Journal Pre-proof

Capacity and Assortment Planning under One-way Supplier-driven Substitution for Pharmacy Kiosks with Low Drug Demand

Gohram Baloch, Fatma Gzara

PII: S0377-2217(19)30745-3
DOI: https://doi.org/10.1016/j.ejor.2019.09.007
Reference: EOR 16035

To appear in: *European Journal of Operational Research*

Received date: 31 May 2018
Accepted date: 3 September 2019

Please cite this article as: Gohram Baloch, Fatma Gzara, Capacity and Assortment Planning under One-way Supplier-driven Substitution for Pharmacy Kiosks with Low Drug Demand, *European Journal of Operational Research* (2019), doi: https://doi.org/10.1016/j.ejor.2019.09.007

Highlights

- An analytics project aimed to determine optimized pharmacy kiosk capacity.
- Models for assortment problem under one-way substitution.
- A column-generation based heuristic approach is proposed.

Capacity and Assortment Planning under One-way Supplier-driven Substitution for Pharmacy Kiosks with Low Drug Demand

Gohram Baloch, Fatma Gzara*

Department of Management Sciences, University of Waterloo

Abstract

MedAvail Technologies Inc. is a healthcare technology company that develops new technologies for self-serve pharmacy solutions. The technology, called MedCenter, is a pharmacy kiosk that provides 24/7, easy, and reliable access to pre-packaged prescription drugs and over the counter medications. To meet its business goals of having the right medication in the right kiosk at the right quantity, MedAvail faces several challenges related to assortment and stocking decisions of medications in the kiosk limited by kiosk capacity. This research addresses these decisions through an analytics project aimed at analyzing pharmaceutical sales, determining optimized kiosk storage capacity and service levels, and recommending assortment, stocking, and supplier-driven product substitution guidelines. We developed several mixed integer optimization models that use sales data to obtain robust solutions with respect to randomness in demand. We perform extensive testing using real as well as randomly generated data, and under multiple substitution rules, replenishment guidelines, and demand prediction strategies. Our results show that supplier-driven product substitution could save up to 9% in storage capacity depending on the desired service level and characteristics of product demand. We also propose a column-generation based heuristic approach that, on average, obtains near optimal solutions within 1.1% of optimality gap while reducing computational times by a factor of three.

Keywords: OR in Health services, capacity planning, assortment, column generation

1. Introduction

A recent innovation in the healthcare space is the automated medication dispensing system where an ATM style kiosk dispenses both prescription and over the counter medications. The global market for such a system enjoys an annual growth rate of 6.7% and is expected to reach $3.6 billion by 2018 (Slawsky, 2015). Several companies are developing self-serve kiosks with the purpose of providing 24 hour service, extending pharmacy operations to remote areas, and reducing both setup and operating costs. One such kiosk, namely MedCenter, shown in Figure 1, was developed by MedAvail Technologies Inc., a healthcare technology company based in Canada.

*Correspondence: F. Gzara, Department of Management Sciences, University of Waterloo, 200 University Ave. West, Waterloo, Ontario, N2L 3G1, E-mail: fgzara@uwaterloo.ca

Email addresses: ggohram@uwaterloo.ca (Gohram Baloch), fgzara@uwaterloo.ca (Fatma Gzara)

Figure 1: MedAvail's MedCenter Kiosk source: (MedAvail, 2017)

Launched in 2013, MedCenters are now successfully deployed in US, Canada, and Switzerland where they are installed in pharmacies, retail stores, hospitals, community clinics, university campuses, and medical office buildings. The MedCenter dispenses prescription drugs and over the counter (OTC) products under the supervision of a remote pharmacist (MedAvail, 2017). It consists of multiple bins, each divided into several slots where a single slot can store various packages each containing a specific drug of a particular quantity. When customers arrive at a MedCenter, they insert their prescription into the kiosk to be scanned and are connected to a live pharmacist who verifies if the medicines are in stock. A medication is considered available only if the drug is stocked in a package with the exact requested quantity. Once the customer pays for the medications, the pharmacist authorizes the release of the prescription and the automated kiosk picks and dispenses it. If the medications are not stocked, a customer may request the pharmacist to call the physician for a substitute, to transfer the prescription to the home pharmacy, or to just cancel the order request.

In comparison to traditional brick-and-mortar pharmacies, MedCenters are significantly less expensive, both in terms of upfront and operating costs. A MedCenter costs around $100,000 while the upfront cost of a traditional pharmacy is around $1,500,000 (HealthcareConference, 2017). Similarly, its annual operating costs are less than $35,000, whereas a conventional pharmacy incurs annual operating costs of at least $100,000. As such, a MedCenter covers its fixed and variable costs with less than 25 dispenses per day (HealthcareConference, 2017). Although MedAvail's dispensing system is cost-effective, it faces inventory challenges at some locations due to its storage capacity. The existing kiosk, developed to complement pharmacy operations, may store up to 1000 packages. This research was conducted in collaboration with MedAvail to optimize kiosk capacity and drug assortment and achieve target service levels. MedAvail provided pharmacy store and MedCenter sales transaction data for the year 2015. As a first step, we performed descriptive

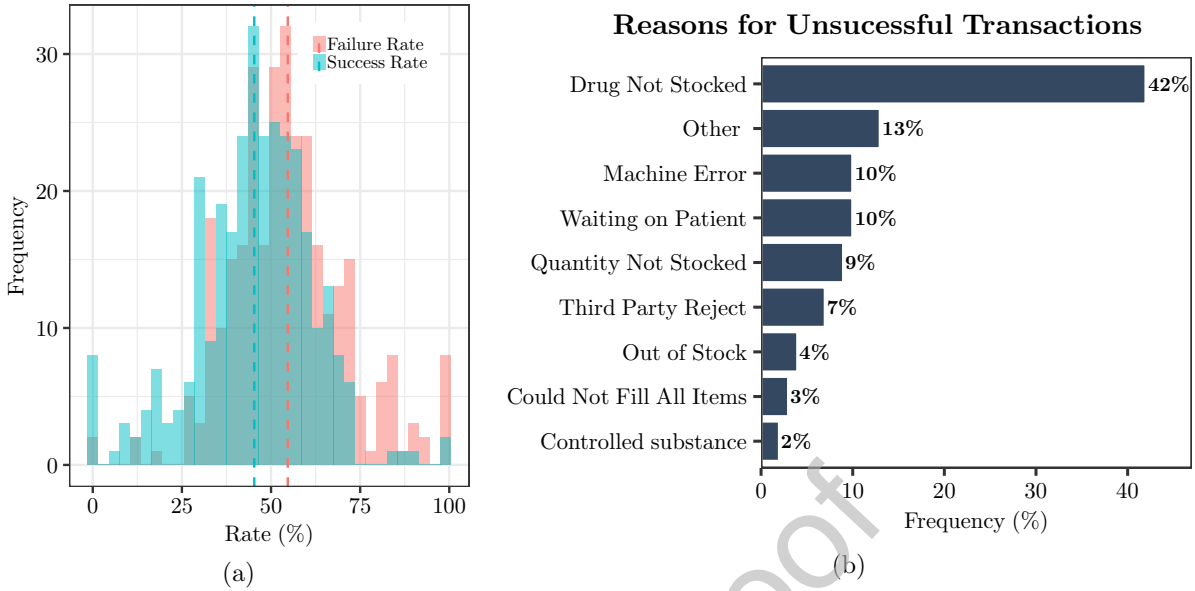


Figure 2: The graphs depict the performance of the existing kiosk in meeting customer requests. Plot (a) illustrates daily success and failure rate distributions. Plot (b) summarizes the main reasons for failed transactions.

analysis to understand demand characteristics at pharmacy stores. Motivated by the findings of the descriptive and predictive analyses, we developed optimization models that use empirical demand distributions to make assortment and substitution decisions and determine optimal kiosk capacity while ensuring that the desired service levels are met.

MedCenters are often located at existing pharmacy stores and provide 24 hour pharmacy access. During working hours, a customer may either buy drugs at the counter or use the kiosk. MedCenters dispense pre-packaged prescription drugs and OTC medications other than controlled-substance drugs, refrigerated drugs, and drugs that need re-pouring. Government regulations prohibit stocking of controlled-substance drugs, but these regulations may be relaxed in the future. Since controlled-substance and refrigerated drugs constitute 17% of total pharmacy sales, MedCenter is developing a new kiosk with refrigeration system and higher capacity to minimize missed opportunities (failed transactions). Figure 2(a) illustrates the distribution of the daily success and failure rates at MedCenters in 2015 where 45% of the transactions are successful, and Figure 2(b) summarizes the factors leading to failed transactions where stocking issues contributed to 60% of the failures.

In the MedCenter, a package is an SKU containing a specific drug of a specific quantity. Since there are thousands of drugs, each ordered in various quantities, inventory decisions are crucial in achieving high service levels when capacity is limited. Past data shows that 60% of the failed transactions occur for three main reasons: (1) the drug is not stocked, (2) the drug is stocked but is currently out of stock, and (3) the drug is stocked but a package with the exact requested quantity is not available. The latter could be partially addressed through *supplier-driven* substitution where the demand for higher quantities could be met by dispensing multiple packages of lower quantity

at the discretion of supplier (pharmacist).

To determine the optimal capacity of a kiosk, the following key questions must be addressed:

1. Which drug should be stocked?
2. In what quantities should each drug be stocked in a package?
3. Which unstocked quantity should be substituted by stocked quantities?
4. What is the stock level of each drug-quantity during the replenishment lead time?
5. What should the replenishment lead time be?

Questions 1, 2, and 3 relate to assortment planning while Questions 3 and 4 relate to inventory planning. All these questions should be addressed simultaneously when deciding on kiosk capacity where the goal is to maximize service level. To make these decisions optimally, one must consider drug demand distributions, seasonal variations, substitution, and co-ordering of the drugs in prescriptions.

The remainder of the paper is organized as follows. In Section 2, an extensive data analysis over pharmacy sales data is carried out. Section 3 reviews the related work in the literature. In particular, we review previous work on newsvendor problem and assortment problem under one-way substitution. Optimization models are formally defined in Section 4. In Section 5, we present a column-generation based heuristic approach to solve large-scale instances. In Section 6, we present model results for the capacity planning problem faced by MedAvail and analyze the effects of supplier-driven drug substitution and replenishment lead time on kiosk capacity. To further generalize model results, the product substitution is studied using randomly generated data and managerial insights are derived. We also compare computational performance of the proposed column generation approach with CPLEX and Benders decomposition. Finally, some concluding remarks and future research directions are presented in Section 7.

2. Analytics of demand

In the US, each drug is assigned a unique 11-digit 3-segment numeric identifier called "National Drug Code (NDC)", denoting manufacturer code, product code, and the package code. Drugs are also assigned a 14-digit hierarchical classification scheme called "Generic Product Identifier (GPI)" that classifies drugs based on their therapeutic use, dosage form, and strength regardless of the manufacturer or package size. Drugs with same ingredients, dosage form, and strength but different manufacturers or package sizes share the same GPI code. Since MedCenter stores a specific quantity of the drug in a standardized package, the manufacturer's package size is irrelevant in this context. Similarly, drugs with the same formula, dosage form, and strength but different manufacturers are pharmaceutically equivalent. It is therefore decided in consultation with the management to consider GPI as a distinct drug identifier.

Analysis of pharmacy sales data shows that most of the GPIs are requested in multiple quantities (QTY). Figure 3 illustrates the distribution of GPIs' distinct quantities requested in the year 2015 over all stores. On average, each GPI is requested in four distinct quantities, while 46% of the GPIs

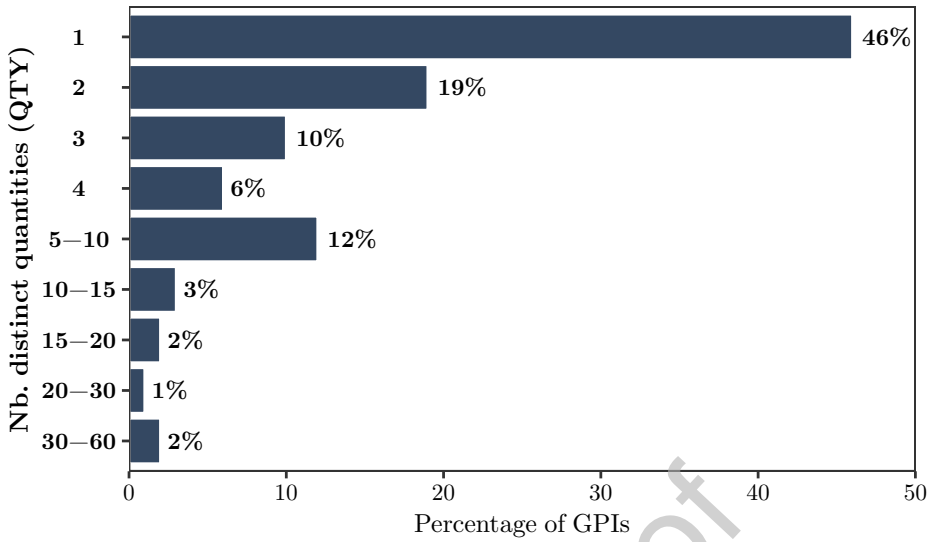


Figure 3: The graph depicts distribution of the GPIs' distinct quantities requested in the year 2015.

are ordered in a single quantity. From a kiosk perspective, a success requires the right drug with the right quantity to be in stock when ordered, so we use GPI-QTY to denote a distinct SKU in the rest of the analysis. We now analyze the significance of product substitution, demand distribution, and co-ordering of drugs using historical data and identify the critical factors to be modelled.

2.1. Product Substitution

Since GPIs are ordered in various quantities, multiple packages of the same GPI with different quantities may need to be stored resulting in higher capacity requirements. One possible solution is to allow supplier-driven substitution between SKUs that share the same GPI code but have a different quantity. We explain the supplier-driven substitution effect using an illustrative example. Consider a GPI that is ordered in five different quantities: $\{20, 28, 40, 56, 60\}$. We may either stock five distinct packages, one of each quantity 20, 28, 40, 56, and 60 or, we may store only packages of 20 and 28 since 40 and 60 are multiples of 20 and 56 is a multiple of 28. As such, GPI-20 may substitute GPI-40 and GPI-60 while GPI-28 may substitute GPI-56. Optimal substitution decisions, however, depend on the demand for each quantity. For instance, if GPI-60 is frequently ordered, we should store it in quantities of 60 rather than 20, which would otherwise result in increased number of packages. On the other hand, when GPI is rarely ordered in quantities of 60, it may be better to stock packages in quantities of 20 to satisfy sales in quantities of 20 and 60. We therefore incorporate the supplier-driven substitution effect in our modelling approach.

Another categorization of substitution is *customer-driven substitution* where customers decide on substitution when their preferred product is not available. For instance, if a customer wanted to buy his/her favorite brand pain reliever that is not available at the pharmacy store, he/she may switch to another pain reliever. However, the data reveals that over the counter drugs constitute only 2.5% of the total pharmacy sales. At pharmacy stores, customer orders predominantly consist

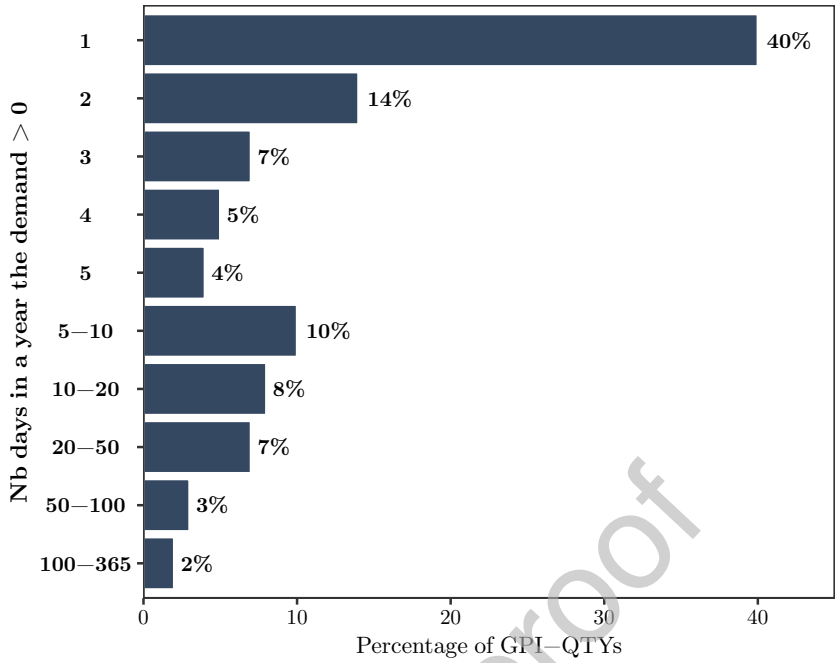


Figure 4: Distribution of the number of days drugs are ordered in a year

of prescribed drugs (97.5% of sales) which cannot be substituted by other drugs at the request of the customer. As such, we do not incorporate customer behavior in our modelling approach.

