Dynamic Agent-based Scheduling of Treatments: Evidence from the Dutch Youth Health Care Sector

Erik Giesen[1(✉)], Wolfgang Ketter[2], and Rob Zuidwijk[2]

[1] INITI8, Rotterdam, The Netherlands
giesen@initi8.nl
[2] Rotterdam School of Management, Rotterdam, The Netherlands
{wketter,rzuidwijk}@rsm.nl

Abstract. We use agent-based simulation to compare the performance of four scheduling policies in youth health care. The policies deploy push/pull and centralized/decentralized concepts. The simulation model represents an authentic business case and is parameterized with actual market data. The model incorporates, among other things, non-stationary Poisson arrival processes, reneging and return mechanisms, and care provider's client preferences. We have identified that performance measurement in youth health care should not be focused on queue lengths alone, which is presently the case, but should include a case difficulty parameter as well. The simulation results, together with contextual data obtained from stakeholder interviews, indicate that a push strategy with a centralized queue suits the sector best, which is different from the current real-world situation. This policy ensures a higher level of fairness in treatment provision because the care providers are compelled to take their share in treating the difficult and economically less attractive cases. The complexity of the case cannot be captured by current queuing theory methods. Our simulation approach incorporates these complexities, which turn out to be relevant for the scheduling policy decision. We validate the model and strategies using real market data and field expert discussions.

Keywords: Agent-based simulation · Resource allocation · Youth health care · Preference behavior · Policy scheduling

1 Introduction

The Dutch youth health care sector is providing care to youths under 19 and their families on a voluntary basis. The scheduling of care includes the allocation of clients to care providers and it features long waiting lists and long waiting times. As in many other countries, the issue is considered an urgent societal problem and has received a lot of media attention [21]. Earlier approaches that solely address the symptom of long waiting lists have proven to be ineffective. The government is funding the sector and it has instituted central bureaus in

J.P. Müller et al. (Eds.): MATES 2015, LNAI 9433, pp. 173–199, 2015.
DOI: 10.1007/978-3-319-27343-3_10

provinces and larger urban areas to manage youth care on a regional level.[1] Each of these institutions operate without regional overlap and act as the gateway to youth care for clients from the region that it serves. Clients in need of care enter the system by visiting the institution for youth health care that diagnoses the situation and provides the client with a diagnosis. This diagnosis can be seen as a entitlement to health care. Typically the institution for youth health care also selects the care provider expected to fit best to the problem and preferences of the client, although it is at the clients' discretion to adhere to this allocation or not. The care providers are compensated by the government for the care that is provided corresponding with the diagnoses from the institution for youth health care; see Fig. 1.

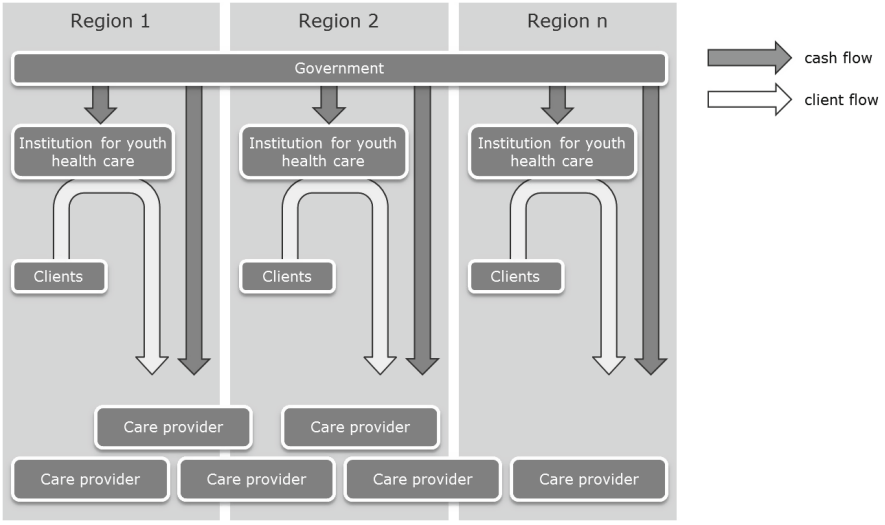

Fig. 1. Overview of the allocation mechanism in the Dutch youth health care system.

Parents, teachers, and other people involved with children have become increasingly aware of potential problems and have also started to signal problems more often. While the question remains whether this can be seen as over-signalling or not, it certainly results in an increase of the amount of clients requesting help [2]. The institution of youth health care acts as a gateway more than as a gatekeeper, as it is not equipped with the legal authority to dismiss a case. As a result, there is not enough capacity at the care providers to deal with the growing number of requests for care. In addition, the provision of care is on a voluntary basis, so clients may renege, i.e. withdraw from the system at any time while waiting for care. This further complicates the management of care

[1] This was the case until 1/1/2015 and reflects the data we used. Today however municipalities are responsible for managing youth care among other types of care.

provision. Reneging may be caused by the fact that clients found other ways to be assisted with their problem, or that the issue at hand resolved itself without professional care. However, reneging may also occur in cases where youth health care should have been provided. This may leave youth health problems to remain unresolved or re-entering of the client in the system while the situation has persisted or even worsened. On the other hand, it has been argued that clients in genuine need of care are willing to wait longer for the requested care [12,13]. In such a manner, reneging would become a sort of natural way of balancing the system and filtering out cases not in genuine need for care.

As care providers are working with under-capacity, they effectively are able to select clients from the queues. In particular, more difficult cases are less attractive from a financial point of view. The selection process depends in an intricate way on a lot of factors such as the age or gender of the child, the type of problem, and the region in which the child lives. As a result, the selection process is not transparent and it allows the care providers to base their selections on financial incentives as well. In order to manage the youth health care system also in this respect, the performance of the system should be expressed both in terms of efficiency and social welfare, where the latter is based on indicators reflecting the actual treatment of difficult cases and waiting times. Such indicators may prevent difficult cases to be disadvantaged and help create a fair scheduling process. We elaborate on such indicators in Sect. 3.

To address the waiting line issues, this paper considers alternative solution directions that not only focus on the handling of contemporary waiting lists, but that may require structural changes in the scheduling of youth health care to clients. We elaborate on such structural changes by presenting an overview of multiple scheduling policies, based on a combination of push/pull and centralized/decentralized scheduling policies. The push and pull scheduling policies define the party which ultimately makes the actual allocation decision. Centralized and decentralized scheduling policies define the moment at which the actual allocation will take place.

Scheduling decision problems, as presented by the youth health care case, suit very well a multi-agent simulation approach for the following reasons. The behavior of stakeholders in the system has a decisive impact on scheduling decisions and therefore needs to be captured well in the decision model. The impact of how communication is organized between the different parties in the system needs to be incorporated as well. Furthermore, institutions and persons have their own objectives, are heterogeneous entities by nature, and the coordination thereof needs to be addressed explicitly. As a result, the actual client flow through the system is the result of a negotiation process between several parties in the supply chain. Indeed, a client scheduling procedure requires input from other parties in the sector on which the final decision can be based. A multi-agent simulation built of individual agents that pursue a specific personal goal can be used in this complex, dynamic setting to evaluate alternative scheduling policies.

To arrive at potentially structural changes that address the problems described above, a systematic approach is required. An analysis of what the various

stakeholders expect from the system, what has presently been achieved, and what can be achieved, needs to encapsulate the rich problem context. The strategic objectives of the system and their target values need to be elicited, and they need to be expressed in terms of Key Performance Indicators (KPIs), which may vary among stakeholders. The actual performance of the system needs to be formalized as a baseline so that the performance gaps can be analyzed and so that performance improvements by alternative scheduling policies can be assessed. In this setting, one should anticipate that one size may not fit all, and that solution directions need to be specified for different contexts, e.g. for different geographical regions and for different care types in the youth care sector. To perform such an analysis in a complex, dynamic environment such as youth care, there is a need for a responsive design paradigm.

We contribute to the research in health care operations management, in particular resource scheduling, by providing a currently unused approach to counter queuing related issues. Simulation of the resource scheduling process helps to understand and test long term effects of a number of alternative scheduling policies and coordination decisions. Although operations research queuing models go a long way in incorporating behavior in queuing systems, such as customer impatience [3], we argue that these models fall short in capturing the behavior required to explain the system behavior in the youth health care sector. Indeed, our simulation approach addresses the complexities of the patient scheduling that were found in the real world case and incorporates, among others, a non-stationary Poisson arrival process, a reneging and return mechanism, and an algorithm to include the preference behavior of the care providers.

