SIMULATING THE IMPACT OF OPTIMIZED DISPATCHING STRATEGIES FOR PATIENT TRANSITS

Nikolaus Furian
Michael O'Sullivan
Cameron Walker
Simon Riezeboz
University of Auckland
Department of Engineering Science
Email: nfur003@aucklanduni.ac.nz

Dietmar Neubacher
Siegfried Vössner
Graz University of Technology
Department of Engineering and Business Informatics
Email: dietmar.neubacher@tugraz.at

KEYWORDS
Health Care, Transportation, Discrete Event Simulation

ABSTRACT

The on-time completion of patient transits can be identified as a bottleneck for the efficiency of health care services in major New Zealand hospitals. Delayed transits of patients between wards and treatment or diagnostic facilities lead to increasing waiting times at clinics and inefficient resource utilization (e.g. surgery teams) as appointment times are not met. Therefore, a long term research project was launched to install optimized and automated dispatching algorithms for staff members performing those transits. In this paper we briefly outline a stochastic simulation model that was designed to evaluate the improvement gained by optimized dispatching and present a set of algorithms for the allocation of transits to orderlies. Initial experimental results on several real-world instances are presented and discussed.

INTRODUCTION

In this article we present the first steps and results towards the installation of automated optimal dispatching strategies in major New Zealand hospitals, characterized by multiple wards and treatment facilities. During their stay, patients frequently require transportation to and from appointments within the hospital (a transit). These transits are usually performed by an orderly. However, patients with additional needs, due to their condition, require the assistance of a nurse. Transits are required to arrive at the destination location within a certain time-window, usually between 15 minutes prior the appointment and the appointment time. However, the actual system performance deviates significantly from this target for two reasons: transit requests are made with not enough notice by the corresponding facilities; the dispatching of transits to available resources (orderlies and nurses) is done in an ad hoc manner. While the first issue may only be overcome by applying and enforcing stricter policies for transit-requests, which is not within the scope of this article, the second issue may be tackled by the installation of automated optimal dispatching strategies, as described in this paper. Therefore, an optimization problem is defined and an algorithm to find an optimal, or at least near-optimal, solution is proposed. To evaluate the effectiveness of the algorithm in a dynamic and stochastic environment a simulation model is implemented for evaluation.

The paper is outlined as follows: in the next section a brief description of the optimization problem is given. In the following sections the simulation model and the proposed algorithms are presented. The paper concludes with experimental results, concluding remarks and suggestions for further research.

PROBLEM DEFINITION

The problem can be described as follows. For a given point in time there is a set of known transits to be performed in the future. They include three different activities: preparation; the actual travel; and a handover activity. Each transit is defined by pickup and drop-off locations, staff requirements and two associated time windows. The first window denotes the desired completion time of the transit, including the handover activity, and is usually chosen as $[-15, 0]$ with respect to the appointment time (note, all times are in minutes). The second time window covers delayed and penalized completion times that are still feasible, usually chosen to be $[0, 15]$ with respect to the appointment time. Note that early arrivals, e.g. $(-\infty, -15)$ to the appointment time rarely occur in the real system but should be avoided by any automated dispatching routine.

Per definition, for transits that require an orderly and a nurse, the nurse is responsible for the transit, i.e. performing preparation and handover tasks. Consequently, orderlies only participate in the actual travel for this type of transit. Furthermore, there are given staff rosters that determine the availability of orderlies and nurses at any point in time. The goal, of any dispatching strategy is to assign transits to staff members so that the overall lateness of all pending transits is minimal, i.e. the sum of all late completion times minus the respective appointment times. This problem is closely related to vehicle routing problem with semi soft time windows (VRPSSTW), as optimal routes through a network of jobs for orderlies and nurses have to be computed. However, the authors are not aware of a previously published problem instance including the need for synchronized routes of multiple staff types. For more detailed information and algorithms on VRPSSTW the reader is referred to (Qureshi et al. 2010). A general description of routing problems with time windows can be found in (Kallehauge 2008). Note that staff members start and finish their shifts at a base, to which they also return in the case of no pending job. As the set of known (or logged) and pending transits to be performed constantly changes (at the start of jobs and arrival of new ones) re-optimization of current solutions have to occur on a regular basis. To allow the evaluation of optimization strategies within such a dynamic environment a simulation model has been implemented.