2.2. Demand distribution

We attempt to determine if demand follows a distribution that could be used in the modelling approach to make stocking and supplier-driven substitution decisions. Pharmacy sales data reveals that demand for the majority of drugs is low as shown in Figure 4. The latter illustrates the distribution of the number of days in a year GPI-QTYs are ordered where 40% of the GPI-QTYs appeared only one day and on average, the number of days GPI-QTYs are requested equals 11. Only 20% of the GPI-QTYs are requested in 10 days or more per year. Figure 5 plots the cumulative demand distribution and yearly demand of the GPI-QTYs. The top 14% (1404) of the GPI-QTYs capture 80% of the pharmacy sales. So to achieve a service level of 80%, it is sufficient to stock the top 14% of drugs. However, at higher service levels, the assortment problem is nontrivial as another 3126 drugs numbered from 1404 to 4530 in Figure 5 represent (31% of drugs) and capture only 15% of the sales. These drugs have yearly demand between 3 and 17 with no particular seasonal trends or patterns throughtout the year. As MedAvail's target service level exceeds 90%, the large number of drugs with low and erratic demand must be considered in making the assortment decisions. Moreover, supplier-driven substitution is expected to have a significant impact on overall stock levels and required kiosk capacity. Due to such random and low demand, fitting theoretical distributions such as Normal and Poisson suffer from over or underestimation of the lead time demand leading to sub-optimal stocking decisions and consequently erroneous service levels. This

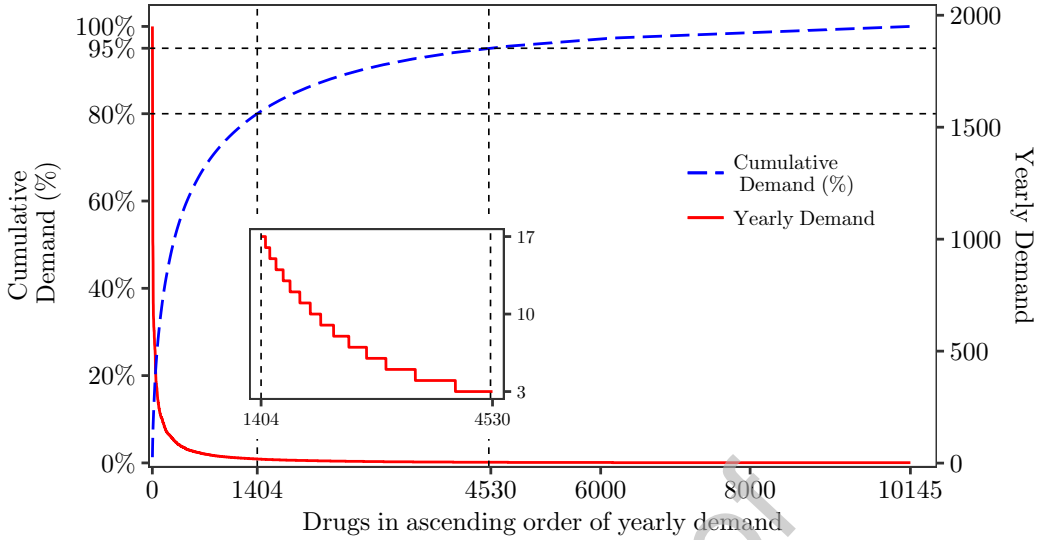


Figure 5: Cumulative demand distribution

motivates the use of empirical distributions of demand in our modelling framework.

2.3. Co-ordering of Drugs

While making stocking decisions, one must consider the possibility of co-ordering of drugs in a prescription. For prescriptions with multiple medications, a customer transaction is less likely to be successful if one of the prescribed drugs is not stocked. Figure 6(a) shows the co-ordering distribution of the transactions recorded in the year 2015 where 82% of the transactions record only one drug, and the average number of drugs in a transaction equals 1.25.

We use the Apriori association rule algorithm (Agrawal et al., 1994) to determine SKUs that frequently appear together in prescriptions. It proceeds by first identifying drugsets that frequently occur in the transactions. A drugset is a set containing one or more drugs. Frequent drugsets are determined using a minimum threshold known as *threshold support*. Support, $supp(X)$ of a drugset X is calculated as the number of times the drugset appears over the total number of transactions in the year 2015. If the support of a drugset is less than the threshold support, it is excluded from further analysis. We set threshold support to be $\frac{15}{D}$, where D is the number of transactions recorded in the year 2015. Once the frequently ordered drugsets are selected based on threshold support, the confidence for all pairs of drugsets is computed. The confidence, $conf\{X \Rightarrow Y\}$ is the probability of purchasing drugset Y when drugset X is purchased. In our case, we select a minimum confidence of 0.5. The results of the algorithm are presented in Figure 6(b) where $lift(X \Rightarrow Y) = \frac{conf(X \Rightarrow Y)}{supp(Y)}$ measures the significance of a rule. A total of 47 association rules between different drugs are found. For better decision making, these association rules should be taken into account when making assortment and stocking decisions. In our modelling approach, we do not explicitly incorporate the effect of association between drugs. We detail the justification in the Section 6 where we note that at higher service levels, i.e., greater than or equal to 80%, all

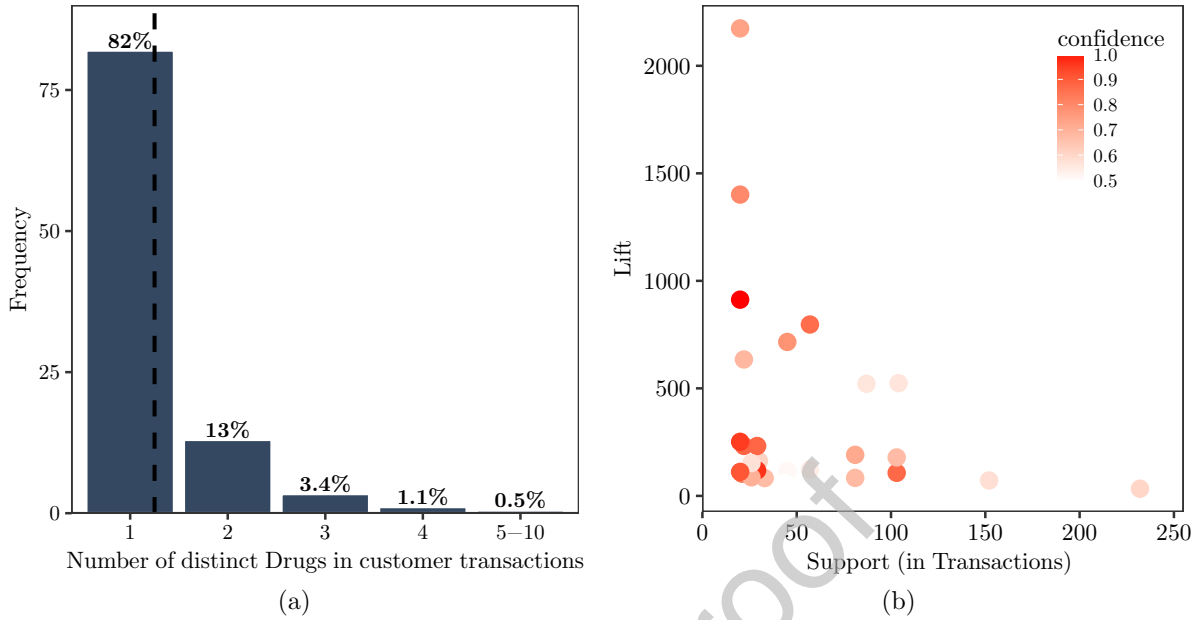


Figure 6: Figure (a) shows the drug co-ordering distribution and Figure (b) compares significance of association rules generated from Apriori association algorithm where threshold support is set to 15 and minimum confidence is 0.5

SKUs with yearly demand greater than or equal to 15 are selected. As such, all drugsets in 47 association rules found are already stocked.

3. Literature Review

Our goal is to develop a modelling framework that determines the required capacity of the kiosk to achieve a desired level of service. Capacity is defined as the total number of packages stored which equals the sum of the stock levels of all GPI-QTYs and therefore depends on the assortment of drugs to be stocked and corresponding stock levels. The stock level of a GPI-QTY is determined by its own demand and the demand of other GPI-QTYs it substitutes as well as the replenishment policy and the target service level. We develop three stochastic optimization models that decide on optimal assortment, inventory, and supplier-driven substitution decisions. These models share similarities to the newsvendor problem and the assortment problem under one-way substitution. In this section, we review previous work on these two problems and position our work accordingly.

3.1. Assortment & Inventory Decisions

MedAvail wants to determine optimal stock levels for GPI-QTYs with stochastic demand to minimize kiosk storage capacity while ensuring that desired service level is achieved. This problem is related to the well-known *Constrained Multi-Product Newsvendor Problem (CMPNP)* where a newsvendor wants to determine single-period optimal stocking policy for multiple products with stochastic demand and resource or budget constraint(s). The literature that deals with stochastic modelling approaches for the newsvendor problem assumes that the demand distribution is known.

In this stream, Hadley and Whitin (1963) are the first to study a CMPNP and propose a Lagrangian-based method to the solve the problem. Fractional stock levels are allowed and to obtain an integer solution, the optimal order quantity is approximated by rounding down to the nearest integer value. Such an approach, however, performs poorly when the demand for products is low. To overcome the issue, Hadley and Whitin (1963) propose a dynamic programming procedure which is computationally inefficient when the products size is large and the largest instance reported in the paper consists of three products only. Nahmias and Schmidt (1984) extends the work of Hadley and Whitin (1963) and propose multiple heuristic approaches to solve the problem efficiently. The approach is however only applicable for moderate-to-high demand items as the proposed solution methodologies use continuous decision variables. The authors argue that for low demand items a discrete model would be more applicable. Lau and Lau (1996) observe that the methodology proposed by (Hadley and Whitin, 1963) may lead to negative optimal order quantities when the capacity is tight. The authors present an extension of the procedure in (Hadley and Whitin, 1963) to deal with general demand distributions including positive lower bounds. Abdel-Malek et al. (2004) propose a closed form expression of optimal order quantities when the demand follows a uniform distribution and present a generic iterative method to find near optimal solutions for other general distributions. To avoid the issue of negative order quantities, Abdel-Malek and Montanari (2005) suggest the use of thresholds to help decision maker remove products with low marginal utilities. A binary search method applicable to both continuous and discrete demand distribution is proposed by Zhang et al. (2009). The proposed solution approach, however, does not guarantee optimality for the discrete distribution. For a comprehensive review on uncapacitated and single newsvendor problems with known demand distribution, we refer the reader to Turken et al. (2012).

In stochastic models, the literature assumes a known distribution and could not be applied in our case where demand is highly erratic and low. Since the demand for each GPI-QTY is erratic and low, experimentation with fitting Negative binomial, Poisson, and Normal distributions reveal that demand does not follow any specific probability distribution. This is true for many real-life problems where the exact distribution is rarely known and is generally approximated based on historical data. This explains the issue of poor out-of-sample performance in stochastic optimization approaches. To address this, robust optimization approaches are proposed in the literature. In this stream, Vairaktarakis (2000) considers a robust CMPNP under the assumption that the demand distribution for each item is completely unknown and only a set of discrete demand scenarios are available. The author presents minmax regret formulations with the objective to minimize expected costs under the worst-case realization of demand. Scenario-based minmax modelling approach is often criticized for being overly conservative as outliers in the historical data are not excluded. Such a minmax approach could be used for the pharmacy kiosk problem but our results show that it performs poorly. The poor performance is not due to overly conservative nature of the model but rather it is unable to provide robust solutions due to fewer number of scenarios for the kiosk problem with thousands of SKUs. To deal with the issue of overly conservative solutions in minmax regret formulations, a standard approach is to assume that the demand for each item

could deviate from its nominal demand while the total deviation for all items is control by a user-defined budget of uncertainty (see for example, Bertsimas and Thiele (2006), Lin and Ng (2011)). However, under service level maximization objective, the adversarial problem in robust optimization is nonlinear and as such, tractable robust counterpart formulation does not exists. In addition, mathematical formulations for such models are complex and difficult to understand for the managers. We therefore adopt a scenario-based stochastic optimization framework where all values of demand for each GPI-QTY recorded in the past data are used. Such an approach does not require the probability associated with each scenario and is therefore appropriate in our case where the probability density functions of GPI-QTYs are not known. In order to obtain robust solutions, we generate robust scenarios using the maximum demand of each GPI-QTY over all stores data in a given time period.

In CMPNP literature, the objectives considered optimize costs, profits, or the probability to achieve a target profit under different criteria (Khouja, 1999). Our objective is to determine minimum kiosk capacity under service level constraints. The modelling approaches in the CMPNP literature do not explicitly model service level constraints and under stocking is penalized through shortage costs that are included in the objective function. Studies that do consider service levels (see Table 1) in CMPNP (Chen and Chuang, 2000, Taleizadeh et al., 2008, 2009, Waring, 2012, Abdel-Aal et al., 2017) include service level constraints for each item and use a well-defined cumulative distribution function of the demand to define the service level as the probability of meeting demand with a given stock level. However, such an approach is not applicable in our case since demand is erratic and low and does not follow a known distribution. We use fill rate to define service level as the proportion of successful transactions with given stock levels of GPI-QTYs over a planning horizon of one year. Moreover, the service level in our problem is defined for the kiosk rather than for each GPI-QTY.