We further contribute to research in agent-based simulation, since our research proves the usability of an agent-based approach in a real world environment by not only matching the current decision making process but also by studying a number of alternatives. The model is loaded with an extensive amount of stochastic distributions based on actual market data and successfully matches the performance of the real world system.

Finally, we contribute to research in information systems by improving the human decision-making process. Our study on the different policies on the youth health care system decreases the information overload which increases the rate of fair child allocations. This will improve socially responsible welfare decision-making.

The paper is organized as follows. In Sect. 2 we review relevant literature. Section 3 describes the foundations and structure of our simulation model which is based on real world data. We present four scheduling policies, the first one serves as a benchmark and represents the current situation, and the other three are potential alternatives for future use. This section also describes the four care types, a balanced score card analysis which serves as a basis for our benchmark, and the four Key Performance Indicators (KPI's) we develop and use to evaluate the different policies. In Sect. 4 we present experimental results using our testbed. Finally, we conclude with directions for future research.

2 Related Literature

A common approach taken by governments to tackle waiting line problems is an ad-hoc supply of monetary resources. This provides only a short term solution to the youth health care sector, as available capacity and queue lengths reach a new equilibrium after a short while [22]. [26] identified five popular approaches to decrease waiting times: monitoring of procedures, using priority scoring tools, setting waiting time targets, using an external advisory body, and registering online. However, [23] argues that such methods do not work by themselves; better coordination and flow control are proposed to increase performance at the public sector. The approach in our paper adheres to this argument by comparing a number of scheduling policies.

Regarding the scope of our research, we emphasize that our discussion on client waiting time in an health care environment distinguishes itself from appointment systems as discussed in for example [20,25]. In such settings, one distinguishes indirect waiting time, i.e. the time between request for treatment and appointment, and the direct waiting time beyond the appointed time at the health care facility, which usually is a result of the emphasis on the utilization of health care resources [15]. In our setting, the waiting time is equal to the time between diagnosis, which includes the identification of the appropriate health care package, and the moment an appointment can be made with a provider of the health care package. Therefore, both the direct and indirect waiting times related to an appointment system will be in effect only after the client has been allocated to the care provider.

Our empirical analysis has revealed that the Dutch youth health care system in which clients are waiting to be allocated to resources is subject to two behavioral patterns. First of all, the scheduling of clients may be subject to prioritization, based on certain client characteristics. Second, reneging is observed, i.e. some clients leave the system spontaneously without treatment after waiting for a certain amount of time. Both behavioral patterns have received some attention in the operations management literature. In the literature on priority classes and queueing models, the optimality of the so-called (generalized) "$c\mu$" priority rule has been established under various circumstances. This rule gives priority to customers with high marginal delay cost (c) and low expected treatment time ($1/\mu$) [27].

[3] explore the optimal capacity and cost of a queueing system in which arriving customers cannot observe their position in the queue and where they show a reneging rate linear in the queue length. However, reneging may be a more complex behavior. For example, several studies showed that the amount of time that a client is willing to wait for care is related to the urgency of the problem [12]. More urgent problems are difficult to treat elsewhere, while they genuinely require attention. These clients will accept longer waiting times. The converse holds for less urgent problems.

Most literature on waiting line management in health care is based on queuing theory and focuses mainly on resource utilization and determination of the minimum required amount of resources while maintaining a high service level [14].

[7] have emphasized the need for detailed data while analyzing queueing systems and have stated that traditional queuing theory does not capture, among other things, more complex customer reneging behavior, time-dependent parameters, and customer heterogeneity. [6] address the incongruence of behavior as modeled in the service operations management literature with the empirical findings from the behavioral literature. We have incorporated the aforementioned characteristics in our agent-based simulation model and we have calibrated the model with detailed, real-life data. Waiting line problems have also been studied using discrete event-based simulation, see for example [4,11,24]. While these studies do include more complex arrival and reneging processes, they still solely focus on utilization issues and capacity planning. For example, [11] use a generic discrete-event simulation model to investigate the feasibility of a particular national service waiting time target and present barriers, some of which related to capacity issues, to meet this target faced by the UK health care system.

Information systems in health care organizations become increasingly instrumental as they drive down the costs of services and support decision-making in complex environments. This is also high-lighted by the current debate of the digital transformation of health care [1]. As the authors point out, it is of paramount importance to learn all the significant institutional knowledge of the health care sector and therefore to collaborate with health care professionals. One of the authors of our team is a health care professional and we completely second their opinion. This has allowed us to gain deep insights into the health care sector, which would have been impossible otherwise. Furthermore, [10] show that investing in IT in the health care industry does lead to organizational profitability. Our research follows a design-oriented approach, as laid out by [16]. With the design and implementation of an agent-based [30] resource allocation decision support system we have created a valid artifact, which is relevant and necessary to solve existing problems in the health care IS domain, because it has the potential to address each of the desired features identified in this section. Agent-based approaches have successfully been applied to manufacturing supply-chain management scenarios, such as [8,9], but have not yet been used in health care systems.

Agent-based simulations, such as ours, TAC SCM [9], or Power TAC [19] along with many related computational tools are driving research into a range of interesting and complex domains that are both socially and economically important [5]. Since such experimental platforms allow market structures to be evaluated under a variety of real-world conditions and competitive pressures, they can also be used to effectively uncover potential hazards of proposed market designs in the face of strategic behaviors on the part of the participating agents. This can help policy makers in policy and regulation design.

3 The Simulation Model

In this section we describe our research framework, the different simulation model parameters and the overall model structure. Furthermore, we describe

our scheduling policies, the care types, and list the different key performance indicators that we developed to evaluate our model.

3.1 Research Framework

The research framework aims at eliciting given characteristics of the decision context and the system design requirements at various decision levels (Fig. 2). The given characteristics are retrieved and validated based on real world data from the Dutch health care sector. The model is initiated with seven youth care institutions and eight care providers in particular regions. The design requirements at the various levels are elicited from interviews and workshops. Our approach comes down to the establishment of an active modeling paradigm for system redesign that evaluates alternative strategies in a risk-free test environment, while incorporating real-world data and expert interviews ("docking").

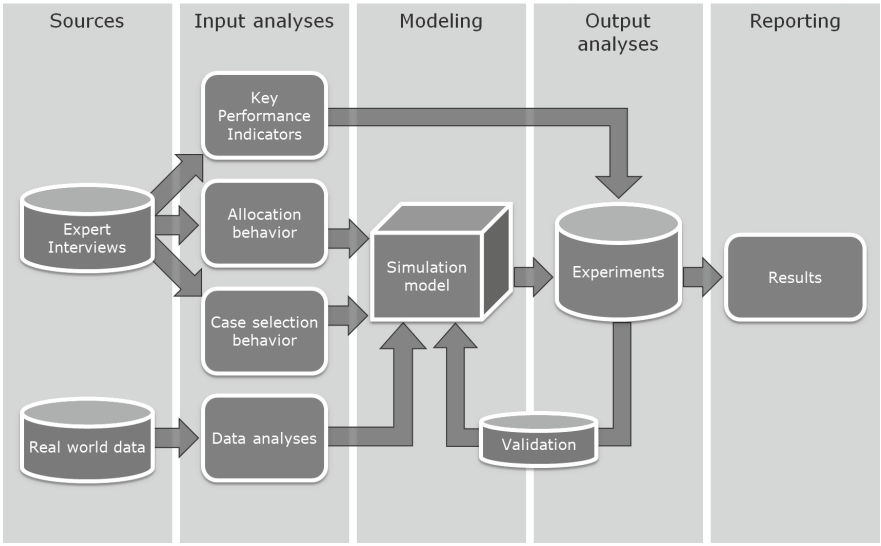

Fig. 2. Research framework.

The model is initiated as a non-terminating system since decisions and performance measures depend on long lasting developments. The model is pre-filled at start in a fully utilized state at the care providers while there are no waiting lists. This procedure will decrease the required warm-up time of the model. Warm-up time has been determined by the method of [28,29] to be 4 years simulation time. The replication length has been set to 20 years simulation time in total being 5 times the warm-up time.

