OR Spectrum (2021) 43:59–87
https://doi.org/10.1007/s00291-020-00605-w

REGULAR ARTICLE

Preventive healthcare facility location planning with quality-conscious clients

Ralf Krohn[1] · **Sven Müller**[2] · **Knut Haase**[1]

Received: 13 July 2018 / Accepted: 18 August 2020 / Published online: 13 October 2020

Abstract

Pursuing the overarching goal of saving both lives and healthcare costs, we introduce an approach to increase the expected participation in a preventive healthcare program, e.g., breast cancer screening. In contrast to sick people who need urgent medical attention, the clients in preventive healthcare decide whether to go to a specific facility (if this maximizes their utility) or not to take part in the program. We consider clients' utility functions to include *decision variables* denoting the waiting time for an appointment and the quality of care. Both variables are defined as functions of a facility's utilization. We employ a segmentation approach to formulate a mixed-integer linear program. Applying GAMS/CPLEX, we optimally solved instances with up to 400 demand nodes and 15 candidate locations based on both artificial data as well as in the context of a case study based on empirical data within one hour. We found that using a Benders decomposition of our problem decreases computational effort by more than 50%. We observe a nonlinear relationship between participation and the number of established facilities. The sensitivity analysis of the utility weights provides evidence on the optimal participation given a specific application (data set, empirical findings).

Keywords Discrete choice · Healthcare · Random utility · Configurations of facilities · Facility location · Benders

✉ Ralf Krohn
ralf.krohn@uni-hamburg.de

Sven Müller
sven.mueller@ovgu.de

Knut Haase
knut.haase@uni-hamburg.de

[1] Faculty of Business Administration (HBS – Hamburg Business School), Institute of Transport Economics, Universität Hamburg, Moorweidenstraße 18, 20148 Hamburg, Germany

[2] Center of Operations Research and Business Analytics, Otto-von-Guericke-University, Universitätsplatz 2, 39106 Magdeburg, Germany

1 Introduction

Preventive healthcare is beneficial to society because it facilitates early detections of diseases and helps to interrupt its development, which is less costly and harmful than curing diseases at advanced stages (Güneş et al. 2015). Established preventive healthcare programs (e.g., breast or colorectal cancer screenings) help to increase the number of early detections (Madadi et al. 2015). To save both lives and potential medical treatment costs, policy makers aim at raising participation in preventive healthcare programs (BreastScreen Australia 2015). Whereas patients have little influence on where they get medical help in urgent cases, preventive measures can be planned and allow for patients' choices. Empirical studies show that accessibility is a significant factor that drives people's participation. The location of facilities is decisive: Our objective is to develop a methodology to set up a limited number of facilities so that patients' expected participation in a preventive screening program can be maximized. For realistic and evidence-based planning, it is crucial to not only rely on deterministic influence factors like proximity between the patient and healthcare facilities. From a microeconomic perspective, multiple drivers of human behavior have to be incorporated to create a thorough planning model based on strong theoretical foundations. Supply (with preventive healthcare services) is not only designed to fulfill demand, demand also depends on the provided supply. For instance, doctor's quality of care and the waiting time for an appointment are both functions of a preventive healthcare facility's utilization, i.e., demand. A policy maker has to pay attention to these endogenous factors to forecast participation rates correctly (Haase and Müller 2015).

The waiting time we consider is not that at the facility, but the time between a patient's call for an appointment and the date of that appointment. This denotes a major difference to what other authors addressed before in a large sequence of papers. We do not explicitly consider operational waiting times at the doctor's surgery.

Our concept for quality of care focuses on medical competence to correctly identify a patient's illness or healthiness. Since this is positively related to the treatment volume, we can label special screening centers with an above-average success rate.

We consider a strategic planning horizon and expect the predicted demand to unfold in the long term, if facilities are located according to our solution. Throughout the paper, we assume that new centers are labeled by the authorities, so clients are aware of the centers' quality. To the best of our knowledge, the literature only covers related problems and no formulation of our problem has been proposed. We incorporate

1. individual demand as dependent on all other individual demands and
2. quality of care, and waiting time for an appointment as decision variables within clients' utility functions
3. a discrete choice model (clients' choices) in a mixed-integer linear program.

The advantage of a linear reformulation is the more tractable solution behavior by standard optimization software. We provide new insights into the trade-offs between waiting time, travel-time, and quality of care, and how these trade-offs contribute to a given preventive healthcare network.

The problem of locating (preventive healthcare) facilities is closely related to other fields of service operations, especially product line design planning (see, e.g., Belloni et al. 2008) and assortment optimization (see, e.g., Rusmevichientong et al. 2014; Besbes and Sauré 2016). Without loss of generality, we motivate our contribution within a preventive healthcare facility setting because of the relevance and topicality. We show how manifold effects on client choice behavior can be included in a facility location planning framework while ensuring a tractable model formulation. The application to other sectors is straightforward, e.g., housing development (Espinoza Garcia and Alfandari 2018) or facility network design for schools (Haase and Müller 2013), public offices, garages, and others.

Our paper is structured as follows: first, we discuss the related literature in Sect. 2. In Sect. 3, we present the relevant theoretical background for our contribution. Afterwards, we formally develop and extensively explain different modeling frameworks (Sect. 4). The results of our computational studies as well as the underlying synthetic data set are depicted in Sect. 5. We apply our model to a real-world case study in Sect. 6 and conclude with a summary and an outlook in Sect. 7.

2 Literature review

In a substantial body of the facility location literature, it is assumed that clients' choices merely depend on the distances or travel time between the facilities and the locations of residence (Harper et al. 2005; Zhang et al. 2012; Tiwari and Heese 2009). Proximity and travel time have a substantial impact on people's choices (Train 2009, p. 72). However, empirical studies illustrate that patients are influenced by many more attributes (controllable by the decision maker/planner) and that stochastic choice models are preferable to deterministic ones (Ayer et al. 2016; Benati and Hansen 2002; Gerard et al. 2008; Hol et al. 2010; van Dam et al. 2010; Feldman et al. 2014). Detailed points of clients' interest may be the proximity to an outpatient service center, its connection to public transport, and information about the center's quality of treatment from word of mouth or official ratings. There is evidence that patients are sensitive to the practitioner's technical experience (quality of care) and waiting time for an appointment (Castillo et al. 2009; Moscelli et al. 2016; Elhedhli 2006; Chen et al. 2015). Osadchiy and Diwas (2017, p. 2) say that a patient only makes an appointment if the waiting time to see a doctor is acceptable.

The importance of waiting time for an appointment is prevalent in manifold clinical settings (Wang et al. 2018). In Killaspy et al. (2000), psychiatric patients reported reasons for not consulting a doctor (no-show). Among the revealed reasons, up to 48% were cases whose probabilities of incidence rises with increasing waiting time for an appointment. Furthermore, patients can feel disrespected when they are offered time slots in the distant future, and they have more time to develop worries and fears (Lacy et al. 2004). When patients have to wait long for an appointment with a specialist, e.g., orthopedics, patients' discomforts can be gone in the meantime so that the cause of their occurrence cannot be traced back anymore. In pediatric care or during pregnancy, waiting time can be critical, because certain medical examinations have to be carried out before the due date. Gerard et al. (2008) provide empirical evidence that waiting time for an appointment has an impact on clients' choices to participate in breast cancer screenings. For convenience reasons, we refer to a cancer screening setting in the following sections.

Facilities' quality of care and waiting time for an appointment can be viewed as attractiveness states, which impact demand (i.e., clients' evaluation of a facility). Moreover, both are a function of the number of patients that access service, i.e., demand (Güneş et al. 2004). Endogeneities and feedbacks are usually managed by queueing theory approaches. In Baron et al. (2008) the problem is to choose the number, location, and facilities' capacities. In contrast, neither do we need to assume certain distributions for arrival and service processes or demand nor that each client goes to the nearest facility. This is due to a different time horizon as well as the general and probabilistic nature of utility theory we apply. Wang and Wang (2017) endogenize network effects in an MNL-based choice model, where utility and choice probability, respectively, for product i at time t depend on the choice probability at time $t-1$. Under the assumption that the utility function is a second-order continuously differentiable function defined on [0, 1], they show that there is a steady state condition for the choice probabilities. In our contribution, we consider a non-continuously differentiable utility function where utility depends on the demand and thus also on the choice probability of an alternative. We use a binary variable—denoted as "dummy" in empirical research papers—for the quality of a healthcare center: If the expected demand reaches a certain threshold (= endogenously given minimum quantity requirement), quality is accredited by authorities (variable equals 1).

Zhang et al. (2009) and Zhang et al. (2010) provide a nonlinear location-allocation model with respect to congestion in the context of preventive healthcare network designs. They employ a so-called optimal choice model that assumes that clients go to the nearest facility (shortest travel time, which is a parameter). Verter and Lapierre (2002) present a preventive healthcare facility location problem with an emphasis on distance and implement a minimum quantity requirement via hard constraints to ensure quality of care. Population centers are deterministically allocated to exactly one established facility. Elhedhli (2006) provides a linearization for the facility location problem with immobile servers, stochastic demand, and congestion and also accounts for waiting time at the facility. Customers' demand is allocated to exactly one facility (single sourcing). Vidyarthi and Kuzgunkaya (2015) investigate the trade-off between waiting and travel time while simultaneously determining both the facilities' locations and their capacities as well as the allocation of clients.

3 Theoretical framework

To model the client behavior, we use random utility theory: Discrete choice models are a reliable tool to both analyze and predict individual choice behavior based on utility maximization. It has received great attention especially in the fields of marketing and transportation research for decades (Liu et al. 2018). An individual chooses exactly one alternative from a finite set of available, mutually exclusive, and collectively exhaustive alternatives (Train 2009, p. 11). In our case, the choice set includes all available facility locations and the alternative of not going to any facility, i.e., the "no-choice" alternative (not participating in the healthcare program).

