allows us to investigate unstable settings while maintaining the slotted queueing model structure. In the simulation, for each day t we sample from the distribution and generate a random variate for each of our random variables described above (A_t^P, A_t^E, $A_{t,r}^E$, $D_{2,t}$). Using these instances of the random variable we compute the state of the system and then advance to day $t + 1$.

2.3 Scenarios, Data, and Validation

The model is calculated first when a traditional ED is in operation and then when the traditional ED is replaced with a CEC. When the traditional ED scenario is run, patients do not go from the night time slot to the daytime slot as a physician is in the hospital in this scenario. In the CEC scenario p_2' percent of patients return the next day to see a physician. In the traditional ED scenario, physicians reduced the availability of daytime appointments in order to recover from overnight care hours. In the model this is managed by reducing the supply of daytime appointments by four, representing one hour of cancelled care [9]. In the CEC scenario the supply of appointments is the full value of S_t^P.

The model is run for four different population sizes: 4100 and 7700 patients (used as two baseline comparisons to towns with existing CECs) and 10,000 patients (a moderate-sized town, approximately 25

Several performance metrics are used to compare these two ED setups in these different communities. The proportion of patients who get an appointment on the first day they asked is a measure of primary care access and is found by computing

$1 - \frac{\sum(L_{t+1,1}+A^E_{t,r})}{\sum L_{t,0}}$ where the sums are taken over all t. The daytime provider utilization is found by taking the total number of appointments filled during the daytime care hours divided by the total number of appointments offered during the daytime. Finally, we compute the physician cost per overnight patient using the costs described in [2]. In the non-CEC scenario the physician cost per overnight patient is found by taking the average cost of overnight care, and dividing by the average number of overnight arrivals. The annual physician cost associated with running an overnight ED was reported to be $350,000 for an eight hour ED and $700,000 for a 14 h ED [2]. Prorating this to a 12 h ED leads to a cost of $657,000 per year or $1800 per night. For the CEC scenario, there is a flat fee of $150 per night for the Medical Oversight Physician, as well as a fee of $60 per patient referred back the daytime CEC, for the cost of the patient's primary care appointment. The difference between support staff costs for the two model scenarios is the same [2] and therefore ignored in the model.

Data for the model can be roughly classified into three categories; (1) primary care demand and service capacity, (2) reneging from the primary care queue to the ED/night time CEC and (3) ED/night time CEC demand and service capacity. Data for (1) comes from the Department of Health and Wellness' billings database. Data for (2) comes from health services literature and data for (3) comes from the electronic patient care record of the local ambulance service provider. Primary care demand and service capacity data for this research was extracted from the Department of Health and Wellness' billings database for the period April 1, 2012 to March 31, 2013. The data showed 8.49 visits per thousand people per day in one CEC community and 7.33 visits per thousand people per day in the other. These numbers are comparable to typical daily primary care appointment rates cited in the literature of approximately eight per thousand people [10]. We assume that a physician sees four patients per hour during the daytime clinics and that clinic hours are not cancelled.

An important characteristic of the model is patient reneging. While waiting for a primary care appointment, patients may decide to instead visit the night time CEC or ED. Previous research on reneging asked patients arriving at the ED how long they had been experiencing an issue [11]. Table 1 gives reneging probabilities used in this paper, with p_0 being the probability that a patient reneges on their day of arrival if they are not seen, and p_r being the probability that a patient who is still waiting for an appointment after r days reneges that day. For numerical purposes we assume the maximum number of days waiting for an appointment is finite and 7 days.

The electronic patient care record (ePCR) database records all visits to the overnight CECs since their inception (among other things). The data for this study

Table 1 Reneging probabilities

Day (r)	0	1	2	3	4	5	6	>7
Renege probability (p_r) (%)	16	6.5	7.0	7.5	8.1	8.9	9.7	100

were collected from April 1, 2012 to March 31, 2013. From this data we determined that after a patient has been assessed they are transferred to another hospital, treated and released with follow-up, or treated and discharged home with probabilities $p_1' = 14.9$, $p_2' = 56.2$, and $p_3' = 28.9$, respectively. The arrival rate of patients to the ED was found to be 2% of the arrival rate to the daytime primary care. This is slightly lower than the 4% reported in [12]. The arrival rate of patients to the night time CEC is made up of patients reneging from the day clinic and patients arriving directly to the night clinic. The former is an output from the daytime slot and the latter is assumed to be 1% of the arrival rate to the daytime primary care.

The warm-up period was found using Welch's graphical procedure and the run length was set to ten times the warm up period [13]. The number of replications required was found by performing a number of pilot runs and then calculating the number of replications required to keep each metric's confidence interval within certain bounds.

To validate the model, the number of overnight arrivals to the modelled ED/CEC was compared to historical data. In the ED version, the model found 1.479 ± 0.049 ($\pm$ denotes the size of the 95% confidence interval) arrivals per night, compared with 1.44 in the historical data. With the CEC operating, the model found there were 0.635 ± 0.021 arrivals per night, compared with 0.68 in the historical data. It is felt that the slight under-estimation of overnight arrivals is consistent with the finding that the night time arrival rate declined as the CEC program progressed [1]. Based on this information, it is felt that the model provides a valid approximation of the overnight arrivals to emergency care in both the CEC and ED settings.

3 Results

In this section we review the numerical results for the 16 scenarios—two baselines representing the existing CECs with 95 appointment requests per 100 offered and two larger communities each with 90, 99 and 110 appointment requests per 100 offered. For simplicity this ratio is expressed as a fraction e.g. 90 appointments requested per 100 offered is referred to as the "90/100 scenario". The summary of results for each scenario is in Table 2.

In the baseline scenario, with 4100 residents, we see that 79.2% of patients can get an appointment on the first day they ask, compared with 94.6% when the CEC opens. In the slightly larger Baseline 2 scenario, with 7700 residents, we see that 74.2% of patients receive care on the first day they ask, increasing to 96.3% when the CEC opens. Big improvements are seen in cost, where the physician cost per overnight patient drops from $1,218 in the baseline to $270 when the CEC opens; a drop of 77.8%.

In the 90/100 scenario and a community of 10,000 residents, 93.7% of patients can get an appointment on the first day they ask, which increases to 97.0% when the CEC opens. In the 99/100 scenario, 72.8% of patients can get an appointment on the first day they ask, which increases to 81.4% when the CEC opens. In the 110/100

Table 2 Summary of numeric results

Population	Appointment requests per 100 offered	Facility	Proportion of patients who get first day appointments	Daytime provider utilization	Physician cost per overnight patient	$\mathbb{E}[D_{2,t}]$
4100	95	ED	0.792	0.952	1218	0
4100	95	CEC	0.946	0.884	270	0.35
7700	95	ED	0.741	0.982	497	0
7700	95	CEC	0.963	0.91	145	0.73
10,000	90	ED	0.937	0.946	1067	0
10,000	90	CEC	0.97	0.918	155	0.68
10,000	99	ED	0.728	0.989	403	0
10,000	99	CEC	0.814	0.973	78	1.83
10,000	110	ED	0.166	1	149	0
10,000	110	CEC	0.007	1	41	9.94
20,000	90	ED	0.97	0.951	743	0
20,000	90	CEC	0.98	0.938	102	1.19
20,000	99	ED	0.777	0.994	237	0
20,000	99	CEC	0.747	0.993	51	4.64
20,000	110	ED	0.263	1	84	0
20,000	110	CEC	0	1	37	21.18

scenario, 16.6% of patients will get appointments on the first day they ask, dropping to 0.3% when the CEC opens. This demonstrates that primary care is less accessible after a CEC opens; a phenomenon which is discussed in Sect. 4. In all three scenarios the CEC reduces the physician cost per overnight patient. For the 90/100 scenario the cost decreases from $1067 to $155 when the CEC opens. In the 99/100 scenario it falls from $403 to $78 and in the 110/100 scenario it falls from $149 to $41.

With 20,000 people and the 90/100 scenario, the CEC increases the proportion of patients seen on the first day from 97.0 to 98.0%. When the CEC opened, 1.19 patients were referred back to daytime care each night. The CEC reduced the physician cost of overnight care from $743.21 to $102.35 in this scenario. In this community the CEC model showed modest performance advantages compared with EDs in the 99/100 scenario. The proportion of patients who were seen on the first day fell from 77.7 to 74.7% when the CEC opened. 4.64 patients were referred back to primary care when the CEC opened and the physician cost per overnight patient declined from $238 to $51.

Advantages of the CEC were not found when there was a shortage of primary care appointments. In the 110/100 scenario, 26.3% patients were seen on the first day they asked for an appointment when the ED was open, but none were seen on the first day they asked once the CEC opened. When the ED was open, the physician

saw 21.34 patients per night. The CEC saw 38.44 patients per night. Furthermore, 21.18 were sent back to primary care for follow-up, which would require 2/3 of a physician's daytime appointment capacity to treat. Despite this, the physician cost per patient still decreases, from $85 to $37. The findings from the 20,000-person community show that the CEC remains modestly beneficial when there is an excess of primary care capacity, but that it is detrimental to care when there is a shortage of primary care appointments.

4 Discussion

The simulation results show that the benefits anticipated from the CECs do occur under certain conditions. Specifically, the average cost per patient is reduced and primary care capacity is increased so long as there is not a shortage of primary care capacity prior to the CEC implementation. Furthermore, the simulation model demonstrates that the CEC program is prone to diminishing returns. As the catchment population grows, the proportional cost savings and primary care access improvements decrease. Importantly, any advantage is dependent on oversupply of primary care, and is lost in all scenarios in which there is undersupply.

To understand why CECs exacerbate shortages of primary care appointments one must understand that while CECs free physicians to provide more daytime primary care they also create rework. We observed that 56% of patients who visit the night time CEC are referred back to the daytime CEC. This means these patients are seen twice and had a physician been present at the night time CEC they would only have been seen once. If the number of rework appointments is greater than the number of extra daytime primary care appointments created by the CEC program, then primary care access becomes worse. This situation occurs primarily if the referral back to primary care rate becomes large or if the arrival rate to the night time CEC is large. The latter is expected when there is a shortage of primary care appointments leading to high reneging rates. Furthermore, in areas with large catchment populations the arrival rate to the night time CEC is expected to be large, meaning we expect a large number of patients referred back to daytime primary care. In this case, the large amount of rework may be reason enough not to convert to the CEC model.

Another reason we see diminishing returns in the benefits of CECs as communities get larger is because larger catchments are more tolerant to the loss of one provider for a portion of the day. Consider a population of 5000 that has access to 100 primary care appointments per day. If a physician cancels four appointments to recover from an overnight ED shift, primary care capacity is decreased by 4%. The same four canceled appointments for a population of 10,000 with 200 primary care appointments per day represents only a 2% decrease in primary care capacity. Such economy of scale results are also evident in the performance metric results.

The model, being the first to analytically evaluate CECs, has a number of limitations which may reduce its utility in measuring certain aspects of CEC performance. Data limitation exists in two ways. First, the model is calibrated using a population

subgroup found in rural Nova Scotia over a predefined time period. This may not be representative of some larger catchment populations and ignores behavioral changes and system feedback which may occur after CEC implementation. Second, a lack of Canadian data, as well as a lack of rural data, could also impact model prediction. We have used data for rural Canada when available, however some data was derived from American sources. Finally, data on reneging probabilities is limited. Note that these data limitations do not affect the model's ability to make comparative assessments of EDs and CECs in a particular community when configured with appropriate input parameters. The chosen model abstraction is overly efficient with regards to primary care access due to daytime provider pooling. The model accounts for all daytime primary care demand in a single queue. This pooling likely leads to underestimated wait times. However, underestimates of wait time due to pooling is small when systems operate at a high load [14], as is the case in our setting.

Acknowledgements This study was supported by the Nova Scotia Health Research Foundation (NSHRF) and Te Punaha Matatini.

References

1. Hampton, M.J.: Care right now. Technical report, Stylus Consulting for the Nova Scotia Government (2014)
2. Ross, J.: The Patient Journey Through Emergency Care in Nova Scotia. Government of Nova Scotia (2010)
3. Hayden, J., Babineau, J., Killian, L., Martin-Misener, R., Carter, A., Jensen, J., Zygmunt, A.: Collaborative Emergency Centres: Rapid Knowledge Synthesis. Nova Scotia Cochrane Resource Centre (2012)
4. Wedge, B.R.: Using a slotted queuing model to predict the operational performance of collaborative emergency centres (2016)
5. Creemers, S., Lambrecht, M.: Modeling a Hospital Queueing Network, chap. 18. Springer (2011)
6. Vanberkel, P.T., Boucherie, R.J., Hans, E.W., Hurink, J.L., Litvak, N.: Efficiency evaluation for pooling resources in health care. OR Spectr. **34**(2), 391–405 (2012)
7. Cohen, J.W.: The single server queue. In: North-Holland Series in Applied Mathematics and Mechanics, 2 nd edn, vol. 8. North-Holland Publishing Co., Amsterdam (1982)
8. Hall, R.W.: Patient Flow: Reducing Delay in Healthcare Delivery, 1st edn. Springer, New York (2006)
9. Ely, J.W., Burch, R., Vinson, D.C.: The information needs of family physicians: case-specific clinical questions. J. Family Pract. (1992)
10. Green, L.V., Savin, S., Murray, M.: Providing timely access to care: what is the right patient panel size? Joint Comm. J. Qual. Patient Saf. **33**(4), 211–218 (2007)
11. Grumbach, K., Keane, D., Bindman, A.: Primary care and public emergency department overcrowding. Am. J. Public Health **3**, 372–378 (1993)
12. Canadian Institute for Health Information.: Understanding emergency department wait times: access to inpatient beds and patient flow. Canadian Institute for Health Information (2007)
13. Law, A.M., Kelton, D.M.: Simulation Modeling and Analysis, 3rd edn. McGraw-Hill Higher Education (1999)
14. van Dijk, N.M., van der Sluis, E.: Pooling is not the answer. Eur. J. Oper. Res. **197**(1), 415–421 (2009)

A Meta Algorithm for Reinforcement Learning: Emergency Medical Service Resource Prioritization Problem in an MCI as an Example

Kyohong Shin and Taesik Lee

Abstract We present a finite-horizon Markov Decision Process (MDP) model for a patient prioritization and hospital selection problem, which is a critical decision-making problem in emergency medical service operation. Solving this model requires reinforcement learning (RL) due to its large state space. We propose a novel approach with an aim to significantly enhance the scalability of RL algorithms. Our approach, which we call a State Partitioning and Action Network, SPartAN in short, is a meta-algorithm that offers a framework an RL algorithm can be incorporated into. In this approach, we partition the state space into smaller subspaces to construct a reliable action network in the downstream subspace. This action network is then used in a simulation to approximate values of the upstream subspace. Using temporal difference (TD) learning as an example RL algorithm, we show that SPartAN is able to reliably derive a high-quality policy solution, thereby opening opportunities to solve many practical MDP models in healthcare system problems.

Keywords Emergency medical service · Patient prioritization · Hospital selection · Reinforcement learning

1 Introduction

One of the important decision problems in the operation of emergency medical service (EMS) system is a patient prioritization and hospital selection problem. This problem is particularly relevant in the aftermath of a mass casualty incident (MCI) when an EMS system experiences a severe resource shortage [1, 2]. While many prior studies on the medical decision-making problem including EMS resource operations

K. Shin · T. Lee (✉)
Department of Industrial and Systems Engineering, KAIST, 291 Daehak -ro, Yuseong -gu, Daejeon 34141, Republic of Korea
e-mail: taesik.lee@kaist.edu

K. Shin
e-mail: hong906@kaist.ac.kr

V. Bélanger et al. (eds.), *Health Care Systems Engineering*,
Springer Proceedings in Mathematics & Statistics 316,
https://doi.org/10.1007/978-3-030-39694-7_9

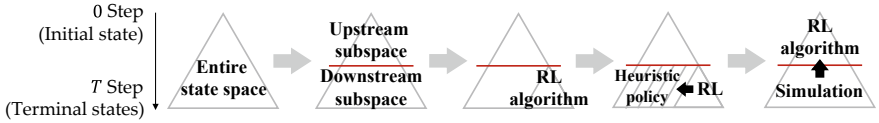


Fig. 1 The conceptual framework of SPartAN

have been based on scheduling or job assignment optimization formulation such as mixed integer program or stochastic optimization, more recent studies formulate the problem as a sequential decision making model, i.e. MDP model [3].

Solutions to an MDP model can be derived by using dynamic programming (DP). However, MDP models of complex systems, such as EMS operations under MCI, often require a large state space, making it difficult to use a dynamic programming method. This problem is well-known as curses of dimensionality [4]. Reinforcement learning (RL) addresses this problem by using an approximate value function. RL uses a simulation model to generate sample paths and calculates the value function of states along the generated sample path. Progressing forward from an initial state along a sample path, RL updates the value function of the states on the sample path. This process is repeated on thousands of sample paths to update the values of as many states as a computational budget allows. Despite the continuous development of various value function approximation techniques [5–10], the scalability problem is far from being conquered, and the need for new ideas and techniques is ever increasing.

With a goal of solving the patient prioritization and hospital selection problem under MCI, we propose a novel approach that significantly enhances the scalability of RL algorithms. Our approach, which we call a State Partitioning and Action Network, SPartAN in short, is a meta-algorithm in that it works as a framework any existing RL algorithm can be incorporated into. In SPartAN, we partition a system's state space into the upstream and downstream subspace, and obtain a policy solution in the downstream subspace first, and then passes the results to the upstream subspace through simulation. See Fig. 1, that is, we solve two RL problems within one big RL problem. Key ideas in SPartAN are threefold: reducing the size of an original RL problem by partitioning the state space into smaller compartments, directly obtaining values of the terminal states of the upstream compartment by using a simulation model, and constructing a quality heuristic policy in the downstream subspace by an action network (e.g. Deep Neural Network (DNN)) to use in the simulation.

This paper is structured as follows. In Sect. 2, we use temporal difference (TD) learning as an example of an RL algorithm to solve a patient transport decision problem under MCI. We show that the TD learning quickly becomes ineffective as the size of the problem increases, which motivates our development. In Sect. 3, we provide a detailed description of SPartAN, and the experiments and discussions on the results are presented in Sect. 4. Then we conclude our work in Sect. 5 with the areas to be addressed in future work.

2 Motivation

2.1 Patient Transport Decision Problem

Our patient transport decision problem is as follows. A large number of victims have just been rescued from a large-scale accident and need to be transported to nearby hospitals by one ambulance. There are two classes of victims, referred to as *immediate* and *delayed*, denoted by p_I and p_D respectively. These patients have a limited survival time within which they must receive care at a hospital to survive. Survival times for the patients in each class follow an exponential distribution with mean survival time of $1/r_I$ and $1/r_D$ respectively (r is referred to as an abandonment rate). While both classes of patients require medical intervention for their survival, immediate class patients are more critical than delayed class in that their mean survival time is shorter than that of delayed class patients ($1/r_I < 1/r_D$).

Two hospitals, H_1 and H_2, are accessible from the accident site with the mean travel time by the ambulance $1/w_1$ and $1/w_2$, respectively. Comparing the two hospitals, H_2 is *better* in that they can treat a patient quicker with a higher quality of care. This is modeled as a single server queue with a higher service rate ($\mu_2 > \mu_1$). For higher quality, we use an adjustment factor δ to give a higher probability of survival in computing immediate reward.

The decision problem is to decide which class of the patients and to which hospital the ambulance should transport next. The objective is to maximize the total sum of survival probability for the entire patients. For this problem, a system state is defined as $S = (p_I, p_D, h_1, h_2, \mathbf{A})$; p_I (p_D) denotes the number of immediate class (delayed class) patients remaining on the accident site, $h_j (j = 1, 2)$ is the number of patients in queue at hospital H_j and $\mathbf{A}$ represent one of the four ambulance's states – to/from (0/1) hospital j, $a_j^{0/1}$. When an ambulance arrives at the site and there are patients to be transported, emergency medical technicians make a decision and this decision is denoted by $X(s) = \{(i, j) : i \in \{I, D\}, j \in \{1, 2\}\}$. Naturally, the decision epoch for this problem is the moment when the ambulance arrives at the accident site. The time horizon of the problem runs through the moment all patients have received care at a hospital or died at the accident site.

Immediate reward of a decision to send a patient in class i to hospital H_j is defined as the patient's probability of survival. This is computed as the probability that a patient will start receiving care at a hospital before his survival time expires. Time until a patient receives care consists of waiting time at the accident site, transport time, and waiting time at the hospital. Then, the immediate reward is specified as:

$$R(s, x) = \delta_{i,j}\Bigg[\sum_{k=0}^{h_j-1}\{P(kZ < Y < (k+1)Z) \times P(T > Y + (h_j - k)Z)\} + P(h_j Z < Y) \times P(T > Y)\Bigg], \quad (1)$$

where $Y \sim \exp(w_j)$ is the travel time distribution, $Z \sim \exp(\mu_j)$ is the service time distribution at the hospital H_j and $T \sim \exp(r_i)$ is the survival time distribution for patients in class $i \in \{I, D\}$. The adjustment factor, $\delta_{i,j}$, represents that, with all other things being equal, the probability of survival at H_2 is higher than H_1.

Because $P(Y < Z) = \frac{w_j}{\mu_j + w_j}$, $P(Y < kZ) = \frac{kw_j}{\mu_j + kw_j}$, and $P(T > Y + h_j Z) = \frac{w_j \mu_j}{(h_j r_i + \mu_j)(r_i + w_j)}$, the final reward is as follows:

$$R(s,x) = \delta_{i,j}\left[\sum_{k=0}^{h_j-1}\left\{\left(\frac{(k+1)w_j}{\mu_j + (k+1)w_j} - \frac{kw_j}{\mu_j + kw_j}\right) \times \frac{w_j\mu_j}{((h_j-k)r_i + \mu_j)(r_i + w_j)}\right\} + \left(1 - \frac{h_j w_j}{\mu_j + h_j w_j}\right) \times \frac{w_j}{r_i + w_j}\right]. \tag{2}$$

The system state transitions to other states when one of three events occurs: patient death, ambulance arrival and patient discharge. Since the model assumes that the inter-arrival times of the three events follow an exponential distribution, the event generation process follows the Poisson process [11]. Thus, the state transition probability of the model is obtained by dividing the occurrence rate of each event by the sum of the rates of all events that can occur in the current state, which is:

$$\gamma = (w_j) + (p_I r_I + p_D r_D) + (1_{h_1>0}\mu_1 + 1_{h_2>0}\mu_2). \tag{3}$$

Each parenthesis represents the rate at which the ambulance arrives at the accident site or the hospital j, the rate at which patients die at the accident site and the rate at which the patient is discharged from the hospital. The indicator function, $1_{h_j>0}$, has the value 1 when there is at least one patient in hospital j and the value 0 when there is no patient.

Finally, the objective function of state $s = (p_I, p_D, h_1, h_2, A_j^0)$, $j \in \{1, 2\}$ is formulated using Bellman's equation:

$$\begin{aligned} V(s) = \frac{1}{\gamma}\Big[& w_j \max\Big(R(s,(I,1)) + V(p_I - 1, p_D, h_1, h_2, A_1^1), \\ & R(s,(I,2)) + V(p_I - 1, p_D, h_1, h_2, A_2^1), \\ & R(s,(D,1)) + V(p_I, p_D - 1, h_1, h_2, A_1^1), \\ & R(s,(D,2)) + V(p_I, p_D - 1, h_1, h_2, A_2^1)\Big) \\ & + p_I r_I V(p_I - 1, p_D, h_1, h_2, A_j^0) + p_D r_D V(p_I, p_D - 1, h_1, h_2, A_j^0) \\ & + 1_{h_1>0}\mu_1 V(p_I, p_D, h_1 - 1, h_2, A_j^0) + 1_{h_2>0}\mu_2 V(p_I, p_D, h_1, h_2 - 1, A_j^0)\Big], \end{aligned} \tag{4}$$

and the objective function of state $s = (p_I, p_D, h_1, h_2, A_j^1), j \in \{1, 2\}$ is formulated:

$$V(s) = \frac{1}{\gamma}\Big[w_j V(p_I, p_D, h_j + 1, h_{3-j}, A_j^0) + p_I r_I V(p_I - 1, p_D, h_1, h_2, A_j^1) + p_D r_D V(p_I, p_D - 1, h_1, h_2, A_j^1) + 1_{h_1>0}\mu_1 V(p_I, p_D, h_1 - 1, h_2, A_j^1) + 1_{h_2>0}\mu_2 V(p_I, p_D, h_1, h_2 - 1, A_j^1)\Big]. \tag{5}$$

2.2 Reinforcement Learning (Temporal Difference Learning)

To demonstrate the scalability in using RL as a solution approach for MDP problems, we use a TD learning as an example of an RL algorithm to solve the patient transport decision problem.

TD learning is one of the first learning algorithms to solve RL problems. Unlike backward DP, TD learning approximates the value functions of system states, $\bar{V}^n(s)$, by moving forward from its initial state to terminal states. It iteratively generates sample paths from a simulation model, where the value functions of the states along each sample path are updated by using the sample estimate [4]. In nth iteration, the sample estimate $\hat{v}^n$ of the current state s^n is computed by,

$$\hat{v}^n = \max_{a \in A(s)} [R(s^n, a^n) + E\{\bar{V}^{n-1}(s'^n)|s^n, a^n\}] \tag{6}$$

where $R(\cdot)$ is the immediate reward and $E\{\cdot\}$ is the expected value from the transition states s'^n. In (6), $\hat{v}^n$ depends on $\bar{V}^{n-1}(s'^n)$, which is the approximate value of state s'^n updated at $(n-1)^{th}$ iteration. With $\hat{v}^n$, the approximate value $\bar{V}^n$ of state s^n is updated as follows:

$$\bar{V}^n(s^n) = (1 - \alpha_n)\bar{V}^{n-1}(s^n) + \alpha_n \hat{v}^n.$$

This process of updating $\bar{V}^n(s^n)$ is continuously repeated until a predefined computational budget (e.g., iteration number) is reached.

2.3 Results from TD Learning

For the patient transport decision problem, we set the total number of patients to be 80, half of which are immediate class patients and the other half delayed class patients. A specific setting for the problem is shown in Table 1. Note that under this

Table 1 Experiment scenario

Patients	Immediate	Delayed	Hospitals	H_1	H_2
Number of patients	40	40	Mean travel time, $1/w_j$	1/8 hr	1/4 hr
Abandonment rate, r_i	0.5 ppl/hr	0.2 ppl/hr	Service rate, μ_j	3 ppl/hr	5 ppl/hr
			$(\delta_{I,j}, \delta_{D,j})$	(0.2 , 1)	(1, 1)

setting, the model can be exactly solved by dynamic programming, and the optimal policy for this problem is known.

We solve the example problem by using TD learning. We use $e^{-0.005n}$ as an exploration rate to balance exploitation and exploration. To control the learning rate, a harmonic stepsize rule is adopted $\alpha_n = \frac{50}{50+n'-1}$, and n' is the number of visits to state s [4]. We set the maximum number of iteration at 1×10^6. We solve this problem five times to obtain five policy solutions.

We examine the quality of the policy solution from the TD learning by comparing the policy with the optimal policy. Specifically, we divide the state space $\mathbf{S}$ into smaller compartments to see how the performance of the TD learning solution varies with respect to the depth of state transition. Since our problem has a finite horizon and the system state transitions from $S_0 = (40, 40, \cdot, \cdot, \cdot)$ toward $S_T = (0, 0, \cdot, \cdot, \cdot)$, a natural choice for compartmentalization is by the total number of patients remaining on the accident site, $p_I + p_D$. $\mathbf{S}_{(a,b)}$ then denotes the subset of $\mathbf{S}$, which is defined as $\mathbf{S}_{(a,b)} = \{S | a \times 10 < (p_I + p_D) \leq b \times 10\}$. For each compartment, we select states that have been visited more than 500 times, and compare the policy solution (i.e., action) from the TD learning with the optimal action computed from dynamic programming. Table 2 shows the number of cases where TD learning solution agrees with the optimal solution.

Table 2 shows that within $\mathbf{S}_{(6,7)}$ and $\mathbf{S}_{(4,6)}$, actions recommended by TD learning significantly differ from the optimal actions; in more than 60% of the states therein, TD learning gives a sub-optimal action. In fact, we observe a clear tendency in the results that, with an exception of $\mathbf{S}_{(7,8)}$, % agreement decreases toward the upstream

Table 2 % of states where TD learning policy agrees with the optimal policy

Compartment	$\mathbf{S}_{(7,8)}$	$\mathbf{S}_{(6,7)}$	$\mathbf{S}_{(4,6)}$	$\mathbf{S}_{(2,4)}$	$\mathbf{S}_{(1,2)}$	$\mathbf{S}_{(0,1)}$
# of states[a] (A)	206	433	1057	891	268	123
# of states in agreement (B)	110	157	395	478	186	96
% agreement (B/A × 100)	53.3	36.3	37.4	53.6	69.4	77.8

[a] These are the states that have been visited more than 500 times

of the state space. [1] That is, the TD learning solution yields fairly high % agreement in $\mathbf{S}_{(0,1)}$, and it drops toward $\mathbf{S}_{(6,7)}$.

In TD learning, to get a good approximation of the value for a state s, s needs to be repeatedly visited many times. In general, however, as the number of states explodes towards the downstream of the state space, chances of generating the same sample path many times thereby visiting same states along the sample path repeatedly are shallow. This makes it difficult for the algorithm to reflect the downstream reward in the value approximation for the states in the upstream state space. Due to this inherent difficulty, while TD learning generates a well-performing policy in the downstream state compartments (e.g., $\mathbf{S}_{(1,2)}$ and $\mathbf{S}_{(0,1)}$), rewards obtained from such a policy are not effectively propagated to the upstream state space.

In summary, this example illustrates the main cause of the limitation of the TD learning algorithm. Our proposed meta-algorithm addresses this problem by taking a different approach to better transfer values of downstream state space to upstream state space. Instead of relying on repeated visits (hopefully) by generating large number of sample paths from top to bottom, we use a simulation to directly obtain approximate value of the interim states, given a well-performing policy in the downstream state subspace. Details of our meta-algorithm are discussed next.

3 State Partitioning and Action Network (SPartAN)

We propose a meta-algorithm that aims to address the scalability problem of RL algorithms. This meta-algorithm is suitable for a finite horizon MDP. We call this meta-algorithm as State Partitioning and Action Network (SPartAN).

Suppose a state space $\mathbf{S}$ for a finite-horizon MDP model with step $2L$ is divided into $\mathbf{S}^U$ and $\mathbf{S}^D$. $\mathbf{S}^U$ is the subset of states in the upstream of the state space (say, step 1 through L). Likewise, $\mathbf{S}^D$ is the set of states in the downstream (step $L+1$ through T). Let us also suppose that we happen to know the optimal policy (Ω^D) for $\mathbf{S}^D$. Then, we can solve this problem for its entire state space $\mathbf{S}$ in the following fashion. Starting from the initial state s_0, we apply an RL algorithm (e.g., TD learning) to advance and update values of system states down to s_L, the terminal states of $\mathbf{S}^U$. $\bar{V}(s_L)$, the approximate value of s_L is then obtained by running a simulation model (starting from s_{L+1}) that uses the optimal policy Ω^D. This allows us to compute a policy for $\mathbf{S}^U$, which we would call Ω^U. [2] Finally, a policy solution for the entire state space is constructed by $\Omega^U \cup \Omega^D$. Clearly this gives an advantage over solving for the entire state space because we have a better approximation for $\bar{V}(s_L)$ than when we have to approximate $\bar{V}(s_L)$ by continued application of the RL algorithm.

[1] For $\mathbf{S}_{(7,8)}$, the reason why % agreement is high in the compartment is that the optimal policy in $\mathbf{S}_{(7,8)}$ is close to a greedy policy. In other words, in the compartment, the optimal policy simply chooses an action that maximizes the immediate reward.

[2] To be precise, if TD learning is used in this process, the policy solution obtained at this point is not defined for complete $\mathbf{S}^U$, and needs to be augmented by an additional step such as DNN.

The problem with the above description is that we do not know Ω^D before solving the entire problem. Our solution to this problem is to construct an approximate policy $\hat{\Omega}^D$. First, we build a reference policy, ω^D for a partial problem within $\mathbf{S}^D$. A partial problem for this construction can take any state $s^0_{(L+1)}$, and it is an MDP with a horizon length of $(T - L)$ on its own. For this partial problem with a reduced scale, an RL algorithm would yield a good quality solution ω^D. Now, we design a DNN to create an action network from ω^D for the rest of the states in $\mathbf{S}^D$. We use this policy as an approximation for Ω^D, and denote it as $\hat{\Omega}^D$.

A pseudo-code for SPartAN is shown in Algorithm 1. SPartAN takes the partitioning location L as an argument.

Algorithm 1 State Partitioning and Action Network: SPartAN(L).

1: $\bar{V}^0(s) \leftarrow 0 \quad \forall s$
2: Set the maximum iteration number N, the partitioning location L, and an initial state s^0_1
3: **for** $t = 1$ to L **do**
4: $\quad x^0_t \leftarrow$ randomly choosing $x \in X(s)$
5: $\quad s^0_{t+1} \leftarrow Sim(s^0_t, x^0_t)$
6: **end for**
7: Set a new initial state to s^0_{L+1}
8: **for** $n = 1$ to N **do**
9: $\quad$ Initialize the simulation configuration to match the new initial state s^0_{L+1}
10: $\quad$ **for** $t = L + 1$ to T **do**
11: $\quad\quad$ Apply an RL algorithm
12: $\quad$ **end for**
13: **end for**
14: Construct an action network $\hat{\Omega}^D$ by training ω^D
15: **for** $n = 1$ to N **do**
16: $\quad$ Initialize the simulation configuration to match the original initial state s^0_1
17: $\quad$ **for** $t = 1$ to T **do**
18: $\quad\quad$ **if** $t \leq L$ **then**
19: $\quad\quad\quad$ Apply an RL algorithm
20: $\quad\quad$ **else**
21: $\quad\quad\quad x^n_t \leftarrow \hat{\Omega}^D(s^n_t)$
22: $\quad\quad\quad \bar{V}^n(s^n_{L+1}) \leftarrow \bar{V}^n(s^n_{L+1}) + R(s^n_t, x^n_t)$
23: $\quad\quad\quad s^n_{t+1} \leftarrow Sim(s^n_t, x^n_t)$
24: $\quad\quad$ **end if**
25: $\quad$ **end for**
26: **end for**

After initialization, we randomly choose s^0_{L+1} as an initial state for a partial problem in $\mathbf{S}^D$ (line 2–7). $Sim(s, x)$ denotes that the action x is performed in the current state s and the simulation is proceeded to the next decision epoch (line 5, 23). In line 8–13, we solve the partial problem by using an RL algorithm to compute the reference policy, ω^D. Then we build an action network using ω^D as its training set (line 14). This is used as a heuristic policy $\hat{\Omega}^D$ for the state subspace $\mathbf{S}^D$. From line 15, we go back to the initial state of the original problem, s^0_1, and we apply an RL algorithm to $\mathbf{S}^U$. As shown in line 18–19, the RL algorithm proceeds down to step

L. At step $(L+1)$, we use the heuristic policy $\hat{\Omega}^D$ to run a simulation (line 21–23). Line 21 shows that an action x_t is determined from the heuristic policy $\hat{\Omega}^D$. During this process, approximate value of state s_{L+1} is obtained from the simulation result (Line 22).

Before closing this section, we want to emphasize that although we have motivated and explained SPartAN by using TD learning as an example, SPartAN works with any RL algorithm. As seen in line 11 and 19 of Algorithm 1, any RL algorithm can be used within the framework of SPartAN, and this is why we refer SPartAN as a meta-algorithm. SPartAN addresses the problem associated with a large state space MDP model, which is a common challenge to all RL algorithms.

4 Numerical Experiments

In this section, we revisit the patient transport decision problem discussed in Sect. 2.1. Again using TD learning as an example of an RL algorithm, we compare the quality of TD learning solution with SPartAN solution in Sect. 4.1. Then in Sect. 4.2, we conduct additional experiments to examine the effect of the relative size of the partitioned state space, $\mathbf{S}^U$ versus $\mathbf{S}^D$.

4.1 Effect of State Partitioning

Problem description and the experimental setting remain the same as presented in Sect. 2, and we only need to mention how we construct a DNN in SPartAN (line 14 in Algorithm 1). The DNN used in SPartAN consists of an input layer, four hidden layers, and an output layer. The input layer consists of four nodes: p_I, p_D, h_1, and h_2. Four hidden layers have 10, 30, 20, and 10 nodes, respectively, and the output layer has four nodes. Four nodes in the output layer describe a possible action in the problem.

For this experiment, we set $L = \frac{T}{2}$ as a partitioning step location for the state space $\mathbf{S}$. In this problem, the number of decisions (steps) made during the transition from the initial state to the final state is not fixed because the patients may die at the accident site. Therefore, we run the simulation using a random policy and set the state to the initial state of S^D when we visit the state where the sum of the number of patients remaining on the accident site is less than half the initial number of patients. Note that the TD learning algorithm is a special case of SPartAN with no partitioning, which is equivalent to partitioning at $L = 0$ or $L = T$. We obtain five policy solutions from SPartAN and TD learning, and under each policy we run 100 simulations respectively. As a performance measure, the average number of survivors from the simulations is reported. For each simulation run, common random numbers are used to test several policies using the same random input streams [12].

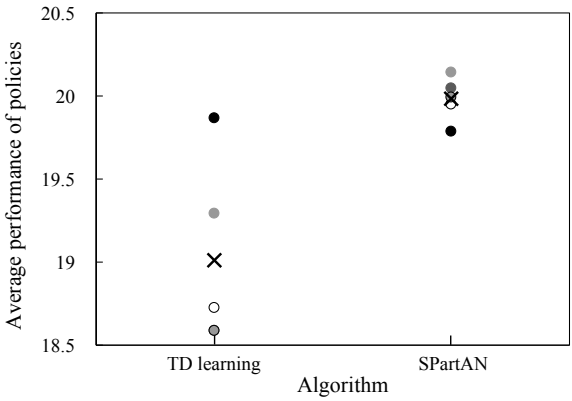


Fig. 2 Performance comparison between TD learning and SPartAN

Figure 2 shows the results from the simulation experiments. Each point in Fig. 2 is the average number of survivors under each policy solution obtained by SPartAN and TD learning. As expected, the performance of the policy solutions by SPartAN is higher than TD learning with the grand average of 19.98 versus 19.01 (X marks in Fig. 2). In addition, the variation in the performance from the five policy solutions is significantly smaller in SPartAN than TD learning. This result is confirmed to be statistically significant by t-test ($p < 0.001$). This means that SPartAN produces policies with more consistent performance, whereas the variance in the policy performance from TD learning is relatively large, rendering TD learning much less reliable.

4.2 Effect of State Partitioning Step Location, L

SPartAN takes L, a step location for state partitioning. To further understand how SPartAN works, we test difference partitioning location by varying L: $\frac{T}{8}$, $\frac{T}{4}$, $\frac{T}{2}$, $\frac{3T}{4}$, $\frac{7T}{8}$, and T. Note that $L = T$ corresponds to SPartAN with no partitioning and it is equivalent to the TD learning, and so is SPartAN with $L = 0$.

We look at the average performance of policy solutions obtained under each setting of L. Figure 3 shows the results. X marks in Fig. 3 mean total average performance. We conduct t-test with $p < 0.001$ to examine the differences among the results from different partitioning locations. $L = \frac{3T}{4}$ and $L = \frac{7T}{8}$ have a higher performance than the other locations with statistical significance. Although we can not make a conclusive remark about optimal partitioning location L^* with this single problem instance, Fig. 3 seems to suggest that larger L works to the advantage of SPartAN. In order words, when L is large (i.e. $\mathbf{S}^D$ is small), the performance of the policy solution is better and the variation in the computed policies is small.

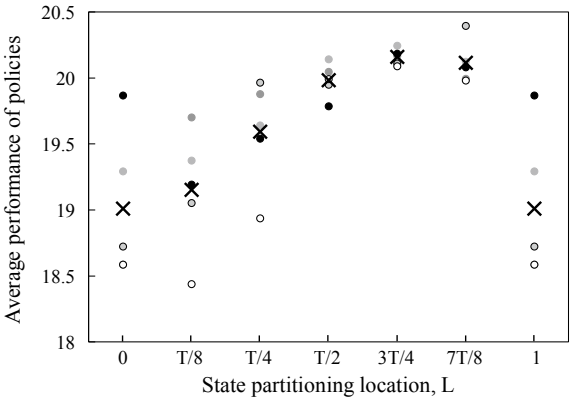


Fig. 3 Performance comparison of the depth size for state partitioning

Table 3 Performance of policy for each state subspace $\mathbf{S}^D$ and $\mathbf{S}^U$

L	$T/8$		$T/4$		$T/2$		$3T/4$		$7T/8$	
Subspace	$\mathbf{S}^D$	$\mathbf{S}^U$	$\mathbf{S}^D$	$\mathbf{S}^U$	$\mathbf{S}^D$	$\mathbf{S}^U$	$\mathbf{S}^D$	$\mathbf{S}^U$	$\mathbf{S}^D$	$\mathbf{S}^U$
# of states[a] (A)	2804	191	2186	638	1179	1844	501	2768	157	3093
# of states in agreement (B)	1473	9	1140	188	613	789	360	1273	122	1541
% agreement (B/A × 100)	53	5	52	30	52	43	72	46	78	50

[a] These are the states that have been visited more than 500 times

We examine the performance of the policies obtained for different L values in further detail. In Table 3, we show % agreement of actions determined by the obtained policies with respect to the optimal policy. The results are separately presented for $\mathbf{S}^D$ and $\mathbf{S}^U$ to compare the relative performance in each state compartment across L. Column $\mathbf{S}^D$ is the performance of the reference policy ω^D obtained for $\mathbf{S}^D$. Likewise, shown in column $\mathbf{S}^U$ is the performance of Ω^U, where Ω^U refers to the policy obtained in SPartAN for $\mathbf{S}^U$.

We see that the quality of ω^D is better for larger L. For example, when $L = \frac{T}{8}$, % agreement of ω^D is 53%, and for $L = \frac{7T}{8}$, it increases to 78%. As L gets larger (i.e. the size of $\mathbf{S}^D$ becomes smaller), an RL applied to solve a partial problem in $\mathbf{S}^D$ is able to derive a good reference policy ω^D. For smaller L (larger $\mathbf{S}^D$), ω^D deviates further away from the optimal policy due to the increased state space of the partial problem. The quality of the reference policy ω^D determines the quality of the heuristic policy $\hat{\Omega}^D$ to be used in the simulation step that approximates the value of the terminal state s_L in $\mathbf{S}^U$. In this sense, a larger value of L is a positive factor for the quality of Ω^U as well. On the other hand, larger L also poses a negative impact for the quality of Ω^U. Larger L means smaller $\mathbf{S}^D$, which in turn means larger $\mathbf{S}^U$. An RL needs to make a Ω^U at a larger scale when L is large. Nevertheless, the results suggest that the benefit from better $\hat{\Omega}^D$, thereby better approximation of s_L, outweigh the negative of having to solve a larger scale problem in $\mathbf{S}^U$.

Similar logic explains the pattern observed in Table 3 in the other way around. We would expect that an RL will better solve for Ω^U when $\mathbf{S}^U$ is small (i.e. L is small). On the contrary, the results show that the performance of Ω^U is in fact worse for smaller L. This is most likely due to the fact that larger $\mathbf{S}^D$ makes the quality of the reference policy ω^D worse, leading to poor approximation of values for s_L. Hence, even though small L should work to the advantage of obtaining Ω^U, its benefit is limited by the low quality of the policy solution in $\mathbf{S}^D$.

5 Conclusion

Recent advances in reinforcement learning techniques enable to solve MDP models that have been unsolvable before. Yet, there still is a great need to further improve their scalability. To address this problem, we propose a novel approach that can significantly enhance the scalability for solving a large-scale finite-horizon MDP model.

Our approach, which we call SPartAN in short, is a meta-algorithm in that it works as a framework the existing RL algorithms can be incorporated into. Three key ideas in SPartAN are partitioning the state space into smaller compartments to reduce the size of an original problem, using a simulation to directly obtain values of the terminal states of the upstream subspace, and constructing a quality heuristic policy in the downstream subspace by an action network to use in the simulation. Using the patient transport decision problem, we show that SPartAN is able to reliably derive a high-quality policy solution.

While we have a proof of concept for SPartAN using TD learning as its RL component, it remains to be confirmed that SPartAN works with other types of RL algorithm, for example DQN. Through several experiments, we have confirmed that SPartAN derives a high-quality policy solution even when using DQN as an RL component. The detailed results will be reported in a forthcoming article.

A vast range of problems in healthcare system operation have been and will be modeled by MDP. For these MDP problems, in most previous operations research literature, their main effort was on understanding the optimal policy structure and properties rather than actually computing the optimal policy solution itself due to computational difficulties. While finding such properties have a great value on its own, obtaining optimal policy and studying them carries practical significance. We hope our work can contribute to making MDP models a more practically viable approach to healthcare systems decision problems.

Acknowledgements This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (No. 2016R1A2B4014323).

References

1. Dean, M.D., Nair, S.K.: Mass-casualty triage: Distribution of victims to multiple hospitals using the save model. Eur. J. Oper. Res. **238**(1), 363–373 (2014)
2. Sung, I., Lee, T.: Optimal allocation of emergency medical resources in a mass casualty incident: Patient prioritization by column generation. Uropean J. Oper. Res. **252**(2), 623–634 (2016)
3. Bennett, C.C., Hauser, K.: Artificial intelligence framework for simulating clinical decision-making: A markov decision process approach. Artif. Intell. Med. **57**(1), 9–19 (2013)
4. Powell, W.B.: Approximate Dynamic Programming: Solving the Curses of Dimensionality. Wiley, New York (2007)
5. Anschel, O., Baram, N., Shimkin, N.: Averaged-dqn: Variance reduction and stabilization for deep reinforcement learning. In: International Conference on Machine Learning, pp. 176–185 (2017)
6. George, A., Powell, W.B., Kulkarni, S.R.: Value function approximation using multiple aggregation for multiattribute resource management. J. Mach. Learn. Res. **9**, 2079–2111 (2008)
7. Jia, Q.S.: On state aggregation to approximate complex value functions in large-scale markov decision processes. IEEE Trans. Autom. Control. **56**(2), 333–344 (2011)
8. Maxwell, M.S., Restrepo, M., Henderson, S.G., Topaloglu, H.: Approximate dynamic programming for ambulance redeployment. INFORMS J. Comput. **22**(2), 266–281 (2010)
9. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al.: Human-level control through deep reinforcement learning. Nature **518**(7540), 529–533 (2015)
10. Van Roy, B.: Performance loss bounds for approximate value iteration with state aggregation. Math. Oper. Res. **31**(2), 234–244 (2006)
11. Ross, S.M., Kelly, J.J., Sullivan, R.J., Perry, W.J., Mercer, D., Davis, R.M., Washburn, T.D., Sager, E.V., Boyce, J.B., Bristow, V.L.: Stochastic Processes, vol. 2. Wiley, New York (1996)
12. Wright, R.D., Ramsay Jr., T.E.: On the effectiveness of common random numbers. Manag. Sci. **25**(7), 649–656 (1979)

Facing Implementation Barriers to Healthcare Simulation Studies

Clio Dosi, Manuel Iori, Arthur Kramer and Matteo Vignoli

Abstract Implementation barriers to simulation studies are a reality in today's healthcare organizations. This work proposes a novel framework to use simulation to maximise successful implementation by (1) framing the right problem to face; (2) using what-if scenarios as an exploration tool for users' value; (3) supporting knowledge integration in giving tangible results to discuss among different professionals. We successfully tested the framework in an 18-month Emergency Department overcrowding case study, by developing a Discrete Events Simulation model and using it as a decision making tool for a multi-disciplinary group of 21 professionals (doctors, nurses, aid nurses, hospital management and engineers expert in simulation). Results show that the framework helps finding the most implementable solutions in the context of study, under the rationale that a small implemented improvement is preferable than a big one on paper. In the presented case study, after 15 years of absence of organisational change, the hospital was able to implement three new simulated solutions in 18 months.

Keywords Simulation · Implementation barriers · Design thinking · Facilitated modeling · Problem framing · What-if scenarios · Users' value · Knowledge integration

C. Dosi (✉) · M. Vignoli
Economic-Management Engineering Study Centre (CIEG), University of Bologna, 28, Via Terracini, Bologna 40131, Italy
e-mail: clio.dosi@unibo.it

M. Iori
DISMI, Department of Sciences and Methods for Engineering, University of Modena and Reggio Emilia, Modena, Italy

A. Kramer
Departamento de Engenharia de Produção, Universidade Federal do Rio Grande do Norte & DISMI, University of Modena and Reggio Emilia, Modena, Italy

V. Bélanger et al. (eds.), *Health Care Systems Engineering*,
Springer Proceedings in Mathematics & Statistics 316,
https://doi.org/10.1007/978-3-030-39694-7_10

1 Introduction

Healthcare problems received a lot of attention in the literature in many different contexts. Due to their growing complexity, these problems are becoming even more relevant in simulation modeling studies. With regard to the application of operations research techniques, it is possible to identify the use of simulation modeling to address problems usually found in healthcare units, as shown in the surveys by Brailsford et al. [5], Gunal and Pidd [13] and Katsaliaki and Mustafee [14]. In Brailsford et al. [5], due to the vast literature on the application of operations research techniques in healthcare, the authors used a new review methodology. With this new methodology, they identified and analysed the frequency of some characteristics of the works in the studied domain. For example, the authors analysed the frequency concerning the year of publication, methodology adopted and the level of implementation of the solution methods. Gunal and Pidd [13] and Katsaliaki and Mustafee [14], in turn, focused on the works involving the use of simulation techniques in the healthcare context. In particular, Gunal and Pidd [13] focused on papers proposing discrete event simulation models. These works show the growing interest in the application of simulation modeling techniques to improve the efficiency of healthcare units. Despite the number of studies, and the fact that the medical and managerial community agree on the importance of simulation studies, nowadays the simulation community recognizes that there are still implementation barriers.

The implementation issues of simulation have been discussed intensively during the last decades, both from practitioners' (since Lowery et al. [17]) and academic point of view (since Wilson [23]). Wilson explicitly addressed the problem in his contribution "Implementation of computer-simulation projects in healthcare", reporting that only 16 simulation projects out of the 200 considered reported successful implementations. Fone et al. [9, p. 333] systematically reviewed the use of healthcare simulation models and shed doubt on the value of the implementation "*we were unable to reach any conclusions on the value of modelling in health care because the evidence of implementation was so scant.[…] Further research to assess model implementation is required to assess the value of modelling.*". Brailsford [4, p. 1446] confirmed that "*Countless projects are carried out by academics and published in academic journals, but these models are not widely taken up by other health providers.*". Gunal and Pidd [13, p. 48] stated that *"Even after 25 years of this [Wilson's] review, all these barriers to the successful implementation of simulation still exist to some degree in all domains, including health care"*. This paper addresses techniques and methods to support simulation experts and decision makers in finding ways that maximise chances of implementation, and looks for suggestions in disciplines that already faced implementation issues such as design and management.

The paper is organised as follows: in Sect. 2, starting from management and design literature we identify three constructs that impact on simulation implementation and develop propositions that explain how, given those constructs, simulation could be used in a different way to maximise the chances of simulation results' implementation. We then present, in Sect. 3, the framework that we designed to support those

propositions, and show, in Sect.4, how that framework has been tested in a real case study. Our concluding remarks are presented in Sect. 5.

2 Constructs Impacting on Simulation Implementation.

We identified three constructs that strongly impact on the successful implementation of a simulation project.

2.1 Problem Framing

The classical approach of simulation only starts when a specific problem and a specific unit of analysis are given. Most of the time, simulation studies are used once the general conceptual organisational prototype is identified as the solution to design, and simulation models helps to detail the solution processes and resources, or to simulate that result in the actual context (see for example Starnino et al. [21] for a Rapid Assessment Zone solution in an adult ED). On the contrary, in the design practice, designers are used to identify problems throughout the process [6, 7]. The slogan is "Problem first, solution later." Designers usually produce several low-fidelity prototypes and test them with users to collect their critical comments. In this way, designers can better frame the problem, identifying the critical success factor. Interpreted in the light of the design approach, simulation can help experts to explore the problem before focusing on finding the best solution.

Proposition #1. *With the classical approach to simulation, the use of simulation aims at finding the best solution to solve a given problem. To maximise simulation results' implementation, we argue that the use of simulation needs to aim first at framing the right problem to be solved.*

2.2 Innovation-Value Fit

In the classical approach to simulation, users are involved in the process to have a deeper shared knowledge of the context and to consider their point of view into the decisional system. Still, the final aim of the simulation and of what-if scenarios is to find the scenario that has the most significant improvement of the defined Key Performance Indicators (KPIs). For example Banditori et al. [3] developed a combined optimization-simulation approach that allows to take decisions based on accurate estimates of the performance that the hospital will actually achieve if the investigated solutions are implemented.

Innovation literature explains why, even after the formal organisational decision to adopt an innovation, the organisation frequently fails the implementation. The construct "innovation-values fit" describes "the extent to which targeted users perceive that use of the innovation will foster (or, conversely, inhibit) the fulfillment of their values" [15], that is how much involved users value the degree of fit of that innovation with their own value. It is renown that, no matter how good the innovation is and how effective the top management of the organisation will be in pushing that implementation in the organisation, if the innovation has a poor fit with the users main values then its implementation will likely fail. Being able to evaluate if a potential innovation fits users' high-intensity values means that the decision maker can explore the organisational resistance to organisational change.

Proposition #2. *With the classical approach to simulation, the use of what-if scenarios aims at finding the scenario that improves the most the defined KPIs of the process. We argue that to maximise simulation results' implementation, the use of what-if scenarios should also be used to understand what is the users' perceived value of the proposed innovation.*

2.3 Knowledge Integration

The classical approach of simulation studies concentrates the interaction of the stakeholders around the simulation model itself, considering the implementation as a later stage that happens after the group has decided what is the best what-if scenario to adopt [20]. After the decision, the project is usually considered finished and the organisation is usually left alone in the process of implementing the solution of the identified what-if scenario.

The organisational complexity is such that the activity of defining how to implement the identified scenario at the organisational level is a design activity per se. The same stakeholder involvement that simulation experts know is needed in the simulation definition has to be put into the implementation part.

Management scholars call "knowledge integration" the concept of putting together diverse knowledge and ideas, that is to say the ability of a multidisciplinary team to attain knowledge integration. Gardner et al. [11, p. 999] define knowledge integration as "a reliable pattern of team communication that generates joint contributions to the understanding of complex problems". The higher the integration of different sources of knowledge (e.g., medical, nurse-related or technical knowledge), the higher the knowledge integration.

Proposition #3. *With the classical approach to simulation, knowledge integration among experts happens along with the simulation and ends once the best scenario has been identified. We argue that to maximise simulation results' implementation, the stakeholders involvement has to be extended to the organisational design of solutions related to the identified scenario.*

3 The Framework

Based on the three constructs presented, we designed a framework that leverages on simulation tools to maximise innovation implementation. The framework includes 3 sequential phases, each of them using the simulation tools in specific ways to enable the consideration of the previously identified constructs.

Such a framework could be used in organisational contexts where decisions are taken from a multidisciplinary group of stakeholders, composed by simulation programmers (engineers, mathematicians, process designers, organisational consultants), healthcare professionals (doctors, nurses, aid nurses, head of physicians and head nurse), and hospital top and middle management. For a shorter reference, from now on we will refer to: *engineers* as simulation experts; *medical staff* for healthcare professionals; *management* for top and middle management roles; *group* when referring to the whole group of decision makers.