3.2. Substitution Decisions

Another challenge is to make substitution decisions along with stocking decisions under stochastic demand. Product substitution in general is defined as the act of using one product to meet the demand of another product. In inventory and assortment planning literature, substitution is categorized as either *supplier-driven* or *customer-driven* (Shin et al., 2015). In customer-driven substitution, customers decide on substitution when their preferred product is not available. In such problems, customer behavior is modelled within the optimization framework, see for example, (Gaur and Honhon, 2006, Kök and Fisher, 2007, Aydin and Porteus, 2008). In this stream of literature, Gaur and Honhon (2006) consider an uncapacitated multiproduct assortment planning problem where the demand follows a known distribution and the goal is to decide on the stock level for each product such that the expected profits are maximized. A utility-based locational choice model is used to estimate the customer demand where substitution between the products is allowed based on the substitution rate. For each customer, the utility it derives from product j is calculated and it is assumed that a customer prefers the product that maximizes his/her utility. If such a product is not available, he/she may select the second highest utility product with a

Table 1: Literature on Newsvendor Problem

Paper	Multi Product	Capacitated	Uncertainty(1)	Distribution(2)	Service level	Objective(3)	Variable type(4)	Methodology(5)	Problem size
Our work	✓	✓	**S**	**E**	✓	**Max CP/S**	**D**	**MILP**	**30000**
Hadley and Whitin (1963)	✓	✓	S	K		Max P	C	L+DP	3
Nahmias and Schmidt (1984)	✓	✓	S	K		Max P	C	L+H	5000
Aardal et al. (1989)			S	K	✓	Min C	C	CF	***
Moon and Choi (1994)			S	F	✓	Min C	C	L+IA	***
Lau and Lau (1996)	✓	✓	S	K		Max P	C	L	1000
Erlebacher (2000)	✓	✓	S	K		Max P	C	CF	***
Vairaktarakis (2000)	✓	✓	S	E		Minmax R	D	DP	***
Chen and Chuang (2000)			S	K	✓	Min C	C	CF	***
Abdel-Malek et al. (2004)	✓	✓	S	K		Min C	C	GIM	6
Bertsimas and Thiele (2006)		✓	R	I		Max P	D	MILP	1
Taleizadeh et al. (2008)	✓	✓	S	K	✓	Min C	D	GA	15
Zhang et al. (2009)	✓	✓	S	K		Max P	C/D	BSM	6
Taleizadeh et al. (2009)	✓	✓	S	K	✓	Max P+S	D	GA	15
Choi et al. (2011)	✓		S	K		Max PR	C	CF	10
Lin and Ng (2011)	✓	✓	R	I		Minmax R	C	L	50
Waring (2012)			S	K	✓	Max P	C	L	1
Jammernegg and Kischka (2013)			S	K	✓	Max MDR	C/D	CF	1

Acronyms:
(1) S - stochastic, R - Robust; (2) K - known, U - unknown, E - Empirical, F - distribution free, I - interval data
(3) P - profit, C - Cost, R - Regret, PVDI - Value of perfect distribution information, MDR - mean deviation rule, PR - profits under risk, S - Service level, CP-Capacity
(4) C - continuous, D - Discrete
(5) L - lagrangian, H - heuristc, MILP - mixed integer linear programming, CF - closed form, DP - Dyanmic programming, GA - Genetic algorithm
(5) GIM - Generic iterative method, BSM - binary search method, IA - iterative algorithm

Table 2: Literature on Assortment Problem under one-way supplier-driven substitution

Paper	Multi Product	Capacitated	Uncertainty(1)	Distribution(2)	Service level	Objective(3)	Variable type(4)	Methodology(5)	Problem size
Our work	✓	✓	**S**	**E**	✓	**Max CP/S**	**D**	**MILP**	**30000**
Sadowski (1959)			S	K		Min L	C/D	DP	***
Pentico (1974)			S	K		Min C	D	DP	***
Pentico (1976)			D	K		Min C	D	MILP	***
Tryfos (1985)			S	K		Max P	C	CF	10
Leachman and Glassey (1987)	✓	✓	S	K		Min C	C	***	***
Bagchi and Gutierrez (1992)	✓		S	K	✓	Max P	C	CF	3
Wollmer (1992)	✓	✓	S	K		Max R	C	H	5
Chand et al. (1994)			D	K		Min C	D	MINLP + DP	***
Bassok et al. (1999)	✓		S	K		Max P	D	IA	2
Rajaram and Tang (2001)	✓		S	K		Max P	C	H	&
Rao et al. (2004)	✓		S	K		Max P	D	MILP+H	25
Dutta and Chakraborty (2010)	✓		F	U		Max P	C	NSP	2
Deflem and Van Nieuwenhuyse (2013)	✓		S	K		Min C	C	CF	2
Ahiska et al. (2017)	✓		S	K		Max P	C	LSA	3
Hsieh and Lai (2019)	✓		D	K		Max P	C	GTM	2

Acronyms:
(1) D - deterministic, S - stochastic, F - fuzzy (2) K - known, E - Empirical, U - unknown
(3) P - profits, C - Costs, L - loss, R - Revenue, CP - Capacity, S - Service level (4) C - continuous, D - Discrete
(5) H - heuristc, MILP - mixed integer linear programming, CF - closed form, DP - Dyanmic programming, IA - Iterative Algorithm
(5) MINLP - mixed integer nonlinear programming, NSP - Numerical search procedure, LSA - Local search algorithm, GTM - Game-theoretical model

probability defined by substitution rates. Kök and Fisher (2007) model the assortment problem using an exogenous demand model where the demand and substitution rates are precomputed using regression models, and are then used to decide on the number of facings allocated to each product under a capacity constraint.

These models do not make substitution decisions but rather consider customers' substitution behavior to decide on assortment and stock levels. We do not incorporate such customer behavior in our modelling approach since customer orders at pharmacy stores predominantly consist of prescribed drugs (97.5% of sales) which cannot be substituted by other drugs at the request of the customer. However, incorporating customer substitution behavior within a model making supplier-driven substitution decisions is a promising future research work. We refer the reader to Kök et al. (2015) and Shin et al. (2015) for a comprehensive review of literature on customer-driven substitution. From here onward, term "substitution" refers to supplier-driven substitution unless explicitly mentioned otherwise.

At a pharmacy kiosk, a pharmacist may dispense multiple packages of one GPI-QTY to satisfy the demand of another GPI-QTY as long as they share the same GPI code, and the quantities match. This is known as supplier-driven substitution where the supplier makes stocking decisions while taking into account product substitution (Shin et al., 2015). More specifically, such quantity based substitution is referred to in the literature as *one-way substitution* and is common in manufacturing and service industries such as semiconductor industry (Bassok et al., 1999), computer hardware industry (Leachman and Glassey, 1987), and airline industry (Wollmer, 1992). One-way substitution may improve overall service level due to pooling. Potential benefits of one-way substitution in inventory management are detailed in Fuller et al. (1993).

The term assortment problem was first introduced by Sadowski (1959) who considers a problem of determining n steel beams of different strengths where the demand of a lesser strength beam is substitutable by a beam with greater strength. A similar problem in apparel industry in considered by Tryfos (1985) where the manufacturer has to decide on the set of m sizes. In these two papers, demand patterns are described by continuous distributions. The modelling approach in these works only decides on whether a quantity is stocked or not. On the other hand, Pentico (1974) considers a single product ordered in different quantities following a discrete probability distribution. The goal is to decide on the stock levels for each size while taking into account one-way substitution where a smaller stocked size can meet the demand of a larger unstocked size while incurring a substitution cost. The demand for each size is assumed to be probabilistic and some strong substitution assumptions are made in the paper. The author assumes that to meet the demand of a larger stocked size, only the smallest stocked size could be used. It is also assumed that demand is realized in descending order of size. Moreover, capacity is incorporated implicitly as a fixed charge cost of stocking a given size. A dynamic programming approach is proposed to formulate and solve the problem. These assumptions greatly limit the applicability of the proposed model. Pentico (1976) relaxes the linear cost functions and substitution cost assumption in Pentico (1974) but considers deterministic demand. Chand et al. (1994) generalizes the problem in Pentico (1976) with infinite

planning horizon. A different variant of demand uncertainty in the assortment problem is studied by Dutta and Chakraborty (2010) where the demand is fuzzy and lies within an interval data. Bassok et al. (1999), Rao et al. (2004), and Deflem and Van Nieuwenhuyse (2013) study multi-product assortment problem under downward substitution without incorporating storage or resource constraints. Bassok et al. (1999) present a two-stage profit maximization formulation with N products and N demand classes under full downward substitution. Rao et al. (2004) consider a similar problem but take into account setup costs while Deflem and Van Nieuwenhuyse (2013) derive optimality conditions where substitution outperforms separate stock levels for the two-item case. Ahiska et al. (2017) and Hsieh and Lai (2019) study one-way substitution for manufacturing industry problem where high quality products substitute low quality ones. Ahiska et al. (2017) formulate the problem using Markov decision process while Hsieh and Lai (2019) use a game-theoretical modelling framework.

Pharmacy kiosk assortment problem poses new research questions within the assortment optimization literature that have not been studied before. As such, our work differs from existing literature in the following aspects.

1. Substitution rules considered in our work have not been studied before. The literature on supplier driven substitution deals with problems where a high-quality product may substitute a lower quality one with one to one substitution i.e., to meet the demand of a single unstocked unit, only one unit of higher quality item is dispatched. On the other hand, in our case, to meet the demand of a single unit, multiple packages must be dispensed to fulfill the demand while ensuring that the quantity dispensed is equal to the requested quantity. Such requirements are not handled by the models in the literature. From a modelling perspective, the exact requested quantity requirement leads to extremely complex mathematical models.
2. The models in the literature explicitly include substitution costs in the objective function. For instance, in a computer hardware industry if a customer order of 4GB memory chip is not available, an 8GB memory chip may fulfill the demand with substitution cost equals to the difference between the prices of the two different memory chips. In our case, there are no explicit substitution costs. The latter are captured implicitly within the service level expression to avoid over-substitution that may lead to lower service levels. To the best of our knowledge, our work is the first to consider fill rate in assortment planning problems with one-way substitution. As shown in Table 2, other than (Bagchi and Gutierrez, 1992), no work considers service level. In Bagchi and Gutierrez (1992), however, service level constraints are added for each item using a well-defined cumulative distribution function and fractional stock levels are also allowed in the optimal solution.
3. A common assumption in assortment planning problems under one-way substitution is that demand for all items is realized at the same time. The problem is then formulated as a two-stage stochastic program. In the first stage, when demand is not realized, the formulation decides on stock levels of each item while taking into account substitution. In the second stage when the demand is realized for all items, substitution decisions are made based on the

given stock levels to meet the demand for all items. However, for a pharmacy kiosk, demand is realized in a dynamic fashion where customers arrive one at a time. Rao et al. (2004) correctly point out that dynamic substitution models are extremely complex. Such complex models are intractable for the large-scale capacity planning problem faced by MedAvail with around 30,000 GPI-QTYs. We therefore employ a stationary substitution policy i.e., same substitution rules are employed throughout the planning horizon irrespective of the stock levels at any given time. However, to deal with the problem of dynamic customer arrivals, our models make robust substitution decisions which guarantee that the desired service level is always achieved irrespective of the sequence of demand realization for substitutable products.

4. The proposed models in this paper are computationally tractable for the pharmacy kiosk problem with 30,000 GPI-QTYs and could be solved using a commercial solver. Other models in the literature are too complex for the large-scale instances with thousands of GPI-QTYs.

4. Modelling Stocking and Assortment Decisions

The problem is to decide on the single period (replenishment lead time) stock level x_i, for each product $i \in I$ using the empirical distribution that is generated from historical data. When $x_i = 0$, product i is not stocked and the assortment is defined by $i \in I$ such that $x_i > 0$. We adopt a scenario-based stochastic optimization model that uses past data to generate T demand scenarios by dividing the planning horizon into $T = \lceil \frac{365}{h} \rceil$ lead time intervals, where h is the lead time. The demand A_{it} for $i \in I$, during time period $t \in \Theta = \{1, ..., T\}$, is calculated using historical sales data. Products are grouped in classes if they only differ by quantities. In the presence of substitution, the demand d_{it}, depends on the substitution variable s_{ij}, which equals 1 if product i substitutes product j. The latter is only possible if products i and j belong to the same product class and quantity q_j is a multiple of quantity q_i. A $0-1$ incidence matrix $\mathbf{b} = [b_{ij}]$ is computed where $b_{ij} = 1$ if product j is substitutable by i. As such, $d_{it} = \sum_{\substack{j \in I: \\ b_{ij}=1}} m_{ij} A_{jt} s_{ij}$, where $m_{ij} = \frac{q_j}{q_i}$ units of product i are required to meet the unit demand for product j. Substitution variables may be either predetermined or optimized within a mathematical model. In pharmacy kiosk application, each GPI is a product class containing GPI-QTYs sharing the same GPI code. Multiple packages have to be dispensed to meet the demand of a higher quantity. Such a substitution arises for a variety of other industry applications where a requested quantity could be substituted by multiple packages of smaller quantities. For instance, in case of a Bank ATM, customer request for $100 could be met by dispensing five currency notes of $20. Similarly, for grocery store/vending machine, a customer may be willing to accept six 250ml bottles of Coke if 1.5 liter family pack is not available.

Since there are no backorders, any unsatisfied demand is a lost sale. The lost sales for a product $i \in I$ during time period $t \in \Theta$ is $\max\{0, d_{it} - x_i\}$. Lost sales occur either because the drug is not stocked, i.e., $x_i = 0$, or observed demand exceeds the stock level, i.e., $x_i < d_i$. At a pharmacy kiosk, unsatisfied demand is lost because a customer is most likely going to use another pharmacy and not wait for the medication to be back-ordered. The same applies for other kiosk applications such

as Bank ATM and vending machines, etc. Since unsatisfied demand is lost, we model the problem with no backorders which also justifies single period stock planning. The expected service level or fill rate is calculated as $1 - \frac{\sum_{i\in I}\sum_{t\in\Theta}\max\{0, d_{it} - x_i\}}{D}$ where D is the total yearly demand. Our goal is to determine the capacity such that a desired service level α is met.

We develop three optimization models to solve the capacity planning problem and address management's questions under three different substitution rules: (1) no substitution, i.e., $s_{ii} = 1$ and all other substitution variables take value 0, (2) management's substitution rule, (3) optimized substitution. In rules (1) and (2), substitution is predefined. We now discuss the models under predefined and optimized substitution.

4.1. Predefined substitution

The first model [M1] decides only on optimal stock levels for products using one of the predefined substitution rules, and minimizes the capacity under service level constraint. Given the substitution rule, demand scenarios d_{it} for each product $i \in I$ are precomputed and serve as input data to the model. The formulation is

$$\text{[M1]:} \quad \min \sum_{i\in I} x_i \tag{4.1.1}$$

$$\text{s.t.} \quad 1 - \frac{\sum_{i\in I}\sum_{t\in\Theta}\max\{0, d_{it} - x_i\}}{D} \geq \alpha \tag{4.1.2}$$

$$x_i \in \mathbb{Z}^+, \qquad \forall \quad i \in I, t \in \Theta, \tag{4.1.3}$$

where the objective function (4.1.1) minimizes the total number of packages stocked i.e., required capacity of the kiosk. Constraint (4.1.2) ensures that the expected service level, $1 - \frac{\sum_{i\in I}\sum_{t\in\Theta}\max\{0, d_{it} - x_i\}}{D}$, is greater than or equal to the desired service level, α. Finally, constraint (4.1.3) is the nonnegative integer requirement on x_i. The above formulation is nonlinear due to max functions in constraint (4.1.2). The latter may be linearized by introducing auxiliary variables (f_{it}, y_{it}) and replacing constraint (4.1.2) with the following set of constraints

$$1 - \frac{\sum_{i\in I}\sum_{t\in\Theta} f_{it}}{D} \geq \alpha, \tag{4.1.4}$$

$$f_{it} \geq 0, \qquad \forall \quad i \in I, t \in \Theta, \tag{4.1.5}$$

$$f_{it} \geq d_{it} - x_i, \qquad \forall \quad i \in I, t \in \Theta, \tag{4.1.6}$$

$$f_{it} \leq (d_{it} - x_i) + M \times y_{it}, \quad \forall \quad i \in I, t \in \Theta, \tag{4.1.7}$$

$$f_{it} \leq 0 + M \times (1 - y_{it}), \quad \forall \quad i \in I, t \in \Theta, \tag{4.1.8}$$

$$f_{it} \geq 0, \ y_{it} \in \{0, 1\}, \qquad \forall \quad i \in I, t \in \Theta, \tag{4.1.9}$$

where M is a significantly large number. If $d_{it} > x_i$, y_{it} must be equal to 0 for the problem to be feasible. Constraint (4.1.6) is then $f_{it} \leq d_{it} - x_i$ and constraint (4.1.7) is $f_{it} \leq M$. As such, $f_{it} = d_{it} - x_i$. On the other hand, if $d_{it} < x_i$, $y_{it} = 1$ for the problem to be feasible and $f_{it} = 0$. The problem, however, becomes challenging to solve due to binary variables y_{it}. We therefore present a relaxed formulation [R1] where constraints (4.1.7) and (4.1.8) are dropped

$$\text{[R1]:} \quad \min \sum_{i \in I} x_i \tag{4.1.10}$$

$$\text{s.t.} \quad f_{it} \geq d_{it} - x_i \qquad \forall \quad i \in I, t \in \Theta, \tag{4.1.11}$$

$$1 - \frac{\sum_{i \in I} \sum_{t \in \Theta} f_{it}}{D} \geq \alpha, \tag{4.1.12}$$

$$x_i \in \mathbb{Z}^+, f_{it} \geq 0, \quad \forall \quad i \in I, t \in \Theta, \tag{4.1.13}$$

and prove in Lemma 1 that its optimal solution $\mathbf{x}^* = [x_i^*]$ is also optimal to the original model [M1].