An individual's utility for an alternative is the result of the alternative's attributes as well as the individual's characteristics. Following the utility maximization choice rule, an individual chooses the available alternative that dominates all other alternatives by having the highest utility (Koppelman and Bhat 2006, p. 18). That is, individual $i \in \mathscr{I}$ chooses alternative j from the choice set $\mathscr{J}$ if

$$U_{i,j} > U_{i,k} \quad \forall\, k \in \mathscr{J} \land k \neq j, \tag{1}$$

where $U_{i,j}$ is individual i's perceived utility value for alternative j. Since the choice-making process is often not entirely understood by the analyst, probabilistic choice models are used to account for unobserved characteristics and incomplete information. For that reason, the utility is decomposed into two components:

$$U_{i,j} = v_{i,j} + \epsilon_{i,j}, \tag{2}$$

where $v_{i,j}$ denotes the deterministic utility part for an individual i and alternative j, which is observed by the analyst. $\epsilon_{i,j}$ is a stochastic error term that equals the difference between the known deterministic utility and the utility $U_{i,j}$ used by the individual, which is generally unknown to the observer. Since $U_{i,j}$ is a random quantity, the analyst can only make probability statements about (1). The probability $p_{i,j}$ that individual i chooses alternative j is

$$p_{i,j} = \Pr\big(U_{i,j} > U_{i,k} \quad \forall\, k \in \mathscr{J} \land k \neq j\big). \tag{3}$$

The multinomial logit model (MNL) is based on the assumption that the error components $\epsilon_{i,j}$ are independent and identically type I extreme-value (also: Gumbel) distributed (i.i.d. EV) across alternatives as well as across individuals (McFadden 2001). Therefore, the MNL choice probability

$$p_{i,j}^{\mathrm{MNL}} = \frac{\mathrm{e}^{v_{i,j}}}{\sum_{k \in \mathscr{J}} \mathrm{e}^{v_{i,k}}} \tag{4}$$

can be derived from (3) (Train 2009, p. 74). A fundamental property of the MNL is the irrelevance of independent alternatives (IIA). When we consider the ratio of any two alternatives' choice probabilities (odds):

$$\frac{p_{i,j}^{\mathrm{MNL}}}{p_{i,h}^{\mathrm{MNL}}} = \frac{\mathrm{e}^{v_{i,j}} / \sum_{k \in \mathscr{J}} \mathrm{e}^{v_{i,k}}}{\mathrm{e}^{v_{i,h}} / \sum_{k \in \mathscr{J}} \mathrm{e}^{v_{i,k}}} = \frac{\mathrm{e}^{v_{i,j}}}{\mathrm{e}^{v_{i,h}}} = \mathrm{e}^{v_{i,j} - v_{i,h}} \tag{5}$$

we can see that the *ratio* of two alternatives' choice probabilities does not depend on any other alternative. The most important advantage of this property is the implication that alternatives (here: facility locations) can be added to or removed from the choice set without changing the structure of the MNL (Koppelman and Bhat 2006, p. 39). For our purpose, we only need the ratios of the choice probabilities for visiting that facility and the no-choice alternative (Haase and Müller 2015). The ratios serve as input parameters for our mathematical model. Its objective is to determine the clients' choice probabilities with respect to the optimal facility setting. While the probabilities can vary, the ratios between any of them are constants and do not depend on the setting. This observation allows for their correct calculation for any constellation of established facilities, i.e., solution.

In our approach, the deterministic utility is a linear function of an individual's travel time to a location, the facility's quality of care, and the waiting time for an appointment at that facility. Travel time to a facility location is a parameter while waiting time for an appointment and quality of care are dependent on other participants' choices, i.e., demand. Since demand depends on the locational decisions, demand is treated as an auxiliary variable in our approach. Therefore, waiting time and quality of care—as functions of demand—are auxiliary variables as well. Both travel and waiting time will generally have a negative impact on utility. A high quality level can be accredited by the authority in charge (e.g., a certain number of cases performed per period is expected for such an accreditation) and is assumed to be positively rated by the clients (BreastScreen Australia 2015). Even if we consider a "green field", without any already established preventive healthcare services, using discrete choice theory needs no scenario warm-up (see, e.g., marketing research for new products). The strategically adjusting equilibrium can be determined beforehand and justifies labeling special treatment centers in advance.

4 Model development

In the following, we define our location problem, illustrate the mathematical representation and develop a linear reformulation of the initially nonlinear mixed-integer program.

4.1 Problem definition

Given are the locations of client nodes (demand points), the number of eligible patients per node, candidate locations for preventive healthcare facilities and a set of feasible facility modes. Different modes represent a facility's attractiveness (utility) levels caused by varying properties (waiting times for an appointment and quality of care). The problem is to determine the locations and modes of established facilities in a way that maximizes the target population's expected participation in the preventive healthcare program. On one hand, both quality of care and waiting time for an appointment are dependent on demand for healthcare service measured as the number of arriving patients at service facilities. On the other hand, the demand side is also suspected to be dependent on supply as quality and waiting time are parts of the clients' utility function. We visualize these relationships in Fig. 1.

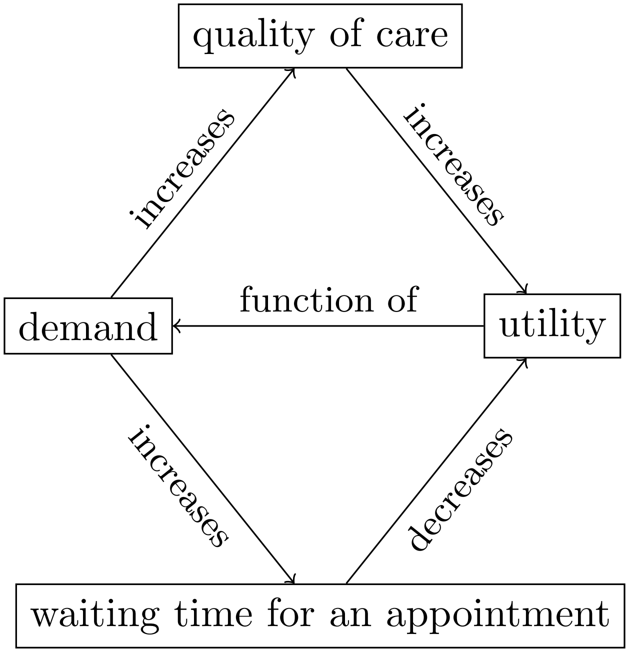


Fig. 1 Utility concept: a higher utility increases the expected demand for a facility. On one hand, higher demand leads to higher waiting times for appointments, which clients are confronted with, and decreases their utility for the facility. On the other hand, it also increases the quality of care in the form of higher practitioner's experience, which increases utility

Our model formulations are based on the following two assumptions:

1. High quality of care is achieved if a healthcare service facility exceeds a certain number of carried out screenings. A corresponding label "specialty treatment center" is then accredited by authorities. The label is lacking if that number is not reached (Zhang et al. 2010, pp. 867–868). The threshold value might come from medical guidelines or case-based research, which is justified by real existing minimum quantity requirements for some medical treatments (Blum and Offermanns 2004; BreastScreen Australia 2015) and operations as well as by the findings of Schrag et al. (2000); Güneş et al. (2004); Harewood (2005) and Blankart (2012). Quality of care denotes medical treatment quality (being qualified to detect diseases, i.e., the likelihood of a true positive / true negative diagnosis). Quality of care does not mean service quality (denoting the personnel's kindness) at a facility. However, this quality type is aggregated in the alternatives' alternative-specific constants (a model parameter).
2. Waiting times for an appointment positively correlate with the number of clients going to a facility (given its capacity). As more patients ask for appointments, the waiting time will increase (Güneş et al. 2004; Viiala et al. 2007; Anand et al. 2011). We do not explicitly consider waiting times at the doctor's surgery (i.e.,

> sitting in the waiting room) since we consider a strategic planning horizon. Differences between facilities regarding the appointment management can be depicted by alternative-specific constants in the utility function. However, we do not want to mix up strategic and operational planning levels. Furthermore, clients do not know the waiting time in the waiting room beforehand at the time they make a choice.

Both high quality of care and short waiting times for an appointment are desirable for patient satisfaction (Liu et al. 2018; Elhedhli 2006; Feldman et al. 2014; Truong 2015; Batt and Terwiesch 2015). Our model's objective is to decide where preventive healthcare service facilities have to be located, whether they provide high quality, and how long patients have to wait for an appointment to maximize participation.

Wang et al. (2002); Marianov et al. (2008) and Dan and Marcotte (2019), inter alia, developed mixed integer nonlinear programs that account for probabilistic clients' choices based on linear continuous waiting times. In contrast, we avoid the continuous nature of queuing theory since it is difficult for clients to evaluate a difference between, e.g., 2 and 2.1 weeks. Furthermore, assuming utility to be linear in waiting time for an appointment is inconsistent with empirical evidence (van der Pol et al. 2014). Thus, we discretize waiting time. We assume known facilities' client demand thresholds that determine a change in the expected waiting time for an appointment faced by a patient. The resulting categories could be separated into e.g., "two weeks," "four weeks," etc.

