Journal Pre-proof

A nested Benders decomposition-based algorithm to solve the three-stage stochastic optimisation problem modeling population-based Breast Cancer Screening

Tine Meersman, Broos Maenhout, Koen Van Herck

PII: S0377-2217(23)00310-7
DOI: https://doi.org/10.1016/j.ejor.2023.04.027
Reference: EOR 18448

To appear in: *European Journal of Operational Research*

Received date: 7 July 2022
Accepted date: 13 April 2023

Please cite this article as: Tine Meersman, Broos Maenhout, Koen Van Herck, A nested Benders decomposition-based algorithm to solve the three-stage stochastic optimisation problem modeling population-based Breast Cancer Screening, *European Journal of Operational Research* (2023), doi: https://doi.org/10.1016/j.ejor.2023.04.027

Highlights

- Modelling population-based Breast Cancer Screening
- Patient rescheduling and cancellation to mitigate patient no-shows
- Formulation of three-stage stochastic programming problem
- Integrating resource planning and patient scheduling
- A SAA heuristic based on nested Benders decomposition and diving

A nested Benders decomposition-based algorithm to solve the three-stage stochastic optimisation problem modeling population-based Breast Cancer Screening

Tine Meersman*[1], Broos Maenhout[1], and Koen Van Herck[2]

[1]Faculty of Economics and Business Administration, Ghent University, Tweekerkenstraat 2, 9000 Gent (Belgium)
[2]Belgian Cancer Registry, Koningsstraat/Rue Royale 215 Box 7, 1210 Brussels (Belgium)

Abstract

Population-based cancer screening programmes invite high-risk population groups to screenings in order to increase the probability of an early diagnosis. In this paper, we study the organisation of preventive breast cancer screening, accounting for both scheduling of patients and planning of resources. Mammography screening comprises a two-stage healthcare process, encompassing a patient scan in a mammography unit and a scan examination by a central coordination center. Objectives are to minimise patient flow times and maximise resource efficiency and number of patients treated. Performance of population-based screening programmes is hampered due to large rates of patient no-shows, which is partially remedied by giving patients the option to cancel or reschedule their appointment. We model the scheduling problem as a three-stage stochastic optimisation problem and propose a diving heuristic relying on Sample Average Approximation and nested Benders decomposition to find high-quality integer solutions. Computational experimentation is performed on real-life instances to benchmark the proposed method to alternative methodologies. Results demonstrate that the proposed heuristic yields a stable performance for instances of different sizes. However, integrating the three decision stages increases significantly the complexity, such that larger-sized instances are preferably solved via two separate solution stages to improve resource efficiency. In addition, insights are provided in the value of stochastic optimisation and strategies to cancel or reschedule appointments mitigating the impact of no-show uncertainty.

Keywords: OR in health services; Multi-stage scheduling; Population-based cancer screening; Sample Average Approximation; Nested Benders decomposition

*Corresponding author - Email: Tine.Meersman@UGent.be

1 Introduction

In order to decrease the incidence and mortality of breast cancer in women, most countries in Europe have set up population-based periodic breast cancer screening that targets all women between 50 and 69 years. Cancer screening programmes invite the most high-risk population group to get a screening on a regular basis in order to increase the probability of an early diagnosis. The breast cancer screening programmes in Europe in particular have lead to a significant decrease in the disease's mortality rate with studies reporting decreases between 28% and 36% (Broeders et al., 2012). Central to the success of population-based breast cancer screening are careful planning and a well-organised programme that ensure coordination, continuity and quality of care operations. Breast cancer screening is a two-stage healthcare process, encompassing a scan appointment and a scan examination. In this paper, we study the organisation of the Breast Cancer Screening Programme (BCSP) relative to the region of Flanders, Belgium, centrablacklly coordinated by the Centre of Cancer Research. The organisation of the population-based programme is very complex and costly since it is resource-intensive and subject to a large degree of patient no-show uncertainty, which is partially countered by giving patients the option to cancel or reschedule their appointment. Coordination between different care providers is required by the Central Coordinator (CC) as the scan taking and examination of the scan take place in different organisations at distant locations. The screening is typically undertaken in outpatient settings of a local care centre, i.e. an acknowledged Mammography Unit (MU) in the proximity of the patient. Figure 1 displays the process flow related to the treatment and analysis relative to a single patient. In advance, the CC invites patients for a mammography scan and assigns these to a specific time and date as well as an MU based on a few demographic parameters. When a patient is informed of her scan appointment, she has three weeks to respond by requesting the rescheduling or cancellation of the appointment. This can be done by contacting the CC via the indicated phone number. When choosing to reschedule, the patient is given the opportunity to indicate her preferences regarding date and MU for rescheduling her appointment. If the patient does not respond, she is expected to show up at the designated time and place. If the patient shows up to the appointment, the scan taken is examined twice, i.e. on the very same day by a radiology specialist from the visited MU and by a specialist attached to the CC centre who is experienced in mammography quality control. Note that, due to required administration the scan examination by the specialist of the CC centre can only be carried out at the earliest one day later than the associated scan appointment and examination at the MU. In case they come to a different conclusion, the scan is examined again by a second specialist from the CC centre. However, as the latter is rarely needed, the third examination is not considered further in our study. After all necessary examinations are completed, the patient is informed about the result.

The coordination of the BCSP is entangled by uncertainty, i.e. patients can cancel or request to reschedule their appointment or may simply not show up at the time of the appointment. Because of this uncertainty, the CC may be forced to adapt devised schedules based on realised uncertainties. Correspondingly, three major decision phases can be discerned, i.e. the resource planning and baseline scheduling, the decisions necessary to adapt schedules based on patient cancellations and rescheduling requests, and the decisions to adapt schedules based on the patient (no-)shows. The optimisation objectives are to minimise the time between scan appointment and examination in order to improve patient satisfaction with respect to the waiting time for results while organising the screening programme in an efficient manner by minimising resource costs and maximising the number of patients treated. These objectives are facilitated by the integration of resource planning and demand scheduling of fig.png

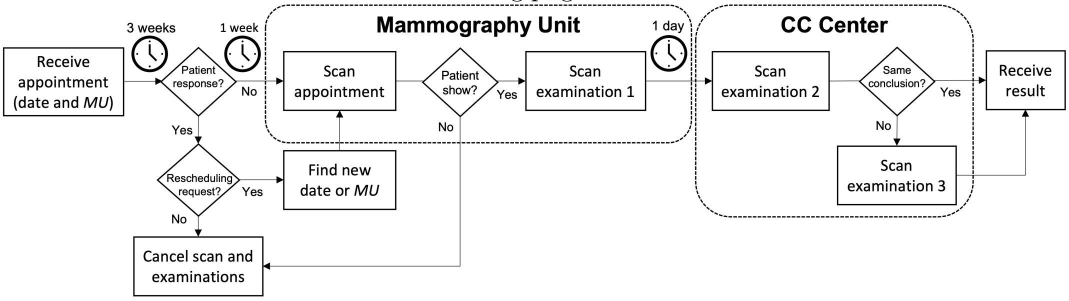


Figure 1: Illustration of the process flow relative to the screening and examination of an individual patient.

patient screenings and scan examinations. Both phases are interconnected since the patient scheduling is a driver for the resource requirements and vice versa, when scheduling patients, assumptions are typically made with respect to the resource availability. Isolated reasoning leads to unbalanced schedule outcomes leading to significantly higher personnel costs and patient waiting times and may have a negative impact on the clinical outcome (Chen et al., 2008). The aims of this research study are to (i) model the organisation of the real-life multi-stage healthcare process relative to the planning of required resources, which is inherently a static budgeting decision, and the scheduling of appointments for scan screenings and subsequent scan examinations under uncertainty; (ii) design an efficient algorithm that finds high-quality solutions in an acceptable time span for the underlying stochastic optimisation problem; and (iii) provide insights into specific features of the BCSP, relative to the process organisation, the scheduling of patients and the planning of resources.

In correspondence to the different decision phases, the problem is solved as a three-stage stochastic programming problem. The proposed solution methodology applies Sample Average Approximation (SAA) in order to reformulate the three-stage stochastic program into a Deterministic Mixed-Integer Linear Program (DMILP). In this way, we approximate the expected objective value in each stage with the respective average objective value over a restricted set of scenarios generated based on underlying probability distributions. To solve this large deterministic problem, we propose a three-step heuristic solution methodology. In the first step, we apply a Nested Benders Decomposition (NBD) algorithm to solve the Linear Programming Relaxation (LPR) of the DMILP. In the second step, a greedy diving heuristic, intertwined with the NBD algorithm, is used to derive a high-quality integer solution. Given a relaxed solution, the diving heuristic repetitively fixes a set of binary variables to integer values. When no integer solution is yielded in the second step, a post-processing procedure is applied in a third step that solves the different stages consecutively in a forward manner. The proposed methodology is tested using real-life data from the BCSP in Flanders. In the computational experiments, we validate the design choices of the heuristic and benchmark the algorithm to (i) multi-stage solution methodologies that optimise the DMILP exactly via mathematical programming, and (ii) decomposed solution methodologies that optimise different single or multiple decision stages separately in a sequential manner. In addition, we provide insights in the benefits of stochastic optimisation and consideration of mitigation strategies to cancel or reschedule appointments to reduce the impact of no-show uncertainty.

The remainder of the manuscript is organised as follows. In Section 2, relevant literature is discussed. In Section 3, the problem description and formulation are provided. Section 4 describes the proposed solution methodology. Section 5 discusses performed computational experiments and validates the proposed procedure. Concluding remarks and directions for future research are discussed in Section 6.

2 Related literature

Healthcare planning and demand scheduling problems have been the subject of many studies (cf. overview papers of Cayirli and Veral (2003) and Marynissen and Demeulemeester (2019)). However, no studies have investigated the coordination of population-based screening programmes. In Section 2.1, we categorise related research based on the (multi-stage) patient process, considered objectives and decisions taken. In Section 2.2, we discuss the considered types of uncertainty, relevant stochastic modelling and solution methodologies.

2.1 Integrated multi-objective and multi-Stage healthcare scheduling problems

Scheduling problems can be divided into single-stage and multi-stage problems (Marynissen and Demeulemeester, 2019). In single-stage scheduling, patients go through one process (e.g. Dogru and Melouk, 2019; Jiang et al., 2019; Kim and Mehrotra, 2015; Wu and Zhou, 2022). Multi-stage scheduling problems involve processes in which patients are served via multiple treatments or stages, and are commonly represented via well-known job-shop and flow-shop models, making distinction between non-sequential and sequential types of patient processes, respectively (Zhou and Yue, 2019). Sequential problems in healthcare consider a stream of patients who follow the same sequence of services, resembling flow-shop problems. This sequence often starts with an appointment, followed by multiple service processes, possibly requiring the presence of patients. In surgery scheduling, for example, different authors (Aissaoui et al., 2020; Zhang et al., 2020) study a multi-stage flow-shop problem in which patients only make an appointment for the first surgery stage and then go through subsequent stages (e.g. recovery) in the same order as their appointments. Zhou and Yue (2019) model a similar four-stage sequential patient scheduling problem. Alike job-shop problems, non-sequential multi-stage problems in healthcare involve patients who may follow different pathways. Pham and Klinkert (2008) study a surgical-case scheduling problem with multiple patient types as a multi-mode blocking job-shop problem. White et al. (2011) optimise an appointment scheduling problem where the appointment is followed by additional stages or tasks depending on the patient's needs on a First-Come-First-Served (FCFS) basis. Gocgun and Puterman (2014) and Diamant et al. (2018), amongst others, model a dynamic patient scheduling problem that considers the scheduling of patients who require multiple treatments. The problem divides the infinite planning horizon into multiple periods and account for the correlation between consecutive decision periods, for example, because of patient deferrals, appointment cancellations and patient no-shows. Diamant et al. (2018) consider a non-sequential problem as the evolution of required treatments is dependent on completed assessments.

Healthcare planning and scheduling problems can also be categorised based on decisions to be taken, i.e. resource capacity planning, patient demand scheduling, or both types of decisions in an integrated manner. The latter problems have not been frequently studied. Zhang et al. (2020) integrate demand and capacity scheduling in a two-stage surgery scheduling problem by determining which surgical blocks to open, selecting the surgeries to be performed from a waiting list, and assigning the selected surgeries to available surgical blocks. White et al. (2011) determine the number of examination rooms and the appointment scheduling rule for planning patients in a radiology department. They found a significant relationship between the two decisions, confirming there is interaction between capacity and demand scheduling decisions. Wu and Zhou (2022) integrate the optimisation of number of working duties and appointment scheduling for a service with multiple servers. Aissaoui et al. (2020) consider simultaneously the sequencing of patients and resource assignment to schedule surgeries.

In multi-stage scheduling problems, it is common practice to consider multiple objectives simultaneously, which are optimised via a single weighted objective function (Marynissen and Demeulemeester, 2019). A distinction can be made between objectives related to patient scheduling and to resource capacity planning. Common objectives for scheduling patients are patient waiting time and flow time, resource idle time and overtime, and total completion time. For example, Zhou and Yue (2019) minimise expected costs related to patient waiting time and resource idle times. Dogru and Melouk (2019) minimise expected costs of patient waiting time, physician idle time and overtime to schedule patient appointments. Common resource objectives are resource utilisation, idle time, overtime and amount of resources used. Kim and Mehrotra (2015), for example, minimise nurse wages, costs for adding shifts, shift cancellation costs and the penalty for under- and overstaffing. Problems that integrate resource planning and patient scheduling oftentimes combine patient- and resource-related objectives. Zhang et al. (2020) maximise the number of surgeries performed and minimise surgery postponement costs based on patient priority. Moreover, they minimise opening and overtime costs of the surgical blocks. White et al. (2011) account for patient waiting time and the amount of idle time and overtime as performance measures. Wu and Zhou (2022) minimise number of working duties, amount of idle time and overtime, and patient waiting time. Aissaoui et al. (2020), however, focus only on patient-related objectives such as patients' length of stay and number of patient overnight stays. The latter objective is related to the patient flow time as an overnight stay for a patient is mainly dependent on the scheduled moment for the surgery and the uncertain recovery duration.

In this paper, we study a multi-stage sequential scheduling problem that integrates capacity planning and demand scheduling. We optimise number of patients scheduled, patient flow time, fixed and variable resource costs, resource-schedule adjustment costs and resource idle time and overtime. Table 1 in Online Appendix A indicates the characteristics discussed above for the most relevant references and our study involving number of processes, process type, type of decisions (patient- or resource-related), and studied objectives.

2.2 Stochastic scheduling

Stochastic optimisation problems take at least one source of parameter uncertainty into account. Section 2.2.1 discusses different types of uncertainty considered in the domain of healthcare planning and demand scheduling. Section 2.2.2 discusses the common techniques to model multi-stage stochastic problems. Section 2.2.3 discusses relevant solution methodologies. Table 2 in Online Appendix A provides a literature synthesis of the most relevant studies regarding the considered types of uncertainties and solution methods.

2.2.1 Sources of uncertainty

The literature on stochastic healthcare planning and scheduling oftentimes only includes stochastic service durations (e.g. White et al., 2011). However, in appointment scheduling, no-shows, patient punctuality, length of stay and arrival processes are also commonly studied sources of uncertainty. Dogru and Melouk (2019), amongst others, consider patient preferences as a source of uncertainty. These preferences are related to the appointment day and time and physician resource, and are used to determine the set of appointment slots offered to the patient. In capacity planning literature, the demand for services is regularly considered as a stochastic variable (e.g. Kim and Mehrotra, 2015). In this paper, we consider three types of parameter uncertainties, i.e. patient cancellations, patients' requests for rescheduling based on their preferences, and no-shows. Patient preferences and especially no-show behaviour have already been widely studied, whereas appointment cancellations and reschedules are less common topics. Though, patient cancellations can be seen as a type of demand uncertainty and, more precisely, as the amount of demand that is overestimated. To the best of our knowledge, patient preferences for rescheduling requests have not been considered in the literature before.

2.2.2 Modelling of stochastic multi-stage problems

Multi-stage stochastic optimisation problems assume some recourse decisions can be taken after uncertainty is disclosed. The stages in a stochastic problem are related to before and after a random experiment and may comprise different decision phases at different moments, though they can coincide. Most studies focus on two-stage programming problems (e.g. Dogru and Melouk, 2019; Jiang et al., 2019; Kim and Mehrotra, 2015; White et al., 2011; Wu and Zhou, 2022; Zhang et al., 2020; Zhou and Yue, 2019). However, in reality, multiple scheduling decisions are often made at different points in time and uncertainties are often realised in between those decision stages, encompassing a multi-stage stochastic problem, which can be represented via either a nodal or stage-scenario formulation with implicit or explicit nonanticipativity constraints (Vitali et al., 2021). Many studies rely on a nested decomposition structure that divides the planning horizon into multiple periods (e.g., days, weeks) and enables the definition of a dynamic programming model or Markov Decision Process (MDP) model (e.g. Diamant et al., 2018; Gocgun and Puterman, 2014). The downside of these techniques is that they may suffer from the curse of dimensionality and only small-sized stochastic problems can be solved efficiently. In an attempt to reduce the number of states, proposed MDP models in the patient scheduling literature rely on the aggregate numbers of patients related to specific classes, which is not suitable for the problem under study because of individual patient preferences for rescheduling. Another common approach is to reduce the number of stochastic stages. The latter has been observed in the literature, for example in the domain of stochastic production planning and scheduling (Drexl and Haase, 1995; Karimi et al., 2003), according to which so-called big-bucket and small-bucket time-period models have been proposed considering the level of uncertainty respectively in an aggregated manner relative to a larger period or in a detailed manner relative to smaller time periods. For example, Gocgun and Puterman (2014) increase their time bucket size by aggregating all patient arrivals of one day and, at the end of that day, assigning them to appointments. In line with a big-bucket time-period model, an often-used benchmark for multi-stage stochastic problems is the static approximation assuming perfect information (e.g. Dogru and Melouk, 2019). In this study, we adopt a similar approach and rely on a big-bucket time-period model with three stochastic stages to derive insights into specific features of the BCSP (e.g. value of no-show mitigation strategies).