1. **Problem Framing: Find the right problems to solve with simulation set up**

   The first phase of the framework aims at understanding what is the general problem to be tackled. The objective is to set a common language and mindset, where data and facts should inform and help understand the relevant problems that have to be faced. Such a phase, that is obvious for the engineers, is rarely shared by healthcare professionals.

   Setting up the simulation and the analysis of data that will be input of the simulation is a way to interact with the medical staff. During this interaction, the engineers need to highlight dynamics that are related to the department considered. While showing some dynamics that are renown from data, engineers can anticipate the medical staff that not all of the dynamics they expect to see will be confirmed, and use this opportunity to challenge medical staff assumptions, so that the group can understand the right problem to face.

   It is interesting to mention that this phase is preliminary to the classical face validation, that happens after the problem framing phase, when all the actors are aligned about what is the problem that will be addressed.
2. **Innovation-Value Fit: What-if to learn about organisational resistance**

   The second phase of the framework is meant to explore organisational resistance to change before deciding what is the solution decision makers will possibly implement. The what-if scenario is used to understand what is the users' perceived value of the proposed innovation, and if the KPIs improvement are worth their possible organisational resistance in the implementation phase. Simulation is thus the tool through which we learn about organisational resistance.

   Engineers present several what-if scenarios, aware that it is just an explorative phase and that they are leveraging on those scenarios to know if they fit with users' main values. In this phase, the engineers should produce the highest possible number of what-if scenarios, check them with the medical staff and management and eventually adjust them, in an iterative use of the simulation. What-if scenarios are used as a medium of continuous feedback with the actors involved in the processes. In plenary discussions, the engineers ask the actors to express their

feelings and thoughts regarding the presented what-if scenario, with explorative questions aiming at creating critical feedback. The phase is repeated until decision makers find a satisfactory number of what-if scenario with good innovation-value fit. In the final meeting, the engineers recap possible what-if scenarios. The whole group decide whether each scenario should receive a *go*, a *kill* or an *explore further.*

3. **Knowledge Integration: Transform scenarios in potential viable solutions**

   The last phase of the framework deals with the post-simulation design effort. Given some chosen scenarios, decision makers need to translate them into organisational potential solutions.

   To do this, the group designs multiple possible organisational solution that make the identified scenario viable. By designing, we mean that possible organisational solutions are benchmarked, ideated and tested with the users in the organisation.

   The learnings that happen during this phase will assure a knowledge integration in the group. The group can then support the implementation stage with a more grounded awareness of the perceived benefits and problems.

4 Case Study: Testing the Framework in a Major Italian EDs

In the last 40 years several studies have approached the Emergency Department (ED) overcrowding topic by using simulation tools, although with different aims, simulation techniques and approaches. Abo-Hamad and Arisha [1], for example, proposed a framework based on a simulation model to improve the efficiency of EDs and applied it to an ED in Dublin. The works by Aboueljinane et al. [2], Gul and Guneri [12] and Salmon et al. [19] present an overview of the use of simulation modeling in EDs. We developed a Discrete Event Simulation (DES) model of the ED processes to understand the current system and to investigate possible organisational changes in order to attain performance improvements.

The ED considered in our study case is located in the north of Italy and admits more than 80,000 patients per year. Given the general dissatisfaction of ED employees (doctors, nurses, and aid nurses) and high conflicts among professionals, the hospital top management and the head physician asked the authors for a support. The project lasted 18 months. The ED redesign had to improve the actual processes, to find possible ways to improve the ED system in general and ED professionals working habits in particular, so that professionals could be supported in their everyday routines. It is important to note that the organizational context was renowned as a conservative and hard-to-manage. In the last 15 years, most of the interventions proposed by different actors failed to be implemented. The hospital top management

involved the authors as mixed team between simulation experts and process engineers, to maximise the chances of implemented organisational changes that would improve the ED performances while increasing the internal professionals' quality of work. The general director and the head of ED would retire soon after the end of the project, and their hope was to leave an improved and more stable organisation.

We tested the proposed framework in this context developing a DES model and using it as a group decision making support instrument. The DES model simulates the flow of patients through the ED considering important features such as patients arrival rate and the distributions of patients urgency codes, service times and exams requirements. Main KPIs are: Wait time to first visit (WT_1V), Wait time to Last visit (WT_LV), Length of Stay (LoS), Outliers (percentage of patients with a WT_1V longer than 240 min—applied to white and green codes only) and number of patients waiting for the first visit (N). Every KPI is expressed both as average number on the total patients and average by color code. Among the possible organisational changes that have been evaluated as what-if scenarios there are: Equipe scheduling, Triage out, Anticipating exams at triage, Act on bottleneck lead times. Further details related to the numerical development of the simulation can be found in Dosi et al. [8]. We created an ad-hoc group of ED professionals that were involved in the design process and decision making. The plenary group involved 5 ED doctors, 6 nurses and 4 aid nurses plus the ED head of Physician and 2 professionals from the Managerial staff (1 from the Medical Direction, 1 from the Nursing Direction). This group had the responsibility to take decisions (with the four authors as simulation experts and process designers) and to inform and receive feedback from all the other 80 colleagues of the ED. The group met once every 10 days and the hospital top management was involved once every 3 months.

4.1 Test of the Framework 1st Phase: Problem Framing

When entering the ED, we realized that doctors, nurses and top management already had ideas regarding the problems they face every day, but none of those ideas were ever confirmed by data or discussed with other departments. The result was that professionals had in mind different problems to tackle and different priorities about them. For example, ED doctors blamed general practitioners' absence for the fact that patients with no need of ED interventions use to reach the ED rather than their general practitioner; ED professionals blamed the radiology department for the long waiting to have results of requested exams; nurses blamed the wards' professionals for the long wait for admission of patients in need of a hospital bed; doctors and nurses had relational problems with aid nurses, who were not always present when needed, and so on.

We realized that, to let the whole group being engaged in the design and implementation of ED improvements, it was necessary to design a pre-simulation phase where professionals were guided to explore different perspectives of the "ED improvement" problem. We used the simulation to discuss with the medical staff which problems

were potential problems that could lead to a successful implementation. In doing so, we were already screening some points of view of the problem.

Discussions related to the best level of analysis arose. The group had different perspectives on what level was more appropriate to study. Some nurses wanted to highlight triage practices so as to show how abilities of triage nurses affected the actual process, improving the *Length of Stay* (LoS) KPI. *"Our head nurse does not emphasize the efforts we have made to make the triage process more efficient. But it's a huge experience my triage colleagues and I have built over the last years. We think our informal triage practice has quite an impact to improve the general system flexibility. If we could show this, we could finally formalize the procedure."*—Triage nurse #2. Most of the professionals wanted to study the connected services (radiologies of the hospital and laboratory exams). "*We have 3 radiology units in this hospital, each of them has 3, 4 or more machines. Until two years ago the nearest radiology could only receive our patients, then the procedure changed, to support wards and special therapies, and this means that we have no more dedicated resources. We need to demonstrate that this choice has drastically affected our performances*"—Doctor #1 Other doctors wanted to demonstrate the impact of wards' bed management on their activities. The head nurse and her staff were more interested in mapping the Short Observation Unit inside the ED (the unit where critical patients should wait for a bed or for exams results after the first visit), to demonstrate that the Short Observation Unit was undersized. "*If we can show with numbers that this ED area has 12 patients slot and we usually have 30 patients or more, it won't be possible to avoid our request of one extra nurse anymore.*" Nurse #7, supported by others.

In this phase, the engineers need to highlight multiple dynamics that emerge from data, and use those dynamics to present different angles of the problem. To do this, we used a consistent database containing data from January to September 2017. It provided input quantitative information concerning the distribution of patients arrival rates, urgency and exams requirements, as well as information about service times of additional exams such as laboratory and x-ray. Among the shared dynamics, some of them will surprise the medical staff and challenge their assumptions. For example, in the next passage, one engineer of the group is challenging the fact that since radiology exams have long waiting times, then that is the problem that needs to be addressed. "*I am not sure the problem is radiology here. We will have to study data that we don't have now, we will understand when we have them, but - even if time for radiology is long - is not obvious that that's the bottleneck we need to act on*"—Engineer #1. In another passage, the healthcare professionals were discussing about the Short Observation Unit inside the hospital, as they initially wanted to demonstrate that the Short Observation Unit was undersized. Engineers explained "*Your IT systems could track the moment in which a patient enters the Short Observation Unit and the moment in which he exits from it, but no doctor fill that information. If you want to analyse that part of the ED process, as a group we need to make an effort and manually collect data. It's an extra effort for the nurses, and it has to last at least 2 weeks.*"—Engineer #1. Beside the input database, part of the simulation input data have been collected by in loco observation and data collection, and by interviewing

ED's staff. Among those input data we can recall queue rules currently used in the ED, some services execution times, resource availability and personnel schedules.

At the end of the first phase, the following major problems were confirmed: (i) many patients of white code arriving to the ED actually do not need any emergency service; (ii) the transport of blood samples required by the laboratory exams usually requires a large amount of time and constitutes a bottleneck in the process. On the other hand, the simulation disconfirmed that the following were relevant problems to face: (i) ED équipes have no way to impact on patients' LoS; (ii) laboratory exams are continuously required by the équipes, as we discovered that laboratory exams are only required during the first visit of the ED équipes; (iii) radiology exams are the bottleneck of the process.

4.2 Test of the Framework 2nd Phase: Innovation-Value Fit

In the second phase, the group develops the what-if scenarios and glimpse if there is an innovation-value fit among the scenarios and the users' main values. We developed 7 main what-if scenarios, named A, B, C, D, E, F and Comb (combination of the previous scenarios), inspired from literature and from discussions we had in the group. Scenarios aimed at improving the ED performance, given the identified KPIs, and identified several renown solutions in the ED literature to improve the ED process. Scenario A worked on team shifts by simulating the system under different shifts for the équipes. Scenarios B, C and F consider organisational solutions to improve low priority patients, by changing their priorities according to the current state of the system. Scenario D focuses on the laboratory exams and Scenario E in the triage process aiming at reducing the number of patients at the ED that are not in need of an ED intervention.

This phase define what scenarios could turn into a potentially successful implemented solution. Scenario D works on the laboratory exams sub-process to reduce the general lead time (i.e. the sum of waiting time, specimen transportation time—from the ED to the lab, and examination times in the lab). The idea of the scenario is to anticipate the request of lab exams: instead of waiting for the doctor that requests lab exams during the first visit, the scenario analyses what happens if triage requests laboratory exams. If triage can request exams, the lab result waiting time happens while the patients wait to see the doctor. Scenario results show that if 50% of the patients have their lab exams anticipated by the triage request, you gain almost 7% on the LoS KPI (5% with 20% of patients). The idea is renown in the literature and it was hiddenly suggested by expert nurses of triage. The engineer group knew that this was probably aligned with nurses values, although they sensed that this solution could stumble across resistance of doctors. This is what happened as most of the doctors said "*You know, you have to go easy on saying that a nurse can be a doctor and prescribe laboratory tests.*"—Doctor #2. At the same time a senior doctor that was sponsoring the idea confided to the engineers that "*Consider that this hospital is a university hospital, and it's not easy to accept that a nurse can have such a*

power compared to the doctors.", with a senior nurse saying "*Let's not forget about repercussion from register of doctors. I already know of bad experiences that happened to colleagues that have touched doctors' privileges, especially in conservative environment as this one is*"—Nurse #2. In front of such an ambiguous feedback of innovation-value fit (fit for nurses, unfit for doctors), the group decided to *explore further* the scenario.

The difficult part of this phase for the engineers is to shift from a perspective on KPIs to a perspective on users values. Among the scenarios identified, scenarios A, D and E improved LoS in a very similar way (respectively of 6.3%, 6.8% and 6.5%), but—given their different impact on users values—they respectively received a *kill* and two *explore further*.

4.3 Test of the Framework 3rd Phase: Knowledge Integration

This phase translate the what-if scenarios that received a *go* or *further explore* in multiple potential organisational solutions. The difficult part is that no one in the group can know with certainty what implementable organisational solution best fits the ED. To reach such an awareness different professionals need to integrate their knowledge while designing and testing that potential solutions. By testing the potential solutions, they will understand if the potential organisational solution can be considered as an implementable prototype for the final solution that will be implemented. To explain what this means we report the story of Scenario D. Scenario D wanted to anticipate laboratory exams request from the first visit to the triage moment. This could improve the LoS KPI by making the waiting time for lab exams results parallel to the first visit waiting time. Despite the innovation-value fit was ambiguous and the what-if scenario received just a "further explore" vote the scenario was finally implemented. The group analysed other case studies where triage could anticipate laboratory exam requests and discovered that not only ED literature accept this option but also that other EDs in the Region had successfully experienced that practice. The group estimated that 20% of patients exams could be intercepted by the triage request. The group realised that the competence of the nurses was critical in this new design as doctors and triage knowledge needed to be integrated into the role of the triage nurse. The group could not say whether their ED could be able to anticipate exam requests. To discover this, the group created an organisational prototype providing answers to such a critical function. Two senior doctors and two triage nurses defined a panel of symptoms and the connected list of lab exams to request. Moreover, the group thought to use the role of the area-doctor as a consultant of triage in case of necessity, so that nurses could always count on an expert medical suggestion. The group decided to prototype the solution and to implement it for one month. At the end of the month, the group realised that only 5% of the exams had been intercepted. The result was that nurse and doctor's knowledge was integrated and the organisation became aware of the fact that the ED could bear such a solution, and that to

improve its efficacy the whole organisation had to foster triage and doctors knowledge integration. If the engineers had suggested the top management to implement lab exams anticipation from triage starting from the simulation only, the chances of implementing such a solution would have been very low as most of the professionals were not aware of the complex organizational shift. Even if the organisation where we tested the framework is renowned for being a difficult-to-manage organization, extremely change resistant, with the use of this framework 3 process improvements were implemented in just 18 months.

5 Conclusion and Discussion

Several conditions impact on the success of the implementation of a new organizational solution: for example, how strongly the organisational climate support the implementation, how good designers are in finding the right solutions, how much the scenario improves the actual process KPIs and motivates professionals. However, given those conditions, how engineers use the simulation tool makes a difference. We designed and tested a framework that leverage on the simulation tool to maximise organisational implementation.

First, the framework changes the approach to the simulation tool, finding the most implementable solutions in the context of study, not the best ones in terms of KPIs improvement. The rationale behind this choice is that—with the words of the medical staff *"a small implemented improvement is better than a great solution on paper"*. Second, the three phase framework proposes a novel framing of the simulation in the whole change process, helping to frame the right problem to face, using what-if scenarios as an exploration tool for users' value and supporting knowledge integration in giving tangible results to discuss among different professionals. At last, with this framework the engineer changes her role. While she is still an expert of the simulation tool, she also becomes a facilitator of the whole organisational redesign. Her work is far more complex and takes more time if compared to the classical simulation process, but this will increase the chances that the result of the simulation project will turn into reality.

In the last years some recent contributions developed frameworks that proposed a different use of simulation techniques, under the keywords of facilitated modelling, soft methodologies or problem structuring (e.g. Franco and Montibeller [10]). Several contributions involve users in simulation modelling, for example Kotiadis et al. [16] report techniques on how to involve stakeholder participation during the first stage of simulation modelling (conceptual modeling), while Robinson et al. [18] involve participants in the experimentation phase.

However, to the authors knowledge, none of the frameworks found in literature use simulation as a tool to foster solution implementation. In our framework, in fact, we believe that implementation should not be considered as the last phase of the process (e.g. Robinson et al. [18], Tako and Kotiadis [22]) but as a goal to pursue in every phase. Future research should build on this framework and develop techniques,

define the type of interaction and role of actors, and assess the best tools to be used in each phase.

Acknowledgements This research was partially funded by CNPq—Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brazil, grant No. 234814/2014-4 and by University of Modena and Reggio Emilia, under grant FAR 2018 Analysis and optimization of healthcare and pharmaceutical logistic processes.

References

1. Abo-Hamad, W., Arisha, A.: Simulation-based framework to improve patient experience in an emergency department. Eur. J. Oper. Res. **224**(1), 154–166 (2013)
2. Abouelji nane, L., Sahin, E., Jemai, Z.: A review on simulation models applied to emergency medical service operations. Comput. Ind. Eng. **66**(4), 734–750 (2013)
3. 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)
4. Brailsford, S.C.: Advances and challenges in healthcare simulation modeling: tutorial. In: Proceedings of the 39th Conference on Winter Simulation: 40 years! The Best Is Yet to Come, IEEE Press (2007)
5. Brailsford, S.C., Harper, P.R., Pate, B., Pitt, M.: An analysis of the academic literature on simulation and modelling in health care. J. Simul. **3**(3), 130–140 (2009)
6. Dorst, K.: The core of 'design thinking' and its application. Des Stud **32**(6), 521–532 (2011)
7. Dosi, C., Rosati, F., Vignoli, M.: Measuring design thinking mindset. In: 15th International Design Conference—DESIGN 2018, Dubrovnik, Croatia (2018)
8. Dosi, C., Iori, M., Kramer, A., Vignoli, M.: Computational Simulation as an organizational prototyping tool. In: Proceedings of the Design Society: International Conference on Engineering Design, Vol. 1, No. 1, pp. 1105–1114. Cambridge University Press (2019)
9. Fone, D., Hollinghurst, S., Temple, M., Round, A., Lester, N., Weightman, A., Palmer, S., et al.: Systematic review of the use and value of computer simulation modelling in population health and health care delivery. J. Public Health **25**(4), 325–335 (2003)
10. Franco, L.A., Montibeller, G.: Facilitated modelling in operational research. Eur. J. Oper. Res. **205**(3), 489–500 (2010)
11. Gardner, H.K., Gino, F., Staats, B.R.: Dynamically integrating knowledge in teams: transforming resources into performance. Acad. Manag. J. **55**(4), 998–1022 (2012)
12. Gul, M., Guneri, A.F.: A comprehensive review of emergency department simulation applications for normal and disaster conditions. Comput. Ind. Eng. **83**, 327–344 (2015)
13. 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)
14. Katsaliaki, K., Mustafee, N.: Applications of simulation within the healthcare context. J. Oper. Res. Soc. **62**(8), 1431–1451 (2011)
15. Klein, K.J., Speer, S.J.: The challenge of innovation implementation. Acad. Manag. Rev. **21**(4), 1055–1080 (1996)
16. Kotiadis, K., Tako, A.A., Vasilakis, C.: A participative and facilitative conceptual modelling framework for discrete event simulation studies in healthcare. J. Oper. Res. Soc. **65**(2), 197–213 (2014)
17. Lowery, J.C., Hakes, B., Lilegdon, W.R., Keller, L., Mabrouk, K., McGuire, F.: Barriers to implementing simulation in health care. In: Proceedings of Winter Simulation Conference, pp. 868–875, IEEE (1994)
18. Robinson, S., Worthington, C., Burgess, N., Radnor, Z.J.: Facilitated modelling with discrete-event simulation: reality or myth? Eur. J. Oper. Res. **234**(1), 231–240 (2014)

19. Salmon, A., Rachuba, S., Briscoe, S., Pitt, M.: A structured literature review of simulation modelling applied to emergency departments: current patterns and emerging trends. Oper. Res. Health Care **19**, 1–13 (2018)
20. Sargent, R.G.: Verification and validation of simulation models. J. Simul. **7**(1), 12–24 (2013)
21. Starnino, A., Dosi, C., Vignoli, M.: Designing the future, engineering reality: prototyping in the emergency department. In: ServDes, 2016: Service Design Geographies, vol. 125, pp. 574–579. Linköping University Electronic Press, Linköpings universitet (2016)
22. Tako, A.A., Kotiadis, K.: PartiSim: a multi-methodology framework to support facilitated simulation modelling in healthcare. Eur. J. Oper. Res. **244**(2), 555–564 (2015)
23. Wilson, J.T.: Implementation of computer simulation projects in health care. J. Oper. Res. Soc., 825–832 (1981)

Operating Room

Reallocating Operating Room Time: A Portuguese Case

Mariana Oliveira, Luísa Lubomirska and Inês Marques

Abstract Health care providers face a continuous increase in the complexity of organizations mainly due to the increasing demand and to the development of new and expensive technologies. The operating room (OR) is a major challenge in the hospital and is crucial for the institution financial health. Moreover, the OR has a large impact in several units of the hospital and on the workforce of the immediate up- and downstream units. In the last decades, surgery demand has been increasing with restrictive resources, forcing ORs to be more efficiently and effectively managed. This work is developed under a partnership with a Portuguese public hospital and aims to achieve a major social impact, which is increasing surgical access and thus reducing the patients waiting lists. Given the hospital restrictions in terms of space and staff, this work focuses on the reallocation of the available OR time among the surgical services, proposing new master surgical schedules—aggregate production planning consisting of timetables with specific timeslots assigned to each specialty. The main objective is to match demand and the existing capacity while maximizing OR efficiency. This work proposes a mathematical programming model, with three objectives: to maximize the allocated slots weighted by aggregated staff preferences; to match supply and demand; and to level the workload of up- and downstream units. A comparison of the actual allocation of slots with the one suggested by this approach is performed. Results show that the workforce is one of the major bottlenecks, suggesting a new distribution of the workforce among the specialties.

Keywords Operating room · Master surgery schedule · Capacity allocation · Tactical decisions · Mathematical model

M. Oliveira (✉) · L. Lubomirska · I. Marques
Centre for Management Studies, Instituto Superior Tócnico,
University of Lisbon, Av. Rovisco Pais 1, 1049-001 Lisbon, Portugal
e-mail: mariana.m.oliveira@tecnico.ulisboa.pt

L. Lubomirska
e-mail: luisalubomirska@tecnico.ulisboa.pt

I. Marques
e-mail: ines.marques.p@tecnico.ulisboa.pt

V. Bélanger et al. (eds.), *Health Care Systems Engineering*,
Springer Proceedings in Mathematics & Statistics 316,
https://doi.org/10.1007/978-3-030-39694-7_11

1 Introduction

Managing health care organizations is increasingly complex and challenging. This complexity can be related with the evolution of two main factors: the ageing and comorbility population which increases demand and the development of expensive technologies in the health sector. Within the services provided by hospitals, surgical activity is a major center of costs and revenues. Operating rooms (ORs) represent more than 40% of hospitals costs and profits and are often considered as the engine of the organization [6, 8]. Surgery involves high specialized medical staff and equipment which have high associated costs. Among the complexity factors, this activity includes a high level of variability and uncertainty related to demand and resources consumption, stakeholders perspectives and material availability. Moreover, surgical activities have not only an intrinsic high complexity but represent a large social responsibility as directly impacting the health status of the patients waiting for surgery. To guarantee quality of health care, surgical activities should be held in a certain time frame measured by the waiting time. A higher service level is achieved with a lower number of days that a patient waits to the procedure at least in compliance with a predetermined maximum waiting time. This work aims to reduce the waiting list for surgical procedures by focusing on the tactical decision level of OR planning and scheduling, namely a resource allocation problem. Thus, a mathematical model is developed to optimize the available OR time while matching surgical supply and demand, balancing the workload of up- and downstream resources and considering stakeholders' preferences.

OR time allocation can follow three different strategies: block scheduling, open scheduling, and modified block scheduling. This paper considers the block scheduling strategy where time slots (i.e., a combination of an OR, a day and a time period) are assigned to a specialty or to a surgeon group [16]. This is the strategy followed by a master surgery schedule (MSS) and this is the most used approach by several authors and in several hospital contexts (e.g. [6, 7, 17]) as providing a higher stability to managers and to the medical staff, especially when dealing with a cyclic MSS. This stability affects managers by giving a more predictable pattern of bed occupancy in the up- and downstream units, and also the required staff and material [5]. Open scheduling allows surgeons to use any of the time slots according to their needs. This is based on the idea that no time slot is reserved for a particular surgeon and, therefore, surgeons can use all available time slots and compete for OR time. This approach increases flexibility, which means that it might better and quicker adjust to waiting lists dynamic, and it can also increase the efficient utilization of the ORs [5]. Liu et al. show that open scheduling is able to increase OR efficiency and decrease overtime cost, especially in large-scale OR cases [12]. Modified block scheduling tries to achieve the benefits of both previous strategies: stability and flexibility. Few authors report on the application of this approach [1].

Different problem characteristics, objectives and solution methods are applied for OR planning problems under the block scheduling strategy. The first papers only focus on the OR. Blake and Donald [8] focus on the equitable distribution of time

slots among the surgical groups while Agnetis et al. [5] assign time slots to surgical groups with the possibility to dynamically adapt the number of hours allocated to each surgical group according to the evolution of the waiting list. Day et al. combine open access scheduling and dedicated slots, and minimizes underused OR time and overtime [9]. In 2007, downstream units are introduced to the OR time allocation problem (e.g. [6, 17]) justified by the high impact that the OR has on many other units inside the hospital. Zhang et al. study the impact of the OR on upstream units [20]. The main objective is to minimize inpatients' length of stay in the wards when waiting for surgery, to build a more robust MSS. Moreover, Banditori et al. [2] address the MSS problem considering the number of patients in the waiting list with due date over the planning horizon. Cappanera et al. [3] propose a surgery scheduling method (operational decision level) in which conflicting stakeholders' interests are considered, namely the compliance with patients' surgery due date, OR utilization, beds' utilization and number of scheduled surgeries. Visintin et al. [18] solve a MSS problem by evaluating three flexible practices and concluded that introducing variable surgical team assignments (every time a new MSS is produced) and mixed sessions (long-stay and short-stay surgeries performed in the same session) increase the number of scheduled surgeries. Recently, Guido and Conforti propose a multi-objective integer linear programming model to consider trade-offs among underutilization of OR capacity, balanced distribution of OR time among surgical groups, waiting time and overtime [11]. Moreover, Visintin et al. [19] create a flexible tool to schedule patientsǵroups for surgery, in which stakeholders are involved in the development process by commenting in the resulting schedules, changing gradually the model and facilitation its implementation in the hospital. Marques et al. also develop a multi-objective mixed integer programming model to build cyclic MSSs for a private hospital where time slots can be either assigned to surgical specialties or to individual surgeons [14].

This work proposes a mathematical model which combines objectives already considered by other authors but never studied together: the allocation of the slots according to the waiting list evolution, as e.g. [13]; the distribution of the slots considering the preferences of the stakeholders (e.g. surgeons and anesthesiologists), as e.g. [15]; and the workload balance on up- and downstream units as e.g. [10]. Although these objectives have already been considered individually in the literature, their integration allows not only to encompass the preferences of the surgical staff but also to respond to patients' needs, while taking into account the characteristics and capacity of the hospital under study. Moreover, to compute the target allocation value, the duration of the surgeries in the waiting lists are considered. In the literature, this values is usually calculated taking into account only the waiting list length and the surgeries due date. This work is motivated by the case of a public hospital in Portugal. The OR time assigned to each surgical service has been kept almost constant over the last couple of years regardless of the changes in the surgical demand pattern and of the raising numbers in waiting list for surgery. The hospital needs to adapt supply to its constantly changing demand, by managing the OR time as efficiently and effectively as possible. Although presenting a general model which can be applied

in other contexts, this work uses this hospital as a case study to validate the model and perform the computational experiments.

The remainder of this paper is structured as follows. Section 2 introduces the MSS problem and the model formulation. In Sect. 3, the model is applied to the case study. Finally, Sect. 4 concludes the paper.

2 Mathematical Model

To build new MSSs, a large (i.e. 1-year) planning horizon is considered and represented by a set of weeks which are composed by a set of working days (in most cases, surgeries are only performed between Monday and Friday) organized in shifts. In each day, there is a set of available ORs equally equipped with all the necessary material and equipment to perform the surgery. When constructing the MSSs, specialties must be assigned to shifts, days and ORs considering the availability of medical staff (e.g. surgeons and anesthesiologists). Besides availability, their preferences are also considered. In order to guarantee workload balance among the staff, the number of slots that each staff member can use must be lower than established maximum values. Each specialty has a predefined weekly target number of slots which is based on the estimated weekly demand measured in number of patients, the average surgery duration and the total number of available slots. However, sometimes it is not possible to comply with the target values and the MSS incur in under or overallocation of slots which should be minimized. Besides ORs, up- and downstream units are also considered (e.g. pre-ward, intensive-care unit and wards). These resources have a limited daily capacity which can influence the ORs throughput. To avoid variability, a target value for utilization of each resource is also defined. Depending on the ORs management and performance, the up- and downstream units can be under or overutilized which it is aimed to be minimized. The number of patients in each unit in a day depends on the slots allocation and on estimated probabilities for a patient of each specialty being in each unit before or after a number of days. The relative importance of each unit is used to weight the minimization of under and over utilization. To keep staff routines and timetable stability but allow for flexibility (e.g. to match variations in demand and to adjust to changes in staff availability over the planning horizon), a non-cyclic MSS approach is considered although a maximum number of monthly and weekly changes is allowed. The monthly stability parameter considers the number of changes on the MSS from one week relatively to its corresponding week of the first month in the planning horizon. Moreover, the weekly stability accounts for the differences between each week and the first week of the considered month. Routines are indeed very important for the staff satisfaction and therefore for the receptiveness of such approaches.

A mathematical model for this MSS problem is formulated in (1)–(23). The notation is summarized in Table 1. Functionality of the ORs is modeled by constraints (2)–(9), supply and demand balance is formulated by constraints (10)–(12), stability

Table 1 Sets, parameters and variables for the mathematical model

Sets and indices	
$s \in \mathrm{S}$	Specialties
$m \in \mathrm{M}$	Months
$w \in \mathrm{W}$	Weeks
$d \in \mathrm{D}$	Week days
$k \in \mathrm{K}$	Days in the planning horizon; the first day of the planning horizon is $k = 1$
$r \in \mathrm{R}$	Operating rooms
$b \in \mathrm{B}$	Shifts
$i \in \mathrm{I}$	Surgeons
$a \in \mathrm{A}$	Anesthesiologists
$z \in \mathrm{Z}$	Up- and downstream units
Subsets	
W_m	Weeks of month m; the first week of month m is w_{1m}
S_z	Specialties that use unit z
I_s	Surgeons of specialty s
Parameters	
κ_{idb}^{surg}	Preference score for surgeon i
κ_{adb}^{anest}	Preference score for anesthesiologist a
t_{sw}	Target slots allocation for specialty s in week w
w_z	Relative weight of unit z
a_{swdb}^{surg}	Number of surgeons of specialty s available on week w, day d and shift b
a_{iwd}^{surgD}	1, if surgeon i is available on at least one shift on week w and day d; 0, otherwise
a_{wdb}^{anest}	Number of anesthesiologists available on week w, day d and shift b
a_{awd}^{anestD}	1, if anesthesiologist a is available on at least one shift on week w and day d; 0, otherwise
ww_i^{surg}	Maximum weekly workload for surgeon i
ww_a^{anest}	Maximum weekly workload for anesthesiologist a
mw_{sm}	Minimum workload for specialty s on month m
Δ_m^M	Monthly stability for month m
Δ_w^W	Weekly stability for week w
e_{zsk}	Probability that a patient of specialty s is in unit z on day k
λ_s	Average number of patients operated per slot by specialty s
n_{zs}	Maximum number of days that a patient of specialty s stays in unit z
u_{zk}	Target utilization for unit z on day k
c_{zk}	Available capacity of unit z on day k
δ^{surg}	Minium requested number of surgeons available per slot assigned
δ^{anest}	Minium requested number of anesthesiologists available per slot assigned
G	Large number

(continued)

Table 1 (continued)

Decision variables	
x_{swdbr}	1, if specialty s is assigned to OR r on week w, day d and shift b 0, otherwise
t_{sw}^{-}, t_{sw}^{+}	Negative and positive deviations of the number of allocated slots to the target value for specialty s on week w, respectively
u_{zk}^{-}, u_{zk}^{+}	Under and overutilization of beds on unit z on week w and day d (compared to the target utilization value), respectively
Auxiliary variables	
y_{swdbr}	0, if specialty s is assigned on week w to the same OR r, day d and shift b as the first week of the same month; 1, otherwise
j_{swdbr}	0, if specialty s is assigned on week w to the same OR r, day d and shift b as in the corresponding week on the first month of the planning horizon; 1, otherwise
f_{zk}	Expected number of patients in unit z on day k
v_{sw}^{t}	0, if $t_{sw}^{-} > 0$; 1, if $t_{sw}^{+} > 0$
v_{zk}^{u}	0, if $u_{zk}^{-} > 0$; 1, if $u_{zk}^{+} > 0$

of the MSS is promoted by constraints (13)–(16), whereas constraints (17)–(21) model workload in up- and downstream units.

Function (1) sums up three objectives. The first objective maximizes the number of slots assigned to each specialty, weighted by the relative aggregate preferences of surgeons and anesthesiologists. The second and third objectives relate with undesired deviations to target values for the OR time assigned to each speacialty and the workload in the up- and downstream units, respectively. The latter (units workload) is measured by the number of occupied beds and weighted by a relative importance given by the stakeholders to each unit.

$$\max \quad \sum_{s\in \mathrm{S}} \sum_{w\in \mathrm{W}} \sum_{d\in \mathrm{D}} \sum_{b\in \mathrm{B}} \sum_{r\in \mathrm{R}} \left(\frac{\sum_{i\in \mathrm{I_s}} \kappa_{idb}^{surg}}{|\mathrm{I}|} + \frac{\sum_{a\in \mathrm{A}} \kappa_{adb}^{anest}}{|\mathrm{A}|} \right) x_{swdbr} - \sum_{s\in \mathrm{S}} \sum_{w\in \mathrm{W}} \frac{t_{sw}^{-} + t_{sw}^{+}}{t_{sw}} - \sum_{z\in \mathrm{Z}} w_z \sum_{k\in \mathrm{K}} \frac{u_{zk}^{-} + u_{zk}^{+}}{u_{zk}} \tag{1}$$

$$\text{s.t.:} \quad \sum_{s\in \mathrm{S}} x_{swdbr} \leq 1 \quad \forall w \in \mathrm{W}, d \in \mathrm{D}, b \in \mathrm{B}, r \in \mathrm{R} \tag{2}$$

$$\delta^{surg} \sum_{r\in \mathrm{R}} x_{swdbr} \leq a_{swdb}^{surg} \quad \forall s \in \mathrm{S}, w \in \mathrm{W}, d \in \mathrm{D}, b \in \mathrm{B} \tag{3}$$

$$\delta^{surg} \sum_{b\in \mathrm{B}} \sum_{r\in \mathrm{R}} x_{swdbr} \leq \sum_{i\in \mathrm{I}_s} a_{iwd}^{surgD} \quad \forall s \in \mathrm{S}, w \in \mathrm{W}, d \in \mathrm{D} \tag{4}$$

$$\delta^{surg} \sum_{d\in \mathrm{D}} \sum_{b\in \mathrm{B}} \sum_{r\in \mathrm{R}} x_{swdbr} \leq \sum_{i\in \mathrm{I}_s} ww_i^{surg} \quad \forall s \in \mathrm{S}, w \in \mathrm{W} \tag{5}$$

$$\delta^{anest} \sum_{s\in \mathrm{S}} \sum_{r\in \mathrm{R}} x_{swdbr} \leq a_{wdb}^{anest} \quad \forall w \in \mathrm{W}, d \in \mathrm{D}, b \in \mathrm{B} \tag{6}$$

$$\delta^{anest} \sum_{s\in \mathrm{S}} \sum_{b\in \mathrm{B}} \sum_{r\in \mathrm{R}} x_{swdbr} \leq \sum_{a\in \mathrm{A}} a_{awd}^{anestD} \quad \forall w \in \mathrm{W}, d \in \mathrm{D} \tag{7}$$

$$\delta^{anest} \sum_{s\in \mathrm{S}} \sum_{d\in \mathrm{D}} \sum_{b\in \mathrm{B}} \sum_{r\in \mathrm{R}} x_{swdbr} \leq \sum_{a\in \mathrm{A}} ww_{a}^{anest} \quad \forall w \in \mathrm{W} \tag{8}$$

$$\sum_{w\in \mathrm{W}_m} \sum_{d\in \mathrm{D}} \sum_{b\in \mathrm{B}} \sum_{r\in \mathrm{R}} x_{swdbr} \geq mw_{sm} \quad \forall s \in \mathrm{S}, m \in \mathrm{M} \tag{9}$$

$$\sum_{d\in \mathrm{D}} \sum_{b\in \mathrm{B}} \sum_{r\in \mathrm{R}} x_{swdbr} + t_{sw}^{-} - t_{sw}^{+} = t_{sw} \quad \forall s \in \mathrm{S}, w \in \mathrm{W} \tag{10}$$

$$t_{sw}^{-} \leq G\left(1 - v_{sw}^{\mathrm{t}}\right) \quad \forall s \in \mathrm{S}, w \in \mathrm{W} \tag{11}$$

$$t_{sw}^{+} \leq G\, v_{sw}^{\mathrm{t}} \quad \forall s \in \mathrm{S}, w \in \mathrm{W} \tag{12}$$

$$|x_{swdbr} - x_{sw_{1m}dbr}| = y_{swdbr} \quad \forall s \in \mathrm{S}, w \in \mathrm{W}_m \setminus \{w_{1m}\}, m \in \mathrm{M}, d \in \mathrm{D}, b \in \mathrm{B}, r \in \mathrm{R} \tag{13}$$

$$\sum_{s\in \mathrm{S}} \sum_{d\in \mathrm{D}} \sum_{b\in \mathrm{B}} \sum_{r\in \mathrm{R}} y_{swdbr} \leq \Delta_w \quad \forall w \in \mathrm{W} \tag{14}$$

$$|x_{swdbr} - x_{sldbr}| = j_{swdbr} \quad \forall s \in \mathrm{S}, w \in \mathrm{W}_m, m \in \mathrm{M} \setminus \{1\}, l = w - \sum_{g<m} |\mathrm{W}_g|, d \in \mathrm{D}, b \in \mathrm{B}, r \in \mathrm{R} \tag{15}$$

$$\sum_{s\in \mathrm{S}} \sum_{w\in \mathrm{W}_m} \sum_{d\in \mathrm{D}} \sum_{b\in \mathrm{B}} \sum_{r\in \mathrm{R}} j_{swdbr} \leq \Delta_m \quad \forall m \in \mathrm{M} \tag{16}$$

$$0 \leq f_{zk} - \sum_{s\in \mathrm{S}_z} \sum_{b\in \mathrm{B}} \sum_{r\in \mathrm{R}} \sum_{l=0}^{\mathrm{n}_{zs}-1} \lambda_s e_{zsk} x_{s,w,d\pm l,b,r} \leq 1 \quad \forall z \in \mathrm{Z}, k \in \mathrm{k} : k \rightarrow (w, d), w \in \mathrm{W}, d \in \mathrm{D} \tag{17}$$

$$f_{zk} + u_{zk}^{-} - u_{zk}^{+} = u_{zk} \quad \forall z \in \mathrm{Z}, k \in \mathrm{K} \tag{18}$$

$$u_{zk}^{+} \leq c_{zk} - u_{zk} \quad \forall z \in \mathrm{Z}, k \in \mathrm{K} \tag{19}$$

$$u_{zk}^{-} \leq G\left(1 - v_{zk}^{\mathrm{u}}\right) \quad \forall z \in \mathrm{Z}, k \in \mathrm{K} \tag{20}$$

$$u_{zk}^{+} \leq G\, v_{zk}^{\mathrm{u}} \quad \forall z \in \mathrm{Z}, k \in \mathrm{K} \tag{21}$$

$$t_{sw}^{-}, t_{sw}^{+}, u_{zk}^{-}, u_{zk}^{+}, f_{zk} \geq 0 \quad \forall s \in \mathrm{S}, w \in \mathrm{W}, z \in \mathrm{Z}, k \in \mathrm{K} \tag{22}$$

$$x_{swdbr}, y_{swdbr}, j_{swdbr}, v_{sw}^{\mathrm{t}}, v_{zk}^{\mathrm{u}} \in \{0, 1\} \quad \forall s \in \mathrm{S}, w \in \mathrm{W}, d \in \mathrm{D}, b \in \mathrm{B}, r \in \mathrm{R}, z \in \mathrm{Z}, k \in \mathrm{K} \tag{23}$$

Constraints (2) force a maximum of one specialty assigned to each slot available. Constraints (3) state that a slot can only be allocated to a specialty if there is a minimum number of available surgeons. These are important constraints since in many cases a surgery must be performed by two surgeons (e.g. it is regulated by Portuguese legislation [4]). Constraints (4) prevent a surgeon to be allocated in consecutive slots. Indeed, it guarantees that, in one day, the number of slots allocated to some specialty is not higher than the number of slots potentially assigned given the number of available surgeons and the minimum number of required available surgeons. Constraints (5) define the weekly workload for surgeons. With these constraints, the number of times that a surgeon can operate and, therefore, the number of

times that a specialty can be assigned to a slot, is restricted. Similar constraints are modeled to anesthesiologists, (6)–(8). Constraints (9) impose a minimum monthly workload for each specialty. Constraints (10) model the weekly target number of allocated slots to each specialty where under and over allocation are allowed at a penalty cost minimized in the second term of the objective function. Constraints (11) and (12) guarantee that either the positive or the negative deviation is higher than zero. The multi-objective effect requires these constraints which are not necessary if only the second objective is minimized. Constraints (13) and (15) define auxiliary variables y_{swdbr} and j_{swdbr}, respectively. These variables account the differences in the MSS of some week when compared with the first week of the considered month and of some week when compared with the corresponding week in the first month of the planning horizon, respectively. These constraints are non-linear although they can be easily linearized. Constraints (14) and (16) limit the number of weekly and monthly changes, respectively, promoting stability in the MSSs. Constraints (17) define auxiliary variables f_{zk} (i.e. number of patients in each unit and day) by linking with variables x. The expected number of patients in unit z on day k of the planning horizon, f_{zk}, depends on the probability e_{zsk}, which is obtained based on the average length of stay of patients in the corresponding units in the hospital under study. Constraints (18) model the units workload as a result of OR slots assignment and deviations are penalized in the third term of the objective function. Constraints (19) bound the positive deviation to the capacity of each unit. Similar to constraints (11) and (12), constraints (20) and (21) guarantee positive or negative deviations to the target workload for each unit and day. Constraints (22) and (23) represent the domain constraints for decision and auxiliary variables.

This mathematical programming model, although motivated by the problem at the hospital under study, intends to be as general as possible. Indeed, a very common problem in hospital is the allocation of slots to specialties (resource allocation) considering the demand and incorporating the workload of other facilities when constructing the MSS. This model maximizes the total number of slots assigned, considering staff preferences and target values, while complying with the upstream and downstream units' capacity.

3 Computational Experiments

The model is applied in the context of the central surgical suite of a public hospital in Portugal. The actual MSS was kept almost unchanged over the last years. Nevertheless, the workforce availability and the surgical demand are dynamic parameters over time. The OR comprises 8 specialties (|S|=8), namely general surgery, urology, orthopedics, ophthalmology, plastic surgery, pediatric surgery, otorhinolaryngology (ORL) and stomatology. Before surgery, all patients stay in a pre-ward for a maximum of one day. Surgeries are performed from Monday to Friday (|D|=5) in two daily shifts (|B|=2). Elective surgeries are performed in four ORs (|R|=4). For each specialty, a target number of allocated slots is defined according to the demand and

Table 2 Results for the case study. Time limit: 1 h

max	Obj1Value	Obj2Value	Obj3Value	Gap	Time
Obj1	1404	−147.46	−531.48	0	1
Obj2	1248	−147.46	−532.35	0	1
Obj3	1196	−147.46	−523.25	1.01	60
(1) capacityR	1404	−147.46	−524.61	0.65	60
(1) capacity+	2288	−19.5	−438.20	0.35	60

to the average surgery duration.[1] The workforce is composed by ten anesthesiologists (|A|=10) and forty-three surgeons (|I|=43). To assign a slot to a specialty, there must be, at least, two available surgeons (δ^{surg}=2) and one available anesthesiologist (δ^{anest}=1). The minimum monthly workload in the OR for each specialy is one slot (mw_{sm}=1). Moreover, since medical staff share OR time with appointments, internment visits and the emergency unit, a maximum of one or two weekly slots is applied to each surgeon (ww_i^{surg}=1,2). Anesthesiologists share OR time with appointments and thus the maximum number of weekly slots is set to four or five (ww_a^{anest}=4,5). After surgery, patients go directly to a ward or through the intensive care unit (ICU). There are 6 down- and upstream units ($|Z| = 6$). The MSS is designed for a 1-year planning horizon.

The mathematical model is implemented in Java, with Eclipse Java Mars.2, using the callable library of ILOG CPLEX 12.8.0. All tests ran on a portable computer with an Intel Core i7-3632QM CPU of 2.20 GHz and 6 GB of RAM under Windows 10 operating system.

The results for the case study instance are shown in Table 2. In this table, 'Obj#Value' is the value for objective #, 'Gap' is an upper bound on the optimality gap (in %) based on the LP relaxation solved by CPLEX and 'Time' is the computation time (in minutes). Each single objective is considered individually ('Obj#') and function (1) is used as optimization objective. An additional case with extended surgeons capacity is also tested ('capacity+'). The model performs well in all cases obtaining an optimal solution in less than one minute when optimizing the first and second objectives individually, and providing a solution with up to 1.01% gap within 1 h computing time for the remaining cases.

Table 2 shows that the total deviation of the number of slots assigned to target values ('Obj2Value') is the same regardless of the optimized objective when the real capacity is considered. This can be explained by the maximum weekly number of OR slots that a surgeon can use in practice. Each department takes full advantage of the workforce, but the surgeons' maximum capacity is not enough to answer the demand, being a strong bottleneck in the system and avoiding achieving the target for the number of allocated slots for all specialties (Table 3). When the maximum number of slots a surgeon can be in the OR is doubled ('capacity+'), the total number

[1] $t_{sw} = slots\ available_w \dfrac{nb\ patients\ in\ the\ waiting\ list_{sw} \times av\ surgery\ duration_s}{\sum_{s \in S} nb\ patients\ in\ the\ waiting\ list_{sw} \times av\ surgery\ duration_s}$.

Table 3 Average compliance rate (in %) of target slots and average expected number of weekly operated patients

	Specialty	GEN	PLS	PDT	OPT	ORT	ORL	URO	STO	Average
Av compliance (%)	CapacityR	53	67	100	71	25	50	50	100	64.5
	Capacity+	100	100	100	100	62	100	100	100	95.3
Av expected patients per week	Actual	18	4	9	36	15	10	4	5	12.6
	CapacityR	16	8	9	30	6	10	4	5	11.0
	Capacity+	30	12	9	42	15	20	8	5	17.6
	Target	30	12	9	42	24	20	8	5	18.7

Table 4 Average utilization of up- and downstream units (in %) in number of beds with respect to the target value

	Pre-ward	ICU	Ward 1	Ward 2	Ward 3	Ward 4
CapacityR	69.85	14.43	21.63	27.30	23.35	17.69
Capcity+	92.56	18.76	33.49	46.77	33.67	17.71

of slots assigned is much closer to the target. Table 3 allows to compare the results on the compliance rate of target slots allocation when the real surgeons' capacity is considered and when an increased capacity is granted, and also the average expected number of weekly operated patients in both cases. For the latter, the actual approach and the target are also shown.

Regarding beds' utilization in the up- and downstream units, the percentage of compliance with the target value for each unit in both solutions using objective (1) is summarized in Table 4. Ward 1 is shared by general surgery, plastic surgery and stomatology, ward 2 is for general surgery and urology, ward 3 includes ophthalmology, orthopedics and ORL and ward 4 is used by pediatric specialty. This table shows that all units are underused regarding the target utilization (which is a percentage of the total capacity) meaning that the capacity in ORs and in the wards is not leveled. The wards present excess capacity, mostly in downstream units, comparing to the ORs which reinforces the idea of a bottleneck in the ORs.

4 Conclusions

This work focuses on the tactical level of decisions of OR management and proposes an integrated approach of three objectives that were already studied individually but never considered together: to adjust the allocation of slots to the surgical demand considering workforce preferences; and to balance the workload in up- and downstream units. A Portuguese public hospital is used as case study and to validate the model. The model is tested using instances based on real data provided by the hospital. Results show that the workforce is a major bottleneck in the OR management system disallowing a proper response to the increasing surgical demand. The solution is only able to assign 64.55% of the OR slots with respect to the target value, which is calculated according to the demand and the OR time capacity. Moreover, increasing the surgeries workload capacity of each surgeon allows most specialties to comply with their allocation target value (95.3%).

For further studies, the dynamic behaviour of the waiting list should be considered. Moreover, the impact of the proposed solutions on the operational level and on key performance indicators for the OR management should be tested through a simulation model. A more extensive discussion of the preferences system is also suggested.

Acknowledgements The authors acknowledge the support provided by FCT and P2020 under the project PTDC/EGE OGE/30442/2017, Lisboa-01.0145-Feder-30442.

References

1. Abdelrasol, Z., Harraz, N., Eltawil, A.: Operating room scheduling problems: a survey and a proposed solution framework. In: Kim, H.K., Ao, S.-I., Amouzegar, M.A. (eds.) Transactions on Engineering Technologies, pp. 717–731. Springer, Dordrecht (2014)
2. 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)
3. Cappanera, P., Visintin, F., Banditori, C.: Addressing conflicting stakeholders' priorities in surgical scheduling by goal programming. Fex. Serv. Manuf. J. **30**(1–2), 252–271 (2018)
4. ERS: Deliberação do Conselho de Administração da Entidade Reguladora da Saúde (ERS). Entidade Reguladora da Saúde (2016) https://www.ers.pt/uploads/document/file/14103/Vers_o_n_o_Confidencial__ERS_61-17.pdf
5. 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)
6. Beliën, J., Demeulemeester, E.: Building cyclic master surgery schedules with leveled resulting bed occupancy. Eur. J. Oper. Res. **76**(2), 1185–1204 (2007)
7. 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**(1), 143–148 (2002)
8. Blake, J.T., Donald, J.: Mount Sinai hospital uses integer programming to allocate operating room time. Interfaces **32**(2), 63–73 (2002)
9. Day, R., Garfinkel, R., Thompson, S.: Integrated block sharing: A win-win strategy for hospitals and surgeons. Manuf. Serv. Oper. Manag. **14**(4), 567–583 (2012)
10. Dellaert, N., Jeunet, J.: A variable neighborhood search algorithm for the surgery tactical planning problem. Comput. Oper. Res. **84**, 216–225 (2017)
11. Guido, R., Conforti, D.: A hybrid genetic approach for solving an integrated multi-objective operating room planning and scheduling problem. Comput. Oper. Res. **87**, 270–282 (2017)
12. Liu, Y., Chu, C., Wang, K.: A new heuristic algorithm for the operating room scheduling problem. Comput. Ind. Eng. **61**(3), 865–871 (2011)
13. Malik, M.M., Khan, M., Abdallah, S.: Aggregate capacity planning for elective surgeries: a bi-objective optimization approach to balance patients waiting with healthcare costs. Oper. Res. Health. Care. **7**, 3–13 (2015)
14. Marques, I., Captivo, M.E., Barros, N.: Optimizing the master surgery schedule in a private hospital. Oper. Res. Health. Care. **20**, 11–24 (2019)
15. Penn, M.L., Potts, C.N., Harper, P.R.: Multiple criteria mixed-integer programming for incorporating multiple factors into the development of master operating theatre timetables. Eur. J. Oper. Res. **262**(1), 194–206 (2017)
16. Samudra, M., Van Riet, C., Demeulemeester, E., Cardoen, B., Vansteenkiste, N., Rademakers, F.E.: Scheduling operating rooms: achievements, challenges and pitfalls. J. Sched. **19**(5), 493–525 (2016)
17. 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**(3), 269–282 (2007)
18. Visintin, F., Cappanera, P., Banditori, C.: Evaluating the impact of flexible practices on the master surgical scheduling process: an empirical analysis. Flex. Serv. Manuf. J. **28**(1–2), 182–205 (2016)

19. Visintin, F., Cappanera, P., Banditori, C., Danese, P.: Development and implementation of an operating room scheduling tool: an action research study. Prod. Plan. Control. **28**(9), 758–775 (2017)
20. Zhang, B., Murali, P., Dessouky, M.M., Belson, D.: A mixed integer programming approach for allocating operating room capacity. J. Oper. Res. Soc. **60**(5), 663–673 (2009)

Evaluating Replenishment Systems for Disposable Supplies at the Operating Theater: A Simulation Case Study

Karen Moons, Geert Waeyenbergh, Paul Timmermans, Dirk De Ridder and Liliane Pintelon

Abstract Ensuring cost containment while providing high quality patient care is of paramount concern to hospitals. The operating theater in particular is a major cost driver for any hospital, and is among the most critical resources in terms of both capacity and patient care. Effective inventory and distribution systems are a prerequisite for realizing efficiency improvements in the internal operating theater supply chain. In this work, discrete-event simulation is used to model part of the internal distribution process in the operating theater at a Belgian Hospital and to identify improvements by focusing on the replenishment process. A logistics performance measurement framework based on Analytic Network Process (ANP), as a popular Multi-Criteria Decision-Making (MCDM) technique, is adopted to assess three replenishment scenarios. The best performing scenario is selected using the Internal Logistics Efficiency Performance (ILEP) index as an evaluation basis. This research indicates that industrial engineering techniques, such as simulation and MCDM, which are successfully applied in industrial sectors, can also be adopted to realize efficiency opportunities in healthcare logistics.

Keywords Healthcare · Supply chain management · Replenishment systems · Simulation · Performance management · Multi-criteria decision-making

K. Moons (✉) · L. Pintelon
Center for Industrial Management/Traffic and Infrastructure, KU Leuven, Louvain, Belgium
e-mail: Karen.Moons@kuleuven.be

G. Waeyenbergh
Research Group Sustainable Engineering, Department of Engineering Technology, KU Leuven, Louvain, Belgium

P. Timmermans
Logistics and Supply Chain Management Department, UZ Leuven, Louvain, Belgium

D. De Ridder
Department of Development and Regeneration, UZ Leuven, Louvain, Belgium

V. Bélanger et al. (eds.), *Health Care Systems Engineering*,
Springer Proceedings in Mathematics & Statistics 316,
https://doi.org/10.1007/978-3-030-39694-7_12

1 Introduction

Efficiently controlling and distributing medical-surgical supplies to the operating theater (OT) is essential to lower costs without sacrificing patient care quality. The OT takes up 40–60% of total hospital supply expenditures [1], and is thus a major cost center and a primary target for materials management improvement. The internal supply chain of the OT, involving product and information flows from receiving, picking and replenishing, is identified as the weak link in supply chain integration [2, 3]. The main goal of healthcare Supply Chain Management (SCM) is to achieve a well-coordinated system and support clinical activities by providing the right materials in the right quantity at the predestined place and time, while reducing costs. Assessing efficiency in healthcare logistics is gaining attention as demonstrated by the growing body of literature [4]. Volland et al. [5] provide an overview of quantitative methods applied to four streams of hospital SCM: supply and procurement, inventory management, distribution and scheduling, and holistic SCM. Camp et al. [1] suggest strategies to increase healthcare supply chain performance, including inventory control, standardization and physician preference card management. From a distribution perspective, resource coordination, effective delivery strategies and efficiently managing care delivery services are a necessity to provide high-quality patient care [6].

In this paper, we investigate the internal distribution practices at the OT at the University Hospital in Leuven (UZ Leuven) by focusing on replenishment of decentral storage locations within the OT. The goal is to assess alternative replenishment policies in the OT to identify efficiency targets. Discrete-event simulation is used to model the replenishment scenarios.