THE SIMULATION MODEL

Simulation, and in particular Discrete Event Simulation (DES) is a commonly used tool for analysis and decision support in the health care industry. For reviews the reader is referred to (J.B. et al. 1999, Brailsford et al. 2009, Günal and Pidd 2010). The main purpose of the implemented simulation model presented here is to allow the evaluation of automated and optimized dispatching strategies within the changing environment of the actual system. This includes: available information on future transits at any given point in time; and the stochastic nature of the real system (random travel, preparation, and hand-over times). This means that the optimized schedule of task allocation will deviate from the real system as time progresses. To address this issue, we conceptualized a DES model according to the Hierarchical Control Conceptual Modeling (HCCM) framework, see (Furian et al. 2015). The main feature of HCCM models are that the model control policies (e.g. triggering of conditional activities and/or dispatching decisions) are centralized in control units, rather than being nested in de-centralized process definitions or activity conditions. This methodology perfectly suits the underlying problem where dispatching is done by a centralized control using automated algorithms.

In the remainder of this section the basic components of the simulation model are briefly outlined. First, the underlying movement model is introduced as it is the basis for all activities that include travelling of staff members, with or without patients, between any two locations in the model. Second, individual behavioral paths of patients and staff members are described. Based on these paths and corresponding requests the control structures of the model are defined. Finally, some notes on the model implementation are made.

Movement Model

The first step of the design of the simulation model is the generation of a network consisting of all facilities within Auckland City Hospital, as well as paths and elevators between different locations. Based on this network a movement model of staff with and without patients is created (some paths may not be used when traveling with patients). In particular, when dispatched, a staff member calculates the shortest path from her/his current location to the destination using Djikstra's-algorithm and starts the movement. The travel time for the move is computed based on the total non-vertical distance to be traversed and the number of floors that have to be passed. Different velocities for movements with or without patients determine the time required to move along non-vertical paths. Note, that we assume that all staff members travel at the same speeds. The usage of elevators to change floors is simplified by sampling an exponentially distributed waiting time for the elevator and constant times for each floor to be passed.

Behavior of Staff Members and Patients

As previously mentioned, a transit request consists of three activities to be performed: preparation; the transits itself; and a handover activity. The generation of transit requests is based on historical transit data. Each request includes the following data: the log time (historic time when the request enters the system); start and end location; staff requirements; appointment time; and historical dispatch time. Note that preparation and handover durations are sampled from exponential distributions, whereas the time required to perform a transit is a result of the movement model described in the previous section. Hence, the process of a patient can be summarized by "Arrival - Wait for Dispatch Time - Wait for Staff - Preparation - Wait for Additional Staff (if required) - Transit - Handover". The dispatch time of transits is computed based on expected durations of activities associated with the transit and denotes the time when staff members are dispatched to perform tasks (note staff members may need to complete current tasks first).

The behavioral path of staff members is not as simple as for patients, since in the case of multi-staff transits, preparation and handover is performed by the responsible nurse only. Further, when there are no transits pending in the near future, idle staff members return to the base location. As this results in decisions with higher complexity to be made (due to multiple staff potentially being involved in multiple tasks across the hospital), this logic is located in a designated and centralized control unit. Figure 1 shows the process for staff members (orderlies and nurses). It includes all activities outlined above. The “Wait for Order” activity indicates that decisions on further actions of staff members are made by the control unit instead of by a priori logic nested within conditions embedded within the process of orderlies and nurses (as in typical DES queuing models).