2.2.3 Solution methodologies

Different heuristic and exact methods have been proposed to solve multi-stage stochastic scheduling problems that incorporate some parameter uncertainty. Aissaoui et al. (2020)

present a DMILP that maximises a mitigation measure to schedule slack time between consecutive patients in order to make a robust schedule without the need for sampling. White et al. (2011) and Dogru and Melouk (2019) propose both a simulation-optimisation method. Simulation-optimisation methods use simulation as evaluation method, and can be applied with many optimisation methods such as statistical ranking and selection methods, heuristics, stochastic approximation, etc. Diamant et al. (2018) and Gocgun and Puterman (2014) both apply Approximate dynamic programming to solve their formulated MDP for scheduling patients in a dynamic manner. Most methods to solve stochastic problems, however, comprise the L-shaped method and the SAA method, which try to formulate a deterministic model by sampling the stochastic variables. Kim and Mehrotra (2015), for example, apply the L-shaped method enhanced with a heuristic branch-and-cut strategy to solve their two-stage stochastic problem. Jiang et al. (2019) and Zhang et al. (2020) rely on SAA to reformulate their two-stage stochastic problems and then apply a column-generation-based heuristic and a Benders decomposition-based heuristic, respectively. It is also possible to combine SAA and the L-shaped method (e.g. Wu and Zhou, 2022; Zhou and Yue, 2019). Wu and Zhou (2022) enhance their L-shaped method with a Benders decomposition-based heuristic and a variable-neighbourhood descent heuristic. Note these studies emphasise that the quality of yielded solutions is highly dependent on the selected sample size. In this regard, Kim and Mehrotra (2015) consider 1000 scenarios for a planning horizon of 18 weeks and Zhou and Yue (2019) consider 5000 scenarios for smaller-sized instances of only 10 to 14 jobs. Zhang et al. (2020) rely on 150 scenarios to solve instances up to 150 tasks and 5 workdays. Wu and Zhou (2022) sample 100 scenarios to schedule at most 20 customers with 5 servers. Jiang et al. (2019) consider 20000 scenarios with, on average, 20 patients who have to be assigned an appointment time. Similar to Jiang et al. (2019), we reformulate our three-stage stochastic optimisation problem using SAA and then solve it via a Benders decomposition algorithm and a diving heuristic.

3 Problem description and formulation

In this paper, we study a multi-objective, multi-phase healthcare optimisation problem under uncertainty integrating resource capacity planning and advance patient scheduling. The problem determines the required resource budget accurately by constructing a baseline resource schedule and an operational patient schedule that assigns offline a set of elective patients over a medium-term horizon to particular days and specific MU and CC resources. Section 3.1 presents a general problem description and the modelling assumptions. Section 3.2 provides the stage-scenario formulation.

3.1 General problem description and assumptions

The preventive BCSP in Flanders (Belgium) involves two sequentially connected and interrelated service processes, i.e. (i) the taking of a scan during an appointment with the patient followed by (ii) the examination of the scan at a later moment. Accordingly, the underlying patient scheduling problem encompasses a flexible flow-shop problem as each patient goes through the service processes in the same order. In contrast to the scan appointment, the scan examination does not require the presence of the patient. In order to complete these two services, different resources are needed. Taking a scan requires a mammography machine at a local MU. We assume that a MU disposes of enough personnel or other resources to handle any appointment planned. A scan examination is performed by a specialist at the

CC. Evidently, the output of the first service, i.e. the made scans, determines the demand for the second service, i.e. examination of these scans. Because of their interrelationship, these processes and associated resources should be scheduled and budgeted in an integrated manner, such that the capacity is adequately coordinated to handle service demand. The problem under study determines the required (strategic) capacity via the composition of a (tactical) baseline resource schedule that is integrated with the planning of patients following a multi-phase decision framework, accounting for different types of uncertainty, i.e. (i) cancellation of scans by patients, (ii) rescheduling of an appointment on the request of patients, and (iii) no-show of patients. The realisation of these uncertainties (or stochastic variables) at different points in time trigger three main decision phases T (index $t \in T, T = \{1, 2, 3\}$). The first decision phase composes the baseline schedule before any of the uncertainties are realised. The second and third decision phases represent adaptations to the schedule based on changed input data resulting from the realised uncertainties. Section 3.1.1 comprises a more detailed description of the phases and relevant assumptions taken.

The goal of this multi-phase scheduling problem is to optimise the expected values relative to both patient- and resource-related objectives. The patient objectives encompass maximising the number of patients planned and minimising the patient flow time. The resource objectives aim to minimise (i) the number of resources budgeted and the number of work duties, (ii) adjustments made to the baseline resource schedules, (iii) idle time and overtime of both the MU and the CC resources. The idle time and overtime are optimised through patient overbooking and resource (re-)scheduling decisions throughout the decision phases. The overbooking of scan appointments and examinations is possible because capacity restrictions are treated as soft constraints, and is used to counteract the negative effects of cancellations and no-shows on resource usage. The budgeted number of resources and scheduled work duties, together with the consideration of minimising the amount of idle time and overtime in the objective function introduces a natural upper bound to the used capacity. The problem formulation of each phase includes related realised and expected objectives. The realised objectives of a given phase can be calculated directly as they are determined based on already-known uncertainty outcomes and decision variables of the given phase. The expected objectives depend on uncertainties that will be realised in the future, and/or on decisions that will be made in future phases.

3.1.1 Decision phases

Figure 2 gives an overview of the decision phases and timeline, i.e. relevant horizons and moments when decisions are made and uncertainties are realised. The figure shows the hierarchy in the decisions from top to bottom, whereas the expected objective values related to lower hierarchical phases are taken into account in the decision-making of higher hierarchical phases, following a bottom-up approach. The figure illustrates the three-phase decision process by displaying an example schedule for the different phases in response to the encountered uncertainty. The first main decision phase ($t = 1$) encompasses the composition of an integrated baseline schedule relative to both resources and patients in a static manner. The CC determines the appointments (time, date, and place) for a set of patients together in an offline manner approximately 30 days in advance, for a future planning horizon. Apart from the initial schedule of scan appointments, the baseline scheduling involves the preliminary construction of a schedule for associated scan examinations together with the capacity planning of required resources and scheduling of the budgeted resources for the involved MUs and CC. The resource schedule determines which and how many MU and CC resources are used, and schedules the work duties for these resources over the entire planning horizon resulting in an accurate estimation of the resource budget. The following assumptions are made related to the first decision phase:

(1) Patients are homogeneous in terms of resource usage.
We assume individual patients require the same service time related to their scan appointment and subsequent examination. As patients are interchangeable, it makes no difference (i) which individual patients are planned on a particular day from the perspective of the resource organisation, and especially (ii) in which sequence patients are scheduled and assigned to a time slot. Therefore, the problem under study only considers the advance planning of patients, assigning them to a particular day and resource, and not to a particular time slot.

(2) The budgeted resources can be determined based on a static decision with a single (reference) period as the planning horizon.
The considered set of patients to be scheduled over a specific planning horizon is determined by the CC based on very few parameters. Accordingly, the patient sets relative to different planning horizons are more or less invariant in terms of size and composition. Taking into account the homogeneity of patients (cf. supra), the required number of resources can be adequately determined based on a static decision that considers a single planning horizon as reference period. Note that small fluctuations in the number of patients (e.g., due to the rescheduling of patients to a later planning horizon) or in the realised uncertainties (e.g., patient no-shows) are mitigated via changes in the resource schedules and not via the postulated resource capacities.

After the first decision phase, patients are informed of their scan appointment and are allowed to respond. Based on the patient cancellations and rescheduling requests submitted during the patient response horizon, the second main decision phase ($t = 2$) is induced. In this phase, the budgeted number of resources as well as the baseline MU schedule are fixed conforming to the decisions made in the first stage. Adjustments to the schedules involving patient scan appointments and examinations and the work duties of CC specialists, however, are possible. The appointment rescheduling takes individual patient preferences into account, whereas scan examinations at the CC can be rearranged almost freely as they do not require the patient's presence. The schedule of the CC specialists is modified by removing or adding work duties. The following assumptions are made in view of this second decision phase:

(3) The probability of a cancellation or rescheduling request is independent of the patient and of the appointment day and resource assigned in the baseline schedule.
Patients are heterogeneous with respect to their appointment preferences, which may define a patient-specific probability to cancel or request a rescheduling. However, as we assume the selected set of patients to be homogeneous (cf. supra), patient preferences are considered as patient independent when constructing the initial baseline schedule, similar to most studies in the literature (e.g. Dogru and Melouk, 2019; Wang and Fung, 2014). Correspondingly, the probability to cancel or request a rescheduling is considered as independent of patients and their initial appointment set.

(4) Patients who request a rescheduling accept the first new appointment that is proposed.
When choosing to reschedule, patients are given the opportunity to indicate their preferences for rescheduling their appointment relative to the date and MU and a new appointment is set based on a negotiation between the CC and the patient. The CC will always offer a new appointment fitting the preferences of the patient, even if this means the patient must be planned in the next planning horizon. Conforming to previous literature (e.g. Dogru and Melouk, 2019), we assume patients will therefore accept the first offered appointment.

(5) Rescheduling requests and cancellations are treated in a batch-wise static manner.
Although cancellations and requests for rescheduling of different patients may arise at different moments in time, the adaptations to the patient schedules and the resource schedules for the relevant period are done for all patients simultaneously in a static manner, i.e., a couple of days before the start of the considered period taking all information into account. Accordingly, we assume patients who request a rescheduling receive their new appointment at the end of the response period. This assumption is similar to many studies on surgery and specialty scheduling (Gupta and Denton, 2008). In this way, (unnecessary) changes to the resource schedules for a particular period are kept to a minimum and personnel and patients can still be informed in a timely manner (Marynissen and Demeulemeester, 2019; Ouelhadj and Petrovic, 2009).

(6) Patients cannot concurrently request a rescheduling and cancel their appointment.
The CC does not allow patients who cancel their appointment, to request a new appointment in the same planning horizon. If this happens, patients are included in a later planning horizon. Additionally, it very rarely occurs that a patient who requests a rescheduling, also cancels the appointment later during the three-week response period. This is assumed to be a consequence of the fact that the CC always offers a new appointment consistent with the patient's preference. Therefore, we assume patients cannot request a rescheduling and cancel in the same patient response period. Note, however, we do assume patients who requested to reschedule, may not show at the time of the appointment.

During the execution of the planning horizon, patients may not show up to their appointment. Based on the realised patient (no-)shows, the third main decision phase ($t = 3$) is triggered. In this phase, the CC can adapt the schedules involving the patient scan examinations and the CC specialists by adding or removing work duties. The following assumption is made to manage this third main decision phase:

(7) The CC has perfect information relative to the realized patient no-shows at the start of the considered planning horizon.
Due to the minimum time lag between the scan appointment and scan examination, the CC can adjust the scan examination and specialist schedules as the planning horizon progresses, based on the (no-)shows of patients at the end of each day when the scan appointments realised on that day are known. To lower the computational complexity for solving the problem under study, we approximate this complex MDP via its static lower bound (cf. Section 2.2.2) and assume that the CC has complete knowledge, i.e. perfect information, relative to the realized patient no-shows at the start of the planning horizon. In this way, the consequential decisions in this phase are taken in a static manner via a single scheduling decision, rather than tackling this phase via multiple (daily) decisions and the patient no-show information becomes only gradually known. This assumption is tested and further investigated in Online Appendix B.

The identification of the decision phases has been specifically based on assumptions (5) and (7), which provide the rationale for the definition of only three decision phases, aggregating the realised uncertainty over a larger period. Given the posed research questions, this modelling approach has been found most efficient to derive relevant estimations for the complex multi-stage or dynamic process.

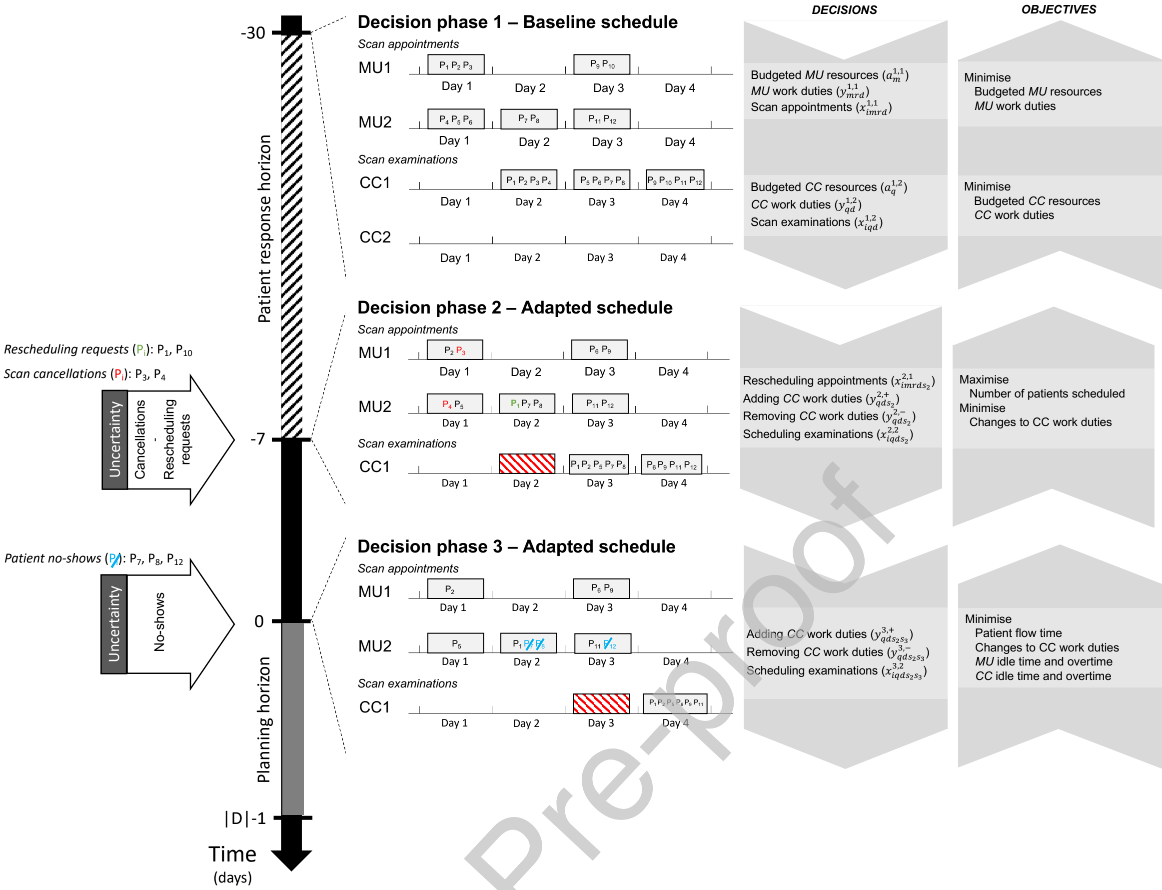


Figure 2: Illustration of the three-phase decision process.

3.2 Stage-scenario formulation

We present a stage-scenario formulation for the stochastic problem under study in which the above-discussed decision phases are modelled as ‘decision stages’ and a ‘scenario’ denotes a possible realisation of all stochastic variables (Birge and Louveaux, 2011). More precisely, a scenario indicates for each patient if they cancel, request a rescheduling and/or do not show up at the appointment time. Let ξ be the set of all possible scenarios for the stochastic variables. As mentioned, these stochastic variables are realised at different points in time. Patient cancellations and requests for rescheduling are realised between the first and second decision phases and no-shows between the second and third decision phases. We assume some dependency between uncertainties, i.e. (i) a patient cannot cancel and reschedule and (ii) a patient cannot cancel and no-show, realised at the same or a different point in time respectively during the decision process. Accordingly, the realised scenarios can be represented in a discrete manner via a so-called scenario tree illustrated in Figure 3 and the following notation. The scenario tree divides into branches corresponding to different realisations of the random sources of uncertainty. Set S_t is the set of realised samples for stage t (index $s_t \in S_t$, $S_t = \{0, ..., |S_t| - 1\}$). We define a realised sample of stage t as a collection of one randomly sampled value per stochastic variable realised between decision stages $t-1$ and t, given stages $1, ..., t-1$. To formulate the dependencies between uncertainties realised at different points in time, we denote a node of stage t in the scenario tree with a vector

$\overrightarrow{s_t} = (s_1, ..., s_t)$ with parameters $s_1 \in S_1, ..., s_t \in S_t$ indicating the history of realisations. As no uncertainties are realised before the first stage, there is only one realised sample in stage 1, i.e. $S_1 = \{0\}$. A node of stage t can therefore be represented as $\overrightarrow{s_t} = (0, ..., s_t)$, representing a (partial) path in the scenario tree between the root node and a node at level t. The probability node $\overrightarrow{s_t} = (0, ..., s_t)$ occurs is equal to $\rho_{s_2 ... s_t}$. In this way, a scenario $\overrightarrow{s_3} = (0, s_2, s_3)$ corresponds to a complete path from the root node to a last-stage node of the scenario tree and has the probability $\rho_{s_2 s_3}$. The number of possible scenarios is equal to $|S_2| \times |S_3|$.