4.2 Nonlinear preventive healthcare facility location planning problem

An initial and probably most intuitive modeling approach leads to a nonlinear formulation:

Sets and indices

$\mathscr{I}$ Set of demand nodes, index i
$\mathscr{J}$ Set of candidate facility location nodes $\mathscr{J} \subseteq \mathscr{I}$, indices j and k

Parameters

β^{Q} Utility coefficient for quality of care attribute ($\beta^{\mathrm{Q}} > 0$)
β_l^{W} Utility coefficient for waiting time level l ($0 > \beta_1^{\mathrm{W}} > \beta_2^{\mathrm{W}} > \cdots > \beta_L^{\mathrm{W}}$)
$\underline{\mu}_l$ Demand threshold above which waiting time level l applies
$\bar{\mu}_l$ Demand threshold below which waiting time level l applies
B Sufficiently big number
$\bar{c}_{i,j}$ Constant part of deterministic utility function for demand node i and facility at location j with, for example, $\bar{c}_{i,j} = \beta_j^{\mathrm{ASC}} + \beta^{\mathrm{D}} \cdot d_{i,j}$ where β_j^{ASC} is the alternative-specific constant for alternative j, $\beta^{\mathrm{D}} < 0$ the coefficient for distance and $d_{i,j}$ the distance from node i to node j

g_i Number of clients in node i that are eligible to require health service
L Number of waiting time levels $l = 1, \dots, L$
$q^{\min}$ Minimum quantity requirement
r Total number of available facilities
$v_{i,\mathrm{no}}$ Deterministic utility for demand node i of not attending any facility ("no-choice" or "opt-out" alternative)

Variables

Q_j = 1 if facility at location j satisfies the minimum quantity requirement; 0, otherwise
$W_{j,l}$ = 1 if waiting time level l applies for facility at location j; 0, otherwise
$X_{i,j}$ Choice probability of clients in i going to facility at location j
Y_j = 1 if location j is specified to offer healthcare service; 0, otherwise
$v_{i,j}$ Deterministic utility of clients in i going to facility at location j
F^{NLP} Objective function value of nonlinear program (expected healthcare service participation)

$$\text{Maximize} \quad F^{\mathrm{NLP}} = \sum_{i \in \mathcal{I}} g_i \cdot \sum_{j \in \mathcal{J}} X_{i,j} \tag{6}$$

subject to

$$X_{i,j} = \frac{\mathrm{e}^{v_{i,j}} \cdot Y_j}{\mathrm{e}^{v_{i,\mathrm{no}}} + \sum_{k \in \mathcal{J}} \mathrm{e}^{v_{i,k}} \cdot Y_k} \quad \forall\, i \in \mathcal{I}; j \in \mathcal{J} \tag{7}$$

$$v_{i,j} = \bar{c}_{i,j} + \beta^{\mathrm{Q}} \cdot Q_j + \sum_{l=1}^{L} \beta_l^{\mathrm{W}} \cdot W_{j,l} \quad \forall\, i \in \mathcal{I}; j \in \mathcal{J} \tag{8}$$

$$\sum_{i \in \mathcal{I}} g_i \cdot X_{i,j} \geq q^{\min} \cdot Q_j \quad \forall\, j \in \mathcal{J} \tag{9}$$

$$\sum_{i \in \mathcal{I}} g_i \cdot X_{i,j} \leq \bar{\mu}_l + B \cdot (1 - W_{j,l}) \quad \forall\, j \in \mathcal{J}; l = 1, \dots, L \tag{10}$$

$$\sum_{i \in \mathcal{I}} g_i \cdot X_{i,j} \geq \underline{\mu}_l \cdot W_{j,l} \quad \forall\, j \in \mathcal{J}; l = 1, \dots, L \tag{11}$$

$$\sum_{l=1}^{L} W_{j,l} = Y_j \quad \forall j \in \mathcal{J} \tag{12}$$

$$\sum_{j \in \mathcal{J}} Y_j = r \tag{13}$$

$$X_{i,j} \geq 0 \quad \forall i \in \mathcal{I}; j \in \mathcal{J} \tag{14}$$

$$v_{i,j} \in \mathbb{R} \quad \forall i \in \mathcal{I}; j \in \mathcal{J} \tag{15}$$

$$W_{j,l} \in \{0;1\} \quad \forall j \in \mathcal{J}; l = 1, \dots, L \tag{16}$$

$$Q_j \in \{0;1\} \quad \forall j \in \mathcal{J} \tag{17}$$

$$Y_j \in \{0;1\} \quad \forall j \in \mathcal{J} \tag{18}$$

The objective function (6) maximizes the expected participation measured as the number of patients that are expected to access preventive healthcare service. (7) are the MNL choice probabilities for an individual to access service at a certain facility. By (8), we calculate the deterministic utilities $v_{i,j}$ with some constant $\bar{c}_{i,j}$ as well as a facility's weighted levels of quality of care and waiting time for an appointment.

(9) ensures that a facility's binary quality indicator variable is set to zero if the number of served patients (left-hand side) is lower than the minimum quantity requirement $q^{\min}$. Instead, if the requirement is exceeded, Q_j can be set to 1. This means that the authority can then label that facility as a high volume center. If the number of such labels is limited to s facilities, we will add

$$\sum_{j} Q_j \leq s \tag{19}$$

to the model. If a label "high volume center" is mandatory, if the minimum quantity requirement is reached then, we will add

$$\sum_{i \in \mathcal{I}} g_i \cdot X_{i,j} \leq q^{\min} + B \cdot Q_j \forall j \in \mathcal{J}. \tag{20}$$

In an analogous manner, a facility's waiting time level is determined by (10) and (11). If the expected demand is greater than a threshold $\bar{\mu}_l$, a sufficiently big value B has to be added on the right-hand side to make (10) valid forcing the waiting time level variable to become 0. B could also be indexed with j and defined by the maximum possible demand per facility. Out of the set of variables $W_{j,l}$ that can become 1 by those constraints, the one corresponding to the lowest feasible waiting time level, meaning highest utility, is used due to (11) (see visualization in Fig. 2). The utilization of the optimization objective (6) may be insufficient to determine the correct

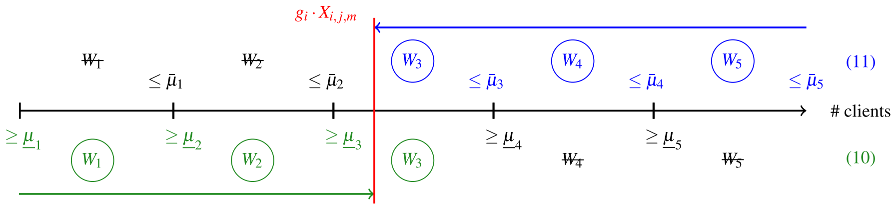


Fig. 2 Waiting time level determination: given the expected demand $g_i \cdot X_{i,j,m}$, each of both constraints (10) and (11) excludes certain levels (crossed out) and leaves the others valid. In combination, only one level remains valid (here: W_3). Schematic view for five waiting time categories. Ticks denote level thresholds

waiting time level just by their upper level thresholds in (10). Due to the feedbacks with the quality level determination in (9), the lower level threshold control via (11) is also required.

(12) ensures that an established facility has exactly one waiting time level. (13) ensures that r facilities are established. (14)–(18) define the variables' domains.

In our model, all facilities are suspected to have the same establishing costs. Although it is easy to extend the objective function by establishing costs f_j to

$$\text{Maximize} \quad F^{\text{NLP}} = \sum_{i \in \mathscr{I}} g_i \cdot \sum_{j \in \mathscr{J}} X_{i,j} - \sum_{j \in \mathscr{J}} f_j \cdot Y_j \tag{21}$$

to control the number of facilities as well, it is (ethically) difficult to compare a number of patients with costs, even with a scale parameter. A new constraint

$$\sum_{j \in \mathscr{J}} f_j \cdot Y_j \leq \lambda \tag{22}$$

with a budget λ additional to or to replace (13) might be an useful alternative. We could also include capacity levels $l = 1, \dots, L$ in (7), (12), (18), (21), (22) and $\sum_{l=1}^{L} Y_{j,l} \leq 1 \ \forall \, j \in \mathscr{J}$.

4.3 Mixed-integer linear preventive healthcare facility location planning problem (PHCFLPP)

It is desirable to transfer the nonlinear model (6)–(18) into a more tractable linear form. Haase and Müller (2014) provide a comparison of different linear reformulations for related problems. Notice also the developments of Ljubić and Moreno (2018). Our approach makes use of the MNL's IIA property (Haase 2009; Aros-Vera et al. 2013; Haase and Müller 2015).

We integrate quality and waiting time into a deterministic mixed-integer linear problem via discretization of the clients' utility function (Gerard et al. 2003; Street and Burgess 2007). We compress the representation of quality and waiting time levels in a more general "mode" structure (see Fig. 3). To reformulate (7) in a linear way, we consider each combination of a facility's location and its mode as a separate choice alternative, e.g., a single facility with two possible modes results in two alternatives within the

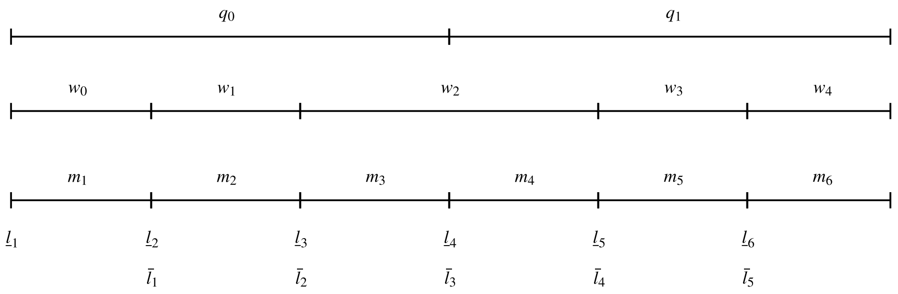


Fig. 3 Interrelation between quality categories q, waiting time categories w, and modes m with their corresponding lower and upper mode thresholds l (all measured in number of clients). m_1, or $m = 1$, means $q = 0$ and $w = 0$; $m = 2$ means $q = 0$ and $w = 1$; … ;$m = 6$ means $q = 1$ and $w = 4$. Schematic view for two quality and five waiting time categories. Ticks denote level thresholds