Lemma 1. *An optimal solution $\mathbf{x}^*$ for model [R1] is also optimal to the original model [M1].*

Proof. Let $(\mathbf{x}^* = [x_{ij}^*], \mathbf{f}^* = [f_{it}^*])$ be an optimal solution to model [R1]. Rearranging constraint (4.1.12),

$$\sum_{i \in I} \sum_{t \in \Theta} f_{it}^* \leq (1 - \alpha) \times D$$

In model [R1], f_{it}^* may take a value greater than the max term $\max\{0, d_{it} - x_i^*\}$ in constraint (4.1.2). As such, $\sum_{i \in I} \sum_{t \in \Theta} f_{it}^* \geq \sum_{i \in I} \sum_{t \in \Theta} \max\{0, d_{it} - x_i^*\}$ and

$$(1 - \alpha) \times D \geq \sum_{i \in I} \sum_{t \in \Theta} f_{it}^* \geq \sum_{i \in I} \sum_{t \in \Theta} \max\{0, d_{it} - x_i\}$$

This implies $\sum_{i \in I} \sum_{t \in \Theta} \max\{0, d_{it} - x_i^*\} \leq (1 - \alpha) \times D$ and constraint (4.1.2) holds for $\mathbf{x}^*$. This proves that solution $\mathbf{x}^*$ is feasible to the original model [M1].

Let z_{M1}^* and z_{R1}^* be the optimal objective function values for models [M1] and [R1], respectively. Since [R1] is a relaxed formulation of model [M1], $z_{\text{R1}}^* \leq z_{\text{M1}}^*$. Since the original model [M1] can not have a solution superior than z_{R1}^*, $\mathbf{x}^*$ is also optimal for [M1]. □

Note that f_{it} is simply an analysis variable used to linearize model [M1]. One could adjust its value after solving the model [R1] by setting $f_{it}^* = \max\{0, d_{it} - x_i^*\}$.

Model [R1] is a new variant of the well known single period newsvendor problem under a service level constraint and could be applied to any inventory problem where the service level needs to be

considered while making stocking decisions. In addition to capacity minimization objective, the model is easily extendable for profit maximization or cost minimization objectives.

4.2. Optimized substitution

We develop two additional models that extend [M1] to optimize both stocking and substitution decisions. Model [M2] decides on substitution and stock levels to minimize storage capacity under a service level constraint. The parameter d_{it} in model [M1] is now a decision variable in [M2] as the model makes substitution decision s_{ij}. As such, model [M2] has two additional decision variables : d_{it} and s_{ij}. The formulation is then as follows.

$$\text{[M2]:} \quad \min \sum_{i \in I} x_i \tag{4.2.1}$$

$$\text{s.t.} \quad d_{it} = \sum_{\substack{j \in I: \\ b_{ij}=1}} m_{ij} A_{jt} s_{ij} \qquad \forall i \in I, t \in \Theta, \tag{4.2.2}$$

$$\sum_{\substack{i \in I: \\ b_{ij}=1}} s_{ij} = 1 \qquad \forall j \in I, \tag{4.2.3}$$

$$1 - \frac{\sum_{i \in I} \sum_{t \in \Theta} \max\{0, d_{it} - x_i\}}{D} \geq \alpha, \tag{4.2.4}$$

$$s_{ij} \in \{0, 1\} \qquad \forall i \in I, j \in I, \tag{4.2.5}$$

$$x_i \in \mathbb{Z}^+, \; d_{it} \in \mathbb{Z}^+ \qquad \forall i \in I, t \in \Theta, \tag{4.2.6}$$

where the objective function (4.2.1) is the same as (4.1.1). Constraint (4.2.2) computes demand d_{it} of a product $i \in I$ in period $t \in \Theta$ taking into account the demand of products it substitutes. Constraint (4.2.3) ensures that each product $j \in I$ is substituted by exactly one product. If $s_{ii} = 1$, it implies that product $i \in I$ is not substituted by any other product. Constraint (4.2.5) is the binary requirement on variable s_{ij} and constraints (4.2.6) are nonnegative integer requirements on variables x_i and d_{it}. Constraint (4.2.4) defines the service level and is the same as constraint (4.1.2) in model [M1]. It may be linearized using the same approach discussed earlier for model [M1]. Note that substitution variables s_{ij} only change d_{it} to a decision variable and constraints (4.1.4) - (4.1.9) are valid for model [M2]. As such, the relaxed formulation [R2] for model [M2] is

$$\text{[R2]:} \quad \min \sum_{i \in I} x_i \tag{4.2.7}$$

$$\text{s.t.} \quad (4.2.2), (4.2.3), (4.2.5), (4.2.6)$$

$$f_{it} \geq d_{it} - x_i, \qquad \forall i \in I, t \in \Theta \tag{4.2.8}$$

$$1 - \frac{\sum_{i \in I} \sum_{t \in \Theta} f_{it}}{D} \geq \alpha, \tag{4.2.9}$$

$$f_{it} \geq 0, \qquad \forall i \in I, t \in \Theta. \tag{4.2.10}$$

Lemma 1 holds trivally and an optimal solution $(\mathbf{x}^*, \mathbf{s}^*)$ to model [R2] is also optimal for [M2].

Model [M3] is developed to maximize the expected service level of a kiosk under a capacity constraint. The decision variables are the same as in [M2], and the mathematical formulation is as follows:

$$\text{[M3]:} \quad \max \quad \alpha = 1 - \frac{\sum_{i\in I}\sum_{t\in\Theta} \max\{0, d_{it} - x_i\}}{D} \tag{4.2.11}$$

$$\text{s.t.} \quad (4.2.2), (4.2.3), (4.2.5), (4.2.6),$$

$$\sum_{i\in I} x_i \leq C, \tag{4.2.12}$$

where the objective function (4.2.11) maximizes the expected service level α and constraint (4.2.12) ensures that the total number of packages stored is restricted to capacity, C. As in model [M2], [M3] is also nonlinear due to the max terms in the objective function. However, to linearize it, we only introduce analysis variable f_{it}. The linear formulation is

$$\text{[R3]:} \quad \max \quad \alpha = 1 - \frac{\sum_{i\in I}\sum_{t\in\Theta} f_{it}}{D} \tag{4.2.13}$$

$$\text{s.t.} \quad (4.2.2), (4.2.3), (4.2.5), (4.2.6), (4.2.12),$$

$$f_{it} \geq d_{it} - x_i, \qquad \forall i \in I, t \in \Theta \tag{4.2.14}$$

$$f_{it} \geq 0, \qquad \forall i \in I, t \in \Theta \tag{4.2.15}$$

where constraint (4.2.14) along with nonnegativity constraint (4.2.15) ensure that $f_{it} \geq \max\{0, d_{it} - x_i\}$. At optimality, $f^*_{it} = \max\{0, d^*_{it} - x^*_i\}$ $\forall i \in I, t \in \Theta$ and is proven in Lemma 2.

Lemma 2. *For model [R3], given an optimal solution* $(\mathbf{x}^*, \mathbf{f}^*, \mathbf{s}^*, \mathbf{d}^*)$, $f^*_{it} = \max\{0, d^*_{it} - x^*_i\}$ $\forall i \in I, t \in \Theta$.

Proof. Note that constraints (4.2.14) and (4.2.15) ensure that $f^*_{it} \geq \max\{0, d^*_{it} - x^*_i\}$ $\forall i \in I, t \in \Theta$.

We now prove by contradiction that at optimality, f^*_{it} can not take a value greater than the max term. Assume that $(\mathbf{x}^*, \mathbf{f}^*, \mathbf{s}^*, \mathbf{d}^*)$ is optimal with objective function value z^* and $f^*_{it} > \max\{0, d^*_{it} - x^*_i\}$ $\exists i \in I, t \in \Theta$. Let $(\mathbf{x}^*, \mathbf{f^a}, \mathbf{s}^*, \mathbf{d}^*)$ be the adjusted solution with objective function value z^a where $f^a_{it} = \max\{0, d^*_{it} - x^*_i\}$. As such,

$$\sum_{i\in I}\sum_{t\in T} f^*_{it} > \sum_{i\in I}\sum_{t\in T} f^a_{it} \Longrightarrow \left(1 - \frac{\sum_{i\in I}\sum_{t\in T} f^*_{it}}{D}\right) < \left(1 - \frac{\sum_{i\in I}\sum_{t\in T} f^a_{it}}{D}\right) \Longrightarrow z^* < z^a$$

which contradicts the assumption that z^* is optimal. This proves that if $f^*_{it} > \max\{0, d^*_{it} - x^*_i\}$ $\exists i \in I, t \in \Theta$, there always exists a better solution $f^a_{it} = \max\{0, d^*_{it} - x^*_i\}$ for which $z^a > z^*$. □

Models [R2] and [R3] are extensions of the capacitated newsvendor problem under supplier-

driven substitution. Since 97.5% of customer orders consist of prescribed drugs which cannot be substituted by other drugs at the request of the customer, we only model one-way supplier-driven substitution where a pharmacist may dispense multiple packages of one GPI-QTY to meet the demand of another sharing the same GPI code. As such, the proposed models are specific to supplier driven substitution and do not readily handle customer driven substitution.

Note that the proposed models are generic and apply for any demand values. However, when demand is less sporadic, a single period model that uses moments of the demand distribution may become useful. Under dynamic customer arrivals, our models make robust substitution decisions which guarantee that the desired service level is always achieved irrespective of the sequence of demand realization for substitutable products. This is detailed next.

4.3. Substitution under dynamic customer arrivals

Models [R2] and [R3] make substitution decisions that are robust against the sequence of demand realization for substitutable products. We first explain this using an illustrative example and present a formal proof in Lemma 3. Consider two products i and j, in the same product class and let $q_i = 20$ and $q_j = 60$. Since q_j is a multiple of q_i, assume that product i substitutes j, and $m_{ij} = \frac{60}{20} = 3$. In a given period t, let $A_{it} = 10$, $A_{jt} = 1$, $D = 10 + 1 = 11$, and stock level $x_i = 10$. If product j is requested when less than three packages of product i are available, then the number of failed transactions equals 1 and demand for product i is fully met. As such, service level $\alpha = 1 - \frac{1}{11} = 91\%$. However, if product j is requested when at least three packages of product, i are available, the demand for product j is fulfilled and there is a shortage of three packages to meet the demand for product i. In this case, $\alpha = 1 - \frac{3}{11} = 73\%$. Depending on the sequence of demand realization, the service level either equals 91% or 73%. Proposed mathematical models calculate service level as $\alpha = 1 - \frac{f_{it}}{D} = 1 - \frac{3}{11} = 73\%$ if product i substitutes j. We now show that substitution decisions are robust against the sequence of demand realization.

Lemma 3. *Substitution decisions are robust against the sequence of demand realization for substiutable products and guarantee that desired service level is achieved.*

Proof. We first present the exact formula to compute the number of failures f_{it}^E. Then, we show that $f_{it} \geq f_{it}^E$ for any sequence of demand realization.

Given a solution $\mathbf{s}$, let $K_i = \{1, 2, ..., n-1, n\}$ be the set of products substituted by product $i \in I$ i.e., $s_{ij} = 1 \forall j \in K_i$. Without loss of generality, assume that the set K_i is ordered such that the sequence of demand realization is

$$A_{nt} \rightarrow A_{n-1,t} \rightarrow \cdots \rightarrow A_{2t} \rightarrow A_{1t} \tag{4.3.1}$$

Let f_{it}^j be the number of failures and x_i^j be the number of packages available for product $j \in K_i$. The exact formula for the number of failures is

$$f_{it}^j = \lceil \max\{0, A_{jt} - \frac{x_i^j}{m_{ij}}\} \rceil \tag{4.3.2}$$

where $\frac{x_i^j}{m_{ij}}$ computes the demand that could be met for product j using product i. Note that since $\frac{x_i^j}{m_{ij}}$ can take fractional values, the value $max\{0, A_{jt} - \frac{x_i^j}{m_{ij}}\}$ needs to be rounded up to the nearest integer value. Within an optimization model, one may linearize constraint (4.3.2) as

$$f_{it}^j \geq 0 \tag{4.3.3}$$

$$f_{it}^j \geq A_{jt} - \frac{x_i^j}{m_{ij}} \tag{4.3.4}$$

$$f_{it}^j \in \mathbb{Z} \tag{4.3.5}$$

Constraints (4.3.3) and (4.3.4) ensure that $f_{it}^j \geq \max\{0, A_{jt} - \frac{x_i^j}{m_{ij}}\}$ while integer requirement (4.3.5) rounds up f_{it}^j to the nearest integer value.

Given the demand sequence (4.3.1), $x_i^n = x_i$ and $x_i^j = x_i - \sum_{k=j+1}^{k=n} \left(A_{kt} - f_{it}^k\right) \times m_{ik}$ where $\left(A_{kt} - f_{it}^k\right) \times m_{ik}$ is the number of packages of product i already used for product k. As such,

$$\begin{aligned}
&f_{it}^n \geq A_{nt} - \frac{x_i}{m_{in}} && \\
&f_{it}^j \geq A_{j,t} - \frac{x_i - \sum_{k=j+1}^{k=n} \left(A_{kt} - f_{it}^k\right) \times m_{ik}}{m_{ij}} && \forall \quad j \in K_i \setminus \{n\} \\
&f_{it}^j \in \mathbf{Z}^+ && \forall \quad j \in K_i
\end{aligned} \tag{4.3.6}$$

The exact total number of failures is then

$$f_{it}^E = \sum_{j \in K_i} f_{it}^j \tag{4.3.7}$$

Given solution $s_{ij} = 1 \ \forall j \in K_i$, we rewrite constraint (4.2.2) as $d_{it} = \sum_{j \in K_i} m_{ij} A_{jt}$. Constraints (4.2.8) and (4.2.14) in models [R2] and [R3] are then

$$f_{it} \geq \sum_{j \in K_i} m_{ij} A_{jt} - x_i \tag{4.3.8}$$

We now show that $f_{it} \geq f_{it}^E$ for any sequence of demand realization. Let $\widetilde{f}_{it}^j = m_{ij} f_{it}^j$ and

rearranging constraints (4.3.6),

$$\begin{aligned} m_{in}f_{it}^{n} &= \widetilde{f}_{it}^{n} \geq m_{in}A_{nt} - x_i & & \\ m_{ij}f_{it}^{j} &= \widetilde{f}_{it}^{j} \geq m_{ij}A_{j,t} - \left(x_i - \sum_{k=j+1}^{k=n}\left(A_{kt} - f_{it}^{k}\right) \times m_{ik}\right) & \forall \quad j \in K_i \setminus \{n\} \\ & \widetilde{f}_{it}^{j} \geq 0 & \forall \quad j \in K_i \end{aligned} \tag{4.3.9}$$

Note that since $\mathbf{A}, \mathbf{m}, \mathbf{x}$ are integers, $\widetilde{f}_{it}^{j}$ always takes an integer value. The integrality requirement on $\widetilde{f}_{it}^{j}$ is therefore dropped. Since $m_{ij} \geq 1$,

$$\sum_{j \in K_i} \widetilde{f}_{it}^{j} \geq \sum_{j \in K_i} f_{it}^{j}. \tag{4.3.10}$$