2 Literature Review

Manufacturing and retailing industries have successfully applied SCM, resulting in substantial benefits and cost savings. In contrast, the healthcare sector is struggling to adopt these concepts [7, 8]. The supply chain activities represent 30% of total hospital costs and inventory costs take up 10–18% of net revenues [1, 9]. Efficiently managing the supply chain can significantly reduce costs and waste, while meeting patient care service levels [9]. The main challenge in healthcare SCM is trading-off costs with inventory levels to sustain quality and timely patient care [10].

Effective inventory control lowers the cost of internal distribution by for instance improving inventory turnovers. Optimal replenishment policies need to be developed to maximize service levels in hospitals. Bijvank and Vis [11] investigate the impact of inventory models on medical supplies by establishing simple replenishment rules to determine reorder levels and reorder quantities. Rosales et al. [12], Di Martinelly [13] combine periodic and continuous replenishments in a hybrid two-bin inventory policy. The authors analyze the impact on costs, inventory levels and replenishment

frequency at point-of-use locations. Landry and Beaulieu [14] use the (de)centralized material management system and the review period to determine the appropriate distribution policy for replenishment of point-of-use locations. Centralizing the logistics processes will lower both costs and inventory levels due to more frequent replenishments, reduce workload for logistics staff and provide a higher service level [15]. The impact of decentral inventory locations on the performance of replenishment systems in nursing units is investigated in the work of Bélanger et al. [3], by focusing on logistics staff productivity. The objective of workload balancing is considered by Lapierre and Ruiz [16], who develop a scheduling approach to coordinate distribution activities while respecting inventory capacities. The schedules determine the delivery frequency, the products to be replenished, as well as task assignment to improve personnel management.

The organization of the internal distribution within the OT is complex, as different types of medical-surgical supplies are stored in various central and decentral storage locations. The models discussed in literature use several decision variables to determine the impact of distribution policies at point-of-use locations. In this work, we will integrate these indicators to assess the performance of the distribution system which depends on multiple criteria, such as cost, service level, personnel satisfaction, degree of centralization, etc. The complex problem structure, however, makes analytical models impractical as they require many assumptions to simplify modelling, making models unrepresentative of reality [8]. Therefore, simulation is used as a tool to support decision-making in healthcare logistics. This work contributes to literature as simulation is mostly applied to study patient flows rather than hospital material logistics, and multiple performance indicators are identified to evaluate the multi-dimensional character of logistics in healthcare.

3 Methodology

The OT supply chain is a complex system featuring many elements and parameters. A simulation case study is an effective decision-support approach to understand the impact of replenishments systems on the performance of the OT and to identify efficiency improvement opportunities.

3.1 Internal Replenishment Process at Operating Theatre

This study is conducted at the OT of UZ Leuven. The OT is divided into seven clusters, depending on medical disciplines, counting 33 operating rooms (ORs) in total. One cluster is typically equipped with four ORs. Commonly used disposable supplies are held in decentral storage locations within the OR and are replenished from central storage rooms within the hospital. Currently, a periodic review replenishment policy is used with a fixed stocking limit, based on gut feeling. However,

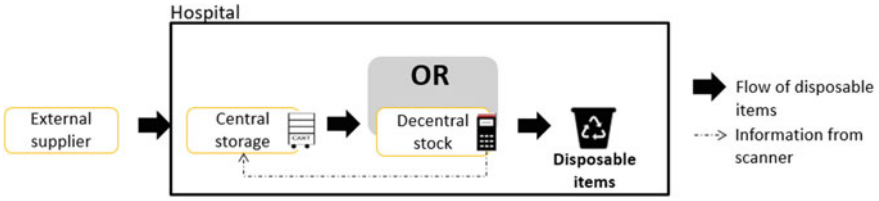


Fig. 1 Internal operating theatre supply chain of disposable supplies

stock-outs, imperfect order fulfilment, stock duplication, lack of standardization and centralization, and ergonomic burden for logistics staff are typical problems. With this study, we aim to improve the efficiency of the internal replenishment process at the OT, focusing at the flow of disposable supplies (e.g. surgical drapes, gloves, syringes). Figure 1 shows the scope of this research.

3.2 Experimentation: Scenario Analysis

Three scenarios, representing alternative replenishment systems for disposable supplies, are introduced. In the As-Is scenario, the decentral stock is replenished using copy carts from the central stock (see Fig. 1). These carts contain a duplicate of all disposables stored in decentral stock, meaning that the carts are a permanent double stock. The great amount of items on these carts cause heavily loaded carts, which increases ergonomic burden for logistics staff. Furthermore, all information exchange is paper-based, meaning that no record is made of what has been consumed. This complicates effective inventory control, requiring consumption data of each surgery. The As-Is scenario serves as the baseline scenario throughout the rest of the paper. The alternative scenarios represent two potential logistics improvements, both featuring barcode scanners. The Standard scenario eliminates the copy carts to reduce the amount of stock in circulation. Barcode scanners are introduced to scan decentral stock (see Fig. 1). Based on the scanning data, the requested items are picked in central stock to replenish decentral stock. Moreover, scanning enables collecting daily consumption data in ORs, which is essential for inventory optimization in a next research phase. A disadvantage of this scenario, however, is that decentral locations are not immediately restocked, resulting in more visits and higher incomplete refills (i.e. location cannot be refilled to the stocking limit). This drawback is countered by the Copy carts scenario, which features both barcode scanners and copy carts. Three copy carts are used to immediately replenish decentral stock. The copy carts hold a selection of commonly used items to avoid heavily loaded double carts. In addition, scanning enables daily consumption data on the copy cart level. Table 1 gives an overview of the replenishment scenarios.

Table 1 Overview of replenishment scenarios

	As-Is	Standard	Copy carts
No double stock	X	✓	X
Barcode scanner	X	✓	✓
Immediate replenishment	✓	X	✓
Consumption data	X	✓ (OR level)	✓ (copy cart level)

3.3 Prioritization of KPIs

Simulation is used to assess the replenishment scenarios based on the logistics performance management framework developed by Moons et al. [17]. Input from all stakeholders (medical staff, materials managers, etc.) is required to select key performance indicators (KPIs) for assessing replenishment policies, since they may have conflicting objectives regarding efficiency [13]. MCDM is a useful approach for prioritizing KPIs when multiple, possibly conflicting, quantitative and qualitative criteria must be considered [2]. Several authors propose Analytic Hierarchy or Network Process (AHP/ANP), developed by Saaty [18–20] as a valuable MCDM tool for performance management in hospitals [17, 21, 22]. In Moons et al. [17], the weights attributed to the KPIs are obtained by eliciting expert judgements through the ANP methodology. One expert, the OT logistics manager, is included in this case study serving as a pilot for further research. For more detailed information on the ANP methodology, the interested reader is referred to the papers by Saaty [18–20] and Moons et al. [17].

The KPIs are related to four main objectives—quality, time, financial and productivity/organization—to assess the replenishment process according to the logistics performance management framework [17]. More information on the KPIs can be found in Appendix 1. According to Table 2, quality (0.32) and productivity/organization (0.48) are the most important objectives with the greatest contribution to improving the internal distribution process. The most critical indicators are Distribution Service Level (DSL), Personnel Management (PM) and Delivery Frequency (DF). Finally, the ILEP (Internal Logistics Efficiency Performance) index is introduced serving as an evaluation basis to choose the scenario that is best performing, based on performance scores on the four objective dimensions.

3.4 Simulation Model

We use simulation to model replenishment policies in the OT, because of the advantages of simulation in solving complex problems [8]. Discrete-event simulation provides the flexibility of evaluating multiple KPIs, and can be used as a decision-support tool for evaluating the efficiency of the logistics processes. The models are developed

Table 2 KPI scores and ANP weights

Metric/scenario		Standard	Copy carts	As-Is	Normalized values		
					Standard	Copy carts	As-Is
Quality (0.32)	Distribution service level (0.136)	12.86	16.04	13.32	97.0	115.3	100
	Delivery accuracy (0.092)	78.67	35.33	139.69	56.3	25.3	100
	Centralization (0.091)	5963.31	8836.84	7480.47	39.9	109.1	100
	Total quality score				69.0	87.6	100
Time (0.15)	Replenishment lead time (0.057)	219.53	202.21	373.46	58.9	54.2	100
	Response time (0.053)	11 h 41	11 h 29	14 h 13	80.7	79.9	100
	Clinical staff involvement (0.036)	171.42	197.14	105.11	163.1	187.6	100
	Total time score				92.5	96.4	100
Financial (0.06)	Inventory cost	2798.05	3907.12	3416.97			
	Total financial score				81.9	114.3	100
Productivity/organization (0.48)	Delivery frequency (0.121)	307.89	279.03	293.67	83.7	49.8	100
	Standardization (0.16)	291	291	291	100	100	100
	Personnel management (0.121)	0.46	0.42	1.78	25.8	23.8	100
	Total productivity score				72.8	61.9	100
Total	Total KPI score				75.8	79.1	100
	ILEP index				24.2	20.9	0

using Arena® Simulation Software. The scenarios are translated into three distinct models, using a similar modular logic and sharing most variables. The models contain entities representing copy carts, resources or logistics staff, attributes (e.g. type of disposable supply), variables (e.g. current inventory level) and queues. Appendix 2 displays the Standard scenario in its conceptual form. Four sections are highlighted representing the core activities: (1) scanning decentral stock (i.e. calculate reorder quantity by comparing the maximum stock limit and the actual stock), (2) uploading scanning data (i.e. waiting for the printer to print the list with scanned items and transfer the list to the central storage room), (3) assembling case carts (i.e. picking items in central storage room) and finally (4) replenishing decentral stock.

Data collection

The simulation model requires two types of data. During a ten-day observation period, a time study is performed on replenishment jobs (e.g. scanning, picking and replenishing times). Due to the stochastic nature of these process times related to human-performed tasks, the time data fit statistical distributions (e.g. Beta distribution). The second type of data deals with daily item usage of decentral stock. The introduction of barcode scanners allows collecting data over a year (August 2017–July 2018). In total, 264 disposable supplies stored at 454 locations are included in the simulation study. A Poisson distribution is fit to the data, expressing the probability of items consumed in one day.

Verification and validation

Attention must be paid to verification and validation, ensuring that the model performs as intended to the modelling assumptions and behaves the same as the real system, respectively [23]. Due to lack of accurate data, we focus on best judgment of the expert to explore the model as thoroughly as possible. First, adding animation, by showing the (de)central storage rooms, logistics staff, copy carts and a clock, facilitates communication with various stakeholders and increases confidence in the simulation results. Second, graphical visualizations aid in monitoring the values of various sub-indicators, such as stock-outs, incomplete refills and items in stock per item per storage room (see Fig. 2). Finally, task completion times and workload distribution are validated by comparing them to the real system. The logistics manager is knowledgeable about the direction of the output behavior and knows the acceptable value range of the magnitudes. Graphs are commonly used to check for operational validity when statistical assumptions cannot be satisfied or when there are insufficient data. Based on these graphs, the model developer and system expert decide the model accuracy is within its acceptable range for its intended purpose, namely evaluating three replenishment scenarios to identify the best-performing one.

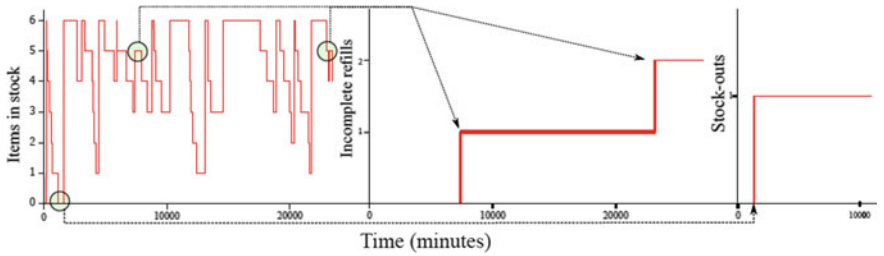


Fig. 2 Simulation output visualizing items in stock, incomplete refills and stock-outs for item FA366823 in OR 2 of Cluster D

4 Results and Interpretation

The computer models are used to assess three replenishment policies in order to realize efficiency opportunities. The simulation output provides a score for the indicators, which is normalized across all scenarios using the As-Is scenario as a baseline. By aggregating the normalized values into a single score for each objective based on ANP weights, the best-performing scenario is selected on the level of one objective. A lower score is better. Table 2 shows the values for these KPIs in each scenario as well as the ILEP index to select the appropriate policy.

4.1 Results

The Standard scenario provides the best quality, due to the high weight attributed to the number of stock-outs, as this is a critical factor in measuring distribution service level. Furthermore, delivery accuracy is performing badly in the As-Is scenario due to the time gap between making picking lists and replenishing stock, which causes incomplete refills. The Copy carts scenario eliminates this time gap by immediately restocking decentral locations. Finally, the impact of centralization is investigated to free up space for primary services in the OT. The Standard scenario eliminates copy carts, and benefits from adjusting decentral stocking limits.

In addition, the model keeps track of the time required to perform replenishment activities by logistics staff. The time related KPIs in Table 2 indicate the best (i.e. lowest) score for the Standard scenario. However, the results should be carefully interpreted. Although all items are delivered before the end of the day, the average delivery time in the As-Is scenario is 14 h 13. Both the response time and the replenishment lead time are considerably higher in the As-Is scenario due to discontinuities in the replenishment process. These interruptions are mainly caused by the involvement of Central Sterilization Department (CSD) employees, who are not included in the study. This explains the low value for clinical staff involvement, which refers

to logistics employees. The Standard and Copy carts scenarios only use logistics employees to perform replenishing activities.

From a financial perspective, an important indicator for assessing replenishment policies is the inventory holding cost. Table 2 shows a 20% reduction in holding cost in the Standard scenario, whereas the Copy carts scenario incurs a 15% increase. The lower holding cost in the former scenario is a result of eliminating copy carts, which carry permanent double stock. The Copy carts scenario requires a careful selection of items to be held on the copy cart in order to reduce holding costs. This should be done in collaboration with the nursing staff. Although costs are an important indicator to evaluate inventory policies, the financial factor has the least impact (0.06) on the replenishment process performance according to the ANP ranking.

Furthermore, productivity and organizational KPIs aid in streamlining processes and improving staff satisfaction. Delivery frequency impacts productivity by the number of items that need replenishment and the visits to decentral locations. On average, 40% of decentral item locations are daily replenished. The Copy carts scenario has the best total score for productivity, mainly due to less frequent visits to decentral stock. A recent trend in healthcare management emphasizes process standardization to simplify workflows. Standardization stimulates increasing the number of items that can be scanned, creating uniform replenishment practices throughout the entire OT. Based on accurate consumption data, the logistics team can use evidence-based arguments for standardizing items in consultation with surgeons, thus decreasing product variety and facilitating inventory control.

Finally, an index is developed by converting the scores of the four KPIs into a single score using ANP priorities, demonstrating the Internal Logistics Efficiency Performance for each scenario. This ILEP index provides an objective evaluation basis for logistics managers to identify the best strategy for managing internal healthcare logistics processes. The higher this index, the better the performance of the process. The ILEP index is added to Table 2. Overall, we observe that the Standard scenario provides the best service quality at the lowest cost, although it is slightly more demanding for logistics staff compared to the Copy carts scenario. However, making decisions based on a single score may be misleading. Therefore, the logistics manager should always consider the trade-off between multiple criteria when choosing an appropriate replenishment policy.

4.2 Interpretation

The performance management framework supports decision-making when selecting the most appropriate scenario for organizing the internal OT replenishment process. KPIs are identified in the framework and serve as a guideline for monitoring the performance of internal healthcare logistics processes. Both the Standard and Copy carts scenarios demonstrate better performance in terms of quality, time and productivity/organization compared to the As-Is scenario. From a financial perspective, the Standard scenario is preferred due to lower inventory holding costs.

When analyzing the findings of this case study, some limitations must be considered. One limitation is related to the sample size of the ANP application. This work presents preliminary results of a healthcare logistics performance management framework as only one stakeholder is included, namely the OT logistics manager expressed his preferences for KPIs. However, multiple stakeholders with possibly conflicting goals are involved in the OT supply chain, which complicates efficient performance management. A single decision-maker's attitude introduces bias in the outcome, which need to be resolved in further research by including different stakeholders allowing for group decision-making and improving the robustness of the framework. The main challenge in group decision-making will be to deal with possibly conflicting perspectives among different stakeholders. Consensus, compromise, or geometric mean are techniques for aggregating group member judgments [20, 24]. Finally, ANP sets priorities according to the different stakeholders' preferences [12, 17]. Second, the accuracy of the data used in the simulation model must be critically assessed. The input data contain 264 of 404 (65%) unique SKUs stored at 454 of 743 (61%) supply locations in decentral stock, representing 98% of total items ordered. The remaining items are excluded due to limited scanning data for various reasons, such as scanning problems, new items, backorders or IT-related problems. For inventory optimization purposes, accurate data is essential to determine appropriate stock limits and reorder points in order to minimize costs and maximize service levels. Despite these limitations, this case study provides insights on the impact of three replenishment scenarios on the workflow, quality and costs of the operating theater.

This section proposes recommendations to improve the internal OT supply chain processes in further research. An interesting follow-up study will focus more on simulation of system parameters to decide on appropriate stock limits, as 16% of stock-outs are caused by 12% of items with a stock limit of one. A "What-if?" analysis is performed to examine the impact on the inventory cost and number of stock-outs. By increasing the stock limit of those items by one, the number of stock-outs decrease on average with 13% while inventory costs increase with roughly 5% due to higher average stock levels. However, having excess items in stock increases both the workload for logistics employees and the costs. Further research should address this main trade-off in inventory control, namely balancing costs and service level, by focusing on inventory parameter optimization in central and decentral locations. In addition, dealing with physician preferences, causing large product variety, by implementing item standardization and engaging physicians in updating preference cards may increase OT supply chain efficiency and facilitates inventory control [25]. In further research, the simulation model can be extended by including other clusters, thereby improving the generalizability of the findings. The ILEP index for one cluster shows a 24% improvement for the Standard scenario compared to the As-Is situation. Extension to all clusters in the OT will reveal even higher improvement opportunities. In addition, the methodology can be customized to healthcare logistics problems using the ILEP index as an objective evaluation basis for assessing logistics scenarios. Hospitals using a performance management framework have

a competitive advantage. The corresponding ILEP index enables hospitals to control their supply chain strategy, implement continuous improvement programs and improve decision-making by focusing on relevant KPIs [26].

5 Conclusion

Cost containment in healthcare is of increasing concern. Streamlining OT supply chains is critical to provide superior, low-cost patient care. This study analyzed the current logistics workflow in one operating room cluster by focusing on replenishment of disposable supplies. Overall, we demonstrate that two replenishment scenarios both result in efficiency improvements compared to the current As-Is situation. The Standard scenario provides a better service than the As-Is at lower costs, whereas the Copy carts scenario improves the logistics staff productivity. However, the latter scenario contains a higher inventory cost due to improper selection of items carried on copy carts. In addition, both these scenarios increase data collection through scanning, resulting in valuable information for inventory control.

Ultimately, the logistics manager should consider the trade-off between service, time, costs and productivity when defining efficiency targets, and discuss this trade-off with medical staff, as they have critical insights into the challenges of achieving an operationally efficient workflow, supporting care delivery. In addition, enabling accurate data collection through scanning is important to monitor performance. Decision support models based on ANP provide useful information for managerial decision-making in relation to performance management. The logistics performance management framework by Moons et al. [17] enables hospitals to control their supply chain strategy, improve supporting services and realize significant efficiency opportunities.

With this research, we prove that industrial engineering concepts applied in industrial sectors are also useful to be adopted by the healthcare sector to identify efficiency opportunities in the internal supply chain processes. Further research will focus on other dimensions determining the replenishment process performance such as optimizing inventory levels, reorder frequency and stock-out rate. In addition, the robustness of the framework will be validated by including multiple stakeholders with possibly conflicting goals for efficiency improvements.

Appendix 1

Quality	Quality specifies how well a specific activity has been performed, ensuring that patients receive care service in a safe manner and that problems such as medical errors are minimized

(continued)

(continued)

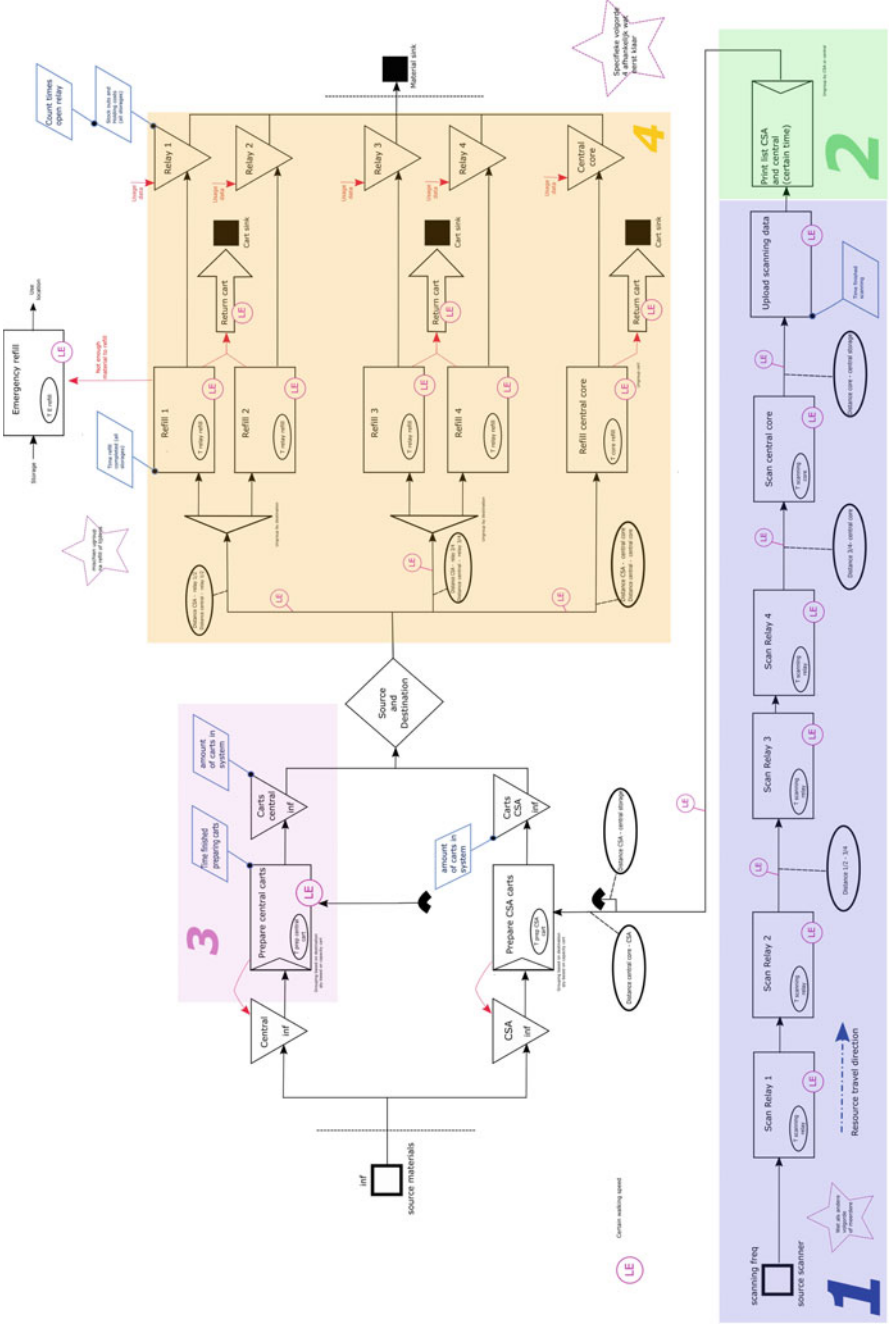
	Distribution service level (DSL)	The availability of logistics services to support clinical care processes	
		Urgent delivery rate	Daily stock − out rate $= \frac{\sum_{i=1}^{i=454}(\text{Stockout}_i)}{730\,\text{days}}$
		Additional items needed	Average replenishment per item = AvgMaxStock − $\frac{\text{Average daily replenishment}}{\text{Average replenishment items per day}}$
	Delivery accuracy (DA)	The ability to pick and deliver the correct items and quantities from storage to point-of-use location	
		Perfect order fulfilment	Daily number of incomplete refills = $\frac{\sum_{i=1}^{i=454}(\text{MaxStock}_i - \text{ItemsInStock}_i)}{730\,\text{days}}$
	Centralization impact (CI)	The ability to locate items only at a central storage room, or also at decentral storages.	
		Permanent double stock	Number of copy carts, containing duplicate of items in decentral storages
		Impact of centralization	Adjust max stock in decentral locations
Time	Time involves the time to complete the logistics operations to ensure that the right items are at the right place and time		
	Replenishment lead time (RLT)	The total amount of time that elapses from the moment an item is ordered until the item is back on the shelf	
		Transport time	Time to move items to the right place
		Replenishing time	Time to replenish decentral stock
		Scanning time	Time to scan items in decentral stock
		Preparing time	Time to pick requested items from central stock
		Other activities time	Time spent on other activities than replenishing due to interruptions
		Replenishment lead time	= Transport + replenishing + scanning + preparing + other activities
	Response time (RT)	The ability to deliver items on time, preventing delays in surgical procedures	
		On-time delivery	Average delivery time = finish time of replenishing decentral stock
	Clinical staff involvement (CSI)	The amount of time clinical staff is busy with logistics tasks, rather than their core activities	
		Logistics employees involvement	Time spent by logistics employees (=RLT—other activities)
Financial	Financial indicators identify supply chain cost drivers, such as expenses incurred by departments for providing services, including direct and overhead costs for inventory and internal distribution		
	Distribution cost (DCo)	Total cost of handling and transporting to move supplies from storage rooms to point-of-care locations	
		Replenishing cost	Related to RLT

(continued)

(continued)

	Personnel cost (PCo)	The cost related to the time personnel is involved with logistics activities	
		Personnel cost	Related to CSI
	Inventory cost	The annual cost of holding inventory at a specific storage room	
		Holding cost	Average holding cost $= [\sum_{i=1}^{i=454} \left(\frac{\text{ItemsInStock}_i}{\text{InventoryCount}}\right)] * \text{UnitCost}_i * 0.25$
Productivity/organization	Productivity/organization involves operational control metrics for logistics departments used for streamlining processes, reducing costs, facilitating information flow and enhancing provided care services		
	Case cart efficiency (CCE)	The availability and utilization of case carts to provide surgeons with the required supplies	
		Not applicable for replenishment process	
	Delivery frequency (DF)	The number of visits to decentral storage locations to deliver or replenish items in these locations	
		Percentage of items replenished	Daily percentage of item replenishment $=$ Total items $- \frac{\text{Daily number of items replenished}}{\text{Total items included}}$
		Scanning frequency	Use of scanner (0/1)
		Visits to decentral locations	Number of opening relay cabins
	Standardization (S)	The ability to simplify workflows between operating rooms and improve working conditions	
		Percentage of scannable items for replenishment	$=$ Total items $- \frac{\text{Number of scannable items}}{\text{Total items}}$
	Personnel management (PM)	A measure of how to obtain, use and maintain a satisfied workforce	
		Personnel utilization	$= \frac{\text{Time busy replenishing (RLT)}}{480\,\text{min}}$
		Ergonomics friendliness	Use of double carts (0/1)
		Workload distribution	Timeline of logistics employees interrupted by other activities

Appendix 2


Emergency refill
Refill 1
Refill 2
Refill 3
Refill 4
Refill central core
Relay 1
Relay 2
Relay 3
Relay 4
Central core
Return cart
Cart sink
Material sink
Source and Destination
Prepare central carts
Prepare CSA carts
Central inf
CSA inf
Carts central inf
Carts CSA inf
amount of carts in system
time finished preparing carts
inf
source materials
Scan Relay 1
Scan Relay 2
Scan Relay 3
Scan Relay 4
Scan central core
Upload scanning data
Print list CSA and central (certain time)
scanning freq
source scanner
Resource travel direction
1
2
3
4

References

1. Camp, M., Pfister, J., Reeves, D., Kneedler, J.: Effective Operating Room Inventory Management. Pfiedler Enterprises 26 (2014)
2. Baltussen, R., Niessen, L.: Priority setting of health interventions: the need for multi-criteria decision analysis. Cost Effectiveness Resour. Allocation **4**, 14 (2006). https://doi.org/10.1186/1478-7547-4-14
3. Bélanger, V., Beaulieu, M., Landry, S., Morales, P.: Where to locate medical supplies in nursing units: An exploratory study. Supply Chain Forum Int. J. (2018). doi: https://doi.org/10.1080/16258312.2018.1433438
4. Moons, K., Waeyenbergh, G., Pintelon, L.: Measuring the logistics performance of internal hospital supply chains—a literature study. Omega (United Kingdom) (2018)
5. Volland, J., Fügener, A., Schoenfelder, J., Brunner, J.O.: Material logistics in hospitals: a literature review. Omega (United Kingdom) **69**, 82–101 (2017). https://doi.org/10.1016/j.omega.2016.08.004
6. Rohleder, T., Bailey, B., Crum, B., Faber, T., Johnson, B., Montgomery, L., Pringnitz, R.: Improving a patient appointment call center at Mayo Clinic. Int. J. Health Care Qual. Assur. **26**, 714–728 (2013). https://doi.org/10.1108/IJHCQA-11-2011-0068
7. Hicks, C., McGovern, T., Prior, G., Smith, I.: Applying lean principles to the design of healthcare facilities. Int. J. Prod. Econ. **170**, 677–686 (2015). https://doi.org/10.1016/j.ijpe.2015.05.029
8. Hu, Q., Boylan, J.E., Chen, H., Labib, A.: OR in spare parts management: a review. Eur. J. Oper. Res. (2018)
9. Lanckzweirt, J.: Een Analyse van de Materiaalstromen in het Operatiekwartier (2010)
10. Abukhousa, E., Al-jaroodi, J., Lazarova-molnar, S., Mohamed, N.: Simulation and modeling efforts to support decision making in healthcare supply chain management. Sci. World J. **2014**, 16 (2014). https://doi.org/10.1155/2014/354246
11. Bijvank, M., Vis, I.F.A.: Inventory control for point-of-use locations in hospitals. J. Oper. Res. Soc. **63**, 497–510 (2012). https://doi.org/10.1057/jors.2011.52
12. Marsh, K., Goetghebeur, M., Thokala, P., Baltussen, R.: Multi-criteria decision analysis to support healthcare decisions (2017)
13. Di Martinelly, C.: Proposition of a framework to reengineer and evaluate the hospital supply chain. Department of Management 139 (2008)
14. Landry, S., Beaulieu, M.: The challenges of hospital supply chain management, from central stores to nursing units. In: International Series in Operations Research and Management Science, pp 465–482 (2013)
15. Carrus, P.P., Marras, F., Pinna, R.: The performance measurement of changes in the logistics of health goods: a theoretical model. In: Proceedings of the 18th Toulon-Verona International Conference 85–100 (2015)
16. Lapierre, S.D., Ruiz, A.B.: Scheduling logistic activities to improve hospital supply systems. Comput. Oper. Res. **34**, 624–641 (2007). https://doi.org/10.1016/j.cor.2005.03.017
17. Moons, K., Waeyenbergh, G., Pintelon, L., Timmermans, P., De Ridder, D.: Performance indicator selection for operating room supply chains: an application of ANP. Oper. Res. Health Care Under Revi., 1–25 (2019)
18. Saaty, T.L.: A scaling method for priorities in hierarchical structures. J. Math. Psychol. **15**, 234–281 (1977). https://doi.org/10.1016/0022-2496(77)90033-5
19. Saaty, T.L.: How to make a decision: the analytic hierarchy process. Eur. J. Oper. Res. **48**, 9–26 (1990). https://doi.org/10.1016/0377-2217(90)90057-I
20. Saaty, T.L., Vargas, L.G.: Decision making with the analytic network process. Economic, political, social and technological applications with benefits, opportunities, costs and risks. Int. Ser. Oper. Res. Manag. Sci. (2006). https://doi.org/10.1007/978-1-4614-7279-7
21. Hariharan, S., Dey, P.K., Moseley, H.S.L., Kumar, A.Y., Gora, J.: A new tool for measurement of process-based performance of multispecialty tertiary care hospitals. Int. J. Health Care Qual. Assur. **17**, 302–312 (2004). https://doi.org/10.1108/09526860410557552

22. Soriya Hoeur, Duangpun Kritchanchai (2015) Key Performance Indicator Framework for Measuring Healthcare Logistics in ASEAN. Toward Sustain. Oper. Supply Chain Logistics Syst. doi: https://doi.org/10.1007/978-3-319-19006-8
23. Sargent, R.G.: Verification and validation of simulation models. J. Simul. **7**, 12–24 (2013). https://doi.org/10.1057/jos.2012.20
24. Dyer, R. F., Forman, H.: Group decision support with the analytic hierarchy process. Decis. Support Syst. **8** (2), 99–124 (1992)
25. Harvey, L.F.B., Smith, K.A., Curlin, H.: Physician Engagement in improving operative supply chain efficiency through review of surgeon preference cards. J. Minim. Invasive Gynecol. (2017)
26. Maestrini, V., Luzzini, D., Maccarrone, P., Caniato, F.: Supply chain performance measurement systems: a systematic review and research agenda. Int. J. Prod. Econ. (2017)

Stochastic Master Surgical Scheduling Under Ward Uncertainty

Asgeir Orn Sigurpalsson, Thomas Philip Runarsson and Rognvaldur J. Saemundsson

Abstract In this work, we address the elective surgery scheduling problem and the risk of last-minute cancellations. This risk is associated with the likelihood of operating rooms going into overtime and ward beds exceeding their limit. The risk of overtime is constrained by considering only feasible combinations of operating room days schedules. To account for the feasibility, we restrict the number of surgeries assigned to each combination and force it to maintain the correct ratio between in- and out-patients for each operator. Furthermore, the probability of running into overtime is bound and verified using Monte-Carlo simulation. The risk of exceeding the ward limit is solved by a mixed-integer programming model where the probability of going over the available ward beds downstream is bound. The approach is inspired by real challenges and tested on real-life hospital data.

Keywords Surgery scheduling · Stochastic integer programming · Monte Carlo simulation

1 Introduction

Due to the aging of the population and increasing cost of care, hospital managers seek to maximize the use of existing resources. One such resource that has received considerable attention are operating rooms [1, 2]. Operating rooms (ORs) are expensive resources but also a significant source of income [3]. However, the capacity of downstream resources, such as intensive care units and wards, constrain the patient flow from the ORs [4, 5]. Therefore, to gain from increased utilization, surgeries need to

A. O. Sigurpalsson (✉) · T. P. Runarsson · R. J. Saemundsson
University of Iceland, Hjardarhagi 6, Reykjavik 107, Iceland
e-mail: aos13@hi.is

T. P. Runarsson
e-mail: tpr@hi.is

R. J. Saemundsson
e-mail: rjs@hi.is

V. Bélanger et al. (eds.), *Health Care Systems Engineering*,
Springer Proceedings in Mathematics & Statistics 316,
https://doi.org/10.1007/978-3-030-39694-7_13

be scheduled in such a way to assure a balanced patient flow for avoiding last-minute cancellations due to downstream bottlenecks. This is a challenging problem due to the stochastic nature of patient admittance, surgery times, length of stay [4, 6] and the competing objectives of the various stakeholders involved in the surgery process [7].

In practice, one can divide surgery scheduling into three decisions levels: strategic, tactical, and operational [8]. At the strategic level, decisions are made about the overall surgical capacity, including ORs and equipment. These decisions are long term. At the tactical level, decisions are made about the availability of OR time, e.g., by temporarily closing ORs or increasing staffing, and how the available OR time is allocated to surgical specialties and operators. These allocations, which may be cyclic, are usually set months in advance. Lastly, at the operational level, patients waiting for elective surgeries are scheduled to a room, day, and time, based on the allocation of the OR time to the specialty or operator in question. This may be done some weeks in advance.

The schedule created at the tactical level is referred to as a Master Surgical Schedule [4, 9]. These schedules may refer to finding an ideal mix of surgical procedures for each day [9, 10]. Furthermore, many approaches have been presented for building them. In [11] an integer programming (IP) model is applied and a post-solution heuristic to minimize the difference between the target and assigned OR time for all specialties, however, without taking the post-operation patient flows into consideration. In [3, 12] and [9] models take into account a varying degree of the stochastic nature of the problem and the constraints posed by downstream resources. Nevertheless, they either assume a given flow of patients reducing their practical use for maximizing throughput while avoiding last-minute cancellations. In [4] a stochastic IP model is proposed, which both accounts for the stochasticity of surgery times and length of stay. The model is solved with a sample average approximation where the goal is to maximize the expected throughput. To maximize the throughput, the authors assign surgery groups to days and rooms for each specialty. However, the assignment of the surgical operators is not considered as it might be no issue in their study.

In this paper, we suggest a new approach for building a Master Surgical Schedule (MSS). First, the schedule refers to a cyclic allocation of OR time to surgical operators. In effect, this takes both the specialty and the mix of elective surgical procedures into account. Second, the schedule is based on historical data, which allows the continuous monitoring of the effectiveness of the schedule in use. Third, to create the schedule, we use an approach to the operational surgery scheduling, we termed Pattern Scheduling [13]. This allows us to account for the stochastic nature of the problem in a practical way and precisely. Fourth, we use an approximation that allows us to bound the likelihood of exceeding the ward capacity. To the best of our knowledge, this has not been attempted before. Often, the expected ward numbers are used. To demonstrate and test our approach, we build and evaluate a MSS using historical data from the National University Hospital of Iceland (Landspitali).

The paper is organized as follows. In the next section, we provide a problem description. The following two sections specify the two steps of our pattern scheduling approach, followed by an experimental study based on historical data. In the final section, we discuss the results and provide conclusions.

2 Problem Description

Landspitali hospital has pre-assigned ORs $r \in R$ for elective surgeries and days $d \in D$ under a weekly planning horizon T consisting of five days. Each day and room represents a single *session* allocated to one operator. The goal is to find the optimal assignments of the surgical operators $o \in O$ such that throughput is maximized. To tackle this problem, the elective surgical case assignment problem is solved for each session using the patients most likely to appear within the planning horizon.

Each patient is assigned to a single operator, and each operator is assigned to at least one day in the weekly planning horizon. Patients may then be assigned to one of these days assigned to their operator. Planning too many patients for any given session may result in overtime, but this is not necessarily undesirable. However, too many sessions in overtime the same day may result in last-minute cancellations. The same applies to days exceeding the available ward beds. Bounding the expected number in the ward can be formulated in a straight forward manner. However, formulating the risk or likelihood of exceeding the ward capacity is more challenging and will be attempted here.

A two-phase approach is used to build a cyclic MSS, for elective surgeries, that minimizes the risk of cancellations due to overtime and ward capacity. In the first phase, all feasible patterns of operator room day schedules are created. In the second phase, these patterns are used by a Mixed Integer Programming (MIP) model that addresses the ward restriction in a probabilistic manner.

3 Pattern Generation

The first step in our approach is the pattern generation. In this context, we define a pattern as a feasible session or one-day assignment of patients to each surgical operator. To generate the patterns p, one can use enumeration. However, not all of them are feasible. Thus, we set up a three-step procedure to account for feasibility. Let us define a binary indicator $z_{i,p}$ taking the value 1 if patient i is assigned to pattern p, otherwise 0. The first step is to put an upper bound M on the number of surgeries assigned to each pattern p,

$$\sum_{i \in L_o} z_{i,p} \leq M, \quad \forall\, p \in P, \quad o \in O \tag{1}$$

where L_o is the set of patients for operator o. Second, we force each pattern to maintain a balance h_o between in- and out-patients for each operator o, determined by their waiting list,

$$|\sum_{i \in L_o} g_i z_{i,p} - \lceil h_o \sum_{i \in L_o} z_{i,p} \rceil| \leq 1, \quad \forall o \in O, \quad \forall p \in P \tag{2}$$

The first part of the constraint counts the number of in-patients assigned to the pattern p, where $g_i \in \{0, 1\}$ indicates when patient i is an in-patient. The second part calculates the desired number of in-patients rounded up to the nearest integer for operator o. The absolute difference between those two parts should preferably be zero. However, we allow the flexibility of one patient. A similar strategy is found in [4, 6] where scheduling balance is forced for different surgery groups. In the last step, we employ a Monte-Carlo simulation to verify each pattern concerning overtime. That is, they will not result in overtime with more than δ probability for a given capacity Cap

$$Pr[f(z_p) \geq Cap] \leq \delta \tag{3}$$

where $f(z_p)$ denotes the distribution function for the stochastic sum of all surgical procedures in pattern p, including the times between the surgeries.

This method is comparable to the work by [9] where column generation is used to create feasible operating room days. However, when generating the sub-problem, the stochastic constraint (3) is approximated using a planned slack, which is not desirable for a practical application. In practice, the number of feasible patterns is limited. Thus, generating all feasible patterns is an attractive exact approach and tractable up to some limited patient list length.

4 Pattern Scheduling with Probabilistic Ward Restrictions

Now all feasible patterns have been generated that satisfy stochastic constraint (3). The decision is now reduced to determine which pattern p should be assigned to day d and room r. Let us introduce the binary variable $x_{d,p,r}$ taking the value 1 if pattern p is assigned on day d and room r. The objective is to maximize the throughput of patients

$$\max_{\mathbf{x}} \sum_{d \in D, p \in P, r \in R} C_p x_{d,p,r} \tag{4}$$

where C_p is the number of patients assigned to pattern p. As each pattern represent a whole session, only one pattern can be in any room at any given day,

$$\sum_{p \in P} x_{d,p,r} \leq 1, \quad \forall r \in R, \quad d \in D \tag{5}$$

and any given surgery can only be performed once

$$\sum_{p\in P_s, d\in D, r\in R} x_{d,p,r} \leq 1, \quad \forall s \in S \tag{6}$$

where $P_s \subseteq P$ are patterns containing surgery s. Each operator is only working according to one pattern per day,

$$\sum_{p\in P_o, r\in R} x_{d,p,r} \leq 1, \quad \forall d \in D, \quad o \in O \tag{7}$$

where $P_o \subseteq P$ are the patterns of operator o. Similarly, we force an assignment of each operator to at least one day for a given week v with working days D^v,

$$\sum_{p\in P_o, d\in D^v, r\in R} x_{d,p,r} \geq 1, \quad \forall o \in O \tag{8}$$

Now we turn our attention to the ward assignments. By using historical data, it is possible to estimate the expected number of patients on each day for any given pattern. However, this will not give the probability of exceeding the ward capacity. Thus, we approximate the problem as follows. We assume that there are only three possible scenarios for a patient, either the patient will be in the ward on the day j with 100% certainty, will have a 50% chance of being discharged or will not be in the ward on the day j. Thus, we define two parameters $w_{j,p}^{50}$ and $w_{j,p}^{100}$ denoting the number of patients with 50% and 100% likelihood respectively of being in the ward on day j for pattern p. At any given day d the number of patients in the ward known with certainty is

$$\bar{w}_d^{100} = \sum_{r\in R, p\in P, j\in\{0,\ldots,m\}: d-j\in D} w_{j,p}^{100} x_{d-j,p,r}, \quad \forall d \in D \tag{9}$$

over a m day period. Similarly, $\bar{w}_d^{50}$ can be formulated as shown in equation (9). Now, we know how many beds are occupied with certainty. Thus we should not exceed the given limit,

$$\bar{w}_d^{100} \leq w^T, \quad \forall d \in D \tag{10}$$

where w^T is the absolute upper limit on the number of beds in the ward. Now we are left with the decision to determine how many patients with 50% chance of being discharged to have each day. First, let us determine how many beds are still available by $w_d^a = w^T - \bar{w}_d^{100}$ for each day d. Now let us introduce an auxiliary binary indicator $y_{d,i} \in \{0, 1\}$ for day d and $i \in \{0, \ldots, w^T\}$,

$$w_d^a = \sum_{i\in\{0,\ldots,w^T\}} i y_{d,i}, \quad \forall\, d \in D \tag{11}$$

In this context, the variable $y_{d,i}$ is used as a look-up for the allowed number of patients on the day d, which will be given by look-up index i. Note that i takes the same value as the remaining beds for each day. Clearly, $y_{d,i}$ can only take one value as only one possibility is available for the day,

$$\sum_{i \in \{0,\ldots,w^T\}} y_{d,i} = 1, \quad \forall d \in D \tag{12}$$

Now we can find an upper bound on the number of patients with 50% chance of being discharged in the ward,

$$\bar{w}_d^{50} = \sum_{\substack{r \in R, p \in P, j \in \{0,\ldots,m\}: \\ d-j \in D}} w_{j,p}^{50} x_{d-j,p,r} \leq \sum_{i \in \{0,\ldots,w^T\}} F_i^{50} y_{d,i}, \quad \forall d \in D \tag{13}$$

where F_i^{50} denotes the maximum number of patients with a 50% chance of being discharged when there are i beds available. Let us assume there exist a patient $\bar{w}_d^{50}$ with 0.5 chance of being on the day d in the ward or not. Then the number of such patients the ward can hold without exceeding its limits w_d^a is computed by the quantile function for the Binomial distribution. The values are for example (i, F_i^{50}) : (0, 0), (1, 1), (2, 2), (3, 3), (4, 4), (5, 4), (6, 5), when using the confidence 0.05.

Although the patterns maintain a balanced ratio between in- and out-patients for any given operator, this value has been approximated to the nearest integer. Thus, the overall balance over the whole set of operators may still be biased towards selecting more out-patients since they usually require less surgery time. To combat this, a global in- out-patient constraint is forced for all operators. The global ratio h_G of in- and out-patients each week can be forced using the following two constraints,

$$h_G \sum_{(d,p,r) \in DPR} C_p x_{d,p,r} - 1 \leq P^{in} \tag{14}$$

$$P^{in} \leq h_G \sum_{(d,p,r) \in DPR} C_p x_{d,p,r} + 1 \tag{15}$$

where P^{in} is a continuous variable denoting the total number of in-patients and is defined by

$$P^{in} = \sum_{p \in P, d \in D, r \in R} G_p x_{d,p,r} \tag{16}$$

where G_p denotes the number of in-patients for pattern p.

5 Experimental Study

For the computational experiments we will create several different MSS using our approach. For the sake of simplicity, we focus on one surgical specialty, General Surgery. The experiments are performed on a 32GB memory Intel Core i7-7700 3.60 GHz with 4 cores. The model is programmed in Python 3.6 using Gurobi version 8.1.0. The time for each experiment is limited to 6 hours.

5.1 Instance Generation

We have obtained an extensive data set from Landspitali from the last ten years. This data set is used to create several different instances. For each instance, we sample surgeries that are most likely to occur for each operator using different sizes of waiting lists $L_o \in \{10, 20, 30\}$. In Table 1, we provide a summary of the main characteristics of the most frequently performed surgeries by each operator.

Based on the current MSS, there are two ORs (r_1 and r_2) available each day both with capacity of $Cap = 450$ minutes. As previously discussed, running into overtime is not undesirable. Thus we assume that $\delta \approx 0.3$. From historical data, we select an upper bound of $M = 6$ on the number of patients assigned to each pattern and that global ratio between in- and out-patients scheduled is $h_G = 0.38$. The total number of beds available in the ward is $w^T = 6$. Different MSS will be found under these parameter settings by varying the cycle length in days $T \in \{7, 14, 28\}$.

Table 1 A summary of the main characteristics of the operators and their surgeries

Operator	Number of surgery types	Mean surgery time	Mean ward probability	Mean ward length of stay
A	9	184 (min)	0.40	2.4 (days)
B	6	137	0.50	2.9
C	9	188	0.50	2.6
D	10	167	0.70	3.2
E	11	162	0.60	3.4
F	5	214	0.40	3.7
G	7	90	0.10	2.0
H	12	143	0.30	2.5
I	13	84	0.10	2.3

5.2 Results

Table 2 provides a summary for the computational results under different parameter settings. In the table, one can find the proportion of scheduled in- and out-patients. By increasing the values of T and L_o substantially makes the problem more difficult to solve. As a result, a sub-optimal solution is provided for the parameter settings of $T = 28$ and $L_o = 30$ (with a small gap of 0.73%). Comparing the throughput under each setting, including more patients on the waiting list leads to more throughput. This is expected since more options are available and so too the surgeries with short surgery time and length of stay. One might have expected to schedule four times what was optimally planned for one week. However, only 136 surgeries are scheduled. It indicates that the optimization has selected operations of a specific type (with short times and length of stay). Hence, more complicated surgeries may be deferred. Comparing the results to the actual MSS (fixed roster days) for $T = 7$, one can see that the same throughput is achieved, but fewer in-patients are scheduled.

Table 3 illustrates the historical ratio h_o and the scheduled ratio for each surgical operator along with the number of planned surgeries in the parenthesis behind. Similarly, the global ratio h_G is provided. It is apparent from the table that each operator's ratio is often achieved. For short planning horizons, it might, however, get challenging to satisfy this ratio completely. Turning now to the global ratio, one can see, that it is satisfied for all settings (with a small deviation). However, the results suggest that planning beyond one week is required to fulfill the ratio of each operator when maximizing the throughput. One could add more restrictions to meet each operator ratio even further, i.e., by taking into account the types of surgeries similar to [4, 6]. However, it might come at the cost of the flexibility of the provided schedules and so the throughput.

Different MSS are provided in Tables 4, 5 and 6 for $L_o = 30$ and $T = \{7, 14, 28\}$ along with the actual MSS. Each table shows the assignments of the operators to days and rooms along with their daily proportion of in-patients from the throughput.

Table 2 Comparison of the throughput under different parameter settings

T	L_o	Number of surgeries	Number of in-patients	Number of out-patients	Duality gap (%)	CPU MIP (sec)	
7	10	35/90	13/27	22/63	0.00	1	
	20	39/180	14/67	25/113	0.00	28	
	30	41/270	16/104	25/166	0.00	202	
7	30	41/270	15/104	26/166	0.00	123	**Actual MSS**
14	10	60/90	22/27	38/63	0.00	4	
	20	72/180	27/67	45/113	0.00	320	
	30	78/270	29/104	49/166	0.00	986	
28	20	124/180	47/67	77/113	0.00	948	
	30	136/270	51/104	85/166	0.73	21600	

Table 3 Comparison of the historical ratio h_o and the scheduled ratio under different planning horizon T and number of patients L_o

Operator	h_o		$T = 7$			Actual ↓	$T = 14$			$T = 28$	
		$L_o \rightarrow$	10	20	30	30	10	20	30	20	30
A	0.40		0.75 (4)	0.25 (4)	0.75 (4)	0.50 (4)	0.57 (7)	0.86 (7)	0.75 (8)	0.58 (12)	0.50 (16)
B	0.50		0.67 (3)	0.25 (4)	0.25 (4)	0.25 (4)	0.50 (8)	0.33 (6)	0.25 (8)	0.46 (13)	0.64 (11)
C	0.50		0.50 (2)	0.33 (3)	0.25 (4)	0.25 (4)	0.50 (4)	0.33 (6)	0.29 (7)	0.60 (10)	0.55 (11)
D	0.70		0.50 (2)	1.00 (3)	1.00 (3)	0.67 (3)	0.50 (4)	0.67 (6)	0.67 (6)	0.70 (10)	0.73 (11)
E	0.60		1.00 (2)	0.33 (3)	1.00 (3)	0.33 (3)	0.57 (7)	0.33 (6)	0.67 (6)	0.54 (13)	0.58 (12)
F	0.40		0.33 (3)	0.67 (3)	1.00 (3)	1.00 (3)	0.20 (5)	0.33 (6)	0.67 (6)	0.20 (10)	0.33 (12)
G	0.10		0.00 (5)	0.50 (4)	0.00 (5)	0.20 (5)	0.00 (9)	0.25 (12)	0.22 (9)	0.20 (20)	0.16 (25)
H	0.30		0.50 (4)	0.25 (4)	0.00 (3)	0.67 (3)	0.67 (6)	0.43 (7)	0.33 (6)	0.31 (16)	0.33 (12)
I	0.10		0.10 (10)	0.18 (11)	0.17 (12)	0.17 (12)	0.10 (10)	0.19 (16)	0.14 (22)	0.15 (20)	0.12 (26)
$h_G \rightarrow$	0.38		0.37 (35)	0.36 (39)	0.39 (41)	0.37 (41)	0.37 (60)	0.38 (72)	0.37 (78)	0.38 (124)	0.38 (136)

Table 4 Optimal MSS for $T = 7$ and number of patients $L_o = 30$ along with the actual MSS

	Actual MSS							Optimal MSS						
d	1	2	3	4	5	6	7	1	2	3	4	5	6	7
r_1	H (2/3)	F (3/3)	E (1/3)	D (2/3)	G (1/5)	–	–	A (3/4)	H (0/3)	B (1/4)	I (1/6)	G (0/5)	–	–
r_2	I (1/6)	B (1/4)	C (1/4)	A (2/4)	I (1/6)	–	–	F (3/3)	E (3/3)	I (1/6)	C (1/4)	D (3/3)	–	–
$\Pr[w_d \geq w^T]$	0.00	0.00	0.03	0.00	0.04	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00

Table 5 Optimal MSS for $T = 14$ and number of patients $L_o = 30$

	Week I							Week II						
d	1	2	3	4	5	6	7	8	9	10	11	12	13	14
r_1	E (2/3)	A (3/4)	F (2/3)	I (1/6)	I (1/6)	–	–	A (3/4)	H (1/3)	B (1/4)	C (1/3)	I (1/5)	–	–
r_2	H (1/3)	C (1/4)	D (2/3)	G (0/5)	B (1/4)	–	–	D (2/3)	G (2/4)	F (2/3)	I (0/5)	E (2/3)	–	–
$\Pr[w_d \geq w^T]$	0.00	0.01	0.10	0.00	0.00	0.00	0.00	0.10	0.06	0.08	0.01	0.09	0.00	0.00

Table 6 Sub-optimal MSS for $T = 28$ and number of patients $L_o = 30$

	Week I							Week II						
d	1	2	3	4	5	6	7	8	9	10	11	12	13	14
r_1	E (3/3)	A (1/3)	F (1/3)	I (1/5)	H (1/3)	–	–	E (2/3)	B (1/4)	A (2/4)	C (2/2)	I (0/6)	–	–
r_2	G (0/4)	D (2/3)	B (2/3)	G (0/4)	C (1/3)	–	–	G (2/4)	I (1/5)	F (1/3)	H (0/3)	D (2/3)	–	–
$\Pr[w_d \geq w^T]$	0.00	0.00	0.02	0.00	0.03	0.00	0.00	0.06	0.01	0.09	0.02	0.02	0.00	0.00
	Week III							Week IV						
d	15	16	17	18	19	20	21	22	23	24	25	26	27	28
r_1	H (2/3)	A (1/3)	E (1/3)	A (2/3)	D (2/2)	–	–	H (1/3)	F (1/3)	D (2/3)	I (1/5)	E (1/3)	–	–
r_2	C (2/3)	F (1/3)	B (2/2)	I (0/5)	G (1/4)	–	–	A (2/3)	C (1/3)	G (0/5)	G (1/4)	B (2/2)	–	–
$\Pr[w_d \geq w^T]$	0.02	0.01	0.06	0.07	0.23	0.00	0.00	0.01	0.00	0.00	0.00	0.05	0.01	0.00

In our MIP model, we have approximated the probability of being in the ward to three rounded probabilities 0, 0.5 and 1. The true probability of going over the set ward limit can, however, be found off-line with Monte Carlo simulation to estimate the true probability of going over the absolute ward limit w^T. This is denoted by $\Pr[w_d \geq w^T]$ in the tables below. In this case, we observed that this likelihood is less than or equal to 0.10 in all cases but one, where on the day 19 in Table 6, it is 0.23. In general, the results are nevertheless promising. Comparing the results to the actual MSS, one can see that the assignments of the operators is different than in reality and that no clear patterns are found. However, some parts remain similar. For $T = \{7, 14\}$ it is suggested that operator I is assigned to two days each week. That is in accordance with the actual MSS however, his days are different except for the Fridays. When $T = 28$ one can see that operator I is no longer assigned to two patterns each week anymore. Instead, operator G is assigned to two patterns in weeks I and IV. Furthermore, operators I and A are assigned to two patterns in week II and III, respectively. As predicted these operators, who have the shortest operating time (see Table 1), complete almost all of their available patients within the planning horizon of $T = 28$ when the throughput is maximized.