The verification of the model is split in two types: First, the introduction of state-transition control and the implementation of numerous checks during the simulation which ensure a correct flow of cases through the system. Second, in-depth source review by others who didn't participate in the design of the model verified the correct coding of the model.

The validation process is split into three phases: First, the input analysis in which the input parameters of the model are calibrated with real world data. Second, two of the most important but less understood parameters of the model are analyzed for sensitivity. Third, a user validation by field experts is done. For the input analysis a comparison of stochastic variables with real world data is performed by analyzing the resulting distribution of values from the model. The theoretical and empirical distributions are visually compared. Analysis has been performed on direct parameters of the model like the arrival, age and difficulty distribution and indirect behavior of the model like reneging and return mechanisms which are partly set by parameters but are also a result of the operational behavior of the system as a whole. Table 1 lists the measures used for validation of the simulation model. At the core of the model lies the political influenced decision algorithm which makes the selection for the next to be treated case at the care provider. The algorithm chooses the case based on a trade-off between efficiency being a shorter estimated duration of care and an acceptable waiting time for the remaining cases. Since the political influenced decision algorithm is essential to the model and it cannot be directly validated against the data, we performed sensitivity analysis by adjusting the threshold values of the algorithm. The results showed moderate sensitivity on these values (further explained in Sect. 3.7). For the user validation phase we included consultation with several experts from the field of youth health care over different fields of expertise: one youth health care consultant with a high level of experience in the sector, one case manager at the institution of youth health care with operational experience, one financial director at a care provider with operational experience and some strategic experience, one director at a care provider with strategic experience and some operational experience. The results show that the model mimics expected behavior accurately. The field experts recognized much of the real world system in the model's output. For example, the arrival distribution including seasonal effects and the construction of treatment trajectories, in which a client can have simultaneous cases and return cases with crisis attribute, were found realistic representations of reality.

3.2 Model Parameters

The model takes as an input the given characteristics of the decision context, i.e. of the health care domain. These characteristics include client population characteristics such as demography and population density, the pattern of client arrivals into the system, which may include seasonal effects, the distribution of diagnostics and required medical care of the client population, and the characteristics of resources such as geographical location of the care providers and the medical expertise offered. The client arrival processes are generated during the time of the simulation by a non-stationary Poisson distribution to include a seasonal influenced arrival effect. Additional client attributes are specified such as age, home location, and case difficulty.

The non-uniform age distribution of the arriving clients is included in the simulation, as it is taken into account while allocating a client to a care provider.

Further, some cases are marked as a'crisis' and are allocated at once. These cases bypass the allocation strategy but do influence the usage of capacity in the model. A crisis denotes a case of extreme urgency and its level of difficulty can be of any kind. Each arriving child will be diagnosed with a varying amount of care needs. These needs can be indicated simultaneously at the first indication or re-indicated after reneging or a successful treatment. This also involves the analysis for reneging probabilities during the waiting phase and return probabilities after reneging or ending care. A return probability on reneging tends to be significantly higher than the probability for return after treatment. A return further involves a return interval since the child will not return immediately but after a varying amount of time. A case is provided with an identifier indicating the difficulty of the case, which is assumed to be uniformly distributed. The (expected) treatment time distributions depend on the difficulty identifier and the care provider. Table 1 lists the parameters and types of distributions as they are used in the model.

Validation of the parameters and model has been split up in three stages:

1. **Direct Validation:** Direct input parameters (like probabilities of multiple simultaneous care tracks, crisis distributions and geographical distributions) have been validated by extracting them back from the result set of the simulation. This stage ensures a correct working of the innermost basics of the model which in turn validates a correct outcome for the upcoming indirect measurements. The parameters were found to be behaving as expected.
2. **Indirect Validation:** Indirect output measures (like waiting times and lines, return rates and actual duration after being pitched on a cases' difficulty) have been validated by comparing these system measures against real world data. Specific waiting times and care durations were behaving significantly off in comparison to the real world data. The model is not capable of reflecting the same treatment and waiting trends as present in the real world. This is most likely a result of simplification of the system whereby the model smoothes results. While these measurements were off, the system as a whole functions as expected and generates comparable behavior as the real world system. The system was found to behave sufficient enough as expected.
3. **User Validation:** The model design and output measures have been validated by field experts who recognized and confirmed the behavior of the model, although the values in detail did not exactly match.

3.3 Structure of the Model

The structure of the model can be further explained while reflecting on design requirements at the strategic, tactical, and operational level. The design requirements at the operational level are supported by performance outcomes of an agent simulation model, in which behaviors have been specified that are established at the tactical level. The strategic decisions and requirements have been taken into account in the overall design and scenario analyses of the agent system, including sensitivity analyses. The scenario and sensitivity analyses ultimately serve as a tool to evaluate and compare the tactical and strategic decisions.

Table 1. Model parameters with type of distribution and short description

Parameter	Type of distribution	Description
Capacities	Absolute value	Maximum number of treatment positions available
Arrival distribution	Non-stationary Poisson distribution	Client arrivals including seasonal effects like the impact of summer holidays
Age distribution	Empirical distribution	The age of the child at arrival on which the birthdate is selected
Crises distribution	Probability (%)	The probability that a case is marked as crisis and will be allocated for immediate care
Parallel tracks	Probability (%)	The probability that there are multiple simultaneous types of care allocated at first arrival and the type of care they are
Difficulty	Uniform distribution	Classification of urgency, this is the base for all further comparisons between cases
Geographical distribution	Uniform distribution	The studied region is mapped to include distances between client and care provider, the clients are uniformly distributed over the map
Geographical range limitations	Probability (%)	The chance that a client is willing to travel mediate of high distances for his care
Care duration	Empirical distribution	Care duration per care type per care provider, the implementation in the model includes the difficulty factor (described above) to pitch the simulated durations towards the easiness or difficulty of a specific case
Reneging ratios	Calculated probability (%)	The chance that a case reneges during the waiting phase, the implementation in the model chooses the reneging date beforehand. If a case is still waiting at that date the case will be withdrawn from waiting
Return rate	Calculated probability (%)	The chance that a case will return for additional care (or care at all in case of withdrawal). In case of withdrawal the difficulty of the withdrawn case is considered relevant, the higher the difficulty the higher the chance on return. For an end of treatment the difficulty isn't considered relevant
Return interval	Uniform distribution	The interval between reneging or end of care and the return if applicable, the interval is chosen to be within the 0-180 days range

The *strategic level* decisions under consideration are push, pull and centralized, decentralized scheduling policies. The push, pull decision defines whether the care providers perform the allocation or that the decision is left to the discretion of the central youth health care bureau. The centralized, decentralized decision concerns the timing of the allocation which results in a queue only at

the central youth health care bureau or at the care providers as well. The design requirements that constrain decisions at the strategic level concern the support of basic roles and responsibilities of the stakeholders involved and how they are related, and include requirements of the methods of communication and the scope of information sharing between actors in the system.

At the *tactical level*, the design of the health care system involves the establishment of policies of several stakeholders, given the queuing structure. The decisions of the client allocation system, i.e. the output of the decision process, need to be made considering the given domain characteristics mentioned above, and design requirements at the strategic, tactical and at the operational level. The design requirements at the tactical level constrain the behavior of the stakeholders (or agents). For example, the client preferences set allocation constraints based on geographical position or other relevant data. Client urgency is based on client diagnostics and other relevant data. The way that medical experts specify acceptance factors based on urgency and other relevant factors may be constrained as well.

At the *operational level*, a control mechanism is being specified that provides a work flow in which activities and decision moments are embedded, based on decision rules established at the tactical level. The work flow establishes paths through the system consisting of activities such as application, allocation, waiting, reneging, start of care and end of care. Table 2 summarizes the structure of the model as discussed above.

We now discuss some technical aspects of the model structure. The agent-based model is written on DSOL [17]. The model features three basic agent roles: a case manager agent, a care provider agent, and a child agent. The description of the agents involves the role they represent, and the types of data that they use. We first explain these types of data and then we describe the agent roles.