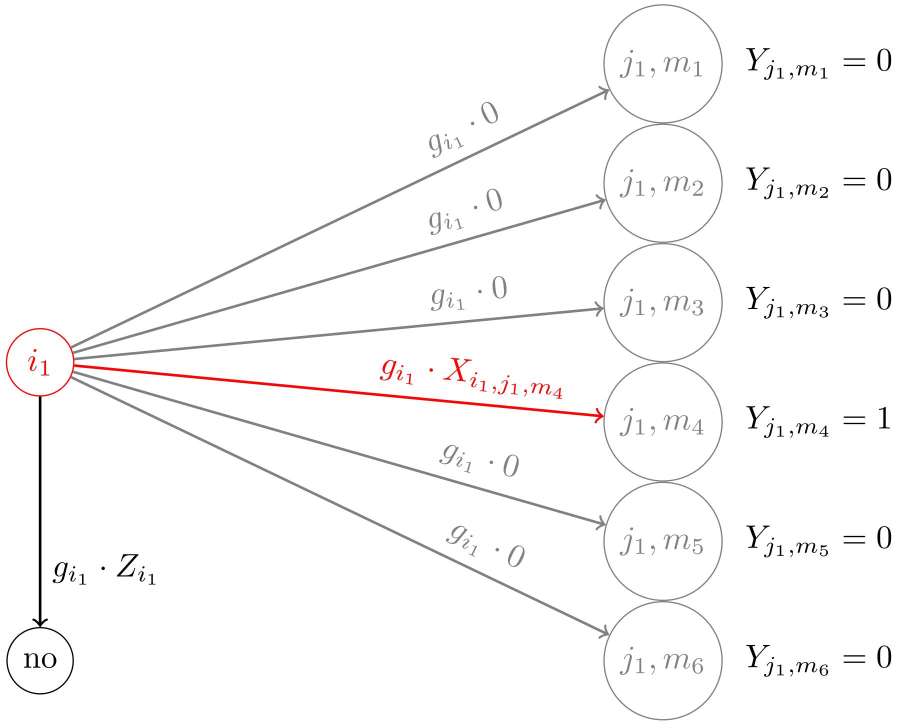


Fig. 4 Modes as virtual facilities: even if we only consider one demand node i_1 and one candidate facility location j_1, it has to be determined in which mode m the facility is established ($r = 1$). Let, in this example, the participation-maximizing solution be j_1 in mode m_4. The expected demand lies between the lower and upper thresholds of mode m_4, i.e. $\underline{l}_4 \leq g_{i_1} \cdot X_{i_1,j_1,m_4} \leq \bar{l}_4$ (see Fig. 3). All other modes do not exist (see $Y_{j,m}$ values), and the corresponding expected demand is 0. The no-choice alternative with its probability Z_{i_1} is always available

choice set. However, those two virtual facilities cannot be established simultaneously, because only exactly one mode is assigned to the facility. Hence, in the solution for this example only one alternative (the facility location in a specific mode) remains in addition to the no-choice alternative that is always present (see Fig. 4).

Now, consider the additional

Sets and indices

$\mathscr{M}$ Set of modes, in which a facility can be established (quality of care and waiting time for an appointment), indices m and n (might also contain capacity levels, for example)

Parameters

$p_{i,j,m}$ Choice probability of clients in i to access service at a facility located at j being in mode m given that (j, m) is the only facility established, i.e., the choice set consists of the two alternatives $\{(j,m);\text{no}\}$, which results in $p_{i,j,m} = \frac{e^{v_{i,j,m}}}{e^{v_{i,\text{no}}}+e^{v_{i,j,m}}}$ where $v_{i,j,m}$ is the deterministic utility of clients in i going to a facility located at j being in mode m

$\underline{l}_m$ Lower threshold for mode m measured in number of clients

$\overline{l}_m$ Upper threshold for mode m measured in number of clients

Variables

$X_{i,j,m}$ Choice probability of clients in i to access service at a facility located at j being in mode m

Z_i Choice probability of clients in i to refuse to access any facility ("no-choice")

$Y_{j,m}$ $= 1$ if location j is specified to offer healthcare service in mode m; 0, otherwise

F Objective function value of PHCFLPP (expected healthcare service participation)

$$\text{Maximize} \quad F = \sum_{i\in\mathscr{I}} g_i \cdot \sum_{j\in\mathscr{J}} \sum_{m\in\mathscr{M}} X_{i,j,m} \tag{23}$$

subject to

$$Z_i + \sum_{j\in\mathscr{J}} \sum_{m\in\mathscr{M}} X_{i,j,m} \leq 1 \quad \forall\, i \in \mathscr{I} \tag{24}$$

$$X_{i,j,m} \leq p_{i,j,m} \cdot Y_{j,m} \quad \forall\, i \in \mathscr{I}; j \in \mathscr{J}; m \in \mathscr{M} \tag{25}$$

$$X_{i,j,m} \le \frac{p_{i,j,m}}{1 - p_{i,j,m}} \cdot Z_i \quad \forall\, i \in \mathcal{I}; j \in \mathcal{J}; m \in \mathcal{M} \tag{26}$$

$$\sum_{i \in \mathcal{I}} g_i \cdot X_{i,j,m} \ge \underline{l}_m \cdot Y_{j,m} \quad \forall\, j \in \mathcal{J}; m \in \mathcal{M} \tag{27}$$

$$\sum_{i \in \mathcal{I}} g_i \cdot X_{i,j,m} \le \bar{l}_m \cdot Y_{j,m} \quad \forall\, j \in \mathcal{J}; m \in \mathcal{M} \tag{28}$$

$$\sum_{m \in \mathcal{M}} Y_{j,m} \le 1 \quad \forall\, j \in \mathcal{J} \tag{29}$$

$$\sum_{j \in \mathcal{J}} \sum_{m \in \mathcal{M}} Y_{j,m} = r \tag{30}$$

$$X_{i,j,m} \ge 0 \quad \forall\, i \in \mathcal{I}; j \in \mathcal{J}; m \in \mathcal{M} \tag{31}$$

$$Z_i > 0 \quad \forall\, i \in \mathcal{I} \tag{32}$$

$$Y_{j,m} \in \{0;1\} \quad \forall\, j \in \mathcal{J}; m \in \mathcal{M} \tag{33}$$

To avoid nonlinearity, we enumerate variable $v_{i,j}$ of (8) to generate parameter $v_{i,j,m}$ and hence parameter $p_{i,j,m}$. It would be the same if we used $v_{i,j,q,w}$ instead where (q, w) is a feasible combination of quality level q and waiting time level w.

The objective function (23) maximizes the expected participation (measured as the number of patients that are expected to access preventive healthcare service). Analogous to Haase and Müller (2015, p. 277), (24)–(26) in combination with the objective function (23) are a linear reformulation of the MNL choice probabilities (7). (24) ensures that a demand node i's final choice probabilities to go to service facilities as well as non-attendance sum up to at most 1. The case where the sum is smaller than 1 can be interpreted as rejecting patients at certain facilities, which is a later demand manipulation. This formulation guarantees feasible solutions if mismatches between mode thresholds and mode demand exist. From computational experiments, we know that infeasibilities can occur with synthetic data. As an alternative, we might consider a finer mode structure with much more mode levels instead of only a few coarse ones to avoid infeasibility. This is also a possibility to approximate continuous waiting times.

The linking constraints (25) allow choice probabilities for a facility to be greater than 0 only if the facility is established. Allowing for $p_{i,j,m}$ yields a smaller upper bound by the LP-relaxation than just using $X_{i,j,m} \le Y_{j,m}$ and tighter bounds for $X_{i,j,m}$ (Haase and Müller 2015), because $p_{i,j,m}$ is distinctly smaller than 1. (26) ensures that the pre-calculated constant substitution ratios between the choice probabilities for any two alternatives are obeyed. They are derived from

$\frac{X_{i,j,m}}{Z_i} = \frac{p_{i,j,m}}{1-p_{i,j,m}}$. However, $X_{i,j,m} \neq p_{i,j,m}$ and $Z_i \neq (1 - p_{i,j,m})$ (unless j is the only established facility).

The correct mode in which a facility is established is selected by (27) (lower mode interval threshold) and (28) (upper threshold). Both restriction blocks are necessary, because there might not be a monotonous utility ascent or descent in patronage with consecutive modes, so that one cannot exploit the optimization direction. If a certain facility j is established in mode m, $\sum_{i \in \mathcal{I}} g_i \cdot X_{i,j,m}$ has to be between the lower and the upper mode thresholds. The most complicated feature in this model is probably that the election probabilities are otherwise 0. In the nonlinear formulation, the left-hand side of the constraints always stays the same for one facility and its mode can be selected on the right-hand side [compare with (9), (10) and (11)]. Within this model, demand is only positive for exactly one mode if that facility is established at all. For all other modes, demand is 0, because the corresponding facility-mode combinations do not even exist (see example in Fig. 4).

Constraints (29) ensure that a facility can either only be established in exactly one mode or not at all. (30) provides that r facilities are established. Again, we might instead use a budget constraint, with a parameter denoting fixed establishing costs per facility and mode on the left-hand side and replacing the number of desired facilities r with a budget based on (22). (31)–(33) define the variables' domains.

4.4 Solution strategies

Since Haase and Müller (2014) report promising solution quality and computational effort for the uncapacitated version of our problem using CPLEX, a first shot is to tackle our problem with CPLEX as well. Additionally, we provide tighter bounds on choice probabilities $X_{i,j,m}$ compared to (25) in Appendix 1. Instead of considering only one established facility, we consider the $r-1$ least attractive facilities to compute $p_{i,j,m}$ in (25). We developed bounds on the no-choice alternative to further strengthen the corresponding LP-relaxation (Appendix 1).

The procedure to determine lower bound (Appendix 2) first selects the r most attractive facilities (i.e., facilities with the highest choice probabilities) for each demand point. Based on this solution, we select the r weighted most attractive facilities over all demand points. The weights are the cumulative attractivities of the first step. So far, no decision has been made about the mode m. Rather we have assumed always the most attractive mode is established. However, in the final step we determine the modes of the located facilities. If the solution is integer feasible, we might use the solution for a MIP start.

Since the PHCFLPP obtains numerous fractional variables and their values are given once the decision variable ($Y_{j,m}$) values are known, Benders decomposition might be a worthwhile solution strategy. Furthermore, the procedure to determine a lower bound yields values for a warm start, while the tight bounds help to strengthen the problem. Therefore, both have the potential to improve the performance of the Benders algorithm (Rahmaniani et al. 2017). We consider a (relaxed) Benders Master Problem that contains the locational decision variables Y_{ijm} and the corresponding constraints (29) and (30). The objective of the master problem is to maximize the bound of the Benders optimality cut (generated by the sub-problem). The solution to this mixed-integer problem Y^*_{ijm} is then used in a single sub-problem that only contains the non-integer variables Z_i and X_{ijm}, as well as (23) and the constraints (24)–(28). The values of Y_{ijm} in the sub-problem are given by Y^*_{ijm} from the master problem. The result of the sub-problem (i.e., the solution of the dual of the sub-problem) is used to generate Benders Cuts (feasibility and optimality), which are then included in the master problem. The CPLEX Benders MIP Solver used here first performs the outlined decomposition and uses default presolve and heuristic strategies. The Benders loop then solves the LP relaxation of the master (to optimality). The benders loop is incorporated in a branch-and-cut to determine the (optimal) integer solution. Benders decomposition has been proven to be quite efficient for facility location problems (Cordeau et al. 2019; Codato and Fischetti 2006).