6 Discussion and Conclusions

In this paper, we have provided an effective way to account for risk associated with last-minute cancellations for the surgery scheduling problem, namely going into overtime and exceeding the limit of the ward beds. To tackle the problem, we used a two-phase approach. In the first phase, we created feasible one-day sessions (patterns) using a three-step procedure. It allowed us to tackle the uncertainty in surgery times in a practical and precise way by generating feasible patterns. In the second step, the patterns are scheduled using a MIP model that addresses the ward restrictions in a probabilistic manner. To illustrate this, we built several MSS for various planning horizons and waiting lists length, where the goal was to find the optimal assignments of the operators. The results suggest that by using the proposed approach, one can bound the risk of exceeding wards effectively and tackle the uncertainty in surgery times. Thus, we avoid the risk of cancellation. Further results indicate that flexibility in the rosters of the operators and planning beyond one week is required to achieve the best possible results in terms of balanced flow of in- and out-patients.

In this study, we have focused on maximizing the throughput of patients. By doing so, the optimization will favor patients with shorter surgery times and ward days. It unfortunately, will leave the more difficult patients for a later date, especially for the shorter planning periods, which is not a desired result of the optimization. Additional criteria will be needed to combat this effect, for example, the patient priority. However, finding the right balance between patient priority and throughput poses new challenges. These issues are the focus of our current work.

Acknowledgements The authors would like to acknowledge the staff and the managers at Landspitali for giving insights and support to this project.

References

1. Cardoen, B., Demeulemeester, E., Beliën, J.: Operating room planning and scheduling: a literature review. Eur. J. Oper. Res. **201**(3), 921–932 (2010)
2. Samudra, M., Van Riet, C., Demeulemeester, E., Cardoen, B., Vansteenkiste, N., Rademakers, F.E.: Scheduling operating rooms: achievements, challenges and pitfalls. J. Sched. **19**(5), 493–525 (2016)
3. Beliën, J., Demeulemeester, E.: Building cyclic master surgery schedules with leveled resulting bed occupancy. Eur. J. Oper. Res. **176**(2), 1185–1204 (2007)
4. M'Hallah, R., Visintin, F.: A stochastic model for scheduling elective surgeries in a cyclic master surgical schedule. Comput. Indus. Eng. **129**, 156–168 (2019)
5. Min, D., Yih, Y.: Scheduling elective surgery under uncertainty and downstream capacity constraints. Eur. J. Oper. Res. **206**(3), 642–652 (2010)
6. Banditori, C., Cappanera, P., Visintin, F.: A combined optimization-simulation approach to the master surgical scheduling problem. IMA J. Manag. Math. **24**, 155–187 (2013)
7. Marques, I., Captivo, M.E.: Different stakeholders' perspectives for a surgical case assignment problem: deterministic and robust approaches. Eur. J. Oper. Res. **261**(1), 260–278 (2017)
8. Wachtel, R.E., Dexter, F.: Tactical increases in operating room block time for capacity planning should not be based on utilization. Anesth. Analg. **106**(1), 215–26 (2008)
9. van Oostrum, J.M., Van Houdenhoven, M., Hurink, J.L., Hans, E.W., Wullink, G., Kazemier, G.: A master surgical scheduling approach for cyclic scheduling in operating room departments. OR Spectr. **30**(2), 355–374 (2008)
10. Adan, I., Bekkers, J., Dellaert, N., Vissers, J., Yu, X.: Patient mix optimisation and stochastic resource requirements: a case study in cardiothoracic surgery planning. Health Care Manag. Sci. **12**(2), 129 (2008)
11. Blake, J.T., Donald, J.: Mount sinai hospital uses integer programming to allocate operating room time. Interfaces **32**(2), 63–73 (2002)
12. Adan, I., Vissers, J.: Patient mix optimisation in hospital admission planning: a case study. Int. J. Oper. Prod. Manag. **22**(4), 445–461 (2002)
13. Runarsson, T.P., Sigurpalsson, A.O.: Towards an evolutionary guided exact solution to elective surgery scheduling under uncertainty and ward restrictions. In: 2019 IEEE Congress on Evolutionary Computation (CEC), pp. 419–425. IEEE (2019)

Home Health Care

Integration of User's Preferences into the Home Healthcare Routing and Scheduling Multi-objective Problem: A Hierarchical Approach with Pareto-Optimal Alternative Solutions

Laura Musaraganyi, Simon Germain, Nadia Lahrichi and Louis-Martin Rousseau

Abstract Home health care structures provide medical and paramedical services in patient homes rather than in facilities, such as hospitals. From these activities emerge various operational problems including the Home Health Care Routing and Scheduling Problem (HHCRSP). In this paper we are especially interested in the multi-objective aspect of the HHCRSP. In fact, while assigning patient visits to caregivers, multiple conflicting criteria must be simultaneously considered in order to find the best trade-off. However the evaluation of these trade-offs depends greatly on the decision maker's judgment, which includes a diverse range of HHCS provide home health care services. This implies that in order for a decision tool to be adopted by the largest number it must be able to adapt to its user's preferences. In this paper, we present a method that takes into account the decision maker's priorities, within the context of an automatic scheduling assistant. The algorithm, based on a heuristic, uses a hierarchical approach to find the best multi-objective solution according to strict priorities. It also suggests near-equivalent Pareto-optimal solutions, selected by means of tolerance parameters, to provide the user with relevant alternative choices.

Keywords Home health care · Multi-objective optimization · Scheduling optimization · Routing problem · Heuristic · User's preferences

L. Musaraganyi (✉) · N. Lahrichi · L.-M. Rousseau
Mathematical and Industrial Engineering, Polytechnique Montreal, Montreal, Canada

S. Germain
AlayaCare Labs, Montreal, Canada

N. Lahrichi · L.-M. Rousseau
HANALOG, Canada Research Chair in Healthcare Analytics and Logistics, Montréal, Canada

V. Bélanger et al. (eds.), *Health Care Systems Engineering*,
Springer Proceedings in Mathematics & Statistics 316,
https://doi.org/10.1007/978-3-030-39694-7_14

1 Introduction

Home health care services offer a wide range of services such as helping individuals cope with long-term medical conditions, illness or injury. While the patient remains in the comfort of their own home, the administrators see a decrease in hospital congestion, which also reduces costs [1]. For these reasons, and due to multiple factors (e.g. aging population, chronic diseases,…), this type of service quickly grew in popularity. Most of home health care providers operate a fleet of mobile care workers that travel from place to place providing care to patients under the company supervision. In this situation, at an operational level, arises the Home Health Care Routing and Scheduling Problem (HHCRSP). The HHCRSP is interested in determining the assignment of home visits to a set of caregivers over the course of a planning horizon, and the routing of these caregiver workdays. In our application, we consider time windows and time-dependent travel issues, as well as constraints on caregivers' skills, patient requirements, and specific caregiver contracts/union rules e.g. work time limits. The home health care context also comes with an important concern, which is the continuity of care. In fact, once a patient-caregiver relationship is established, there are strong benefits to maintain that match.

In the process of assigning routes to home health care workers, schedulers consider simultaneously various conflicting criteria (worker skill-set, availability / time of day, distance traveled, patient-worker relations,…) and try to make the best schedule possible, minimizing time spent in the car and maximizing time with patients, while also limiting costs for the home health care service provider. In this paper, we present a method tackling the challenging task of finding the right balance between these contradictory criteria. In particular, our work is centered around the users' perspective as, for a method to be beneficial to a large number of different home health care structures, it has to be easily adaptable to their preferences. To do so, we combine a hierarchical optimization technique with a Pareto based approach. Tolerance parameters are introduced in order to display only relevant Pareto-optimal choices to the decision maker. For this project, we collaborated with the company *AlayaCare*. It offers a cloud-based platform (SaaS) for home health care service providers to improve their efficiency in their different tasks. We are especially interested in *AlayaCare*'s Schedule Optimizer. It is an optimization tool built to provide full daily schedules for every care worker over a given period. We will focus on a significant overhaul of the Schedule Optimizer that is the integration of users' preferences. In fact, our industrial partner has to deal with a considerable number of home health care structures with their own characteristics and their own policies. To address this issue, the proposed method contributes to a much better decision making experience for the end-users as it offers priority-oriented solutions and a varied choice of alternative solutions.

A growing interest in home health care over the last few decades has given rise to a number of publications dealing the HHCRSP. Two complete reviews on the HHCRSP are available [2, 3]. As mentioned in Cissé et al.[3], although the HHCRSP has gained popularity in recent years, few research has focused on tackling its multi-

objective aspect. However, multi-objective resolution methods can help home health care schedulers find the best trade-off according to their experience and knowledge of home health care environments. In addition, an optimization method able to adapt to the user and provide good decision support have a higher probability to be used effectively. Nevertheless, solving the HHCRSP is usually done by using a multi-objective weighted sum model [3–5]. While it does definitely work, we found a significant problem when the algorithm was faced with real users (the schedulers in the HHCS). Translating the intuitions of human beings into weights to allocate the importance in the objective function ended up to be a challenging task. Every scheduler wanted to have slightly different weights, and even with full control of them, they were often unsatisfied with the results. Regarding the estimation of the weights for each objective, the analytic hierarchy process (AHP) method, proposed by Saaty [6] has been considered. It was used, for example, in Jafari et al.[7] in order to maximize nurses' preferences. However, in this application the process would need to be repeated several times and this would imply an unnecessary burden for the end-user.

More generally, many models solving nurse scheduling problems use a hierarchical or goal programming approach [8–11] whereas the Pareto-based approach is often used in vehicle routing problems [12–15] and also in some scheduling problems [16, 17]. Therefore it seems natural to combine both approaches to solve the HHCRSP. Furthermore, Drechsler et al.[18] introduced an idea of ϵ-limit to avoid considering solutions with excessive values for less preferred objectives.

This paper is outlined as follows: in Sect. 2, we describe our formulation of the HHCRSP, in particular the constraints and the objectives that are considered as well as the resolution approach. Then Sect. 3 focuses on how different multi-objective techniques have been combined to better adapt to the user preferences. Especially, 3.1 presents how multi-objective aspects were handled during the generation of solutions and 3.2 tackles the issue of choosing the most suitable solution regarding user priorities and selecting alternative solutions. Results are discussed in Sect. 4 and finally, we offer our conclusion.

2 Problem Statement and Resolution Approach

The HHCRSP can be modeled in a few different ways [3]. We chose the approach of the multiple depot traveling salesman problem with time windows (MDTSPTW). In addition to the classical routing and assignment constraints, constraints specific to the home health care context are taken into account:

- *Planning horizon*: Period over which the routing and scheduling decisions are made. The planning horizon considered in this application is one or two weeks.
- *Continuity of care (patient–nurse loyalty)*: For the patient, it involves consistency and trust in the experience of care. For caregivers, it is related to collecting sufficient information and knowledge about the patient in order to give the best possible care.

- *Time windows*: Time interval in which the patient can receive care. It is linked to patient availability and to the type of care that needs to be provided (e.g medication intake). Only hard time windows were considered.
- *Preferences*: They are divided into hard constraints (if a patient rejects a particular caregiver) and soft constraints (preferences related to the caregiver's gender or language skills). The optional preferences were not considered in this application, although the algorithm is able to support these constraints.
- *Time-dependent travel times*: For real-world applications such as this one, it is essential to include the time-dependent aspects in travel times as they change considerably over the course of the day.
- *Work time*: Can be a soft or a hard constraint. When it is set as a soft constraint, overtime work refers to working hours exceeding those specified in union rules or work contracts; and under-time work refers to working hours missing to meet those specified in union rules or work contracts.
- *Qualifications/skills*: To satisfy a patient's need, the caregiver's qualifications must match the patient's care requirements.

More details about these constraints are given in Table 1. To facilitate comparisons with other HHC models we used terms from Cissé et al. [3].

Within these constraints, the goal is to minimize the following objectives :

- *Unscheduled visits*: A service that corresponds to a specific care type that needs to be provided to a patient. It is composed of one or several visits that have to be carried out. In the context of this application, all the visits cannot always be scheduled, therefore the constraint related to scheduling visits has been relaxed.
- *Travel Time*: Total travel time over all caregivers and over the planning horizon. It is expressed in minutes.
- *Wait Time*: Sometimes a caregiver has to wait for a patient to be available. This objective corresponds to the total waiting time over all caregivers and over the planning horizon. It is expressed in minutes.
- *Assigned Nurses*: Continuity of care measured within a single care type.
- *Loyalty*: Continuity of care measured across care types.

Table 1 Classification of the specific constraints

Constraints	Type of constraint	Related to
Planning horizon	Temporal	HHC structure
Continuity of care	Assignment	
Time windows	Temporal	Patients
Preferences	Assignment	
Time-dependent travel times	Geographic	
Worktime	Temporal	Caregivers
Qualifications/skills	Assignment	

- *Daily Overtime*: Total working hours exceeding a daily limit. It is expressed in minutes.
- *Daily Under-Time*: Total working hours missing to meet a daily limit. It is expressed in minutes.
- *Weekly Overtime*: Total working hours exceeding a weekly limit. It is expressed in minutes.
- *Weekly Under-Time*: Total working hours missing to meet a weekly limit. It is expressed in minutes.
- *Used Nurses*: Number of caregivers serving all patients.

As mentioned previously, these objectives are conflicting. More specifically, scheduling a visit often increases all the other objectives. Trying to minimize the travel time can deteriorate the continuity of care. Reducing the number of caregivers serving all patients potentially increases the overtime.

The HHCRSP, is solvable to optimality [19]. However, the complexity of the problem leads to scalability issues [20, 21]. To bypass this obstacle, methods based on heuristics or meta-heuristics have been developed [2]. Some use a mix of heuristic and Mixed Integer Linear Programming [4, 5], while others rely only on heuristic [22]. In this work, the solution developed is based around an Adaptive Large Neighbourhood Search (ALNS). Extension of the Large Neighborhood Search, this method destroys parts of the current solution through *destroy operators* then recreates a new solution with *repair operators* as shown in procedure 1. It allows for the exploration of promising areas more easily than with the use of local search heuristics.

3 Introducing a Hierarchical Approach

3.1 Generating Solutions

During the problem resolution, the method used to generate the solutions can influence their quality, especially with regards to how they conform to a certain order of priority. This is particularly true for the ALNS method. In fact, some ALNS operators rely on a cost function (e.g. a weighted sum) to remove or insert visits; meaning that in our case, the definition of the cost function should be easily adaptable to different preference orders. Most of the operators used in our application are the same as those described in Grenouilleau et al. [4], namely, *WorstRemoval*, *RandomRemoval*, *ServiceRemoval* and *FlexibleAvailRemoval* for the destroy procedure and *Greedy Heuristic*, *Regret - 2* and *Regret - 3* to repair the solution. Two other repair operators are applied:

- *TightVisitInsertion*: The visits are ordered in increasing order of tightness i.e.

$$\frac{NumberOfAvailableDays}{NumberOfVisitsToSchedule} \times \frac{TimeWindowLength}{VisitDuration}$$

Then, a greedy allocation method is called to insert the visits at the lowest-cost position.

- *WorkTimeRoutesInsertion*: The unscheduled visits are ordered in decreasing order of their average feasible routes' congestion. Then, a greedy allocation method is called to insert the visits at the lowest-cost position.

Some of these operators use a cost function to destroy or repair solutions (e.g when comparing the potential insertions of a visit). The creation of the initial solution also requires a cost function as it is the same procedure as in Grenouilleau et al. [4], following a lowest-cost insertion logic. The idea of a weighted sum was kept to compute the costs, but the objectives were transformed with this upper-bound approach, as follows:

$$\forall\, i,\ \tilde{F}_i = \begin{cases} \frac{F_i}{F_i^{max}} & \text{if } F_i^{max} \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

Where $\tilde{F}_i$ is non-dimensional and ≤ 1. Next, the weights are set in such a way that, if the objectives 1,…, n are ranked in order of priority, 1 being the most preferred, then $w_i = 5 \times w_{i+1}$.

Algorithm 1: Generate ALNS Solution

Input: Initial solution S_i
Destroy part of the solution S_i ;
Repair & create a new solution S_{i+1} ;
Output: Next solution S_{i+1}

We will not go into more detail about the ALNS, more information can be found in [22–24]. Any suitable heuristic that generates valid solutions could be used.

3.2 Guiding the Optimization Process

In order to guide the optimization process and handle the multiple objectives, a hierarchical approach is proposed. This can be achieved through defining the weights so that they do not interfere one with the other or through using a lexicographic comparison. In this application, a lexicographic comparison is used to compare solutions according to a hierarchical order, set by the user, as shown in procedure 2. This approach has the advantage of being easily understandable and intuitive for the end-user. It is simpler to sort one's priorities rather than weighting them one against the other.

Algorithm 2: Update Best Solution

Input: Solution S, current best solution S^*
if $S \leq S^*$ *(as per hierarchical order)* **then**
  Set $S^* = S$
end
Output: $\emptyset$

In traditional scalarized multi-objective methods, the model will select one solution and others will be discarded. However, some of the rejected solutions may still be interesting. The existing algorithm was already offering alternative choices that were the previous best weighted-sum solutions found during the search process. Figure 1 shows an example of choices suggested to the user when a new service needs to be scheduled. We can see that option 6 offers the least travel time but comes with a substantial number of conflicts. Option 7 requires an extra 14 min of travel time and an extra caregiver compared to option 6, but with much less conflict. Although option 8 brings an extra 11 min of travel time to the user and has the same number of conflicts as option 7, it assigns only one caregiver to the service, leading to better continuity of care.

Besides the fact that some potentially interesting solutions were not considered, there was no guarantee for the suggested solutions not to be dominated by others. To address this issue, the concept of Pareto front is introduced. The idea is to record, in a list, all the near Pareto-optimal solutions generated during the search process, even if they are not the best according to the hierarchical order. The list is dynamically updated as shown in procedure 3.

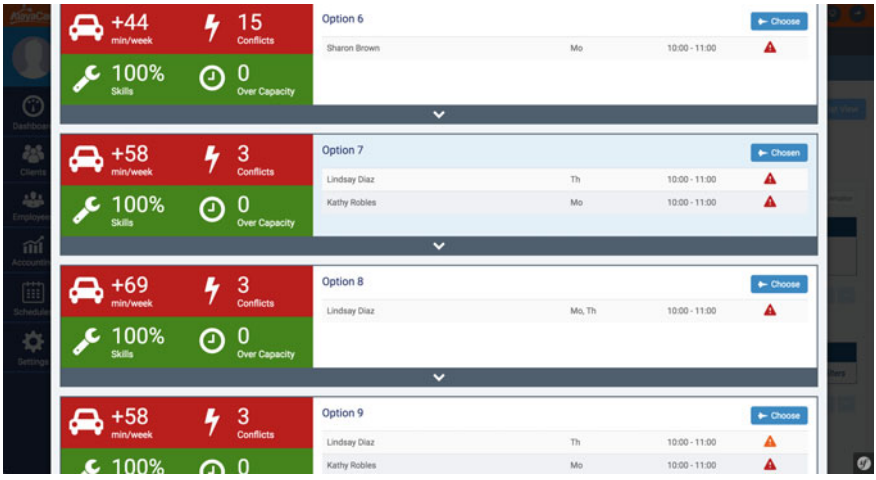


Fig. 1 Example of alternative solutions from *AlayaCare* software

Algorithm 3: Update Pareto List

Input: Solution S, current Pareto list $\mathbb{P}$
if *S non-dominated in* $\mathbb{P}$ **then**
  for S_i *in* $\mathbb{P}$ **do**
    if *S dominates* S_i *in* $\mathbb{P}$ **then**
      Remove S_i from $\mathbb{P}$;
    end
  end
  Add S to $\mathbb{P}$;
end
Output: $\emptyset$

Nevertheless, due to the number of objectives, the size of the approximate Pareto front could potentially be very large. Thus, as we only want to provide the user with relevant solutions, a procedure is applied after the search process to prune the Pareto list. To do so, the user is asked to set some tolerance parameters so that only the interesting solutions, near-equivalent to the best hierarchical solution, are displayed (see procedure 4). In fact, an experienced user could look at the set of near-equivalent solutions and decide that the trade-off between the different criteria is better on some solutions than others. More formally:

Let $F_1, F_2, \ldots, F_n$ be the objective functions and $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ their associated tolerance. Let $\mathbb{P}$ be the set of Pareto-optimal solutions and S^* be the best hierarchical solution. The solution $S_i \in \mathbb{P}$ will be considered as near-equivalent to S^* if and only if :

$$\forall\, k \in \{1, \ldots, n\},\ F_k(S_i) \leq \begin{cases} F_k(S^*) + \epsilon_i, & \text{if } \epsilon_i \text{ expressed in absolute value} \\ F_k(S^*) \times (1 + \epsilon_i), & \text{if } \epsilon_i \text{ expressed in percentage} \end{cases}$$

Algorithm 4: Define ϵ-Neighbourhood

Input: Solution S, current Pareto list $\mathbb{P}$
for S_i *in* $\mathbb{P}$ **do**
  if S_i *ϵ-nondominated by S* **then**
    Add S_i to $\mathbb{P}^*$;
  end
end
Output: $\mathbb{P}^*$

The global process is summarized in Algorithm 5.

Algorithm 5: Main procedure

Generate Solution S_0 ;
while *Stopping criteria not reached (e.g. time)* **do**
  Generate ALNS Solution ;
  Update Best Solution ;
  Update Pareto List ;
end
Define ϵ-Neighbourhood ;
Output: Best Solution S^* and its ϵ-Neighbourhood $\mathbb{P}^*$

3.3 Scenarios

In order to evaluate our method, we considered only the first 6 objectives in Sect. 2 because, due to the young age of its optimisation tool, our industrial partner was not able to provide us with more complex and realistic instances. Then, we looked at 4 different priority rankings believed to be the most probable scheduler's choices. The scenarios are displayed in table 2. The tolerance parameters are the same for all the tests and are shown in table 3.

Table 2 Scenarios for the set of instances

Criteria	Scenario 1		Scenario 2		Scenario 3		Scenario 4	
	Order	Weights	Order	Weights	Order	Weights	Order	Weights
Unscheduled visits	1	10,000	1	10,000	1	10,000	1	1,000,000
Travel time	2	15	3	1	3	1	3	10
Wait time	5	1	5	1	5	1	5	1
Assigned nurses	3	10	2	200	4	5	4	10
Loyalty	4	5	4	5	2	200	–	–
Daily overtime	–	–	–	–	–	–	2	1000

Table 3 Tolerance parameters

Criteria	ϵ Tolerance
Unscheduled visits	2
Travel time	max(200, 15% of $F_{TravelTime}$)
Wait time	max(200, 15% of $F_{WaitTime}$)
Assigned nurses	max(10, 10% of $F_{AssignedNurses}$)
Loyalty	max(30, 10% of $F_{Loyalty}$)
Daily overtime	max(50, 5% of $F_{DailyOvertime}$)

4 Results

We ran our algorithm against the current weighted sum used in *AlayaCare* software. The tests were performed using instances provided by our industrial partner. We had a set of 8 instances from the same home health care structure, divided equally into instances with tight time-window and instances with large time-window. An additional instance was generated to show examples of alternative solutions and how changing criteria order can affect the solutions proposed by the method.

All the instances of the set had between 311 and 340 visits to schedule and between 11 and 16 caregivers available. The planning horizon is a week and overtime is not allowed. The algorithm was run during 600 seconds. The results for the 3 scenarios are expressed as average relative variation from the best weighted sum solution and shown in Fig. 2. The trends are quite similar for both types of instances and for the two first scenarios. The average variation from the weighted sum baseline remains between -14 and 6% for the number of unscheduled visits and between -1, and 6% for the travel time and the loyalty objectives. For the third scenario, the number of unscheduled visits is about 10% higher in the hierarchical solutions than in the weighted sum solutions for both types of instances. The new method is able to greatly decrease the number of assigned nurses, around -30% for the three scenarios and for both types of instances. It is, therefore, able to improve the continuity of care. This

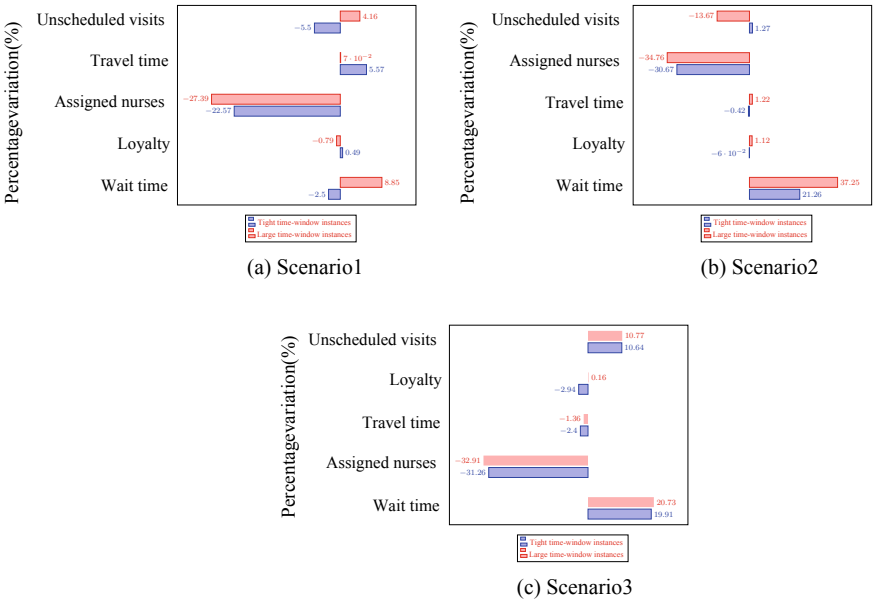


Fig. 2 Average relative variation of best hierarchical solution from baseline

Table 4 Baseline, best hierarchical solution and alternative solutions

Order	Objectives	Weighted sum baseline	Best hierarchical solution	Alternative solution 1	Alternative solution 2
1	Unscheduled visits	2	1	2	3
2	Daily overtime (min)	142	354	176	65
3	Travel time (min)	4394	4299	4205	4478
4	Assigned nurses	99	112	99	105
5	Wait time (min)	311	317	315	310

improvement is done at the expense of the waiting time (it deteriorates, on average, up to 37%) which is not surprising given the fact that this objective is ranked last in all scenarios.

To present examples of alternative solutions, an instance of 7 nurses, 217 visits to schedule and 2 weeks planning horizon has been generated. The CPU time is 600 seconds. As a result, 12 alternative solutions were kept in the pruned Pareto list but for the sake of simplicity only two are presented in Table 4. In this example, we can see that the new method is able to schedule one additional visit with less total travel time and similar values for the objectives. assigned nurses, and wait time compared to the weighted sum solution but with twice as much daily overtime. Regarding the alternative solutions, the alternative solution 1 is quite similar to the baseline with around 30 more minutes of overtime and 200 less minutes of travel time. The alternative solution 2 offers to increase the number of unscheduled visits by 2 and the travel time by approximately 180 min in order to save almost 300 min of overtime. These different trade-offs suggested by the new method are represented in Fig. 3 where we can see that the best hierarchical and the second alternative solution are two extremes, one with the lowest number of unscheduled visits and the other the least daily overtime. As for alternative solution 1, it sits between the two other solutions as a compromise. In this situation, only the decision makers are able to choose the right solution from their experience and their knowledge.

5 Conclusion

Our industrial partner *AlayaCare* is offering a scheduling tool in the home health care industry and has to deal with a significant amount of home health care service providers, each of which has different priorities. The ease of use of this technique and the speed at which the system provides an interesting set of solutions were important factors when choosing to develop and implement this method. In fact, the users are now relieved from the difficult task of setting weights to their priorities and this was achieved without degrading the quality of the proposed solution. With this technique, we bring a method based around existing algorithms to provide a set of interesting

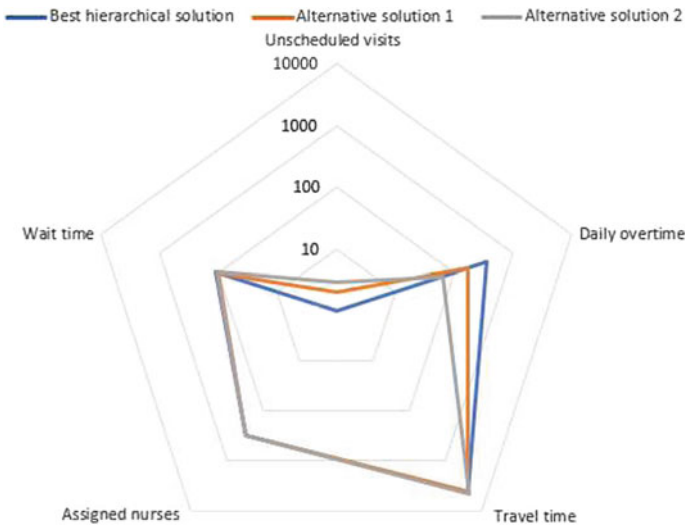


Fig. 3 Best hierarchical solution and alternative solutions—graphic representation

alternatives to a solution. This algorithm can be tweaked in many ways to be adapted to various problems dealing with multi-objective criteria. In particular, the algorithm that generates solutions should be adapted to fit the context of the problem. Further studies could be conducted using machine learning techniques to best predict the user choice (e.g. preferences for a certain type of solution).

References

1. Canadian Healthcare Association: Home Care in Canada: From the Margins to the Mainstream. (2009). ISBN: 978-1-896151-33-5
2. Fikar, C., Hirsch, P.: Home health care routing and scheduling: a review. Comput. Oper. Res. **77**, 86–95 (2017)
3. Cissé, M., Yalçndağ, S., Kergosien, Y., Şahin, E., Lenté, C., Matta, A.: OR problems related to home health care: a review of relevant routing and scheduling problems. Oper. Res. Health Care **13**, 1–22 (2017)
4. Grenouilleau, F., Legrain, A., Lahrichi, N., Rousseau, L.-M.: A set partitioning heuristic for the home health care routing and scheduling problem. Eur. J. Oper. Res. **275**(1), 295–303 (2019)
5. Decerle, J., Grunder, O., Hajjam El Hassani, A., Barakat, O.: A general model for the home health care routing and scheduling problem with route balancing. In: IFAC PapersOnLine 50-1, pp. 14662–14667 (2017)
6. Saaty, T.L.: How to make a decision: the analytic hierarchy process. Eur. J. Oper. Res. **48**(1), 9–26 (1990)
7. Jafari, H., Salmasi, N.: Maximizing the nurses' preferences in nurse scheduling problem: mathematical modeling and a meta-heuristic algorithm. J. Indus. Eng. Int. **11**(3), 439–458 (2015)

8. Oughalime, A., Ismail, W.R., Yeun, L.C.: A tabu search approach to the nurse scheduling problem. In: 2008 International Symposium on Information Technology, vol. 1, pp. 1–7. IEEE (2008)
9. Morizawa, K., Hirabayashi, N.: A heuristic approach for nurse scheduling under two and three-shifts workers mixed situation. DEStech Trans. Eng. Technol. Res. (icpr) (2017)
10. Ferland, J.A., Berrada, I., Nabli, I., Ahiod, B., Michelon, P., Gascon, V., Gagné, É.: Generalized assignment type goal programming problem: application to nurse scheduling. J. Heuristics **7**(4), 391–413 (2001)
11. Saji, Y., Riffi, M.E., Ahiod, B.: Multi-objective ant colony optimization algorithm to solve a nurse scheduling problem. Int. J. Adv. Res. Comput. Sci. Softw. Eng.**3**(8), (2013)
12. Ke, L., Zhai, L.: A multiobjective large neighborhood search for a vehicle routing problem. In: International Conference in Swarm Intelligence, pp. 301–308. Springer, Cham (2014)
13. Hsu, W.H., Chiang, T.C.: A multiobjective evolutionary algorithm with enhanced reproduction operators for the vehicle routing problem with time windows. In: 2012 IEEE Congress on Evolutionary Computation, pp. 1–8. IEEE (2012)
14. Ghoseiri, K., Ghannadpour, S.F.: Multi-objective vehicle routing problem with time windows using goal programming and genetic algorithm. Appl. Soft Comput. **10**(4), 1096–1107 (2010)
15. Song, Q,. Zilecky, P., Jakob, M., Hrncirv J.: Exploring Pareto routes in multi-criteria urban bicycle routing. In: 2014 IEEE 17th International Conference on Intelligent Transportation Systems (ITSC). 8–11 Oct 2014. Qingdao, China
16. Burke, E.K., Li, J. Qu, R.: A Pareto-Based Search Methodology for Multi-Objective Nurse Scheduling. Springer (2009)
17. Guo, Z.X., Wong, W.K., Li, Z., Ren, P.: Modeling and Pareto optimization of multi-objective order scheduling problems in production planning. Comput. Indus. Eng. **64**(4), 972–986 (2013)
18. Drechsler, N., Sülflow, A., Drechsler, R.: Incorporating user preferences in many-objective optimization using relation ϵ-preferred. Nat Comput. **14**, 469–483 (2015)
19. Hall, R. (Ed.) Handbook of Healthcare System Scheduling. International Series in Operations Research and Management Science, Series (2012)
20. Borsani, V., Matta, A., Beschi, G., Sommaruga, F.: A homecare scheduling model for human resources. Serv. Syst. Serv. Manag. **1**, 449–454 (2006)
21. Torres-Ramos, A., Alfonso-Lizarazo, E., Reyes-Rubiano, L., Quintero-Araujo, C.: Mathematical model for the home health care routing and scheduling problem with multiple treatments and time windows. In: Proceedings of the 1st International Conference on Mathematical Methods and Computational Techniques in Science and Engineering, pp. 140–145 (2014)
22. Ribeiro, G.M., Laporte, G.: An adaptive large neighborhood search heuristic for the cumulative capacited vehicle routing. Problem. Comput. Oper. Res. **39**(3), 728–735 (2012)
23. Aksen, D., Kaya, O., Salman, F.S.: Tüncel, Ö.: An adaptive large neighborhood search algorithm for a selective and periodic inventory routing problem. Eur. J. Oper. Res. **239**(2), 413–426 (2014)
24. Ropke, S., Pisinger, D.: An adaptive large neighborhood search heuristic for the pickup and delivery problem with time windows. Transp. Sci. **40**(4), 455–472 (2006)
25. Braekers, K., Hartlc, R.F., Parraghc, S.N., Tricoirec, F.: A bi-objective homecare scheduling problem : analyzing the trade-off between costs and client inconvenience. Eur. J. Oper. Res. **248**, 428–443 (2016)
26. Martinez, C., Espinouse, M. L., Di Mascolo, M.: Continuity of care in home services: a client-centered heuristic for the home health care routing and scheduling problem. In: 2018 5th International Conference on Control, Decision and Information Technologies (CoDIT), pp. 1045–1050. IEEE (2018)
27. Colette, Y., Siarry, P.: Optimisation Multiobjectif. Eyrolles (2002). ISBN: 2-212-11168-1
28. Schaus, P., Hartert, R.: Multi-objective large neighborhood search. In: International Conference on Principles and Practice of Constraint Programming, pp. 611–627. Springer, Berlin, Heidelberg (2013)

A Two-Phase Method for Robust Home Healthcare Problem: A Case Study

Mahdyeh Shiri, Fardin Ahmadizar, Houra Mahmoudzadeh and Mahdi Bashiri

Abstract The adoption of home healthcare is occurring rapidly because it decreases pressure on in-patient hospital beds by providing care to patients at home. This paper proposes a novel hybrid approach to solving the uncertain problems associated with a robust home healthcare model. In the first phase of this two-phase robust approach, the qualified candidate locations for health facilities are selected by using a fuzzy analytic hierarchy process and grey rational analysis. In the second phase, a robust model considering several aspects such as over-qualification cost, violation of service time, and overtime is proposed. The objective function minimizes the total cost. Finally, a case study from the city of Sanandaj, Iran, is utilized to validate the solution in a real-world situation.

Keywords Home healthcare · Routing · Scheduling · Fuzzy analytic hierarchy process · Grey rational analysis · Robust optimization

1 Introduction

In the home healthcare problem, health services are provided by operators (e.g., nurses, physicians, social workers) in the patients' home to reduce hospitalization expenses and increase patient satisfaction. Home healthcare is growing in impor-

M. Shiri (✉) · F. Ahmadizar
University of Kurdistan, Sanandaj, Kurdistan, Iran
e-mail: Mahdyeh.shiri@uwaterloo.ca

F. Ahmadizar
e-mail: F.ahmadizar@uok.ac.ir

H. Mahmoudzadeh
University of Waterloo, Waterloo, ON, Canada
e-mail: Houra.Mahmoudzadeh@uwaterloo.ca

M. Bashiri
Shahed University, Tehran, Iran
e-mail: Bashiri@shahed.ac.ir

V. Bélanger et al. (eds.), *Health Care Systems Engineering*,
Springer Proceedings in Mathematics & Statistics 316,
https://doi.org/10.1007/978-3-030-39694-7_15

tance, especially with the increase in elderly people prefer to receive health services at home [1]. The process of planning human resources in home healthcare is of great importance and led to the formulation of the Home Healthcare Routing and Scheduling Problem (HHRSP) which is receiving increasing attention in the literature. Uncertainty in patient health status causes uncertainty in service times, which in turn affects routing and scheduling decisions. All patients need to receive routine service and, therefore, in addition to considering the optimality of solutions in terms of cost, we must also examine the feasibility of the model in terms of patient services. The purpose of this paper is to simultaneously minimize the overall costs of establishing and operating a home healthcare system and to ensure each patient is served within a reasonable time window while also considering all possible scenarios of patient service times.

In recent literature of HHRSP, facility locations are assumed to be fixed in most papers (see [2–14]). Some authors consider only the cost factor when selecting the best facilities [15, 16]. To the best of our knowledge, no work has considered all different criteria in addition to the cost of selecting potential facility locations. Most recent work in the literature focuses on a single-period problem using deterministic approaches [3–9, 12–16]. However, total cost, overqualified skill levels, overtime, and violation of service time have not been considered in the home healthcare routing and scheduling problem simultaneously. Some papers have considered stochastic optimization for facing inherent uncertainty in parameters [2, 4, 9, 11]. Robust optimization has also been applied to these problems to address uncertainty in demand and travel time [10, 13]. The Mulvey approach [17] is a specific type of robust optimization that tackles scenario-based uncertainty that can be evaluated by two measures: optimality (solution robustness) and feasibility (model robustness). Solution robustness shows how "close" a solution is to optimality under all scenarios, and model robustness denotes that the model stays "almost" feasible under all scenarios. To the best of our knowledge, no work has considered the trade-off between optimality and feasibility in the HHRSP using this approach.

This paper proposes a Two-Phase Multi-criterion Robust Home Healthcare Problem (2P-MRHHP). The specific contributions of this paper are as follows:

1. We use a hybrid Fuzzy Analytic Hierarchy Process and Grey Rational Analysis (FAHP-GRA) method to select facilities based on multiple criteria.
2. We propose a mixed-integer programming model for HHRSP considering overqualified skill levels and overtime.
3. We provide a robust optimization model based on the Mulvey approach [17] that could immunize the optimality of the solution and the feasibility of the model simultaneously under uncertainty in service times.
4. We demonstrate the results of our proposed approach on a real case study.

The rest of the paper is structured as follows: The 2P-MRHHP methodology is defined in Sect. 2 along with problem definition and mathematical formulation. Section 3 provides the numerical results of our case study and sensitivity analyses on critical parameters. Finally, conclusions and future research directions are shown.

2 Solution Methodology

In this section, we introduce the 2P-MRHHP approach. In the first phase, a hybrid FAHP-GRA is used for selecting the candidate facility locations. These locations then serve as the input data for the second phase. In the second phase, a model for a robust routing and scheduling problem is proposed to select the top facilities based on criteria, and the optimal routing and scheduling of the healthcare team are determined.

2.1 First Phase: Selecting the Candidate Locations

In this phase, a strategic-level decision is made based on long-term qualitative criteria (other than cost). We use a hybrid FAHP-GRA [15, 18] for ranking the candidates based on these criteria. We make pairwise comparisons between the candidate locations under each criterion. Triangular fuzzy numbers are considered to capture the fuzzy expert opinion. More specific details are provided with the numerical case study in Sect. 3.3.

The few top candidates, which are all considered *good* strategic options, are then selected as the input of the second phase where the optimal location among these top candidates is selected.

2.2 Second Phase: Robust Optimization Model

In the second phase, two decisions are made jointly with the objective of minimizing total cost: (i) selecting a facility from the top options identified in the first phase; and (ii) finding the optimal routing and scheduling for all nurse teams. In this phase, the only monetary cost is considered in making these simultaneous decisions.

The model considers teams of nurse team that depart the healthcare center and visit multiple patients within a set period according to a given schedule and pre-specified route. At the end of each route, the nurse teams go to the laboratory to deliver all samples (e.g., blood) taken. Team structures stay fixed for each period and route. The nurses have fixed a working hour and we consider a penalty for overtime cost. There is also an upper bound on how much overtime each nurse can have per period. The qualification (proficiency) level of nurses and the skills required for each patient are considered to differ according to patient conditions. We consider a penalty cost for assigning an overqualified nurse to a patient.

We consider uncertain service time and use three scenarios, each of which corresponding to a different service duration: *pessimistic*, *most likely*, and *optimistic*. These three scenarios are set by nurse experts and are decided based on patients' health conditions. Ensuring the constraints are met for all possible scenarios may not

be feasible; however, we would like to make the solution as "near-feasible" as possible to avoid adverse effects. Therefore, in addition to considering the optimality of solutions in terms of cost, the feasibility of the model in terms of patient services, and the trade-off between the two are of great importance. Therefore, we use the Mulvey approach, which fits well with the home healthcare routing and scheduling problem tackled in this paper. Applying the Mulvey approach to our model, solution robustness is achieved by minimizing the expected value and standard deviation of the objective function among all scenarios. Model robustness is achieved by minimizing the number of violations in the service time constraint [17].

The notations used in the robust model (indices, sets, parameters, decision variables) and all problem assumptions are summarized in Table 1.

The model formulation is presented below. Note that for teams with no assignments, empty routes (from the healthcare center to laboratory) are allowed.

$$\min\ Z = \sum_{s\in\mathscr{S}} p_s\xi_s + \lambda \sum_{s\in\mathscr{S}} p_s[(\xi_s - \sum_{s'\in\mathscr{S}} p_{s'}\xi_{s'}) + 2\theta_s] + \omega \sum_{s\in\mathscr{S}}\sum_{r\in\mathscr{I}} p_s\delta_{rs} \tag{1}$$

s.t.

$$\xi_s = w^c(\sum_{h\in\mathscr{I}} c_h u_h + \sum_{l\in\mathscr{I}} c_l u_l \sum_{i,j\in\mathscr{A}}\sum_{k\in\mathscr{K}}\sum_{t\in\mathscr{T}} cv_{ij}x_{ijkts} + \sum_{k\in\mathscr{K}}\sum_{t\in\mathscr{T}} c_k o_{kts})$$

$$+w^q p \sum_{r\in\mathscr{I}}\sum_{k\in\mathscr{K}}\sum_{t\in\mathscr{T}} \phi_{rkts}, \qquad \forall s \tag{2}$$

$$\xi_s - \sum_{s'\in\mathscr{S}} p_{s'}\xi_{s'} + h_s \geq 0, \qquad \forall s \tag{3}$$

$$\phi_{rkts} = \sum_{n\in\mathscr{N}} \beta_{nk} z_{nkt} - \alpha_{rs} y_{rkts}, \qquad \forall r,k,t,s \tag{4}$$

$$\sum_{h\in\mathscr{I}} u_h = 1, \tag{5}$$

$$\sum_{l\in\mathscr{I}} u_l = 1, \tag{6}$$

$$x_{hrkts} \leq u_h, \qquad \forall h,r,k,t,s \tag{7}$$

$$x_{rlkts} \leq u_l, \qquad \forall l,r,k,t,s \tag{8}$$

$$\sum_{k\in\mathscr{K}}\sum_{t\in\mathscr{T}} y_{rkts} = 1, \qquad \forall r,s \tag{9}$$

$$\sum_{j:(r,j)\in\mathscr{A}} x_{rjkts} = y_{rkts}, \qquad \forall r,k,t,s \tag{10}$$

$$\sum_{j:(h,j)\in\mathscr{A}} x_{hjkts} = 1, \qquad \forall k,t,s \tag{11}$$

$$\sum_{i:(i,l)\in\mathscr{A}} x_{ilkts} = 1, \qquad \forall k,t,s \tag{12}$$

Table 1 Indices, parameters, variables, and assumptions

Sets		β_{nk}	The skill level of nurse n in team k
$i, j \in \mathscr{I}$	The set of all nodes($r, h, l \in \mathscr{I}$)	p	Penalty cost for overqualified skill levels
$k \in \mathscr{K}$	The set of teams		
$t \in \mathscr{T}$	The set of time periods	*Indices*	
$s \in \mathscr{S}$	The set of scenarios	r	Patient homes
$n \in \mathscr{N}$	The set of nurses	h	Healthcare center candidates
$\mathscr{A}$	The set of arcs	l	Laboratory candidates
Parameters		*Decision variables*	
$o^{\max}$	The maximum overtime	u_h	Equal to 1 if a healthcare center is established at nodes h, 0 otherwise
v_{ij}	Traveling time from node i to node j	u_l	Equal to 1 if a laboratory is established at nodes l, 0 otherwise
d_{js}	The service time for node j under scenario s	z_{nk}	Equal to 1 if nurse n is assigned to team k in period t, 0 otherwise
c_k	The cost of overtime for team k	x_{ijkts}	Equal to 1 if team k goes directly from node i to node j in period t under scenario s, 0 otherwise
m	The number of nurses in a team	y_{rkts}	Equal to 1 if team k is assigned to patient r in period t, 0 otherwise
f	The upper bound of the working hours	δ_{rs}	Violation of service time in the patient home r under scenario s
e	The lower bound of the working hours	θ_s	Non-negative deviational variable of objective function under each scenario s
w^c	The weight for costs	ξ_s	Total cost under scenario s
w^q	The weight for overqualified skill levels	o_{kts}	Team's overtime k in period t under scenario s
a_{rt}	The earliest starting time window for patient r in period t	Φ_{rkts}	the overqualified level of team k for patient r in period t
b_{rt}	The latest starting time window for patient r in period t	s_{ikts}	Start time for patient i by team k in period t under scenario s
p_s	The probability of scenario s		
λ	The variability weight	*Assumptions*	
ω	The risk-aversion weight	Two nurses in each team	
c	The unit traveling cost per kilometer	Three scenarios for service time	
c_h	Cost of establishing a healthcare center	Time horizon is a week	
c_l	Cost of establishing a laboratory	One healthcare center needed	
α_{rs}	The proficiency required for patient r under scenario s	One laboratory needed	

$$\sum_{i:(i,r)\in\mathscr{A}} x_{irkts} - \sum_{j:(r,j)\in\mathscr{A}} x_{rjkts} = 0, \qquad \forall r,k,t,s \tag{13}$$

$$x_{rjkts}(s_{rkts} + v_{rj} + d_{rs} - s_{jkts}) \le 0, \qquad \forall r,j:(r,j)\in\mathscr{A},k,t,s \tag{14}$$

$$y_{rkts}(a_{rt} - s_{rkts}) \le 0, \qquad \forall r,k,t,s \tag{15}$$

$$y_{rkts}(s_{rkts} - b_{rt} - \delta_{rs}) \le 0, \qquad \forall r,k,t,s \tag{16}$$

$$x_{hrkts}(s_{rkts} - e - v_{hr} - d_{hs}) \ge 0, \qquad \forall r,h,k,t,s \tag{17}$$

$$x_{rlkts}(s_{rkts} + v_{rl} + d_{rs} - f - o_{kts}) \le 0, \qquad \forall r,l,k,t,s \tag{18}$$

$$\sum_{k\in\mathscr{K}} z_{nkt} \le 1, \qquad \forall n,t \tag{19}$$

$$\sum_{n\in\mathscr{N}} z_{nkt} = m, \qquad \forall k,t \tag{20}$$

$$o_{kts} \le o^{\max}, \qquad \forall k,t,s \tag{21}$$

$$u_h, u_l, z_{nk}, x_{ijkts}, y_{rkts} \in \{0,1\}, \qquad \forall h,l,n,k,i,j,t,s \tag{22}$$

$$\Phi_{rkts}, o_{kts}, \delta_{rs}, \theta_s, \xi_s, s_{ikts} \ge 0, \qquad \forall i,r,k,t,s \tag{23}$$

The model aims to determine the optimal routing and scheduling under each scenario that minimizes the total home healthcare cost, consisting of the cost of establishing the healthcare center and laboratory, the cost for traveling between patient home and health facilities, the cost of nurse overtime, and the penalty for overqualified skill levels.

According to the Mulvey approach for scenario-based models [17], we formulate the objective function for the proposed robust model in (1)–(3). In constraint (4), the nurses can serve a patient if their skill level is above the minimum required level. Constraints (5) and (6) show that only one candidate must be selected for each of the health facilities. Constraints (7) and (8) allow the use of the facility only when it is selected. Constraints (9) and (10) state that a patient is allocated exactly once to a team of nurses in each period. In constraint (11), each team of nurses starts its route from the healthcare center. Constraint (12) guarantees that each team of nurses ends its route at the laboratory. Constraint (13) shows flow conservation for each team of nurses. Constraint (14) states that the service for the next patient can be started after the previous patient has been served and the team has traveled to the next patient's home. Constraint (15) implies that a patient should be served within a certain time window. Constraint (16) is *control* constraint that calculates any violation of service time. Constraints (17) and (18) define the duration of routes and calculate overtime by considering the working hours of a nurse. Constraint (19) implies that a nurse is assigned to at most one team per period. Constraint (20) determines the number of teams of nurses available. Constraint (21) ensures that the overtime is less than the allowed limit. Constraints (22) and (23) indicate binary and sign constraints, respectively. Note that the nonlinear constraints can easily be linearized based on the method explained in reference [19].

3 Implementation and Evaluation

3.1 Case Explanation

To show the practicality and validity of the model, we present a real case study. Our data was provided by the Kosar hospital in the city of Sanandaj, Iran. Sanandaj is divided into 10 districts, shown in Fig. 1, each of which is a candidate location for establishing a healthcare center. Districts 1, 3, 4, 8 and 9 are potential candidates for establishing a laboratory.

The models were solved using GAMS and CPLEX on a Core i5-5257U laptop with 2.70 GHz CPU and 8 GB RAM. The problem was solved to optimality (zero optimality gap) and the solution time for a problem with 5 periods, 20 patients, 3 healthcare center candidates, 3 laboratory candidates, 3 teams and 6 nurses was 29:59 min.

3.2 Selecting the Top Facility Locations

In the first phase, the strategic criteria considered for selecting the top candidates for the healthcare center were: city planning, security, road access, cost level, and social impact. For the laboratory, these criteria were: city planning, security, climate, cost level, and political impact. Triangular fuzzy numbers for each criterion were provided for each district based on expert knowledge and documentation from Kosar hospital and the municipality of Sanandaj. We use the hybrid FAHP-GRA method [15, 18] to rank all healthcare center and laboratory candidates. First, the fuzzy numbers for each criterion of candidates are changed to interval values using an α-cut of 0.5. Next, the normalized values are achieved by interval values. The weighted normalized value matrix is then calculated by the normalized interval multiplying of

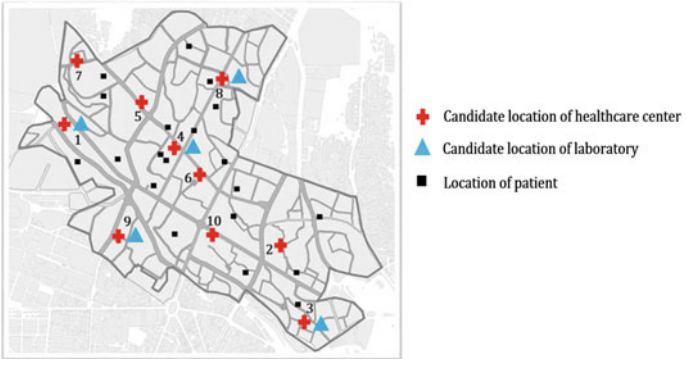


Fig. 1 Geographical dispersion and location of MRHHP facilities

each candidate under each criterion. The weights of the criteria are determined by experts. Next, the reference value vector under every criterion is acquired using two measures: (1) maximum of minimum and (2) maximum of maximum. The maximum start value of the weighted normalized interval is represented for all candidates, and the second step is the maximum end value of all candidates from the first step. Then, we calculate the maximum distance for every candidate from its reference value. We also obtain the minimum and maximum distance for every criterion of candidates. Finally, we calculate the weighted distance from the reference value vector and the average score for every candidate.

3.3 Optimal Route Visualization

Based on the ranking approach explained in the first phase, three top scenarios were selected by the municipality as the potential locations for consideration in the second phase. The robust model then finds the optimal routes as well as optimal facility locations for all scenarios. Note that the robust problem provides all these solutions simultaneously, and therefore the locations of the facilities are fixed regardless of the scenario whereas the routes will change depending on the scenario.

In Fig. 2, we show a small sample to illustrate the solution for a problem with 20 patients, 2 nurse teams, 2 periods and 3 scenarios. Districts 8 and 4 are selected as the optimal locations for the healthcare center and laboratory, respectively. The optimal routes are illustrated in Fig. 2.

3.4 Optimality and Feasibility of Mulvey Approach

To explore the trade-off between optimality and feasibility, a risk-aversion weight (ω) is used. A risk-averse DM who strictly avoids a violation of service time selects a higher ω. Figure 3a explains the trade-off between optimality and feasibility for different values of risk-aversion weight—as ω increases, the violation of service time (feasibility) decreases, while the overall cost (optimality) increases. The expected violation will eventually stabilize at a higher weighting penalty. This trade-off can help in determining a suitable value for the weighting penalty. Mulvey and Ruszczynski [17] report similar results. Figure 3b, on the other hand, looks at variations in the *variability weight* λ. A higher value for λ does not necessarily result in a higher cost and also does not result in reduced feasibility. The lowest service time violation is observed at $\lambda = 0.5$. Feasibility and optimality are insensitive to changes in λ after a certain threshold.