In the following, we formulate the objective function and the constraints related to each decision stage. The multi-stage problem under study considers minimisation objective (1) subject to first-stage constraints (2)-(13), second-stage constraints (14)-(27), and third-stage constraints (28)-(41). To define the variables associated with each stage, we specify the stage number in the upper index and the related node is indicated by the lower indices. Note that the first-stage decisions are taken without any information on the random events, whereas decisions relative to decision stages 2 and 3 depend on realised uncertainties or samples. Accordingly, we index both the realisations of uncertainties and the decisions related to stages 2 and 3 by a sample index s_2 and/or s_3 representing the involved node $(0, s_2)$ or $(0, s_2, s_3)$, respectively. The decision variables related to stages 1 and 2 do not contain all scenario indices as these decisions cannot anticipate every possible outcome. In this way, the staged recourse decisions are modelled via a nodal representation and the non-anticipativity of the decisions is implicitly embedded in the variable definition. Hence, no explicit non-anticipativity constraints are required to model the multi-stage problem (Barro et al., 2006; Menon et al., 2021). The following notation of symbols is used to formulate the stage-scenario model for the problem under study:

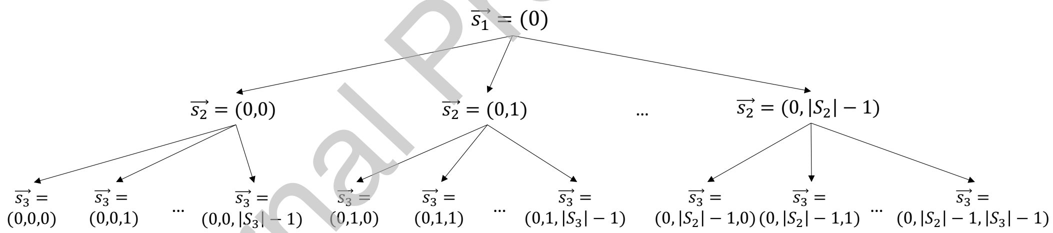


Figure 3: Scenario tree representation defining the set of possible scenarios ξ.

Deterministic sets

I	Set of patients (index $i \in I$, $I = \{0, 1, ..., |I| - 1\}$)
M	Set of mammography units (MUs) (index $m \in M$, $M = \{0, 1, ..., |M| - 1\}$)
M_i	Set of MUs accessible to patient i (index $m \in M_i$, $M_i \subseteq M$)
R_m	Set of resources available in MU m (index $r \in R_m$, $R_m = \{0, 1, ..., |R_m| - 1\}$)
Q	Set of resources available at the CC (index $q \in Q$, $Q = \{0, 1, ..., |Q| - 1\}$)
D	Set of days within the planning horizon (index $d \in D$, $D = \{0, 1, ..., |D| - 1\}$)
S_2	Set of realised samples for stage 2 (index $s_2 \in S_2$, $S_2 = \{0, 1, ..., |S_2| - 1\}$)
S_3	Set of realised samples for stage 3, per realised sample of stage 2 (index $s_3 \in S_3$, $S_3 = \{0, 1, ..., |S_3| - 1\}$)

Deterministic parameters

c_1	Cost per MU used (in euro)
c_2	Cost per budgeted CC specialist (in euro)
c_3	Cost per scheduled mammography machine per day (in euro)
c_4	Cost per scheduled CC specialist per day (in euro)
c_5	Cost per adaptation to the CC resource schedule in stage 2 (in euro)
c_6	Cost per adaptation to the CC resource schedule in stage 3 (in euro)
c_7	Cost per hour idle time for MU resources (in euro)
c_8	Cost per hour overtime for MU resources (in euro)
c_9	Cost per hour idle time for CC resources (in euro)
c_{10}	Cost per hour overtime for CC resources (in euro)
w^{AP}	Weight to maximise number of scheduled patients in stage 2
w^{FT}	Weight per day of patient flow time in stage 3
w^I	Weight assigned to all patient-related objectives
w^R	Weight assigned to all resource-related objectives
g^{CC}	Maximum number of CC specialists planned per day
l	Minimum number of days between the scan appointment and the scan examination
k_m^{MU}	Session duration at MU m (in hours)
k^{CC}	Session duration at the CC (in hours)
θ^1	Average service duration of a scan appointment (in hours)
θ^2	Average service duration of a scan examination (in hours)
Sampled values from stochastic sets and variables	
$\overline{\Gamma}_{is_2}$	Set of appointment MUs and days acceptable to patient i in node $(0, s_2)$ (tuple (m, d), with $m \in M_i$ and $d \in D$)
$\overline{\delta}_{is_2}^{canc}$	1, if patient i wants to cancel the scan appointment in node $(0, s_2)$; 0, otherwise
$\overline{\delta}_{is_2}^{resch}$	1, if patient i wants to reschedule the scan appointment made in stage 1, in node $(0, s_2)$; 0, otherwise
$\overline{\delta}_{is_2s_3}^{show}$	1, if patient i showed up to the scan appointment in node $(0, s_2, s_3)$; 0, otherwise
Decision variables related to decision stage 1	
$a_m^{1,1}$	1, if appointments are planned for MU m; 0, otherwise
$a_q^{1,2}$	1, if CC specialist q is budgeted; 0, otherwise
$y_{mrd}^{1,1}$	1, if machine r of MU m is deployed on day d; 0, otherwise
$y_{qd}^{1,2}$	1, if CC specialist q is scheduled on day d; 0, otherwise
$x_{imrd}^{1,1}$	1, if patient i is scheduled a scan appointment on day d on machine r of MU $m \in M_i$; 0, otherwise
$x_{iqd}^{1,2}$	1, if the scan examination of patient i is scheduled on day d for specialist q; 0, otherwise
Decision variables related to decision stage 2	
$y_{qds_2}^{2,+}$	1, if an extra work duty is scheduled for specialist q on day d given node $(0, s_2)$; 0, otherwise
$y_{qds_2}^{2,-}$	1, if the work duty of specialist q on day d is removed given node $(0, s_2)$; 0, otherwise
$x_{imrds_2}^{2,1}$	1, if patient i is scheduled a scan appointment on day d on machine r of MU $m \in M_i$ given node $(0, s_2)$; 0, otherwise
$x_{iqds_2}^{2,2}$	1, if the scan examination of patient i is scheduled on day d for specialist q given node $(0, s_2)$; 0, otherwise
$z_{is_2}^2$	1, if patient i is scheduled given node $(0, s_2)$; 0, otherwise
Decision variables related to decision stage 3	
$y_{qds_2s_3}^{3,+}$	1, if an extra work duty is scheduled for specialist q on day d given node $(0, s_2, s_3)$; 0, otherwise
$y_{qds_2s_3}^{3,-}$	1, if the work duty of specialist q on day d is removed given node $(0, s_2, s_3)$; 0, otherwise
$x_{iqds_2s_3}^{3,2}$	1, if the scan examination of patient i is scheduled for specialist q on day d given node $(0, s_2, s_3)$; 0, otherwise
$it_{mrds_2s_3}^{3,1}$	The amount of idle time on machine r of MU m on day d given node $(0, s_2, s_3)$ (in hours)
$ot_{mrds_2s_3}^{3,1}$	The amount of overtime on machine r of MU m on day d given node $(0, s_2, s_3)$ (in hours)
$it_{qds_2s_3}^{3,2}$	The amount of idle time for CC specialist q on day d given node $(0, s_2, s_3)$ (in hours)
$ot_{qds_2s_3}^{3,2}$	The amount of overtime for CC specialist q on day d given node $(0, s_2, s_3)$ (in hours)

3.2.1 Objective function

The objective of this multi-stage stochastic problem (eq. (1)) is to optimise the expected objective value Z^*, which is calculated by taking the weighted average over all possible scenarios (ξ). The different terms in the objective are dependent on different stochastic variables and decision variables related to particular decision stages. In this way, the objective can be split based on the stage after which they are realised (cf. Figure 2). The components Z_1^*, $Z_2^*(\overrightarrow{s_2})$ and $Z_3^*(\overrightarrow{s_3})$ denote the (expected) objective values resulting from first, second and third-stage decisions, respectively, with $Z_t^*(\overrightarrow{s_t})$ the objective value impacted by decisions made in stages $1, ..., t$, given the realised samples related to node $\overrightarrow{s_t}$. The expected value of $Z_t^*(\overrightarrow{s_t})$ can be calculated by taking the weighted average over all nodes of stage t with the weight equal to the node probability $\rho_{s_2 \ldots s_t}$.

$$\min \quad Z^* = \quad Z_1^* + \sum_{s_2 \in S_2} \left(\rho_{s_2} \times Z_2^*(\overrightarrow{s_2})\right) + \sum_{s_2 \in S_2} \sum_{s_3 \in S_3} \left(\rho_{s_2 s_3} \times Z_3^*(\overrightarrow{s_3})\right) \tag{1}$$

Equations (2)-(4) indicate the functions of objective components Z_1^*, $Z_2^*(\overrightarrow{s_2})$ and $Z_3^*(\overrightarrow{s_3})$, respectively. Z_1^* minimises fixed and variable resource costs for both MUs and CC specialists. $Z_2^*(\overrightarrow{s_2})$ maximises the total number of patients scheduled to stimulate rescheduling of patients and minimises variable costs involving the number of work duties scheduled for CC specialists and the number of adjustments made to the specialist schedule. $Z_3^*(\overrightarrow{s_3})$ aims to minimise patient flow time, variable costs involving the number of work duties scheduled for CC specialists, the number of adjustments made to the specialist schedule, idle time and overtime relative to the usage of MU and CC resources. The flow time is defined as the positive difference in days between day of scan examination and day of scan appointment if the patient showed up. Note that the eventual day of a patient's scan appointment is determined in stage 2, whereas the day of scan examination can still be adjusted in stage 3.

$$Z_1^* = \quad w^R \times \left(c_1 \times \sum_{m \in M} a_m^{1,1} + c_2 \times \sum_{q \in Q} a_q^{1,2} + c_3 \times \sum_{m \in M} \sum_{r \in R_m} \sum_{d \in D} y_{mrd}^{1,1} + c_4 \times \sum_{q \in Q} \sum_{d \in D} y_{qd}^{1,2}\right) \tag{2}$$

$$\begin{aligned} Z_2^*(\overrightarrow{s_2}) = \quad & w^I \times \left((-w^{AP}) \times \sum_{i \in I} z_{is_2}^2\right) \\ & + w^R \times \left(c_4 \times \sum_{q \in Q} \sum_{d \in D} (y_{qds_2}^{2,+} - y_{qds_2}^{2,-}) + c_5 \times \sum_{q \in Q} \sum_{d \in D} (y_{qds_2}^{2,+} + y_{qds_2}^{2,-})\right) \qquad \forall s_2 \in S_2 \end{aligned} \tag{3}$$

$$\begin{aligned} Z_3^*(\overrightarrow{s_3}) = \quad & w^I \times \left(w^{FT} \times \sum_{i \in I} \left(\overline{\delta}_{is_2 s_3}^{show} \times \Big(\sum_{q \in Q} \sum_{d \in D} (d \times x_{iqds_2 s_3}^{3,2}) - \sum_{m \in M} \sum_{r \in R_m} \sum_{d \in D} (d \times x_{imrds_2}^{2,1})\Big)\right)\right) \\ & + w^R \times \left(c_4 \times \sum_{q \in Q} \sum_{d \in D} (y_{qds_2 s_3}^{3,+} - y_{qds_2 s_3}^{3,-}) + c_6 \times \sum_{q \in Q} \sum_{d \in D} (y_{qds_2 s_3}^{3,+} + y_{qds_2 s_3}^{3,-})\right. \\ & + c_7 \times \sum_{m \in M} \sum_{r \in R_m} \sum_{d \in D} it_{mrds_2 s_3}^{3,1} + c_8 \times \sum_{m \in M} \sum_{r \in R_m} \sum_{d \in D} ot_{mrds_2 s_3}^{3,1} \\ & \left. + c_9 \times \sum_{q \in Q} \sum_{d \in D} it_{qds_2 s_3}^{3,2} + c_{10} \times \sum_{q \in Q} \sum_{d \in D} ot_{qds_2 s_3}^{3,2}\right) \qquad \forall s_2 \in S_2, \forall s_3 \in S_3 \end{aligned} \tag{4}$$

3.2.2 First-stage constraints

Constraints (5) - (7) are related to the scheduling of scan appointments for all considered patients together. Equation (5) indicates each patient i must receive exactly one scan appointment, relying on variables $x_{imrd}^{1,1}$. These variables are only defined for the set of geographically accessible MUs (M_i) for a particular patient. Further, the baseline schedule for MU resources reveals the required budget, i.e. number and specific MUs employed, relying on variables $a_m^{1,1}$, and the specific mammography machines brought into service on certain days, relying on the variables $y_{mrd}^{1,1}$. Constraint (6) stipulates that a machine can only be scheduled ($y_{mrd}^{1,1} = 1$) if the associated MU is elected ($a_m^{1,1} = 1$). Moreover, constraint (7) imposes that an appointment can only be scheduled ($x_{imrd}^{1,1} = 1$) if an (accessible) MU is available and operating ($y_{mrd}^{1,1} = 1$), linking the appointment schedule to the baseline MU

schedule. A mammography machine accommodates up to one session per day. In this way, the MU resource schedule determines the fixed costs for opening MUs and the variable costs for deploying machines. In the scheduling of scan appointments, we take into account the available MU machines per day. There are no other (hard) capacity constraints imposed restricting the number of patients planned per day.
Constraints (8)-(11) are related to the scheduling of scan examinations. Scan examinations are planned by assigning each patient i to a specialist q and a day d via variables $x^{1,2}_{iqd}$, which is enforced via equation (8). Constraint (9) indicates scan examinations can be scheduled no sooner than l days after the associated scan appointment due to administrative reasons. The scheduling of scan examinations drives the staff planning problem of CC specialists, which determines the required number of specialists, relying on variables $a^{1,2}_q$. To that purpose, a baseline schedule is composed by assigning specialists to days using variables $y^{1,2}_{qd}$. Constraint (10) stipulates that a CC specialist can only be scheduled ($y^{1,2}_{qd} = 1$) if (s)he is staffed ($a^{1,2}_q = 1$). Constraint (11) indicates that at most g^{CC} specialists can be assigned to the same day due to limited workspace. Constraint (12) ensures that a scan examination can only be assigned, if a CC specialist is scheduled, linking the examination schedule to the baseline specialist schedule. In this way, the CC resource schedule determines fixed costs for staffing specialists and variable staffing costs. Constraint (13) represents variable domains, stating that all variables considered in this stage are binary.

$$\sum_{m\in M_i}\sum_{r\in R_m}\sum_{d\in D} x^{1,1}_{imrd} = 1 \qquad \forall i \in I \tag{5}$$

$$y^{1,1}_{mrd} \le a^{1,1}_m \qquad \forall m \in M, \forall r \in R_m, \forall d \in D \tag{6}$$

$$x^{1,1}_{imrd} \le y^{1,1}_{mrd} \qquad \forall i \in I, \forall m \in M_i, \forall r \in R_m, \forall d \in D \tag{7}$$

$$\sum_{q\in Q}\sum_{d\in D} x^{1,2}_{iqd} = 1 \qquad \forall i \in I \tag{8}$$

$$\sum_{m\in M_i}\sum_{r\in R_m}\sum_{d\in D} \left((d+l)\times x^{1,1}_{imrd}\right) \le \sum_{q\in Q}\sum_{d\in D}(d \times x^{1,2}_{iqd}) \qquad \forall i \in I \tag{9}$$

$$y^{1,2}_{qd} \le a^{1,2}_q \qquad \forall q \in Q, \forall d \in D \tag{10}$$

$$\sum_{q\in Q} y^{1,2}_{qd} \le g^{CC} \qquad \forall d \in D \tag{11}$$

$$x^{1,2}_{iqd} \le y^{1,2}_{qd} \qquad \forall i \in I, \forall q \in Q, \forall d \in D \tag{12}$$

$$a^{1,1}_m \in \{0,1\} \qquad \forall m \in M$$

$$y^{1,1}_{mrd} \in \{0,1\} \qquad \forall m \in M, \forall r \in R_m \forall d \in D$$

$$x^{1,1}_{imrd} \in \{0,1\} \qquad \forall i \in I, \forall m \in M_i, \forall r \in R_m \forall d \in D$$

$$a^{1,2}_q \in \{0,1\} \qquad \forall q \in Q$$

$$y^{1,2}_{qd} \in \{0,1\} \qquad \forall q \in Q, \forall d \in D$$

$$x^{1,2}_{iqd} \in \{0,1\} \qquad \forall i \in I, \forall q \in Q, \forall d \in D \tag{13}$$