Setting $x_i^j = x_i - \sum_{k=j+1}^{k=n}\left(A_{kt} - f_{it}^{k}\right) \times m_{ik} = 0 \; \forall j \in K_i \setminus \{n\}$, we have

$$\begin{aligned} \overline{f}_{it}^{n} &\geq m_{in}A_{nt} - x_i, & & \\ \overline{f}_{it}^{j} &\geq m_{ij}A_{j,t} & \forall j \in K_i \setminus \{n\}, \\ \overline{f}_{it}^{j} &\geq 0 & \forall j \in K_i, \end{aligned} \tag{4.3.11}$$

and

$$\sum_{j \in K_i} \overline{f}_{it}^{j} \geq \sum_{j \in K_i} \widetilde{f}_{it}^{j} \tag{4.3.12}$$

$$\sum_{j \in K_i} \overline{f}_{it}^{j} \geq \sum_{j \in K_i} m_{ij}A_{jt} - x_i \tag{4.3.13}$$

Since $f_{it} \geq \sum_{j \in K_i} m_{ij}A_{jt} - x_i$, then by inequalities (4.3.7), (4.3.10), (4.3.12), and (4.3.13)

$$f_{it} = \sum_{j \in K_i} \overline{f}_{it}^{j} \geq \sum_{j \in K^i} \widetilde{f}_{it}^{j} \geq \sum_{j \in K^i} f_{it}^{j} = F_{it}^{E} \Longrightarrow f_{it} \geq f_{it}^{E}$$

This shows that for any given sequence of demand realization, $f_{it} \geq f_{it}^{E}$. Let z^E be the service level achieved when the exact number of failures are computed while z^* be the service level using f_{it}. Then,

$$\sum_{i \in I}\sum_{t \in T} f_{it} \geq \sum_{i \in I}\sum_{t \in T} f_{it}^{E} \Longleftrightarrow \left(1 - \frac{\sum_{i \in I}\sum_{t \in T} fit}{D}\right) \leq \left(1 - \frac{\sum_{i \in I}\sum_{t \in T} f_{it}^{E}}{D}\right) \Longleftrightarrow z^* \leq z^E \tag{4.3.14}$$

which proves that desired service level z^* is always achieved irrespective of the sequence of demand realization. □

5. A Column-Generation Based Heuristic Approach

In many practical problems, the optimization models are of large-scale and it may be impossible to explicitly include all variables in the initial formulation or it may consume too much memory. Column generation is a well-known procedure to solve such large-scale problems where columns are added at each iteration of the simplex method. The idea of column generation was first suggested by Ford Jr and Fulkerson (1958) for *multicommodity network flow* problem and have been successfully applied to many real-life problems including cutting stock problems (Gilmore and Gomory, 1961, 1963), crew scheduling (Desaulniers et al., 1997), and vehicle routing (Agarwal et al., 1989). Oğuz (2002) show that column generation may even be efficient for some problems where the number of variables are low enough to be explicitly included in the model. In our problem, all variables can be explicitly included in the formulation but it consumes too much memory and thus slowing down CPLEX. In particular, the pharmacy kiosk problem consists of too many GPI-QTYs and only a few could be stocked due to limited capacity. As such, one may include variables x_i and s_{ij} only for products that are most likely to be stocked.

We present a column-generation based heuristic approach (CGA) to solve model [R3] to near optimality by selecting only a subset of products in the initial formulation. Other products are then added iteratively. The approach is also applicable for the other two models, [R1] and [R2]. Let $\widetilde{I} \subseteq I$ be the set of products selected for the initial formulation. We rewrite model [R3] and drop integer requirements to formulate the restricted master problem [RMP] as

$$\text{[RMP]:} \quad \max \quad 1 - \frac{1}{D} \times \sum_{i \in I} \sum_{t \in \Theta} f_{it} \tag{5.1}$$

$$\text{s.t.} \quad \sum_{i \in I} x_i \leq C, \qquad [\lambda] \tag{5.2}$$

$$\sum_{\substack{j \in I: \\ b_{ij}=1}} m_{ij} A_{jt} s_{ij} - x_i - f_{it} \leq 0 \quad \forall i \in I, t \in \Theta, \qquad [u_{it}] \tag{5.3}$$

$$\sum_{\substack{i \in I: \\ b_{ij}=1}} s_{ij} = 1 \qquad \forall j \in I, \qquad [\omega_j] \tag{5.4}$$

$$x_i, s_{ij} \geq 0 \qquad \forall i \in \widetilde{I}, j \in I : b_{ij} = 1, \tag{5.5}$$

$$f_{it} \geq 0, \qquad \forall i \in I, j \in I, t \in \Theta, \tag{5.6}$$

where [.] are dual variables for each constraint. Constraint (5.4) along with nonnegativity constraint (5.6) ensures $s_{ij} \leq 1$, and we therefore do not include this constraint to the model. We select a subset of products to initialize the algorithm. A subset should be selected such that it minimizes the number of iterations required to add the columns. To do so, we sort products in decreasing

order of the number of substitution a product can make, and the yearly demand. We then select top $\frac{C}{2}$ products with highest yearly demand and the number of substitutions. This allows us to start off with products that are likely to be stocked due to higher demand and their ability to substitute the demand for other products. Variables x_i and s_{ij} $\forall j \in I$ are introduced for the selected products. Products that are not selected, we only include variable s_{ii} i.e., either its demand is met using one of the selected products or $s_{ii} = 1$, and the demand for such product is never met. Once model [RMP] is solved, its dual information is used to determine potential products to be added to the model. Taking the dual,

$$\text{[RMP-D]:} \quad \min \quad 1 + C\lambda + \sum_{j \in I} \omega_j \tag{5.7}$$

$$\text{s.t.} \quad -u_{it} \geq -\frac{1}{D} \qquad i \in I, t \in \Theta, \qquad [f_{it}] \tag{5.8}$$

$$\lambda - \sum_{t \in \Theta} u_{it} \geq 0 \qquad \forall i \in I, \qquad [x_i] \tag{5.9}$$

$$\sum_{t \in \Theta} m_{ij} A_{jt} u_{it} + \omega_j \geq 0 \qquad \forall i \in I, j \in I : b_{ij} = 1 \qquad [s_{ij}] \tag{5.10}$$

$$\lambda \geq 0, \ u_{it} \geq 0, \omega_j \rightarrow \text{urs} \qquad \forall i \in I, j \in I, t \in \Theta \tag{5.11}$$

Given λ and u_{it}, the reduced cost is $RC_i = \lambda - \sum_{t \in \Theta} u_{it}$ for product i. Let $\mathcal{I}$ be the set of products not included in the initial formulation. The pricing problem $\min_{i \in \mathcal{I}} \{\lambda - \sum_{t \in \Theta} u_{it}\}$ determines the product with most negative reduced costs which is then added to [RMP]. For most of the problems in the literature, enumerating over all possible columns is computational impractical and therefore a pricing subproblem is solved to determine the column to be added. In our case, however, one could easily calculate reduced costs for all products. Instead of selecting the product with most negative reduced cost, we select all products with reduced cost $RC_i < 0$ and columns x_i and s_{ij} $\forall j \in I$ are added to [RMP] which is solved again. This procedure terminates when $RC_i \geq 0 \forall i \in I$, and the latest [RMP] solution provides a lower bound to the original model [R3]. The other approach could be to first solve model [RMP] with all variables. Then, for products with positive RC_i, variables x_i and s_{ij} $\forall j \in I$ are removed. However, it turns out that such an approach is computationally inefficient compared to the proposed column generation approach.

To obtain a feasible solution, [RMP] is solved with integrality constraints on x_i and s_{ij} and its objective function value is an upper bound to model [R3]. Note that this approach does not guarantee optimality. To solve to optimality, one needs to apply the CGA at each node of the branch-and-bound tree. However, we implement CGA only at the root node and computational results in Section 6.3 show that optimality gap is 1.1%, on average.

6. Results

We perform numerical testing over several datasets including seven pharmacy store sales data and randomly generated instances. In Section 6.1, we use the proposed optimization models to determine the optimized storage capacity for MedAvail's pharmacy kiosk and recommend assortment and stocking guidelines using pharmacy sales data. To further generalize model results, we solve model [R2] using randomly generated instances in Section 6.2 and derive managerial insights. Finally, the proposed column generation solution approach is compared against CPLEX and Benders decomposition in Section 6.3.

6.1. The case of MedAvail

In this section, we first use models [R1] and [R2] to analyze the effects of substitution and replenishment lead time on the capacity of a kiosk using single pharmacy store data for the year 2015. The data records 2,355 GPIs (or product classes) and 10,145 GPI-QTYs (or products). The goal is to assess the savings in kiosk capacity through drug substitution and through reducing replenishment lead time from two days to one day. The management suggested that it is useful to explore the effect of capacity on service level, as it may not be possible to build a machine of an optimized capacity. Therefore, we use model [R3] to determine maximum service level achieved at different capacity levels as suggested by the management. We then perform several experiments using multiple datasets generated from seven 24/7 pharmacy store sales data to provide bounds on the service level that management should expect to achieve at a given capacity. All optimization models are coded in C++ and solved using CPLEX version 12.6.3 on a 64-bit Windows 10 with Intel(R) Core i5-5300U 2.30GHz processors and 4.00GB RAM. We solve all instances to an optimality gap of 0.5% since solving the problem to optimality may only reduce the required capacity by at most 57 which is not significant from the management perspective. They were of the view that such an exact machine could not be built and the optimized capacity values be rounded off to the nearest 100. Finally, we evaluate the computational efficiency of the proposed column generation approach against solving model [R3] directly using CPLEX.

6.1.1. Effects of substitution

Service level, α	[R2]	[R1]-MedAvail's substitution		[R1]-no substitution	
	Capacity, C	Capacity, C	$\triangle$% to [M2]	Capacity, C	$\triangle$% to [M2]
80%	2,542	2,710	6.6%	2,618	3.0%
85%	3,261	3,449	5.8%	3,385	3.8%
90%	4,375	4,606	5.3%	4,583	4.8%
95%	6,485	6,856	5.7%	6,938	7.0%
96%	7,233	7,604	5.1%	7,686	6.3%
97%	7,980	8,506	6.6%	8,690	8.9%
98%	9,460	10,002	5.7%	10,186	7.7%
99%	10,956	11,497	4.9%	11,681	6.6%

Table 3: Kiosk storage capacity to achieve desired service level α under different substitution rules.

The management was inclined towards a predefined substitution criterion rather than a complex mathematical model. So we optimized capacity under various substitution strategies to see whether substitution plays a role in deciding on the capacity of the kiosk. We generate a dataset using one pharmacy store sales data for the year 2015 with replenishment lead time, $h = 2$. Model [R1] is solved under two distinct substitution rules: (1) MedAvail's substitution rule, and (2) no substitution. MedAvail's substitution rule was suggested by the management where a GPI-QTY i substitutes GPI-QTY j with the same GPI code, if the quantity of j is twice that of i and its average lead time demand is less than 25% of that of i, or if the quantity of j is three times that of i and its average lead time demand is less than 15% of that of i. An iterative procedure is used to assign values to the substitution variables s_{ij} based on this rule, and d_{it} is calculated apriori. Model [R2] is solved to determine optimized substitution. Each model is solved repeatedly by varying the desired service level α between 80% and 99%. Table 3 summarizes the results.

At 95% service level, the capacity under optimized substitution is 6,485. It increases by 5.7% when MedAvail's rule is used and by 7.0% when substitution is not allowed. As the service level decreases, the effect of MedAvail's substitution rule decreases. In fact, the effect becomes negative relative to no substitution when the service level is 90% or lower. This is due to over substitution by MedAvail's substitution rule at lower service levels. At lower service levels, fewer GPI-QTYs should be substituted to optimize the capacity. Optimized substitution, as expected, is always better than both no substitution and apriori rules. This comes at the expense of larger solution times. Given the potential improvements in capacity under optimized substitution, model [R2] is used in subsequent analysis.

6.1.2. Effect of replenishment lead time

Before the start of the project, kiosks were being replenished every other day. MedAvail management wanted to investigate the effect of replenishment lead time on capacity and assortment decisions. A larger lead time is expected to increase capacity since lead time demand would be higher, so we experimented with 1 and 2 day lead times. Table 4 summarizes the results where [R2] is solved at eight different service levels. At 90% service level, the capacity is reduced by 14% when the replenishment lead time is reduced from two days to one day. Although a one day lead time may increase operating costs of the kiosk due to frequent replenishment, management believes that the significant reduction in capacity is much more important when taking into account the technical challenges in designing a kiosk with higher capacity. Testing in subsequent sections is

Service level, α	Capacity			Threshold demand	
	C_1	C_2	$\triangle\%$	$h = 1$	$h = 2$
80%	2,093	2,542	18%	14	15
85%	2,743	3,261	16%	11	11
90%	3,769	4,375	14%	6	7
95%	5,745	6,485	11%	3	4
99%	10,161	10,956	7%	2	2

Table 4: We compare storage capacity C_h at one day ($h = 1$) and two day ($h = 2$) lead time. Threshold demand is the highest yearly demand among all GPI-QTYs that are not stocked.

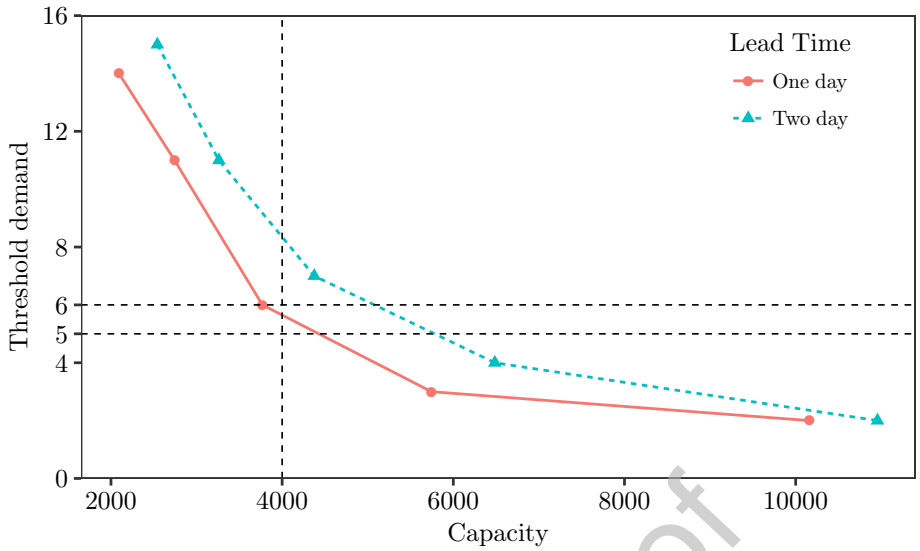


Figure 7: The graph plots threshold demand against storage capacity under different lead times.

based on daily replenishment.

To study the significance of co-ordering, we report the highest yearly demand among all GPI-QTYs that are not stocked in Table 4. For a one day lead time ($h = 1$) and service level $\alpha = 80\%$, every GPI-QTY with a yearly demand greater than 15 are stocked. Recall that in the Section *Co-ordering of drugs*, we set threshold support to $\frac{15}{D}$. The results in Table 4 show that demand threshold is always less than or equal to the threshold support used in the Apriori association rule algorithm. As such, GPI-QTYs that frequently appear together are already stocked when the service level is $\alpha \geq 80\%$. As expected, threshold demand decreases as α increases. Therefore, we do not need to incorporate association rules explicitly in our modelling framework. Given the lead time and kiosk capacity, Figure 7 may be used as easy to use guidelines to decide on which medications to store without solving the assortment problem. For instance, if MedAvail decides on one day replenishment lead time for a kiosk with capacity $C = 4,000$, threshold demand is 5.6 based on Figure 7. As such, MedAvail should stock all GPI-QTYs with yearly demand greater or equal to 6.