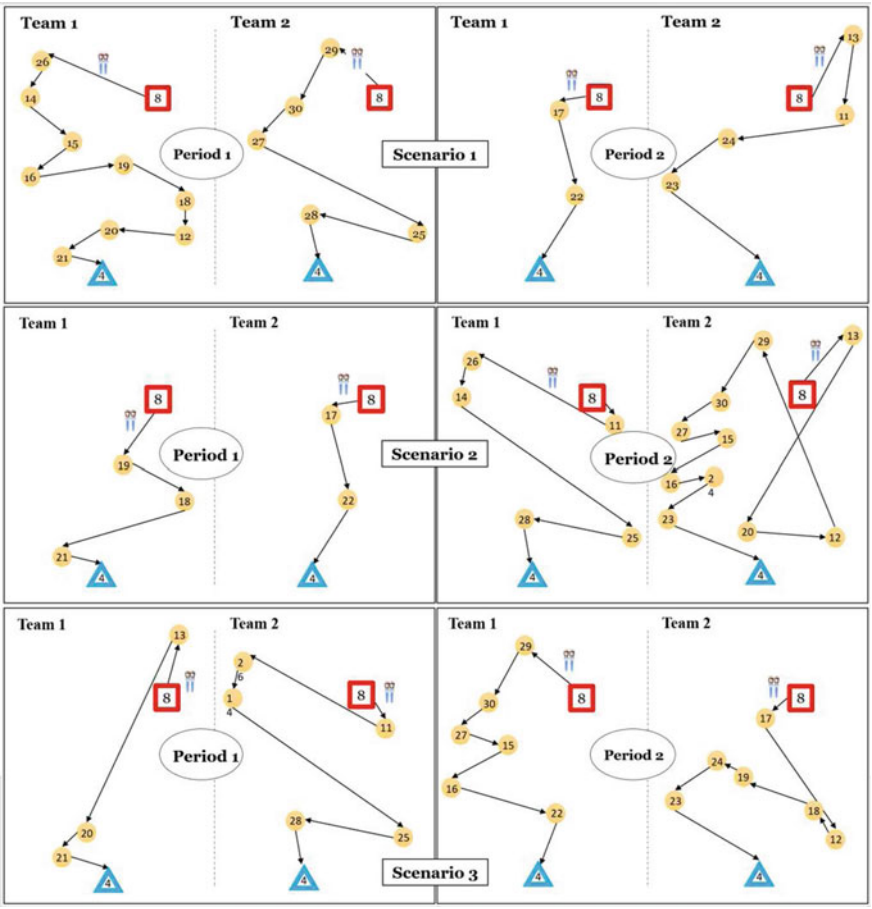
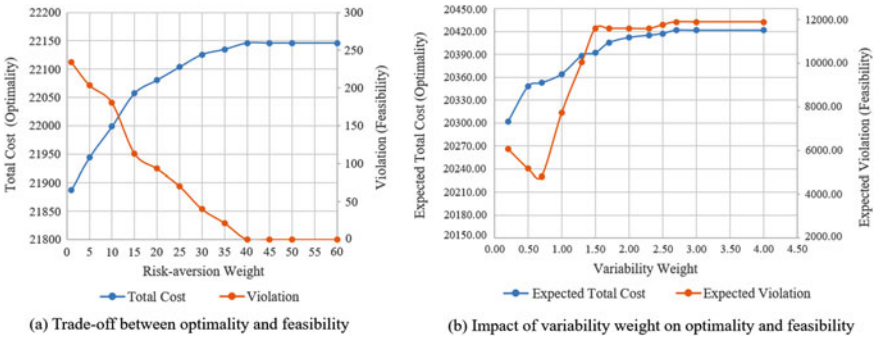


Fig. 2 The HHRSP optimal routes



(a) Trade-off between optimality and feasibility

(b) Impact of variability weight on optimality and feasibility

Fig. 3 The impact of **a** risk-aversion weight, **b** variability weight, on optimality and feasibility

4 Conclusions

In this paper, we developed a two-phase multi-criterion robust home healthcare problem. In the first phase, we used a hybrid method for selecting health facilities in which decision makers consider a variety of criteria to choose the top strategic candidates. In the second phase, we proposed a robust approach to select the optimal facilities among the top candidates and make optimal routing and scheduling decisions simultaneously. To consider a trade-off between feasibility and optimality, we used the Mulvey approach for scenario-based uncertainty. Numerical experiments are shown for a real-life case study. Our future research direction includes considering a cardinality-constrained robust optimization approach for considering uncertainties in individual patient's service times.

References

1. Fikar, C., Hirsch, P.: Home health care routing and scheduling: a review. Comput. Oper. Res. **77**, 86–95 (2017)
2. Rodriguez, C., Garaix, T., Xie, X., Augusto, V.: Staff dimensioning in homecare services with uncertain demands. Int. J. Prod. Res. **53**(24), 7396–7410 (2015)
3. Fikar, C., Hirsch, P.: A matheuristic for routing real-world home service transport systems facilitating walking. J. Clean. Prod. **105**, 300–310 (2015)
4. Yuan, B., Liu, R., Jiang, Z.: A branch-and-price algorithm for the home health care scheduling and routing problem with stochastic service times and skill requirements. Int. J. Prod. Res. **53**(24), 7450–7464 (2015)
5. Rest, K.D., Hirsch, P.: Daily scheduling of home health care services using time-dependent public transport. Flex. Serv. Manuf. J. **28**(3), 495–525 (2016)
6. Braekers, K., Hartl, R.F., Parragh, S.N., Tricoire, F.: A bi-objective home care scheduling problem: analyzing the trade-off between costs and client inconvenience. Eur. J. Oper. Res. **248**(2), 428–443 (2016)
7. Redjem, R., Marcon, E.: Operations management in the home care services: a heuristic for the caregivers' routing problem. Flex. Serv. Manuf. J. **28**(1–2), 280–303 (2016)
8. Yalçındağ, S., Matta, A., Şahin, E., Shanthikumar, J.G.: The patient assignment problem in home health care: using a data-driven method to estimate the travel times of care givers. Flex. Serv. Manuf. J. **28**(1–2), 304–335 (2016)
9. Shi, Y., Boudouh, T., Grunder, O.: A Home health care routing problem with stochastic travel and service time. IFAC-PapersOnLine **50**(1), 13987–13992 (2017)
10. Cappanera, P., Scutellà, M.G., Nervi, F., Galli, L.: Demand uncertainty in robust Home Care optimization. Omega (United Kingdom). **80**, 95–110 (2018)
11. Shi, Y., Boudouh, T., Grunder, O., Wang, D.: Modeling and solving simultaneous delivery and pick-up problem with stochastic travel and service times in home health care. Expert Syst. Appl. **102**, 218–233 (2018)
12. Fathollahi-Fard, A.M., Hajiaghaei-Keshteli, M., Tavakkoli-Moghaddam, R.: A bi-objective green home health care routing problem. J. Clean. Prod. **200**, 423–443 (2018)
13. Hosseini-Motlagh, S.M., Samani, M.R.G., Cheraghi, S.: Robust and stable flexible blood supply chain network design under motivational initiatives. Socio-Econ. Plan. Sci. 100725 (2019)
14. Decerle, J., Grunder, O., Hajjam El Hassani, A., Barakat, O.: A memetic algorithm for a home health care routing and scheduling problem. Oper. Res. Heal. Care **16**, 59–71 (2018)
15. Samani, M. R. G., Hosseini-Motlagh, S.M.: An enhanced procedure for managing blood supply chain under disruptions and uncertainties. Ann. Oper. Res. **283**(1), 1413–1462 (2019)

16. Rodriguez-Verjan, C., Augusto, V., Xie, X.: Home health-care network design: location and configuration of home health-care centers. Oper. Res. Heal. Care **17**, 28–41 (2018)
17. Mulvey, J.M., Vanderbei, R.J., Zenios, S.A.: Robust optimization of large-scale systems. Oper. Res. **43**(2), 264–281 (1995)
18. Samvedi, A., Jain, V., Chan, F.T.S.: An integrated approach for machine tool selection using fuzzy analytical hierarchy process and grey relational analysis. Int. J. Prod. Res. **50**(12), 3211–3221 (2012)
19. Norouzi, N., Tavakkoli-Moghaddam, R., Ghazanfari, M., Alinaghian, M., Salamatbakhsh, A.: A new multi-objective competitive open vehicle routing problem solved by particle swarm optimization. Netw. Spat. Econ. **12**(4), 609–633 (2012)

Adverse Event Prediction by Telemonitoring and Deep Learning

Antoine Prouvost, Andrea Lodi, Louis-Martin Rousseau and Jonathan Vallee

Abstract Home health care comes as a potential solution to increasing stress on health-care systems, as well as concerns for medical patients comfort. However, additional distance from the care workers to the patients lead to more challenges, some of which can be addressed with machine learning (ML) and operations research (OR) algorithms. In this paper, we focus on automating a risk assessment of remote patients. Namely, we describe a risk prediction framework for home telemonitoring patients and show that learning a risk from weak signals in the patient's data outperforms simple risk threshold proposed by care workers to automate the task. We combine recurrent neural networks with a ranking objective from survival analysis to evaluate the risk of patient's adverse events. Training and testing of our methodology is achieved on a retrospective dataset gathered by an Ontario home health care agency during the course of a multi-year pilot home telemonitoring program. Results are benchmarked against alerts that were manually engineered by registered nurses, and against a simple linear baseline. This is an additional step in the application of machine learning in health care for patient-centered personalized treatments.

Keywords Home health care · Telemonitoring · Time-series prediction · Deep learning

A. Prouvost (✉) · A. Lodi · L.-M. Rousseau
École Polytechnique de Montréal, Montreal, Canada
e-mail: antoine.prouvost@polymtl.ca

A. Lodi
e-mail: andrea.lodi@polymtl.ca

L.-M. Rousseau
e-mail: louis-martin.rousseau@polymtl.ca

J. Vallee
AlayaCare, Montreal, Canada
e-mail: jonathan.vallee@alayacare.com

V. Bélanger et al. (eds.), *Health Care Systems Engineering*,
Springer Proceedings in Mathematics & Statistics 316,
https://doi.org/10.1007/978-3-030-39694-7_16

1 Introduction

Population ageing comes with increased care needs since 85% of elderly will develop chronic conditions [39]. From 6.9% in 2000 to an estimated proportion of 19.3% of the global population in 2050 [10], the elderly account for a growing proportion of the health care costs.

Keeping elderly healthy and at home longer is thus a critical endeavour. Home Health Care (HHC) is starting to be widely adopted since it is seen as a cost effective alternative to traditional care and because patients often prefer it.

Home telemonitoring (HT), a specialized form of HHC is a potential alternative that empowers patients to take charge of their health, generates reliable data that can be leveraged to better assess the patients' states and that may improve the patient's medical condition [27].

Given the growing demand for HHC and HT, data is accumulating at an extreme velocity, in a great volume and in a variety of forms. The advancements in monitoring devices is also contributing to the velocity and volume of data generated by HT programs. Valuable information lies in this data. There is thus a pressing need for improved decision systems that can use the information.

When a patient is admitted to a home care agency, she generally gets visited by a registered nurse who will perform an initial needs assessment [33]. If the agency offers a HT program, patients can be admitted to it. While on a HT program, the patient answers a periodic questionnaire during which she will be asked to take some vital signs readings. This information is then transmitted to the HHC agency where a nurse monitors a HT case load. Based on the patient diagnosed conditions and initial assessment, the care workers create alerts based on acceptable range of each of the measured vital signs as shown in [35]. Sometimes, with more advanced systems, complex rules can be developed to get alerted based on combinations of suspect readings.

In all cases, care workers bear the weight of setting up patients with the right set of alerts based on their conditions. The manually engineered rules then need to evolve with the patient's condition in order to remain reliable. When a vital sign reading is out of the acceptable range, the monitoring nurse can perform one or two of the following actions: (1) call the patient to determine next steps, and/or (2) schedule an in-person visit. The challenge is to prevent costly hospital readmissions and emergency room visits, but there is also a cost to each intervention. To add complexity, most of the alarms are false positives, not leading to adverse events.

Early detection of these events serves the purpose of the triple aim of improving outcomes: (1) quality of health services, (2) improving health of populations and (3) reducing costs [3].

The contribution of this paper is to propose a patient ranking approach to adverse events prediction and to show that it performs well on a retrospective patient cohort.

The remainder of the paper is organized as follows: In Sect. 2, we review the body of work related to adverse events predictions. In Sect. 3, we review the technical

background required for our proposed solution to adverse events prediction. Section 4 details our proposed framework and Sect. 5 reports our results. We conclude the paper with a discussion and perspectives in Sect. 6.

2 Adverse Event Prediction Related Work

Since about 20% of Medicare patients are readmitted within 30 days of discharge [18] and since [17] established financial penalties to hospital with the highest readmission rates 30 days after discharge, prediction of adverse events such as hospital readmissions has been extensively done in health care research.

Linear models such as multivariate logistic regression and Cox Proportional Hazard [12, 20, 34, 38] are often used because of their understandable nature. Indeed, most of the work so far has been interested in understanding the significant factors that lead to adverse events. Conversely, neural networks have not been used as much because they are seen as hard to interpret black-boxes [42] despite their success in many industries, from computer vision to market finance.

In health care, some examples of neural network use are as diagnostic tool such as in [2], as prediction tools in [14, 24], in emergency states detectors [36], and in psychology [31]. In particular, [2] hypothesizes that neural networks could outperform linear models because of their capacity to capture relationships between input variables that are not seen by simpler models. More recently, additional work has been done using neural networks to anticipate patient outcome (mortality, readmission, extended stay, etc.) from their electronic health records [1, 32], including using recurrent neural networks [9].

3 Technical Background

Neural networks are parametric functions approximators built by composition. The network is built by alternatively composing matrix multiplications and a non linear (element-wise) function, called activation function (such as the Rectified Linear Units (ReLU), $x \mapsto \max(x, 0)$). This type of architecture is that of a multi-layer perceptron. Finding the coefficients of the matrices building the network is an optimization problem, also called "*learning*" or "*fitting*" the model. The loss function of this problem depends on the input data. In supervised learning, neural networks are optimized on a training set to minimize a loss function between their prediction and an observed target. The nature of the loss function depends on the task. Usually, mean square error is used for regression and cross-entropy for classification. To minimize the training error, neural networks are designed differentiable, and optimized using Stochastic Gradient Descent (SGD), a gradient descent approach where the loss is approximated over a subset of the training examples (called a "*mini-batch*"). Optimizing on a training set eventually leads to degrading performances on unseen

examples (this is known as *over-fitting*). To curtail this, the performance of the neural network is monitored on a dataset of unseen examples, known as the *validation* set, and optimization is stopped once the validation metrics worsen. This is done in combination with regularization techniques: any method that empirically reduces the test error, at the expense of the training error.

Recurrent neural networks, reuse their internal parameters sequentially to be able to process and learn over time series data of arbitrary length. At every step, they combine information coming from the current time step with past information condensed by the neural network into a hidden state vector. Popular models include the long short term memory (LSTM) [15] and Gated Recurrent Units (GRUs) [5]. When time series are long, recurrent neural networks can be trained with truncated back propagation through time [41]. Namely, the gradients are approximated, and optimization steps are taken, over consecutive subsets of data along the time dimension. Data global to a time series can be combined with the neural network, for instance through the initial hidden vector. The reader is referred to [11] for an extensive textbook on deep learning.

4 Adverse Events Prediction

Because predicting adverse events, either as a regression (for the time to the next adverse event), or as a very in-balanced classification (classifying if an adverse event will occur in the next k days), is hard, we chose to model the problem as a survival task. Namely, we predict a *latent* patient risk of experiencing an adverse event. It is important to understand that this risk is interpreted only relative to other patient risks, that is a patient risk is higher than another if the former is more likely to experience an adverse event than the latter. In other words, this risk is only introduced to output a ranking of patients. Given a predicted ranking, and the true ranking (computed from the times to the next adverse event), we can compute a score metric called *concordance index* (C-index) [13]. This measure is not differentiable and therefore cannot be optimized through gradient based methods (as it is done for neural networks). Hence, we use a surrogate loss derived from the maximum likelihood of the Cox proportional hazard model for survival analysis [7]. This likelihood loss function is used to compute gradients for the neural network, but model selection and final scores are expressed in terms of C-index. The detail of these loss function can be burdensome and we deliberately omit it here. In short, the C-index is a measure counting the percentage of pairs (of patients here) properly ranked. The Cox model makes more assumptions on the mathematical form of the risk function in order to derive a likelihood. Both are able to deal with censored data, i.e. patients exiting the program (or the program ending). The interested reader is referred to the aforementioned literature, as well as the adaptation for neural networks introduced in [21]. Unlike in Cox proportional hazard model, the problem contains a strong time component as we wish to predict a risk for every patient *on every day* (with new information coming in). The metrics are therefore evaluated on a daily basis across all

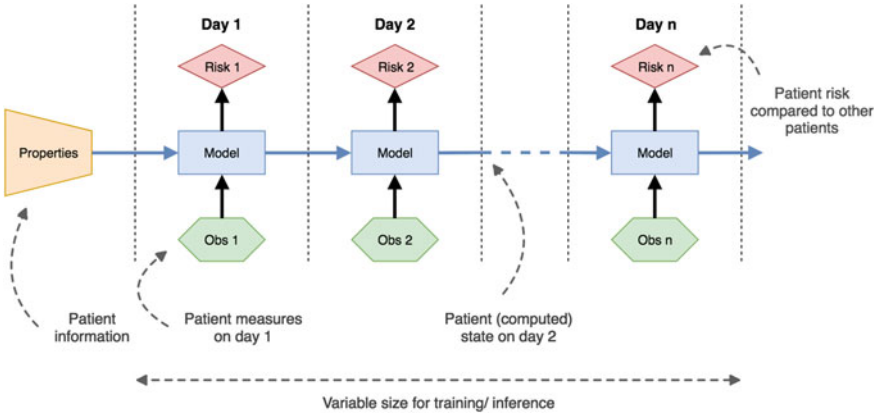


Fig. 1 A recurrent neural network is used to combine the patients *a priori* information with their daily vital signs and output a risks

patients. Modeling the problem as a ranking problem is in line with the application pursued here. Indeed, on every day, it is sufficient for the care giver to be able to rank the patients by risk, in order to provide an intuition on where to prioritize, as care workers cannot visit all patients on a daily basis.

As depicted in Fig. 1, we use a recurrent neural network to process patient data, both static (patient information and diagnoses) and time distributed (vital signs measured on a daily basis). The network outputs a risk for every patient, on every day. We use the Cox log-likelihood as a loss function to train the network, as was previously done by Katzman et al. [21], and report the C-index as well.

Static patient data contain medical diagnoses codes from the International Statistical Classification of Diseases and Related Health Problems (ICD9, ICD10) [26], which are not very informative on their own. To tackle this issue [6] uses an unsupervised deep learning approach to embed these codes into a more meaningful vector space, using additional information from the disease, as well as other types of codes. The embedded codes show desirable properties such as diseases with similar symptoms or prescriptions are close together in Euclidean distance. We used their pre-trained embeddings to represent this part of our data. Because patients have a variable number of diseases, we need to use another (small) recurrent neural network (hidden inside the orange parallelogram in Fig. 1) to process these diseases before passing them on to the main (larger) recurrent neural network. Alternatively, we also try not to include that information, and simply pass a null vector (as usually done), with the intuition that learning should be easier.

We face more challenges as we have some missing data in the vital signs observed on a daily basis. Although more complex solutions exist to model this (e.g., [4]), we chose the simple approach of modelling missing data with additional binary variables representing if the data is missing.

The data is split into training (60%), validation (20%) and test (20%) sets. We made sure that no information from the future could be used to make prediction and hence computed the splits as dates: from the beginning to the first split date would constitute the training set etc. Past events (previous targets) however can be used as input after their occurring date. Some patients appear in multiple sets, while other are new in every one. This is because of the nature of the application as we want to keep assessing the risk of patients throughout their participation in the program and not just once.

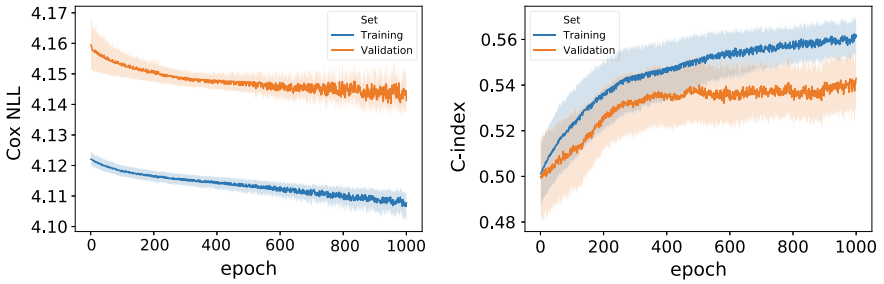
Neural networks and their optimization algorithms come with a number of so called "*hyper-parameters*": parameters that cannot be optimized directly. In our case, these hyper-parameters divide into two groups. The first group controls the optimization algorithm itself: SGD, or the Adam variant [22], the gradient step size, the L_2 regularization multiplier, the number of examples in a SGD mini-batch, the number of time steps considered in the truncated version of back-propagation through time, and the use of dropout (a regularization technique) [22]. The second group controls architectural decisions: number and size of hidden layers (the matrices involved in the neural network), the type of recurrent layers (LSTM or GRU), and whether to use the patient static data as the initial hidden vector or to simply omit it. A pragmatic way to select these hyper-parameters is to generate randomly some configurations, train the networks for each of them and finally keep the one performing best (in C-index) on the validation set.

Expanding on the evaluation of machine learning performance, we compared it to manually engineered alerts (four levels of severity). We also compare to a simple linear survival baseline using the time distributed readings independently (time dependencies are not taken into account) without using static patient data.

5 Results

The dataset we use has been gathered by an Ontario private HHC agency during the course of a multi-year pilot HT program and is fully anonymized. The input data include the patient static information (sex, age, and medical records through ICD codes), and daily vital sign readings (blood glucose, systolic, diastolic, heart beat, SpO_2 oximeter, and weight). The dataset also contains observed adverse events experienced by the patients and used to compute the losses (either Cox log-likelihood or C-index). On any given day, past events are also added as input.[1] The 320 patients in our study were aged from 31 to 101 with an average age of 79. Women represented 56.25% of the patient group. Moreover, 36.25% of patients experienced at least one hospital readmission or emergency room visit while on the HT program. The average number of events per patient was 0.72 with a standard deviation of 1.30. The average number of events for the 36.25% of patients that experienced at least one was 1.98 with a standard deviation of 1.47. Finally, 91.56% of patients had comorbidities,

[1] This is correct as no information from the future is added to the input.



(a) Cox negative log-like lihood. (b) Concordance index.

Fig. 2 Training of the most promising set hyper-parameters, averaged over twenty runs

hypertension being the more frequent at 55.31% followed by chronic heath failure with 46.25%, diabetes with 39.69% and chronic obstructive pulmonary disease at 38.13%. In total, we have access to 76,359 daily patient observations.

The training process of neural networks was highly stochastic, with results often close to random, as there is a strong noise to signal ratio in the data. This made it difficult to differentiate promising models from random luck. Therefore, hyper-parameters configurations were manually selected from across all runs (over a hundred), based not only on validation performance, but as well on stability (reduced stochasticity during training) in addition to small gap between the training and validation scores.

These configurations were then retrained over twenty random seeds to average the results. The training for the best performing set of hyper-parameters is depicted in Fig. 2. Important details of this configuration are: three gated recurrent unit (GRU) layers with eight units each and dropout for the network architecture, an SGD optimizer with a batch size of 64, and a truncation length for back-propagation through time of 15 days for the training procedure. It it worth noting that the neural network presented in Fig. 2 does *not* use static information about the patients (this was a configurable hyper-parameter choice). That is among all the configurations trained, the best performing model was one that did not make use of the static information. This is, further discussed in Sect. 6.

We removed from these twenty models a few that did not perform well on the training or validation set (three of them). Due to the stochasticity in the training process, some trained networks can under-perform. We can legitimately filter them out, as long as we do so on validation or training sets.[2] Then, we computed their final score on the test set. The results are given in Fig. 3. The box plot reports a test concordance index of $58.8 \pm 4.6\%$. We performed a Student T-test against the value of 50% and rejected with p-value 3.7×10^{-7} that our model is equivalent to a random one.

We compares against two baselines. The first ones are the manual alerts set by the care workers. We have a history of four levels of urgency (none, low, medium,

[2]The only difference is that the training procedure now includes a filtering phase.

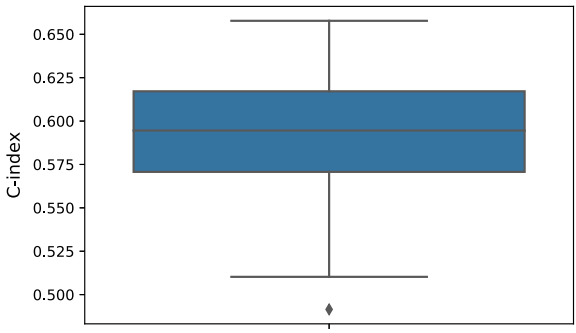


Fig. 3 A box plot of the test error of the most promising set hyper-parameters, over seventeen runs (three dropped due to under-performance on training or validation set)

high) for every patient and every day. As done with the latent risk modeled by the neural network, these alerts yield a natural ranking that can be used to compute a concordance index. These alerts achieve concordance indices of 48.7%, 50.7%, and 51.1% on respectively the training, the validation, and the test set. Note that these alerts were not set based on training data, so these results could be aggregated. However, we provide them separated so that the test error could be compared to the neural network on the same data points. The second baseline is a linear survival regression model, where data points are the vital signs for every patient and every day, as if they were independent (no time dependencies are taken into account). This model achieves a training error (on training and validation sets combined) of 49.7% and a test error of 48.0%. Even if these two baselines are simplistic, the fact that they do not achieve better than a random draw shows the difficulty of the problem.

We implemented the neural network in PyTorch [28], used Lifelines [8] to compute the linear baseline and the C-index, made use of Numpy [37], Scipy [19], Pandas [25], Scikit-Learn [29], IPython [30], and Jupyter [23] for pre-processing of the data and post-processing of the results, and rendered the figures with Matplotlib [16] and Seaborn [40].

6 Discussion

The high stochasticity of the problem makes training hard and long, therefore making comparison between different neural network architectures and training procedures either expensive (through averaging) or unfair (some configuration randomly performing abnormally well or poorly). As stated in the previous section, passing the patient static information as the first hidden vector of the recurrent neural network (as opposed to just passing a null vector), as proposed in Sect. 4, did not improve the performances and was therefore not selected through hyper-parameter search. Further inquiring should be done to find out if a better model could be obtained using the static patient information. Our hypothesis is that this data does have predictive power for this task, but that the specific part of the neural network responsible for it failed

to learn, due to optimization hardships. Indeed, not only this part of the model adds more parameters (layers) to train, these parameters are also updated less frequently in the truncated variant of back-propagation through time. Potential directions could be to focus more on the training of this part of the model (for instance through pertaining), or to include this data at every time-step (explicitly or through an attention mechanism).

Our results do not show that a linear model could not perform as well as the neural network, as less effort was given to this model. Improving the linear baseline would however mean additional engineering of the data to include time dependencies, patient static information, and perform feature selection.

These results suggest that combining, even weak, signals from remote monitoring in the homecare context can outperform simple manual baselines which open the door to better models.

Nonetheless, a self-fulfilling prophecy problem could occur with a better prediction accuracy. Care workers using machine learning generated alerts would prevent events from happening and reduce the observations labeled as events. A potential alternative is to ask care workers if the prediction was useful or not, i.e., if they want such prediction happening again in the future. While far from perfect, this methodology has the advantage of enabling model retraining as the data is gathered. More research is required to evaluate the risk of this problem and performance of the proposed alternative.

In addition, further research is needed to better understand the factors that contribute to higher risk days for home telemonitoring patients. Indeed, the black-box nature of neural networks makes them difficult to implement in the health care industry since physicians and other care workers generally want to understand why an adverse event probability is predicted. For example, what action should a care worker take if manual alerts are triggered but neural networks say that nothing is happening? The model performance suggests that no action should be taken, but this is clearly a difficult call.

References

1. Avati, A., Jung, K., Harman, S., Downing, L., Ng, A., Shah, N.H.: Improving palliative care with deep learning. BMC Med. Inform. Decis. Mak. **18**(4), 122 (2018)
2. Baxt, W.G., Shofer, F.S., Sites, F.D., Hollander, J.E.: A neural network aid for the early diagnosis of cardiac ischemia in patients presenting to the emergency department with chest pain. Ann. Emerg. Med. **40**(6), 575–583 (2002)
3. Berwick, D.M., Nolan, T.W., Whittington, J.: The triple aim: care, health, and cost. Health Aff. **27**(3), 759–769 (2008)
4. Che, Z., Purushotham, S., Cho, K., Sontag, D., Liu, Y.: Recurrent neural networks for multivariate time series with missing values. Sci. Rep. **8**(1), 6085 (2018)
5. Cho, K., van Merrienboer, B., Gulcehre, C., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Conference on Empirical Methods in Natural Language Processing (EMNLP 2014) (2014)

6. Choi, Y., Chiu, C.Y.-I., Sontag, D.: Learning low-dimensional representations of medical concepts. AMIA Summ. Transl. Sci. Proc. **2016**, 41–50 (2016)
7. Cox, D.R.: Regression models and life-tables. J. R. Stat. Soc. Ser. B **34**, 187–220 (1972)
8. Davidson-Pilon, C., Kalderstam, J., Zivich, P., Kuhn, B., Fiore-Gartland, A., Moneda, L., Gabriel, WIlson, D., Parij, A., Stark, K., Anton, S., Besson, L., Jona, Gadgil, H., Golland, D., Hussey, S., Noorbakhsh, J., Klintberg, A., Jordan, J., Rose, J., Slavitt, I., Martin, E., Ochoa, E., Albrecht, D., dhuynh, Zgonjanin, D., Chen, D., Fournier, C., Arturo, Rendeiro, A.F.: Camdavidsonpilon/lifelines: v0.19.2 (2019)
9. Esteban, C., Staeck, O., Baier, S., Yang, Y., Tresp, V.: Predicting clinical events by combining static and dynamic information using recurrent neural networks. In: 2016 IEEE International Conference on Healthcare Informatics (ICHI), pp. 93–101 (2016)
10. Gavrilov, L., Heuveline, P.: Aging of population. Technical report (2003)
11. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016)
12. Hansen, L.O., Young, R.S., Hinami, K., Leung, A., Williams, M.V.: Interventions to reduce 30-day rehospitalization: a systematic review. Ann. Intern. Med. **155**(8), 520–528 (2011)
13. Harrell, F.E.J., Califf, R.M., Pryor, D.B., Lee, K.L., Rosati, R.A.: Evaluating the yield of medical tests. JAMA **247**(18), 2543–2546 (1982)
14. Harrison, R.F., Kennedy, R.L.: Artificial neural network models for prediction of acute coronary syndromes using clinical data from the time of presentation. Ann. Emerg. Med. **46**(5), 431–439 (2005)
15. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**, 1735–80 (1997)
16. Hunter, J.D.: Matplotlib: a 2D graphics environment. Comput. Sci. Eng. **9**(3), 90–95 (2007)
17. Huntington, W.V., Covington, L.A., Center, P.P., Manchikanti, L.: Patient protection and affordable care act of 2010: reforming the health care reform for the new decade. Pain Phys. **14**(1), E35–E67 (2011)
18. Jencks, S.F., Williams, M.V., Coleman, E.A.: Rehospitalizations among patients in the medicare fee-for-service program. N. Engl. J. Med. **360**(14), 1418–1428 (2009)
19. Jones, E., Oliphant, T., Peterson, P., et al. SciPy: open source scientific tools for python [Online; accessed <today>] (2001)
20. Kansagara, D., Englander, H., Salanitro, A., Kagen, D., Theobald, C., Freeman, M., Kripalani, S.: Risk prediction models for hospital readmission: a systematic review. JAMA **306**(15), 1688–1698 (2011)
21. Katzman, J.L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., Kluger, Y.: Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. BMC Med. Res. Methodol. **18**(1), 24 (2018)
22. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint. arXiv:1412.6980 (2014)
23. Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., Willing, C.:. Jupyter notebooks—a publishing format for reproducible computational workflows. In: Loizides, F., Schmidt, B. (eds.) Positioning and Power in Academic Publishing: Players, Agents and Agendas, pp. 87–90. IOS Press (2016)
24. Luck, M., Sylvain, T., Cardinal, H., Lodi, A., Bengio, Y.: Deep learning for patient-specific kidney graft survival analysis. arXiv (1705.1024) (2017)
25. McKinney, W.: Data structures for statistical computing in python. In: van der Walt, S., Millman, J. (eds.) Proceedings of the 9th Python in Science Conference, pp. 51–56 (2010)
26. Organization, W.H.: International Statistical Classification of Diseases and Related Health Problems, vol. 1. World Health Organization (2004)
27. Paré, G., Jaana, M., Sicotte, C.: Systematic review of home telemonitoring for chronic diseases: the evidence base. J. Am. Med. Inform. Assoc. **14**(3), 269–77 (2007)
28. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
29. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**(Oct), 2825–2830 (2011)

30. Perez, F., Granger, B.E.: Ipython: a system for interactive scientific computing. Comput. Sci. Eng. **9**(3), 21–29 (2007)
31. Price, R.K., Spitznagel, E.L., Downey, T.J., Meyer, D.J., Risk, N.K., El-Ghazzawy, O.G.: Applying artificial neural network models to clinical decision making. Psychol. Assess. **12**(1), 40 (2000)
32. Rajkomar, A., Oren, E., Chen, K., Dai, A.M., Hajaj, N., Hardt, M., Liu, P.J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G.E., Irvine, J., Le, Q., Litsch, K., Mossin, A., Tansuwan, J., Wang, D., Wexler, J., Wilson, J., Ludwig, D., Volchenboum, S.L., Chou, K., Pearson, M., Madabushi, S., Shah, N.H., Butte, A.J., Howell, M.D., Cui, C., Corrado, G.S., Dean, J.: Scalable and accurate deep learning with electronic health records. NPJ Digit. Med. **1**(1), 18 (2018)
33. 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**(3), 598–610 (2012)
34. Ross, J.S., Mulvey, G.K., Stauffer, B., Patlolla, V., Bernheim, S.M., Keenan, P.S., Krumholz, H.M.: Statistical models and patient predictors of readmission for heart failure: a systematic review. Arch. Intern. Med. **168**(13), 1371–1386 (2008)
35. Suh, M.K.K., Evangelista, L.S., Chen, C.-A.A., Han, K., Kang, J., Tu, M.K., Chen, V., Nahapetian, A., Sarrafzadeh, M.: An automated vital sign monitoring system for congestive heart failure patients. In: Proceedings of the 1st ACM International Health Informatics Symposium, pp. 108–117. ACM (2010)
36. Swiercz, M., Mariak, Z., Lewko, J., Chojnacki, K., Kozlowski, A., Piekarski, P.: Neural network technique for detecting emergency states in neurosurgical patients. Med. Biol. Eng. Compu. **36**(6), 717–722 (1998)
37. Van Der Walt, S., Colbert, S.C., Varoquaux, G.: The numpy array: a structure for efficient numerical computation. Comput. Sci. Eng. **13**(2), 22 (2011)
38. Wallace, E., Stuart, E., Vaughan, N., Bennett, K., Fahey, T., Smith, S.M.: Risk prediction models to predict emergency hospital admission in community-dwelling adults: a systematic review. Med. Care **52**(8), 751 (2014)
39. Ward, B.W., Schiller, J.S., Goodman, R.A.: Peer reviewed: multiple chronic conditions among us adults: a 2012 update. In: Preventing Chronic Disease, vol. 11 (2014)
40. Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Ostblom, J., Lukauskas, S., Gemperline, D.C., Augspurger, T., Halchenko, Y., Cole, J.B., Warmenhoven, J., de Ruiter, J., Pye, C., Hoyer, S., Vanderplas, J., Villalba, S., Kunter, G., Quintero, E., Bachant, P., Martin, M., Meyer, K., Miles, A., Ram, Y., Brunner, T., Yarkoni, T., Williams, M.L., Evans, C., Fitzgerald, C., Brian, Qalieh, A.: mwaskom/seaborn: v0.9.0, July 2018 (2018)
41. Williams, R.J., Peng, J.: An efficient gradient-based algorithm for on-line training of recurrent network trajectories. Neural Comput. **2**(4), 490–501 (1990)
42. Zhu, M., Cheng, L., Armstrong, J.J., Poss, J.W., Hirdes, J.P., Stolee, P.: Machine Learning in Healthcare Informatics. Springer, Berlin (2014)

Mass Casualty Events: A Decision Making Tool for Home Health Care to Discharge Conventional Hospitals

Alain Guinet and Eric Dubost

Abstract We need to admit in a Home Health Care structure a massive influx of patients requiring an early discharge from conventional hospitals, due to a terrorist attack. Such situation requires to free hospitalization beds at the earliest for the victims. The early discharge patients are transferable to the Home Health Care (HHC) structure from a given release date and until a due date in order to reach the patient's home and find the caregiver. The home Health care structure must plan the patient admissions with the objective to admit as soon as possible the most victims in conventional hospitals using the least amount of HHC human resources during the discharges of hospitalized patients. An admission-planning model is proposed, the bi-objective problem modelled is solved with CPLEX.

Keywords Admission planning · Home Health Care · RCPSP · Bi-objective mixed linear model

1 Introduction

In the framework of the project PrHoDom (Protection of Home Health Care structures) we are working with the 3rd biggest HHC structure in France i.e. the hospital centre "Soins et Santé", in order to develop decision making tools to support the different processes of its crisis management plans. In this paper, we focus on the response to a terrorist risk after a bombing attack in the framework of a collaborative emergency management plan with conventional hospitals. Contexts are first presented and secondly the HHC contribution is introduced to face such situations.

A. Guinet (✉)
Institut National Des Sciences Appliquées de Lyon, Université de Lyon, DISP, 21 Av. Jean Capelle, 69621 Villeurbanne, France
e-mail: alain.guinet@insa-lyon.fr

E. Dubost
Centre Hospitalier Soins Et Santé, 325 Rue Maryse Bastié, 69141 Rillieux-la-Pape, France

V. Bélanger et al. (eds.), *Health Care Systems Engineering*,
Springer Proceedings in Mathematics & Statistics 316,
https://doi.org/10.1007/978-3-030-39694-7_17

1.1 Mass Casualty Contexts: A Large Number of Hospitalizations

On 11 September 2001, four coordinated terrorist attacks tacked place in the east cost of the United States. The attacks killed 2996 persons and injured over 6000 people. Four planes were hijacked by terrorists who crashed two planes on the south and north towers of the World Trade Centre in New-York. Saint Vincent's Hospital was the closest trauma hospital of the World Trade Centre. It received 844 patients over the 3 days following the terrorist attacks [4].

On 11 March 2004, ten terrorist bombs exploded in trains in Madrid killing 177 people and injuring 2062 others. From injured people, 312 patients were treated in hospitals and 91 persons were hospitalized [15].

On 13 November 2015, four bombing attacks tacked place in Paris at four different sites. These assaults wounded more than 300 people and killed 129 others on sites [9]. 256 wounded people were safely transferred and treated in APHP (Assistance Publique des Hôpitaux de Paris) hospitals, 44 others arrived at hospitals by their own means.

On 22 March 2016, two bombing attacks tacked place in Brussels at two different sites. The terrorist attacks killed at least 35 people including the three suicide bombers and injured 340 people [11].

On 14 July 2016, a truck crashed into the crowd in Nice during the French national celebration day. This terrorist attack wounded more than 400 people and killed 86 others [2].

1.2 The Home Health Care Contribution to Mass Casualties

Home Health Care (HHC) refers to a health structure that provides care at home for patients requiring complex postoperative treatments, or patients suffering from a chronic illness, a disability, or patients needing palliative care. In France, a HHC structure has the same rights and duties than a conventional hospital, for example, the requirement to establish emergency management plans in collaboration with other hospitals if possible. As long as the collaboration between HHCs and conventional hospitals is concerned, HHCs might provide restorative care while the conventional hospitals deliver mainly acute cares. Such collaboration could offer many mutual benefits. Home rehabilitation shorters expensive care length of stay of the patients, improves the patient physical health and family reintegration. Another improvement resulting from the combination of HHC and the conventional hospitalization is the emergency department freeing, HHC discharging directly the emergency system without requiring hospitalization beds [5]. Cooperation between Home Health Care structures and Conventional Hospitals can be benefit to face crisis with efficiency, by sharing acute patients and rehabilitation patients. Considering the studied scenario, we retain the hypothesis of an approved emergency management plan between a

conventional hospital and a home health care structure. The objective of such an emergency plan is to improve the efficiency of resources (beds and physicians) by orienting victims to the conventional hospital and the early hospitalization discharges generated to the home health care structure.

2 The Investigated Problem

2.1 The Massive Admission Scenario

A terrorist attack occurs in St Exupery Airport in Lyon. Around 8:00 AM, a bomb explodes in the main terminal killing 14 people and injuring more than 100 others, similar to the Zaventem airport attack on 2016. The ORSEC NOVI plan in St Exupery Airport is activated. The ORSEC NOVI plan is a French emergency management plan used for a mass casualty incident in a limited area. Around 8:15 AM, the HCL (Hospices Civils de Lyon) is alerted for an eventual massive influx of injuries and it activates its external emergency management plan. The HCL tries to empty the Emergency Departments from the regular patients, and to free the hospitalization beds with early discharges to the Home Health Care structure "Soins et Santé". The HHC structure activates its external emergency plan to admit a large number of early hospitalization discharges from HCL. A patient admissions planning is required.

2.2 Previous Works on Patients' Admissions

Our review of the literature investigates the field of admission planning and the field of early discharge planning regarding conventional hospitals and Home Health Care structures. The discharge process must favour the continuing of care for the patient in order to preserve the quality of cares, to reduce readmissions to conventional hospital and to avoid patient psychological distress. Collaboration with other care providers (e.g. the HHC), patient and his family are essential [13]. In a usual context, the discharge planning resulting from the discharge process helps to reduce the length of stay for patients in the hospital and to promote adherence of treatment [10]. In a mass admission context, the discharge process must help the care coordinator to identify the early discharges freeing the hospitalization beds while transferring the patient in safe conditions to another health care structure [3]. Preplanning with other care providers must be prepared and Home Health Care is an appropriate candidate [8]. The HHCs and conventional hospitals require a good coordination between them in terms of drug administration (i.e. a common pharmaceutical booklet) and of nursing technical cares (i.e. care protocols) [1]. Less literature has been found on admission planning. Granja et al. [6] define a patient planning, a control process, and propose a simulated annealing algorithm to optimize the patient admission process. The

authors defined a task as a group of activities using the same resources in the same time range. Resources are characterized by their capacity, their availability and their ability to perform tasks. The patient pathway makes necessary to specify precedence relations between tasks. The patient waiting times and the total completion time were minimized respectively for quality of cares and resource efficiency. In a mass admission context, there is no admission planning for a conventional hospital and victim triage is used facing to patient arrivals in the emergency department [12].

2.3 The Patient Admission Planning Problem

We need to admit in a Home Health Care structure a massive influx of patients requiring an early discharge from conventional hospitals, due to a terrorist attack (Mass Casualty Incident) which requires freeing hospitalization beds at the earliest. The early discharge patients are transferable from a given release date and can be managed by the Home Health Care structure until a due date in order to reach on time the patient's home and find the caregiver. Both dates are specific to each patient and define time windows. Patients wait in their hospital beds for their transportation, so they immobilize the bed. The schedule of anticipated exits must be specified by an early discharge planning of the conventional hospital, which depends on the admission planning of the HHC. Both planning can be the same. The HHC admission activities are: the establishment of the care order by the HHC coordinating physician and the hospital physician; the assignment of the salaried/liberal nurse for the cares; the establishment of the medication order by the HHC pharmacist and the hospital pharmacist; the preparation of the delivery of medicines; the delivery of medicines and medical equipment (medical bed, syringe pump, mechanical ventilator …); the transport of the patient by ambulance; the patient's entrance to his home and the information of the caregiver (family member) with the patient by the nurse. Human resources associated with admission activities are limited to: coordinating physicians, head nurses, pharmacists, pharmacy technicians, deliverymen and paramedics. The objective is to admit as soon as possible the most victims in conventional hospitals, i.e. to free hospital beds at the earliest so that they can accommodate victims; using the least amount of HHC human resources during the transfer of hospitalized patients. This will result in a HHC admission planning. Our problem is close to the investigated problem by Granja et al. but with a bi-objective function i.e. planning the admissions at the earliest with the smallest HHC resource employment.

2.4 The Admission Problem Complexity

We have a set of N projects (patient admissions) to plan with due-dates (latest hospitalization dates) and release-dates (early hospitalization dates). Each patient admission requires (is composed of) M non-preemptive tasks linked with precedence constraints. Medical staffs (renewable resources) with limited availabilities (capacities) perform the admission tasks according to their skills, i.e. the task processing requires dedicated resources (physician, pharmacist, head-nurse, paramedic…). A resource constraint multi-project scheduling problem (RCPSP) [7] is then defined with two objectives to optimize. The first objective is to admit the patients at the earliest, i.e. to minimize the sum of flow-time with a complexity similar to a hybrid flow-shop [16], which is NP hard. The second objective is to minimize the number of employees used, with a complexity of a resource investment problem [14], which is also NP hard. In the next section, we model such problem according to a lexicographic approach to integrate both objectives. Therefore, we model an admission planning problem and a resource sizing problem. A planning problem is favoured rather than a scheduling problem for complexity reasons. We verify the resource capacities roughly. Another simplifying hypothesis is the approximated duration of tasks, because the HHC knows generally only the patient medical speciality before planning.

3 The Mixed Linear Programs for the Bi-objective Optimization

3.1 The Admission Planning Model

Data:

- T: number of periods (hours),
- N: number of admissions (patients), each of them are composed of M tasks,
- Dur (i,j): duration of task j for patient i in minutes,
- Dtot (i): early hospitalization date (hour) in the HHC structure for patient i,
- Dtar (i): latest hospitalization date (hour) in the HHC structure for patient i,
- Pred (j,k): kth predecessor of task j,
- Cap (j,t): number of available resources associated to task j for period t.

Variables:

- X (i,j,t): binary variable equal to 1 if task j of admission i ends on period t,
- Y (j,t): real variable equal to the deferred resource (stock) for task j on period t, for a resource employment overlapping two periods,
- Tach (i,j): real variable equal to the completion date of task j for patient i,
- Cre (j,t): number of resources associated to task j and used on period t,

– Dfin (i): completion date of patient admission i.

Objective function:

$$Minimize(Z1) = \sum_{i=1}^{N} (Dfin(i) - Dtot(i)) \tag{1}$$

We minimize the waiting times sum of the patients to be admitted at home i.e. the flow-time. The last task of a patient admission is the last completed task. The waiting time is the difference between this admission completion date and the early hospitalization date.

Constraints:

$$\sum_{t=1|t<Dtot(i)}^{T} X(i, j, t) = 0 \quad \forall i = 1, \ldots, N \quad \forall j = 1, \ldots, M \tag{2}$$

$$\sum_{t=1|t>\ Dtar(i)}^{T} X(i, j, t) = 0 \quad \forall i = 1, \ldots, N \quad \forall j = 1, \ldots, M \tag{3}$$

$$\sum_{t=1}^{T} X(i, j, t) = 1 \quad \forall i = 1, \ldots, N \quad \forall j = 1, \ldots, M \tag{4}$$

A task of a patient admission cannot be realized before the early hospitalization date of the patient. A task of a patient admission cannot be realized after the latest hospitalization date of the patient. Each task of a patient must be realized.

$$Tach(i, j) - Dur(i, j) \geq Tach(i, Pred(j, k)) \quad \forall i = 1, \ldots, N \quad \forall j, k = 1, \ldots, M | Pred(j, k) > 0 \tag{5}$$

We must respect the task precedence. The completion date of a task minus the task duration must be greater than the completion date of the precedent task.

$$\sum_{t=1}^{T} (\mathrm{X(i, j, t)*t*60}) \geq Tach(i, j) \quad \forall i = 1, \ldots, N \quad \forall j = 1, \ldots, M \tag{6}$$

$$\sum_{t=1}^{T} (\mathrm{X(i, 1, t)*t*60}) \leq Tach(i, 1) \quad \forall i = 1, \ldots, N \tag{6bis}$$

The task completion date of a patient must be linked with the binary variable which specifies the period where the task is achieved.

$$\sum_{i=1}^{N} (X(i, j, 1) * Dur(i, j)) = (Cre(j, 1) - Y(j, 1)) * 60 \quad \forall \mathrm{j} = 1, \ldots, \mathrm{M} \quad (7)$$

$$\sum_{i=1}^{N} (X(i, j, t) * Dur(i, j)) = (Cre(j, t) + Y(j, t-1) - Y(j, t)) * 60$$
$$\forall t = 2, \ldots, T \ \forall j = 1, \ldots, M \quad (7\text{Bis})$$

For each period, the sum of admission tasks must be equal to the resource capacity used in minutes plus or minus the differed resources in minutes. The differed resources are acting as a stock to respect the continuous employment of resources.

$$Y(j, t) \leq Cre(j, t) \quad \forall j = 1, \ldots, M \quad \forall \mathrm{t} = 1, \ldots, \mathrm{T} \quad (8)$$

The stock cannot exceed the amount of used resources of one period regarding to the resource employment overlapping two periods.

$$Cre(j, t) \leq Cap(j, t) \quad \forall j = 1, \ldots, M \quad \forall t = 1, \ldots, T \quad (9)$$

The resource capacity used is limited by the resource capacity available.

$$Dfin(i) \geq X(i, j, t) * t \quad \forall i = 1, \ldots, N \quad \forall j = 1, \ldots, M \quad \forall t = 1, \ldots, T \quad (10)$$

The completion date of patient admission i is equal to the period where the last task is completed. The completion date accuracy is given in hours.

3.2 The Resource Sizing Model

The mixed linear program to minimize the human resource utilization during the patient admissions (i.e. resource sizing), is similar than the mixed linear program for admission planning. It differs from the objective function which minimizes the maximums of the used resources. Data, variables and constraints are the same.

Second objective function Z2 for resource sizing replaces function Z1:

$$\text{Minimize (Z2)} = \sum_{\mathrm{j}=1}^{\mathrm{M}} (\mathrm{Rmax(j)}) \quad (11)$$

Added Constraints:

$$R\max(j) \geq Cre(j, t) \quad \forall t = 1, \ldots, T \quad \forall j = 1, \ldots, M \quad (12)$$

We calculate the maximum resource used over the horizon per task to minimize it.

3.3 The Lexicographic Model to Integrate Both Objectives

Two types of constraints are added in order on one hand to calculate the selected objective function and on the other hand, to keep the best solution found by the mixed linear program for the non selected objective function.

$$\sum_{i=1}^{N}(Dfin(i) - Dtot(i)) \leq Z1 \tag{13}$$

We keep the best solutions found for Z1 minimizing Z2.

$$\sum_{j=1}^{M}(R\max(j)) \leq Z2 \tag{14}$$

Equation 14 replaces Eq. 13. We keep the best solutions found for Z2 minimizing Z1.

4 Scenario Study

Table 1 defines the tasks, the task precedence, the task durations, the task resources and the number of resource exemplars. The early hospitalization dates in the HHC are set from 1 to 6 in equal proportion regarding the number of patient admissions. The latest early hospitalization dates in the HHC are set from 8 to 13 in equal proportion regarding the number of patient admissions. The horizon is set to 13 periods.

The early hospitalization date of an admission is always smaller than its latest hospitalization date. These parameters have been defined by both hospitals according to patient medical specialities. The problems are solved with the Cplex Solver

Table 1 Patient admission parameters

S. No.	Task	Predecessors	Task duration	Type of resource	Resource capacity
1	Care order		{10,15}	Physician	2
2	Nurse assignment	1	{25,35,45}	Head nurse	4
3	Medication order	1	{10}	Pharmacist	2
4	Drug preparation	2,3	{20}	Technician	3
5	Drug delivery	4	{40}	Delivery man	6
6	Transportation	4	{90}	Ambulance	10

(https://www-01.ibm.com/software/commerce/optimization/cplex-optimizer/) limited to one hour of computation time when the optimal solution is not proved. Optimal solutions are found in less than 10 min.

Table 2 presents: the results of the 4 different approaches, specifying: the problem solved, the number of admissions (N.), the mean patient waiting time (i.e. the objective function of the planning program Z1 divided by N), the sum of the maximums of resources used (Z2. i.e. the objective function of the sizing program) and the completion time of the last admission (Cmax). Six sizes of problems (N varying from 10 to 60) are studied for four problem configurations (i.e. minimizing Z1 only, Z2 only, Z1 before Z2, Z2 before Z1). The left side of Table 2 shows the results considering only one criteria. The right side of Table 2 presents the results of the lexicographic approach considering both criteria. Only solutions in bold have been proved to be optimal.

Minimizing first the objective function Z1 before Z2 (1 » 2) comparing to minimizing first the objective function Z2 before Z1 (2 » 1), improves the admission planning dividing the mean waiting time of patients by 1.5. The resources in '1 » 2' approach are used during a shorter time comparing to the '2 » 1' approach. The completion time of the last admission (Cmax) for the '1 » 2' approach is around 3 periods less than for the '2 » 1' approach. Most solutions have been proved to be optimal (numbers in bold). Non optimal solutions are higher of less than 1% of the solver lower-bound.

Minimizing first the objective function Z2 before Z1 (2 » 1), improves the resource sizing reducing the maximum numbers of resource required. The number of resources used is divided by 1.3–2.5 regarding to the problem size. Using '2 » 1', less resources are used but for a longer time (see Cmax).

Considering the two used bi-objective approaches, minimizing first the objective function Z1 before Z2 seems to us the most suitable approach. On one hand, the two approaches require either fewer resources for a longer time or more resources for less time and on the other hand, the mean waiting time of patients is 1.5 times shorter for the '1 » 2′ approach.

The two previous experiments show that an optimized mean waiting time is contradictory with a minimized amount of the maximums of resources used. The left side of Table 2 presents the results of the planning program with the initial parameters of Table 1 i.e. without optimizing the maximums of resources used (Z1 only) or without optimizing the mean waiting time (Z2 only). All the solution found are here optimal. Results for minimizing 'Z1 only' are 25–50% more costly than minimizing first the objective function Z1 before Z2 for the sum of the maximums of resources used. Results for minimizing 'Z2 only' are 15–20% more costly than minimizing first the objective function Z2 before Z1 for the mean patient waiting time. Our lexicographic approach seems justified and the '1 » 2′ approach is better.

Table 2 Patient admission results

Type	N	Z1	Z2	Cmax	Type	N	Z1	Z2	Cmax
1 only	10	**3.10**	21.75	5	1 » 2	10	**3.10**	**13.02**	5
1 only	20	**3.10**	26.67	6	1 » 2	20	**3.10**	**18.68**	6
1 only	30	**3.20**	26.67	7	1 » 2	30	**3.20**	**19.82**	7
1 only	40	**3.33**	26.83	8	1 » 2	40	**3.33**	**21.19**	8
1 only	50	**3.54**	27.00	10	1 » 2	50	**3.54**	**21.25**	10
1 only	60	**3.73**	26.83	11	1 » 2	60	3.73	21.85	11
2 only	10	5.80	**5.22**	8	2 » 1	10	**5.00**	**5.22**	8
2 only	20	5.75	**9.04**	9	2 » 1	20	**4.80**	**9.04**	9
2 only	30	5.80	**11.98**	10	2 » 1	30	**4.84**	**11.98**	10
2 only	40	5.68	**14.33**	11	2 » 1	40	4.83	14.33	11
2 only	50	5.70	**16.19**	12	2 » 1	50	4.82	16.19	12
2 only	60	5.65	**17.76**	13	2 » 1	60	4.82	17.76	13

5 Conclusion

We have modelled a mass admission planning problem for HHC structures optimizing two objective functions (patient waiting times and resource employments). Cplex solver seems suitable to solve our models. The decision tool will be implemented in the Hospital centre "Soins et Santé". Model parameters Table 1 are defined in an Excel sheet and results Tables 2 are written in another Excel sheet. The home care director knows very few information about the solver, just enough to lunch the calculus. Our tool can also be transferable to conventional hospitals.