5 Computational study

We have implemented the mixed-integer linear model PHCFLPP (23)–(33) in the General Algebraic Modeling System 24.8.4 (GAMS Development Corporation 2017) and solved it with IBM ILOG CPLEX 12.7.1.0 on a 64-bit workstation running under Windows 10 Pro with two Intel® Xeon® CPU E5-2667 v3. We used synthetic data for our computational study. All data and results can be found on GitHub.[1]

5.1 Data setting

The actual number of candidate facilities, i.e., choice alternatives, is $|\mathscr{J}| \cdot |\mathscr{M}|$. Candidate facility locations are a proper subset of the set of demand nodes, and the stipulated number of established facilities is $r = \lceil 0.5 \cdot |\mathscr{J}| \rceil$. This kind of limitation for r can be understood as being hard to solve (Müller and Haase 2014).

All models consider two quality levels and up to five waiting time levels following Ryan et al. (2008c). The modes are constructed as $\mathscr{M} = \mathscr{Q} \times \mathscr{W}$, where $\mathscr{Q}$ is the set of quality levels and $\mathscr{W}$ the set of waiting time levels, and the corresponding demand thresholds for quality and for waiting time levels are determined in such a way that one waiting time category is divided into two due to the demand requirement to ensure quality (see Fig. 3). Basically, there are $|\mathscr{M}| = |\mathscr{Q}| \cdot |\mathscr{W}| = 2 \cdot 5 = 10$ modes, but only six of them are feasible, because the necessary demands to determine the quality level do not suit all demands needed to set the correct waiting time category. Very high quality in conjunction with very low waiting times, e.g., contradicts each other by definition. The interval thresholds $\underline{l}_m$ and $\overline{l}_m$ are scaled by the instance size. The prevailing discrete waiting time values in weeks are elements of the set $\{0, 2, 4, 6, 8\}$, and the deterministic utility of not taking part ("no-choice")

[1] https://github.com/Urwolfen/ORSP-Preventive-Healthcare-Facility-Location-Planning-with-Quality-Conscious-Clients.

Table 1 Average computational results for ten pseudo-random instances per problem set of PHCFLPP each

	$\lvert\mathscr{J}\rvert$	$\lvert\mathscr{I}\rvert$	r	$\lvert\mathscr{M}\rvert = 4$		$\lvert\mathscr{M}\rvert = 6$	
				GAP (%)	CPU (s)	GAP (%)	CPU (s)
	5	50	3	$0.00^{(10)}$	0.6	$0.00^{(10)}$	0.7
	5	100	3	$0.00^{(10)}$	1.3	$0.00^{(10)}$	1.5
	5	200	3	$0.00^{(10)}$	6.1	$0.00^{(10)}$	4.9
	5	400	3	$0.00^{(10)}$	21.0	$0.00^{(10)}$	26.5
CPLEX Benders	5	400	3			$0.00^{(10)}$	10.6
	10	50	5	$0.00^{(10)}$	5.9	$0.00^{(10)}$	8.7
	10	100	5	$0.00^{(10)}$	19.9	$0.00^{(10)}$	41.8
	10	200	5	$0.00^{(10)}$	73.4	$0.00^{(10)}$	226.7
	10	400	5	$0.00^{(10)}$	329.3	$0.00^{(10)}$	1012.7
CPLEX Benders	10	400	5			$0.00^{(10)}$	325.0
	15	50	8	$0.00^{(10)}$	8.8	$0.00^{(10)}$	89.9
	15	100	8	$0.00^{(10)}$	25.1	$0.00^{(10)}$	582.9
	15	200	8	$0.00^{(10)}$	202.3	$0.10^{(7)}$	1911.0
	15	400	8	$0.00^{(10)}$	327.1	$5.40^{(1)}$	3501.7
CPLEX Benders	15	400	8			$0.01^{(0)}$	3600.0

The GAP superscript in parenthesis stands for the number of proven optimal solutions found for the ten instances. We provide the solutions for selected instances (largest problem sets) using CPLEX's automatic benders decomposition

is $v_{i,\mathrm{no}} = 0$. The integer number of potential clients is stipulated to follow a uniform distribution within a given interval. We arbitrarily set $\beta^{\mathrm{Q}} = 0.24$, $\beta^{\mathrm{W}} = -0.04$, $\beta^{\mathrm{D}} = -0.005$. The distance $d_{i,j}$ is calculated as the Euclidean distance between nodes $i \in \mathscr{I}$ and $j \in \mathscr{J} \subseteq \mathscr{I}$, which are placed on a 170×170 plane although $d_{i,j} = 2$ for $j = i$.

5.2 Results

We solved ten pseudo-random instances per problem set in a row and applied the lower bound procedure for the objective function value (Appendix 2). Thereby we observed that the solutions of Step I's (38)–(41) and Step II's (42)–(44) LP-relaxations were already integer. We document the results obtained for our main model PHCFLPP in Table 1.

Additionally, we design a second setting with only four modes by leaving out the first and last waiting time levels. "CPU" gives the time in CPU seconds CPLEX used to solve PHCFLPP. "GAP" displays the integrality gap. The maximum computational time for a model solution was limited to one hour.

When we left the CPLEX default branch-and-bound settings enabled, sometimes the solver found solutions faster when no lower bound was provided by Step III (45), (24)–(28), (31)–(33) and (46), but PHCFLPP was solved alone. This circumstance

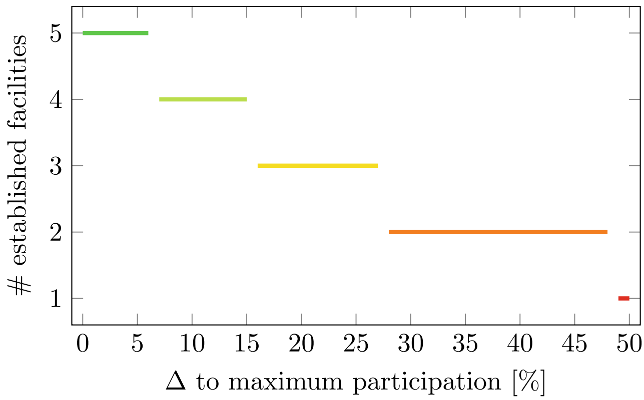


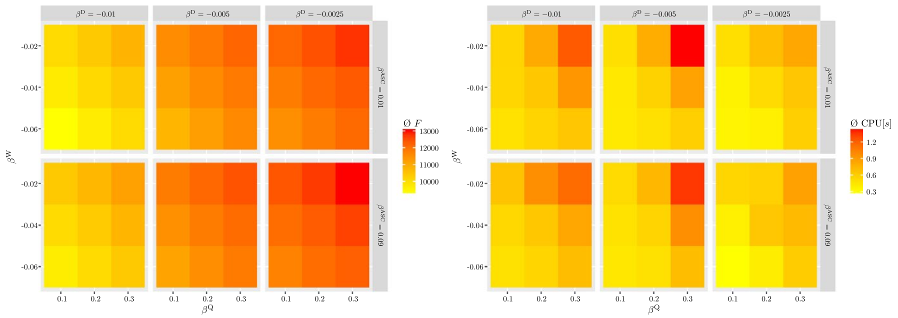
Fig. 5 Participation subject to the number of facilities: if fewer facilities are required due to financial issues, the manager has to decide how much participation he is willing to sacrifice compared to the optimal solution. With more relative deviation of patients' participation, a smaller number of established facilities is allowed. After calculating the optimal objective function value, we changed the objective function to minimize $\sum_{j\in\mathscr{J}}\sum_{m\in\mathscr{M}} Y_{j,m}$ and used $\sum_{i\in\mathscr{I}}\sum_{j\in\mathscr{J}}\sum_{m\in\mathscr{M}} g_i \cdot X_{i,j,m} \geq \frac{100-\Delta}{100} \cdot F$. The problem size is $|\mathscr{J}| = 10$, $|\mathscr{I}| = 50$, $|\mathscr{M}| = 6$

makes us question the theoretically beneficial auxiliary heuristic. In Haase and Müller (2015), it performed well. Note that we always performed a CPLEX MIP start, i.e., we provided a solution vector (variable values) to start from. When we required just $F \geq LB$ instead, the solution performance was worse.

Our experiments indicate that most of the computational effort is needed to prove optimality after an integer feasible solution was found. We also observed that solving performance can be strongly dependent on the data setting. However, as we are considering a strategic planning horizon, the necessary CPU times are satisfying. Additional experiments revealed that using GAMS/CPLEXD and utilizing the CPLEX built-in Benders Algorithm with fully automatic decomposition mode, the calculation times can be further reduced. We used Benders here only on the largest instances that we considered in this study. To be specific, we applied Benders only to instances with 400 demand points and 6 modes, which are the largest instances considered here. As long as we are able to solve the instances to optimality, the Benders approach reduced computation times by more than 50%. For the largest problem set, Benders found no optimal solution, i.e., our approach performed better on one of the ten random instances. However, over all ten instances the Benders approach found better solutions (i.e., smaller integrality gaps).

5.3 Analysis and managerial insights

To make the effect of the number of open facilities visible, we investigated the results if one wants to minimize the number of established facilities subject to a certain level of participation (see Fig. 5). Thus, we see a trade-off between establishing costs per facility and expected participation. We start from five established facilities. When the manager is willing to lose 7% of the maximum participation, which can be



(a) Objective function values

(b) Calculation times

Fig. 6 Average objective function values F and CPU calculation times for ten pseudo-random instances of a $|\mathscr{J}| = 5$, $|\mathscr{M}| = 6$, $|\mathscr{I}| = 50$, $r = 3$ problem set where β^{ASC}, β^{D}, β^{Q} and β^{W} were varied each time. Generated with R (R Core Team 2017)

achieved with five facilities, he could establish one facility less. A further reduction to only three facilities would result in a participation loss of at least 16%. At least 49% of the maximum participation is about to be sacrificed by the establishment of only one facility.