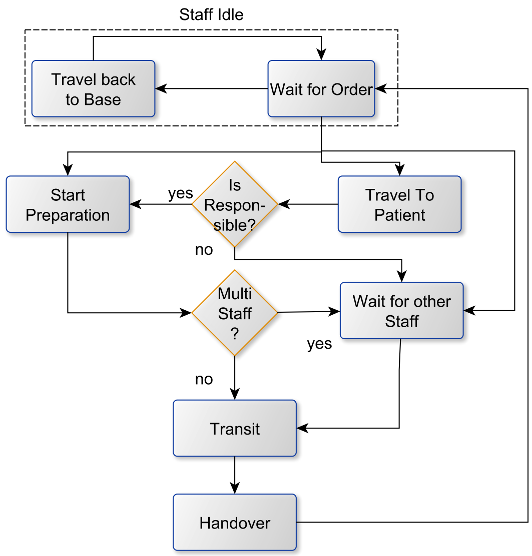

Figure 1: Behavioral Path of Staff members

Model Control Structure

As described above, the model control mechanism that handles requests for transits is nested in a centralized control unit. It includes a set of rules that reflects the decision logic of the model. The rule set is evaluated each time the system changes, which occurs in DES models upon the triggering of events. Events occurring in this model are: start and end events of activities; arrival events of patients; and events representing the dispatch time of transit requests. The control logic can be summarized as follows:

1. If a new transit has arrived or been completed dispatching decisions are updated by calling the optimization algorithm;

2. For each idle staff member:

   (a) check if she/he has jobs pending;

   (b) If no jobs are in the personal queue and the staff member is located outside the base, then she/he is sent to the base;

   (c) If there are existing jobs in the personal queue, then

      i. For transits that require immediate dispatching of the staff member, she/he is sent to that location;

      ii. If there is not enough time to return to the base, then the staff member is sent to the pick-up location of the next job and waits;

      iii. If there is enough time before the next job to return to base, then the staff member is sent to the base;

   (d) When located or arriving at the pickup location it is checked if the staff member is responsible for the transit: if responsible the preparation of the transit is triggered; if not, the staff member is sent in a “Waiting for Other Staff” activity;

3. For each transit with a completed preparation it is checked if all required staff members are present at the pickup-location; if so the travel is launched.

Note, with regards to dispatching and the update of staff positions: non-idle staff members are assigned an earliest available time based on expected remaining durations of their current transit as an input for the dispatching algorithm; the optimization is allowed to entirely change previous assignments, if the transits have not started; staff members that are waiting for their next jobs to become due may be re-assigned to other transits by the dispatching algorithm; non-responsible orderlies that wait for the nurse to complete the preparation activity are not re-assigned to other jobs.

Model Implementation

The model implementation was done using an object-oriented and activity based DES library, previously developed by the authors in the C# programming language. Its main functionality is closely coupled with the principles of the HCCM framework, that is centralizing model execution logic in designated control units. In particular it consists of two event triggering mechanisms: first, a scheduled event list, that is standard to almost any DES library; second, conditional behavior (events and activities) that is only triggered within the proposed control units according to the rule sets specified by the user. Hence, the usual DES scanning of activity or process activation conditions is replaced by the execution of centralized rule sets, as described by Furian et al. (2014b;a).

OPTIMIZATION ALGORITHMS FOR TRANSIT DISPATCHING

In this section an optimization method for the optimal, or near optimal, transit dispatching problem is proposed. It consists of several sub-algorithms of increasing complexity, that are called sequentially if the previous one failed to find an optimal solution. Therefore, it has to be noted that the maximum of unavoidable delays for all transits considered individually, which itself is given by the delay that occurs when sending the set of staff members that are able to complete the job earliest, is a valid lower bound. In many cases this denotes a tight lower bound which reduces the use of more runtime-intensive algorithms. Next, the proposed algorithms are briefly outlined individually, followed by a description of their combined use embedded in the simulation.

Next Possible Heuristic

The Next Possible Heuristic (NPH) is not only the simplest algorithm presented, but also approximately describes the ad-hoc policy that is currently in place to dispatch transits in the real system. It loops through the set of transits, ordered with respect to appointment times, and assigns staff members that can start the job earliest (based on expected availability times).

Enumerative Search Algorithm

The Enumerative Search Algorithm (ESA) makes use of the fast and simple NPH algorithm. As explained above NPH loops through the ordered set of transits (with respect to appointment time) for assignment. ESA breaks this ordering and defines its search space by a sub-set of all possible sequences of jobs. The subset is defined in such a way that the violation of ordering by appointment time is only allowed to some extent (i.e. jobs may only be switched in the sequence if their appointment times deviate not more than 5 minutes). Based on that search space a set of sequences is enumerated and each sequence is evaluated with the same assignment strategy as NPH. The best solution found is returned.