There are several types of data identified in the model. First, some data define fixed values like agent names, the theoretical distributions, and the geographical home location of an agent. These parameters are mined from real world health care data and health care expert interviews. Second, there are dynamic data stores which hold process information upon which an agent can make decisions. This type of data can be divided in two groups; the transactional data store and the decision data store. The transactional data store holds records of the overall process of an agent. For example, the agents that represent the institution for youth health maintain an internal care database holding all relevant client information. The data store holds factual information emulating historical record keeping. On the other hand, the decision data stores hold time specific data relevant to the execution of the allocation strategy. The value of this data in the decision process decays over time. For example, the decision on the most appropriate care location for a particular client, as determined by the institution for youth health, is based on available information at a particular point in time. Moreover, the agent-based model provides a communication platform enforcing straightforward message based communication between the agents. All inter-agent communication passes this platform such that only those pieces of information that are passed through becomes available to other agents.

Table 2. Structure of the model.

Decision level (stakeholders)	Design requirements Decisions (model structure)
Strategic level (policy makers)	*Policy maker preferences* Organizational roles and responsiveness (model scenarios: push, pull and centralized, decentralized)
Tactical level (care providers)	*Preference behavior of actors in the system* Care types offered Acceptance ratios (decision rules, either normative or descriptive)
Operational level (all actors in the system)	*Control mechanisms and interactions* - (multi-agent system structure)

We now describe the agent roles. The case manager agents act on behalf of the institution for youth health care and they maintain a shared transactional data store for record keeping and private data stores for allocation decisions. The care provider agents all operate on their own on behalf of a care provider. They use private transactional and decision data stores for record keeping and client selection. The client agents operate on behalf of individual clients and while they use both shared and private data stores, they merely initiate the process of care inquiry at the institution for youth health care. A client agent may choose to wait for care or may decide to renege after a certain amount of time.

The process of care provision is implemented on the case level rather than the client level. A single arriving client can be signed up for multiple types of care at the same time and for each of these types a new case is generated. Each of these cases can independently renege or get care and each of these cases are independently considered for returns after reneging or care provision. There is a strict activity path that is followed by all cases in the system as illustrated in Fig. 3. The activity path includes allocation, waiting phase, and treatment mechanisms. It includes client reneging during the waiting phase and client returns after treatment or reneging. An important step in the activity path is the client allocation process for treatment at the care providers which takes place during the waiting phase of a case. It is this specific point in the process were the different allocation scenarios in this research are focused on (see Sect. 3.5). When a care provider selects the next client for treatment, he will evaluate the clients in the queue based on certain characteristics in order to match the client with the available treatment location. While clients are to be selected on a first come, first serve basis, this is often violated by the care providers because they prefer clients that are easier to treat. Easier clients lead to higher throughput which increases profit.

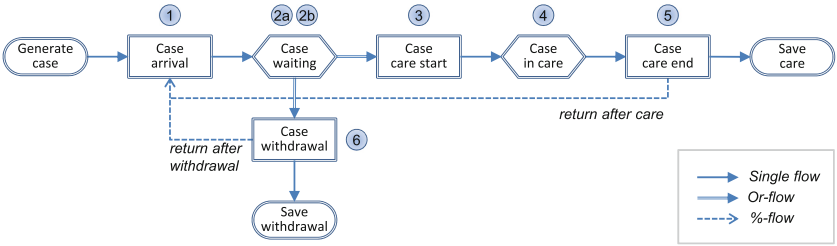

Fig. 3. Life cycle of a case with system measure points.

3.4 Model Measures

At the operational level, the model is about the cases and the events that take place to handle them, this is at the granularity level on which measurements take place. The case events are the base for measuring system performance. As Fig. 3 illustrates there is a strict path for each case implemented in the model. The figure also shows the measuring points of the system relative to the status of a case. There are two types of measures: (1) event counts; the amount of occurrences of a specific event and (2) time averages; the average amount of time spent in a specific state. We have the following system measures which are saved:

1. **Case Arrivals:** The amount of cases that are created during the replication. This includes the amount of cases created by the case generator, the amount of cases created due to returns after care and the amount of cases created due to returns after withdrawal without care. The case generator is identical for all scenarios which simulates the demand for care from the region throughout the replication and includes a correction for seasonal effects. The returns for both after care as well as withdrawal are implemented identical for all scenarios since the probability of return is related to the outcome of a case not the way the system is modeled. The outcome for the measure however can differ for both these returns since it depends on the amount of cases ending care or withdrawing. Note that a shift in the treatment portfolio from less to more difficult cases leads to higher average treatment times and therefore less treatment ends and probably more withdrawals due to capacity constraints. Simply put, one must choose to spend time on fewer difficult cases or more easier cases, while the available capacity stays the same. Returns after care are solely based on the probability of returning whilst the probability of returns after withdrawal also includes the difficulty factor which ensures that the more difficult cases tend to return more often than less difficult ones. In the end, the implementation of a scenario will have its effect on the outcome of case arrivals by influencing the returns as opposed to first arrivals.
2. **Average Waiting Time:** The average waiting time of a case until the next event, being either a start of care or a withdrawal during the replication. The measure has been split into two sub-measures to point out the difference

between a wait time resulting from waiting until a care position became available and a wait time resulting from an early withdrawal. Note that the second wait time doesn't reflect the actual waiting time of the system at that point in time but rather the amount of time the client was willing to wait for care.

3. **Starts of Treatment:** The amount of cases that started treatment during the replication. These are the cases that actually use the system resources.
4. **Average Treatment Time:** The average treatment time of a case until the end of treatment during the replication. The generation of the treatment time per case is implemented identical for all scenarios. The actual treatment time however is influenced by the difficulty factor. On average, a higher difficulty factor will yield higher treatment times and will therefore block the resource for a longer period than a lower difficulty factor would. The composition of cases that get treatment therefore influences the average treatment time and throughput on the resources.
5. **Ends of Treatment:** The amount of cases that ended treatment during the replication. Note that this measure will be equal to the starts of treatment with the absence of the cases that were still in treatment at replication end.
6. **Case Withdrawals:** The amount of cases that withdrew from waiting before a treatment position became available. Note that on average, a case with a higher difficulty factor will be willing to wait longer than a case with a lower difficulty factor. The selection behavior for who's getting the treatment of the model will therefore influence the composition of the withdrawals.

3.5 Scheduling Policies

The set of simulation experiments covers a number of variations of the model structure as exhibited in Table 2, i.e. push, pull and centralized, decentralized decision policies, the four care types, the stakeholder behavior expressed in terms of an acceptance ratio function, and sensitivity analyses.

We first consider the decision strategies.

1. **Decentralized Pushing:** Pushing cases to decentralized queues. As soon as a child has been diagnosed, the institution for youth health care pushes the case to one of the care providers. This strategy is currently implemented in the youth care sector. In this case, the care providers maintain and control their own queues. Workshops with professionals from the field revealed that the selection of children was biased by financial considerations, amongst other things. We have performed an analysis on real life selection data and have estimated a functional relationship between expected treatment time and selection likelihood (details are provided in Sect. 3.7). The institution for youth health care pushes a case to the applicable care provider with the shortest queue. This decision is based on incomplete information since the actual queue lengths at the care providers at runtime are unknown as updates are provided only periodically or upon a limited amount of requests during the allocation process.

2. **Centralized Pushing:** Pushing cases from a centralized queue. When a case has been diagnosed at the institution for youth health care, it is held in a centralized queue until capacity for the required treatment becomes available. The institution for youth health care maintains and controls the central queue while the care providers have no queue at all. The care provider announces its available capacity, and the institution of youth health care pushes the cases for treatment. Observe that any preference bias at the care providers has no impact on the allocation of cases, which is solely done by the institution for youth health care.
3. **Decentralized Pulling:** Pulling cases to decentralized queues. When a case is diagnosed at the institution for youth health care, it is published on a bulletin board in the model until it is selected by a care provider who commits future capacity to the case. The bulletin board is a passive intermediary whose sole function is to provide information to the involved agents to enable the allocation process. Both the institution for youth health care and the care providers hold queues in this strategy. In case a care provider wishes to select an easy case, it must also select all comparable cases in the queue that entered the system before the preferred case. Waiting for the preferred case to be first in line bears the risk of losing the case to another care provider. Therefore, the care providers need to balance the burden of accepting unfavorable cases against the risk of not utilizing their capacity to the full extent.
4. **Centralized Pulling:** Pulling cases from a centralized queue. When a case is diagnosed at the institution for youth health care, it is held in a centralized queue until it is pulled by a care provider which has available capacity. The institution for youth health care publishes the waiting list on a bulletin board for evaluation by the care providers. The care providers do not have queues themselves. The institution for youth health care monitors selection behavior and enforces a'first come, first serve' policy among comparable cases. Care providers have some discretion to exercise their bias by selecting favorable cases at the expense of cases that are, strictly speaking, not comparable.