6.1.3. Capacity planning over multiple pharmacies

A crucial question we faced in deriving demand distributions from the data is whether to use individual store data or multiple stores data and whether to use average or maximum observed demands in the latter case. Each of these approaches may have merits and drawbacks. We carried several tests to answer this question. At this point, management suggested that it is useful to explore the effect of limited capacity on service level, as it may not be possible to build a machine of an optimized capacity. Hence, we modified the objective to service level maximization and added a constraint that limits capacity to obtain model [R3]. The results presented next are based on service level maximization where capacity is varied between 2,000 and 7,000 with an increment of

1,000.

Individual store data (IAS). The IAS approach makes stocking and substitution decisions for each store individually using its yearly demand data. The expected service levels achieved at seven pharmacy stores are shown in Table 5a. On average, setting the capacity to 5,000 achieves a service level of 92.5%. The drawback of IAS approach is that it may lead to overestimation of the service level due to over-fitting, also referred to as *optimizers curse* in the Operations Research literature. Overfitting leads to stocking decisions that are susceptible to small changes in demand which could lead to much worse service levels. IAS approach therefore provides an upper bound on the service level achieved.

Capacity C	Store ID							Average
	S1	S2	S3	S4	S5	S6	S7	
2,000	78.9%	78.3%	77.4%	77.1%	75.9%	75.9%	74.5%	76.9%
3,000	86.3%	86.0%	85.2%	85.1%	84.1%	84.0%	83.0%	84.8%
4,000	90.6%	90.4%	89.8%	89.7%	89.0%	88.9%	88.0%	89.5%
5,000	93.4%	93.4%	92.8%	92.7%	92.1%	92.0%	91.3%	92.5%
6,000	95.3%	95.4%	94.7%	94.1%	94.4%	94.2%	93.6%	94.5%
7,000	96.5%	96.5%	96.2%	96.1%	95.8%	95.7%	95.1%	96.0%

(a) Service level achieved at different stores at different capacities (IAS)

Capacity C	Store ID						
	S1	S2	S3	S4	S5	S6	S7
2,000	69.4%	69.6%	71.3%	70.1%	70.3%	71.1%	74.5%
3,000	76.5%	76.6%	78.1%	77.3%	77.8%	78.3%	83.0%
4,000	80.5%	80.7%	82.0%	81.4%	82.1%	82.3%	88.0%
5,000	83.8%	83.4%	84.6%	84.2%	85.0%	85.0%	91.3%
6,000	85.6%	85.1%	86.5%	86.0%	86.9%	86.9%	93.6%
7,000	87.1%	86.6%	88.1%	87.4%	88.6%	88.4%	95.1%

(b) Service level achieved using most-active store data to make stocking decisions (MSD)

Capacity C	Store ID							Average
	S1	S2	S3	S4	S5	S6	S7	
2,000	69.5%	69.9%	69.7%	68.7%	67.3%	67.3%	66.0%	68.3%
3,000	75.6%	76.4%	75.9%	75.2%	73.8%	73.7%	72.9%	74.8%
4,000	78.9%	79.7%	79.1%	78.7%	77.3%	77.3%	76.6%	78.2%
5,000	80.6%	81.5%	80.7%	80.8%	79.2%	79.4%	78.7%	80.1%
6,000	82.0%	82.5%	82.0%	82.1%	80.6%	81.2%	80.0%	81.5%
7,000	83.4%	84.1%	83.3%	83.7%	82.1%	82.6%	81.5%	83.0%

(c) Service level achieved using average demand over all stores to make stocking decisions (ADS)

Capacity C	Store ID							Average
	S1	S2	S3	S4	S5	S6	S7	
2,000	73.7%	74.6%	74.4%	73.8%	73.1%	71.9%	71.7%	73.3%
3,000	81.6%	82.2%	81.6%	81.5%	81.1%	79.6%	79.9%	81.1%
4,000	85.8%	86.6%	86.1%	85.9%	85.7%	84.4%	84.9%	85.6%
5,000	88.6%	89.1%	88.8%	88.6%	88.6%	87.2%	87.8%	88.4%
6,000	90.5%	91.0%	90.7%	90.6%	90.8%	89.7%	90.1%	90.5%
7,000	92.2%	92.6%	92.3%	92.4%	92.5%	91.6%	91.9%	92.2%

(d) Service level achieved using the highest demand across all stores in a given period $t \in \Theta$ to make stocking decisions (HDS)

Table 5: Capacity Planning using different demand prediction strategies

Most-active store data (MSD). To avoid overfitting, we make stocking and substitution decisions using the most active store data, i.e., the one with highest yearly sales. The optimized decisions are then applied to all other stores data to calculate their achieved service levels. The results in Table 5b highlight the problem of overfitting with IAS approach. The expected service levels are substantially reduced when the optimal solution from the most active store is applied to other stores. On average, service level achieved at capacity $C = 5,000$ is 84.3%.

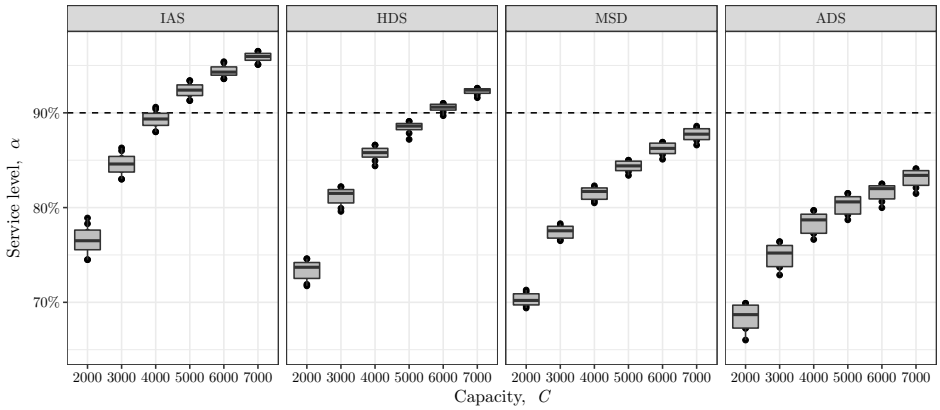
Although MSD approach addresses the problem of overfitting, it ignores GPI-QTYs ordered at other stores. The number of distinct GPIs recorded in the year 2015 at a store varies between 2,316 and 2,509. However, when the data is aggregated for all stores, the total number of distinct GPIs equals 3,579. Similarly, the number of distinct GPI-QTYs recorded at the most active store equals 12,014. This number increases to 29,626 when all store data is analyzed. As such, an optimal solution derived based on one store may be suboptimal for other stores and only provides a lower bound on the service level.

Average demand over all stores (ADS). Both IAS and MSD approaches use a single store data and ignore GPI-QTYs ordered at the other stores. To overcome this, we generate a new dataset by calculating the average demand of a GPI-QTY in time period $t \in \Theta$, over all stores. We use this new dataset containing all GPI-QTYs to make stocking and substitution decisions, which are then applied to stores data to calculate their achieved service levels. ADS approach results in poor stocking and substitution decisions as shown in Table 5c. When capacity $C = 5,000$, the average service level over all stores is 80.1%. This is due to the aggregation of demand which leads to reduced uncertainty. Consider a GPI-QTY i, with demand on a specific day at four stores as $\{0, 1, 0, 3\}$. If the stock level $x_i = 1$, then the number of failures at store 4 equals $3 - 1 = 2$. However, the average demand equals $\frac{0+1+0+3}{4} = 1$, and the calculated number of failures equals $1 - 1 = 0$. This example shows that averaging demand over all stores does not capture variability among stores, leading to suboptimal solutions and lower service levels.

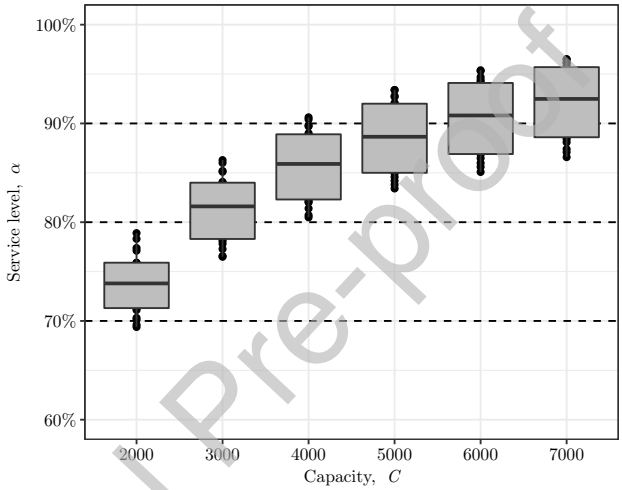
Highest demand over all stores (HDS). Another approach is to use the maximum demand of each GPI-QTY in a given time period $t \in \Theta$ across all stores. The HDS approach provides better stock levels that are robust for all stores by making stocking and substitution decisions under the worst-case scenario. The results are summarized in Table 5d. At capacity $C = 5,000$, the service level achieved is 88.4% on average. The drawback of HDS is that it may overestimate stock levels for some SKUs as the decisions are made under worst-case scenario. It is also possible that demand characteristics may vary from store to store and some GPI-QTYs ordered at one store may never be ordered at other stores.

6.1.4. Recommendations

Computational results show that substitution and daily replenishment guidelines significantly reduce the capacity required to achieve a desired service level. We observe that MedAvail's substitution rule is not as effective as optimized substitution which may save up to 9% of capacity. On



(a) Comparative analysis of the stocking decision approaches



(b) Capacity planning using top 3 approaches (IAS,MSD,HDS)

Figure 8: Capacity Planning using different approaches

the other hand, daily replenishment saves up to 18% of capacity compared to two day replenishment. The results also show that the marginal benefit of additional capacity decreases at higher capacities as illustrated in Figure 8a where the service level increases at a decreasing rate as the capacity increases. We also observe that the service level achieved at a fixed capacity is roughly the same across all stores as shown by the boxplots in Figure 8a.

To present robust results, we perform several experiments at different service levels using four different demand prediction strategies. Other than ADS which results in suboptimal solutions, management may use any of the other three approaches discussed earlier. The approaches IAS and MSD provide upper and lower bounds on the service level, respectively. On the other hand, the HDS approach offers a more realistic expectation of the service level and gives stock levels that are robust against small changes in demand. Management may also make capacity decisions using a combination of the three approaches as illustrated in Figure 8b. The boxplot represents the uncertainty in the service level achieved at a fixed capacity. For instance, when the capacity is set to 5,000, MedAvail should expect a service level between 84% and 93% depending on the store

under consideration and the level of conservatism when making stocking decisions.

6.2. Numerical Analysis over Randomly Generated Instances

In this section, we solve model [R2] using randomly generated data instances to generalize the findings of the case study. Section 6.2.1 details the procedure employed to generate random data and in Section 6.2.2, we discuss model results and derive managerial insights.

6.2.1. Data Generation

To generate data instances, we consider 200 distinct product classes and randomly generate products for each class from a uniform distribution, $Unif[1, 10]$. To study the effect of substitution, three distinct substitution patterns are defined: (1) "None", QTYs $= \{2, 3, 5, 7, ...\}$, where product substitution is not possible as no product quantity is a multiple of another, (2) "Single", QTYs $= \{1, 2, 3, 5, ...\}$, where only the smallest quantity product is able to substitute all other quantities, and (3) "All", QTYs $= \{1, 2, 4, 8, ...\}$, where all smaller quantity products can substitute larger quantity product. To generate demand values, we randomly generate yearly demand for each product from an exponential distribution $Exp(\frac{1}{\mu})$ where μ is varied between 10 and 50 with increments of 10. Mean daily demand μ_i for product i is calculated as $\mu_i = \frac{Exp(\frac{1}{\mu})}{365}$ which is used to generate 200 demand scenarios from Poisson distribution, $Poi(\mu_i)$. Figure 9 plots the cumulative distribution of yearly demand for different values of μ. As μ increases, the product's probability of having high yearly demand increases. As such, increasing μ reduces the number of products with low yearly demand. Sensitivity analysis over μ allows us to study the effect of substitution under different demand settings where low values of μ implies low and erratic demand while setting higher values for μ implies less sporadic demand. Service level α is also set at eight different levels between 80% to 99%. For a given μ, substitution pattern, and service level, 5 random instances are generated, resulting in a total of 600 instances.

6.2.2. Results on Random Instances

Tables 6 and 7 summarize computational results for substitution patterns "Single" and "All", respectively. Average values over 5 randomly generated instances are reported in the tables. Column "value" under "Capacity" records the minimum capacity required to achieve desired service level α and column "Δ% imp" denotes percentage reduction in required capacity due to substitution. The latter is calculated as the percentage difference in optimal capacity under a given substitution pattern ("Single" or "All") and substitution pattern "None". Column "Possible" counts the total number of products that can be substituted by other products, while Column "Optimized" counts the number of products substituted by other products in the optimal solution, i.e., $\sum_{\substack{i \in I: \\ b_{ij}=1}} \sum_{\substack{j \in I: \\ i \neq j}} s_{ij}$. Column "% Substituted" is the ratio of "Optimized" to "Possible". Total number of products considered is given in column "Nb. Products" out of which, "Nb. Covered" number of products are stocked or substituted by stocked products in the optimal solution. Column "%. Covered" is the percentage of products covered in each instance.

Mean	Service	Capacity		Product Substitution			Product Coverage		
Demand	level	Value	Δ% imp	Possible	Optimized	% Substituted	NbProducts	Covered	% Covered
10	80%	461	0.3%	872	21	2.4%	1072	460	42.9%
	85%	532	0.4%	872	30	3.4%	1072	529	49.4%
	90%	623	0.5%	872	42	4.9%	1072	616	57.4%
	95%	751	0.7%	872	71	8.2%	1072	730	68.1%
	96%	781	1.1%	872	73	8.3%	1072	760	70.8%
	97%	833	1.6%	872	113	13.0%	1072	796	74.2%
	98%	891	1.5%	872	121	13.9%	1072	819	76.4%
	99%	949	1.4%	872	115	13.2%	1072	849	79.1%
Average		728	0.9%	872	73	8.4%	1072	695	64.8%
20	80%	508	0.3%	872	15	1.7%	1072	502	46.8%
	85%	590	0.5%	872	25	2.9%	1072	580	54.1%
	90%	696	0.7%	872	36	4.1%	1072	676	63.0%
	95%	856	1.3%	872	69	7.9%	1072	796	74.2%
	96%	902	1.8%	872	80	9.2%	1072	834	77.8%
	97%	960	1.7%	872	93	10.6%	1072	860	80.2%
	98%	1022	2.4%	872	99	11.3%	1072	904	84.3%
	99%	1136	2.4%	872	128	14.7%	1072	942	87.8%
Average		834	1.4%	872	68	7.8%	1072	762	71.0%
30	80%	559	0.5%	872	18	2.0%	1072	546	50.9%
	85%	648	0.7%	872	25	2.9%	1072	626	58.3%
	90%	766	1.1%	872	37	4.2%	1072	725	67.6%
	95%	944	1.5%	872	58	6.6%	1072	846	78.9%
	96%	999	1.7%	872	74	8.5%	1072	869	81.1%
	97%	1063	2.1%	872	79	9.0%	1072	907	84.6%
	98%	1151	2.0%	872	91	10.5%	1072	948	88.5%
	99%	1283	2.6%	872	119	13.7%	1072	977	91.1%
Average		927	1.5%	872	63	7.2%	1072	806	75.1%
40	80%	587	0.5%	872	16	1.9%	1072	565	52.7%
	85%	685	0.7%	872	24	2.7%	1072	647	60.3%
	90%	815	1.0%	872	39	4.5%	1072	746	69.5%
	95%	1013	1.6%	872	60	6.9%	1072	871	81.2%
	96%	1073	1.7%	872	68	7.8%	1072	897	83.6%
	97%	1146	1.9%	872	74	8.5%	1072	924	86.1%
	98%	1241	2.2%	872	90	10.3%	1072	955	89.1%
	99%	1381	3.0%	872	118	13.5%	1072	989	92.2%
Average		993	1.6%	872	61	7.0%	1072	824	76.9%
50	80%	619	0.3%	872	14	1.7%	1072	583	54.4%
	85%	723	0.5%	872	19	2.2%	1072	666	62.1%
	90%	860	0.7%	872	32	3.7%	1072	762	71.0%
	95%	1075	1.1%	872	51	5.9%	1072	877	81.8%
	96%	1139	1.3%	872	58	6.7%	1072	909	84.8%
	97%	1221	1.6%	872	69	7.9%	1072	939	87.6%
	98%	1327	1.9%	872	80	9.2%	1072	970	90.4%
	99%	1481	2.6%	872	94	10.8%	1072	999	93.2%
Average		1056	1.3%	872	52	6.0%	1072	838	78.2%