Combining Optimal Individual Solutions

The Combining Optimal Individual Solutions (COIS) algorithm makes use of the observation that the share of transits that require the assistance of nurses is rather small for practical instances. Hence, it optimally solves the problem reduced to orderlies and corresponding transits only. Therefore, a Mixed Integer Program (MIP) was formulated that represents the orderly only problem. This is then solved using a typical column generation approach for the VRPSSTW extended by some problem-specific enhancements. However, the exact description of the methodology is beyond the scope of this paper.
Based on the optimal solution for orderlies COIS attempts to fit nurse routes to the given allocation schedule. Therefore, critical time windows for nurses are computed. Basically, they denote the maximum delay of jobs with respect to the completion times given by the orderly solution without causing additional delays. If it is possible to find a solution with zero delay for the nurse problem with respect to the updated time windows, both solutions can be combined to an optimal overall dispatching decision. If it is not possible, either a feasible solution has been found (which may not be optimal) or no solution could be generated by the COIS algorithm. In that case, the best solution found by other algorithms is used as a basis for dispatching decisions.

Insert Job to Path

The last algorithm presented, Insert Job to Path (IJTP), is a simple heuristic that attempts to insert additional jobs to the routes found at the solution of the last call of the optimization method.

Optimization Framework

As mentioned above, the outlined algorithms are used in sequential manner to make dispatching decisions within the simulation model. Starting with the attempt to make use of the previously found solution (checks of potential for improvement in case of removed jobs due to stochasticity and IJTP), NPH, ESA and COIS are called after checks if the currently best solution is optimal, as illustrated by figure 2. The resulting overall algorithm is denoted by Optimization Framework (OptFram). Note that it may occur that no feasible solution with respect to hard time windows could be found by any of the algorithms in OptFram. In the real system, this transit would be still performed and the infeasibility would be accepted. To allow for this behavior we enlarge the hard time windows in the case that ESA was not able to find a feasible solution. This decision was motivated by the observation that the runtime of COIS increases dramatically if an empty set of feasible starting solutions is provided.

RESULTS

In this section we outline the process of simulation fitting and provide experimental results for a selected set of instances based on data collected at the Auckland City Hospital.

Simulation Fitting

Data available on historic transits includes locations, staff requirements, appointment time, dispatching time and performing staff members. However, we do not have any knowledge on the decision when to dispatch tasks and at what location staff members were when they received their dispatch order or respectively started a job. The lack of this information, combined with the inaccurate nature of the available data led to the following approach:
First, traveling speeds of staff members, preparation and handover times, as well as elevator waiting times and traveling speeds were fitted on a filtered set of transits.