3.6 Care Types

The model facilitates four types of care present in the youth care system being ambulatory care (AH), day care (DH), foster care (PZ) and residential care (RH). First we'll discuss some of the main characteristics of these care types, followed by an overview of the main differences as the main reason to study them separately.

1. **Ambulatory Care:** A child is attended at home or at at the location of a care provider by a professional social worker. It includes a series of sessions between client (and parents when useful) and a professional from the care provider. Compared to the other care types the treatment time is on the low end. This is the most basic and cheapest type of care since it only involves little time of the professional. Since the client or the professional has to travel for each session it is preferred that treatment is provided locally. The capacity

is rather high compared to arrivals and clients do not have to wait very long for treatment, since this care type is provided by all care providers.
2. **Day Care:** A child stays at the care provider during the day so that a secure and stable setting can be provided to treat the client. Care is mainly provided to a group under close professional surveillance. Due to the relative lower costs of this type of care longer treatment times are possible. Capacity is sufficient, although waiting times are generally higher, since the care is provided locally and not all care provider provide this type of care.
3. **Foster Care:** A child is actively moved from his/her parental home into a stable and secure setting at a foster family. The care is provided by foster parents who are contracted by the care provider. This care is not provided locally; in certain cases it is even preferred to get clients away from their familiar region. Treatment times are on the high end compared to the other care types, and treatment is focused on longer term solutions in which it is necessary to separate clients from the home region. The care is cost friendly, capacity is sufficient, and waiting times are at the lower end.
4. **Residential Care:** A child is moved from his/her parental home into a stable and secure setting at a location of the care provider. Residential care is seen as the most drastic intervention since it acknowledges that the child requires additional attention from professionals above the basic need to get him/her away from the parental home. Treatment time can range up to months or even years. Due to the complex nature of the treatment it is the most expensive type of care making it important to limit treatments to only the cases who genuinely require it. In practice it often happens that a child receives a combination of several care types; many children who receive residential care are also supported with an ambulatory track, which sometimes even is used as a partial substitution for the more intensive type of care. Multiple care types may also be offered simultaneously in order to reduce the queue length. It is much easier to get a child into an ambulatory track than a residential one, and by doing so a child is already receiving basic care and is considered less an urgent problem than a child who isn't getting care at all. Capacity is sufficient and waiting times are at the lower end.

3.7 Acceptance Factors

The behavior of the health care providers is partly captured by their preferences for specific types of cases. Interviews with field experts revealed that care providers in addition show a preference behavior which is not consistent with a first come first serve principle. In fact, some cherry picking is taking place. In order to capture this behavior, a preference function is introduced in the simulation model. Equation (1) describes the actual preference order of cases which has been elicited by means of a balanced scorecard technique [18], based on interviews with field experts and the evaluation of real world data containing over 30,000 care trajectories. The parameter α_{bench} is called the acceptance factor. A case with a lower acceptance value factor is preferred by the care provider. The observed behavior is parameterized into the resulting equation which consists of

two terms. The first term describes the impact of the waiting time t_{wait} of the case at the moment of evaluation and the second term describes the impact of the expected treatment time of the case t_{treat}. Equation (1) contains two fixed threshold values ϵ_{wait} and ϵ_{treat} which are estimated in such a way that the equation resembles the selection behavior as discovered during the interviews. Fine tuning has been done by visual inspection of the function' output. To illustrate the strategic behavior defined by Eq. (1), Fig. 4 shows an example of four potential cases [B,C,E,G] which is a subset of the actual waiting line [A-H] obtained by filtering on characteristics of both the clients and the open treatment position. The order of the clients in terms of waiting time (horizontal axis) differs from the preference order based on the acceptance values (right vertical axis). In Fig. 4, the order of decreasing waiting times is B-C-E-G, while the acceptance value increases along C-E-G-B. When solely looking at waiting time, client B would be selected, however the acceptance function describes a preference for client C.

$$\alpha_{bench} = \frac{\epsilon_{wait}}{t_{wait} + 1} + \frac{E(t_{treat})^2}{2\epsilon_{treat}^2} \tag{1}$$

The policies where studied with the same strategic decision making algorithm in place, given by Eq. (1). It was recognized that the algorithm would be ineffective in certain scenarios where the care providers are not able to exercise their preferences. Moreover, it can be assumed that a high level of control, exercised by the institution of youth health care in the "Centralized Pull" strategy, decreases the freedom to select clients at will. Nevertheless, the care providers will exercise this type of behavior when the design of the system permits them to do so and this phenomenon should therefore be studied accordingly. We perform a sensitivity analysis of this algorithm applied to the "Decentralized Push" allocation strategy by measuring the direct effects on waiting time. The approach is based on the continuum between a focus on the waiting time of clients, as promoted by the institution of youth health care and government, and a focus on the expected treatment time, which aligns with the economic incentives felt by the care providers. Indeed, governmental policies require that care is provided

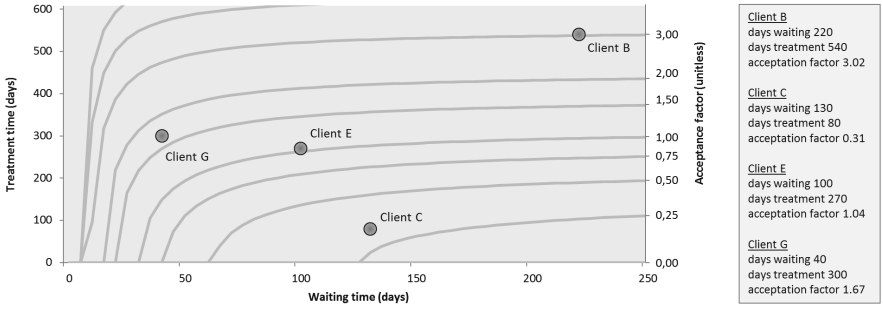

Fig. 4. Indifference curves of the strategic decision algorithm for specific acceptance factors including an example subset of clients ready for allocation.

first to the clients that have been waiting the longest. These policies are based on the recognition that clients cannot be distinguished based on urgency, so that waiting time serves as a proxy. The rationale represents a"first come-first serve" approach, which is in conflict with monetary incentives that favor clients that require the least treatment time. In our model, we study balances between acceptance rationales following governmental policies, i.e. which are based on waiting time, and acceptance rationales which are based on efficiency, i.e. treatment time. We analyze convex combinations of the two extreme rationales as described in Eq. (2). By increasing β step by step, we introduce unfairness between client selection by the care providers.

$$\alpha_{linear} = \frac{\beta\ \epsilon_{treat}}{E(t_{twait}) + 1} + (1 - \beta)\ t_{treat} \quad \text{for } \beta\epsilon[0, 1] \tag{2}$$

We also added a benchmark rationale in which the allocation is fully random as defined in Eq. (3).

$$\alpha_{random} = RAND \tag{3}$$

where RAND follows the homogeneous distribution on [0,1].

3.8 Key Performance Indicators (KPIs)

A number of Key Performance Indicators (KPIs) have surfaced while researching the interests of stakeholders in workshop discussions and desk research on professional publications, publications from the youth care sector and field data. First of all, the public health care system is bound by law to treat children in need within a reasonable amount of time, so waiting times are under scrutiny. On the other hand, it has also been recognized that reneging from queues, i.e. clients spontaneously leaving the queue after a certain period of time, may filter out those clients that are able to resolve issues by themselves. Children that need extensive care are likely not to belong to this category. However, beyond utilizing their capacity to the full extent, care providers have financial incentives to avoid the treatment of difficult cases, so there is a tendency to prioritize less difficult cases. To properly manage queues in YHC under these circumstances, we will study the system measures as discussed in Sect. 3.4 and shown below per scenario.