3.2.3 Second-stage constraints

Constraints (14)-(18) are related to the scheduling of scan appointments. Constraint (14) ensures variable $z^2_{is_2}$ is equal to 1 if patient i is scheduled within the planning horizon (i.e. $\exists m, r, d$: $x^{2,1}_{imrds_2} = 1$). Constraints (15)-(17) implement the cancellation and rescheduling of patients. A cancellation and request for rescheduling are indicated by respective binary input parameters $\overline{\delta}^{canc}_{is_2}$ and $\overline{\delta}^{resch}_{is_2}$, which are sampled from the underlying probability distributions δ^{canc}_i and δ^{resch}_i. As a patient cannot cancel and request a rescheduling together,

the probability $Pr(\delta_i^{canc}, \delta_i^{resch})$ is modelled following a joint Bernoulli distribution for which $Pr(\delta_i^{canc} = 1, \delta_i^{resch} = 1) = 0$. In case of a cancellation ($\overline{\delta}_{is_2}^{canc} = 1$), constraint (15) imposes the patient is not scheduled ($z_{is_2}^2 = 0$). In case of a rescheduling request ($\overline{\delta}_{is_2}^{resch} = 1$), constraint (16) ensures the new scan appointment (date and MU) is conform to the patient's preference set $\overline{\Gamma}_{is_2}$. If this is not possible, the patient cannot be scheduled within this time horizon as new preferences need to be queried ($z_{is_2}^2 = 0$). For patients who do not wish to reschedule or cancel, constraint (17) ensures their scan appointments set in the baseline schedule remain unchanged regarding date and MU. In other words, if both $\overline{\delta}_{is_2}^{resch}$ and $\overline{\delta}_{is_2}^{canc}$ are 0 for patient i, the variable $x_{imrds_2}^{2,1}$ should be equal to the variable $x_{imrd}^{1,1}$ for all MUs and days. The mammography machine at the MU of the appointment, however, can be changed as this does not impact the schedule of the patient. Constraint (18) connects the scan appointments with the MU resources.
Constraints (19)-(26) are related to the scheduling of scan examinations. Scan examinations at the CC can be rescheduled almost without restrictions. Constraints (19) and (20) are equivalent to constraints (8) and (9) of the first stage. In contrast to the budgeted number of specialists that is fixed, adjustments to the CC resource schedule are possible by removing or adding work duties for CC specialists, which is modelled via constraints (21)-(25). Constraints (21) and (22) are the equivalent of constraints (10) and (11) but include variables $y_{qds_2}^{2,+}$ and $y_{qds_2}^{2,-}$. Constraints (23)-(25) assure the correct addition and removal of specialist work duties. Constraint (26) connect the scan examination schedule to the CC resource schedule. The domain constraints (27) indicate all decision variables of this stage are binary. Note, there are no specific capacity restrictions limiting the number of scan appointments (examinations) planned per day and MU machine (CC specialist), so that the second-stage decision always returns a feasible solution when the first-stage decision is feasible. As a result, the scheduling of scan appointments and examinations is primarily directed via objective function Z^*, so that patients requesting to revise their planned appointments are rescheduled as best as possible within the planning horizon.

$$\sum_{m\in M_i}\sum_{r\in R_m}\sum_{d\in D} x_{imrds_2}^{2,1} = z_{is_2}^2 \qquad \forall i \in I, \forall s_2 \in S_2 \tag{14}$$

$$z_{is_2}^2 \leq 1 - \overline{\delta}_{is_2}^{canc} \qquad \forall i \in I, \forall s_2 \in S_2 \tag{15}$$

$$\overline{\delta}_{is_2}^{resch} \times z_{is_2}^2 \leq \sum_{(m,d)\in\overline{\Gamma}_{is_2}}\sum_{r\in R_m} x_{imrds_2}^{2,1} \qquad \forall i \in I, \forall s_2 \in S_2 \tag{16}$$

$$\sum_{r'\in R_m} x_{imr'ds_2}^{2,1} \geq (1-\overline{\delta}_{is_2}^{resch}) \times (1-\overline{\delta}_{is_2}^{canc}) \times \sum_{r\in R_m} x_{imrd}^{1,1} \qquad \forall i \in I, \forall m \in M_i, \forall d \in D, \forall s_2 \in S_2 \tag{17}$$

$$x_{imrds_2}^{2,1} \leq y_{mrd}^{1,1} \qquad \forall i \in I, \forall m \in M_i, \forall r \in R_m, \forall d \in D, \forall s_2 \in S_2 \tag{18}$$

$$\sum_{q\in Q}\sum_{d\in D} x_{iqds_2}^{2,2} = z_{is_2}^2 \qquad \forall i \in I, \forall s_2 \in S_2 \tag{19}$$

$$\sum_{m\in M_i}\sum_{r\in R_m}\sum_{d\in D}\left((d+l)\times x_{imrds_2}^{2,1}\right) \leq \sum_{q\in Q}\sum_{d\in D}(d \times x_{iqds_2}^{2,2}) \qquad \forall i \in I, \forall s_2 \in S_2 \tag{20}$$

$$y_{qd}^{1,2} + y_{qds_2}^{2,+} - y_{qds_2}^{2,-} \leq a_q^{1,2} \qquad \forall q \in Q, \forall d \in D, \forall s_2 \in S_2 \tag{21}$$

$$\sum_{q \in Q}(y^{1,2}_{qd} + y^{2,+}_{qds_2} - y^{2,-}_{qds_2}) \le g^{CC} \qquad \forall d \in D, \forall s_2 \in S_2 \qquad (22)$$

$$y^{2,+}_{qds_2} \le 1 - y^{1,2}_{qd} \qquad \forall q \in Q, \forall d \in D, \forall s_2 \in S_2 \qquad (23)$$

$$y^{2,-}_{qds_2} \le y^{1,2}_{qd} \qquad \forall q \in Q, \forall d \in D, \forall s_2 \in S_2 \qquad (24)$$

$$y^{2,+}_{qds_2} + y^{2,-}_{qds_2} \le 1 \qquad \forall q \in Q, \forall d \in D, \forall s_2 \in S_2 \qquad (25)$$

$$x^{2,2}_{iqds_2} \le y^{1,2}_{qd} + y^{2,+}_{qds_2} - y^{2,-}_{qds_2} \qquad \forall i \in I, \forall q \in Q, \forall d \in D, \forall s_2 \in S_2 \qquad (26)$$

$$z^2_{is_2} \in \{0,1\} \qquad \forall i \in I, \forall s_2 \in S_2$$

$$x^{2,1}_{imrds_2} \in \{0,1\} \qquad \forall i \in I, \forall m \in M_i, \forall r \in R_m, \forall d \in D, \forall s_2 \in S_2$$

$$y^{2,+}_{qds_2}, y^{2,-}_{qds_2} \in \{0,1\} \qquad \forall q \in Q, \forall d \in D, \forall s_2 \in S_2$$

$$x^{2,2}_{iqds_2} \in \{0,1\} \qquad \forall i \in I, \forall q \in Q, \forall d \in D, \forall s_2 \in S_2 \qquad (27)$$

3.2.4 Third-stage constraints

Constraints (28) - (41) are related to the scheduling of scan examinations. In this stage, it is specified which patients did not show up to their scan appointment, indicated via the binary parameter $\overline{\delta}^{show}_{is_2s_3}$, and should not be considered in the planning of scan examinations. Parameter $\overline{\delta}^{show}_{is_2s_3}$ is a sample from the underlying probability distribution δ^{show}_i. Of course, a patient who cancels her appointment cannot realise a no-show. Accordingly, probability $Pr(\delta^{canc}_i, \delta^{show}_i)$ is modelled following a joint Bernoulli distribution for which $Pr(\delta^{canc}_i = 1, \delta^{show}_i = 0) = 0$. Taking the patients' no-show behaviour into account, the scan examination schedule and CC resource schedule may be adapted, affecting only patient flow times and resource utilisation at the CC. Constraints (28) and (29) relate to the scheduling of scan examinations for patients that showed (i.e. $\overline{\delta}^{show}_{is_2s_3} = 1$), which are similar to constraints (8) and (9) and constraints (19) and (20). Constraints (30)-(34) model the adjustments to the CC resource schedule resulting from adding or removing work duties. These are similar to constraints (21)-(25), but now modelling variables $y^{3,+}_{qds_2s_3}$ and $y^{3,-}_{qds_2s_3}$.
Constraint (35) connects the scan examination and resource schedule and is similar to constraint (12) related to stage 1 but incorporates resource adjustment variables from stages 2 and 3. Constraints (36)-(39) calculate overtime and idle time for the MU and CC resources via variables $it^{3,1}_{mrds_2s_3}$, $ot^{3,1}_{mrds_2s_3}$, $it^{3,2}_{qds_2s_3}$ and $ot^{3,2}_{qds_2s_3}$, representing the negative or positive difference between session duration (k^{MU}_m and k^{CC}) and scheduled number of working hours. In this regard, we account for the service duration of a scan appointment (θ^1) or scan examination (θ^2), which are perceived as deterministic and known because these are standard processes. Constraints (36) and (37) calculate the amount of idle time and overtime for machine r of MU m on day d. Note that the MU resource schedule has been determined via the first-stage decisions so that, throughout stages 2 and 3, the MU idle time and overtime are solely influenced by the encountered uncertainty. Similar to the MU machines, constraints (38) and (39) determine idle time and overtime for budgeted CC specialists, taking into account all adjustments to the resource schedule. Constraints (40)-(41) state the domains of the decision variables. This third stage provides always a feasible solution if decisions of stages 1 and 2 are feasible, thanks to the flexibility in the CC resource schedule.

$$\sum_{q \in Q}\sum_{d \in D} x^{3,2}_{iqds_2s_3} = \overline{\delta}^{show}_{is_2s_3} \times z^2_{is_2} \qquad \forall i \in I, \forall s_2 \in S_2, \forall s_3 \in S_3 \qquad (28)$$

$$\overline{\delta}^{show}_{is_2s_3} \times \Big(\sum_{m \in M_i}\sum_{r \in R_m}\sum_{d \in D}\big((d+l) \times x^{2,1}_{imrds_2}\big)\Big) \le \sum_{q \in Q}\sum_{d \in D}(d \times x^{3,2}_{iqds_2s_3}) \qquad \forall i \in I, \forall s_2 \in S_2, \forall s_3 \in S_3 \qquad (29)$$

$$
\begin{aligned}
&y^{1,2}_{qd} + y^{2,+}_{qds_2} - y^{2,-}_{qds_2} + y^{3,+}_{qds_2s_3} - y^{3,-}_{qds_2s_3} \leq a^{1,2}_q && \forall q \in Q, \forall d \in D, \forall s_2 \in S_2, \forall s_3 \in S_3 \quad (30)\\
&\sum_{q\in Q}(y^{1,2}_{qd} + y^{2,+}_{qds_2} - y^{2,-}_{qds_2} + y^{3,+}_{qds_2s_3} - y^{3,-}_{qds_2s_3}) \leq g^{CC} && \forall d \in D, \forall s_2 \in S_2, \forall s_3 \in S_3 \quad (31)\\
&y^{3,+}_{qds_2s_3} \leq 1 - y^{1,2}_{qd} - y^{2,+}_{qds_2} + y^{2,-}_{qds_2} && \forall q \in Q, \forall d \in D, \forall s_2 \in S_2, \forall s_3 \in S_3 \quad (32)\\
&y^{3,-}_{qds_2s_3} \leq y^{1,2}_{qd} + y^{2,+}_{qds_2} - y^{2,-}_{qds_2} && \forall q \in Q, \forall d \in D, \forall s_2 \in S_2, \forall s_3 \in S_3 \quad (33)\\
&y^{3,+}_{qds_2s_3} + y^{3,-}_{qds_2s_3} \leq 1 && \forall q \in Q, \forall d \in D, \forall s_2 \in S_2, \forall s_3 \in S_3 \quad (34)\\
&x^{3,2}_{iqds_2s_3} \leq y^{1,2}_{qd} + y^{2,+}_{qds_2} - y^{2,-}_{qds_2} + y^{3,+}_{qds_2s_3} - y^{3,-}_{qds_2s_3} && \forall i \in I, \forall q \in Q, \forall d \in D, \forall s_2 \in S_2, \forall s_3 \in S_3 \quad (35)
\end{aligned}
$$

$$
it^{3,1}_{mrds_2s_3} \geq k^{MU}_m \times y^{1,1}_{mrd} - \sum_{i\in\{I|m\in M_i\}} \left(\bar{\delta}^{show}_{is_2s_3} \times \theta^1 \times x^{2,1}_{imrds_2}\right) \qquad \forall m \in M, \forall r \in R_m, \forall d \in D, \forall s_2 \in S_2, \forall s_3 \in S_3 \quad (36)
$$

$$
ot^{3,1}_{mrds_2s_3} \geq \sum_{i\in\{I|m\in M_i\}} \left(\bar{\delta}^{show}_{is_2s_3} \times \theta^1 \times x^{2,1}_{imrds_2}\right) - k^{MU}_m \times y^{1,1}_{mrd} \qquad \forall m \in M, \forall r \in R_m, \forall d \in D, \forall s_2 \in S_2, \forall s_3 \in S_3 \quad (37)
$$

$$
it^{3,2}_{qds_2s_3} \geq k^{CC} \times (y^{1,2}_{qd} + y^{2,+}_{qds_2} - y^{2,-}_{qds_2} + y^{3,+}_{qds_2s_3} - y^{3,-}_{qds_2s_3}) - \sum_{i\in I}\left(\theta^2 \times x^{3,2}_{iqds_2s_3}\right) \qquad \forall q \in Q, \forall d \in D, \forall s_2 \in S_2, \forall s_3 \in S_3 \quad (38)
$$

$$
ot^{3,2}_{qds_2s_3} \geq \sum_{i\in I}\left(\theta^2 \times x^{3,2}_{iqds_2s_3}\right) - k^{CC} \times (y^{1,2}_{qd} + y^{2,+}_{qds_2} - y^{2,-}_{qds_2} + y^{3,+}_{qds_2s_3} - y^{3,-}_{qds_2s_3}) \qquad \forall q \in Q, \forall d \in D, \forall s_2 \in S_2, \forall s_3 \in S_3 \quad (39)
$$

$$
\begin{aligned}
&it^{3,1}_{mrds_2s_3}, ot^{3,1}_{mrds_2s_3} \geq 0 && \forall m \in M, \forall r \in R_m, \forall d \in D, \forall s_2 \in S_2, \forall s_3 \in S_3\\
&it^{3,2}_{qds_2s_3}, ot^{3,2}_{qds_2s_3} \geq 0 && \forall q \in Q, \forall d \in D, \forall s_2 \in S_2, \forall s_3 \in S_3 \quad (40)\\
&y^{3,+}_{qds_2s_3}, y^{3,-}_{qds_2s_3} \in \{0,1\} && \forall q \in Q, \forall d \in D, \forall s_2 \in S_2, \forall s_3 \in S_3\\
&x^{3,2}_{iqds_2s_3} \in \{0,1\} && \forall i \in I, \forall q \in Q, \forall d \in D, \forall s_2 \in S_2, \forall s_3 \in S_3 \quad (41)
\end{aligned}
$$

4 Methodology

The multi-stage stochastic programming problem described in the previous section is characterised by large numbers of decision variables and stochastic variables, leading to an exponential number of possible scenarios. Because solving the problem under study using an exact stochastic approach is intractable, we approximate the stochastic problem by reducing the scenario set using Monte Carlo simulation and estimating the expected objective function by the sample average. However, even when relying on SAA, reasonably sized instances with hundred patientsblack, two MUs, a planning horizon of five days and 100 different scenarios per stage are characterised by large dimensions, counting nearly 11×10^6 (mostly binary) decision variables and more than 13×10^6 constraints, of which a significant number are equality constraints (cf. Online Appendix C.1). For that reason, we developed a heuristic solution method that employs NBD to derive an optimal solution for the LPR of the problem under study, and a diving algorithm and post-processing procedure to retrieve a high-quality integer solution. Deriving an integer solution based on the (proxy) LPR solution has been demonstrated to be a common and valuable optimisation strategy in literature (Sadykov et al., 2019). Time limits are imposed to truncate the search in different steps and accelerate the algorithm. Figure 4 denotes the flowchart of the proposed methodology.
When initialising the solution procedure, we apply Monte Carlo simulation to replace random variables in the stochastic model with a limited set of randomly sampled values from the underlying probability distributions to obtain a deterministic SAA problem, which corresponds

to the model presented in Section 3 with only a subset of the possible scenarios. In other words, the denumerable sets S_2 and S_3 are replaced by restricted sets $\overline{S}_2 = \{0, ..., |\overline{S}_2| - 1\}$ and $\overline{S}_3 = \{0, ..., |\overline{S}_3| - 1\}$. We assume each sample has the same probability of occurrence, i.e $\rho_{s_2} = \frac{1}{|\overline{S}_2|}$ and $\rho_{s_2 s_3} = \frac{1}{|\overline{S}_2| \times |\overline{S}_3|}$. The SAA model approximates the expected objective value of the stochastic problem by optimising the sample average estimate derived from these random scenarios introduced in the multi-stage decision model. The quality of the solution is dependent on the number of scenarios, which is indicated by the sample size ($|\overline{S}_2| \times |\overline{S}_3|$).