The results of a sensitivity analysis can be found in Fig. 6. We generated ten instances of a problem set and varied all utility function coefficients each time. For every single parameter setting, the average objective function values or computational efforts are illustrated, respectively, to get a deeper insight into their interrelation. When we distinguish between a parameter's weight and its direction (positive/negative), we see that higher participation F is possible with a decreasing (negative) influence of proximity β^{D}. People are willing to travel longer distances, and the spatial facility setting becomes less important. When the weight of waiting time β^{W} becomes less negative, participation is higher as well due to a decreasing importance of waiting for an appointment, which clients consider. Thus, growing waiting time caused by higher participation is less influential. A larger impact of quality β^{Q} also leads to a higher level of attractiveness and higher participation. Note that this only affects the higher quality category. If there are enough participating patients to ensure the minimum quantity requirement, this quality level generates a higher utility (all other things being equal). Furthermore, the negative influence of higher waiting time levels can be compensated more easily. A higher alternative-specific constant increases utility and participation as illustrated by darker rectangles in the lower half of Fig. 6a. The alternative-specific constant can represent the service quality at a facility or the waiting time in the waiting room.

As expected, the alternative-specific constant is not responsible for different CPU calculation times. However, we see a higher computational effort if the weight of waiting time decreases. In those scenarios, it is more difficult to distinguish between facility modes, i.e., solutions, regarding the objective function value. With growing importance of quality of care, the trade-off between higher quality and competing longer waiting times for an appointment seems to complicate optimal solution finding. Interestingly, while the effect of proximity follows the waiting time pattern,

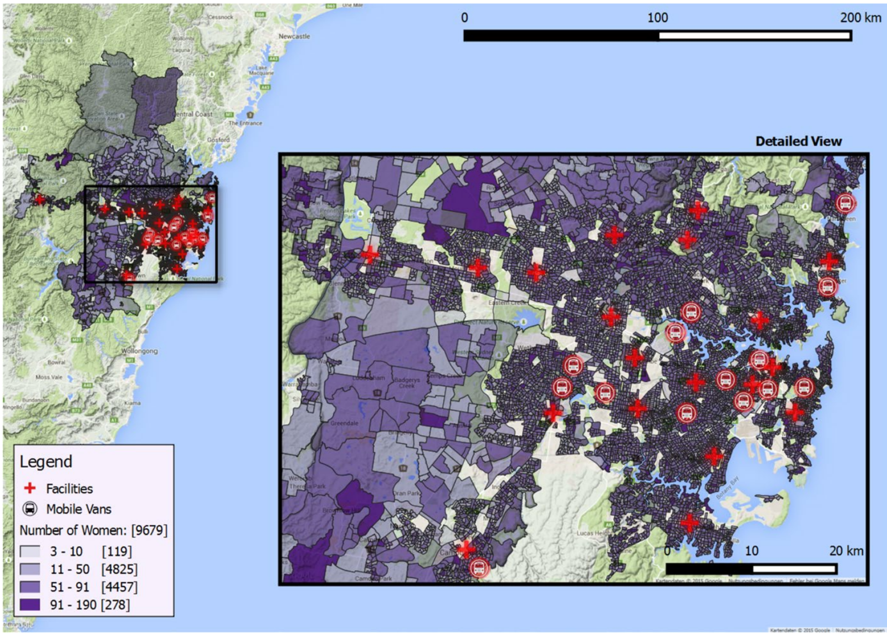


Fig. 7 Map with selected statistical areas of Greater Sydney and locations for breast cancer screening. The detail shows the city of Sydney. The statistical areas (SA1) are colored with respect to the number of resident women that are eligible for the breast cancer screening program. The numbers in brackets stand for the number of respective SA1. Facilities are buildings, whereas mobile van locations are dynamic. The vans change their locations every few months

there is a distance-waiting time combination that results in extraordinary high computational effort. Instead, with $|\beta^{\mathrm{D}}|$ being lower or higher, its effect vanishes or overlays, respectively.

6 Case study

For our real-world case study, we use the results of Gerard et al. (2008), who investigated women's preferences for breast cancer screening arrangements. Since their discrete choice experiments were conducted in Sydney, Australia, we chose the area of Greater Sydney for the application of our model. The task is to find a facility location setting that maximizes women's participation. All data and results can be found on GitHub.[2]

To represent the geographic area, we combined publicly available geographical and census data processed with QGIS (QGIS Development Team 2015) and provided by the Australian Bureau of Statistics (Australian Bureau of Statistics 2011a,

[2] https://github.com/Urwolfen/ORSP-Preventive-Healthcare-Facility-Location-Planning-with-Quality-Conscious-Clients.

Table 2 Elements of constant part $\bar{c}_{i,j}$ of deterministic utility $v_{i,j,m}$

Variable	Explanation	Value
Constant	Base utility	− 3.5390
TTFX0	travel time coefficient 0–20 min	0.4020
TTFX1	travel time coefficient 20–40 min	0.3370
TTFX2	travel time coefficient 40–60 min	0.0020
TTFX3	travel time coefficient 60–120 min	− 0.7410
HINFFX0	Personal letter from service	0.2270
ISHT	Information sheet included	0.1960
STAFF	Staff relates in a welcoming manner	0.1720
PRIV	Privacy of changing facilities: private	0.4100
SCTFX1	Screening time: 30 min	0.2260
WRFX3	Results in 14 working days	0.0370
ACCFX25	Accuracy of test: 99% (interpolated)	1.1142
ISHHINFX0	ISHT * HINFFX0	0.0090
STAFF_PR	STAFF * PRIV	0.1110
GETBC1	Sometimes thinks of breast cancer	− 0.0330
MARITAL	Married	− 0.2510
AGE1	Aged between 50 and 69 years	− 0.4360
ACT1	Part-time employment	0.2110

b). We selected statistical areas (SA1) in the area around Sydney (see Fig. 7) and determined the number of women aged between 50 and 74 years living there, because they are eligible for screening (BreastScreen Australia 2014). To generate the demand node locations, we calculated the spatial units' centroids on a plane UTM coordinate system. The locations of 20 existing screening facilities and 14 candidate mobile screening van locations were taken from the internet presence of BreastScreen New South Wales (BreastScreen New South Wales 2015) and assigned to the nearest SA1 each. We approximated real distances between all centroid nodes by Manhattan distance and assume a constant travel speed of 25 km/h to transform distances into travel times (plus a constant of 5 min). In addition, we manually identified 22 further areas as candidate locations. To be consistent with Gerard et al. (2008) we excluded connections with travel times of 120 min or more. The constant part of individuals' utility functions $\bar{c}_{i,j}$ consists of the elements listed in Table 2. With increasing travel time, women are less likely to participate. Furthermore, women are interested in a comfortable environment at a facility and prefer short screening times as well as quick results (Gerard et al. 2008). The positive sign for part-time employment indicates the potential to improve the offer for full-time employees. Waiting one week for an appointment is evaluated with 0, four weeks with −0.083. Since the quality of care could not be appropriately presented with the available empirical data, we only considered those two modes of shorter or longer waiting time. If the travel time between i and j is 30 min and i has to wait four weeks for an appointment, e.g., $v_{i,j,m_2} = -0.083 - 3.539 + 0.337 + 0.227 + 0.196 + 0.172 + 0.41 + 0.226 + 0.037 + 1.1142 + 0.009 + 0.111 - 0.033 - 0.251 - 0.436 + 0.211 = -1.2918$. The threshold for a waiting time of four weeks was set to 10,000 participants. The resulting problem size is $|\mathscr{J}| = 56$, $|\mathscr{M}| = 2$, $|\mathscr{I}| = 9{,}679$, $r = 20 + 14 = 34$.

We found a facility location setting with an objective function value of approximately 417,742.9 expected participants and an average no-show rate of 25.5% after 18 min CPU time (3.7% gap). The same parameter constellation evaluates the real situation with 412,728.7 (26.5% no-show), which is an improvement of more than five thousand expected patients (about 1.2%).

7 Concluding remarks

Previous location planning papers, which consider congestion, end up with nonlinear programs, that are usually solved by heuristics. Alternatively, they formulate mixed-integer linear programs, but in turn limit the number of arriving clients. In contrast, we showed how to determine facilities' optimal locations as well as their capacity levels simultaneously, when customers' choice behavior is probabilistic and individual choices are dependent on other individuals' choices. The objective of our mixed-integer linear program is to maximize market share. For planning managers in charge it gets possible to account for manifold properties of interest while locating facilities, far beyond minimizing travel distances, i.e., accounting for clients' complete utility function so that managers gain more realistic solutions to maximize market share. Our model is directly applicable to various fields, e.g., to facility location problems, assortment optimization or product line design planning. Within the healthcare context, our model provides a framework for policy makers. The non-urgent nature of preventive medical examinations is predestined for using our model, but other problems can be displayed, too, as long as there is a choice situation for clients. The vital requirements are that 1. there is a demand and a supply side, 2. clients face a discrete choice situation, and 3. available alternatives are induced by demand. There is a straightforward link to the research fields assortment optimization or product line design planning (Müller and Haase 2016; Gallego and Topaloglu 2014). On one hand, the product price is a determinant for customers' choice. On the other hand, it can be fixed subject to the demand level and potentially decreases with growing customer interest and increased production if economies of scale apply. Locating branches with different hours of business is another possible application.

The large number of parameters that are part of the deterministic utility function does not pose a problem. As broadly described in the literature and conducted in reality for decades, discrete choice experiments can be used to estimate those parameters (Liu et al. 2018; de Bekker-Grob et al. 2012; Lancsar and Louviere 2008; Ryan et al. 2008a, b, c; Street et al. 2008; Viney et al. 2002). For real applications, insight is needed regarding the choice-making process in practice as well as the relationship between the number of clients accessing service and corresponding facility capacity levels, i.e., modes. The data can be generated employing a survey or a simulation that helps to understand that mechanism.