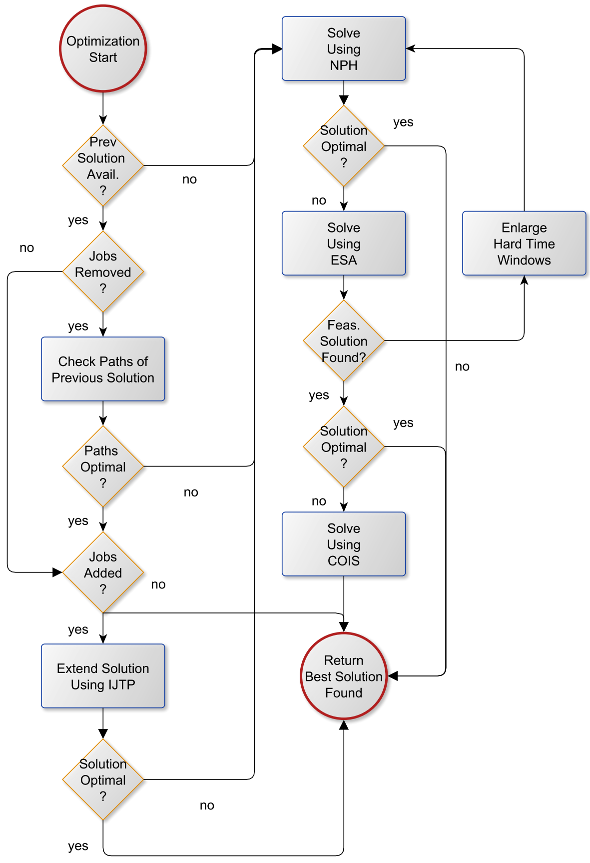

Figure 2: Optimization Framework

Experimental tests have shown that the following values for parameters lead to a Pareto-minimum over the mean deviation of transit times and the standard deviation: traveling speeds (with and without patients) of (50,70) [m/minute]; exponentially distributed preparation, handover and elevator waiting times with means (1,2,0.25)[minute]; and elevator floor-change times of 0.5 seconds.

Second, we compared the historic overall system performance of a large data set with the simulated performance using only NPH as a dispatching routine, as shown by figure 3. As results were satisfactory we do not need to evaluate OptFram on small instances of historic dispatching decisions data that most likely include high uncertainty of quality, but are able to compare its performance with delays observed using NPH. Note that, for this analysis early arrivals were allowed (as we used historic dispatching times), even if they must be avoided by any future automatic dispatching policy.

Experimental Results

To evaluate the performance of OptFram we created 10 instances based on weekday transit data with corresponding appointment times between 8:00am and

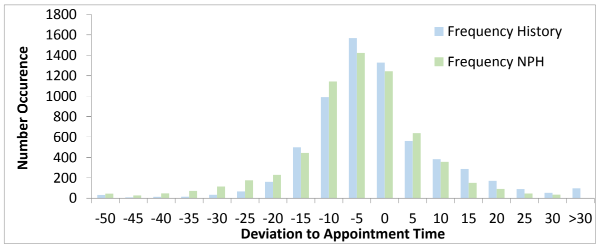

Figure 3: Parameter Fitting-Simulation

8:00pm. The average number of transits performed during this period was 443.3 out of which on average 47.0 required a nurse to be present. It was assumed that 4 orderlies and 2 nurses were present the whole period and an additional orderly and nurse were present between 9:00am and 6:00pm. This denotes a realistic, but rather low, staffing level. Each instance was simulated 20 times and the average and worst performance of both each mode (NPH only and OptFram) was reported. Note that the same historic arrival streams were used with the exact same activity durations for each mode to ensure comparability. Table 1 shows the collected results, where l_t denotes the aggregate lateness of all transits, l_n the number of transits completed outside the desired time window and il_n the number of transits completed outside the hard time window. One can clearly see that the OptFram yields a reduction in the average and maximum overall lateness and number of late arrivals over all simulation runs for all but one instance.

Table 2 shows performance measures of OptFram, including: the average instance size of optimization calls $\varnothing n$; the maximum instance size of a single optimization call $\max n$; the average share of calls that resulted in an optimal solution $\varnothing opt$; the average runtime of optimization calls $\varnothing rt$; and the average runtime of the most complex COIS algorithm $\varnothing rt_{cois}$. Note that a three minute time limit was applied to COIS. In case this limit was reached the solution obtained by ESA is used.

Results show that optimality is reached in a vast majority of cases and computation times are within a range that allows the usage of OptFram in the real world operating system where decisions must be obtained within minutes due to the constant re-planning.

CONCLUSION AND FURTHER RESEARCH

In this paper we proposed a simulation model to reproduce patient transits within major New Zealand hospitals. Further, we briefly outlined a optimization methodology for dispatching of staff members to transits. The algorithm was integrated into the simulation model and its performance compared to an ad-hoc heuristic that represents the policy currently used. Results show that significant reductions of overall lateness