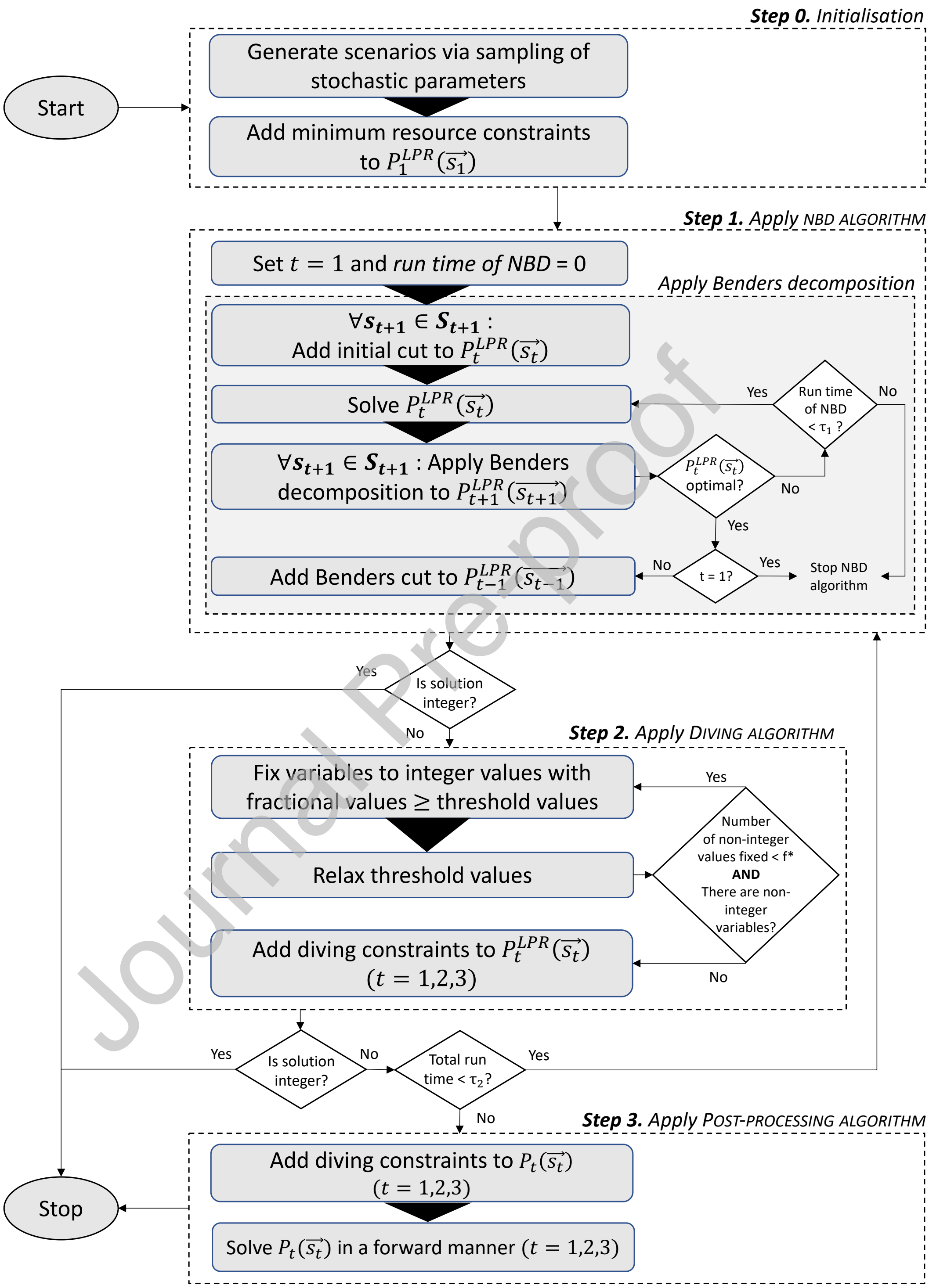


Figure 4: Overview of the solution methodology.

To solve the resulting complex deterministic model, we apply three steps. In the first step, we rely on the NBD method, also known as stochastic dual dynamic programming, which has been found suitable to solve the continuous LPR of multi-stage stochastic problems (Rebennack, 2016). The NBD method applies Benders decomposition to each pair of adjacent stages in a recursive manner. NBD views the scenario tree as a set of nested two-stage problems for which the integer domain constraints have been relaxed. We refer to the dependency between two nodes as a parent-child relation, i.e. each child node is dependent on one parent node. Correspondingly, every problem associated with an inner node in the tree is both the master problem to its children and a sub-problem of its parent. After solving the problems at a given stage, primal information, i.e. a proposed higher-stage solution, is pushed down toward the leaf nodes and dual information is passed up toward the root node, defining an optimality cut to improve the solution yielded by higher-stage problems. In this way, NBD disintegrates a large-scale problem into sub-problems that can be solved individually and then compounded together. Note that to accelerate the NBD algorithm, we derive a good initial solution at the start of each recursively applied Benders decomposition based on problem-specific knowledge by introducing (i) minimum resource constraints and (ii) initial optimality cuts. Section 4.1 discusses the applied NBD algorithm in more details. The NBD algorithm is stopped either when the gap between the calculated upper and lower bounds is closed or when the run-time limit τ^1 is reached. The best LPR solution found is used as input to the diving algorithm, which comprises the second step of the algorithm and is detailed in Section 4.2. The diving algorithm repetitively fixes particular variables to integer values and is alternated with the NBD algorithm to find an (optimal) LPR solution given those fixed integer values. For a particular LPR solution, the variable fixing is done based on specific threshold values for the different sets of variables, which are gradually relaxed during the progress of the diving algorithm, until a certain number of non-integer values (f^*) are fixed. The alternation between the NBD and diving algorithm is stopped when an integer solution is found or when the run-time limit τ^2 is reached. Whenever an integer solution is not yielded in time, a post-processing procedure is invoked in the third step to obtain an integer solution. This post-processing procedure solves the different stages with the diving constraints yielded in step 2 in a consecutive manner, pushing a proposed higher-stage solution forward toward the leaf nodes without recursion. Section 4.3 describes this procedure in more details.

4.1 The nested Benders decomposition

Decomposing the SAA model via Benders decomposition leads to one optimisation problem per node of the restricted scenario tree. Hence, the formulated SAA model is split into (i) a master problem (P_1) representing the first decision stage involving constraints (5)-(13), (ii) $|\overline{S}_2|$ second-stage sub-problems ($P_2(\overrightarrow{s_2})$) associated with second-stage nodes $\overrightarrow{s_2} = (0, s_2)$ involving constraints (14)-(27), and (iii) $|\overline{S}_2| \times |\overline{S}_3|$ third-stage sub-problems ($P_3(\overrightarrow{s_3})$) associated with third-stage nodes $\overrightarrow{s_3} = (0, s_2, s_3)$ involving constraints (28)-(41). The LPR of a specific node, denoted as $P_t^{LPR}(\overrightarrow{s_t})$, is obtained by relaxing the relevant binary domain constraints and replacing them by constraints bounding the equivalent relaxed variables to the interval $[0, 1]$. The objective of each problem $P_t(\overrightarrow{s_t})$ minimises the objective components associated with node $\overrightarrow{s_t}$ in model (1)-(41) and the average objective value relative to all the relevant child nodes. More precisely, the objective values of the master problem P_1 (Z) and the sub-problems $P_t(\overrightarrow{s_t})$ ($Z_t(\overrightarrow{s_t})$) are calculated as follows:

$$P_1: \quad Z = \quad Z_1^* + \frac{1}{|\overline{S}_2|} \sum_{s_2 \in \overline{S}_2} E[Z_2((0, s_2))]) \tag{42}$$

$$P_2((0,s_2)): \quad Z_2((0,s_2)) = \quad Z_2^*((0,s_2)) + \frac{1}{|\overline{S}_3|}\sum_{s_3\in\overline{S}_3} E[Z_3((0,s_2,s_3))] \qquad \forall s_2 \in \overline{S}_2 \tag{43}$$

$$P_3((0,s_2,s_3)): \quad Z_3((0,s_2,s_3)) = \quad Z_3^*((0,s_2,s_3)) \qquad \forall s_2 \in \overline{S}_2, \forall s_3 \in \overline{S}_3 \tag{44}$$

Given a solution of the parent node, a solution for the problem of a child node can be constructed. The primal objective value of a child node provides an Upper Bound (UB) for the expected objective value of the considered node, given the current solution of the parent node. On top of that, dual information relative to the child node can be used to calculate a Lower Bound (LB) for the expected objective value of the child node. This LB is added to the parent-node problem as a Benders optimality cut. Algorithm 1 describes the proposed NBD optimisation procedure. The algorithm iterates forward and backwards between nodes of the scenario tree until all nodes are optimal, i.e. until the optimal objective value Z^{LPR} of the relaxed SAA model is found and each child node's LB and UB are equal. In each iteration, one node $P_t(\overrightarrow{s_t})$ is visited, starting from the master problem (see Algorithm 1, line 1). If any changes were made to the underlying (sub-)problem, we re-optimise the relaxed (sub-)problem $P_t^{LPR}(\overrightarrow{s_t})$ using mathematical programming (see Algorithm 1, lines 3-7). These changes may include changes made to the model's Right-Hand-Side (RHS) values due to a different parent solution, or optimality cuts that were added or removed. Note, in case the master problem $P_1(\overrightarrow{s_1})$ is considered, we solve the integer problem instead of the relaxed problem, i.e. no dual information is retrieved nor optimality cuts are defined. If a new optimal value $Z_t^{LPR}(\overrightarrow{s_t})$ is found for the current node, the following three actions are taken (see Algorithm 1, lines 8-15):

- If (i) $t > 1$, (ii) the expected objective value of this problem in the parent node ($E[Z_t(\overrightarrow{s_t})]$) is smaller than the objective value $Z_t^{LPR}(\overrightarrow{s_t})$, and (iii) all of the child nodes of the current problem are optimal ($E[Z_{t+1}(\overrightarrow{s_{t+1}})] = Z_{t+1}^{LPR}(\overrightarrow{s_{t+1}})$), a Benders optimality cut is added to the problem of the parent node. An optimality cut in problem $P_{t-1}(\overrightarrow{s_{t-1}})$, indicating a LB for the expected objective value $E[Z_t(\overrightarrow{s_t})]$ of the current node, is dependent on (i) the decision variables of problem $P_{t-1}(\overrightarrow{s_{t-1}})$, (ii) the known values of the decision variables of all previous nodes in the scenario tree $P_1(\overrightarrow{s_1})$,...,$P_{t-2}(\overrightarrow{s_{t-2}})$, and (iii) the dual variables of the child node $P_t(\overrightarrow{s_t})$. Online Appendix C.2 discusses the formulation of the Benders optimality cuts in more details.
- The RHS values of all subordinate nodes in the scenario tree that involve resulting values of decision variables related to the current node are updated.
- Optimality cuts in all subordinate nodes in the scenario tree are removed. As the RHS values of subordinate nodes need to be revised, the associated dual information of the cuts previously added to any of the subordinate nodes is no longer relevant, rendering these cuts invalid for a new solution of the considered node.

Note, no feasibility cuts are needed as the master and sub-problems are formulated such that they are always feasible despite being dependent on previous decision stages (cf. Section 3). If the current node is the master problem ($t = 1$), its LB and UB are updated (see Algorithm 1, lines 16-17). The current LB is equal to the master problem's objective value, i.e. Z^{LPR}. The current UB can be calculated using equation (45) as the sum of the first-stage objective components (Z_1^*) and the average objective value of the second-stage sub-problems, given the considered first-stage solution.

$$UB = Z_1^* + \frac{1}{|\overline{S}_2|} \times \sum_{s_2\in\overline{S}_2} Z_2(\overrightarrow{s_2}) \tag{45}$$

The selection of the next node goes as follows (see Algorithm 1, lines 18-24). If not all child nodes are optimal, we select the first child node of the current node and continue the procedure such that all subordinate nodes can be re-optimised if needed. If all of the subordinate nodes of the current problem are optimal, i.e $E[Z_{t+1}(\overrightarrow{s_{t+1}})] = Z_{t+1}^{LPR}(\overrightarrow{s_{t+1}}))$, we move to the next node of the current stage. If all nodes related to the current stage are optimal, we select the parent node. Note, if the current node is the master problem, the optimal value Z^{LPR} is found and the algorithm will terminate because LB is equal to UB. The NBD procedure is accelerated via the construction of a good initial solution, which is discussed in Section 4.1.1.

Algorithm 1 The NBD ALGORITHM

Set $t = 1$, $n = 0$ and $s_t = 0$ such that $\overrightarrow{s_1} = (0)$ ▷ *Initialisation*
While $LB < UB$
  If any changes made to $P_t^{LPR}(\overrightarrow{s_t})$ ▷ *(Sub-)problem re-optimisation*
    If $t > 1$
      Re-optimise $P_t^{LPR}(\overrightarrow{s_t})$
    Else
      Re-optimise $P_1(\overrightarrow{s_1})$
  If new optimal value $Z_t^{LPR}(\overrightarrow{s_t})$
    If $t > 1$
    and $E[Z_t(\overrightarrow{s_t})] < Z_t^{LPR}(\overrightarrow{s_t})$
    and $E[Z_{t+1}(\overrightarrow{s_{t+1}})] = Z_{t+1}^{LPR}(\overrightarrow{s_{t+1}}) \quad \forall \overrightarrow{s_{t+1}} = (s_1, ..., s_t, s_{t+1})$ with $s_{t+1} \in \overline{S}_{t+1}$
      Add optimality cut to $P_{t-1}^{LPR}(\overrightarrow{s_{t-1}})$ with $\overrightarrow{s_{t-1}} = (s_1, ..., s_{t-1})$
    For $\forall \overrightarrow{s_{t+1}} = (s_1, ..., s_t, s_{t+1})$ with $s_{t+1} \in \overline{S}_{t+1}$
      Update the RHS values
      Remove any cuts in $P_{t+1}^{LPR}(\overrightarrow{s_{t+1}})$ and set $Z_{t+1}^{LPR}(\overrightarrow{s_{t+1}}) = +\infty$
  If $t = 1$
    Set $LB = Z_t^{LPR}(\overrightarrow{s_t})$ and re-calculate UB
  If $E[Z_{t+1}(\overrightarrow{s_{t+1}})] = Z_{t+1}^{LPR}(\overrightarrow{s_{t+1}}) \quad \forall \overrightarrow{s_{t+1}} = (s_1, ..., s_t, s_{t+1})$ with $s_{t+1} \in \overline{S}_{t+1}$ ▷ *Next node selection*
    If $s_t + 1 < |\overline{S}_t|$
      Set $s_t = s_t + 1$ in $\overrightarrow{s_t}$
    Else if $t > 1$
      Set $t = t - 1$
  Else
    Set $\overrightarrow{s_{t+1}} = (s_1, ..., s_t, 0)$ and $t = t + 1$

4.1.1 Construction of an initial solution

In order to solve the LPR faster, each recursively applied Benders decomposition is started from a good initial solution, which is accomplished by calculating accurate LBs on the expected objective values and the minimum resources needed. In this way, better optimality cuts can be included when initialising (sub-)problems. More precisely, constraints (46),(47) and (48) are added to the master problem. Constraint (46) indicates a LB for the expected objective values of the second-stage problems by assuming that (i) all patients are planned but have a flow time of zero, and (ii) all specialist duties planned in the first stage are removed in the second stage. Constraints (47) and (48) enforce a minimum number of working duties for the MUs (e^{MU}) and specialists (e^{CC}) in the master problem to avoid an extreme amount of overtime in the third stage. e^{MU} and e^{CC} are based on average service duration (θ^1 and θ^2), (minimum) session duration (k_m^{MU} and k^{CC}), and average number of patients who show up to their appointment ($|I| \times (1 - \hat{\delta}^{canc}) \times \hat{\delta}^{show}$, with $\hat{\delta}^{canc}$ and $\hat{\delta}^{show}$ the average probability of cancellation and a patient show, respectively). In each second-stage problem, constraint (49) is added to indicate the LBs for the expected objective values of the third-stage problems. It assumes that (i) all patients have a flow time of zero, and (ii) all specialist duties planned in the second stage are removed in the third stage. Note, the values of decision variables from stage t are parameters in subordinate stages $t' \in \{t + 1, ..., |T|\}$. This is indicated by overlining the symbols of the decision variables. Constraint (49) is dependent on the first-stage decisions related to variable $y^{1,2}_{qd}$, i.e on parameter $\overline{y}^{1,2}_{qd}$.

$$E[Z(\overrightarrow{s_2})] \geq -w^I \times w^{AP} \times |I| + w^R \times (-c_4 + c_5) \times \sum_{q \in Q} \sum_{d \in D} y^{1,2}_{qd} \qquad \forall s_2 \in \overline{S}_2 \tag{46}$$

$$\sum_{m \in M} \sum_{r \in R_m} \sum_{d \in \{0,\ldots,|D|-1-l\}} y^{1,1}_{mrd} \geq \min\Big(|M| \times |R_m| \times (|D| - l), e^{MU}\Big) \tag{47}$$

$$\sum_{q \in Q} \sum_{d \in \{l,\ldots,|D|-1\}} y^{1,2}_{qd} \geq \min\Big(|Q| \times (|D| - l), e^{CC}\Big) \tag{48}$$

$$E[Z(\overrightarrow{s_3})] \geq w^R \times (-c_4 + c_6) \times \sum_{q \in Q} \sum_{d \in D} (\overline{y}^{1,2}_{qd} + y^{2,+}_{qds_2} - y^{2,-}_{qds_2}) \qquad \forall s_3 \in \overline{S}_3 \tag{49}$$

$$\text{with} \quad e^{MU} = \Big\lceil |I| \times (1 - \hat{\delta}^{canc}) \times \hat{\delta}^{show} \times \theta^1 / \min_m(k^{MU}_m) \Big\rceil$$

$$e^{CC} = \Big\lceil |I| \times (1 - \hat{\delta}^{canc}) \times \hat{\delta}^{show} \times \theta^2 / k^{CC} \Big\rceil$$

4.2 Diving algorithm

The diving algorithm is invoked each time an (optimal) LPR solution is obtained via the NBD algorithm. In each call of the diving algorithm, multiple variables are selected and fixed to binary values via the addition of so-called diving constraints. Every binary variable can be fixed at most once. Note, not all binary variables have to be fixed by the diving algorithm to yield an integer solution as some binary variables may turn out to be integer in the LPR solution thanks to the fixing of other variables. For example, second-stage variables $z^2_{is_2}$ will be integer if all second-stage patient scheduling variables are integer (cf. constraints (14)-(16), (19)). In the following, we denote requirements and conditions for fixing variables and the functioning of the diving algorithm.