1. **Case arrivals**
2. **Average waiting time**
3. **Starts of treatment**
4. **Average treatment time**
5. **Ends of treatment**
6. **Case withdrawals**

KPIs 1 (less returns), 2 (less waiting time) and 6 (less withdrawals) can be seen as *social* indicators, since they relate strongly to the children who need care.

KPIs 3 (more starts of treatment), 4 (less treatment time) and 5 (more ends of treatment) can be seen as *efficiency* indicators, since they relate strongly to the overall efficiency and economic incentives of care providers.

Interviews with field experts indicated that a major shortcoming of the current system is the neglecting of difficult cases. However [2] argue that there are many cases which receive help via the institution for youth health care are not genuine cases requiring professional help. The authors indicate that these cases shouldn't enter the system because either the indication of a problem is falsely recognized or the problem is of such a low level that these are able to help themselves. The field experts support this conclusion. The discussions on these KPIs therefore includes a distinction of judgment on overall performance against a judgement based on a subdivision on the difficulty factor of cases.

4 Discussion of Results and Managerial Insights

The simulation is set to run 20 years of simulation time therefore including over 160,000 clients per replication on which a long running average waiting time is calculated. Each setting is run for 75 different seeds therefore making it possible to calculate reliable means with a 95 % confidence interval per setting.

4.1 Key Performance Indicators

In the Sect. 3.8, a number of KPI's have been studied to select the best scheduling policy. Most importantly we discussed that the difficulty of a case should be taken into account as well. As shown in Figs. 6, 7, 8 and 9, for each of the four scheduling policies, confidence intervals for means are presented for subsets of cases (left side) and all cases (right side). Please note that the scales differ among the figures to ensure comparability of values within a figure. The subsets of cases are created with bins of 10 % difficulty ranging from 0.0-0.1 (less difficult) to 0.9-1.0 (more difficult). The number of cases ending up in the bins is not equal since a bin is created on the difficulty factor itself rather than the resulting set of cases. I.e. the amount of treatment starts for bin 0.0-0.1 at day care in the push to central scenario (S1) with about 1200 treatments differs from the about 700 treatments for bin 0.9-1.0. Note that this also means that the waiting and treatment times are calculated on differently sized subsets. Vertically, the

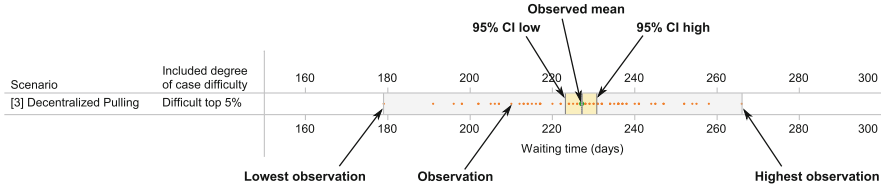

Fig. 5. Example of visualization method for result analysis.

confidence intervals for means are presented using five vertical lines indicating the 95 % confidence levels. See Fig. 5 for further guidance in reading the results.

4.2 Comparison of Scheduling Policies

Based on the KPI analyses as outlined in Figs. 6 to 9, we gain the following insights:

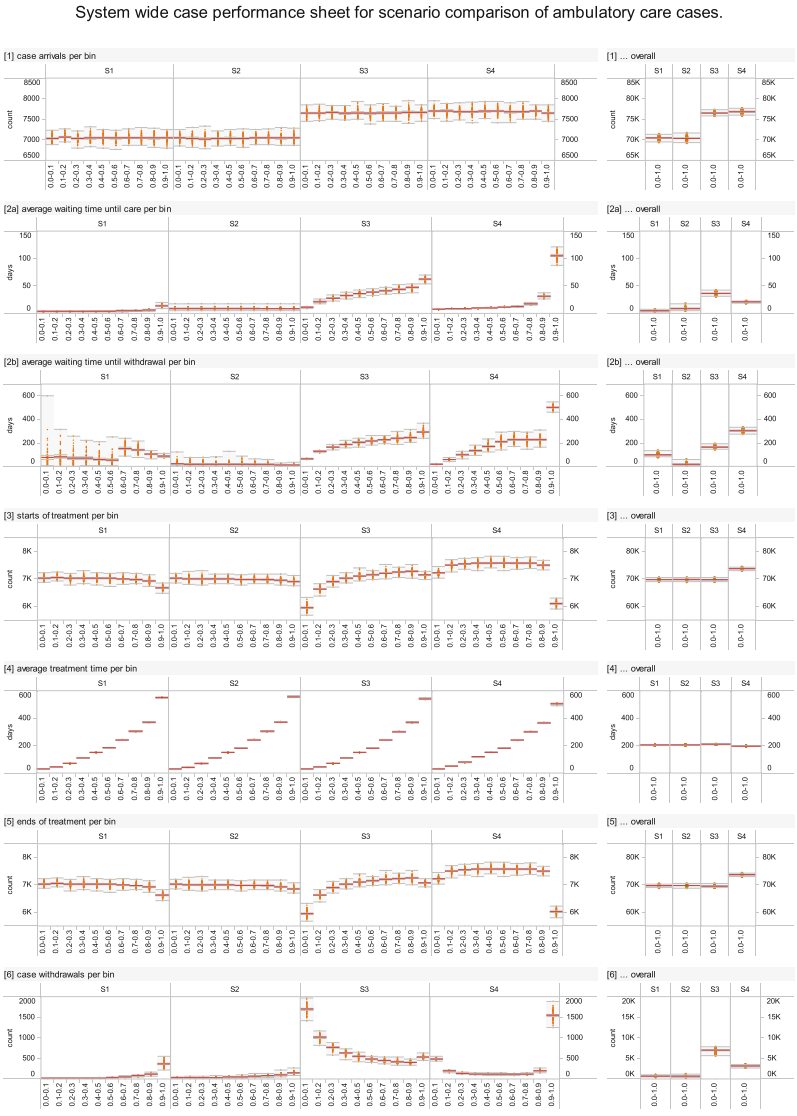

Fig. 6. Comparison 1: System analysis for Ambulatory Care.

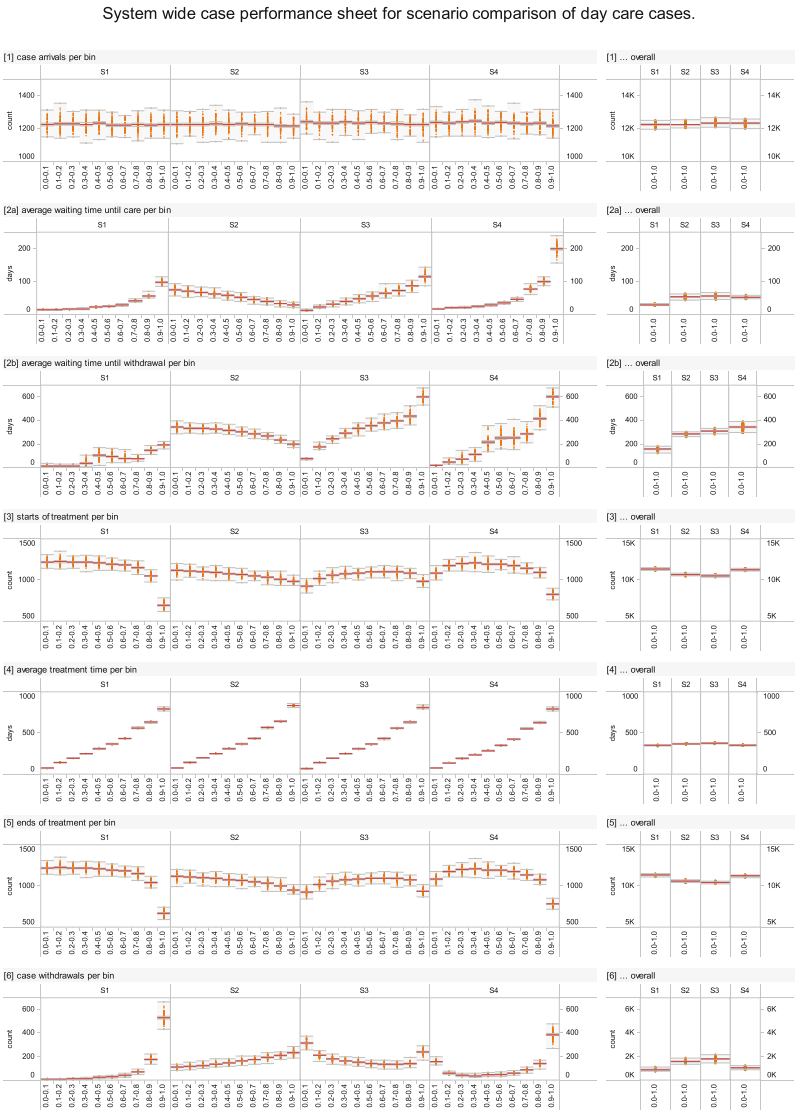

Fig. 7. Comparison 2: System analysis for Day Care.