NPH				
Inst.	$\varnothing l_t$	$\max l_t$	$\varnothing(l_n, il_n)$	$\max(l_n, il_n)$
Mo-I	154.61	221.23	(27.6,2.7	(35,4)
Tu-I	402.75	598.75	(73.65,2.5)	(100,7)
We-I	348.13	641.47	(81.94,2.11)	(118,5)
Th-I	184.64	213.9	(36.1,0.85)	(43,2)
Fr-I	307.77	523.66	(44.58,4.11)	(63,10)
Mo-II	130.4	197.95	(28.9,0.3)	(40,1)
Tu-II	297.86	376.89	(40.5,3.35)	(44,6)
We-II	205.35	327.2	(42.45,.6)	(58,3)
Th-II	141.86	179.76	(23.85,1.35)	(33,2)
Fr-II	553.81	800.17	(84.7,5.7)	(103,15)
OptFram				
Mo-I	155.17	225.84	(30.25,2.3)	(35,5)
Tu-I	222.67	415.85	(50.5,.7)	(100,3)
We-I	181.82	272.39	(50.06,1.33)	(118,3)
Th-I	161.93	206.94	(32.45,.75)	(43,2)
Fr-I	248.99	286.27	(39.21,3.11)	(63,5)
Mo-II	120.07	179.12	(27.15,.3)	(40,1)
Tu-II	232.83	261.34	(37.1,2.85)	(44,4)
We-II	185.42	248.99	(40,.7)	(58,2)
Th-II	129.87	155.2	(25.4,1.35)	(33,3)
Fr-II	329.52	697.94	(60.75,3.7)	(103,11)

Table 1: Simulation-Optimization Results

Inst.	$\|J\|$	$\varnothing n$	$\max n$	$\varnothing opt$	$\varnothing rt$	$\varnothing rt_{cois}$
Mo-I	457	15.09	32	0.96	2.47	1.89
Tu-I	466	14.78	33	0.86	2.47	16.73
We-I	397	11.47	22	0.96	0.66	7.99
Th-I	408	12.18	22	0.97	1.17	8.85
Fr-I	444	13.64	23	0.97	1.08	9.79
Mo-II	451	12.37	26	0.97	0.59	4.3
Tu-II	430	15.94	32	0.96	5.04	5.88
We-II	447	11.7	21	0.92	1.43	10.12
Th-II	491	15.86	28	0.92	2.47	3.63
Fr-II	452	13.49	26	0.96	2.19	8.39

Table 2: Optimization Performance Measurements, times in seconds

and number of patients arriving late can be achieved by the usage of the optimization algorithm. As these are initial results, it is left for further research to analyze a broader set of instances and multiple staffing rosters. Furthermore, a comparison of results achieved under stochastic and deterministic environment may be of interest.

ACKNOWLEDGMENTS

The authors would like to thank Tim Winston and Jonathan Wallace for their inputs and cooperation, as well as Thomas Adams for the basis of floor-plan models of Auckland City Hospital. Further, the first author thanks the **Austrian Science Fund (FWF): Project Nr. J3376-G11**, for funding his research in New Zealand.

REFERENCES

Brailsford S.; Harper P.; Patel B.; and Pitt M., 2009. *An analysis of the academic literature on simulation and modelling in health care. Journal of Simulation*, 3, 130–140.

Furian N.; O'Sullivan M.; Walker C.; and Vössner S., 2014a. *A Generic Integrated Health Care Model.* In *Proceedings of the 48th Annual Conference of the ORSNZ.*

Furian N.; O'Sullivan M.; Walker C.; and Vössner S., 2014b. *HCCM - A Contorl World View for Health Care Discrete Event Simulation.* In *Proceedings ECMS 2014.*

Furian N.; OSullivan M.; Walker C.; Vössner S.; and Neubacher D., 2015. *A conceptual modeling framework for discrete event simulation using hierarchical control structures. Simulation Modelling Practice and Theory*, 56, no. 0, 82 – 96. ISSN 1569-190X.

Günal M. and Pidd M., 2010. *Discrete event simulation for performance modelling in health care: a review of the literature. Journal of Simulation*, 4, 42–51.

J.B. J.; S.H. J.; J.R. S.; and Correspondence, 1999. *Application of discrete-event simulation in health care clinics: A survey. Journal of the Operational Research Society*, 50, no. 2, 109–123.

Kallehauge B., 2008. *Formulations and exact algorithms for the vehicle routing problem with time windows. Computers & Operations Research*, 35, no. 7, 2307 – 2330. ISSN 0305-0548.

Qureshi A.G.; Taniguchi E.; and Yamada T., 2010. *Exact solution for the vehicle routing problem with semi soft time windows and its application. Procedia - Social and Behavioral Sciences*, 2, no. 3, 5931 – 5943. ISSN 1877-0428.