4.2.1 Requirements for fixing variables

In order to yield a feasible integer solution, the following requirements have to apply before a specific variable can be fixed:

- A second-stage scan appointment, i.e. variable $x^{2,1}_{imrds_2}$, can only be fixed if (i) the patient's scan appointment is either scheduled on day d in MU m in the first stage, or the involved patient requested to reschedule the appointment (cf. constraint (17)), and (ii) the respective work duty for MU m, machine r and day d has already been fixed in the first stage (cf. constraint (18)). When a second-stage scan appointment of a patient is fixed, all other assignments of the scan appointment can be set to 0 (cf. constraint (14)). Additionally, if the patient does not request for a rescheduling or does not cancel the appointment, the first-stage scan appointment should be fixed as well (cf. constraint (17)).
- A second- or third-stage scan examination, i.e. variable $x^{2,2}_{iqds_2}$ or $x^{3,2}_{iqds_2s_3}$, can only be fixed if the second-stage scan appointment has already been fixed at least l days earlier (cf. constraint (20) and (29)). When a second- or third-stage scan examination of a patient is fixed, all other assignments of the considered scan examination can be set to 0 (cf. constraints (19) and (28)).
- When a specialist work duty added in the second or third stage, i.e. variable $y^{2,+}_{qds_2}$ or $y^{3,+}_{qds_2s_3}$, is fixed, the specialist should be budgeted in the first stage, i.e. $a^{1,2}_q = 1$ (cf.

constraints (21) and (30)). Moreover, the work duty should also not be selected in the preceding decision stage, i.e. when $y^{2,+}_{qds_2}$ ($y^{3,+}_{qds_2s_3}$) is fixed, $y^{1,2}_{qd}$ ($y^{1,2}_{qd} + y^{2,+}_{qds_2} - y^{2,-}_{qds_2}$) should be fixed to 0 (cf. constraints (23) and (32)).

4.2.2 Conditions for fixing variables

On top of these requirements, following additional conditions are introduced to increase probability of obtaining a high-quality integer schedule:

- A variable related to the scheduling of a scan appointment or examination in the second or third stage, i.e. $x^{2,1}_{imrds_2}$, $x^{2,2}_{iqds_2}$, and $x^{3,2}_{iqds_2s_3}$, can only be fixed if their fractional value is larger than predefined threshold value w. Note, variables are processed in a logical order and we fix the first variable that meets the threshold value to a value of 1. For example, if w is equal to 0.5 and $x^{2,1}_{imrds_2} = 0.8$ and $x^{2,1}_{imrd's_2} = 0.2$, only variable $x^{2,1}_{imrds_2}$ is allowed to be fixed.
- Variables related to the scan examination in the second or third stage, i.e. $x^{2,2}_{iqds_2}$ or $x^{3,2}_{iqds_2s_3}$, can only be fixed if the associated work duty has already been fixed.
- A variable related to the scheduling of a scan appointment or examination in the second or third stage, i.e. $x^{2,1}_{imrds_2}$, $x^{2,2}_{iqds_2}$, and $x^{3,2}_{iqds_2s_3}$, can only be assigned to a work duty for which the average used capacity in the last stage would stay below h times the session duration. Threshold parameter h indicates the maximum relative usage of a work duty in the last stage and therefore ensures overtime is limited. The used capacity of a work duty is averaged over all subordinated third-stage scenarios. For example, the average used capacity of machine r of MU m on day d, given scenario $(0, s_2)$, is equal to $\frac{1}{|\overline{S}_3|} \times \sum_{s_3 \in \overline{S}_3} \sum_{i \in I} (x^{2,1}_{imrds_2} \times \overline{\delta}^{show}_{is_2s_3} \times \theta^1)$. The expected usage is updated each time a variable is fixed.
- A variable $y^{1,2}_{qd}$ related to the work duty planned in the first stage can only be fixed if its average used capacity is at least b times the session duration. Additionally, variables related to the removal of a specialist work duty in the second (third) stage, i.e. $y^{2,-}_{qds_2}$ and $y^{3,-}_{qds_2s_3}$, can only be fixed to 0 when (i) the work duty is fixed in the preceding stage, and (ii) its average used capacity is at least b times the session duration. Threshold parameter b indicates the minimum relative used capacity of work duties in the last stage and therefore ensures idle time is limited.

4.2.3 The diving procedure

Algorithm 2 describes the different steps of the proposed diving algorithm, for which the following notation is used: number of non-integer variables f of the current LPR solution fixed so far via diving, maximum number of non-integer variables f^* to fix, threshold values w, h and b, current diving iteration $iter$, and maximum number of diving iterations $maxIter$. At the beginning of the algorithm, threshold parameters w, h and b are set equal to initial values w^*, h^* and b^*. In subsequent iterations, these thresholds are loosened until at least f^* non-integer variables are fixed to an integer value or no more non-integer values can be fixed. In each iteration, the binary variables are fixed in the following order if necessary requirements and conditions stated above are fulfilled, i.e. (i) fix resources at the MUs ($y^{1,1}_{mrd}$), (ii) fix scan appointments in second stage ($x^{2,1}_{imrds_2}$), (iii) fix CC specialist work duties in first stage ($y^{1,2}_{qd}$) and remove work duties in second and third stages ($y^{2,-}_{qds_2}$ and $y^{3,-}_{qds_2s_3}$), and (iv) fix variables relative to scan examinations in second and third stages ($x^{2,2}_{iqds_2}$ and $x^{3,2}_{iqds_2s_3}$). Algorithms 5, 6, 7 and 8 in Online Appendix C.3 describe the respective procedures for fixing these variables in more details. After fixing variables in this order, one of the threshold values w, h or b is relaxed and set to a lower value in a step-wise linear manner using a constant, whereafter the next diving iteration starts. If, after $(maxIter - 1)$ iterations one or more non-integer values have been fixed, the diving algorithm is stopped. If no variables have been fixed so far, the thresholds are completely relaxed and a last diving iteration is performed. In that case, parameters w and b are set to $\epsilon = 10^{-6}$, such that any strictly positive relevant variable is a candidate to be fixed. Parameter h is set equal to the maximum possible relative usage of a work duty, such that any patient can be added to a planned work duty. If, after this last iteration, still no variables have been fixed ($f = 0$), the specialist work duties added in the second and third stage, i.e. variables $y^{2,+}_{qds_2}$ and $y^{3,+}_{qds_2s_3}$, are candidates to be fixed, for which the procedure is described in Algorithm 9 in Online Appendix C.3. These variables are fixed only if no other variables can be fixed, as they block the respective specialist work duty to be selected in the problem associated with the parent node of the considered node (cf. constraints (23) and (32)).

Algorithm 2 The DIVING ALGORITHM

```
Set f = 0, w = w*, b = b* and h = h*
While f < f* and iter < maxIter
    Apply DIVING ALGORITHM - MU RESOURCES (Algorithm 5)
    Apply DIVING ALGORITHM - SCAN APPOINTMENTS (Algorithm 6)
    Apply DIVING ALGORITHM - CC RESOURCES - A (Algorithm 7)
    Apply DIVING ALGORITHM - SCAN EXAMINATIONS (Algorithm 8)
    If iter < maxIter − 1
        Relax one of the threshold values, i.e. w = w − γ1 , h = h + γ2 or b = b − γ3
        (given constant values γ1, γ2 and γ3)
    Else if f = 0
        Set w = ϵ
        Set b = ϵ
        Set h = |I| × max(θ^1 / max_m(k_m^MU), θ^2/k^CC)
    Else
        iter = maxIter
    iter = iter + 1
If f = 0
    Apply DIVING ALGORITHM - CC RESOURCES - B (Algorithm 9)
```

4.3 Post-processing

Algorithm 3 describes the post-processing steps to yield an integer solution without recursive steps. The algorithm solves all (sub-)problems using mathematical programming but with integer domain conditions and prevailing variable-fixing constraints imposed, derived via the diving algorithm. As the master problem was already solved with integer domain conditions, the post-processing procedure first re-optimises the master problem if any relevant changes have been made (see Algorithm 3, lines 1-5). Next, all sub-problems are optimised, given the integer solution of their parent node (see Algorithm 3, lines 6-10).

Algorithm 3 The POST-PROCESSING ALGORITHM

If any changes made to $P_1(\overrightarrow{s_1})$ ▷ *Re-optimise master problem*
  Optimise $P_1(\overrightarrow{s_1})$
  For $\forall \overrightarrow{s_{t+1}} = (s_1, ..., s_t, s_{t+1})$ with $s_{t+1} \in \overline{S}_{t+1}$
    Update the RHS values
    Remove any cuts in $P_{t+1}(\overrightarrow{s_{t+1}})$
For $\forall P_t(\overrightarrow{s_t})$ with $t \in T \backslash \{1\}$, $\overrightarrow{s_t} = (s_1, ..., s_t)$ and $s_t \in \overline{S}_t$ ▷ *Optimise sub-problems*
  Optimise $P_t(\overrightarrow{s_t})$
  For $\forall \overrightarrow{s_{t+1}} = (s_1, ..., s_t, s_{t+1})$ with $s_{t+1} \in \overline{S}_{t+1}$
    Update the RHS values
    Remove any cuts in $P_{t+1}(\overrightarrow{s_{t+1}})$

5 Computational experiments

In this section, the computational experiments are described to validate the proposed methodology. In Section 5.1, we discuss the test design and input parameters characterising the test data set and methodology. The considered data instances are based on real-life problem settings regarding the BCSP in East Flanders, Belgium. Section 5.2 benchmarks the proposed solution procedure to other methodologies and discusses the value of integrating different stages. In Section 5.3, we investigate the benefits of stochastic optimisation and the options for patients to cancel or reschedule their appointment as mitigation strategies. Section 5.4 discusses the managerial insights. Experiments are carried out using a 1.60 GHz Dual-Core Intel Core i5 processor. The exact optimisation of a (sub-)problem is obtained via mathematical programming using the commercial package Gurobi version 9.0.

5.1 Test design

In this section, we discuss the characteristics of the test instances used in the experiments. This input data is deduced from discussions with stakeholders and aggregated data provided by the CC, involving the years 2018 and 2019. In addition, we define performance measures to benchmark the proposed methodology and set the relevant values for parameters characterising the solution methodology.

5.1.1 Input problem characteristics

In 2019, the CC coordinated the screenings at 41 hospitals and private practices and the scan examinations by 10 specialists (Bevolkingsonderzoek.be, 2019c). The average number of women invited each year, in 2018 and 2019, was 96 944 (Statbel.fgov.be, 2018, 2019). Patients are assigned to an eligible MU based on their residence, taking into account travel distance and public transport. Though, no strict rule exists for this assignment. As the hospitals are geographically dispersed throughout East Flanders (see Online Appendix D.1) and the planning is organised on a monthly basis, smaller instances are generated by decomposing the large-scale BCSP problem with respect to both the planning horizon and the geographical location of patients. This way, different instances are created in a structured and controlled way. In the following, the input instance data related to the resources, care process, patients and objective function are discussed. Online Appendix D.2 summarises these settings and parameter values.

Resource data - Real-life data indicates that 44% of the available MUs are smaller and have only one mammography machine, while the other 56% are larger and dispose of two machines on average (Bevolkingsonderzoek.be, 2019c). MU machines are made available for, on average, three hours per day for the BCSP. Though, larger MUs tend to have more personnel and are able to treat more patients per session than smaller MUs. Therefore, we assume the sessions of the larger MUs to be longer to mimic this difference in efficiency. Instances consider one smaller MU and one larger MU, composing the set M in a particular region. Smaller (larger) MUs dispose of one (two) mammography machines (R_m), accessible for three (four) hours per day (k_m^{MU}). As MUs are three or five days a week available to the BCSP, we consider instances with a planning horizon of three and five days (D). In addition, we also consider smaller instances with a planning horizon of two days to ensure adequate benchmarking with other solution methodologies. Further, at the CC, at most 2 specialists can work at the same time (g^{CC}). A specialist session endures approximately 2.5 hours per day (k^{CC}).

Care process data - A scan appointment has a duration of 30 minutes (θ^1), including preparation time (Bevolkingsonderzoek.be, 2019b). As a CC specialist can examine approximately 150 scans per session, a scan examination only takes 1 minute (θ^2). These durations are assumed to be deterministic. The minimum lag between appointment and examination (l) is 1 day.

Patient data - Real-life data suggests that, on average, one smaller and one larger MU can schedule 40, 61 and 101 patients for a planning horizon of, respectively, 2, 3 and 5 days. Accordingly, we consider 50 or 100 patients (I) depending on the planning horizon. The uncertainty related to the patient cancellations (δ_i^{canc}), rescheduling (δ_i^{resch}) and no-shows (δ_i^{show}) are modelled using a binomial distribution with respective probabilities of 3%, 30% and 33% based upon real-life input data (Bevolkingsonderzoek.be, 2018, 2019a). Other patient characteristics involve sets M_i and Γ_i, which are modelled as follows:

- **Set $\mathbf{M_i}$**: For each patient i, the set M_i is generated based on the city or town they are residents. The number of MUs patients can access depends on the population density of the considered region. Three density settings were tested: high, medium and low. Online Appendix D.1 provides a map of East Flanders (Belgium), the available MUs and examples of these density settings (Bevolkingsonderzoek.be, 2019c). The instances with 'high density' have no geographical limitations, i.e. all patients are inhabitants of a large city and both MUs are located in the vicinity so that all sets M_i include both MUs. The instances with a 'medium density' represent a situation where patients and MUs can be located in a neighbouring city. Based on the real-life data, 42% of the patients can access both MUs whereas 58% of the patients can be assigned to only one geographically close MU. The last setting, 'low density', involves patients and MUs who are all widely spread in the region. For these instances, only 6% of the patients can access both MUs. All other patients can access only a single MU (Statbel.fgov.be, 2018, 2019).
- **Set $\mathbf{\Gamma_i}$**: When a patient requests rescheduling, she denotes her preference set (Γ_i) to the CC, which is composed of the allowable day-MU combinations for rescheduling. These sets include all combinations of days and MUs, with the exception of the combinations associated with either one day or one MU. According to the CC, day preferences occur more often than MU preferences. Therefore, 75% of the preference sets exclude one day of the planning horizon. The other 25% of preference sets exclude an MU.

Objective function data - Based upon the real-life data and estimations from field experts, the variable cost per mammography machine shift for an MU is 400 (c_3) and for specialists 900 (c_4). There are no explicit fixed costs related to capacity decisions, i.e. employing a MU or specialist in the CC. However, to use these resources efficiently, we assign a cost equal to 50 to the use of both types of resources (c_1 and c_2). In real life, no explicit costs are associated with adjusting the scheduled shift of a CC specialist. However, as this determines employee satisfaction significantly, we assign a cost of 100 and 200 per change in the second and third decision stages (c_5 and c_6), respectively. Idle time and overtime negatively impact employee satisfaction as well, such that respective costs are defined as equal to 500 (c_7 and c_9) and 2500 (c_8 and c_{10}) per hour. Regarding the patient objectives, we set weight w^{FT} equal to 10 and weight w^{AP} equal to 100. The CC puts a higher priority on patient objectives compared to resource objectives. Therefore, all resource objectives are multiplied by a weight of 0.1 (w^R) and all patient objectives by 0.9 (w^I). Note the stability of the yielded solution quality and algorithm performance have been evaluated and validated via additional computational experiments, for which the different objective function coefficients have been varied (see Online Appendix D.3).

In summary, 9 instances have been generated, based on real-life data. The instances with 50 patients have a planning horizon of 2 or 3 days, while instances with 100 patients have a planning horizon of 5 days. In addition, each of these instances is tested for a high, medium and low-density area.

5.1.2 Performance measures

To evaluate the proposed methodology, eight performance measures are defined. The two main performance measures are run time (in seconds) and objective value (Z^*) calculated according to eq. (1). Other performance measures involve average number of patients treated, i.e. number of planned patients that show up to their appointment (N), average flow time per patient (FT, in days), average idle time and overtime per MU duty (IT^{MU} and OT^{MU}, in minutes) and average idle time and overtime per CC duty (IT^{CC} and OT^{CC}, in minutes), resulting from optimising objective components relative to different stages.