Other assumptions concerning the distribution of the utility function's error terms would result in different mathematical models. The i.i.d. EV distribution assumption leads to the IIA property, which means constant substitution patterns across alternatives. If correlations are detected empirically, other distributions are necessary. This variation is a valuable challenge for future research. Furthermore, one might investigate under which conditions our model is superior to classical facility location models with a focus on distances or travel times only and when it is sufficient to use simplistic models. Finally, whereas in our model longer waiting times for an appointment mean less utility, we can consider a scenario with a supply cutoff such that a number of clients up to a capacity threshold are treated, while others are not. The latter group then has a 100% no-choice probability. This only slightly different verbal description is likely to result in a very different mathematical model, which is an interesting topic for further research. Moreover, tailored Benders decomposition for our problem (and its extensions in Sect. 4.2) seems a very promising path to follow. However, using nonlinear MIP solver to solve the original nonlinear MIP might be worth considering as well—in particular in combination with an outer approximation scheme. Finding efficient solution procedures remains an open avenue for future research.

Acknowledgements The authors are very grateful for the very helpful suggestions of two anonymous reviewers to improve the paper's quality.

Appendix 1: Tighter bounds on variables

We can improve the upper bound for $X_{i,j,m}$ in (25): If it is certain that r facilities will exist, we do not need to do the calculation as if (j, m) was the only facility, but we can assume that $r-1$ other facilities will be established as well. We replace $p_{i,j,m}$ and get

$$X_{i,j,m} \leq \frac{e^{v_{i,j,m}}}{e^{v_{i,\text{no}}} + e^{v_{i,j,m}} + \sum_{k \in \mathscr{J}_i^{\min}} e^{v_{i,k,\check{m}}}} \cdot Y_{j,m} \quad \forall\, i \in \mathscr{I}; j \in \mathscr{J}; m \in \mathscr{M} \tag{34}$$

where $\mathscr{J}_i^{\min}$ is the set of the $r-1$ least attractive other facility locations for demand node i and $\check{m}$ represents the globally least attractive mode. This change reduces solution times remarkably—a finding that is also discussed in Freire et al. (2016).

It is also helpful to determine bounds for the no-choice probability variables Z_i. They can be easily computed by assuming worst case or best case scenarios. Let $\mathscr{J}_i^{\text{worst}}$ and $\mathscr{J}_i^{\text{best}}$ be the sets of the r least or most attractive alternatives j for demand node i, and $\hat{m}$ the most attractive mode. Then we get upper bounds Z_i^{UB} and lower bounds Z_i^{LB}:

$$Z_i^{\text{UB}} = \frac{e^{v_{i,\text{no}}}}{e^{v_{i,\text{no}}} + \sum_{k \in \mathscr{J}_i^{\text{worst}}} e^{v_{i,k,\check{m}}}}, \tag{35}$$

$$Z_i^{\text{LB}} = \frac{e^{v_{i,\text{no}}}}{e^{v_{i,\text{no}}} + \sum_{k \in \mathscr{J}_i^{\text{best}}} e^{v_{i,k,\hat{m}}}}. \tag{36}$$

If we also decided about the number of established facilities r with a variation of (30) to

$$\sum_{j \in \mathcal{J}} \sum_{m \in \mathcal{M}} Y_{j,m} \leq r, \tag{37}$$

our model is harder to solve and (34) and (35) would not be applicable.

Appendix 2: Modeling framework to derive a lower bound

Since it may take a long time to find a first integer solution within the solving process, it is expedient to determine a lower bound for our linear model PHCFLPP (23)–(33). We present three models that altogether yield a lower bound. This approach is basically introduced in Haase and Müller (2015). It relies on the constant part of individuals' utility functions, which mainly depends on distances.

Step I

With the first auxiliary model the highest choice probabilities for each single demand node are chosen. I.e., the most attractive facilities for them are identified. We continue our nomenclature:

Additional variables

$\Upsilon_{i,j,m}$	= 1 if demand node i is assigned to location j in mode m; 0, otherwise
F^{I}	Objective function value of Model Step I (cumulated r highest choice probabilities)

$$\text{Maximize} \quad F^{\mathrm{I}} = \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} \sum_{m \in \mathcal{M}} p_{i,j,m} \cdot \Upsilon_{i,j,m} \tag{38}$$

subject to

$$\sum_{j \in \mathcal{J}} \sum_{m \in \mathcal{M}} \Upsilon_{i,j,m} \leq r \quad \forall\, i \in \mathcal{I} \tag{39}$$

$$\sum_{m \in \mathcal{M}} \Upsilon_{i,j,m} \leq 1 \quad \forall\, i \in \mathcal{I}; j \in \mathcal{J} \tag{40}$$

$$\Upsilon_{i,j,m} \in \{0;1\} \quad \forall\, i \in \mathcal{I}; j \in \mathcal{J}; m \in \mathcal{M} \tag{41}$$

The objective function (38) maximizes the cumulated chosen choice probabilities subject to (39), which means that the at most r highest choice probabilities for each demand node i are selected. (40) ensures that a facility can only be demanded in exactly one mode m.

Step II

In the second auxiliary model, the r overall most attractive facilities are established. The specific mode is not of interest here and the best one is always chosen. The decision is based on the remaining influences, mainly the distance between demand and supply nodes. An additional parameter that makes use of the solution to Step I sums up for how many demand nodes a certain facility belongs to the most attractive ones. The more demand nodes assigned to a facility the higher its attraction by this definition. We further extend our nomenclature:

Additional parameter

$b_{j,m}$ Attractiveness value for each facility at location j in mode m with $b_{j,m} = \sum_{i \in \mathscr{I}} \Upsilon^*_{i,j,m}$ where $\Upsilon^*_{i,j,m}$ is the optimal solution to Step I

Additional variables

$\tilde{Y}_{j,m}$ = 1 if facility at location j is specified to offer healthcare service in mode m; 0, otherwise

F^{II} Objective function value of Model Step II (attractiveness of located facilities)

$$\text{Maximize} \quad F^{\mathrm{II}} = \sum_{j \in \mathscr{J}} \sum_{m \in \mathscr{M}} b_{j,m} \cdot \tilde{Y}_{j,m} \tag{42}$$

subject to

$$\sum_{j \in \mathscr{J}} \sum_{m \in \mathscr{M}} \tilde{Y}_{j,m} = r \tag{43}$$

$$\tilde{Y}_{j,m} \in \{0;1\} \quad \forall j \in \mathscr{J}; m \in \mathscr{M} \tag{44}$$

The objective function (42) maximizes the cumulated overall attractiveness and thereby chooses the r most demanded facilities according to (43). The result represents a solution that depends on the individuals' constant part of their deterministic utility function disregarding capacities.

Step III

The last auxiliary model is used to determine the final modes for the facility locations previously identified in Step II. We consider

Additional parameter

$\tilde{Y}^*_{j,m}$ optimal solution to Step II

Additional variable

LB objective function value of Model Step III (a lower bound for PHCFLPP)

$$\text{Maximize} \quad LB = \sum_{i \in \mathscr{I}} g_i \cdot \sum_{j \in \mathscr{J}} \sum_{m \in \mathscr{M}} X_{i,j,m} \tag{45}$$

subject to (24)–(28), (31)–(33) as well as

$$\sum_{m \in \mathscr{M}} Y_{j,m} = \sum_{m \in \mathscr{M}} \tilde{Y}^*_{j,m} \quad \forall j \in \mathscr{J}. \tag{46}$$

The first restriction blocks are in fact PHCFLPP without two redundant constraints of which information is already included in Step I and Step II. The information is that only one mode per facility can be present as well as that r facilities are established. (45) maximizes the expected participation by now selecting the final modes of the predestined facility locations via (46). The objective function value is a lower bound for PHCFLPP as it is a feasible integer solution to it. It is allowed to add up the pre-defined locations on the right hand side of (46), because at most one value can equal 1 due to (40). Thus, (46) is a substitute for (29) and a linking constraint in addition. Those constraints can be left out in Step III, as well as (30), because of (43). So we can perform a MIP start with PHCFLPP after solving the Steps I, II and III in a row.

References

Anand KS, Paç MF, Veeraraghavan S (2011) Quality-speed conundrum: trade-offs in customer-intensive services. Manag Sci 57(1):40–56

Aros-Vera F, Marianov V, Mitchell JE (2013) p-Hub approach for the optimal park-and-ride facility location problem. Eur J Oper Res 226(2):277–285

Australian Bureau of Statistics (2011a) Greater Sydney (Statistical Area Level 1), Basic Community Profile, Excel spreadsheet 2011 Census B04 A+B New South Wales SA1, AGE BY SEX. Viewed May 16, 2018. http://www.abs.gov.au/websitedbs/censushome.nsf/home/census?opendocument&navpos=10

Australian Bureau of Statistics (2011b) Statistical Area Level 1 (SA1) ASGS Ed 2011 Digital Boundaries in ESRI Shapefile Format. Viewed May 16, 2018. http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1270.0.55.001July

Ayer T, Alagoz O, Stout NK, Burnside ES (2016) Heterogeneity in women's adherence and its role in optimal breast cancer screening policies. Manag Sci 62(5):1339–1362

Baron O, Berman O, Krass D (2008) Facility location with stochastic demand and constraints on waiting time. Manuf Serv Oper Manag 10(3):484–505

Batt RJ, Terwiesch C (2015) Waiting patiently: an empirical study of queue abandonment in an emergency department. Manag Sci 61(1):39–59

Belloni A, Freund R, Selove M, Simester D (2008) Optimizing product line designs: efficient methods and comparisons. Manag Sci 54(9):1544–1552