We would like to thank the reviewers for their relevant and helpful comments.

References

1. Bangsbo, A., Reg, O.T., Duner, A., Dahlin-Ivanoff, S., Liden, E.: Collaboration in discharge planning in relation to an implicit framework. Appl. Nurs. Res. **36**, 57–62 (2017)
2. Carles, M., Levraut, J., Gonzalez, J.F., et al.: Mass casualty events and health organization: terrorist attack in Nice. Lancet **388**, 2349–2350 (2016)
3. Davis, D.P., Poste, J.C., Hicks, T., Polk, D., Rymer, T.E., Jacoby, I.: Hospital bed surge capacity in the event of a mass-casualty incident. Prehospital Disaster Med. **20**, 169–176 (2005)
4. Feeney, J.M., Goldberg, R., Blumenthal, J.A., Wallack, M.K.: September 11, 2001, revisited: a review of the data. Arch. Surg. **140**, 1068–1073 (2005)
5. Frick, K.D., Burton, L.C., Clark, R., Mader, S.I., Naughton, B., Burl, J.B., Greenough, W.B., Steinwachs, D.M., Leff, B.: Substitutive hospital at home for older persons: effects on costs. Am. J. Manag. Care **15**, 49–56 (2009)
6. Granja, C., Almada-Lobo, B., Janela, F., Seabra, J., Mendes, A.: An optimization based on simulation approach to the patient admission scheduling problem using a linear programing algorithm. J. Biomed. Inform. **52**, 427–437 (2014)
7. Habibi, F., Barzinpour, F., Sadjadi, S.J.: Resource-constrained project scheduling problem: review of past and recent developments. J. Proj. Manag. **3**, 55–88 (2018)
8. Hick, J.L., Hanfling, D., Burstein, J.L., DeAtley, C., Barbisch, D., Bogdan, G.M., Cantrill, S.: Health care facility and community strategies for patient care surge capacity. Ann. Emerg. Med. **44**, 253–261 (2004)
9. Hirsch, M., Carli, P., Nizard, R., Riou, B., Baroudjian, B., Baubet, T., Chhor, V., Chollet-Xemard, C., Dantchev, N., Fleury, N., Fontaine, J.P., Yordanov, Y., Raphael, M., Paugam, B.C., Lafont, A.: The medical response to multisite terrorist attacks in Paris. Lancet **386**, 2535–2538 (2015)
10. Huber, D.L., Mcclelland, E.: Patient preferences and discharge planning transitions. J. Prof. Nurs. **19**, 204–210 (2003)
11. Keren, D.: The Brussels attacks—22/03/2016 what do we know? & insights from ICT experts. International Institute for Counter Terrorism (ICT), Special report, pp. 1–19, Mar 22 (2016)
12. Lax, P., Prior, K.: Major incident pre-hospital care. Surgery **36**, 402–408 (2018)
13. Lin, C.J., Cheng, S.J., Shih, S.C., Chu, C.H., Tjung, J.T.: Discharge planning. Int. J. Gerontol. **6**, 237–240 (2012)
14. Newmann, K., Zimmermann, J.: Procedures for resource levelling and net present value problems in project scheduling with general temporal and resource constraints. Eur. J. Oper. Res. **127**, 425–443 (2000)
15. Peral-Gutierrez de Ceballos, J., Turégano-Fuentes, F., Pérez-Díaz, D., Sanz-Sánchez, M., Martín-Llorente, C., Guerrero-Sanz, J.E.: 11 March 2004: the terrorist bomb explosions in

Madrid, Spain—an analysis of the logistics, injuries sustained and clinical management of casualties treated at the closest hospital. Crit. Care **9**, 104–111 (2005)
16. Vob, S., Witt, A.: Hybrid flow shop scheduling as a multi-mode multi project scheduling problem with batching requirements: a real world application. Int. J. Prod. Econ. **105**, 445–458 (2007)

Radiology, Radiotherapy and Chemotherapy

Simultaneous Optimization of Appointment Grid and Technologist Scheduling in a Radiology Center

Dina Bentayeb, Nadia Lahrichi and Louis-Martin Rousseau

Abstract The objective of this paper is to simultaneously optimize the appointment grid and the technologist scheduling. We develop an integer programming model by integrating the constraints of appointments and technologist schedules. We evaluate the optimization model using a real case of the Magnetic Resonance Imaging in the CHUM radiology department. The proposed approach provides a decision tool for outpatient centers, and improves resource utilization as well as patient access to the service.

Keywords Appointment grid · Staff scheduling · Radiology

1 Introduction

Radiology departments in hospitals contain expensive resources such as Computed Tomography scans (CT-scan) and Magnetic Resonance Imaging (MRI). The management of these services is generally based on appointment scheduling systems. The objective is to match supply and demand. Due to the demand heterogeneity and limited resources, the match is not usually optimal, which generates either server idle time or overtime, machine under-utilization, and patient waiting time. This gives rise to high costs of care services and patient dissatisfaction.

Appointment and staff scheduling for diagnosis resources has been widely studied. Medical staff scheduling is a relevant problem in healthcare. Researchers are proposing an efficient planning by considering different aspects of scheduling issues, such as: preferences, fairness, breaks, etc. We refer the reader to [1, 2] for an extensive review. Considering how this applies to radiology, Chen et al. [3] propose a method for the allocation and scheduling of radiological technologists. Yuura et al. [4] present a scheduling model for radiographers by integrating their skills and

D. Bentayeb · N. Lahrichi (✉) · L.-M. Rousseau
Department of Mathematical and Industrial Engineering, CIRRELT and Polytechnique Montreal, Succursale Centre-ville, C.P. 6079, Montreal, QC H3C 3A7, Canada
e-mail: nadia.lahrichi@polymtl.ca

V. Bélanger et al. (eds.), *Health Care Systems Engineering*,
Springer Proceedings in Mathematics & Statistics 316,
https://doi.org/10.1007/978-3-030-39694-7_18

training. He [5] is conducting a tabu search for radiological resource planning during the examination process.

To present realistic appointment scheduling, the majority of researchers consider patients with multiple priority and different service duration. They have applied several methods to improve access to service and to reduce patient wait time: simple scheduling policies [6], optimization and simulation [7], Markov decision process [8, 9], etc. Cappanera et al. [10] treat the patient scheduling for magnetic resonance imaging, taking into account the number of allocated radiologists. They present two approaches: offline and online. They take into consideration examination overlap and radiologist cross-training.

Two nuclear medicine studies combine patient and resource scheduling. Perez et al. [11] propose two algorithms: in the first, the task resource assignment is fixed; in the second, they carry out the assignment by solving an integer programming model, specifically for days with high demand. Perez et al. [12] develop stochastic online planning, starting with the offline version. In the existing literature, the studies focus only on the patient scheduling, or determine simultaneously the assignment of resources to tasks at the operational level. However, the authors do not consider the problem of the staff scheduling in parallel.

In outpatient healthcare centers based on appointment systems, patient scheduling is more efficient as a result of the timetable redesign that adjusts the service time [13] or the appointment type classification [14]. To the best of our knowledge, the studies in the literature don't address the grid design when doing staff scheduling. However, the appointment types and their number depend on the availability of appropriate resources, as well as on the historical demand. Therefore, scheduling patient appointments separately leads to inefficient resource allocation.

Integrated tactical planning allows dynamic and flexible resource management and is adapted to the variability and the heterogeneity of demand: in the case of high demand, managers can use the maximum capacity of resources, including overtime, in order to furnish more time slots in the grid; and in the case of low demand, they can plan technologist training or accept more holiday requests. In this article, we optimize the patient appointment grid along with the technologists scheduling. We provide an optimal allocation of personal resources to maximize the machine utilization, and the number of patients seen per day. The proposed approach is applied to real data from the Magnetic Resonance Imaging (MRI) in the CHUM radiology center. It combines, simultaneously, the scheduling of appointments and technologists at the tactical level while taking demand into account. For each machine, we decide the following: types, number, and sequence of radiology tests. We also decide the following for each technologist: their shift, room assignment, days off, and breaks. In the next section, we discuss the problem context and our case study. In Sect. 3, we present the associated scheduling model. We analyze the experiments and results in Sect. 4 and finally, conclude in Sect. 5.

2 Problem Statement

The technologist scheduling is performed manually without taking into account the exam planning and other constraints of the hospital, such as the demand variation. In this study, we implement a decision tool that allows the interaction between the human resources and patient scheduling.

In the CHUM radiology department, MRI is divided into the following categories: neuroradiology, abdomen, musculoskeletal, cardiac, breast and vascular. Figure 1 shows that the neuroradiology and the abdomen MRI represent more than 50% of the performed exams; 20% of the capacity is reserved to emergency or research exams, which can be included in any category. The center contains six machines of

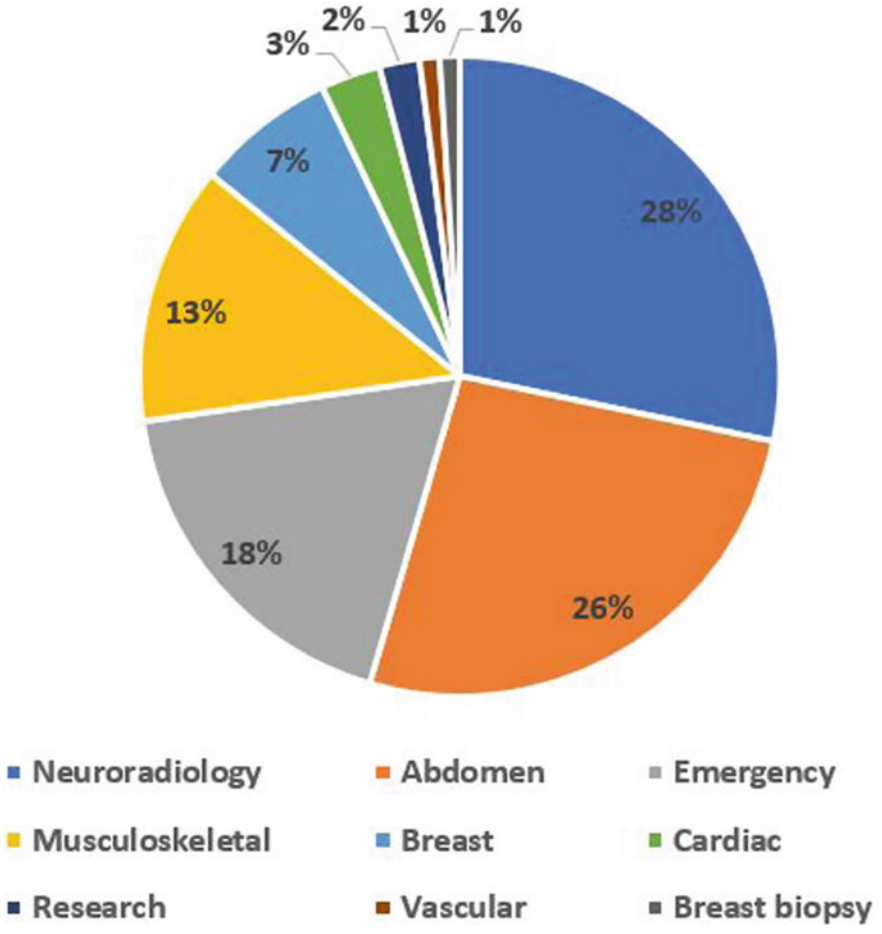


Fig. 1 MRI Categories distribution in the CHUM radiology center

three different types, so assigning an exam to a machine depends on its category. In fact, breast and cardiac exams are performed only on machine A. Neuroradiology is performed on any machine, except machine A. Abdomen MRI is performed on machines C, E and F. The execution of the MRI for a given patient requires the availability of a minimum number of technologists, depending on the exam category.

Twenty four technologists are allocated to this radiology center, they are divided into part-time and full-time.16 technologists work only the morning shift, and, 8 technologists work only the evening shift. We also take into account cross-training, meaning the technologist may treat all exam categories. In our case study, the scheduling period is 28 days. We define for each technologist the assigned machine, the workdays, and the planning. The planning represents the daily schedule: it determines the start and end times of the shift, the dedicated slots for work, and the break times. The planning classification corresponds to the type of shift that the technologist can work.

3 IP Model for Appointment and Technologist Scheduling

In this section, we present an integer programming (IP) model to schedule radiological technologists simultaneously with the grid appointment. We define the mathematical model using the following sets, parameters and variables.

$Sets$:	
H	the set of shifts
P	the set of plannings
P_h	the set of plannings of shift h
T	the set of technologists
T_k	the set of technologists that can't work more than k days per 14 days
M	the set of machines
C	the set of exam categories
D	the set of slot lengths
S	the set of slots
S_d	the set of slots of length d
J	the set of days
J_1	the set of the first half of days in the planning horizon
J_2	the set of the second half of days in the planning horizon
$J_W \subset J$	the set of weekend days
$J_{St} \subset J$	the set of Saturdays
$J_{Sn} \subset J$	the set of Sundays
$W_i \subset J$	the set of days from Monday to Thursday of the week i

Parameters :

a_{tp}	binary parameter, equal to 1 if it is possible to assign planning p to technologist t
b_{cm}	binary parameter, equal to 1 if it is possible to assign category c to machine m
e_{ps}	binary parameter, equal to 1 if planning p covers slot s
n_c	the total minimum number of each category c
nb_c	the minimal required number of technologists to execute the exam category c
ns_h	the minimal number of active machines in the shift h

Variables :

x_{cs}^{mj}	binary variable, equal to 1 if category c is performed on machine m, slot s and day j
y_{pt}^{mj}	binary variable, equal to 1 if planning p is assigned to technologist t and machine m on day j
off_{tj}	binary variable, equal to 1 if technologist t is off on day j
δ_s^{mj}	binary variable, equal to 1 if a change of category is done on slot s day j, machine m

The present optimization model considers two types of constraints: hard and soft. The first type represents the constraints that we must respect; they include the covering (3), the feasibility (6) and the hospital regulation (10). However, if we don't regard the soft constraint (5), we will have a feasible solution, but with low quality.

The objective function is defined in Eq. (1). The first term maximizes the machine utilization by filling the maximum number of available slots. The second term penalizes the change of category on the same machine during a day. The last term leads to minimizing the number of technologists working on the weekend.

$$\text{Max}\lambda = \sum_{c\in C}\sum_{s\in S}\sum_{m\in M}\sum_{j\in J} x_{cs}^{mj} - \sum_{s\in S}\sum_{m\in M}\sum_{j\in J}\delta_s^{mj} + \sum_{t\in T}\sum_{j\in J_W} off_{tj} \tag{1}$$

We consider the following constraints.

We may attribute at most one exam category to a slot on a given machine and day:

$$\sum_{c\in C} x_{cs}^{mj} \leq 1, \quad \forall s \in S,\ m \in M,\ j \in J \tag{2}$$

The execution of each category requires a minimum number of technologists:

$$\sum_{p\in P}\sum_{t\in T} e_{ps}\, y_{pt}^{mj} \geq nb_c x_{cs}^{mj}, \quad \forall\, c \in C,\ s \in S,\ m \in M,\ j \in J \tag{3}$$

The total number of each category throughout the planning horizon has to be at least equal to n_c:

$$\sum_{d\in D} \sum_{s\in S_d} \sum_{m\in M} \sum_{j\in J} d * x_{cs}^{mj} \geq n_c, \quad \forall\, c \in C \tag{4}$$

Penalize the change of category from one slot to the next one on the same machine during a day:

$$x_{cs}^{mj} + x_{c'(s+1)}^{mj} \geq \delta_{s+1}^{mj} + 1, \quad \forall\, c \in C,\ c' \in C,\ c \neq c',\ s \in S,\ m \in M,\ j \in J \tag{5}$$

An exam category can only be assigned to the appropriate machine:

$$x_{cs}^{mj}(1 - b_{cm}) = 0, \quad \forall\, c \in C,\ s \in S,\ m \in M,\ j \in J \tag{6}$$

A technologist can work, at most, on one machine according to one planning per day:

$$\sum_{p\in P}\sum_{m\in M} y_{pt}^{mj} \leq 1, \quad \forall\, t \in T,\ j \in J \tag{7}$$

A technologist is assigned only to the appropriate planning:

$$y_{pt}^{mj}(1 - a_{tp}) = 0, \quad \forall\, p \in P,\ t \in T,\ m \in M,\ j \in J \tag{8}$$

If a technologist does not work one day; then, it's his day off:

$$1 - \sum_{p\in P}\sum_{m\in M} y_{pt}^{mj} = off_{tj}, \quad \forall\, t \in T,\ j \in J \tag{9}$$

The technologist can't work more than five consecutive days:

$$\sum_{p\in P}\sum_{m\in M}\sum_{j'=j}^{j+5} y_{pt}^{mj'} \leq 5, \quad \forall\, t \in T,\ j \in \{1, .., max(J) - 5\} \tag{10}$$

During a week, if the technologist works a day shift, he will have the same assignment for the next day or he will be off:

$$y_{pt}^{mj} \leq y_{pt}^{m(j+1)} + off_{t(j+1)}, \quad \forall\, p \in P,\ t \in T,\ m \in M,\ j \in W_i,\ i \in \{1, 2, 3, 4\} \tag{11}$$

The number of active machines per shift during the weekend has to respect the minimum coverage of resources:

$$\sum_{p \in P_h} \sum_{t \in T} \sum_{m \in M} y_{pt}^{mj} \geq ns_h, \quad \forall\, j \in J_W,\ h \in H \tag{12}$$

If a technologist works on Saturday, he will work on Sunday with the same planning:

$$y_{pt}^{mj} = y_{pt}^{m(j+1)}, \quad \forall\, p \in P,\ t \in T,\ m \in M,\ j \in J_{St} \tag{13}$$

A technologist can't work two weekends consecutively:

$$\sum_{p \in P} \sum_{m \in M} y_{pt}^{mj} + \sum_{p \in P} \sum_{m \in M} y_{pt}^{m(j+6)} \leq 1, \quad \forall\, t \in T,\ j \in J_{Sn} \tag{14}$$

The technologist who belongs to T_k can't work more than k days on the half of the planning horizon:

$$\sum_{p \in P} \sum_{m \in M} \sum_{j \in J_1} y_{pt}^{mj} \leq K, \quad \forall\, t \in T_k \tag{15}$$

$$\sum_{p \in P} \sum_{m \in M} \sum_{j \in J_2} y_{pt}^{mj} \leq K, \quad \forall\, t \in T_k \tag{16}$$

4 Experiments and Results

To evaluate our IP model, we conduct computational experiments based on the case of the CHUM radiology department. We solve the model through the CPLEX Solver using a PC with a processor Intel Core i7 2.80 GHZ and 16 GB RAM.

We consider six machines, seven exam categories, and 24 technologists. All exams may be executed by one technologist, except breast biopsy exams which require more than one. During the weekend, at least two machines are active in the morning shift, and, one machine in the evening shift.

The planning defines working hours, breaks, start and end times of the shift for a technologist. The planning is constructed manually. We split the day into slots of one hour, and slots of a half hour, which are reserved for some breaks. We list all the possible planning based on the different hospital regulations. The morning shift can start between 7 am and 9 am, or at 12 pm. The start time of the evening shift is between 3 pm and 4 pm. The lunch break is taken between 11 am and 3 pm. However, the dinner break is between 6 pm and 8 pm. The technologists have three breaks during a working day: lunch break of one hour, dinner break of a half hour, and two breaks of 15 min that are taken at the end of the shift. Table 1 describes an example of planning. All the time slots are of one hour; though, the length of the last two slots is a half hour. The technologist starts the working day at 8 am, finishes at 3:30 pm, and has his lunch break at 12 pm and the two other breaks of 15 min at 3 pm.

Figures 2 and 3 illustrate respectively the scheduling of technologists and appointments on machine 1 and day 27. Four technologists are assigned to this machine. They are from four different shifts that start at 7 am, 9 am, 12 pm and 4 pm. From 12 pm to 6:30 pm, two or three technologists work on this machine; so, the breast biopsy exam can be executed at this time interval because its realization requires more than one technologist.

Table 2 presents the results. The machine utilization is the ratio between the total number of working hours on this resource and its maximum capacity. The computational time is calculated for the optimality gap of 1%.

We have good results. The proposed optimization model outperforms the current scheduling approach of the CHUM. The total number of category change over the

Table 1 Technologist planning example

slot (s_d)	7_{1h}	8_{1h}	9_{1h}	10_{1h}	11_{1h}	12_{1h}	1_{1h}	2_{1h}	$3_{0.5h}$	$3_{0.5h}$
e_{2s}	0	1	1	1	1	0	1	1	1	0

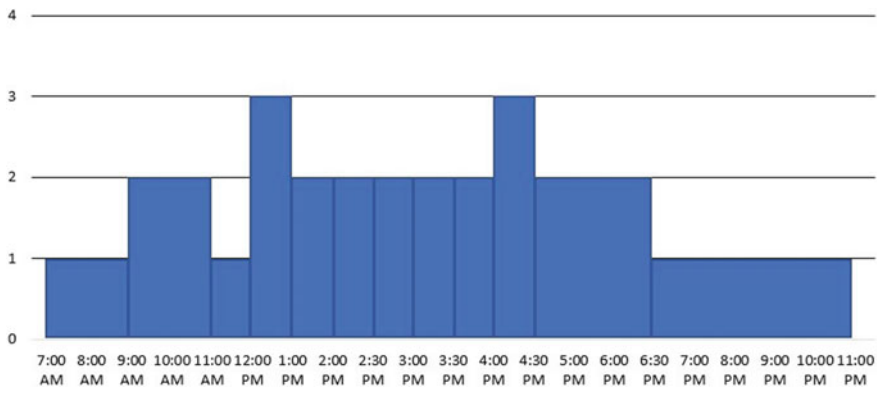


Fig. 2 Technologist allocation on machine 1, day 27

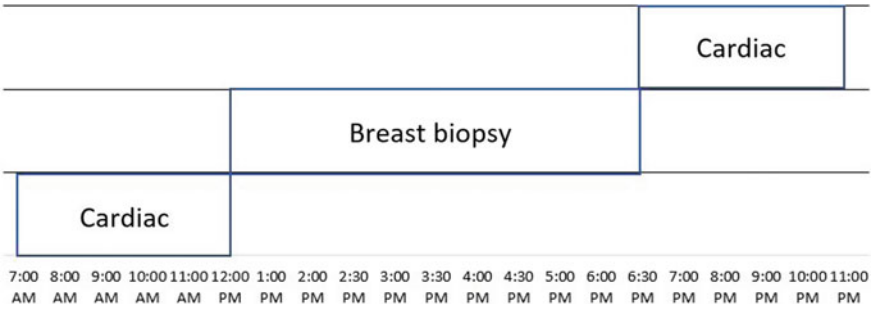


Fig. 3 MRI categories scheduling on machine 1, day 27

Table 2 Results of optimization model run

Scheduling method	Computational time	Number of category change	Machine utilization		Average number of technologists on the weekend
			Weekdays (%)	Weekends (%)	
Current case	–	136	62	20	3
IP model	16,009 s	80	93	74	11.6

scheduling period is reduced by 40%. We increase the number of allocated technologists on the weekend; therefore, the machine utilization reaches 74%. Moreover, we exploit 93% of the machine capacity over the weekdays, the gain compared to the real case is about 30%.

5 Conclusion

The scheduling of the technologist and the grid appointment is usually performed independently and manually without taking into account the demand. This study involves developing a decision tool that leads the scheduling of appointments and technologists at the same time in the tactical level. We present an Integer Programming model which is evaluated based on a real case of the Magnetic Resonance Imaging in the CHUM radiology department. The computational experiments indicate that our approach provides a considerable improvement in the machine utilization, that reaches 50% over the weekend and 30% over the weekdays.

References

1. Burke, E.K., De Causmaecker, P., Berghe, G.V., Van Landeghem, H.: The state of the art of nurse rostering. J. Sched. **7**, 441–499 (2004)
2. Erhard, M., Schoenfelder, J., Fügener, A., Brunner, J.O.: State of the art in physician scheduling. Eur. J. Oper. Res. **265**, 1–18 (2018)
3. Chen, P.-S., Lin, Y.-J., Peng, N.-C.: A two-stage method to determine the allocation and scheduling of medical staff in uncertain environments. Comput. Industr. Eng. **99**, 174–188 (2016)
4. Yuura, H., Miyamoto, T., Hidaka, K.: An integer programming model for radiographer scheduling considering skills and training. In: 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), pp. 889–893. IEEE, New York (2017)
5. He, C.-H.: Tabu search based resource allocation in radiological examination process execution. Front. Inf. Technol. Electr. Eng. **19**, 446–458 (2018)
6. Jiang, Y., Abouee-Mehrizi, H., Diao, Y.: Data-driven analytics to support scheduling of multi-priority multi-class patients with wait time targets. Euro. J. Oper. Res. (2018)
7. Patrick, J., Puterman, M.L.: Improving resource utilization for diagnostic services through flexible inpatient scheduling: A method for improving resource utilization. J. Oper. Res. Soc. **58**, 235–245 (2007)
8. Patrick, J., Puterman, M.L., Queyranne, M.: Dynamic multipriority patient scheduling for a diagnostic resource. Oper. Res. **56**, 1507–1525 (2008)
9. Kolisch, R., Sickinger, S.: Providing radiology health care services to stochastic demand of different customer classes. OR Spectrum **30**, 375–395 (2008)
10. Cappanera, P., Visintin, F., Banditori, C., Di Feo, D.: Evaluating the long-term effects of appointment scheduling policies in a magnetic resonance imaging setting. Flex. Serv. Manuf. J. **1–43** (2018)
11. Pérez, E., Ntaimo, L., Wilhelm, W.E., Bailey, C., McCormack, P.: Patient and resource scheduling of multi-step medical procedures in nuclear medicine. IIE Trans. Healthcare Syst. Eng. **1**, 168–184 (2011)
12. Pérez, E., Ntaimo, L., Malavé, C.O., Bailey, C., McCormack, P.: Stochastic online appointment scheduling of multi-step sequential procedures in nuclear medicine. Health Care Manage. Sci. **16**, 281–299 (2013)
13. Bentayeb, D., Lahrichi, N., Rousseau, L.-M.: Patient scheduling based on a service-time prediction model: a data-driven study for a radiotherapy center. Health Care Manage. Sci. **1–15** (2018)
14. Huang, Y., Verduzco, S.: Appointment template redesign in a womens health clinic using clinical constraints to improve service quality and efficiency. Appl. Clin. Inf. **6**, 271–287 (2015)

A Mathematical Programming Model for Radiotherapy Scheduling with Time Windows

Bruno Vieira, Derya Demirtas, Jeroen B. van de Kamer, Erwin W. Hans, Louis-Martin Rosseau, Nadia Lahrichi and Wim H. van Harten

Abstract In external-beam radiotherapy (RT), high-energy radiation beams are delivered by a linear accelerator in a series of irradiation sessions undertaken over multiple days. In this work, we consider the problem of scheduling and sequencing RT sessions considering time window preferences given by patients for the starting time of their appointments. Most studies in the literature focus on assigning patients to linacs and days, neglecting the sequencing component, and existing sequencing algorithms are only able to solve the problem using approximation methods due to the intractability of the formulated models. We propose a mixed-integer linear programming model and test it using data from a large Dutch RT center. Results show

B. Vieira (✉)
Center for Healthcare Operations Improvement and Research (CHOIR), University of Twente, Enschede, The Netherlands
e-mail: b.vieira@nki.nl

B. Vieira · J. B. van de Kamer
Department of Radiation Oncology, Netherlands Cancer Institute, Amsterdam, The Netherlands
e-mail: j.vd.kamer@nki.nl

D. Demirtas · E. W. Hans
Department of Industrial Engineering and Business Information Systems, Faculty of Behavioural Management and Social Sciences, Center for Healthcare Operations Improvement and Research (CHOIR), University of Twente, Enschede, The Netherlands
e-mail: d.demirtas@utwente.nl

E. W. Hans
e-mail: e.w.hans@utwente.nl

L.-M. Rosseau · N. Lahrichi
Mathematics and Industrial Engineering, Polytechnique Montreal, Montreal, Canada
e-mail: louis-martin.rousseau@polymtl.ca

N. Lahrichi
e-mail: nadia.lahrichi@polymtl.ca

W. H. van Harten
Rijnstate General Hospital, Arnhem, The Netherlands
e-mail: WvanHarten@Rijnstate.nl

V. Bélanger et al. (eds.), *Health Care Systems Engineering*,
Springer Proceedings in Mathematics & Statistics 316,
https://doi.org/10.1007/978-3-030-39694-7_19

that the problem can be solved in reasonable computation time for real-world size instances using our model.

Keywords Mathematical programming · Radiotherapy scheduling · Patient preferences · Sequencing model

1 Introduction

As the number of new cancer cases increases [1], demand for radiotherapy (RT) is expected to grow by an average of 16% until 2025 [2]. In external-beam radiotherapy (RT), treatments are administered by a linear accelerator (linac), which delivers high-energy radiation to kill cancer cells and prevent them from multiplying. While RT resources (staff and machines) are expensive and delays in the start of treatment may induce greater psychological distress patients subject to longer waiting times [3], RT centers are encouraged to manage their linac capacity in the most efficient manner.

The RT treatment is divided into a set of (daily) irradiation sessions. Due to the large variety of possible treatment schemes and the high number of technical constraints involved, the problem of scheduling RT treatment sessions has increasingly been addressed in the literature [4]. As we elaborate in Sect. 2, the majority of the models proposed in the literature assign patients' irradiation sessions to linacs, as done in [5–7], neglecting the sequencing of patients in each day and linac. However, given that the majority of RT centers have enough linac capacity to treat all patients in due time [8], the main problem becomes on how to schedule the irradiation sessions such that not only timeliness constraints are satisfied, but also the fulfilment of patient preferences regarding the starting time of their sessions is maximized. Therefore, optimizing the sequencing of patients when scheduling RT treatments allowing the fulfillment of patient preferences becomes important and relevant for RT centers. Two studies addressed the problem of scheduling RT sessions considering time windows in the literature [9, 10], both on particle therapy (PT). However, in PT there are technical and medical constraints which are not found in conventional RT. For instance, in PT a single beam source is used by multiple treatment rooms, but only one room can use the source at a time. Besides, in PT there is a minimal and maximal number of days allowed between treatment sessions, and there has to be a break from the treatment of at least two consecutive days each week [9]. Therefore, the methods proposed in [9, 10] cannot be applied directly to conventional external-beam RT. The contribution of this paper is to propose a mixed-integer linear programming (MILP) model for the delivery of daily treatment sessions in conventional RT with maximization of time window preferences given by patients for the starting time of their treatment sessions. The applicability of our model is tested with real-world data from the Netherlands Cancer Institute (NKI), a large RT center (4966 treatments per year) based in Amsterdam, the Netherlands. Our model will serve as a basis for finding optimal schemes for the allocation of linac types to patient groups (e.g. the most advanced linacs treat the most complex tumor groups such as brain).

The remainder of this paper is organized as follows: Sect. 2 presents a more detailed description of the problem and related to current literature. The MILP model is presented in Sect. 3. Section 4 describes the computational experiments and corresponding results, and Sect. 5 outlines conclusions and draws future research lines on this project.

2 Problem Description and Background

In the RT scheduling problem, the aim is to schedule a set of treatment sessions for a set of cancer patients $\mathcal{P}$ over a given planning horizon $\mathcal{T}$, discretized in time periods $t = 1, \ldots, |\mathcal{T}|$ of usually one day. Each session of each patient $i \in \mathcal{P}$ has an estimated duration p_i, and sessions are mostly delivered in consecutive days (daily treatments). Each patient is assigned a due date d_i, which is the date by which the patient should start treatment based on his/her urgency level. Treatment sessions are delivered in a set of (technically feasible) linear accelerators $\mathcal{K}$, with a pre-defined number of time slots $|\mathcal{S}|$ available. In our MILP model, we make the following assumptions:

- Treatment sessions of a given patient have the same (estimated) duration. It is known that sessions' duration may vary between different patient groups, but current literature indicates that each particular patient is scheduled sessions with the same duration throughout the course of his/her treatment [5–7], regarding of his/her patient group.
- Sessions are delivered in consecutive days (daily treatments). Although some small patient populations may not be prescribed daily treatment sessions (e.g. hypofractionation schemes, in which high-dose radiation is delivered, may require one day off in between treatment sessions), we know that the great majority of the patient population in RT receive daily treatment sessions, as confirmed by the referred literature studies [5–7].
- There are no pre-allocated linacs to patient types, i.e. all linacs are available to treat all types of patients. While some linacs may be technologically more suited for some cancer types (e.g. brain) than others, literature shows that all irradiation machines are typically able to treat all types of patients.
- Patients can switch linacs during the course of treatment. Although from a patient perspective it is desirable that patients receive their irradiation sessions in the same machine (so they see the same facilities and most likely the same radiation technologists), there are no technical or medical constraints that point it as a necessary condition, as confirmed by the aforementioned studies.

Previous studies have approached different variants of the problem and several methods have been proposed. Sauré et al. [5] formulated the problem as a discounted infinite-horizon Markov decision process, with the percentage of treatments initiating treatment within 10 days increasing from 73 to 96%. Legrain et al. [11] proposed a two-step stochastic algorithm for online scheduling of RT sessions, with results

showing an average decrease in the number of patients breaching the standards of 50% for acute patients and 81% for subacute patients. In their case, consistent appointment times for irradiation sessions is achieved by considering that all treatment sessions of all patients have the same duration. However, in many centers such as the NKI this is not the case. Besides complying with timeliness requirements and technical constraints, those RT centers are interested in finding a schedule that maximizes patient preferences regarding the starting time of irradiation sessions. The goal is that the starting time of these sessions fall within the patients' desired time window $[t_i^{\min}, t_i^{\max}]$ in a consistent basis. To this end, models have been proposed [9, 10] such that the starting time of irradiation sessions do not deviate from a pre-defined target time by more than a certain threshold (30 min in both [9, 10]).

Overall, most models presented in the current literature focus on deciding upon the specific day and linac of each scheduled irradiation session, with the sequence of patients in each linac and each day being either neglected or determined on a second stage. Studies addressing the sequencing problem considering time windows are only able to solve the problem using approximation methods, which often lead to suboptimal solutions. In this paper, we propose a MILP model to solve the RT scheduling problem with time windows to optimality and test its performance for real-world size instances.

3 Methodology

In this section, we present the MILP model for the RT scheduling problem with time windows. We use the notation presented in Table 1.

In our formulation, linacs' daily availability is divided in time slots $s = 1, \ldots, |\mathcal{S}|$ of a fixed duration l. Irradiation sessions are scheduled by assigning a starting time slot in a given day and linac, and by preventing other sessions from being assigned to the remainder slots (needed to achieve the session's duration) in that same linac

Table 1 Notation of the MILP model

Set/Parameter	Description
$\mathcal{P}$	Set of patients to be scheduled ($i, j \in \mathcal{P}$)
$\mathcal{K}$	Set of identical linear accelerators ($k \in \mathcal{K}$)
$\mathcal{S}$	Set of time slots per linac ($s \in \mathcal{S}$)
$\mathcal{T}$	Set of workdays in the planning horizon ($t \in \mathcal{T}$)
I_i	Number of sessions to be delivered to patient i
d_i	Due date: day by which patient i must start treatment
p_i	Duration, in number of time slots, of each session of patient i
$t_i^{\min}, t_i^{\max}$	Lower and upper bound of the time window preference for patient i
l	Length of each time slot of each linac

and day. In our model, binary variables are represented by X^t_{iks}, which take the value 1 if patient i is scheduled a session starting on time slot s of linac k in day t, and 0 otherwise. Real variables Δ^-_{it} and Δ^+_{it} are used to minimize the overall deviation from the preferred time window by means of the objective function (1):

$$\min \sum_{i \in \mathcal{P} \setminus \{1\}} \sum_{t \in \mathcal{T}} (\Delta^-_{it} + \Delta^+_{it}) \tag{1}$$

The set of constraints is as follows:

$$\sum_{k \in \mathcal{K}} \sum_{s \in \mathcal{S}} X^t_{iks} \leq 1, \forall i \in \mathcal{P}, \forall t \in \mathcal{T} \tag{2}$$

$$\sum_{i \in \mathcal{P}} X^t_{iks} \leq 1, \forall k \in \mathcal{K}, \forall s \in \mathcal{S}, \forall t \in \mathcal{T} \tag{3}$$

$$\sum_{k \in \mathcal{K}} \sum_{s \in \mathcal{S}} \sum_{t \in \mathcal{T}} X^t_{iks} \leq I_i, \forall i \in \mathcal{P} \tag{4}$$

$$\sum_{k \in \mathcal{K}} \sum_{s \in \mathcal{S}} X^t_{iks} - \sum_{k \in \mathcal{K}} \sum_{s \in \mathcal{S}} X^{t-1}_{iks} \leq \sum_{k \in \mathcal{K}} \sum_{s \in \mathcal{S}} X^n_{iks}, \forall i \in \mathcal{P}, \forall t \in \mathcal{T},$$
$$\forall n = t, \ldots, \min\{|\mathcal{T}|, t + I_i - 1\}, \{X^0_{iks} = 0, \forall i, k, s\} \tag{5}$$

$$\sum_{k \in \mathcal{K}} \sum_{s \in S} \sum_{t=1}^{d_i} X^t_{iks} \geq 1, \forall i \in \mathcal{P} \tag{6}$$

$$X^t_{iks} \leq 1 - \sum_{i' \in \mathcal{P}} X^t_{i',k,s'}, \forall i \in \mathcal{P}, \forall k \in \mathcal{K}, \forall s = 1, \ldots, |\mathcal{S}| - p_i + 1,$$
$$\forall t \in \mathcal{T}, \forall s' = s + 1, \ldots, s + p_i - 1, p_i \geq 2 \tag{7}$$

$$X^t_{iks} = 0, \forall i \in \mathcal{P}, \forall k \in \mathcal{K}, \forall s = |\mathcal{S}| - p_i + 2, \ldots, S, \forall t \in \mathcal{T}, p_i \geq 2 \tag{8}$$

$$l(s-1)X^t_{iks} \geq t_i^{\min} X^t_{iks} - \Delta^-_{it}, \forall i \in \mathcal{P}, \forall k \in \mathcal{K}, \forall s \in \mathcal{S},$$
$$\forall t \in \mathcal{T} \tag{9}$$

$$l(s-1)X^t_{iks} \leq t_i^{\max} X^t_{iks} + \Delta^+_{it}, \forall i \in \mathcal{P}, \forall k \in \mathcal{K}, \forall s \in \mathcal{S},$$
$$\forall t \in \mathcal{T} \tag{10}$$

$$\Delta^-_{it} \geq 0, \Delta^+_{it} \geq 0, \forall i \in \mathcal{P}, \forall t \in \mathcal{T} \tag{11}$$

Constraints (2) ensure that each patient is scheduled at most one session per day. Restrictions (3) establish that at most one session per day is scheduled in each slot of each linac. Constraints (4) restrict the number of sessions delivered to the number of remaining sessions for that patient. Inequalities (5) ensure that patients receive daily sessions until the number of sessions or the end of planning horizon is reached. Constraints (6) impose that every patient starts treatment before their due date. Inequalities (7) ensure that no slots are booked for the remainder of the session duration p_i after the starting slot, while constraints (8) prevent sessions with two or more slots from being booked in the last slot(s) of the day. Constraints (9)–(10) force variables

Δ_{it}^{-} and Δ_{it}^{+} to take a non-zero value if a session's starting time deviates from the desired lower and upper bounds, respectively. Constraints (11) are the non-negativity constraints associated with the real variables.

4 Case Study and Preliminary Results

We tested our model by generating a set of test instances with various sizes regarding the number of patients ($|\mathcal{P}|$) and available linacs ($|\mathcal{K}|$) maintaining the patient-to-linac ratio of the NKI, i.e. 30 patients to be scheduled per linac, per week. We used historical data from the RT department of the Netherlands Cancer Institute (NKI), a large cancer center operating in the Netherlands, to generate patient-specific data for each instance size.

4.1 Input Data

Patient characteristics are generated according to empirical distributions generated using historical data collected throughout 2017 (number of treatments = 4966). In our algorithm, we start by generating a care plan (i.e. care trajectory) for each patient. There are 63 care plans in total, with the largest being "Bone metastasis" (22.5%), "Breast" (15.7%), "Long > 44 Gy" (5.7%), "Prostate" (4.9%), and "Head-and-neck" (4.7%). Thereafter we generate the number of sessions I_i of each patient, which can vary between 1 and 35 sessions depending, to a large extent, on the care plan. For instance, nearly half of all prostate patients will undergo 35 sessions, while 65% of all bone metastasis patients are prescribed 3 sessions or less. Similarly, the urgency level of each patient, which can be either urgent (34%) or regular (66%), is randomly assigned according to historical data associated with his/her care plan. Urgent patients need to start treatment in the earliest time period possible, thus $d_i = 1$ for all urgent patients. The due date of regular patients ($d_i = 1..5$) is generated per care plan and considers the maximum recommended waiting time for RT [12] (28 days from referral to start of treatment). We calculated, for each patient, the time difference between the waiting time target and the time elapsed from referral to the end of treatment planning to derive the corresponding empirical distributions. Furthermore, the duration of each session p_i is also assigned on a care plan basis, ranging from 10 to 30 min in multiples of 5 min. Data shows that the majority of patients will be scheduled 15-min sessions (60.5%), with 19.9% patients having sessions of 20 min or longer. The daily available time for delivering irradiation sessions in the clinic ranges from 07h30 to 17h30, thus $|\mathcal{S}| = 120$ by considering $l = 5$ min. We solve the problem for a planning horizon of one labour week, discretized in time periods of one day ($|\mathcal{T}| = 5$).

In the NKI, patient preferences, i.e. the preferences given by patients for the desired starting time of their irradiation sessions, are currently considered when scheduling irradiation sessions of regular patients only. In the case of urgent pa-

tients, preferences are not currently considered in order to provide more flexibility for finding earlier appointment slots for their treatment. Regular patients are asked for a 150-min window during which they would like to receive their sessions. However, data regarding patient preferences are not currently recorded in the clinic, thus historical data regarding patient preferences cannot be used. Nevertheless, interviews with the planners revealed that most patients who have a preference either prefer to have their sessions early in the morning (<10 h 00), or later in the day (>15 h 00), while some patients do not have a preference at all. Therefore, we randomly generated time window preferences $[t_i^{\min}, t_i^{\max}]$ for regular patients as follows: 1/3 of regular patients with a preference for the early morning (window = [0, 150]), and 1/3 with a preference for the end of the workday (window = [450, 600]). The remainder 1/3 of regular patients, as well as all urgent patients have no preferred time window ([0, 600]).

4.2 Preliminary Results

The MILP model was coded in C++ using Visual Studio 2017 and the Concert Technology of CPLEX v12.8.0, which was used as a solver. All experiments were conducted on a desktop computer with a processor Intel i7 3.6 GHz and 16 GB of RAM using up to 8 threads, running on a 64-bit version of Windows 10. The maximum allowed CPU time was set to 28,800 s (8 h), which was considered to be a reasonable time for an RT center to wait for an output solution to be implemented in practice. In our case study, in which the goal is to find a weekly schedule (Monday to Friday), RT managers can hypothetically run the model during the last workday of the previous week (i.e. Friday), and have a complete, perhaps optimal solution to be implemented in less than 8 h of computation time. Table 2 shows the results regarding the performance of the model using the described input data for different instance sizes.

Table 2 Results for several instance sizes using NKI patient data

#patients	#linacs	# sessions outside window	MIP gap (%)	CPU time (s)
30	1	0	0	2
60	2	0	0	14
90	3	0	0	37
120	4	0	0	2482
150	5	0	0	12,818
180	6	Out of memory	Out of memory	Out of memory

Analyzing the results of Table 2, we can see that the proposed MILP model is able to find an optimal solution that schedules all sessions within the desired time window for all instance sizes up to 150 patients and 5 linacs. Even for the instance with 150 patients and 5 linacs, the proposed formulation proved effective in finding the optimal weekly schedule in less than 3.5 h of CPU time. However, the computer used in our experiments ran out of memory when attempting to solve the problem for 180 or more patients, most likely due to complexity introduced by the exponentially higher number of constraints (around two million for the instance with 180 patients).

5 Conclusions and Future Research

Previous work on modeling the problem of scheduling RT sessions considering time windows resulted intractable even for small-sized instances and, as a result, (meta)heuristics have been proposed [9, 10]. In this paper, we propose an exact method for the RT scheduling problem with time windows that can be solved via MILP in reasonable computation time for real-world instance sizes. Besides providing automated decision making for scheduling RT treatments, our algorithm is capable of incorporating patient preferences while ensuring that all patients start treatment in due time. Moreover, our model has proved to be efficient in achieving an optimal solution for instances of up to 150 patients, although a feasible solution could not be achieved in due time for the instance with 180 patients and 6 linacs. In further research, we aim to explore alternative solution methods (e.g. row generation methods, constraint programming) for improving the computational time required to solve the problem to optimality for instances with more than 150 patients, and test the models for other patient preference schemes by varying the time window size and proportion of patients "competing" for the same time window. Moreover, we plan to use the model as a basis for finding optimal schemes for the allocation of some linac types to certain patient groups. For instance, RT centers often want the most advanced linacs to treat the most complex tumor sites, such as brain. In addition, we consider making use of prediction models for treatment times [13] to develop data-driven approaches considering the variability inherent to the session's duration for finding more robust schedules.

Acknowledgements The authors would like to thank to Eva Euser, Bram van den Heuvel, and Herman Vijlbrief from the department of radiation oncology of the NKI for providing the necessary clinical information to build the MILP model. We would also like to thank Maarten Broekhof for his help on gathering the data used in the computational experiments. This work was supported by Alpe d'Huzes/KWF under the ALORT project (2014-6078).

References

1. Siegel, R.L., Miller, K.D., Jemal, A.: Cancer statistics, 2017. CA Cancer J. Clin. **67**, 7–30 (2017)
2. Borras, J.M., Lievens, Y., Barton, M., Corral, J., Ferlay, J., Bray, F., Grau, C.: How many new cancer patients in Europe will require radiotherapy by 2025? An ESTRO-HERO analysis. Radiother. Oncol. **119**, 5–11 (2016)
3. Mackillop, W.J.: Killing time: the consequences of delays in radiotherapy. Radiother. Oncol. **84**, 1–4 (2007)
4. Vieira, B., Hans, E.W., van Vliet-Vroegindeweij, C., van de Kamer, J., van Harten, W.: Operations research for resource planning and -use in radiotherapy: a literature review. BMC Med. Inf. Decis. Making **16**, 149 (2016)
5. Sauré, A., Patrick, J., Tyldesley, S., Puterman, M.L.: Dynamic multi-appointment patient scheduling for radiation therapy. Eur. J. Oper. Res. **223**, 573–584 (2012)
6. Conforti, D., Guerruiero, F., Guido, R.: RASON: non-block scheduling with priority for radiotherapy treatments. Eur. J. Oper. Res. **201**, 289–296 (2016)
7. Legrain, A., Fortin, M.A., Lahrichi, N., Rousseau, L.-M., Widmer, M.: Stochastic optimization of the scheduling of a radiotherapy center. J. Phys. Conf. Ser. **616**(1), 012008 (2015)
8. Rosenblatt, E., Izewska, J., Anacak, Y., Pynda, Y., Scalliet, P., Boniol, M., Autier, P.: Radiotherapy in European countries: an analysis of the Directory of Radiotherapy Centres (DIRAC) database. Lancet Oncol. **14**, 79–86 (2013)
9. Maschler, J., Raidl, G.R.: Particle therapy patient scheduling with limited starting time variations of daily treatments. Int. Trans. Oper. Res. **00**, 1–22 (2018)
10. Vogl, P., Braune, R., Doerner, K.F.: Scheduling recurring radiotherapy appointments in an ion beam facility. J. Sched. **22**, 137–154 (2019)
11. Legrain, A., Fortin, M.A., Lahrichi, N., Rousseau, L.-M.: Online stochastic optimization of radiotherapy patient scheduling. Health Care Manag. Sci. **18**, 110 (2015)
12. NVRO: Waiting times, standards and maximum waiting times for radiotherapy (in Dutch). Available at http://www.nvro.nl/kwaliteit/indicatoren/ (2019). Accessed 19 Feb 2019
13. Bentayeb, D., Lahrichi, N., Rousseau, L.-M.: Patient scheduling based on a service-time prediction model: a data-driven study for a radiotherapy center. Health Care Manag. Sci. **22**(4), 768–782 (2018)

Pattern-Based Online Algorithms for a General Patient-Centred Radiotherapy Scheduling Problem

Roberto Aringhieri, Davide Duma and Giuseppe Squillace

Abstract A radiotherapy treatment consists in a given number of radiation sessions, one for each (working) day, which should start before a given due date. Patients are usually classified into classes of urgency having different deadlines and number of sessions. Waiting time is the main critical issue in the management of a radiotherapy health system. After deriving a general problem statement from the case studies reported in the literature, we present three online optimisation algorithms that try to exploit the particular structure of the solution, and we compare their results with two baseline online algorithms.

Keywords Radiotherapy · Scheduling · Online algorithm

1 Introduction

Malignant tumours can be treated by a radiation treatment, which consists in the use of ionising radiation. A radiation therapy, say also radiotherapy, is delivered by a linear accelerator or *linac* in a clinical setting. A linac is a special device whose main function is to concentrate in beams and accelerate the emission of subatomic particles. Radiotherapy is the primary treatment for many type of cancers whilst, for others, it is used in combination with other forms of therapy (surgery and/or chemotherapy). More generally, an accurate treatment plan depends on the cancer type, location and stage, as well as the general conditions of the patient. A radiotherapy consists in a given number of radiation sessions, one for each (working) day, which should start

R. Aringhieri (✉) · D. Duma · G. Squillace
Dipartimento di Informatica, Università degli Studi di Torino,
Corso Svizzera 185, 10149 Torino, Italy
e-mail: roberto.aringhieri@unito.it

D. Duma
e-mail: davide.duma@unito.it

G. Squillace
e-mail: giuseppe.squillace@edu.unito.it

V. Bélanger et al. (eds.), *Health Care Systems Engineering*,
Springer Proceedings in Mathematics & Statistics 316,
https://doi.org/10.1007/978-3-030-39694-7_20

after a given release date and before a given due date. Patients are usually classified into classes of urgency having different deadlines and number of sessions to be delivered.

Waiting time is the main critical issue in the management of a radiotherapy health system. Actually, the delay between the first consultation and the first treatment is typically rather long. Such a delay has the potential to damage the health status of the patients both directly and indirectly. For instance, a delay can affect the radiotherapy outcomes by permitting proliferation of clonogenic cells within the field, leading to a decrease in the probability of local control [7]. Radio-biological principles suggest that prevailing waiting times for radiotherapy, which often approximate the doubling time of a fast growing human tumour, may have a clinically significant effect on local control [8].

From such an analysis, it clearly emerges the room for applying optimisation techniques to the scheduling of the patients needing a radiotherapy treatment, in order to improve both the quality of the health service provided (reducing the waiting times) and the utilisation of the involved resources (linacs and personnel). After deriving a general patient-centred problem statement from the case studies reported in the literature (Sect. 2), we present three online optimisation algorithms that tries to exploit the particular structure of the solution (Sect. 3), and we compare their results with two baseline online algorithms (Sect. 4).

2 Literature Review and Problem Formulation

General patient-centred problem statement. The Radiotherapy Patient Scheduling (RPS) problem falls into the broader class of multi-appointment scheduling problems in hospital in which patients need to visit sequentially multiple or single resource types in order to receive treatment or be diagnosed [9]. In order to provide a general problem statement, we resume the real case studies in the literature and we choose the most general settings taking into account the best ones from a patient-centred perspective.

Addressing any appointment scheduling problem, we have a certain number of machines that could be equals or provide different services, available on a subset of days over the planning horizon, and organised on one or more shifts in such days. In the RPS, shifts are usually one per day or two per day (morning and afternoon). Patients of different categories, that means different urgencies and treatments, should be performed on the shifts minimising their waiting times and, in some cases, trying to respect a deadline called due date.

A first classification about any appointment scheduling problem can be made with respect to the choice of using the blocking or the non-blocking policy. Most of the contributes of the literature use the former, that is dividing each shift in slots of fixed time and assigning them to the patients in accordance with several rules. Some works in the literature use the non-blocking policy for the RPS (see, e.g., [3]) which consists in selecting the sessions to be performed in each shift and ordering such sessions in order to provide an appointment to the patients. We limit our summary to the contributions under the blocking policy, which are reported in chronological

Table 1 Summary of the main problem setting in the literature

Reference	[11]	[2]	[10]	[13]	[14]	[6]	[12]
First author	Petrovic	Conforti	Petrovic	Sauré	Tang	Legrain	Riff
Year	2006	2008	2011	2012	2014	2015	2016
Slots per session	1	1	≥1	≥1	1	1	1
C1	✗	✗	✓	✓	✓	✓	✗
C2	✗	✓	✗	✗	✗	✓	✗
C3	✗	✗	✗	✓	✗	✗	✗
Setup slots	✗	1	✗	1–2	✗	✗	1–2
Machine types	2	1	4	1	1	1	2
Categories	3	4	3	5	2	3	3
Due dates (days)	2, 14, 28	1, 14, 28, n.d.	2, 14, 28	5, 8, 10, 12, 15	n.d.	3, 4, 18	2, 14, 28
Overtime	✗	✗	✗	✓	✗	✓	✗
Problem	NO	CA	NO	CA	CA	PO	CA

order in Table 1. Usually a session is scheduled into a single slot, but in [10, 13] it could require two or three consecutive slots of the same shift, depending on the patient therapy.

Due to medical and patient-centred reasons, there are three common constraints that could be considered. The first constraint (C1) is the absence of interruptions in the series of sessions, that is patients have to be always scheduled in consecutive working days. The second constraint (C2) is the planning of the patient sessions always in the same daily slot, that means to give always the same appointment time to the patient. The third constraint (C3) is the availability of the patient during the shift hours of the day, that is a sort of preference expressed during the booking.

The first session could require one or two additional slots for the setup time, that is the time necessary to set the machine [2, 12, 13]. The setup slot considerably complicates the scheduling of the treatments under the constraints C1 and C2. The reason is that this causes a *hook* shape for the patient sessions into the schedule, as shown in the example in Fig. 1.

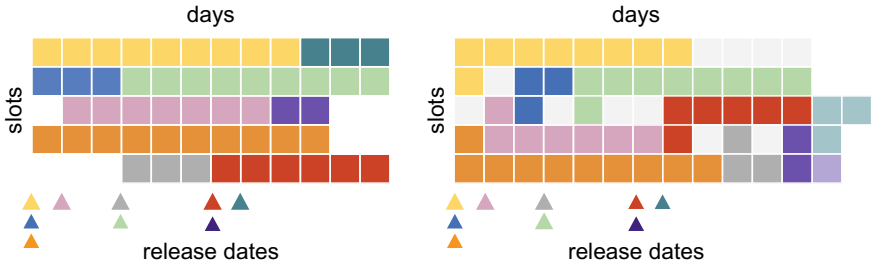


Fig. 1 Comparing RPS with constraints C1–C2 with (right) and without (left) the setup slot: the number of slots for each patient (indicated with a different colour) is the same to highlight how the hook shape complicates the scheduling. Days are ordered from left to right

In [10–12] different slot durations are set for the scheduling of several types of treatments, that is on different shifts since they require the use of two or more different machines.