Insights Ambulatory and Residential Care. For these two care types, the pull policies show a 10 % higher rate of arrivals due to returned withdrawals; this is especially true for the difficult cases. The central pull policy enables enforcement of "fairness", which can be inferred from the number of start events of the cases at the various difficulty levels. On the contrary, the decentralized pull policy generally neglects the most difficult cases, and the total throughput is the highest. Although the system is efficient, returned difficult cases create waiting lines and are not being treated.

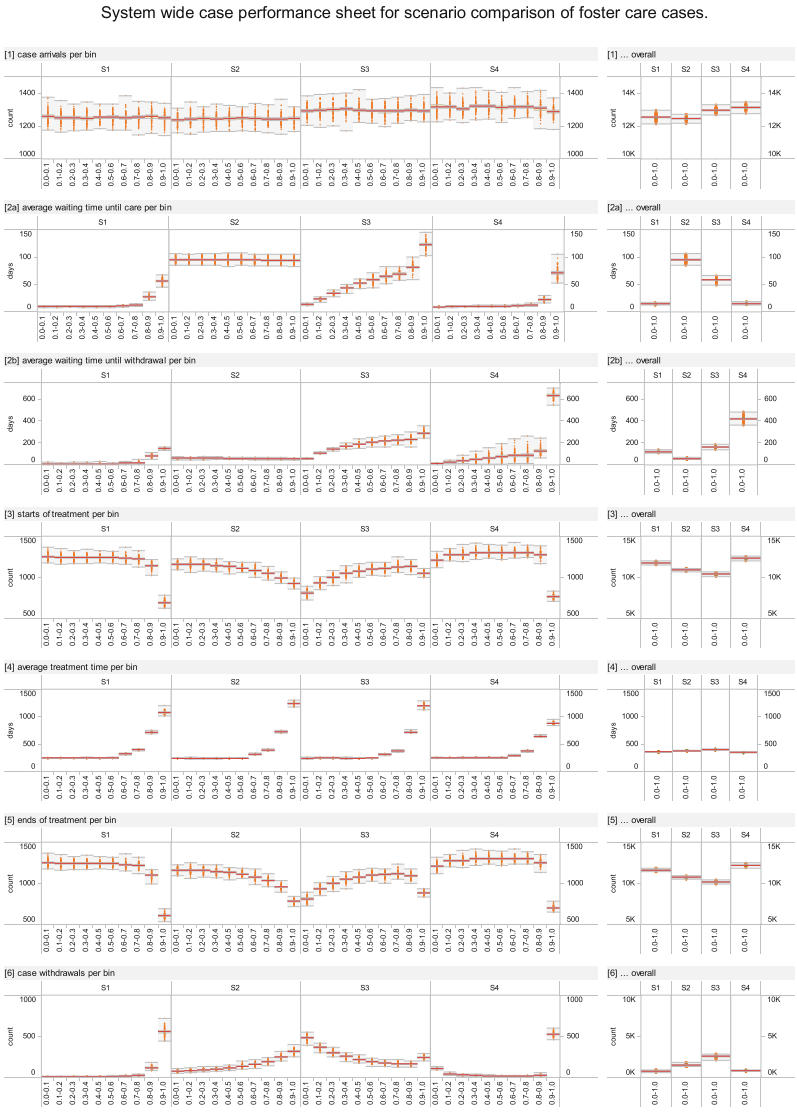

Fig. 8. Comparison 3: System analysis for Foster Care.

Insights Foster and Day Care. Arrivals pull and push policies are almost equal for these two care types, while for Ambulatory and Residential care types, pull and push policies show differences. Centralized policies maintain a certain degree in fairness. Due to increased waiting times, more easy cases withdraw. Therefore, we see that the difficult cases are more often treated than the easy cases. Decentralized policies tend to treat an equal amount of cases on all levels of difficulty,

except the most difficult ones which are treated significantly less. Most of these effects are more pronounced for Day care.

Policy Comparison Over all Care Types. For push scenarios, withdrawal rates positively relate to case difficulty, while for pull scenarios, both easy and difficult cases show higher withdrawal rates.Since all systems demonstrate an increase of withdrawals when waiting times increase, there is less difference in performance

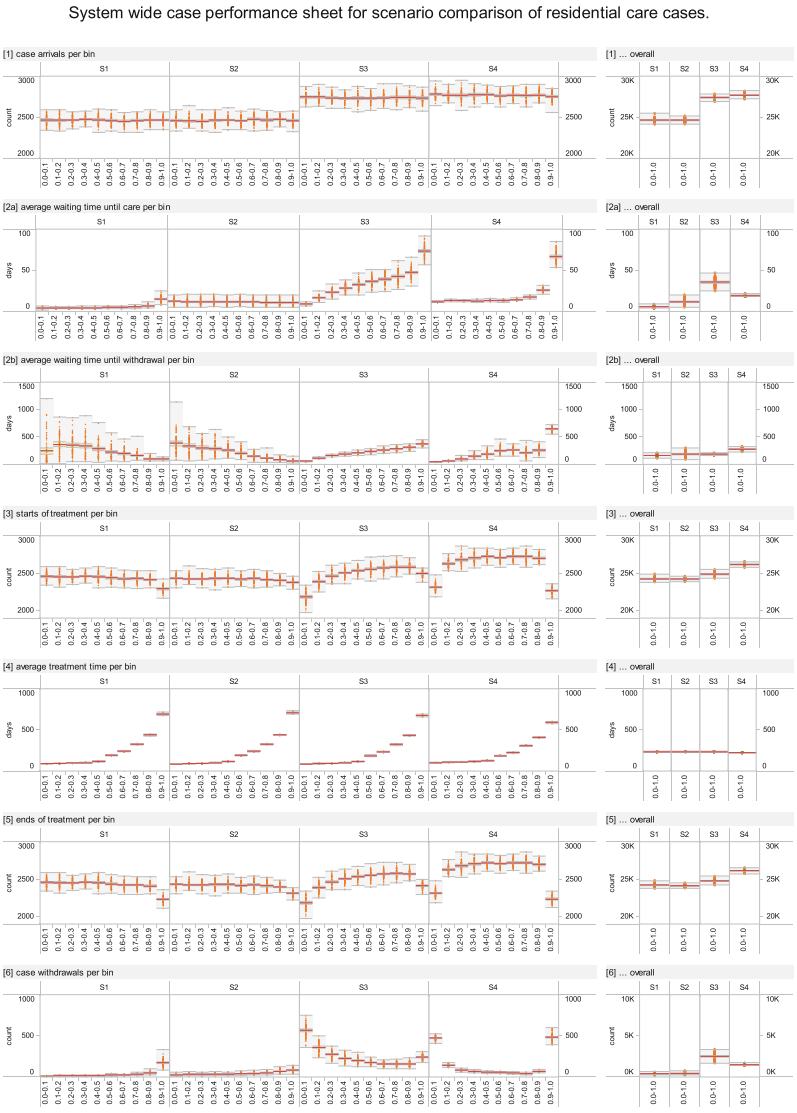

Fig. 9. Comparison 4: System analysis for Residential Care.

among the policies under such circumstances.While the decentralized policies have room to increase throughput by choosing easy cases over difficult cases, the centralized policies maintain fairness in the system, which comes at the cost of lower throughput.

Managerial Takeaways. One may argue that from a fairness viewpoint, push is good, pull is bad, and central is good, decentral is bad. Therefore, central push is good, and decentral pull is bad, while the mixed scenarios are in the middle. Under moderate workload conditions, a bad policy performs (service level) just slightly worse compared to a good policy, on average. On the contrary, under high workload conditions, a bad policy is more efficient than a good policy, on average. A bad policy under moderate workload neglects the difficult cases, and thereby creates additional workload when cases return, while a good policy handles all cases without problems. However, when the workload increases, accepting the difficult cases is affecting general performance.