5.1.3 Input solution methodology parameters

Based upon preliminary computational experiments, the values are set for parameters characterising the methodology. SAA requires only two input parameters, $|\overline{S}_2|$ and $|\overline{S}_3|$. To determine the number of i.i.d. scenarios $|\overline{S}_2| \times |\overline{S}_3|$, we account for the trade-off between computation time and accuracy of the estimated objective values (Hu et al., 2020). A larger number of scenarios leads to a higher accuracy but also induces a higher computation time. Parameters $|\overline{S}_2|$ and $|\overline{S}_3|$ are set equal to $|\overline{S}|$. Based on an analysis of the stability of the objective value and associated confidence intervals (see Online Appendix D.4), we consider two values for $|\overline{S}| \in \{50, 100\}$, leading to 2500 or 10000 scenarios per SAA problem. For these numbers of scenarios, the average margin of error for the objective components and tested methods (cf. Section 5.2) amounts to 1.06%. Moreover, average results are observed to be stable when the number of scenarios is increased from 50 to 100. Note in the presentation and analysis of results below, we only mention results for individual instances with $|\overline{S}| = 100$ as conclusions are similar for instances with $|\overline{S}| = 50$. In addition, calculated averages are displayed for (i) all test instances with $|\overline{S}| = 50$, (ii) all test instances with $|\overline{S}| = 100$ and (iii) all test instances, i.e. with $|\overline{S}| \in \{50, 100\}$. The individual results for $|\overline{S}| = 50$ are provided in Online Appendix D.5.
The diving algorithm is stopped when $f^* = |I| \times |D| \times |\overline{S}|^2/500$ non-integer variables are fixed. In addition, parameters w^* and h^* are set equal to 0.9 and 1, respectively. Parameter b^* is set equal to 0.9 and 0.7 for MU and CC duties, respectively. Parameter $maxIter$ is set equal to 4×10, such that each threshold variable (w, h, b^{MU} and b^{CC}) is modified at most 10

times, plus one last iteration to fix variables under loose threshold values. All parameters are adapted according to a linear function. Parameters w and b decrease in steps equal to their original value (w^* and b^*) divided by 10. Parameter h is increased in steps equal to $h^*/20$. Furthermore, the time limit τ^1 for calculating the LPR using NBD is equal to 2 hours. The time limit τ^2 for the diving heuristic intertwined with NBD is equal to 3.75 hours. On top of that, time limits per (sub-)problem are set for the post-processing procedure, i.e. a time limit of (i) 25 minutes for solving the master problem, (ii) 5 minutes for solving second-stage sub-problems, and (iii) 5 minutes for solving third-stage sub-problems. The total run-time limit for all steps is set equal to 4 hours.

5.2 Benchmark comparison and impact of integration

To benchmark the proposed methodology, the algorithm is compared to (i) solution methodologies that optimise the SAA model exactly, and (ii) decomposed solution methodologies that optimise different single or multiple decision stages separately in a sequential manner. The following five methodologies are compared.

- **Exact $\mathbf{1} \uplus \mathbf{2} \uplus \mathbf{3}$**: This methodology optimises the integrated formulation presented in Section 3 relying on SAA using mathematical programming.
- **Heuristic $\mathbf{1} \uplus \mathbf{2} \uplus \mathbf{3}$**: This is the proposed methodology that optimises the integrated SAA formulation using the proposed NBD and diving heuristic.
- **Exact $\mathbf{1} \uplus \mathbf{2} \rightarrow \mathbf{3}$**: This method solves only the first two stochastic decision stages in an integrated manner. The SAA formulation is decomposed into (i) the two first stages and (ii) the third stage, which are solved as follows. First, the integrated SAA model relative to the first two stages is optimised via mathematical programming. Note that models of these two stages (cf. Section 3) do not take idle time and overtime, capacity utilisation restrictions nor patient flow time explicitly into account. Thanks to the fixed resource costs, idle time will be minimised without needing extra constraints. Overtime is minimised via addition of soft constraints that impose an optimal amount of overbooking for scan appointments and examinations. The maximum amount of overbooking in the first stage is set to 1.53 times the session duration, based on the average number of cancellations and no-shows. Likewise, the maximum amount of overbooking in the second stage is 1.49 times the session duration. The objective is modified accounting for a large penalty of 10000 per minute the maximum amount is exceeded. The patient flow time is minimised by adding it explicitly to the first- and second-stage objective function for all planned patients. Next, the third stage is optimised separately given inputs from the first two decision stages, i.e. for each third-stage scenario, a sub-problem is defined and solved via mathematical programming.
- **Heuristic $\mathbf{1} \uplus \mathbf{2} \rightarrow \mathbf{3}$**: This solution methodology applies decomposition similarly to previous method but solves the first two decision stages in an integrated manner relying on the proposed methodology, which is slightly modified. The NBD only iterates between the master problem and the second-stage sub-problems, while the diving algorithm only fixes variables relative to these problems. After the first two stages are optimised, the third-stage problems are solved separately via mathematical programming.
- **Exact $\mathbf{1} \rightarrow \mathbf{2} \rightarrow \mathbf{3}$**: In this methodology all three decision stages are optimised sequentially. Each (sub-)problem, obtained after decomposing the DMILP into one problem per scenario, is optimised individually via mathematical programming. Note that soft constraints on the amount of overbooking and optimisation of the patient flow time similar to previous methodologies have been added to first- and second-stage models.

Results of the benchmark are provided in Table 1, which displays run time (in seconds), objective value (Z^*) and objective value gap relative to the best objective value found for the respective instance ($\%gap$) between brackets. The best obtained objective value is indicated in bold. The first column indicates type of instances, denoted as $|D|$-population density. Population density is shown as low (l), medium (m) or high (h). The table reveals that methodology *Exact* $1 \uplus 2 \uplus 3$, operating on the integrated problem, cannot solve instances to optimality within 4 hours for any instance. Performance of this method significantly deteriorates when problem size increases. Moreover, the largest instances cannot be solved with this methodology due to the amount of memory needed and, consequently, no integer solution (n.i.s.) has been found for these instances. These findings confirm the appropriate use of heuristics for solving the problem. Required run times increase significantly when (i) instances are larger, (ii) more stages are integrated and (iii) the problem is solved using a MILP solver. The impact on the objective value is highly dependent on instance size. The following conclusions can be made:

- For small instances, i.e. $|D| = 2$, we observe that integration leads to a better average objective value. Integrating all stages ($1 \uplus 2 \uplus 3$) leads to significantly better objective values compared to (partially) sequential solution methodologies ($1 \uplus 2 \rightarrow 3$ or $1 \rightarrow 2 \rightarrow 3$), independent whether mathematical programming or the diving heuristic is applied. For small instances, the proposed heuristic performs only slightly worse (0.84% on average with $|\overline{S}| \in \{50, 100\}$) compared to the use of mathematical programming, which records far larger run times. This confirms that the proposed diving heuristic produces high-quality integer solutions.
- For medium instances, i.e. $|D| = 3$, the proposed heuristic delivers significantly better results than the exact method when all stages are integrated ($1 \uplus 2 \uplus 3$). However, other solution methods relying on partial integration ($1 \uplus 2 \rightarrow 3$) demonstrate a very similar performance as the proposed heuristic ($1 \uplus 2 \uplus 3$). Comparison learns that the method *Exact* $1 \uplus 2 \rightarrow 3$ marginally outperforms the other methods, though needs more computation time.
- For large instances, i.e. $|D| = 5$, (partially) sequential solution methodologies ($1 \uplus 2 \rightarrow 3$ or $1 \rightarrow 2 \rightarrow 3$) yield better results. The best performance is recorded by the method *Heuristic* $1 \uplus 2 \rightarrow 3$ in acceptable run times. The method *Exact* $1 \rightarrow 2 \rightarrow 3$ yields solutions of comparable quality, whereas the fully integrated proposed method *Heuristic* $1 \uplus 2 \uplus 3$ performs significantly worse.

In general, *Heuristic* $1 \uplus 2 \rightarrow 3$ leads to better objective values (Z^*) and lower deviations from the best objective values found ($\%gap$). Moreover, when comparing results for different instance sizes, we observe that this method leads to the most stable performance. Online Appendix D.6 validates the beneficial performance and role of the introduced accelerators for this method.
Figure 5 displays the objective component values averaged over all instances and the average rank for each methodology. The latter is calculated by averaging the rank of each method over the objective components, with the lowest rank number assigned to the method that scores best. Component CC overtime is not reported because it is 0 for all instances. Additionally, methodology *Exact* $1 \uplus 2 \uplus 3$ is not included either, as not for every instance a feasible solution has been found. Note that attained CC idle times are relatively high, which results from the short processing times to examine a scan and limited number of patients considered in an instance. The following conclusions can be made. The average flow times are acceptable and are very similar for the different methods. The heuristic methods yield schedules in which a significantly larger number of patients can be treated and the more when integrating all stages (*Heuristic* $1 \uplus 2 \uplus 3$), because this method takes all types of uncertainty into account

relative to all stages in a unified manner (cf. Section 5.3). In addition, integrating all stages leads to a significantly lower CC idle time, because in (partially) sequential methods this objective is not considered in the first two stages. However, these (partially) sequential methods attain significantly better values for the realised MU overtime, resulting from the large penalty associated with exceeding the maximum amount of overbooking. The method *Heuristic* $1 \uplus 2 \rightarrow 3$ leads to a significantly lower MU idle time compared to other methods. Note, due to the complexity of the problem, a completely integrated method ($1 \uplus 2 \uplus 3$) and a fully sequential method ($1 \rightarrow 2 \rightarrow 3$) obtain the same average rank with respect to the different objective components. Overall, the partially integrated method ($1 \uplus 2 \rightarrow 3$) obtains the best average rank, which finds the right compromise between integration and empirical problem complexity.

	Param-eters	$1 \uplus 2 \uplus 3$ Exact Z^* (%gap)	Time	$1 \uplus 2 \uplus 3$ Heuristic Z^* (%gap)	Time	$1 \uplus 2 \rightarrow 3$ Exact Z^* (%gap)	Time	$1 \uplus 2 \rightarrow 3$ Heuristic Z^* (%gap)	Time	$1 \rightarrow 2 \rightarrow 3$ Exact Z^* (%gap)	Time
Instances ($\|\overline{S}\| = 100$)	2-l	**-2246** (0.00%)	14454	-2208 (1.66%)	2376	-2117 (5.74%)	100	-2179 (2.95%)	58	-2128 (5.25%)	55
	2-m	**-2398** (0.00%)	14449	-2384 (0.59%)	7641	-2172 (9.43%)	66	-2211 (7.79%)	174	-2149 (10.40%)	56
	2-h	-2396 (0.14%)	14451	**-2399** (0.00%)	8884	-2172 (9.46%)	3113	-2208 (7.99%)	1225	-2166 (9.75%)	57
	3-l	-1185 (60.06%)	14630	**-2968** (0.00%)	13566	-2883 (2.85%)	13052	-2856 (3.78%)	210	-2884 (2.85%)	92
	3-m	-1982 (38.10%)	14711	-3170 (1.01%)	10242	**-3203** (0.00%)	13589	-3174 (0.91%)	7307	-3110 (2.89%)	109
	3-h	-995 (69.22%)	14868	-3110 (3.79%)	11652	**-3233** (0.00%)	13590	-3112 (3.72%)	266	-3156 (2.39%)	93
	5-l	n.i.s.	-	-3909 (34.51%)	13617	-5434 (8.95%)	13844	**-5968** (0.00%)	1737	-5704 (4.43%)	353
	5-m	n.i.s.	-	-4524 (29.86%)	13685	-6001 (6.96%)	13881	-6381 (1.08%)	1910	**-6450** (0.00%)	781
	5-h	n.i.s.	-	-3045 (52.59%)	13678	-6058 (5.68%)	13916	-6313 (1.70%)	7457	**-6423** (0.00%)	1035
Average	$\|\overline{\mathbf{S}}\| = \mathbf{50}$	-1879* (38.22%)	12977*	**-3826** (2.55%)	9441	-3763 (5.19%)	7748	-3816 (4.08%)	946	-3798 (5.07%)	199
	$\|\overline{\mathbf{S}}\| = \mathbf{100}$	-1867* (27.92%)	14594*	-3080 (13.78%)	10594	-3697 (5.45%)	9461	**-3823** (3.32%)	2261	-3796 (4.22%)	292
	$\|\overline{\mathbf{S}}\| \in \{\mathbf{50}, \mathbf{100}\}$	-1874* (34.10%)	13624*	-3453 (8.16%)	10017	-3730 (5.32%)	8605	**-3820** (3.70%)	1603	-3797 (4.65%)	246

* Instances with (i) $|D| = 5$, $|I| = 100$, high density and $|\overline{S}| = 50$ and (ii) $|D| = 5$, $|I| = 100$, low, medium or high density and $|\overline{S}| = 100$ not included

Table 1: Benchmark between proposed method and other methodologies (Run time, Z^*, $\%gap$) for individual instances with $|\overline{S}| = 100$ and averaged.

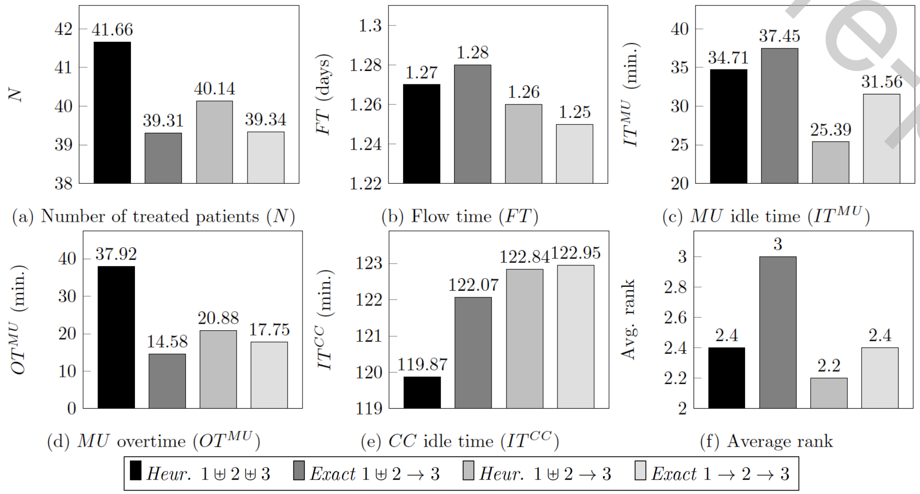


Figure 5: The objective component value and associated average rank per method.

$\|\overline{S}\|$	$\|D\|$	SO	WS	EVPI (%)	EEV	VSS (%)
50	2	-2362	-2570	208 (8.74%)	-2280	82 (3.59%)
50	3	-3196	-3475	280 (8.73%)	-2673	523 (19.72%)
50	5	-6298	-7045	747 (12.00%)	-4810	1488 (31.59%)
100	2	-2348	-2578	230 (9.76%)	-2281	67 (2.92%)
100	3	-3135	-3464	330 (10.56%)	-2675	459 (17.24%)
100	5	-6280	-7040	760 (12.15%)	-4817	1464 (30.95%)
Avg.		-3936	-4362	426 (10.81%)	-3256	681 (17.67%)

Table 2: The calculation of the EVPI and VSS.

5.3 Value of stochastic optimisation and no-show mitigation mechanisms

5.3.1 Value of information and the stochastic solution

To measure the potential benefits of stochastic optimisation, we evaluate the Expected Value of Perfect Information (EVPI) and the Value of Stochastic Solution (VSS), i.e.

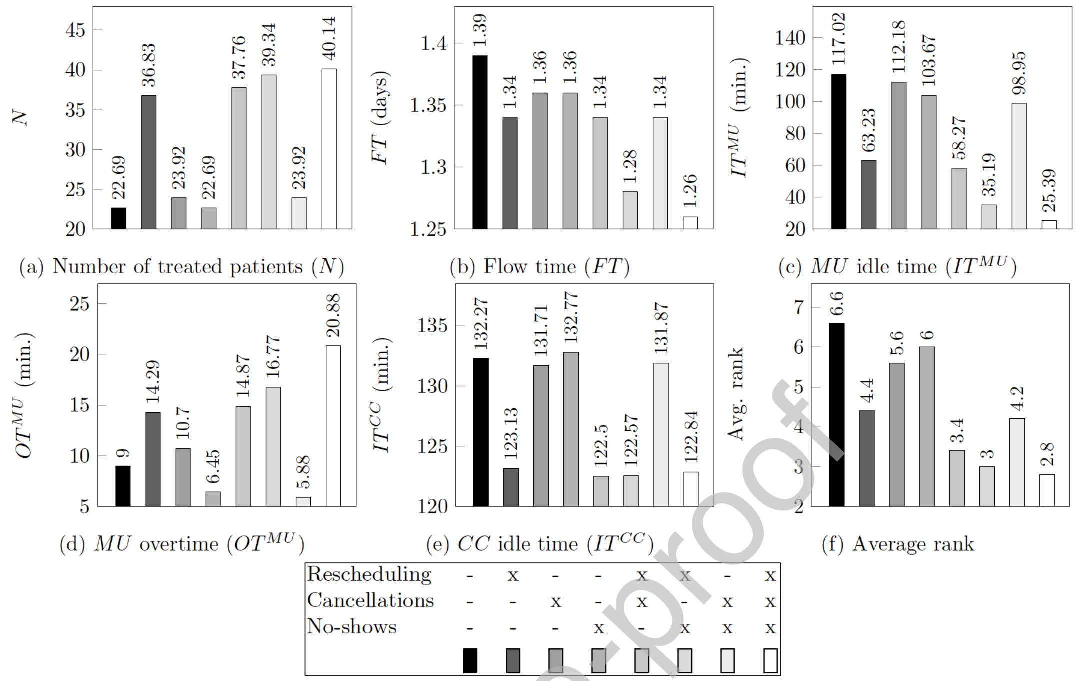
- The EVPI provides the upper bound of the improvement in the objective value if the realisations of the stochastic variables would be known at the time of composing the initial patient and resource schedule, comparing the so-called here-and-now and wait-and-see approaches. More precisely, for minimization problems, EVPI is calculated as the difference between the objective value using Stochastic Optimisation (SO) and the average of the so-called Wait-and-See problems (WS), i.e. EVPI = SO - WS. WS is calculated by optimizing the three-stage decision problem (1)-(41) per (realised) scenario and taking the average of the yielded objective values over all scenarios (Avriel and Williams, 1970; Escudero et al., 2007).
- The VSS indicates the profit from using SO rather than using the expected average values. VSS is calculated as the difference between the Expected result of using the Expected Value (EEV) and the objective value using SO, i.e. VSS = EEV - SO. The EEV of the multi-stage stochastic problem under study is calculated using the approach of Escudero et al. (2007).