Patients are always classified into two or more categories, that is patients with different urgency for which different due dates are usually defined. Most of the works refer to the same three classes of patients: even if they are called with different names, the same due dates are provided, that is 28 days for the less urgents (called *palliative* or *radicals*), 14 days for the middle urgency (*curative* or *palliative*) and 1–3 days for the others (called *emergency*, *urgent* or *others*). The various categories have also a different arrival rate and also a different number of sessions for the whole treatment, that is usually few days for most urgent patients up to 6–9 weeks for the most urgent ones.

Generally, the appointment scheduling approaches can be classified into two categories: *offline* and *online*. The offline appointment scheduling is performed cyclically on a certain planning horizon after collecting the request of treatments in the previous period. Then, at the moment of the scheduling, the demand is known and the schedule is (usually) empty. On the contrary, in the online appointment scheduling the appointment requests are updated over time, then the sessions have to be scheduled (on a not empty schedule) without knowing the future requests. Because of the characteristics of the RPS, which requires a timely planning, an online appointment scheduling approach is necessary, as demonstrated by the contributions in the literature. Online approaches for the multi-appointment scheduling problems can be divided into three levels: *capacity allocation* (CA), *near-online* (NO) and *pure-online* (PO). At the CA level, the decision to be taken is the allocation of the slots to the patients over a certain planning horizon (usually one week). At the NO level, the request of treatments is collected in a short time (e.g. daily) before of allocating the slot to the sessions, then appointments are promptly provided to the patients. Finally, at the PO level appointments are given in real time at the moment of the patient arrival.

Our approach places at the PO level, which presents a lack of attempts in the literature: such a level is addressed only in [6] even if it represents the most common organisation approach in the real life. We use a blocking approach on several machines of the same type. We take into account constraints C1–C2, that is each session is performed in the same shift of the day, on the same machine and in the same slot and it is not possible to interrupt the series of sessions in the working days. We assume that each session requires one slot, except for the first session that includes an additional setup slot determining, together with C1–C2, the hook shape of the patient sessions. For the sake of simplicity, all the slots of a sessions series (that are in the same position of the same shift in different days in accordance to C1 and C2) are denoted as *regular slots*, while the additional setup slot is denoted as *left slot* or *right slot* depending on whether the setup slot precedes (left) or follows (right) the regular slot of the first appointment, respectively. Finally, we consider an operative context in which sessions can be scheduled over 5 days a week, without overtime, to patients divided into different categories.

Integer linear programming model. Let J be the set of all the patients generated over the time horizon indicated as a set of working days K. Patients can start the

session series only in working days $k \in K'$, $K' \subset K$, that are not pre-holiday (i.e., Friday or Christmas eve are not feasible starting days). Let W be the set of the slots of each day, with $m = |W|$. Slots occupied by the sessions of already scheduled patients are indicated by the matrix $S = (s_{kw})_{k \in K, w \in W}$, with $s_{kw} = 1$ if the slot is occupied, and $s_{kw} = 0$ otherwise. Let d_j be the duration of the treatment of the patient $j \in J$, that is the number of sessions to be scheduled, and let ρ_j be the release date, that is the first day in which the patient can start the treatment.

Let us define the following decision variables to give a mathematical programming formulation of the problem: y_{jkw} is equal to 1 if a session of the patient j is scheduled in the slot w of the day k, 0 otherwise; t_{jkw} is equal to 1 if the first session of the patient j is scheduled in the slot w of the day k, 0 otherwise ($t_{jkw} = 1 \Rightarrow y_{jkw} = 1$); r_{jkw} is equal to 1 if the slot w on the day k is the right slot of the patient j, 0 otherwise; l_{jkw} is equal to 1 if the slot w on the day k is the left slot of the patient j, 0 otherwise. Inspired to that reported in [2], the following integer linear programming (ILP) model describes the offline version of our problem on a given shift:

$$\min \quad z = \sum_j \left[\sum_k (k - \rho_j) \sum_w t_{jkw} + \left(1 - \sum_k \sum_w t_{jkw} \right) (|K| - \rho_j + 1) \right]$$

subject to

$$\sum_j \left(y_{jkw} + r_{jkw} + l_{jkw} \right) + s_{kw} \leq 1 \qquad \forall k, w = 2, \ldots, m \qquad (1)$$

$$\sum_j \left(y_{jk1} + l_{jk1} \right) + s_{k1} \leq 1 \qquad \forall k \qquad (2)$$

$$\sum_j \left(y_{jkm} + r_{jkm} \right) + s_{km} \leq 1 \qquad \forall k \qquad (3)$$

$$\sum_k \sum_w t_{jkw} \leq 1 \qquad \forall j \qquad (4)$$

$$t_{jkw} = r_{jkw+1} + l_{jkw-1} \qquad \forall j, \forall k, w = 2, \ldots, m-1 \qquad (5)$$

$$t_{jk1} = r_{jk2}\,, \quad t_{jkm} = l_{jkm-1} \qquad \forall j, \forall k \qquad (6)$$

$$y_{jkw} \geq t_{jkw} \qquad \forall j, \forall k, \forall w \qquad (7)$$

$$\sum_{\tau=k}^{\min\{k+d_j, |K|\}} y_{j\tau w} \geq \min\{d_j, |K| - k\} t_{jkw} \qquad \forall k \qquad (8)$$

$$\sum_{\tau=\max(k-d_j,1)}^{k} t_{j\tau w} \geq y_{jkw} \qquad \forall k \qquad (9)$$

$$t_{jkw} = 0 \qquad \forall j, \forall k = 1, \ldots, \rho_j - 1, \forall w \qquad (10)$$

$$t_{jkw} = 0 \qquad \forall j, \forall k \in K \backslash K', \forall w \qquad (11)$$

Constraints (1)–(3) impose that each slot of each day can be occupy at most by one patient, that can be for a regular slot ($y_{jkw} = 1$), a left slot ($l_{jkw} = 1$), a right slot ($r_{jkw} = 1$) or an already scheduled slot ($s_{jkw} = 1$). In particular, (2) indicate that the first slot of the shift can not have a left slot and likewise (3) impose that the last slot of the shift can not have a right slot, while (1) concern all the other slots. Constraints (4) make sure that all patients have at most one first session. Constraints (5) impose that we have a setup slot in correspondence of the first session, which can be a left slot or a left slot. Constraints (7) combine the variables y_{jkw} and t_{jkw} to make them coherent. The operative constraints C1–C2 are fixed by (8) and (9), respectively, while (10) and (11) impose that the session series can not start before the release date and in a pre-holiday, respectively. Furthermore $y_{jkw}, t_{jkw}, l_{jkw}, r_{jkw} \in \{0, 1\}$.

The objective function that we would minimise is the waiting time. To this purpose we sum the number of days between the release date and the first session for scheduled patients. Otherwise, for all the other patients we sum the number of days between the due date and the last day of the planning horizon plus one.

The ILP model can be easily generalised to the case of two or more shifts and different machines. Although the high complexity of its offline solution [2], we try to solve smaller instances of this model (i.e., 10 slots and 10 working days as planning horizon) to identify a common pattern that can be exploited in the development of new online algorithms.

3 Online Algorithms

We propose several online algorithms for the pure online version of the RSP, that is patients are scheduled one by one in real time. Starting from an adaptation to our operational settings of the heuristics proposed in [11] for the NO level, that is the As-Soon-As-Possible (ASAP) and the Just-In-Time (JIT), we define two baseline online algorithms in order to compare them with our approach.

For the sake of simplicity, we indicate with the term *right hook* and *left hook* a sessions series with a right slot and a left slot, respectively, observing that each sessions series can be scheduled as a right hook or a left hook, since the decision of the slot to dedicate for the setup is a decision that is taken during the scheduling. Further, we suppose to have two shifts on each machine, but the approaches are easily adaptable to scenarios with one or more shifts.

The ASAP and the JIT are adapted to the case of sessions series with a hook shape. In this context, placing only right hooks or only left hooks side by side means a trivial lost of slots, which will be not allocated due to the lack of two adjacent slots needed for the first session. We call Smart ASAP (SASAP) and Smart JIT (SJIT) our adaptions.

Three further online algorithms are conceived on the basis of a pattern, say *fountain effect*, observed analysing the above ILP model solutions on small instances. The fountain effect consist in the nesting of hooks of the same type (right or left) on the two sides of the shift (from the beginning forward or from the end backward,

respectively). A first online algorithm called Fountain On Shift (FOS) has been conceived exploiting this idea. However, we observed that in some cases it is better to choose only one side of the shift for the nesting, then for machines that have two shifts per day we propose a further algorithm called Fountain On Machine (FOM). Another observation is that nesting a small hook between two longer hooks causes a waste of slots, then the Selective Fountain On Machine (SFOM) pre-assigns a category to each shift in proportion to the expected demand and tries to exploit the fountain effect on the selected shift. In all the algorithms, patients with a very short due date are scheduled with the SASAP rule.

We provide in Algorithm 1 a general scheme for the proposed online approaches, which take in input a patient flow F and the schedule of the already assigned slots indicated with the 3-dimensional matrix S, defined on the set of the working days K, the set of the shifts Σ, and the set of the daily slots W. Each algorithm waits for the arrival of new patients for the whole time horizon and it schedule them online one by one. A sequence of decisions is taken on the basis of the schedule S and the patient to be scheduled p. The first choice is the shift σ in which the patient p will be scheduled (*selectShiftFromMachines*), then the first regular slots t and the additional (right or left) slot a are selected (*selectFirstSlots*). Finally, on the basis of such decisions, the appointments are provided to the patient p and the schedule S is updated (*schedulePatient*).

Algorithm 1: General online RPS scheme

Data: Patient flow: F, Scheduling matrix: $S = (s_{k\sigma w})_{k\in K,\sigma\in\Sigma,w\in W}$
Result: Scheduling matrix

```
k=1;
while k ∈ timeHorizon do
    while new patients arrive do
        p := newPatient(id, category, duration, releasedate, delay);
        σ := selectShiftFromMachines(S, p);
        (t, a) := selectFirstSlots(S, σ, p);
        S := schedulePatient(S, σ, t, a, p);
    k=k+1;
return S;
```

The functions introduced in the general scheme are defined in order to provide our five online approaches, which are described below, while an example on a small instance is illustrated in Fig. 2.

SASAP: The function *selectShiftFromMachines* chooses the shift σ that has a feasible solution that minimizes the waiting time of the patient p and, if two or more shifts provides the same waiting time, a pre-fixed order in Σ is followed. Then, the function *selectFirstSlots* selects the first two feasible slots and chooses the hook orientation (left or right) that minimises the number of adjacent empty slots and returns the first regular slot t and the additional slot a accordingly.

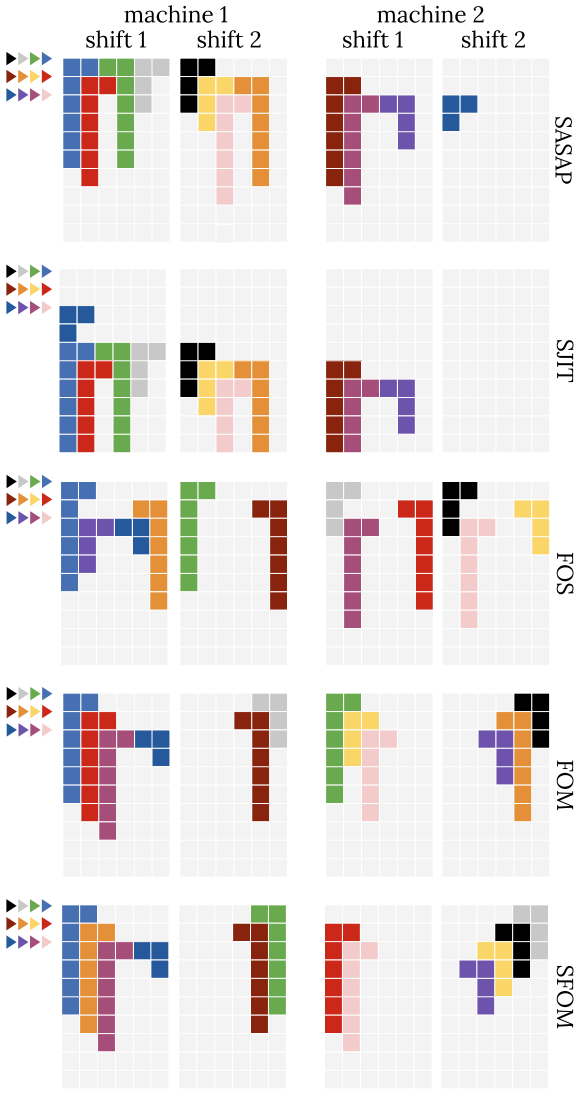


Fig. 2 Solutions provided by the 5 proposed algorithms for a small instance. We suppose to have an empty schedule and 12 patients indicated with different colours, with release dates on the first 3 days and belonging to 3 different categories. The arrival order is the following: light blue, green, grey and black (day 1), red, yellow, orange and brown (day 2), pink, purple, violet and dark blue (day 3). The hooks with duration 3 and 6 have due date after 4 days after the release date, while the one with duration 2 have to be scheduled as soon as possible

SJIT: The function *selectShiftFromMachines* chooses the shift σ that has a feasible solution as late as possible within the due date; if such solution is provided by two or more shifts, a pre-fixed order in Σ is followed, otherwise if any shift has a feasible solution the patient is scheduled using the SASAP. Then, the function *selectFirstSlots* is the same of the SASAP.

FOS: The function *selectShiftFromMachines* selects the less loaded shift σ. Then, *selectFirstSlots* chooses alternately the the left and the right side of the shift to

nest the hook with respect the fountain pattern as soon as possible, and the first slots t and a are set accordingly.
FOM: The function *selectShiftFromMachines* selects the less loaded machine and chooses alternately one of its shifts. Then, *selectFirstSlots* chooses always the same side of the shift to nest the hook (e.g., the left side for the morning shifts and the right side for the afternoon shift) with respect the fountain pattern as soon as possible, and the first slots t and a are set accordingly.
SFOM: The function *selectShiftFromMachines* selects the less loaded shift σ among those pre-assigned to the category of the patient p. Then, *selectFirstSlots* chooses always the same side of the shift to nest the hook with respect the fountain pattern as soon as possible, and the first slots t and a are set accordingly.

4 Preliminary Computational Results

In this section we provide a quantitative analysis for two different scenarios that differs only for the workload determined by a different patient arrival rate. For each scenario, 10 different instances has been randomly generated using *GeneRa*, an instance generator available online [1], capable to take into account specific characteristics inherent to different scenarios and realistic organisational settings.

In accordance with the literature reported in Sect. 2, our scenarios consider 3 categories of patients, that we call *radicals*, *curative* and *urgent* denoted by 1, 2, and 3, respectively. For these categories, we set a frequency of the 67%, 31% and 2%, a duration of 40, 15 and 3 days, and a due date after 28, 14 and 2 days from the release date, respectively. These parameters are consistent with the literature. We suppose to work on two identical machines and over two shifts per machine, composed by 20 slots each one. For each day of a time horizon of 260 days (52 weeks of 5 working days) GeneRa provides a certain number of patients belonging to the 3 categories in accordance with the defined parameters. Such a number is generated using a uniform distribution of minimum 0 and maximum M. We fix such a parameter in order to have the 80% and the 100% as upper bound for the machine utilisation, that is the ideal but geometrically impossible case in which all patients are scheduled and treated within the time horizon. Then, we obtained two different scenario S_1 and S_2 for $M = 4$ and $M = 5$, that is 2 and 2.5 patients per day on average, respectively.

In Table 2 we report the average results about a set of performance indices evaluated on 10 instances for each scenario, excluding the first 20% of patients (in arrival order) considered as warm up. Such instances generated 507 and 653 patients on average, of which only the last 406 and 523 are taken into account for the results, respectively. Due to the limited number of the analysed instances, some indices could have large confidence intervals, but the differences between the performance of the online algorithms is such that the intervals do not overlap, except few cases (e.g., FOS vs. FOM in scenario S_1 for patients scheduled and treated in time). Results give very clear indications about which is the best algorithms for the two analysed scenarios, no trade-offs are indeed found between the performance indices.

Table 2 Results of the quantitative analysis (average values on 10 different instances): columns indicated with 1, 2, and 3 refer to indices of radicals, curative, and urgent patients, respectively

Scen.	Algorithm	Utilisation	% scheduled patients				% treated in time				Waiting time (days)			
			1	2	3	All	1	2	3	All	1	2	3	All
S_1	SASAP	**66.6%**	87.3%	87.1%	95.8%	**87.4%**	93.2%	67.8%	12.4%	**83.5%**	12.3	12.2	13.1	**12.3**
	SJIT	**59.8%**	77.1%	78.2%	84.5%	**77.6%**	22.5%	8.1%	37.0%	**18.1%**	35.7	34.0	25.4	**34.9**
	FOS	**67.5%**	88.5%	88.4%	100%	**88.7%**	96.4%	76.4%	61.7%	**89.3%**	9.6	9.7	3.9	**9.5**
	FOM	**68.1%**	89.3%	88.7%	100%	**89.3%**	97.4%	81.9%	66.5%	**91.9%**	8.1	7.9	3.2	**7.9**
	SFOM	**71.5%**	95.3%	97.5%	100%	**96.1%**	100%	100%	100%	**100%**	3.1	2.2	0.2	**2.7**
S_2	SASAP	**68.6%**	64.3%	64.9%	71.4%	**64.6%**	23.9%	2.2%	0.0%	**16.6%**	42.8	44.0	42.2	**43.2**
	SJIT	**62.3%**	56.9%	58.0%	64.3%	**57.3%**	0.2%	0.0%	0.7%	**0.1%**	62.0	63.5	60.7	**62.5**
	FOS	**71.0%**	66.5%	67.2%	97.6%	**67.2%**	37.6%	7.9%	9.2%	**27.5%**	35.8	37.4	23.4	**36.0**
	FOM	**72.2%**	68.5%	69.4%	97.6%	**69.2%**	40.6%	9.5%	30.3%	**30.5%**	34.2	35.1	10.4	**34.0**
	SFOM	**80.1%**	75.6%	93.3%	97.6%	**81.5%**	59.1%	83.7%	100%	**68.7%**	26.1	8.3	0.1	**19.2**

The SASAP and the SJIT algorithm provide significantly different results: the latter seems to schedule inefficiently because of the hook shape of the sessions series, which hinders the insertion from low to high. On the contrary, SASAP provides a discrete performance for the balanced scenario S_1, while poor results are obtained for the overloaded scenario S_2, for this reason we refer to this online algorithm for the evaluation of our proposed algorithms.

Using the FOS, we are able to provide a first slight reduction of the waiting times, which decreases in average of 2.8 days for the scenario S_1 and 7.2 days in the scenario S_2, with 5.8 and 10.9% more patients treated within the due date. However, exploiting the fountain structure only on one side of the shift, that is using the FOM, we obtain a further improvement of all the indices in both the scenarios: utilisation increases of the 1.9–4.6% compared to the SASAP, while 8.4–13.9% more patients start the sessions series within the due date, and on average patients have the first appointment 4–9 working days earlier. Furthermore, the FOM provides different waiting times for the three patient categories, while using the SASAP they are uniform.

Finally, the SFOM is the best algorithm for both the scenarios S_1 and S_2. When the workload is balanced (scenario S_1), we have almost 4 more occupied slots every day, and all patients are treated in time with very low average waiting times. Considering the overloaded scenario S_2, the SFOM is able to treat within the due date all urgent patients and most of those belonging to the other categories, with average waiting times lower than the maximum allowed delay. As a matter of fact, we remark that the SFOM seems the most robust algorithm as soon as the workload increases.

5 Conclusions

In this paper we provided a general problem statement and a set of new online algorithms for the RSP. We provided three new online algorithms that exploit a special pattern, say fountain effect, observed in the offline solutions of the RPS when the sessions series has a hook shape due to the setup slot in the first appointment. We further provided the adaptation at the PO level of two already proposed heuristics for the NO level, which are used as baseline in our algorithm comparisons.

The quantitative analysis proved the effectiveness of our algorithms, showing on two different scenarios that the proposed online algorithms are able to significantly decrease the waiting times and to increase the machine utilisation. The algorithm that provides the best performance is the SFOM, which is based on the idea of exploiting the special fountain pattern and of pre-assigning different shifts to patients of different categories.

Further developments of this work should consider an experimental competitive analysis [4], look-ahead online algorithms [5], and a better instance generator (arrivals distributed as a Poisson process instead of a discrete uniform). In terms of problem settings, it could be of interest to generalise our algorithms to other operative contexts (e.g., the possibility of interrupting the sessions series, changing the assigned slots along the sessions series, …).

References

1. Cares, J., Riff, M.C., Neveu, B.: Genera: a problem generator for radiotherapy treatment scheduling problems. Ann. Math. Artif. Intell. **76**(1–2), 191–214 (2016)
2. Conforti, D., Guerriero, F., Guido, R.: Optimization models for radiotherapy patient scheduling. 4OR **6**(3), 263–278 (2008)
3. Conforti, D., Guerriero, F., Guido, R.: Non-block scheduling with priority for radiotherapy treatments. Eur. J. Oper. Res. **201**(1), 289–296 (2010)
4. Duma, D., Aringhieri, R.: The real time management of operating rooms. In: Kahraman, C., Topcu I. (eds.) Operations Research Applications in Health Care Management, International Series in Operations Research & Management Science, vol. 262, pp. 55–79. Springer International Publishing AG (2018)
5. Dunke, F., Nickel, S.: A general modeling approach to online optimization with lookahead. Omega **63**, 134–153 (2016)
6. Legrain, A., Fortin, M.A., Lahrichi, N., Rousseau, L.M.: Online stochastic optimization of radiotherapy patient scheduling. Health Care Manag. Sci. **18**(2), 110–123 (2015)
7. Mackillop, W.: Killing time: the consequences of delays in radiotherapy. Radiother. Oncol. **84**(1), 1–4 (2007)
8. Mackillop, W., Bates, J., O'Sullivan, B., Withers, H.: The effect of delay in treatment on local control by radiotherapy. Int. J. Radiat. Oncol. Biol. Phys. **34**(1), 243–250 (1996)
9. Marynissen, J., Demeulemeester, E.: Literature review on multi-appointment scheduling problems in hospitals. Eur. J. Oper. Res. **272**(2), 407–419 (2019)
10. Petrovic, D., Morshed, M., Petrovic, S.: Multi-objective genetic algorithms for scheduling of radiotherapy treatments for categorised cancer patients. Expert Syst. Appl. **38**(6), 6994–7002 (2011)
11. Petrovic, S., Leung, W., Song, X., Sundar, S.: Algorithms for radiotherapy treatment booking. In: Proceedings of 25th Workshop of the UK Planning and Scheduling Special Interest Group, vol. 25, pp. 105–112 (2006)
12. Riff, M.C., Cares, J., Neveu, B.: Rason: a new approach to the scheduling radiotherapy problem that considers the current waiting times. Expert Syst. Appl. **64**, 287–295 (2016)
13. Sauré, A., Patrick, J., Tyldesley, S., Puterman, M.: Dynamic multi-appointment patient scheduling for radiation therapy. Eur. J. Oper. Res. **223**(2), 573–584 (2012)
14. Tang, J., Yan, C., Cao, P.: Appointment scheduling algorithm considering routine and urgent patients. Expert Syst. Appl. **41**(10), 4529–4541 (2014)

Multi-level Heuristic to Optimize the Chemotherapy Production and Delivery

Alexis Robbes, Yannick Kergosien and Jean-Charles Billaut

Abstract The bio pharmaceutical unit of Oncology Clinic (UBCO) of the hospital of Tours (France) produces between 100 and 300 injections per day for three hospital units of Tours. The production of chemotherapy drugs consists of two steps: a sterilization step and a preparation step performed by pharmacists. The production process can be modeled as a hybrid flow shop scheduling problem. Once the drugs are completed, they have to be delivered to the patient at a given due date. The delivery problem is a variant of the Multi-Trip Vehicle Routing Problem. We propose in this paper a multi-level heuristic to solve the integrated production and delivery problem. Computational experiments are conducted on real-life based instances to compare multiple settings and to evaluate the efficiency of the proposed approach.

Keywords Integrated · Scheduling · Routing · Chemotherapy production

1 Introduction

The health care system is a demanding public service with various challenges. This paper focuses on an integrated chemotherapy production and delivery problem. In 2010, a first work with the bio pharmaceutical unit of Oncology Clinic (UBCO) of the hospital of Tours (France) [1] proposed to optimize the preparation of the chemotherapy products by solving a parallel machine scheduling problem. A first integrated solution to the UBCO [2] was presented in 2011. A method for a combined transportation and scheduling version of the problem [3] was proposed in 2017,

A. Robbes (✉) · Y. Kergosien · J.-C. Billaut
Université de Tours, LIFAT EA 6300, CNRS, ROOT ERL CNRS 7002,
64 Avenue Jean Portalis, 37200 Tours, France
e-mail: alexis.robbes@univ-tours.fr

Y. Kergosien
e-mail: yannick.kergosien@univ-tours.fr

J.-C. Billaut
e-mail: jean-charles.billaut@univ-tours.fr

V. Bélanger et al. (eds.), *Health Care Systems Engineering*,
Springer Proceedings in Mathematics & Statistics 316,
https://doi.org/10.1007/978-3-030-39694-7_21

however this study considers only one delivery man and a simplified workshop configuration. This paper is an extension of this study where the scheduling problem is a hybrid flow shop scheduling problem and the delivery problem is a variant of the Multi-Trip Vehicle Routing Problem. Even if the two sub-problems (the scheduling problem and the delivery problem) are considered independently, their resolution remains difficult. Most of the variants of the hybrid flow shop scheduling problem are NP-Hard [4], the same for the Multi-Trip Vehicle Routing Problem [5, 6]. Integrated production and distribution scheduling problems have been studied in several papers [7–9] where the objective functions are to minimize the combination of production and delivery costs or to minimize the makespan. In this study, we propose a model of an integrated chemotherapy drugs production and delivery problem which represents the real-life case (note that many services like UBCO have the same configuration). The objective is to minimize the total tardiness in order to provide a better health service quality. We propose a multi-level heuristic to solve the problem within a reasonable computation time, in order to be applied online and to compute an updated solution every time a new event occurs (e.g. the arrival of a new request).

2 Problem Definition

The process of a chemotherapy treatment requires various steps. First of all the patient receives a medical consultation few days before the treatment. At the end of this consultation, the doctor prescribes the forthcoming treatment. A production order with the prescribed drugs is sent to the UBCO and another consultation is scheduled just before the beginning of the treatment. During this second consultation, the doctor checks the patient health and validates the previous prescription. The preparation of the order by the UBCO can only start after this validation in order to avoid the losses of drugs. Figure 1 illustrates this process.

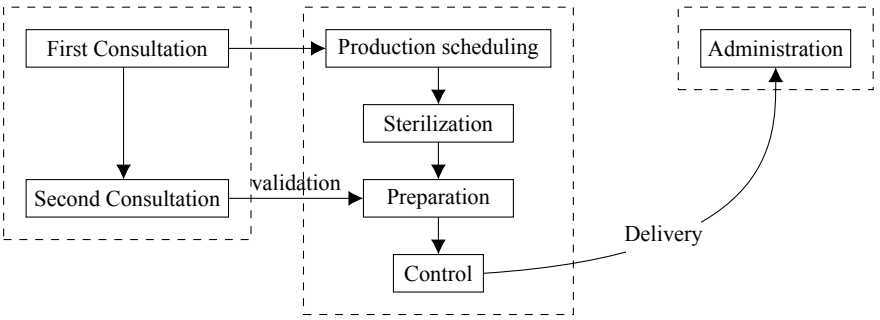


Fig. 1 Chemotherapy treatment process

The chemotherapy production is a 3-step process: `Sterilization`, `Preparation` and `Control`. Due to the isolators design, the `Sterilization` and the `Preparation` steps are done in the same isolator. Each isolator has several work stations where operators handmake the `Preparation` step. When a chemotherapy is completed, a `Control` step is executed on a single automated analyzer.

The delivery part is done by a delivery men team. Each chemotherapy drug has to be delivered to a given patient. All patients are dispatched in different oncology units of several hospital units.

The set of chemotherapy drugs to produce and to deliver is represented by the set of jobs J. Each job j in J has a release date r_j corresponding to the validation time before which the `Preparation` step cannot start, a processing time p_j^O for the `Preparation` step, an assigned oncology unit u_j where it has to be delivered before its due date d_j.

The production is done with $|I|$ identical parallel isolators. An isolator is characterized by a sterilizer capacity Q (i.e maximum number of jobs), a `Sterilization` processing time p^S (which do not depend on the sterilized batch of jobs) and a number of operators m which can work at the same time (i.e. number of work stations).

The `Control` step is proceeded by a single automated analyzer. The `Control` processing time p^A is the same for every job. To deliver the jobs, $|V|$ delivery men can make more than one trip. The objective function is to minimize the total tardiness $\sum_{j \in J} T_j$ where T_j is the delivery tardiness of the job j computed by $T_j = \max(0, D_j - d_j)$ where D_j denotes the delivery date.

We propose a modelization of the chemotherapy production and delivery problem as an integrated scheduling and routing problem. The scheduling part corresponds to a 3-stage Hybrid Flow shop scheduling problem. The routing part corresponds to a variant of the Multi-Trip Vehicle Routing Problem with due dates.

Let consider a given schedule and a given delivery plan, for every job j, c_j^O denotes the completion time of the `Preparation` step, c_j^A denotes the completion time of the `Control` step. The batch `Control` completion time is the maximum `Control` completion time of the jobs in the batch.

Figure 2 is a Gantt chart representing a partial solution of a problem instance with 2 isolators, 2 operators per isolator and 2 delivery men. As an example we highlight the process of the job 20 from the `Sterilization` step to the delivery. First, the job is sterilized in the first batch of isolator 1. Then, it is prepared by operator 1 after its release date r_{20} and after the end of the `Sterilization` step. This job is packed in a delivery batch with jobs 1 and 4. This delivery batch is completed at $\max_{j \in \{1,4,20\}}(c_j^A)$ and is delivered by delivery man 1. The delivery trip is the first one of delivery man 1. The delivery man leaves the UBCO after the delivery batch completion time, then delivers jobs 1 and 4 at their assigned oncology unit ($u_1 = u_4$). The job 20 is then delivered at its oncology unit u_{20}. Finally, the delivery man comes back at the UBCO and is available for another trip.

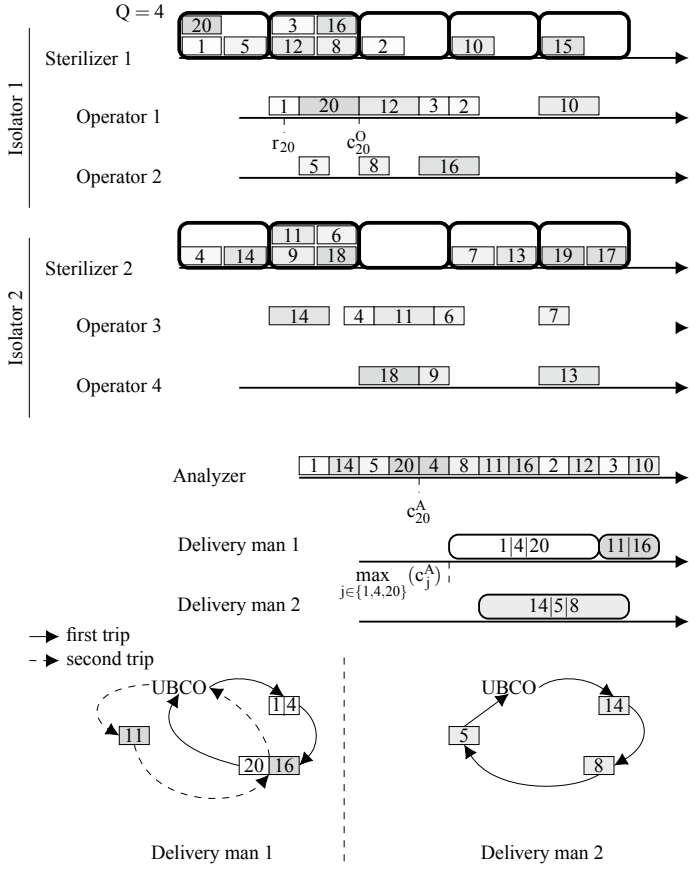


Fig. 2 Illustration of an instance of the problem: 2 isolators, 2 operators per isolator, 2 delivery men

3 Lower Bound

In order to propose a lower bound, we introduce *revised release dates* $\widetilde{r_j} = r_j + p_j^O$ and *revised due dates* $\widetilde{d_j} = d_j - t_{0,j}$ where $t_{0,j}$ represents the shortest possible transportation time to deliver the job j. $\widetilde{r_j}$ is the minimum possible value for c_j^O and $\widetilde{d_j}$ is the maximum possible value for c_j^A to deliver the job without tardiness.

Let consider the single machine scheduling problem with *revised release dates*, *revised due dates*, identical processing times ($p_j = p^A$) and total tardiness minimization, which can be denoted by $1|r_j, p_j = p|\sum T_j$ using the 3-field Graham notation of scheduling problems [10]. Any lower bound of this problem is a lower bound of our problem. Indeed, the $1|r_j = \widetilde{r_j}, p_j = p^A, d_j = \widetilde{d_j}|\sum T_j$ problem is equivalent to our problem considering a large number of operators and delivery men.

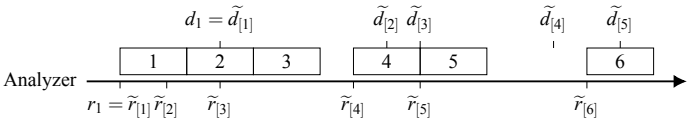


Fig. 3 Illustration of the lower bound definition

The proposed lower bound is computed as follows. We build a pseudo instance where job k has the kth shortest *revised release date* $\widetilde{r}_{[k]}$ and the kth shortest *revised due date* $\widetilde{d}_{[k]}$. The lower bound is given by the evaluation of sequence $(1, \ldots, |J|)$ of this pseudo instance (Figure 3 represents an instance with J = 6 jobs).

4 Multi-level Heuristic

The proposed multi-level heuristic is a constructive heuristic with multi-level decisions. The first decision level is the clustering of the jobs into delivery batches, the second decision level is the job assignment to a `Sterilization` batch, the third decision level is the `Preparation` scheduling, the fourth decision level is the `Control` scheduling and the last decision level is the delivery routing. Figure 4 represents the flowchart of the multi-level heuristic.

Clustering and sort: Each job is assigned to a cluster corresponding to a delivery batch by an Agglomerative Hierarchical Clustering method. This method needs a distance function between two clusters (i.e. batches). We define the distance between two batches B and B' as the maximum Euclidean distance between the jobs which is named "complete link" in [11]: $dist(B, B') = \max_{(j,j') \in B \times B'}(dist(j, j'))$. The Euclidean distance between jobs uses three dimensions: the oncology unit u_j

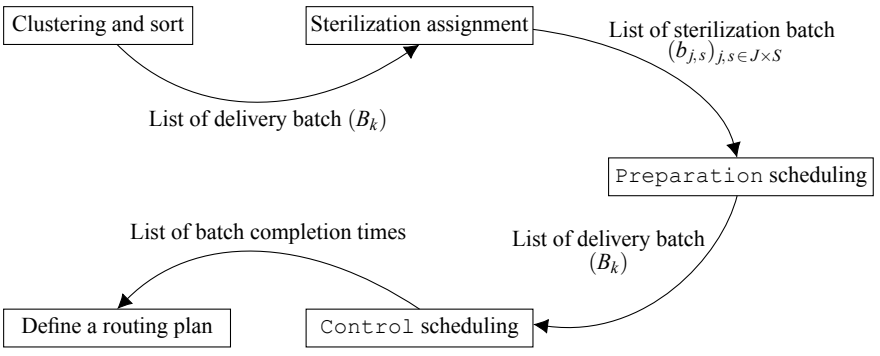


Fig. 4 Multi-level heuristic flowchart

location, the due date d_j and the *revised release date* $\widetilde{r}_j$. This clustering method aims to limit the waiting times before the delivery of each jobs and to reduce the duration of trips.

The number of delivery batches is given as an input. It is an important setting which is related to the maximum acceptable distance between two jobs from a same delivery batch.

The resulting delivery batches are then sorted using one sorting rule based on the due date of jobs of each batch. We propose three sorting rules by increasing order of:

1. $\min_{j \in B}(d_j)$, denoted MIN
2. $\text{mean}_{j \in B}(d_j)$, denoted MEAN
3. $\text{median}_{j \in B}(d_j)$, denoted MED.

A study of the impact of the number of delivery batches and the sorting rule is presented in Sect. 5.2.

Sterilization assignment: The sterilization assignment is an iterative method that consists in assigning jobs one by one to a sterilization batch. The jobs are sorted according to the sequence of delivery batches first (sorted by the selected sorting rule), then they are sorted in each delivery batch in $r_j + p_j^O$ increasing order. Then, each job is assigned to the last unfilled sterilization batch ending before the job's release date. If no such batch exists, the job is assigned to the first unfilled sterilization batch after the job's release date. In case of equality (i.e. two sterilizations batches complete at the same time) the job is assigned to the sterilization batch with the minimum sum of processing times p_j^O of jobs already assigned to that batch. This sterilization batch assignment allocates the jobs to an isolator and its set of operators.

`Preparation` **scheduling**: For each isolator the jobs of the sterilization batches are successively scheduled. All the jobs of a sterilization batch must been scheduled before starting to schedule the jobs of the next sterilization batch. The jobs of each sterilization batch are sorted according to the sequence of delivery batches first, then they are sorted by increasing release date r_j. The jobs are then scheduled as soon as possible on the first available operator.

`Control` **scheduling**: The jobs are sorted first by increasing preparation completion time c_j^O and in case of equality they are sorted according to the sequence of delivery batches. The jobs are successively scheduled as soon as possible.

Routing: Delivery batches are assigned iteratively to the first available delivery man. The trip of a delivery batch is constructed by the Nearest Neighbor heuristic [12]. The delivery man repeatedly delivers to the nearest oncology units until all jobs have been delivered.

5 Computational Experiments

In this section, the performances of the multi-level heuristic with various settings is evaluated on pseudo real life instances. We compare the proposed multi-level heuristic with the following 4-step method currently used at the UBCO:

1. `Preparation`: schedule the jobs by earliest release date first on the first available operator.
2. `Control`: schedule the jobs by earliest `Preparation` completion time first.
3. The clustering of jobs is done by the clustering algorithm described before with the following dimensions: `Control` completion times, due dates and oncology units.
4. The routes are defined by the Nearest Neighbor heuristic.

Note that the steps 3 and 4 are an approximation of the real life behavior of the delivery men who build their trips. This method is similar to a two-phase algorithm (scheduling then routing). This algorithm is called the Reference algorithm. The algorithms are implemented in Python language. Tests have been performed on an Intel(R) Core(TM) i5-7440HQ CPU @2.80GHz with 16 Go of Ram. The computation time of the two algorithms is about 1 second, which is acceptable for online use.

5.1 Datasets

The generation of 100 instances is inspired by the real case of the UBCO and have the following features:

- number of chemotherapy drugs: $|J| = 150$
- due dates: $\forall j \in J, d_j \in [9\,\text{h}, 18\,\text{h}]$
- release dates: $\forall j \in J, r_j \in [d_j - 10\,\text{h}, d_j - 50\,\text{min}]$
- number of isolators: $|I| = 4$
- production hours: [8 h, 18 h]
- `Sterilization` processing time: 15 min
- number of operators per isolator: 2 operators
- `Preparation` processing time: $p_j^O \in [5, 15]\,\text{min}$
- `Control` processing time: $p^A = 5\,\text{min}$
- number of oncology units: 60 units
- number of delivery men: $|V| = 3$
- all oncology units are within 35 min from the UBCO

For each interval, the distribution is uniform.

5.2 Parameters Setting and Evaluation

The number of delivery batches is a parameter that must be determined to find a good compromise:

- few delivery batches would result in long waiting times for the delivery men due to the time required to complete the batches
- a large number of delivery batches will generate a huge number of round trips which would imply waiting times for the completed batches.

To find the best compromise, we tested different number of delivery batches (from 15 to 35) on the 100 instances.

For each instance we determined the gap between the lower bound and the solution found. The gap is computed as $gap = \frac{h-Lb}{Lb}$ where h is the total tardiness found by the proposed multi-level heuristic or the Reference algorithm and Lb is the lower bound defined in Sect. 3.

Figure 5 represents the evolution of the mean gap for each tested number of delivery batches and for each sorting rule. Figure 5 shows that increasing the number of delivery batches increases the quality until a tipping point around 26 for the proposed multi-level heuristic and 24 for the Reference algorithm.

A delivery of 150 chemotherapy drugs in 26 trips means an average delivery batch size of 6 jobs and just under three trips per hour.

The best average gap found is around 52% which seems to be a big value. However, the lower bound, defined in Sect. 3, is clearly weak but is useful to compare the methods with a common reference. The lower bound weakness is mostly due to the assumption of an infinite number delivery men. It seems that the choice of the sorting rule do not have a big impact on the mean gap.

Fig. 5 Impact of the number of delivery batches - Mean Gap

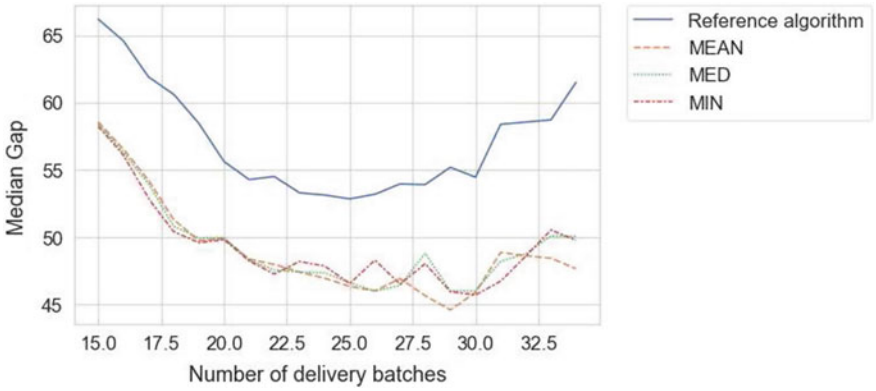


Fig. 6 Impact of the number of delivery batches—median gap

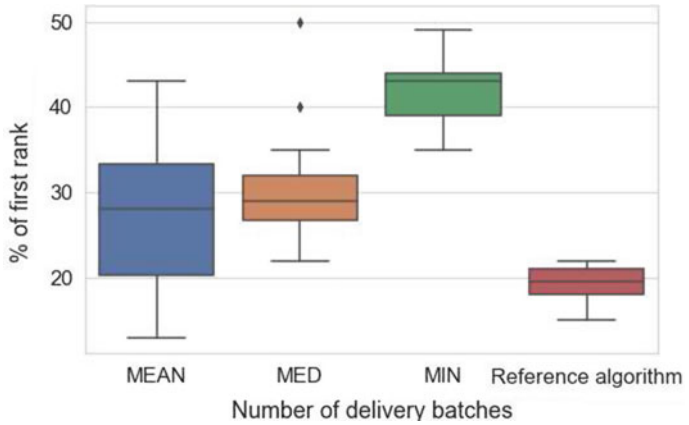


Fig. 7 Boxplots of the percentage of been the best heuristic i.e. first rank (number of delivery batches varying between 15 and 35)

Figure 6 presents the same results as Fig. 5 using the median gap instead of the mean gap. Around the tipping point, we note that the MEAN presents the best results on the median gap whereas there is not much difference between the three rules on the mean gap.

The median gap is smaller than the mean gap. One of the reason would be the existence of few outliers.

Figure 7 illustrates the quality difference of the sorting rules. The size of the boxplots shows that the Reference algorithm is the best algorithm around 20% of the time without depending of the number of delivery batches. However, it is the MIN rule which is the best most of the time (around 45%). This is in accordance with Fig. 5.

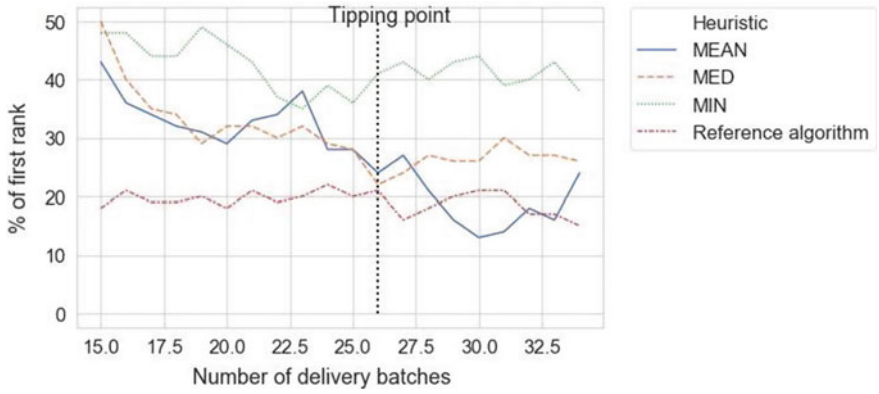


Fig. 8 Percentage of times a heuristic gives the best result i.e. first rank

Figure 8 illustrates the influence of the number of delivery batches on the heuristics ranking. We can see that the MIN rule is more is often the best heuristic whatever the number of delivery batches. While, the MEAN rule ranking is really dependent of the number of delivery batches.

6 Conclusions and Future Works

A real case of production and delivery of chemotherapy drugs was studied and a model of the problem was proposed. The model is based on an integrated version of a hybrid flow shop scheduling problem and a Multi-Trip Vehicle Routing Problem. To quickly solve the problem, we proposed a multi-level heuristic which schedules the production after taking into account the delivery part. The numerical experiments showed the efficiency of the proposed method compared to a Reference algorithm corresponding to the current planning method. A study of the multi-level settings illustrated the importance of the number of delivery batches and of the sorting rule.

Several research perspectives can be considered. First, the lower bound quality may be improved on the routing part. Besides, a local search at each scheduling level could improve the quality of the heuristic.

References

1. Mazier, A., Billaut, J.-C., Tournamille, J.-F.: Scheduling preparation of doses for a chemotherapy service. Ann. Oper. Res. **178**(1), 145–154 (2010)
2. Kergosien, Y., Tournamille, J.-F., Laurence, B., Billaut, J.-C.: Planning and tracking chemotherapy production for cancer treatment: a performing and integrated solution. Int. J. Med. Inf. **80**(9), 655–662 (2011)

3. Kergosien, Y., Gendreau, M., Billaut, J.-C.: Benders decomposition based heuristic for a production and outbound distribution scheduling problem with strict delivery constraints. Eur. J. Oper. Res. **262**(1), 287–298 (2017)
4. Abyaneh, F., Gholami, S.: A comparison of algorithms for minimizing the sum of earliness and tardiness in hybrid flow-shop scheduling problem with unrelated parallel machines and sequence-dependent setup times. J. Ind. Syst. Eng. **8**(2), 67–85 (2015)
5. Cattaruzza, D., Absi, N., Feillet, D.: The multi-trip vehicle routing problem with time windows and release dates. Transp. Sci. **50**(2), 676–693 (2016)
6. Azi, N., Gendreau, M., Potvin, J.-Y.: Optimization and approximation in deterministic sequencing and scheduling: a survey. Eur. J. Oper. Res. **202**(3), 756–763 (2010)
7. Amorim, P., Belo-Filho, M.A.F., Toledo, F.M.B., Almeder, C., Almada-Lobo, B.: Lot sizing versus batching in the production and distribution planning of perishable goods. Int. J. Prod. Econ. **146**(1), 208–218 (2013)
8. Belo-Filho, M.A.F., Amorim, P., Almada-Lobo, B.: An adaptive large neighbourhood search for the operational integrated production and distribution problem of perishable products. Int. J. Prod. Res. **53**(20), 6040–6058 (2015)
9. Devapriya, P., Ferrell, W., Geismar, N.: Integrated production and distribution scheduling with a perishable product. Eur. J. Oper. Res. **259**(3), 906–916 (2017)
10. Graham, R.L., Lawler, E.L., Lenstra, J.K., Rinnooy Kan, A.H.G.: An exact algorithm for a vehicle routing problem with time windows and multiple use of vehicles. Discr. Optim. II **5**, 287–326 (1979)
11. Murtagh, F.: A survey of recent advances in hierarchical clustering algorithms. Comput. J. **26**(4), 354–359 (1983)
12. Rosenkrantz, D.J., Stearns, R.E., Lewis, P.M.: An analysis of several heuristics for the traveling salesman problem. SIAM J. Comput. **6**(3), 563–581 (1977)

Primary Care

Acuity-Based Access Time Evaluation in Primary Care: A Case Study of an Ontario Clinic

Nazanin Aslani, Fariborz Fazileh, Donatus Mutasingwa and Daria Terekhov

Abstract Measuring *access* to primary care is complicated due to a variety of perspectives. In Ontario, Canada, one of the main metrics currently used to evaluate access is the proportion of patients who are able to obtain a same- or next-day appointment with a primary care provider. However, this metric does not accurately reflect patients who do not medically require same- or next-day access. In this study, we demonstrate the need for developing more detailed metrics which capture the urgency of needed care via a case study of an Ontario primary care clinic. Our results show that using the standard metric, the clinic's performance appears unsatisfactory, while using the more detailed acuity-based metrics, the clinic is shown to be performing well for non-urgent requests.

Keywords Access time · Performance evaluation · Primary care · Ontario · Non-urgent patient prioritization

1 Introduction

Primary care has been considered as a main element of a high-performing health system from the beginning of the 20th century [3]. Hence, it is important to evalu-

N. Aslani (✉) · D. Terekhov
Department of Mechanical, Industrial and Aerospace Engineering,
Concordia University, Montreal, QC, Canada
e-mail: na_aslan@encs.concordia.ca

D. Terekhov
e-mail: daria.terekhov@concordia.ca

F. Fazileh · D. Mutasingwa
Department of Family and Community Medicine,
University of Toronto, Toronto, ON, Canada
e-mail: fariborz.fazileh@gmail.com

D. Mutasingwa
e-mail: dmutasin@gmail.com

V. Bélanger et al. (eds.), *Health Care Systems Engineering*,
Springer Proceedings in Mathematics & Statistics 316,
https://doi.org/10.1007/978-3-030-39694-7_22

ate access to primary care, but this evaluation is difficult as *access* can be viewed and measured in multiple ways [19]. In Ontario, Canada, the main method for evaluation of access to primary care is through surveys, such as the Commonwealth Fund International Health Policy Survey and the Quality and Costs of Primary Care (QUALICOPC) Patient Experiences Survey (PES). However, these methods have two limitations. First, due to being survey-based, they are influenced by respondent perceptions and biases. Evaluating access only from patient perception can be misleading due to the weak relationship between care accessibility and patient perception of access [16]. Second, current metrics calculated from survey data, most notably the number of patients who obtain a same-day or next-day appointment, do not consider the urgency of the patient request. However, given the scarcity of healthcare providers in Canada [8], knowing the urgency of the patient could lead to more equitable and effective allocation of available physician time. The need for prioritization of patients in the setting of scarce resources is well-known in medical environments outside of primary care, such as emergency departments [13]; it has also recently been examined in the context of non-emergency settings, such as physiotherapy or rehabilitation services [10].

Our focus is the evaluation of *access time*, which is defined as the interval between the arrival of an appointment request and the scheduled time of appointment [6]. To address the above limitations, we argue for the evaluation of access time (a) through clinic data in order to overcome the potential subjectivity resulting from surveys, and (b) based on detailed metrics related to the various patient groups that primary care serves, akin to how emergency care performance is measured through different access time targets for patients of different acuity. While the first argument has already appeared in previous literature, see e.g., Rao et al. [20], we provide further evidence that there exist discrepancies between performance evaluation from objective data and patient surveys. Our second argument builds on work by Haggerty et al. [9], whose definition of accessibility considers the appropriateness of access time "to the urgency of the problem", and by Premji [18], who demonstrates a limitation of the most-prominent metric used for evaluation of Ontario primary care, i.e., the percentage of patients able to obtain a same-day or next-day appointment. Motivated also by the use of prioritization in other medical contexts with scarce resources, we propose to categorize patients in primary care according to the urgency of their request, a proposal which, to the best of our knowledge, has not been explored in the literature.

Our argument is illustrated by a case study of the Health for All (HFA) clinic, located in Markham, Ontario, Canada. Using a comprehensive data set of patient records from September 2017 to September 2018, we compute both the standard same-day/next-day access metric as well as the proportion of patients obtaining care with access times within targets appropriate to their level of acuity. Our results show a discrepancy between the two evaluation approaches: for non-urgent patients, using the standard metric, the clinic's performance appears unsatisfactory, while using the more detailed metrics, the clinic is shown to be performing well.

2 Current Performance Evaluation in Ontario Primary Care

Health Quality Ontario [11] evaluated performance of primary care in Ontario, Canada, based on the Primary Care Performance Measurement framework, which evaluates nine domains, including *access*. The specific criteria used to evaluate access by Health Quality Ontario [11] include, among others, *timely access at regular place of care*. Prior to 2017, Health Quality Ontario focused on the percentage of patients with same-/next-day access to a primary care provider, based on the question "The last time you were sick, how quickly could you see *any* doctor, nurse practitioner or physician assistant in this clinic?" In 2017, the latest year for which the performance report is currently available, a distribution of access times is presented, showing that 39.9%, 26.5%, 19.2% and 14.5% had access times of <2 days, 2–3 days, 4–7 days, and ≥ 8 days, respectively. The percentage of people waiting for ≥ 8 days ranged from 5.6% in the Central West region to 40.7% in the North West. At the same time, 67.6% of the respondent Ontarians reported that their wait for an appointment was "about right", 18.3% said "somewhat too long" and 14.1% said "much too long", with a range of 10.2% (Toronto Central) to 23.6% (North East) in the "much too long" category. Interestingly, the regions with the highest proportion of appointments with high access times are not necessarily the ones with the highest percentage in the "much too long" wait category. This observation supports our investigation: first, it demonstrates the impact of patient perceptions and expectations; second, it does not capture how many of the appointments were obtained within medically-warranted time frames.

The Quality and Costs of Primary Care (QUALICOPC) Patient Experiences Survey (PES) is a framework for evaluating care quality and outcomes in primary care [23]. For evaluating timely access to care, QUALICOPC-PES asks: "How many days did you wait for this visit from the time that you tried to make an appointment? For patients who had made an appointment for today's visit: Was it easy to get the appointment? Were you able to arrange an appointment with the doctor as soon as you wanted to?" For data collected between 2013 and winter 2014, 32% (of 1379) respondents said that they obtained a same-/next-day appointment; surprisingly, 87% (of 1536) said they were "able to arrange an appointment as soon as [they] wanted to" [15]. These differing statistics again motivate the need for further research into performance evaluation: for instance, were the complaints of patients who did not get a same-/next-day appointment and yet were satisfied less urgent than the ones who were dissatisfied?

The Canadian Institute for Health Information Commonwealth Fund Survey [5] in 2016 reports that "only 43% of Canadians were able to get a same- or next-day appointment at their regular place of care last time they needed medical attention". Furthermore, the study shows a discrepancy between the numbers reported by patients and physicians regarding access to primary care; in particular, the statistic provided for patients who say they could get a same- or next-day appointment for 2016 is 43% while the statistic provided for "primary care physicians who say most (at least 60%) of their patients can get a same- or next-day appointment" is 53% in 2016.