Table 6: Numerical results for Random Instances under Substitution Pattern "Single"

Mean	Service	Capacity		Product Substitution			Product Coverage		
Demand	level	Value	Δ% imp	Possible	Optimized	% Substituted	NbProducts	Covered	% Covered
10	80%	457	1.4%	872	59	6.8%	1072	456	42.5%
	85%	525	1.6%	872	87	10.0%	1072	524	48.9%
	90%	612	2.3%	872	110	12.6%	1072	610	56.9%
	95%	732	3.3%	872	154	17.6%	1072	725	67.6%
	96%	761	3.7%	872	159	18.3%	1072	752	70.1%
	97%	796	6.3%	872	159	18.3%	1072	781	72.9%
	98%	853	6.0%	872	153	17.6%	1072	822	76.7%
	99%	911	5.6%	872	154	17.6%	1072	866	80.8%
Average		706	3.8%	872	129	14.8%	1072	692	64.5%
20	80%	501	1.7%	872	55	6.3%	1072	499	46.5%
	85%	579	2.4%	872	79	9.0%	1072	576	53.7%
	90%	678	3.3%	872	106	12.2%	1072	672	62.6%
	95%	826	5.0%	872	158	18.1%	1072	802	74.8%
	96%	865	6.1%	872	160	18.4%	1072	828	77.2%
	97%	920	6.1%	872	185	21.2%	1072	865	80.7%
	98%	979	6.9%	872	202	23.2%	1072	902	84.1%
	99%	1075	8.3%	872	186	21.3%	1072	939	87.6%
Average		803	5.0%	872	141	16.2%	1072	760	70.9%
30	80%	552	1.8%	872	59	6.7%	1072	546	50.9%
	85%	637	2.5%	872	79	9.0%	1072	625	58.3%
	90%	747	3.6%	872	109	12.5%	1072	724	67.5%
	95%	910	5.4%	872	147	16.8%	1072	849	79.1%
	96%	955	6.4%	872	160	18.4%	1072	878	81.9%
	97%	1014	7.1%	872	179	20.5%	1072	914	85.3%
	98%	1095	7.2%	872	177	20.3%	1072	946	88.3%
	99%	1207	9.1%	872	182	20.9%	1072	974	90.8%
Average		889	5.4%	872	136	15.6%	1072	807	75.3%
40	80%	580	1.7%	872	58	6.6%	1072	566	52.8%
	85%	672	2.6%	872	78	8.9%	1072	649	60.5%
	90%	794	3.6%	872	112	12.9%	1072	750	69.9%
	95%	978	5.3%	872	162	18.5%	1072	876	81.7%
	96%	1033	5.6%	872	163	18.7%	1072	905	84.4%
	97%	1098	6.3%	872	183	21.0%	1072	932	86.9%
	98%	1180	7.5%	872	180	20.6%	1072	961	89.6%
	99%	1301	9.3%	872	190	21.8%	1072	988	92.2%
Average		955	5.2%	872	141	16.1%	1072	828	77.3%
50	80%	610	1.8%	872	58	6.6%	1072	589	55.0%
	85%	707	2.7%	872	82	9.4%	1072	674	62.8%
	90%	836	3.6%	872	117	13.4%	1072	777	72.4%
	95%	1035	5.0%	872	156	17.8%	1072	894	83.4%
	96%	1094	5.5%	872	162	18.6%	1072	920	85.8%
	97%	1167	6.2%	872	176	20.2%	1072	951	88.7%
	98%	1262	7.2%	872	185	21.2%	1072	978	91.2%
	99%	1403	8.3%	872	202	23.2%	1072	1005	93.7%
Average		1014	5.1%	872	142	16.3%	1072	848	79.1%

Table 7: Numerical results for Random Instances under Subsitution Pattern: "ALL"

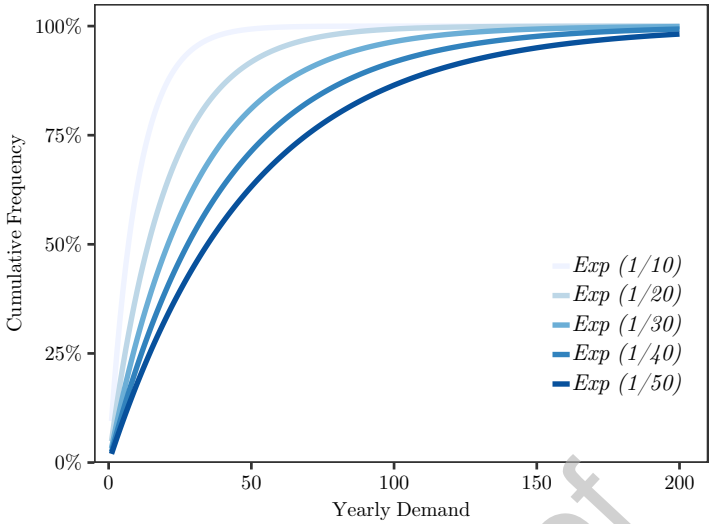


Figure 9: The graph plots exponential distribution under different mean values.

Results in Tables 6 and 7 show that substitution plays an important role in reducing the storage capacity at higher service levels. For instance, under substitution pattern "All" and $\mu = 40$, substitution is able to reduce the storage capacity by 9.3% when desired service level $\alpha = 99\%$. On the other hand, when $\alpha = 80\%$, the capacity is reduced by only 1.7%. This is further illustrated in Figure 10a(i) that plots a boxplot for the percentage reduction in capacity (PRC) at given a service level under each substitution pattern. At higher service levels, more products are substituted as shown in Figure 10a(ii). We observe that the effect of substitution is significant when more products are able to substitute. As shown in Figure 10a(ii), percentage of products substituted is higher under substitution pattern "All" compared to "Single", and as such, PRC is significantly higher for "All". For instance, when $\mu = 30$, the average PRCs under patterns "Single and "All" are 1.5% and 5.4%, respectively. However, when product demand is generated from exponential distribution with $\mu = 10$, PRC starts decreasing at higher service levels. Under substitution pattern "All" and $\alpha = 97\%$, PRC is 6.3% which decreases to 5.6% for $\alpha = 99\%$. This is because substitution negatively impacts the service level of the products substituting other products and at higher service levels, negative affect outweighs the positive affect of substitution in improving service level of the unstocked products. As such, product substitution is less preferred as shown in Table 7 where number of products substituted decreases from 159 to 154 when desired service level is increased from 97% to 99%.

Sensitivity analysis over mean demand μ shows that when the number of products with low demand is high, the effect of substitution on PRC is low. The effect of product demand is illustrated in Figure 10b(i) where PRC increases at a decreasing rate as the product demand increases. Under pattern "All" and $\mu = 10$, the average PRC is 3.8% which increases to 5.0% when $\mu = 20$. When all or most of the products have low demand, products' stock levels are low and cannot substitute higher quantity products that require multiple packages to be dispensed. In contrast, when μ

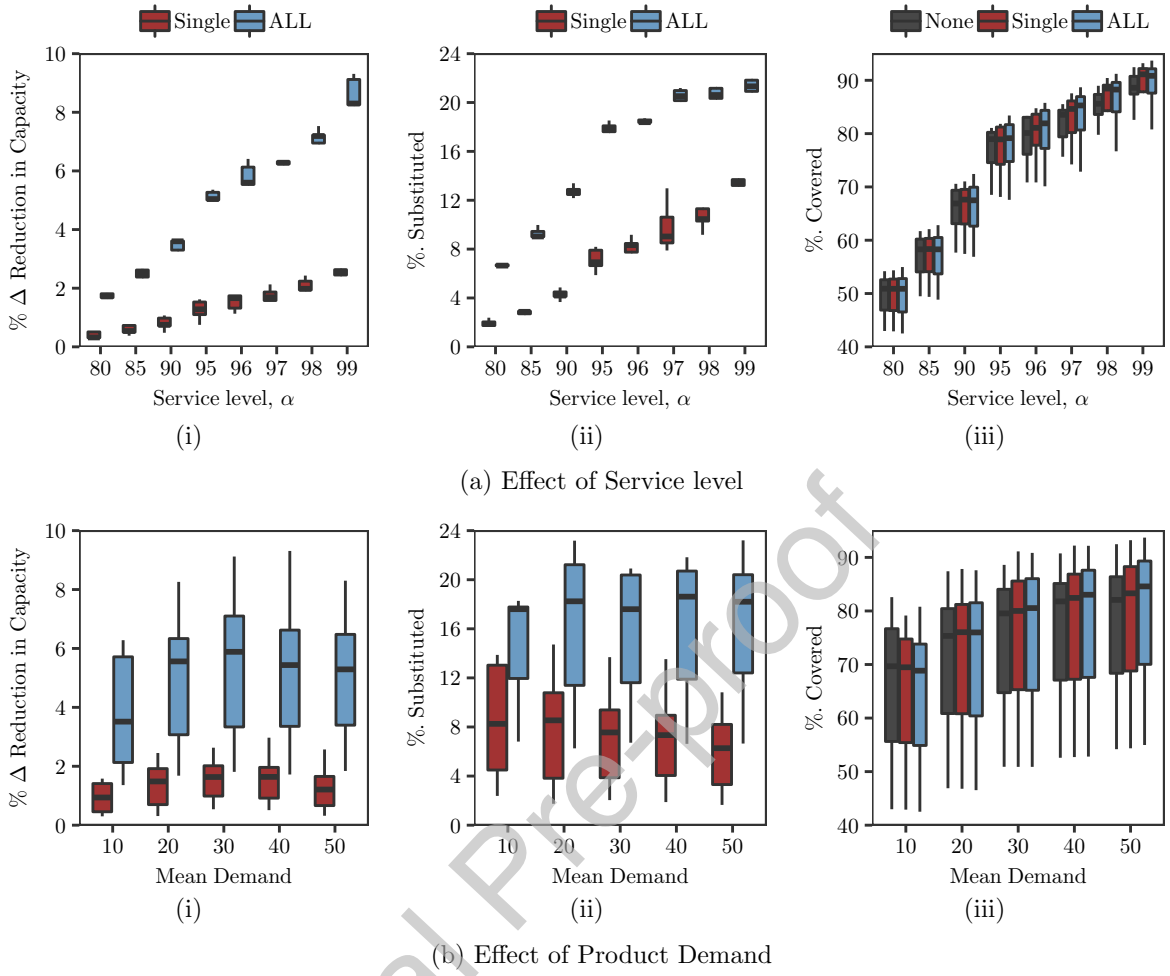
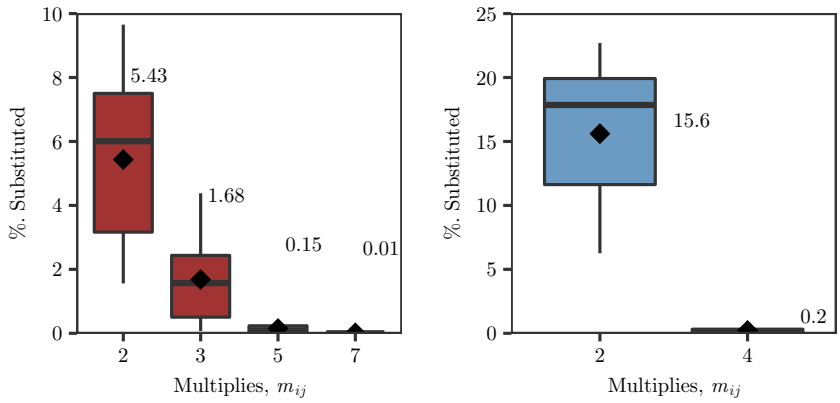


Figure 10: Figures (a) and (b) illustrate the effect of service level and product demand on percentage reduction in capacity (PRC), products substituted, and product coverage, respectively.

changes from 30 to 40, PRC decreases from 5.4% to 5.2%. This is due to the fact at higher values of μ, product demand is high and it is preferred to stock a product rather than substituting it which would result in multiple packages to be dispensed, whenever it is ordered. This is illustrated in Figure 10b(ii) where product substitution does not increase with increasing μ. Figures 10a(iii) and 10b(iii) illustrate how product coverage is effected by service level and mean demand μ under each substitution pattern, respectively. The plots show that the effect of substitution in improving product coverage is not significant. For $\mu = 30$, the percentage of products covered is 75.1%, on average, under pattern "Single" which increases slightly to 75.3% under pattern "All". Analysis over demand shows that product substitution is preferred when there is a right balance between the number of products with low demand and the ones with high demand. Product substitution does not have a significant effect when most of the products have either high or low demand.

We also study the effect of multiplies on product substitution as shown in Figures 11a and 11b. Under substitution pattern "Single", the smallest quantity product only substitutes products with



(a) Single (b) All

Figure 11: Figures (a) and (b) plot the percentage of products substituted with multiples m_{ij} under substitution pattern "Single" and "All", respectively.

multiples m_{ij} less than or equal 7, where 5.3% of the products substituted have multiplies $m_{ij} = 2$ only 0.01% of the products are substituted with $m_{ij} = 7$. For pattern "All", where all smaller quantity products are able to substitute higher quantity products, only the products with $m_{ij} \leq 4$ are substituted. The results show that the cost of substitution is implicitly captured by m_{ij} and the product substitution is less preferred when m_{ij} increases. This is due to multiple packages to be dispensed for the substituted products which would lead to fewer packages available for substituting product.

6.3. Analysis of Solution Approach: CGA

To test the computational efficiency of the column generation approach, we generate five instances using HDS data with 29,626 products by varying the kiosk capacity between 1,000 and 7,000. For each instance, we solve model [R3] using CPLEX and Benders decomposition, and compare their performances against the CGA approach. Column generation alogrithm is coded in C++ Visual Studio 2013 and all optimization problems are solved using CPLEX version 12.6.1 on a 64-bit Windows 10 with Intel(R) core i7-4790 3.60GHz processors and 8.00 GB RAM. For CPLEX and CGA, each instance is executed to an optimality gap of 1e-09 with no time limit.