Benati S, Hansen P (2002) The maximum capture problem with random utilities: problem formulation and algorithms. Eur J Oper Res 143(3):518–530
Besbes O, Sauré D (2016) Product assortment and price competition under multinomial logit demand. Prod Oper Manag 25(1):114–127
Blankart CR (2012) Does healthcare infrastructure have an impact on delay in diagnosis and survival? Health Policy 105(2–3):128–137
Blum K, Offermanns M (2004) Umverteilungswirkungen der Mindestmengenregelung. das Krankenhaus 96(10):787–790
BreastScreen Australia (2014) BreastScreen and you—information about mammography screening. Viewed May 16, 2018. http://cancerscreening.gov.au/internet/screening/publishing.nsf/Content/breastscreen-and-you
BreastScreen Australia (2015) National Accreditation Handbook. Viewed May 16, 2018. http://www.cancerscreening.gov.au/internet/screening/publishing.nsf/Content/acc-hb-2015
BreastScreen New South Wales (2015) Viewed August 18, 2015. http://www.bsnsw.org.au/
Castillo I, Ingolfsson A, Sim T (2009) Social optimal location of facilities with fixed servers, stochastic demand, and congestion. Prod Oper Manag 18(6):721–736
Chen H, Qian Q, Zhang A (2015) Would allowing privately funded health care reduce public waiting time? Theory and empirical evidence from Canadian joint replacement surgery data. Prod Oper Manag 24(4):605–618
Codato G, Fischetti M (2006) Combinatorial Benders' cuts for mixed-integer linear programming. Oper Res 54(4):756–766
Cordeau J-F, Furini F, Ljubic I (2019) Benders decomposition for very large scale partial set covering and maximal covering location problems. Eur J Oper Res 275(3):882–896
Dan T, Marcotte P (2019) Competitive facility location with selfish users and queues. Oper Res 67:479–497
de Bekker-Grob EW, Ryan M, Gerard K (2012) Discrete choice experiments in health economics: a review of the literature. Health Econ 21(2):145–172
Elhedhli S (2006) Service system design with immobile servers, stochastic demand, and congestion. Manuf Serv Oper Manag 8(1):92–97
Espinoza Garcia JC, Alfandari L (2018) Robust location of new housing developments using a choice model. Ann Oper Res 271:527–550
Feldman J, Liu N, Topaloglu H, Ziya S (2014) Appointment scheduling under patient preference and no-show behavior. Oper Res 62(4):794–811
Freire AS, Moreno E, Yushimito WF (2016) A branch-and-bound algorithm for the maximum capture problem with random utilities. Eur J Oper Res 252(1):204–212
Gallego G, Topaloglu H (2014) Constrained assortment optimization for the nested logit model. Manag Sci 60(10):2583–2601
GAMS Development Corporation (2017) General Algebraic Modeling System (GAMS) Release 24.8.4. GAMS Development Corporation, Washington, DC, USA
Gerard K, Shanahan M, Louviere J (2003) Using stated preference discrete choice modelling to inform health care decision-making: a pilot study of breast screening participation. Appl Econ 35(9):1073–1085
Gerard K, Shanahan M, Louviere J (2008) Using discrete choice modelling to investigate breast screening participation. In: Ryan M, Gerard K, Amaya-Amaya M (eds) Using discrete choice experiments to value health and health care, volume 11 of the economics of non-market goods and resources. Springer, Dordrecht, pp 117–137
Güneş ED, Chick SE, Akşin OZ (2004) Breast cancer screening services: trade-offs in quality, capacity, outreach, and centralization. Health Care Manag Sci 7(4):291–303
Güneş ED, Örmeci EL, Kunduzcu D (2015) Preventing and diagnosing colorectal cancer with a limited colonoscopy resource. Prod Oper Manag 24(1):1–20
Haase K (2009) Discrete location planning. Technical Report WP-09-07, Institute of Transport and Logistics Studies, University of Sydney
Haase K, Müller S (2013) Management of school locations allowing for free school choice. Omega 41(5):847–855
Haase K, Müller S (2014) A comparison of linear reformulations for multinomial logit choice probabilities in facility location models. Eur J Oper Res 232(3):689–691
Haase K, Müller S (2015) Insights into clients' choice in preventive health care facility location planning. OR Spectr 37(1):273–291

Harewood GC (2005) Relationship of colonoscopy completion rates and endoscopist features. Dig Dis Sci 50(1):47–51

Harper PR, Shahani AK, Gallagher JE, Bowie C (2005) Planning health services with explicit geographical considerations: a stochastic location-allocation approach. Omega 33(2):141–152

Hol L, de Bekker-Grob EW, van Dam L, Donkers B, Kuipers EJ, Habbema JDF, Steyerberg EW, van Leerdam ME, Essink-Bot M-L (2010) Preferences for colorectal cancer screening strategies: a discrete choice experiment. Br J Cancer 102(6):972–980

Killaspy H, Banerjee S, King M, Lloyd M (2000) Prospective controlled study of psychiatric out-patient non-attendance. Br J Psychiatry 176(2):160–165

Koppelman FS, Bhat C (2006) A self instructing course in mode choice modeling: multinomial and nested logit models. US Department of Transportation, Federal Transit Administration

Lacy NL, Paulman A, Reuter MD, Lovejoy B (2004) Why we don't come: patient perceptions on no-shows. Ann Fam Med 2(6):541–545

Lancsar E, Louviere J (2008) Conducting discrete choice experiments to inform healthcare decision making. PharmacoEconomics 26(8):661–677

Liu N, Finkelstein SR, Kruk ME, Rosenthal D (2018) When waiting to see a doctor is less irritating: understanding patient preferences and choice behavior in appointment scheduling. Manag Sci 64(5):1975–1996

Ljubić I, Moreno E (2018) Outer approximation and submodular cuts for maximum capture facility location problems with random utilities. Eur J Oper Res 266(1):46–56

Madadi M, Zhang S, Henderson LM (2015) Evaluation of breast cancer mammography screening policies considering adherence behavior. Eur J Oper Res 247(2):630–640

Marianov V, Ríos M, Icaza MJ (2008) Facility location for market capture when users rank facilities by shorter travel and waiting times. Eur J Oper Res 191(1):32–44

McFadden D (2001) Economic choices. Am Econ Rev 91(3):351–378

Moscelli G, Siciliani L, Gutacker N, Gravelle H (2016) Location, quality and choice of hospital: evidence from England 2002–2013. Reg Sci Urban Econ 60:112–124

Müller S, Haase K (2014) Customer segmentation in retail facility location planning. Bus Res 7(2):235–261

Müller S, Haase K (2016) On the product portfolio planning problem with customer-engineering interaction. Oper Res Lett 44(3):390–393

Osadchiy N, Diwas K (2017) Are patients patient? The role of time to appointment in patient flow. Prod Oper Manag 26(3):469–490

QGIS Development Team (2015). QGIS Geographic Information System. Open Source Geospatial Foundation Project

R Core Team (2017) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Rahmaniani R, Crainic TG, Gendreau M, Rei W (2017) The benders decomposition algorithm: a literature review. Eur J Oper Res 259(3):801–817

Rusmevichientong P, Shmoys D, Tong C, Topaloglu H (2014) Assortment optimization under the multinomial logit model with random choice parameters. Prod Oper Manag 23(11):2023–2039

Ryan M, Gerard K, Amaya-Amaya M (2008a) Discrete choice experiments in a nutshell. In: Ryan M, Gerard K, Amaya-Amaya M (eds) Using discrete choice experiments to value health and health care, volume 11 of the economics of non-market goods and resources. Springer, Dordrecht, pp 13–46

Ryan M, Gerard K, Watson V, Street DJ, Burgess L (2008b) Practical issues in conducting a discrete choice experiment. In: Ryan M, Gerard K, Amaya-Amaya M (eds) Using discrete choice experiments to value health and health care, volume 11 of the economics of non-market goods and resources. Springer, Dordrecht, pp 73–97

Ryan M, Skåtun D, Major K (2008c) Using discrete choice experiments to go beyond clinical outcomes when evaluating clinical practice. In: Ryan M, Gerard K, Amaya-Amaya M (eds) Using discrete choice experiments to value health and health care, volume 11 of the economics of non-market goods and resources. Springer, Dordrecht, pp 101–116

Schrag D, Cramer LD, Bach PB, Cohen AM, Warren JL, Begg CB (2000) Influence of hospital procedure volume on outcomes following surgery for colon cancer. JAMA 284(23):3028–3035

Street DJ, Burgess L (2007) The construction of optimal stated choice experiments: theory and methods. Wiley, Hoboken

Street DJ, Burgess L, Viney R, Louviere J (2008) Designing discrete choice experiments for health care. In: Ryan M, Gerard K, Amaya-Amaya M (eds) Using discrete choice experiments to value health and health care, volume 11 of the economics of non-market goods and resources. Springer, Dordrecht, pp 47–72

Tiwari V, Heese HS (2009) Specialization and competition in healthcare delivery networks. Health Care Manag Sci 12(3):306–324

Train KE (2009) Discrete choice methods with simulation, 2nd edn. Cambridge University Press, Cambridge

Truong V-A (2015) Optimal advance scheduling. Manag Sci 61(7):1584–1597

van Dam L, Hol L, de Bekker-Grob EW, Steyerberg EW, Kuipers EJ, Habbema JDF, Essink-Bot M-L, van Leerdam ME (2010) What determines individuals' preferences for colorectal cancer screening programmes? A discrete choice experiment. Eur J Cancer 46(1):150–159

van der Pol M, Currie G, Kromm S, Ryan M (2014) Specification of the utility function in discrete choice experiments. Value Health 17(2):297–301

Verter V, Lapierre SD (2002) Location of preventive health care facilities. Ann Oper Res 110(1–4):123–132

Vidyarthi N, Kuzgunkaya O (2015) The impact of directed choice on the design of preventive healthcare facility network under congestion. Health Care Manag Sci 18(4):459–474

Viiala CH, Tang KW, Lawrance IC, Murray K, Olynyk JK (2007) Waiting times for colonoscopy and colorectal cancer diagnosis. Med J Aust 186(6):282–285

Viney R, Lancsar E, Louviere J (2002) Discrete choice experiments to measure consumer preferences for health and healthcare. Expert Rev Pharmacoecon Outcomes Res 2(4):319–326

Wang R, Wang Z (2017) Consumer choice models with endogenous network effects. Manag Sci 63(11):3944–3960

Wang Q, Batta R, Rump CM (2002) Algorithms for a facility location problem with stochastic customer demand and immobile servers. Ann Oper Res 111(1–4):17–34

Wang D, Morrice DJ, Muthuraman K, Bard JF, Leykum LK, Noorily SH (2018) Coordinated scheduling for a multi-server network in outpatient pre-operative care. Prod Oper Manag 27(3):458–479

Zhang Y, Berman O, Verter V (2009) Incorporating congestion in preventive healthcare facility network design. Eur J Oper Res 198(3):922–935

Zhang Y, Berman O, Marcotte P, Verter V (2010) A bilevel model for preventive healthcare facility network design with congestion. IIE Trans 42(12):865–880

Zhang Y, Berman O, Verter V (2012) The impact of client choice on preventive healthcare facility network design. OR Spectr 34(2):349–370