3 Literature Review

Primary Care Performance Evaluation Jones et al. [14] state that there exist two main approaches to evaluating primary care performance: appointment-data-based and survey-based. However, to the best of our knowledge, all existing studies in Canada are survey-based. Haggerty et al. [9] consulted primary healthcare (PHC) experts across Canada to formulate operational definitions of PHC attributes that should be evaluated in the Canadian primary healthcare setting. Importantly, the definition of *first-contact accessibility* as "the ease with which a person can obtain needed care (including advice and support) from the practitioner of choice within a time frame appropriate to the urgency of the problem" received a high level of physician consensus. We highlight in their definition the need to define a "time frame appropriate to the urgency of the problem": for acute patients, the *appropriate* time frame might indeed be same-day or next-day, but for patients requesting a periodic health exam, obtaining an appointment within several weeks of their requests is reasonable. Similarly, by analyzing the data from the QUALICOPC PES, Premji et al. [19] determined that the same-day/next-day access to primary care indicator does not match patients' perceptions of access to primary care.

Patient Classification in Primary Care The aim of patient classification is to prioritize patients objectively based on equitable criteria to ensure that patients with more urgent needs receive services first [7]. In the literature, the classification of primary care patients into different types has been considered in the context of improving primary care payment schemes (e.g., [21]) as well as improving patient access times. In the latter category, Balasubramanian et al. [1] study improving access by redesigning a physician panel based on the patients' age and presence of chronic disease; Ozen and Balasubramanian [17] use the number of simultaneous chronic conditions a patient has to classify patients in primary care and use as a predictor of the number of visits. In the capacity allocation literature, the majority of papers classify patients as urgent and non-urgent, e.g., Wang and Gupta [22]. However, based on our knowledge none of the existing access-time-focused literature in primary care explores the idea of performance evaluation based on acuity levels or prioritization to provide equitable access time.

4 The *Health for All* Clinic Background

The Health for All (HFA) clinic is adjacent to the Markham Stouffville Hospital, located in the City of Markham in the Regional Municipality of York within the Greater Toronto Area of Southern Ontario, Canada. It is located approximately 30 km northeast of Downtown Toronto. The HFA clinic is affiliated with the University of Toronto's Department of Family and Community Medicine. Residents spend their final two years of training with HFA to become family physicians; they see their own patients and go through clinical rotations at the hospital. HFA is a family health

team (FHT)—an inter-professional team of health care providers consisting of family doctors, nutritionists, social workers, and other professionals who provide comprehensive care to patients enrolled within the FHT.

Data We use a comprehensive data set from the HFA clinic for the period from September 2017 until September 2018. The data set contains 60,682 records listing the provider name, booking date, appointment date, appointment type, primary MD, no show, appointment detail, scheduled time, duration, arrival time and departure time. In order to prepare the data for analysis, we remove extra records and outliers. The records that we remove from consideration are those with negative access time (time of appointment in data set was before the time of booking); with no booking or appointment date; with doctor unavailability; home visits; evening and Saturday clinic appointments; and records that were labeled as "deleted", which generally corresponded to an appointment that was rescheduled for later. After removing these records, the remaining data set consists of 39,608 records. In order to choose an appropriate outlier labeling method, we considered whether the underlying data is symmetric or skewed [2]. Histograms of access times for all appointment types in this study were found to be right-skewed. Therefore the Adjusted Boxplot developed by Hubert and Vandervieren [12] is applied as an outlier labeling method. After removing the outliers using this method, we are left with 39,397 records in the data set.

Current Appointment Types at HFA The appointment classification system at HFA is based on patient complaints, i.e., the reason why the appointment has been requested. When a patients calls the clinic, an administrative clerk asks the patient for the reason of their request, their family doctor, their availability, etc., and suggests a time slot. To aid this process, the HFA clinic currently classifies appointments into 14 types: 11 of these are presented in the first column of Table 1, with example conditions given in the second column. Since the focus of this study is on the access time of patients who physically visit the clinic, we do not consider the 'Home-visit' appointment type in our analysis. Table 1 also omits the 'New Patient' category, an appointment for a patient to be introduced to their new family physician. Finally, another category of appointment is referred to as 'Blank'—a category that is meant to encompass all requests that do not clearly fit into the other 13 types. Importantly, 'Blank' appointments comprise 52% of all appointments at HFA; anecdotally, it appears that a large proportion of 'Blank' appointments request same-day or next-day appointments—however, this hypothesis cannot be confirmed by the current data set due to missing description of patient conditions in the data.

5 Proposed Acuity-Based Evaluation

As seen from Table 1, the complaints for which family medicine clinic appointments are requested vary widely in their nature and urgency. This observation suggests that evaluation of access time which considers the urgency of the patient request would

Table 1 Appointment types, example conditions and proposed access time bounds

App type	Condition example	Should be seen within
Follow-up	Soft tissue infection started on an antibiotic	1 week
	Sub-acute abdominal pain with blood work and imaging	2 weeks
	Hypertension with recent medication change	4 weeks
	Thyroid medication dose modification	12 weeks
Injection	Travel medicine injection	1 week
	First visit of a patient who has not started	2 weeks
	Their routine immunization in their infancy	
	Intra-articular injection	4 weeks
	Repeat intra-articular injection	12 weeks
Mental health	Anxiety	2 weeks
	Depression started on new medication	4 weeks
	Mental health condition responded moderately to medication change	12 weeks
First Pre-Natal	Appt requested 1–2 weeks before week 8th of pregnancy	2 weeks
	Appt requested 3–4 weeks before week 8th of pregnancy	4 weeks
Pre-Natal	After week 28th of pregnancy	2 weeks
	Week 12th till week 28th of pregnancy	4 weeks
Well-baby	Baby should be seen in 4, 6, 9, 12, 15, 18m	12 weeks
Pre-op	Request 1–2 weeks before operation	2 weeks
	Request more than 2 weeks before operation	4 weeks
Diabetic management	Patient should be seen every 3 months	12 weeks
Child physical	Child should be seen in 2, 4, 6, 16 year	12 weeks
Periodic health exam	Annual visits for chronic illness and/or health issues	12 weeks
Driver's physical	Every 5 years if < 46 y/o; 3 years if 46–64 y/o; annually if > = 65 y/o	12 weeks

give a more accurate representation of the performance of primary care and lead to more equitable allocation of appointments to patients.

For example, consider patient A who has diabetes and is calling for his/her regular three-month appointment. In this case, as long as the appointment is scheduled close to the target time of three months from the previous check-up, the care needs of the patient and risks of adverse health outcomes are appropriately addressed. In contrast, consider patient B who calls for a follow-up appointment for an eight-month-old baby with five days of fever and a possible viral illness diagnosis. Patient B needs to be seen in the clinic on the same day to ensure they are not at risk for significant adverse health outcomes. Table 1 provides additional examples of conditions of various urgency—the third column of Table 1 gives potential access time upper bounds obtained through discussions with two practitioners from HFA. For the majority of appointments, the access time upper bounds can be defined as the maximum time from the arrival of the patient request until the patient is seen; however, for some periodic appointments such as physicals, or for follow-up appointments, the upper bound is defined as the time between the previous appointment and the next one (e.g., the time between an appointment to resolve an initial complaint and a follow-up appointment to discuss the effectiveness of the care received).

As described in Sect. 2, current methods of primary care performance evaluation in Ontario do not take into account the urgency of the patient request; furthermore, as shown in Sect. 3, a detailed classification of patient urgency in primary care is done for billing purposes or, for the purposes of appointment allocation, is usually limited to two types. Such a performance evaluation approach in primary care is in contrast to performance evaluation in emergency care. In emergency care, the Canadian Triage Acuity Scale (CTAS) is employed to classify patients into five categories in order to "triage patients according to acuity, risk, and care needs based on their presenting signs and symptoms" and to "ensure that the sickest and highest risk patients are seen first when ED capacity has been exceeded" [4]. In addition, ED managers can use CTAS to "capture and analyze ED patient visit based on volume, acuity and by CEDIS presenting complaint" [4].

Considering Table 1 and using an analogy with the CTAS system in emergency care, we propose a five-level classification of patients in primary care based on their urgency, varying from urgent patients that need to be seen within one day to routine patiens that should be seen within 12 weeks, as shown in Table 2. In Table 3 we show how the current HFA appointment types map to our proposed acuity levels. Given the acuity levels defined in Table 2, we can now evaluate the performance of primary care with respect to the urgency of the appointment, by finding the proportion of patients in each acuity category that obtained an appointment within the proposed access time upper bound (referred to as *access time target*).

Table 2 Proposed acuity levels and corresponding access time targets

Acuity level	Description	Should be seen within
1	Same day urgent	1 day
2	Same week urgent	1 week
3	Two weeks non-urgent	2 weeks
4	Four weeks non-urgent	4 weeks
5	Routine	12 weeks

Table 3 Acuity levels per appointment type at HFA

Appointment type	Acuity level	Appointment type	Acuity level
Periodic health exam	5	Injection	2–5
Child physical	5	New patient	1–5
Diabetic management	5	Pre-Op assessment	3.4
Driver's physical	5	Pre-Natal	3.4
First Pre-Natal	3.4	Well baby	5
Follow up	2–5	Blank	1–5
Mental health	3–5		

6 Measuring Timely Access to Care

We first analyze HFA clinic access time performance based on the indicators from the Ontario Ministry of Health and Long-term Care (MOHLTC), followed by performance evaluation based on the acuity levels defined in Table 2.

Evaluation based on MOHLTC indicator MOHLTC considers same-/next-day access as one of the major access time indicators [11]. Patients at HFA reported "same-day or next-day access when they are sick" as 54% and 52% for 2016 and 2017, respectively. Calculating the same metric from appointment data, we see that only 32% (out of 38,367) and 31% (out of 39,608) of patients had same-/next-day appointments for 2016/17 and 2017/18 data, suggesting that the surveyed sample consisted of patients that either actually had faster access times or were influenced by an overall favourable perception of their experience at the clinic. In addition, it is not clear whether patients who are over-due to visit for a chronic condition would classify their request as being in the category "when they are sick". The discrepancy in these numbers supports our argument for objective evaluation from appointment data systems rather than surveys. Furthermore, the values 31 and 32% seem to indicate sub-par performance of the HFA clinic.

In Table 4, we present the cumulative percentage of same-/next-day and same-week appointments for each patient class based on our 2017/18 data set. Looking at the percentages by current appointment types, we see that the percentage of

Table 4 % of patient types with same-day/next-day/same-week appointments

Appointment type	Access time			
	Total #	% same d (%)	% next d (%)	% same w (%)
Periodic health exam	2309	10	13	14
Child physical	910	8	10	12
Diabetic management	1906	6	8	9
Driver's physical	38	18	26	55
First Pre-Natal	136	7	9	13
Follow up	5978	19	24	27
Mental health	965	10	13	16
Injection	1499	48	53	55
New patient	1470	13	17	18
Pre-Op assessment	247	19	24	28
Pre-Natal	708	7	9	11
Well baby	2711	6	9	11
Blank	20,731	37	42	46
All types	39,608	27	31	35

same-/next-day visits ranges from 8% (diabetic management) to 53% (injection) which again seems to suggest lack of timely access to care for many HFA patients. Additionally, these results show substantial variability among patient classes.

Evaluation based on acuity-based indicators Figures 1, 2 and 3 show access time histograms for three appointment types, 'Follow up', 'Diabetic Management' and 'Periodic Health Examination'. Importantly, from these histograms we can observe that access time behaviour of different appointment types is quite different. The histogram for 'Follow up' shows a zero-inflated distribution with a long right tail that extends beyond 120 days; the histogram for 'Diabetic Management' is bi-modal, with peaks at two weeks and 90 days; for the 'Periodic Health Exam', the majority of appointments happen within one month, with a mode of 1 week. In addition to the differences in behaviour among the appointment types, in all three histograms, we see substantial variability in access times among patients within each appointment type. For 'Follow up' appointments, there are peaks, of diminishing magnitudes, at the end of every week, suggesting the existence of multiple "sub-types" in the 'Follow up' category, that is, patients whose follow-up appointments should be in one week, two weeks, etc. For diabetic appointments, the peak at 90 days matches the suggested interval between two regular visits for a patient with diabetes. For the periodic health exam, the high number of patients being scheduled within one day and one week is particularly surprising, given that these are non-urgent appointments.

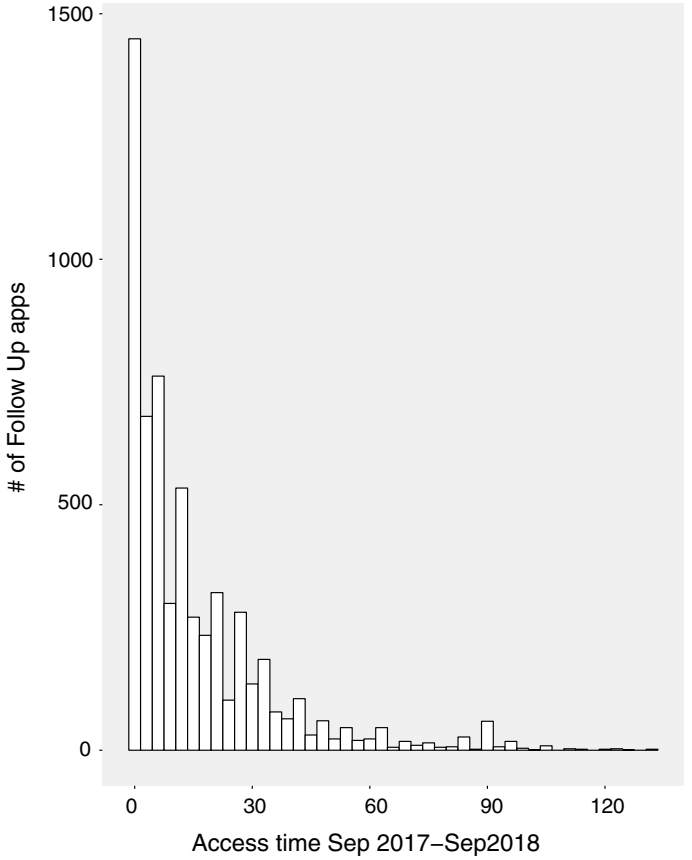


Fig. 1 Access times (days) for follow-up

We now focus on the appointment types identified in Table 3 as having a single acuity level of 5. For these appointment types, we present additional statistics as well as the evaluation of their access times according to our proposed metric in Table 5. In particular, we observe that the majority of patients requesting 'Periodic Health Exam', 'Child Physical', 'Diabetic Management' and 'Well Baby' appointments were able to obtain them within the suggested upper bounds on access time, demonstrating that HFA performs very well for non-urgent appointments, which is not obvious from standard metrics, and in fact contradicts the conclusion one would make from looking at same-/next-day metrics over all appointment types or even for these specific non-urgent appointment types. Furthermore, we can observe that a large number of acuity level 5 patients obtained a same-/next-day appointment, despite being of low urgency. In a setting with scarce resources, this observation effectively implies that same-/next-day appointment times are not being used effectively by the clinic, and can inform new allocation policies.

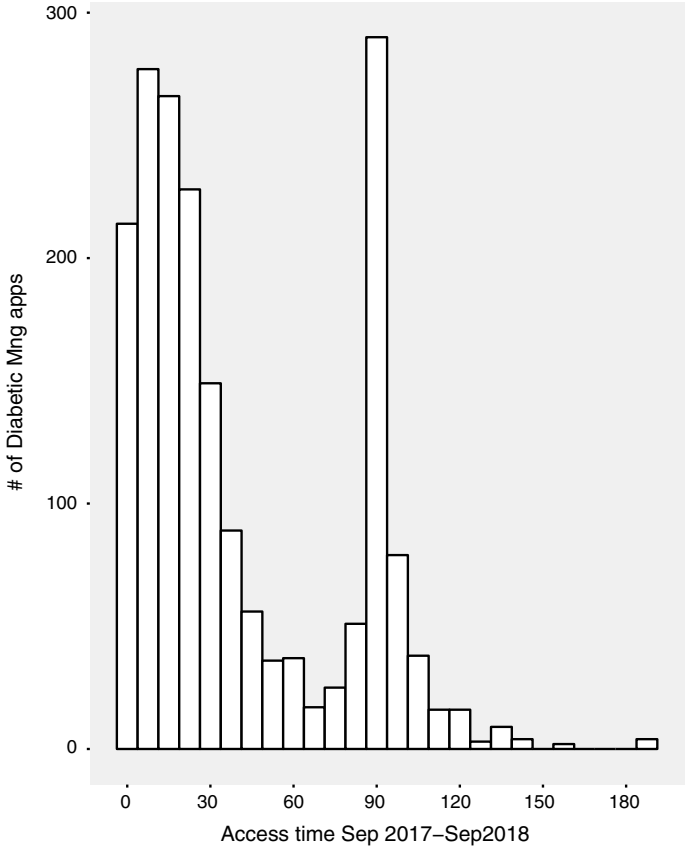


Fig. 2 Access time (days) for diabetic management

Discussion A limitation of our study is that we could not evaluate the performance of the clinic for appointment categories with multiple acuity levels (see Table 3) due to the lack of data regarding patient complaints. Furthermore, we note that the acuity level definitions proposed in Table 2 constitute a proposal that we expect will be refined by researchers and practitioners in future work. Given these (or refined) acuity level definitions, the list of conditions and corresponding access time targets would require substantial work from clinicians, similarly to the work involved in defining and updating CTAS. We observe that any potential implementation of our proposal in practice immediately raises the question of triage in primary care. However, despite the work required for formalizing a prioritization scheme and triage procedures, doing so may lead to more equitable resource allocation in primary care, which so far remains a setting with limited resources.

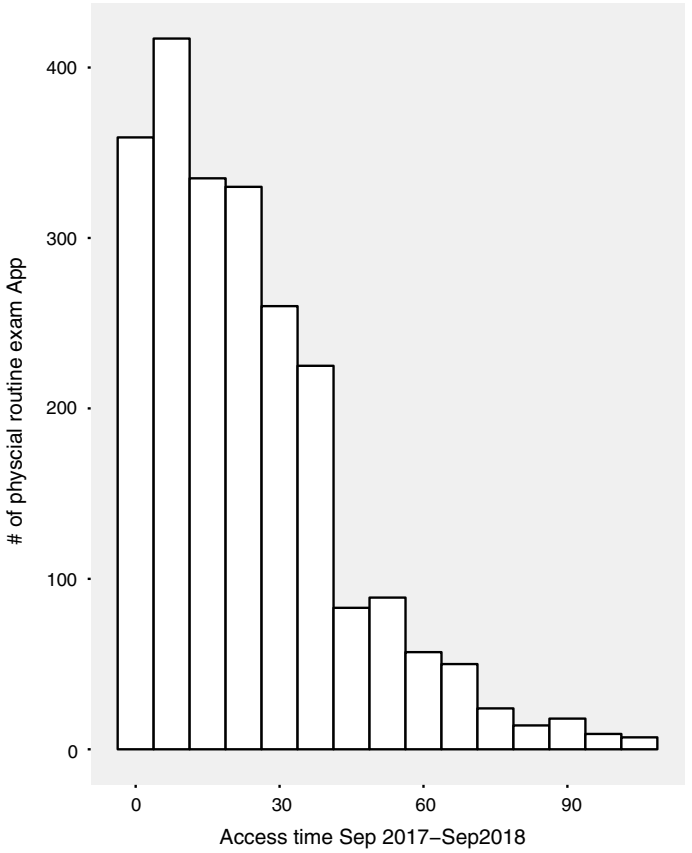


Fig. 3 Access time (days) for periodic health exam

Table 5 Summary of access times for acuity level 5 appointment types at HFA

Appointment type	Range (w)	Mean (w)	Median (w)	Mode (w)	% seen within access target (%)
Periodic health exam	0–15	3.44	3	1	98
Child physical	0–28	4.05	3	2	95
Diabetic management	0–27	5.94	4	2	91
Well baby	0–25	4.93	4	1	91

7 Conclusion

In this paper, we contrasted performance evaluation of primary care using a traditional metric and new metrics via a case study of a primary care clinic, namely the Health for All Clinic in Markham, Ontario, Canada. Inspired by CTAS classification used in emergency departments, we defined acuity levels for the conditions relative to each appointment type in our case study of primary care. Future work needs to focus on developing the exact list of conditions for different acuity levels as well as evaluation of other clinics.

References

1. Balasubramanian, H., Banerjee, R., Denton, B., Naessens, J., Stahl, J.: Improving clinical access and continuity through physician panel redesign. J. Gen. Intern. Med. **25**(10), 1109–1115 (2010)
2. Ben-Gal, I.: Outlier detection. In: Data Mining and Knowledge Discovery Handbook, pp. 131–146. Springer (2005)
3. Bitton, A., Ratcliffe, H.L., Veillard, J.H., Kress, D.H., Barkley, S., Kimball, M., Secci, F., Wong, E., Basu, L., Taylor, C., et al.: Primary health care as a foundation for strengthening health systems in low-and middle-income countries. J. Gen. Intern. Med. **32**(5), 566–571 (2017)
4. Canadian Association of Emergency Physicians. The Canadian Triage & Acuity Scale (CTAS). http://ctas-phctas.ca/?page_id=17. Accessed 2nd Mar. 2019
5. Canadian Institute for Health Information. Commonwealth fund survey 2016: Chartbook. Technical report (2017)
6. CIHI.: National health expenditure trends, 1975 to 2016. Canadian Institute for Health Information (2016)
7. Déry, J., Ruiz, A., Routhier, F., Gagnon, M.-P., Côté, A., Ait-Kadi, D., Bélanger, V., Deslauriers, S., Lamontagne, M.-E.: Patient prioritization tools and their effectiveness in non-emergency healthcare services: a systematic review protocol. Syst. Rev. **8**(1), 78 (2019)
8. Globerman, S., Barua, B., Hasan, S.: The Supply of Physicians in Canada: Projections and Assessment (2018). http://www.fraserinstitute.org
9. Haggerty, J., Burge, F., Lévesque, J.-F., Gass, D., Pineault, R., Beaulieu, M.-D., Santor, D.: Operational definitions of attributes of primary health care: consensus among Canadian experts. Ann. Family Med. **5**(4), 336–344 (2007)
10. Harding, K., Taylor, N.:. Triage in non-emergency services. In: Patient Flow: Reducing Delay in Healthcare Delivery, pp. 229–250. Springer (2013)
11. Health Quality Ontario. Measuring up 2018. Technical report (2018)
12. Hubert, M., Vandervieren, E.: An adjusted boxplot for skewed distributions. Comput. Stat. Data Anal. **52**(12), 5186–5201 (2008)
13. Iserson, K.V., Moskop, J.C.: Triage in medicine, part i: concept, history, and types. Ann. Emerg. Med. **49**(3), 275–281 (2007)
14. Jones, W., Elwyn, G., Edwards, P., Edwards, A., Emmerson, M., Hibbs, R.: Measuring access to primary care appointments: a review of methods. BMC Family Pract. **4**(1), 8 (2003)
15. Laberge, M., Pang, J., Walker, K., Wong, S., Hogg, W., Wodchis, W.P., et al.: QUALICOPC (Quality and Costs of Primary Care) Canada: a focus on the aspects of primary care most highly rated by current patients of primary care practices (2014)
16. Llanwarne, N.R., Abel, G.A., Elliott, M.N., Paddison, C.A.M., Lyratzopoulos, G., Campbell, J.L., Roland, M.: Relationship between clinical quality and patient experience: analysis of data from the english quality and outcomes framework and the national GP patient survey. Ann. Family Med. **11**(5), 467–472 (2013)

17. Ozen, A., Balasubramanian, H.: The impact of case mix on timely access to appointments in a primary care group practice. Health Care Manag. Sci. **16**(2), 101–118 (2013)
18. Premji, K.: In Our Rush to Offer "McMedicine", Do We Even Know What Patients Really Want? (2018). https://healthydebate.ca/opinions/same-day-access-family-doctor
19. Premji, K., Ryan, B.L., Hogg, W.E., Wodchis, W.P.: Patients' perceptions of access to primary care: Analysis of the QUALICOPC patient experiences survey. Can. Family Phys. **64**(3), 212–220 (2018)
20. Rao, M., Clarke, A., Sanderson, C., Hammersley, R.: Patients' own assessments of quality of primary care compared with objective records based measures of technical quality of care: cross sectional study. BMJ **333**(7557), 19 (2006)
21. Starfield, B., Weiner, J., Mumford, L., Steinwachs, D.: Ambulatory care groups: a categorization of diagnoses for research and management. Health Serv. Res. **26**(1), 53 (1991)
22. Wang, W.-Y., Gupta, D.: Adaptive appointment systems with patient preferences. Manuf. Serv. Oper. Manag. **13**(3), 373–389 (2011)
23. Wong, S.T., Chau, L.W., Hogg, W., Teare, G.F., Miedema, B., Breton, M., Aubrey-Bassler, K., Katz, A., Burge, F., Boivin, A., et al.: An international cross-sectional survey on the quality and costs of primary care (QUALICO-PC): recruitment and data collection of places delivering primary care across Canada. BMC Family Pract. **16**(1), 20 (2015)

Blood System

Uncertainty in the Blood Donation Appointment Scheduling: Key Factors and Research Perspectives

Ettore Lanzarone and Semih Yalçındağ

Abstract We consider the management of a blood collection center, which includes the features of both a production system and a service provider. In particular, we analyze the scheduling of donors and the related appointment system, addressing the so-called Blood Donation Appointment Scheduling (BDAS) problem. From the production system viewpoint, the requirement is to balance the production of whole blood units between days in order to meet the requirement of a constant supply of blood to hospitals and transfusion centers; from the service provider perspective, appointments reduce waiting times and improve the service quality perceived by donors. Thus, the goals of the BDAS are to guarantee a quite constant production of whole blood units and to reduce physicians' overtimes while including appointments and free slots for donors without a reservation. A framework for the BDAS problem has been recently proposed, in which slots are first preallocated to the different blood types and then assigned to the donors when they call to make a reservation. However, this framework refers to a deterministic setting in which all input parameters are assumed to be known in advance. On the contrary, the BDAS problem is stochastic in nature and includes stochastic parameters that must be predicted from historical data. In this paper, we first analyze the possible uncertainty sources to determine the most critical ones. Then, we propose research directions to properly include them in the BDAS framework, considering both stochastic programming and robust optimization methodologies.

Keywords Blood Donation · Appointment Scheduling · Uncertainty Factors · Robust Optimization · Stochastic Programming

E. Lanzarone
CNR-IMATI, Milan, Italy
e-mail: ettore.lanzarone@cnr.it

S. Yalçındağ (✉)
Department of Industrial and Systems Engineering, Yeditepe University, Istanbul, Turkey
e-mail: semih.yalcindag@yeditepe.edu.tr

V. Bélanger et al. (eds.), *Health Care Systems Engineering*,
Springer Proceedings in Mathematics & Statistics 316,
https://doi.org/10.1007/978-3-030-39694-7_23

1 Introduction

Blood is necessary to perform several treatments and surgeries, and blood supply is a key problem for all health care systems. However, while the demand for blood is usually high (10 million units per year in the USA, 2.1 in Italy, and 2 in Turkey), blood is a limited resource as it cannot be produced in laboratory but only withdrawn from healthy individuals. Thus, blood is usually collected from volunteer donors all around the world, and the collected blood units are used to satisfy the demand from hospitals and transfusion centers. Blood units are managed by the so-called Blood Donation (BD) supply chain. Together with the high demand to satisfy, one of the main criticalities in managing the BD chain is the short shelf life of the blood units (about 6 weeks), which limits the time period between blood collection and utilization. Thus, the BD system must effectively manage the collection of blood units with respect to the demand, in order to avoid blood shortage and wastage. This is a difficult task, which involves several decision levels, and optimization tools are highly recommended to guarantee effectiveness and efficiency of the BD supply chain.

The BD supply chain can be classified in different ways [6, 25, 35]. According to [35], it is divided in four phases: *collection*, *storage*, *transportation* and *utilization*. In this work, we focus on the blood collection phase, which is the first and the most crucial phase of the entire BD chain. In fact, the management of blood collection impacts all other phases, as an unbalanced number of donations between days with respect to the demand might result in blood shortage or wastage despite a proper management of the other phases. There is no doubt that an increased number of donations has positive consequences on the BD system, but in any case an unbalanced arrival of donors could determine alternate periods of wastage and shortage. Hence, it is crucial to systematically receive donations through a reservation system that balances the daily production of blood units with respect to the demand.

Moreover, as blood collection centers include the features of both a production system and a service provider, a proper reservation system can also reduce waiting times for donors and improve the perceived quality of the service, thus making the potential donors more willing to donate regularly.

Though most of the donors are willing to exploit reservation systems (either manual or based on an optimization tool), it is unavoidable that some other donors do not reserve the donation in advance. As donation is a voluntary activity in several Countries, blood collection centers accept these non-booked donors not to lose blood units and not to discourage future donations from these donors. However, the presence of non-booked donations makes it difficult to perfectly align blood production with demand, as random arrivals of these donors generally cause production imbalances between days. Another issue related with the appointment scheduling is the no-show of booked donors, who make a reservation but do not show up on the scheduled day. As in the case of non-booked donors, these no-shows cause production imbalances between days. Thus, optimization tools for BD appointment scheduling should take into account these randomness sources while planning donation schedules.

Despite its importance, the blood collection phase is marginally addressed in the management literature of the BD supply chain [3, 5] and, even considering the available works that deal with this phase, uncertainty is still neglected.

In this work, in collaboration with a real Italian provider, we discuss the uncertainty sources that affect blood collection to determine the most critical ones and their impact. Then, we consider the deterministic Blood Donation Appointment Scheduling (BDAS) framework proposed in [4] for the collection of whole blood units and propose research directions to adequately include uncertainty in this framework.

The remainder of this chapter is organized as follows. A brief literature review on BD supply chain and optimization approaches that include uncertainty is presented in Sect. 2. Then, the BDAS framework taken as reference is presented in Sect. 3, while the uncertainty sources, the methodologies to address them and future research directions are discussed in Sect. 4. Finally, concluding remarks are reported in Sect. 5.

2 Literature Review

The BD supply chain has been extensively studied, as discussed in recent reviews available in the literature [3, 5, 25]. These works show that most of the studies focus on storage and utilization phases, while the collection has not been adequately considered.

Baş et al. [4] were among the first who contributed to the development of a decision support tool for blood collection via the definition of the BDAS problem and the development of a two-phase solution framework. Though this work filled an important gap in the literature, it did not include any uncertainty source related to blood collection or the rest of the BD supply chain. However, uncertainty should be considered to create more effective tools and decision support systems for the BD system. In the following, we overview all recent works that deal with uncertainty in the BD supply chain, which are also classified in Table 1 according to the addressed BD phase, the included uncertainty source and the adopted optimization approach.

Most of the works that include uncertainty address the inventory management (storage phase), the transportation/distribution problem and the design of the overall supply chain network. This also reflects the trend of all available works, including those that neglect uncertainty [3, 5]. As for inventory management, Van Dijk et al. [36] proposed a multi-step procedure based on Markov dynamic programming, Zhou et al. [40] developed a stochastic dynamic programming approach, Gunpinar and Centeno [16] proposed a stochastic programming model, Dillon et al. [12] proposed a two-stage stochastic programming model, Najafi et al. [24] developed a chance-constraint programming model, and Puranam et al. [27] developed a stochastic dynamic programming approach. As for distribution, Hemmelmayr et al. [17] developed a stochastic programming model, Kazemi et al. [21] proposed a robust possibilistic programming approach, Akhavan Niaki [2] developed a stochastic programming model, and Jafarkhan and Yaghoubi [20] proposed a robust optimization model.

Table 1 Literature works that include uncertainty in the management of the BD supply chain

References	BD supply chain phase	Uncertainty source			Optimization approach
		Demand	Supply	Other	
Akhavan Niaki et al. [2]	Distribution	✓			Stochastic programming
Dillon et al. [12]	Inventory management	✓	✓		Stochastic programming
Ensafian and Yaghoubi [13]	Supply chain network design	✓			Robust optimization
Ensafian et al. [14]	Supply chain network design	✓			Stochastic programming
Fazli-Khalaf et al. [15]	Supply chain network design			✓	Robust possibilistic chance-constraint programming
Gunpinar and Centeno [16]	Inventory management	✓			Stochastic programming
Hemmelmayr et al. [17]	Distribution			✓	Stochastic programming
Jabbarzadeh et al. [19]	Collection/location-allocation	✓	✓		Robust optimization
Jafarkan and Yaghoubi [20]	Distribution	✓	✓		Robust optimization
Kazemi et al. [21]	Distribution	✓			Robust possibilistic programming
Najafi et al. [24]	Inventory management	✓	✓		Chance-constraint programming
Osorio et al. [26]	Collection	✓			Stochastic programming
Puranam et al. [27]	Inventory management		✓		Stochastic dynamic programming
Rabbani et al. [28]	Collection/location		✓		Fuzzy mathematical programming
Ramezanian and Behboodi [29]	Collection/location-allocation	✓			Robust optimization
Salehi et al. [31]	Supply chain network design	✓			Robust stochastic programming
Samani et al. [32]	Supply chain network design	✓			Stochastic and possibilistic programming
Van Dijk et al. [36]	Inventory management		✓		Markov dynamic programming
Zahiri et al. [39]	Collection/location-allocation	✓	✓		Robust possibilistic programming
Zahiri and Pishvaee [37]	Supply chain network design			✓	Robust possibilistic programming
Zahiri et al. [38]	Supply chain network design			✓	Stochastic programming
Zhou et al. [40]	Inventory management	✓			Stochastic dynamic programming

Other works include uncertainty while integrating more than one phase (*Supply Chain Network Design*). In this group, Ensafian et al. [14] developed a stochastic programming model, Ensafian and Yaghoubi [13] presented a robust optimization approach, Fazli-Khalaf et al. [15] proposed a robust possibilistic flexible chance constraint programming model, Salehi et al. [31] developed a robust two-stage multi-period stochastic model, Zahiri and Pishvaee [37] designed two bi-objective robust possibilistic programming models, Samani et al. [32] proposed stochastic programming and possibilistic programming approaches, and Zahiri et al. [38] developed a multi-stage stochastic programming approach.

Finally, few contributions address uncertainty in the collection phase; however, they are mostly related to location/allocation problems and not to the appointment scheduling problem. Among them, Jabbarzadeh et al. [19] developed a robust optimization model to support blood facility location and allocation decisions, Zahiri et al. [39] proposed a robust possibilistic programming model to determine the best locations for fixed and temporary blood facilities, Ramezanian and Behboodi [29] proposed a robust optimization approach for the location-allocation problem, Rabbani et al. [28] addressed a fuzzy mathematical programming model for the mobile blood collection system, and Osorio et al. [26] proposed a multi-objective stochastic programming model for collection technology selection and donor allocation problems.

From this literature classification, we may observe a lack of robust appointment scheduling systems that include uncertainty for the BD collection phase. This gap could be filled by incorporating the uncertainty sources into the existing BDAS deterministic framework [4] or in other deterministic models.

Three main approaches to include uncertainty in optimization problems are studied in the literature and applied in practice [1]. Stochastic programming considers the uncertain parameters as random variables with a known probability distribution, and the problem is solved including a set of scenarios generated with those distributions. Distributionally robust optimization and ambiguous chance-constrained approaches assume that the probability distributions of the uncertain parameters lie within a known family of distributions, and the problem is solved for the worst-case realization compatible with the family. Robust optimization assumes that the uncertain parameters belong to a given convex set (named *uncertainty set*) without any assumption on the probability distributions over the set, and the problem is solved guaranteeing that the solution remains feasible for all parameter values within the uncertainty set. We will also focus on these methods in our discussion.

3 Blood Donation Appointment Scheduling (BDAS) Problem

The BDAS architecture proposed by Baş et al. [4] consists of an offline preallocation of the time slots for donation based on blood type, and an online allocation of each incoming reservation request to a suitable slot preallocated for the donor's blood

type. The offline preallocation is modeled as a deterministic mixed integer linear programming model, which is solved at a fixed frequency (e.g., once per day); the online allocation consists of a prioritization policy of the slots, to propose the best slots each time a donor calls to make the reservation. The number of preallocated slots converted in actually reserved slots is fed back to the next run of the preallocation model, and the process is repeated so on.

The preallocation model of [4] considers a number of slots x_t^b to preallocate (decision variables) and a number of already allocated slots a_t^b from previous reservations (parameters) for each day t of the time horizon T and each blood type $b \in B$. The total number of slots to preallocate $\sum_t x_t^b + a_t^b$ for each $b \in B$ is forced to lie between $(1-\varepsilon)d_b$ and $(1+\varepsilon)d_b$, where d_b is the expected number of booked donors over T and ε is a flexibility parameter. An amount of slots n_t^b is then left empty at day $t \in T$ for non-booked donors of blood type $b \in B$. Thus, the overall number of planned donations for blood type $b \in B$ at day $t \in T$ is $y_t^b = x_t^b + a_t^b + n_t^b$. Moreover, all days $t \in T$ are divided in a set K of periods, and the service time required in period $k \in K$ of day $t \in T$ above the capacity c_{tk} is defined as dispersion penalty p_{tk}. The primary goal of the BDAS preallocation model is to minimize, over the days $t \in T$ and for each blood type $b \in B$, the absolute variation of y_t^b with respect to its average value over the days, which is denoted by z_t^b. The secondary goal is to minimize a weighted sum of p_{tk} over $t \in T$ and $k \in K$. Further details are provided in [4].

4 Uncertainty Sources and Possible Methodologies

In the above described preallocation model, all parameters are assumed to be deterministic and their values are taken from historical expected values; however, several uncertainty factors may affect the actual values of the parameters and, thus, the quality of a solution when applied in practice. On the one hand, it is impossible to include all of them in the problem, as common in health care where the complexity of both problem and data is high. On the other hand, the most critical factors should be included in order to avoid unexpected issues when a solution is executed.

The most critical parameters in the BDAS are those that directly affect the filling of slots and, therefore, the number of produced units. In particular, with respect to the planned layout of slots, variations may occur due to the random arrivals of non-booked donors and the unexpected no-show of booked donors, who do not show up on the scheduled day without notifying. Thus, the most critical uncertain parameters are the random arrivals of non-booked donors and the effective number of already allocated slots (parameters n_t^b and a_t^b in the BDAS preallocation model, respectively).

In the following, we separately address these two parameters. Other secondary uncertain parameters are the demand for blood, the availability of physicians and machinery for blood withdrawing, and the health status of donors, as it may happen that a donor must be excluded from donation or a produced blood unit must be discarded. This classification and prioritization of the uncertainty factors has been discussed and outlined with staff of the Associazione Volontari Italiani Sangue (AVIS), the largest network of BD collection centers in Italy, and in particular with its Milan Department.

4.1 Uncertain Arrivals of Non-booked Donors

This concerns the parameters n_t^b in the BDAS preallocation model. We suggest two possible approaches to deal with their uncertainty, either with stochastic programming or with robust optimization.

A trade-off between the robustness level, which refers to the feasibility of the solution over the possible parameter realizations, and the solution quality, which concerns the deterioration of the planned solution when applied to the uncertain system, must be taken into account [1]. On the one hand, a non-conservative solution might easily become infeasible even for small deviations from the nominal/deterministic problem; on the other hand, very conservative solutions may turn out to be expensive for likely scenarios. Thus, the adopted approach and the level of protection against the uncertain parameters must be tailored based on the specific BDAS problem and the stochastic information available for the uncertain parameters.

Stochastic programming requires a deep knowledge of the problem to get the probability distributions. Its main advantage is not to produce over-conservative solutions, as the level of protection is fine-tuned by the distributions themselves; however, the resulting problem can be difficult to solve and, if the distributions are not reliable, the quality of the solution may be low. On the contrary, the solutions produced via robust optimization may be over-conservative, but the possibility to give a specific shape to the uncertainty set and limit the parameters' movement within it allow us to adjust the level of robustness, and the resulting problems are usually easier to solve than under stochastic programming.

Stochastic programming

Stochastic programming solves the problem over a set S of scenarios generated based on the probability distribution of the unknown parameters [9]. Two approaches are usually adopted to manage the scenarios. In the *Here & Now* approach, the scenarios are included in a single optimization problem by repeating the constraints for each scenario $s \in S$ and reformulating the objective function to include the contribution of each scenario s, weighted by its occurrence probability π_s. In the *Wait & See* approach, the problem is separately solved in each scenario $s \in S$ and the expected value of all obtained solutions is then considered.

In our BDAS problem, scenarios refer to the arrival of non-booked donors. Thus, for the *Here & Now* approach, the number of non-booked donors with blood type $b \in B$ in day $t \in T$ is redefined as n_t^{bs} to be replicated in each scenario $s \in S$, while all other parameters remain the same over the scenarios. Also the main decision variables x_t^b and w_{tk}^b remain the same over the scenarios, providing a single *stochastic solution* to the problem, while the scenarios affect the definition of the other decision variables, whose values depend on scenario s. In particular, they are redefined as follows: y_t^{bs} is the number of planned units for blood type $b \in B$ in day $t \in T$ under scenario $s \in S$; z_t^{bs} of scenario $s \in S$ is the absolute variation of y_t^{bs} with respect its average value over T; p_{tk}^s is the dispersion penalty in period $k \in K$ of day $t \in T$ under scenario $s \in S$.

The objective function is finally given by the sum of the terms in the scenarios, weighted by the occurrence probability π_s of each scenario s. However, from a practical viewpoint, the number of theoretical scenarios to include could be extremely high, considering all possible combinations of n_t^{bs} over $b \in B$ and $t \in T$. Thus, our suggestion is to draw a predefined number $|S|$ of scenarios with a Monte Carlo approach from their probability distributions. Then, as they are generated with the same sampling mechanism, we suggest to assume a uniform probability distribution for their occurrence, by assigning equal probabilities π_s for each s. Finally, a sensitivity analysis can be conducted with respect to the choice of $|S|$, to derive the minimum number of scenarios to include in order to get a good robust solution. This sampling approach has been already applied in the health care literature [22].

However, in our opinion, the classical *Here & Now* approach as it is could not be fully effective for the BDAS preallocation model, because it is based on the so-called *risk-neutral* paradigm in which the objective function is given by the expected value over all considered scenarios. On the contrary, we think that focusing more on the worse realizations could be more effective. In this light, as future research direction, we suggest to consider the *risk-averse* stochastic programming and the Conditional Value-at-Risk (CVaR) risk measure for the objective function, which is directly based on the Value-at-Risk (VaR) measure [30, 33].

Let us consider the cumulative distribution function $F_Z(\cdot)$ of random variable Z, which is minimized in the objective function. VaR at confidence level $\alpha \in (0, 1]$ is equal to the α-quantile of Z, i.e., to $\inf\{\eta : F_Z(\eta) = \alpha\}$ [30, 33]. This is a simple and widely used risk measure with a clear interpretation; however, it only considers the position of the α-quantile. The CVaR risk measure takes into account both the α-quantile (the impact of all extreme values in the tail of distribution) and the conditional expectation of the least favorable outcomes (the expectation of only the worst values above the α-quantile). Indeed, the CVaR of random variable Z at confidence level α is [30, 33]:

$$CVaR_\alpha(Z) = \inf_{\eta \in \mathbb{R}} \left\{ \eta + \frac{1}{1-\alpha} \mathbb{E}\left[(Z-\eta)^+\right] \right\}$$

where $\mathbb{E}$ denotes the expected value and the superscript $^+$ defines the maximum between the argument and 0. This expression is nonlinear, due to the maximum operator $^+$; however, it can be linearized and easily rewritten considering the scenarios $s \in S$ and their occurrence probabilities π_s.

These stochastic programming approaches allow a detailed description of the uncertainty, based on the probability densities. Moreover, when considering a *risk-averse stochastic programming* model, the decision maker's risk-aversion degree can be easily adjusted by tuning the α-quantile value. However, the probability densities of the random arrivals of non-booked donors are not always available or their knowledge is inaccurate. In addition, based on the specific application, we might discover that a huge number of scenarios is required to get effective solutions, thus making it impossible to exploit the stochastic programming in practice. To address this possibility, we also discuss a robust optimization approach.

Robust Optimization

Several robust optimization approaches are available in the literature. One of the first and simplest ones was proposed by Soyster [34], in which all uncertain parameters vary in a predefined range and may assume they worst value together. However, in practice, it is highly unlikely that all parameters assume they worst value together.

Thus, more complex approaches have been developed. For example, Bertsimas and Sim [7] proposed an approach in which only a given subset of parameters for each constraint, whose cardinality is fixed, assume their worst value together; to respect the robustness of the solution, the worst combination of parameters is then chosen in the optimal solution. This approach has been already applied in health care [10, 23] and proved to be an effective tool [1]. However, it is not suitable for the BDAS preallocation problem, where all uncertain n_t^b are typically correlated. In fact, it is not realistic that the value of n_t^b in a day is uncorrelated to the other values in the close days. For this reason, we searched for other robust approaches that take parameter correlations into account.

Due to the characteristics of the BDAS, our suggestion is to apply the so-called *implementor-adversary* approach [8]. Its idea is to obtain the worst-case robust solution over the uncertainty set by iteratively adding scenarios that are critical for the last solution obtained. When no critical scenarios can be added, the set of the generated scenarios defines the uncertainty set. Indeed, the problem is formalized as a two-players game. At each iteration, the *implementor* solves the principal problem while considering the scenarios generated up to that iteration, while the *adversarial* determines a scenario that respects the conditions of the uncertainty set and worsen the solution found by the *implementor*. Then, the two problems are iteratively solved by adding the newly generated scenarios until the *adversarial* is no more able to find a critical scenario that worsen the *implementor* solution. This approach has been rarely applied in the health care literature: to allocate scarce medical staff to medical specialties [18] and to the nurse-to-patient assignment problem in home care [11].

The *implementor-adversarial* approach can be applied to the BDAS preallocation problem in a similar way than in Carello et al. [11]. In particular, rough information on the probability density of each n_t^b can be used to generate a set L of equiprobable levels, i.e., values n_t^{bl} with $l \in L$ that are representative of a band within the support of the distribution, where the bands contain equiprobable parts of the probability density. Then, the uncertainty set can be defined by assigning rules on the relationships between the bands, and scenarios s respecting these rules can be iteratively added.

4.2 Random No-Shows of Already Booked Donors

This concerns the reliability of parameters a_t^b in the BDAS preallocation model, whose values can be lower than planned due to donors who do not actually fill the reserved slot.

From the historical data, we may assume to know the show probability of each booked donor who contributes to each single a_t^b parameter. This show probability can be either derived as the ratio of times in which the specific donor showed up after reserving a donation, or a stochastic prediction model can be developed to determine the probability based on donor's covariates (e.g., age, gender, working status, distance from the collection center).

Thanks to the show probabilities, we can build a probability density function for each a_t^b, whose support ranges from $a_t^b = 0$, when all donors do not show up, to a maximum value when all donors do. The resulting probability density functions of all a_t^b can be employed in different ways. The simplest approach is to consider their expected values, which means to get a_t^b by summing the show probabilities of the already booked donors rather than their amount. In this case, the BDAS preallocation model remains deterministic, but at least the parameters match the expectation of the show probabilities, while in the model of [4] all donors were assumed to show up. Other possible approaches are those already suggested for n_t^b. On the one hand, scenarios can be generated and addressed with stochastic programming, also including the CVaR risk measure; on the other hand, a robust approach can be exploited. Also in this case, the implementor-adversarial approach could be a valid tool.

5 Conclusions

In this chapter, we have discussed the major sources of uncertainty that affect any appointment scheduling system for blood donation, with particular reference to the BDAS preallocation model. In close collaboration with AVIS Milan, we have found out the two most critical sources, i.e., the uncertain arrivals of non-booked donors and the uncertain no-shows of booked donors. For both of them we have suggested possible integrations with the BDAS framework, in order to promote future research and suggest practical solutions. In the future, we will integrate these uncertainty sources and test the benefits on the AVIS Milan case. The final goals are to improve the management of BD collection centers taking into account their twofold function, i.e., production system and service provider.

References

1. Addis, B., Carello, G., Grosso, A., Lanzarone, E., Mattia, S., Tànfani, E.: Handling uncertainty in health care management using the cardinality-constrained approach: advantages and remarks. Oper. Res. Health Care **4**, 1–4 (2015)
2. Akhavan Niaki, S.T.: Presenting a stochastic multi choice goal programming model for reducing wastages and shortages of blood products at hospitals. J. Ind. Syst. Eng. **10**, 81–96 (2017)
3. Baş, S., Carello, G., Lanzarone, E., Ocak, Z., Yalçindağ, S.: Management of blood donation system: literature review and research perspectives. In: Springer Health Care Systems Engi-

neering for Scientists and Practitioners—Proceedings of HCSE 2015 (Springer Proceedings in Mathematics and Statistics), vol. 169, pp. 121–132 (2016)
4. Baş, S., Carello, G., Lanzarone, E., Yalçindağ, S.: An appointment scheduling framework to balance the production of blood bags from donation. Eur. J. Oper. Res. **265**, 1124–1143 (2018)
5. Baş, Güre S., Carello, G., Lanzarone, E., Yalçindağ, S.: Unaddressed problems and research perspectives in scheduling blood collection from donors. Prod. Plan. Control **29**, 84–90 (2018)
6. Beliën, J., Forceé, H.: Supply chain management of blood products: a literature review. Eur. J. Oper. Res. **217**, 1–16 (2012)
7. Bertsimas, D., Sim, M.: The price of robustness. Oper. Res. **52**, 35–53 (2004)
8. Bienstock, D.: Histogram models for robust portfolio optimization. J. Comput. Finance **11**, 1–65 (2007)
9. Birge, J.R., Louveaux, F.: Introduction to Stochastic Programming. Springer Science and Business Media (2011)
10. Carello, G., Lanzarone, E.: A cardinality-constrained robust model for the assignment problem in home care services. Eur. J. Oper. Res. **236**, 748–762 (2014)
11. Carello, G., Lanzarone, E., Laricini, D., Servilio, M.: Handling time-related demands in the home care nurse-to-patient assignment problem with the implementor-adversarial approach. In: International Conference on Health Care Systems Engineering, pp. 87–97 (2017)
12. Dillon, M., Oliveira, F., Abbasi, B.: A two-stage stochastic programming model for inventory management in the blood supply chain. Int. J. Prod. Econ. **187**, 27–41 (2017)
13. Ensafian, H., Yaghoubi, S.: Robust optimization model for integrated procurement, production and distribution in platelet supply chain. Transp. Res. Part E Logistics Transp. Rev. **103**, 32–55 (2017)
14. Ensafian, H., Yaghoubi, S., Yazdi, M.M.: Raising quality and safety of platelet transfusion services in a patient-based integrated supply chain under uncertainty. Comput. Chem. Eng. **106**, 355–372 (2017)
15. Fazli-Khalaf, M., Khalilpourazari, S., Mohammadi, M.: Mixed robust possibilistic flexible chance constraint optimization model for emergency blood supply chain network design. Ann. Oper. Res. **283**, 1079–1109 (2019)
16. Gunpinar, S., Centeno, G.: Stochastic integer programming models for reducing wastages and shortages of blood products at hospitals. Comput. Oper. Res. **54**, 129–141 (2015)
17. Hemmelmayr, V., Doerner, K.F., Hartl, R.F., Savelsbergh, M.W.: Vendor managed inventory for environments with stochastic product usage. Eur. J. Oper. Res. **202**, 686–695 (2010)
18. Holte, M., Mannino, C.: The implementor/adversary algorithm for the cyclic and robust scheduling problem in health-care. Eur. J. Oper. Res. **226**, 551–559 (2013)
19. Jabbarzadeh, A., Fahimnia, B., Seuring, S.: Dynamic supply chain network design for the supply of blood in disasters: a robust model with real world application. Transp. Res. Part E Logistics Transp. Rev. **70**, 225–244 (2014)
20. Jafarkhan, F., Yaghoubi, S.: An efficient solution method for the flexible and robust inventory-routing of red blood cells. Comput. Ind. Eng. **117**, 191–206 (2018)
21. Kazemi, S.M., Rabbani, M., Tavakkoli-Moghaddam, R., Shahreza, F.A.: Blood inventory-routing problem under uncertainty. J. Intell. Fuzzy Syst. **32**, 467–481 (2017)
22. 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.-Part A Syst. Hum. **42**, 1346–1363 (2012)
23. Marques, I., Captivo, M.E.: Different stakeholders' perspectives for a surgical case assignment problem: deterministic and robust approaches. Eur. J. Oper. Res. **261**, 260–278 (2017)
24. Najafi, M., Ahmadi, A., Zolfagharinia, H.: Blood inventory management in hospitals: considering supply and demand uncertainty and blood transshipment possibility. Oper. Res. Health Care **15**, 43–56 (2017)
25. Osorio, A.F., Brailsford, S.C., Smith, H.K.: A structured review of quantitative models in the blood supply chain: a taxonomic framework for decision-making. Int. J. Prod. Res. **53**, 7191–7212 (2015)

26. Osorio, A.F., Brailsford, S.C., Smith, H.K.: Whole blood or apheresis donations? A multi-objective stochastic optimization approach. Eur. J. Oper. Res. **266**, 193–204 (2018)
27. Puranam, K., Novak, D.C., Lucas, M.T., Fung, M.: Managing blood inventory with multiple independent sources of supply. Eur. J. Oper. Res. **259**, 500–511 (2017)
28. Rabbani, M., Aghabegloo, M., Farrokhi-Asl, H.: Solving a bi-objective mathematical programming model for bloodmobiles location routing problem. Int. J. Ind. Eng. Comput. **8**, 19–32 (2017)
29. Ramezanian, R., Behboodi, Z.: Blood supply chain network design under uncertainties in supply and demand considering social aspects. Transp. Res. Part E Logistics Transp. Rev. **104**, 69–82 (2017)
30. Rockafellar, R.T., Uryasev, S.: Conditional value-at-risk for general loss distributions. J. Banking Finance **26**, 1443–1471 (2002)
31. Salehi, F., Mahootchi, M., Husseini, S.M.M.: Developing a robust stochastic model for designing a blood supply chain network in a crisis: a possible earthquake in Tehran. Ann. Oper. Res. **283**, 679–703 (2019)
32. Samani, M.R.G., Torabi, S.A., Hosseini-Motlagh, S.M.: Integrated blood supply chain planning for disaster relief. Int. J. Disaster Risk Reduction **27**, 168–188 (2018)
33. Schultz, R., Tiedemann, S.: Conditional value-at-risk in stochastic programs with mixed-integer recourse. Math. Program. **105**, 365–386 (2006)
34. Soyster, A.: Convex programming with set-inclusive constraints and applications to inexact linear programming. Oper. Res. **21**, 1154–1157 (1973)
35. Sundaram, S., Santhanam, T.: A comparison of blood donor classification data mining models. J. Theor. Appl. Inf. Technol. **30**, 98–101 (2011)
36. Van Dijk, N., Haijema, R., Van Der Wal, J., Sibinga, C.S.: Blood platelet production: a novel approach for practical optimization. Transfusion **49**, 411–420 (2009)
37. Zahiri, B., Pishvaee, M.S.: Blood supply chain network design considering blood group compatibility under uncertainty. Int. J. Prod. Res. **55**, 2013–2033 (2017)
38. Zahiri, B., Torabi, S.A., Mohammadi, M., Aghabegloo, M.: A multi-stage stochastic programming approach for blood supply chain planning. Comput. Ind. Eng. **122**, 1–14 (2018)
39. Zahiri, B., Torabi, S.A., Mousazadeh, M., Mansouri, S.A.: Blood collection management: methodology and application. Appl. Math. Modell. **39**, 7680–7696 (2015)
40. Zhou, D., Leung, L.C., Pierskalla, W.P.: Inventory management of platelets in hospitals: optimal inventory policy for perishable products with regular and optional expedited replenishments. Manuf. Serv. Oper. Manag. **13**, 420–438 (2011)