In relation to current developments in the sector in which allocation is shifting from a central point (province level) to a more distributed point (municipality), these observations become very relevant. Instead of a few connections between care providers and allocators, there will be many. And instead of a few allocators who know each other, there will be more allocators without direct working relationships. This situation increases the level of ambiguity between care providers and allocators and decreases the level of oversight from the allocators. Therefore, these smaller allocation units will be in a weaker position to enforce a pull scenario and there will be more room for cherry picking at the care providers.

The central question is: should the system focus on fairness? If yes, there should be a centralized allocation management. In such a case, however, one needs to be willing to accept the costs of lower throughput of the easier cases. The model shows that in a situation where these cases can leave the system without returning (i.e., the cases resolve themselves), then it will have no noticeable effect on the overall workload of the system. On the other hand, one could argue that focus should be on the throughput to help as much clients as possible. In such a case, one should create a backup option for the neglected cases that would otherwise not be treated at all. This scenario however is not strong in itself, since (1) the case in question will have waited already for too long before it makes use of the backup option, and (2) the mere presence of the backup option legitimates cherry picking. In particular, it will be difficult to decide at what difficulty level the backup option becomes the preferred one.

5 Conclusions and Future Work

We have presented a versatile computational approach for analyzing a number of resource scheduling policies in the youth health care sector while including an extensive set of constraints and behaviors from the real world domain. The model successfully simulated many of the complex and dynamic relations between the involved parties in the healthcare sector. We demonstrated the ability of the

model to incorporate different scheduling policies while maintaining an overall structure which deals with the common tasks outside the scheduling procedure. We discussed the differences between the scenarios and their impact on system performance. The introduction of a case's difficulty into performance measurement leads us to the advice of the push from a centralized scheduling policy for future resource scheduling in the youth health care sector. The postponement of the actual allocation in this policy ensures a higher level of fairness in treatment provision by the care providers because they cannot avoid the difficult cases anymore which increases overall social welfare.

Our approach shows the importance of agent-based modeling in complex, dynamic environments like the youth health care sector where much of the issues are related to coordination and communication between different heterogeneous parties. We contribute to research in service operations management by not only showing its usability in such a setting, but also showing the ability to study alternative scenarios which couldn't be studied otherwise with this level of complexity.

Generally, our findings show that a scheduling system which includes a reneging mechanism can handle a workload that is bigger than the available resources to the system while the system as a whole appears to be stable by using the reneging mechanism as a filter on arrivals. As we see in this health care case the measurement methods for performance (which can be translated as a key parameter for the rewarding structure) are out of balance with the social goals of the system and therefore the filtering effect is indirectly used to increase measured performance while social performance is neglected. When it is not possible to bring the measurement methods in line with the social goals, then it should be assured that there is no room for cherry picking. In this case it can be arranged by postponing the actual allocation towards the point that an independent party can make the final decision and ensure that this decision is in line with the social goals. Furthermore when the decision power is positioned at the party who has to perform upon this decision it becomes even more important that the performance indicators are in line with the social goals otherwise performance based preferences have an even stronger negative effect on the social performance.

The current model incorporates basic methods to emulate interdependencies between the available care types. In future, we plan to study the model in alternative configurations with varying settings for geographical distributions and number of agents in such that we are able to assist in strategic decision making.

References

1. Agarwal, R., Guodong (Gordon), G., DesRoches, C., Jha, A.K.: The digital transformation of healthcare: Current status and the road ahead. Inf. Syst. Res. **20**(4), 796–809 (2010)
2. Andriessen, S., Besseling, J.: Jongeren zijn steeds vaker niet normaal. Jeugd Beleid **2**(1), 87–95 (2008)
3. Armony, M., Plambeck, E., Seshadri, S.: Sensitivity of optimal capacity to customer impatience in an unobservable m/m/s queue (why you shouldnt shout at the dmv). Manufact. Serv. Oper. Manage. **11**(1), 19–32 (2009)

4. Bagust, A., Place, M., Posnett, J.W.: Dynamics of bed use on accommodating emergency admissions: Stochastic simulation model. Br. Med. J. **X**, 319 (1999)
5. Bichler, M., Gupta, A., Ketter, W.: Designing smart markets. Inf. Syst. Res. **21**(4), 688–699 (2010)
6. Britan, G.R., Ferrer, J.C., e Oliveira, P.R.: Managing customer experiences: Perspectives on the temporal aspects of service encounters. Manuf. Servi. Oper. Manage. **1**(1), 61–83 (2008)
7. Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., Zhao, L.: Statistical analysis of a telephone call center. J. Am. Stat. Assoc. **100**(469), 36–50 (2005)
8. Collins, J., Ketter, W., Gini, M.: A multi-agent negotiation testbed for contracting tasks with temporal and precedence constraints. Int. J. Electron. Commer. **7**(1), 35–57 (2002)
9. Collins, J., Ketter, W., Sadeh, N.: Pushing the limits of rational agents: the trading agent competition for supply chain management. AI Mag. **31**(2), 63–80 (2010)
10. Devaraj, S., Kohli, R.: Information technology payoff in the health-care industry: a longitudinal study. J. Manage. Inform. Syst. **16**(4), 41–67 (2000)
11. 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)
12. Goldman, R.D., Macpherson, A., Schuh, S., Mulligan, C., Pirie, J.: Patients who leave the pediatric emergency department without being seen: case-control study. Can. Med. Assoc. J. **172**(1), 39–43 (2005)
13. Goodacre, S., Webster, A.: Who waits longest in the emergency department and who leaves without being seen? Emerg. Med. J. **22**(2), 93 (2005)
14. Gorunescu, F., McClean, S.I., Millard, P.H.: A queueing model for bed-occupancy management and planning of hospitals. J. Oper. Res. Soc. **53**, 19–24 (2002)
15. Gupta, D., Denton, B.: Appointment scheduling in health care: Challenges and opportunities. IIE Trans. **40**(9), 800–819 (2008)
16. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design science in information systems research. Manage. Inf. Syst. Q. **28**(1), 75–106 (2004)
17. Jacobs, P.H.M., Lang, N.A., Verbraeck, A.: D-sol: A distributed java based discrete event simulation architecture. X, ed., In: Proceedings of the 2002 Winter Simulation Conference. San Diego, pp. 793–800. ISBN 0-7803-7614-5 (2002)
18. Kaplan, R.S., Norton, D.P.: The balanced scorecard: Measures that drive performance. Harvard Bus. Rev. **83**(7), 172–180 (2005)
19. Ketter, W., Collins, J., Reddy, P.: Power TAC: A competitive economic simulation of the smart grid. Energy Econ. **39**, 262–270 (2013)
20. Liu, N., Ziya, S., Kulkarni, V.G.: Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. Manuf. Serv. Oper. Manag. **12**(2), 347–364 (2010)
21. Netherlands National News Agency, NANP. 2008. Millions of additional funding for youth care
22. Postl, B.D.: Final report of the federal advisor on wait times. Technical Report, Health Canada (2006)
23. Rachlis, M.: Public solutions to health care wait lists. Technical Report, Canadian Centre for Policy Alternatives (2005)
24. Ridge, J.C., Jones, S.K., Nielsen, M.S., Shahani, A.K.: Capacity planning for intensive care units. Eur. J. Oper. Res. **105**, 346–355 (1998)
25. Robinson, L.W., Chen, R.R.: Estimating the implied value of the customer's waiting time. Manuf. Serv. Oper. Manage. **13**(1), 53–57 (2011)

26. Saulnier, M., Shortt, S., Gruenwoldt, E.: The taming of the queue: Toward a cure for health care wait times. Technical Report, Canadian Medical Association (2004)
27. Van Mieghem, J.: Dynamic scheduling with convex delay costs. Ann. Appl. Probab. **5**(3), 809–833 (1995)
28. Welch, P.D.: On the problem of the initial transient in steady-state simulation. IBM Watson Research Center (1981)
29. Welch, P.D.: The statistical analysis of simulation results. The computer performance modeling handbook 268–328 (1983)
30. Wooldridge, M., Jennings, N.R.: Intelligent agents: Theory and practice. Knowl. Eng. Rev. **10**(2), 115–152 (1995)