Table 2 displays the yielded values SO, WS and EEV, and calculated metrics EVPI and VSS in an absolute and relative manner. The value for SO corresponds to the best solution found over all different versions of the (heuristic) algorithm compared in Section 5.2 because we want to calculate the bounds EVPI and VSS as accurately as possible. The value for WS is computed by solving $|\overline{S}_2| \times |\overline{S}_3|$ scenario problems, which is computationally demanding, with a time limit of 10 seconds per scenario. In this way, the run times for computing the value of WS amount to maximum 28 hours per instance. To calculate the value for EEV, a time limit of 230 minutes is imposed for solving the three-stage average scenario problem to determine the first-stage decision variables. No time limit was imposed on solving the second-stage or third-stage problems. In this way, the run times for computing the value of EEV are observed to amount to maximum 4 hours per instance.
The results denote that the EVPI averaged over all instances amounts to 426 and is similar for all instance sizes, i.e. having perfect information on the patient cancellations, rescheduling and no-show behaviour would lead to a significant improvement in the objective value of 10.81%. This endorses the need to gather more accurate information related to the patients' preferences and no-show behaviour, e.g. based on patient characteristics to better specify individual uncertainties (e.g. da Silva et al., 2021; Yan et al., 2022). Moreover, results relative to VSS show that using stochastic optimisation rather than the average scenario to determine first- and second-stage decisions leads to a significant improvement of 17.67% in the objective value. Obviously, the VSS is dependent on the instance size. More precisely, for smaller instances the model using the expected average values provides a good approximation of the stochastic problem as the VSS is small. The VSS is higher for larger instances having a longer planning horizon and a larger number of patients, better representing real-life complexity.

5.3.2 Impact of proactive mitigation strategies

To evaluate the benefits of the no-show mitigation strategies, we rely on the best-performing method identified in Section 5.2, i.e. *Heuristic* $1 \uplus 2 \rightarrow 3$, and assess the outcome of the yielded schedules after realisation of the uncertainties. As the problem under study considers three sources of uncertainty, 2^3 experiments are conducted for which different types of uncertainty are included and/or excluded. In case the option to cancel (or reschedule) is not considered, we assume a pessimistic but realistic scenario in which patients who would normally cancel (or reschedule), do not show up to the appointment. During the optimisation, we consider these patients as being regular ones, who may or may not show up to their appointment as a result of no-show uncertainty. In case we do not take into account no-show uncertainty, overbooking of resource duties to mitigate the impact of patient no-shows is left out in the design of the appointment and resource schedules. Figure 6 displays the average results related to all instances, for both $|\overline{S}| = 50$ and $|\overline{S}| = 100$. The depicted graphs provide the average objective component values and the associated average rank (cf. Section 5.2). The legend table indicates the included sources of uncertainty per experiment.
The results reveal that considering all sources of uncertainty yields the best schedules in terms of solution quality after the uncertainties have been realised. This is confirmed by the average rank of the different solution methods and all objective components except for the MU overtime. Most importantly, this scenario records the largest number of patients treated and the lowest patient flow time. In contrast, neglecting all sources of uncertainty leads to a schedule with inferior solution quality, showing the lowest number of patients treated, the largest flow time and high resource idle times. Comparing the results of these two experiments reveals the value of the mitigation strategies together.
In the following, we discuss the effect of implementing specific mitigation strategies to reduce (the impact of) patient no-show behaviour, i.e. the options to reschedule or cancel appointments and patient overbooking, and associated sources of uncertainty. Comparison is made to the case neglecting all sources of uncertainty. The results show that if the single option to reschedule an appointment is offered, the number of treated patients increases significantly by 62% (from 22.69 to 36.83) together with a decrease in the resource idle times by 46% at the MUs (from 117.02 to 63.23) and 7% at the CC (from 132.27 to 123.13), whereas the MU overtime increases by 59% (from 9 to 14.29). Installing the single option to cancel appointments only slightly improves all objective components, except for OT^{MU}, but observed improvements are considerably lower than in case of rescheduling. Taking into account no-show uncertainty via overbooking leads to small improvements for each objective component, similar to the ones observed for the option to cancel appointments. Only for the MU overtime, we observe a significant reduction by 28% (from 9 to 6.45). The superior performance of when considering the rescheduling of patients can be explained by the large percentage of patients (30%) requesting a rescheduling that is granted, whereas in the other scenarios rescheduling is not possible and these patients do not show up.
Excluding one type of uncertainty provides insight in the complementarity and added value of a particular mitigation strategy. Results demonstrate that excluding the option to reschedule significantly lowers the number of patients treated by 40% (from 40.14 to 23.92) and increases resource idle times by 290% at the MUs (from 25.39 to 98.95) and 7% at the CC (from 122.84 to 131.87). Not allowing the patients to cancel their appointment or omitting patient overbooking leads to smaller decreases in performance, indicating a smaller added value for these options. However, overbooking to mitigate cancellations and/or no-shows does install additional flexibility to schedule patients who request a rescheduling. As a consequence, considering all mitigation strategies together leads to an increased number of patients treated and lower patient flow times and resource idle times. On the other hand, the amount

of overbooking does have to be monitored closely as it significantly increases MU overtime.



(a) Number of treated patients (N) (b) Flow time (FT) (c) MU idle time (IT^{MU})

(d) MU overtime (OT^{MU}) (e) CC idle time (IT^{CC}) (f) Average rank

Figure 6: The impact of mitigation strategies on the objective components and average rank.

5.4 Managerial findings

This paper provides managers with guidance on how to manage the BSCP to achieve better system performance. In this paper, we have demonstrated the benefits of the proposed approach with respect to the organisation of the process, the patient demand, and the resources. First, the BCSP is more efficiently and effectively organised when managers approach the different decision phases in an integrated manner and proactively mitigate the impact of uncertainty compared to tackling the different decision phases in a sequential manner, composing schedules based on expected values and/or responding to the realised uncertainty in a reactive manner. Neglecting the uncertainty in the patient behaviour significantly reduces the number of patients treated, increases the patient flow time and leads to inefficient utilisation of involved resources. Managers should preserve the possibilities for patients to cancel their appointment and especially formulate rescheduling requests in advance to mitigate the impact of no-shows as these options improve performance, i.e. the number of treated patients increases whereas the idle times of the resources decrease. Moreover, managers should be eager to gather more accurate information on no-show behaviour based on individual patient characteristics to further improve the process organisation. Second, substantial patient overbooking is suggested when composing the initial patient schedule, but the extent should be monitored closely to limit overtime. This overbooking is preferably slightly lower than the expected number of no-shows as a result of the higher costs associated with overtime compared to idle time and the complicated structure of the multi-stage healthcare scheduling problem under study. The overbooking rate is independent of the day in the scheduling horizon and the MU considered. Note that, apart from being a mitigation strategy in response to no-shows, overbooking allows for some flexibility to reschedule patients. Third, managers should link the resource schedules to the patient treatment schedules to improve the resource organisation, realising a more efficient deployment and effective utilisation of resources. In this way, the budgeted number of resources, incurred idle time and overtime of MU resources are kept to a minimum and patients' flow times are acceptable. The latter is important to reduce patient anxiety toward the clinical outcome. For that purpose, scan appointments are planned in a (more or less equally) distributed manner over the horizon as a function of the number of MU sessions scheduled. The duties of CC specialists, which are fewer in number because of the lower capacity unit requirements, are attuned to these and are to some extent balanced in a likewise manner over the time horizon.

6 Conclusions

The contribution of this study is fourfold. First, we studied the coordination of BCSP in Flanders, integrating patient scheduling and resource planning to minimise both patient flow time and resource costs. The provided patient service embodies a two-stage healthcare process, involving both patient scan appointment and subsequent examination by skilled personnel. Second, the problem is modelled as a three-stage stochastic optimisation model as patients are given the option to reschedule or cancel their appointment after the first decision stage to mitigate impact of large patient no-show rates. Additional changes to the CC resource schedule can be done after scan appointments are realised and patients' (no-)shows are known. Third, we propose a diving heuristic operating on the linear relaxation of the problem under study, calculated viblacka sample average approximation and nested Benders decomposition. The diving heuristic fixes the variables in a specific order when problem-specific requirements and conditions are met. Results reveal that integrating only the first two stages and applying the proposed heuristic yields the best and most stable performance. Fourth, we demonstrate the value of stochastic optimisation to tackle the problem under study and the complementarity of different mitigation strategies, i.e. the options to reschedule or cancel appointments and patient overbooking, to reduce the impact of no-shows on system performance.
In future research, this study can be extended in different directions relative to both problem definition and solution methodology. First, the proposed model assumes that patients are homogeneous relative to their resource usage and their cancellation, re-scheduling and no-show probabilities, leaving room for improvement by incorporating patient-specific information based on history. This information could be taken into account during decision-making and the design of efficient and effective overbooking strategies to schedule an appropriate casemix of patients to single MU sessions. Second, the construction of the patient and resource schedules and associated adaptations embody a dynamic process and can be studied in future research using Markov decision processes rather than via static decision phases, because of the interrelationship between subsequent periods. Unfulfilled rescheduling requests may impact the set of considered patients for the next horizon. Adaptations to the schedule are driven by the arrival rate of the cancellations and rescheduling requests during the patient response horizon and the patient no-shows during the processing of the planning horizon. However, because the application of Markov decision processes for scheduling individual patients may severely increase the size and complexity of the model, it may be beneficial to study the decision phases separately as suggested by our experiments. Third, the proposed approach is based on stochastic optimisation, optimising the expected flow time and resource costs, but does not ensure that these objectives may worsen substantially in specific scenarios. A suitable solution approach to prevent the potential deterioration of the realised schedule quality is, for example, adjustable robust optimisation in concordance with the multi-stage

character of the problem. Robust optimisation optimises worst-case performance to protect the system against worst-case scenario realisations.

References

Aissaoui, N. O., Khlif, H. H., and Zeghal, F. M. (2020). Integrated proactive surgery scheduling in private healthcare facilities. *Computers & Industrial Engineering*, 148:106686.

Avriel, M. and Williams, A. (1970). The value of information and stochastic programming. *Operations Research*, 18(5):947–954.

Barro, D., Canestrelli, E., et al. (2006). Time and nodal decomposition with implicit non-anticipativity constraints in dynamic portfolio optimization. *Math Methods Econ Finance*, 1:1–20.

Bevolkingsonderzoek.be (2018). Annual report BCSP Flanders for 2018. `https://borstkanker.bevolkingsonderzoek.be/bk/literatuur` Last accessed: 2023-01-10.

Bevolkingsonderzoek.be (2019a). Annual report BCSP Flanders for 2019. `https://borstkanker.bevolkingsonderzoek.be/bk/literatuur` Last accessed: 2023-01-10.

Bevolkingsonderzoek.be (2019b). General information on the BCSP in flanders. `https://borstkanker.bevolkingsonderzoek.be/en` Last accessed: 2023-01-10.

Bevolkingsonderzoek.be (2019c). Mammography units in east flanders available to the BCSP. `https://borstkanker.bevolkingsonderzoek.be/en/bk/engwhere-can-i-participate` Last accessed: 2023-01-10.

Birge, J. R. and Louveaux, F. (2011). *Introduction to stochastic programming*. Springer Science & Business Media.

Broeders, M., Moss, S., Nyström, L., Njor, S., Jonsson, H., Paap, E., Massat, N., Duffy, S., Lynge, E., and Paci, E. (2012). The impact of mammographic screening on breast cancer mortality in europe: a review of observational studies. *Journal of Medical Screening*, 19(1_suppl):14–25.

Cayirli, T. and Veral, E. (2003). Outpatient scheduling in health care: a review of literature. *Production and Operations Management*, 12(4):519–549.

Chen, Z., King, W., Pearcey, R., Kerba, M., and Mackillop, W. J. (2008). The relationship between waiting time for radiotherapy and clinical outcomes: a systematic review of the literature. *Radiotherapy and Oncology*, 87(1):3–16.

da Silva, R. B. Z., Fogliatto, F. S., Krindges, A., and dos Santos Cecconello, M. (2021). Dynamic capacity allocation in a radiology service considering different types of patients, individual no-show probabilities, and overbooking. *BMC Health Services Research*, 21(1):1–24.

Diamant, A., Milner, J., and Quereshy, F. (2018). Dynamic patient scheduling for multi-appointment health care programs. *Production and Operations Management*, 27(1):58–79.

Dogru, A. K. and Melouk, S. H. (2019). Adaptive appointment scheduling for patient-centered medical homes. *Omega*, 85:166–181.

Drexl, A. and Haase, K. (1995). Proportional lotsizing and scheduling. *International Journal of Production Economics*, 40(1):73–87.

Escudero, L. F., Garín, A., Merino, M., and Pérez, G. (2007). The value of the stochastic solution in multistage problems. *Top*, 15(1):48–64.

Gocgun, Y. and Puterman, M. L. (2014). Dynamic scheduling with due dates and time windows: an application to chemotherapy patient appointment booking. *Health Care Management Science*, 17(1):60–76.

Gupta, D. and Denton, B. (2008). Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions*, 40:800–819.

Hu, Y., Chen, X., and He, N. (2020). Sample complexity of sample average approximation for conditional stochastic optimization. *SIAM Journal on Optimization*, 30(3):2103–2133.

Jiang, B., Tang, J., and Yan, C. (2019). A stochastic programming model for outpatient appointment scheduling considering unpunctuality. *Omega*, 82:70–82.

Karimi, B., Ghomi, S. F., and Wilson, J. (2003). The capacitated lot sizing problem: a review of models and algorithms. *Omega*, 31(5):365–378.

Kim, K. and Mehrotra, S. (2015). A two-stage stochastic integer programming approach to integrated staffing and scheduling with application to nurse management. *Operations Research*, 63(6):1431–1451.

Marynissen, J. and Demeulemeester, E. (2019). Literature review on multi-appointment scheduling problems in hospitals. *European Journal of Operational Research*, 272(2):407–419.

Menon, K. G., Fukasawa, R., and Ricardez-Sandoval, L. A. (2021). A novel stochastic programming approach for scheduling of batch processes with decision dependent time of uncertainty realization. *Annals of Operations Research*, 305(1):163–190.

Ouelhadj, D. and Petrovic, S. (2009). A survey of dynamic scheduling in manufacturing systems. *Journal of Scheduling*, 12(4):417–431.

Pham, D.-N. and Klinkert, A. (2008). Surgical case scheduling as a generalized job shop scheduling problem. *European Journal of Operational Research*, 185(3):1011–1025.

Rebennack, S. (2016). Combining sampling-based and scenario-based nested benders decomposition methods: application to stochastic dual dynamic programming. *Mathematical Programming*, 156(1-2):343–389.

Sadykov, R., Vanderbeck, F., Pessoa, A., Tahiri, I., and Uchoa, E. (2019). Primal heuristics for branch and price: The assets of diving methods. *INFORMS Journal on Computing*, 31(2):251–267.

Statbel.fgov.be (2018). Belgian population in 2018. `https://statbel.fgov.be/open-data/population-place-residence-nationality-marital-status-age-and-sex-8` Last accessed: 2023-01-10.

Statbel.fgov.be (2019). Belgian population in 2019. `https://statbel.fgov.be/open-data/population-place-residence-nationality-marital-status-age-and-sex-9` Last accessed: 2023-01-10.

Vitali, S., Dominguez, I., and Moriggia, V. (2021). Comparing stage-scenario with nodal formulation for multistage stochastic problems. *4OR*, 19:613–631.

Wang, J. and Fung, R. Y. (2014). An integer programming formulation for outpatient scheduling with patient preference. *Industrial Engineering and Management Systems*, 13(2):193–202.

White, D. L., Froehle, C. M., and Klassen, K. J. (2011). The effect of integrated scheduling and capacity policies on clinical efficiency. *Production and Operations Management*, 20(3):442–455.

Wu, X. and Zhou, S. (2022). Sequencing and scheduling appointments on multiple servers with stochastic service durations and customer arrivals. *Omega*, 106:102523.

Yan, C., Huang, G. G., Kuo, Y.-H., and Tang, J. (2022). Dynamic appointment scheduling for outpatient clinics with multiple physicians and patient choice. *Journal of Management Science and Engineering*, 7(1):19–35.

Zhang, J., Dridi, M., and El Moudni, A. (2020). Column-generation-based heuristic approaches to stochastic surgery scheduling with downstream capacity constraints. *International Journal of Production Economics*, page 107764.

Zhou, S. and Yue, Q. (2019). Appointment scheduling for multi-stage sequential service systems with stochastic service durations. *Computers & Operations Research*, 112:104757.