Computational results are summarized in Table 8a where the column generation is compared against CPLEX solution. Column "RMP linear Sol" is the solution to model [RMP] and column "Best found Sol" denotes the best integer solution found by adding integer constraint to [RMP]. Gap is calculated as $\frac{\text{RMP linear Sol-Best found Sol}}{\text{Best found Sol}}$. Column "CG time" is the time spent to generate all columns while "RMP-MIP time" denotes the time spent to solve model [RMP] with integer constraint. The total CPU time (in seconds) spent by column generation approach and CPLEX are denoted by "TCGA time" and "CPLEX time", respectively. Finally, Time Ratio in Table 8a is calculated as $\frac{\text{CPLEX time}}{\text{TCGA time}}$. Overall, CGA is able to solve all instances in less than one hour with optimality gaps less than 2%. At capacity $C = 7,000$, gap value shows that the best solution obtained from CGA can only be improved by at most 1.94% if the original model [R3] is solved

Capacity	Column Generation Approach					Time Comparison		
C	[RMP] linear Sol.	Best found Integer Sol.	Gap	CG time	RMP-MIP time	TCGA time(s)	CPLEX time (s)	Time ratio
1000	38.85%	38.84%	0.02%	82.92	568.35	651.27	3651.55	5.61
2000	58.68%	58.41%	0.47%	143.35	1162.32	1304.67	3968.64	3.04
3000	70.45%	69.63%	1.18%	143.77	1728.56	1872.33	4300.48	2.30
5000	82.95%	81.39%	1.91%	195.48	1754.58	1949.62	4906.26	2.52
7000	89.06%	87.36%	1.94%	268.98	2854.92	3123.90	5880.13	1.88
		Avg	**1.10%**				**Avg**	**3.07**

(a) Column Generation Approach vs CPLEX

Capacity, C	Benders				Column Generation Approach			
	UB	LB	Gap	Iterations	UB	LB	Gap	Time(s)
1000	41.16%	26.00%	58.31%	71	38.85%	38.84%	0.02%	651.27
2000	67.78%	36.70%	84.67%	52	58.68%	58.41%	0.47%	1304.67
3000	88.17%	42.47%	107.60%	38	70.45%	69.63%	1.18%	1872.33
5000	114.67%	53.13%	115.83%	31	82.95%	81.39%	1.91%	1949.62
7000	1.33629	60.31%	121.56%	28	89.06%	87.36%	1.94%	3123.9
		Average	97.59%			**Average**	1.10%	

(b) Column Generation Approach vs L-shaped Benders decompositionL-shaped

Table 8: Computational efficiency of the proposed column generation against CPLEX and L-shaped Benders Decomposition

to optimality. In fact, for all instances, optimal solutions obtained by directly solving model [R3] equals the solution obtained by CGA. This signifies the effectiveness of the CGA in obtaining solutions that are close to optimal while reducing the computational effort by a factor of three.

We also compare our proposed column generation approach against the L-shaped Benders decomposition approach generally applied in stochastic programming where the master problem decides on first stage decision variables while the subproblem decides on second stage decision variables. The overall Benders procedure is based on the general framework in Carøe and Tind (1998) and is detailed in Appendix A. All instances are solved with a time limit of 3600 seconds. Computational results are summarized in Table 8b where Columns "UB" and "LB" denote upper bound and lower bound obtained from respective solution approaches, respectively, while column "Iterations" refers to the number of iterations between the master problem and subproblems in Benders decomposition. Computational results show that the proposed column generation approach outperforms Benders decomposition. The latter fails to solve any instance to optimality within one-hour time window and reports an average optimality gap of 97.6%.

7. Conclusions

In this paper, we addressed the strategic capacity and assortment planning problem faced by MedAvail through extensive descriptive and prescriptive analytics. We developed three optimization models that decide on stock levels and product substitution. In addition, we developed a column-generation based heuristic solution methodology that is able to obtain near-optimal solutions within

1.1% of optimality gap while reducing computational times by a factor of 3. Computational experiments over real and randomly generated data show that product substitution reduces kiosk's capacity requirements by up to 9%. We also show that the effect of product substitution depends on desired service level and the nature of demand data. As an outcome of this work, MedAvail expects to improve its service levels by 30% using a larger capacity kiosk. MedAvail also expects 10% improvement in service levels of the existing kiosks by optimizing assortment and stocking decisions using the suggested optimization models. In the future, MedAvail plans to implement our optimization models in their technology to make better inventory and assortment decisions.

A promising research direction is to use robust optimization to model the uncertainty in demand. Another possible extension could be to model exact substitution. As discussed earlier in Lemma 3, the proposed model obtains a lower bound on the service level due to conservative approximation of the number of failures. As such, some of the substitution rules that could improve the solution are not selected. Also, the model allows only one substitute for each quantity. Modelling the problem with exact substitution and multiple substitutes would be computationally difficult to solve, and developing an efficient solution methodology for such a model is another promising future work.

Acknowledgement

This research was supported by NSERC Engage grant number EGP 500468-16.

Appendix A. L-shaped Benders Decomposition

In this section, we solve model [M3] using L-shaped Benders decomposition based on the general framework by Carøe and Tind (1998) where the master problem decides on first stage decision variables while the subproblem decides on second stage decision variables. For model [M3], $\mathbf{x} = [x_i]$ and $\mathbf{s} = [s_{ij}]$ are the first-stage variables, and second stage consists of variables $\mathbf{f} = [f_{it}]$. The master problem [MP] is

$$\text{[MP]:} \quad \max \quad 0 + z(\mathbf{x}, \mathbf{s}) \tag{A.1}$$

$$\text{s.t.} \quad \textit{Benders Optimality Cuts}$$

$$\sum_{\substack{i \in I: \\ b_{ij}=1}} s_{ij} = 1 \qquad \forall j \in I, \tag{A.2}$$

$$\sum_{i \in I} x_i \leq C \tag{A.3}$$

$$x_i \in \mathbb{Z}^+, s_{ij} \in \{0, 1\} \qquad \forall i \in I, j \in I \tag{A.4}$$

The optimal solution to [MP] is an upper bound to the original problem [M3] and $z(\mathbf{x}, \mathbf{s})$ is the optimal solution to the subproblem [SP] given $(\overline{\mathbf{x}}, \overline{\mathbf{s}})$

$$[\text{SP}]: \quad \max \quad 1 - \frac{\sum_{i \in I} \sum_{t \in \Theta} f_{it}}{D} \tag{A.5}$$

$$\text{s.t.} \quad f_{it} \geq \sum_{\substack{j \in I: \\ b_{ij}=1}} m_{ij} A_{jt} \overline{s}_{ij} - \overline{x}_i \quad i \in I, t \in \Theta, \qquad [\mu_{it}] \tag{A.6}$$

$$f_{it} \geq 0, \qquad \forall i \in I, t \in \Theta. \tag{A.7}$$

where [.] corresponds to dual variable for constraint (A.6). The optimal solution to [SP] provides an upper bound to the lower bound to the original problem [M3]. Note that subproblem [SP] further into sub subproblems for each GPI-QTY $i \in I$ and scenario $t \in \Theta$ as

$$[\text{SP}]_{it} \quad \min \quad f_{it} \tag{A.8}$$

$$\text{s.t.} \quad f_{it} \geq \sum_{\substack{j \in I: \\ b_{ij}=1}} m_{ij} A_{jt} \overline{s}_{ij} - \overline{x}_i \qquad [\mu_{it}] \tag{A.9}$$

$$f_{it} \geq 0, \tag{A.10}$$

Let f^*_{it} be the optimal solution to $[\text{SP}]_{it}$, then the optimal solution to [SP] is $1 - \frac{\sum_{i \in I} \sum_{t \in \Theta} f^*_{it}}{D}$. To solve sub subproblem $[\text{SP}]_{it}$, we take its dual

$$[\text{DSP}]_{it} \quad \max \left(\sum_{\substack{j \in I: \\ b_{ij}=1}} m_{ij} A_{jt} \overline{s}_{ij} - \overline{x}_i \right) \mu_{it} \tag{A.11}$$

$$\text{s.t.} \quad \mu_{ijt} \leq 1, \tag{A.12}$$

$$\mu_{it} \geq 0 \tag{A.13}$$

which is trivial to solve. The optimal solution $\mu^*_{it} = 1$ if $\sum_{\substack{j \in I: \\ b_{ij}=1}} m_{ij} A_{jt} \overline{s}_{ij} - \overline{x}_i > 0$, else $\mu^*_{it} = 0$. Note that since the subproblem [SP] is always feasible for a given $(\mathbf{x}, \mathbf{s})$, we do not need to add feasibility cuts (extreme rays) to the master problem [MP]. Let $\mathcal{E}_{it}$ be the set of the extreme points

to $[\text{DSP}]_{it}$. The master problem could be written as

$$[\text{MP}]: \quad \max \quad 1 - \frac{\sum_{i\in I}\sum_{t\in\Theta} z_{it}}{D} \tag{A.14}$$

$$\text{s.t.} \quad z_{it} \geq \left(\sum_{\substack{j\in I:\\ b_{ij}=1}} m_{ij} A_{jt} s_{ij} - x_i \right) \overline{\mu}^e_{it} \quad i \in I, t \in \Theta, e \in \mathcal{E}_{it}, \tag{A.15}$$

$$\sum_{\substack{i\in I:\\ b_{ij}=1}} s_{ij} = 1 \qquad \forall j \in I, \tag{A.16}$$

$$\sum_{i\in I} x_i \leq C \tag{A.17}$$

$$x_i \in \mathbb{Z}^+, s_{ij} \in \{0,1\} \qquad \forall i \in I, j \in I \tag{A.18}$$

Note that the set of extreme points $\mathcal{E}_{it} = \{0,1\}$. For $e = 0$, $\overline{\mu}_{ijt} = 0$, and Constraint (A.15) is $z_{it} \geq 0$ which corresponds to the nonnegativity constraint (4.2.15) in the original formulation [M3]. On the other hand, when $e = 1$, $\overline{\mu}_{ijt} = 1$ and Constraint (A.15) is $z_{ijt} \geq \sum_{\substack{k\in J_i:\\ b_{ijk}=1}} m_{ijk} A_{ikt} s_{ijk} - x_{ij}$ corresponding to constraint (4.2.14). The approach is equivalent to a cutting plane algorithm where constraints (4.2.14) and (4.2.15) in the original model [M3] are dropped and added iteratively.

To warm-start the algorithm, nonnegativity constraints (A.15) corresponding to $e = 0$ are included in [MP]. To tighten the relaxation, we also add a set of valid inequality constraints

$$x_i \leq \sum_{\substack{i\in I:\\ b_{ij}=1}} d^{\max}_j s_{ij} \quad \forall \quad i \in I \tag{A.19}$$

where $d^{\max}_j$ is the maximum daily demand recorded for GPI j in the sales data. Constraint (A.19) ensures that GPI-QTY i is not stocked if $s_{ij} = 0 \forall j \in I$.

References

Aardal, K., Jonsson, Ö., and Jönsson, H. (1989). Optimal inventory policies with service-level constraints. *Journal of the operational research society*, 40(1):65–73.

Abdel-Aal, M. A., Syed, M. N., and Selim, S. Z. (2017). Multi-product selective newsvendor problem with service level constraints and market selection flexibility. *International Journal of Production Research*, 55(1):96–117.

Abdel-Malek, L., Montanari, R., and Morales, L. C. (2004). Exact, approximate, and generic iterative models for the multi-product newsboy problem with budget constraint. *International Journal of Production Economics*, 91(2):189–198.

Abdel-Malek, L. L. and Montanari, R. (2005). An analysis of the multi-product newsboy problem with a budget constraint. *International Journal of Production Economics*, 97(3):296–307.

Agarwal, Y., Mathur, K., and Salkin, H. M. (1989). A set-partitioning-based exact algorithm for the vehicle routing problem. *Networks*, 19(7):731–749.

Agrawal, R., Srikant, R., et al. (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499.

Ahiska, S. S., Gocer, F., and King, R. E. (2017). Heuristic inventory policies for a hybrid manufacturing/remanufacturing system with product substitution. *Computers & Industrial Engineering*, 114:206–222.

Aydin, G. and Porteus, E. L. (2008). Joint inventory and pricing decisions for an assortment. *Operations Research*, 56(5):1247–1255.

Bagchi, U. and Gutierrez, G. (1992). Effect of increasing component commonality on service level and holding cost. *Naval Research Logistics (NRL)*, 39(6):815–832.

Bassok, Y., Anupindi, R., and Akella, R. (1999). Single-period multiproduct inventory models with substitution. *Operations Research*, 47(4):632–642.

Bertsimas, D. and Thiele, A. (2006). A robust optimization approach to inventory theory. *Operations research*, 54(1):150–168.

Carøe, C. C. and Tind, J. (1998). L-shaped decomposition of two-stage stochastic programs with integer recourse. *Mathematical Programming*, 83(1-3):451–464.

Chand, S., Ward, J. E., and Weng, Z. K. (1994). A parts selection model with one-way substitution. *European Journal of Operational Research*, 73(1):65–69.

Chen, M. and Chuang, C. (2000). An extended newsboy problem with shortage-level constraints. *International Journal of Production Economics*, 67(3):269–277.

Choi, S., Ruszczyński, A., and Zhao, Y. (2011). A multiproduct risk-averse newsvendor with law-invariant coherent measures of risk. *Operations Research*, 59(2):346–364.

Deflem, Y. and Van Nieuwenhuyse, I. (2013). Managing inventories with one-way substitution: A newsvendor analysis. *European Journal of Operational Research*, 228(3):484–493.

Desaulniers, G., Desrosiers, J., Dumas, Y., Solomon, M. M., and Soumis, F. (1997). Daily aircraft routing and scheduling. *Management Science*, 43(6):841–855.

Dutta, P. and Chakraborty, D. (2010). Incorporating one-way substitution policy into the newsboy problem with imprecise customer demand. *European Journal of Operational Research*, 200(1):99–110.

Erlebacher, S. J. (2000). Optimal and heuristic solutions for the multi-item newsvendor problem with a single capacity constraint. *Production and Operations Management*, 9(3):303–318.

Ford Jr, L. R. and Fulkerson, D. R. (1958). A suggested computation for maximal multi-commodity network flows. *Management Science*, 5(1):97–101.

Fuller, J. B., O'Conor, J., and Rawlinson, R. (1993). Tailored logistics: the next advantage. *Harvard Business Review*, 71(3):87–98.

Gaur, V. and Honhon, D. (2006). Assortment planning and inventory decisions under a locational choice model. *Management Science*, 52(10):1528–1543.

Gilmore, P. C. and Gomory, R. E. (1961). A linear programming approach to the cutting-stock problem. *Operations research*, 9(6):849–859.

Gilmore, P. C. and Gomory, R. E. (1963). A linear programming approach to the cutting stock problempart ii. *Operations research*, 11(6):863–888.

Hadley, G. and Whitin, T. M. (1963). Analysis of inventory systems. Technical report.

HealthcareConference (2017). Medavail technologies inc. In *Cowen and Company 37 th Annual Health Care Conference Boston*.

Hsieh, C.-C. and Lai, H.-H. (2019). Pricing and ordering decisions in a supply chain with downward substitution and imperfect process yield. *Omega*.

Jammernegg, W. and Kischka, P. (2013). Risk preferences of a newsvendor with service and loss constraints. *International Journal of Production Economics*, 143(2):410–415.

Khouja, M. (1999). The single-period (news-vendor) problem: literature review and suggestions for future research. *Omega*, 27(5):537–553.

Kök, A. G. and Fisher, M. L. (2007). Demand estimation and assortment optimization under substitution: Methodology and application. *Operations Research*, 55(6):1001–1021.

Kök, A. G., Fisher, M. L., and Vaidyanathan, R. (2015). *Assortment Planning: Review of Literature and Industry Practice*, pages 175–236. Springer US, Boston, MA.

Lau, H.-S. and Lau, A. H.-L. (1996). The newsstand problem: A capacitated multiple-product single-period inventory problem. *European Journal of Operational Research*, 94(1):29–42.

Leachman, R. and Glassey, R. (1987). Preliminary design and development of a corporate level production planning system for the semiconductor industry. In *Robotics and Automation. Proceedings. 1987 IEEE International Conference on*, volume 4, pages 710–710. IEEE.

Lin, J. and Ng, T. S. (2011). Robust multi-market newsvendor models with interval demand data. *European Journal of Operational Research*, 212(2):361–373.

MedAvail (2017). medavail.com. *MedAvail*.

Moon, I. and Choi, S. (1994). The distribution free continuous review inventory system with a service level constraint. *Computers & industrial engineering*, 27(1-4):209–212.

Nahmias, S. and Schmidt, C. P. (1984). An efficient heuristic for the multi-item newsboy problem with a single constraint. *Naval Research Logistics (NRL)*, 31(3):463–474.

Oğuz, O. (2002). Generalized column generation for linear programming. *Management Science*, 48(3):444–452.

Pentico, D. W. (1974). The assortment problem with probabilistic demands. *Management Science*, 21(3):286–290.

Pentico, D. W. (1976). The assortment problem with nonlinear cost functions. *Operations Research*, 24(6):1129–1142.

Rajaram, K. and Tang, C. S. (2001). The impact of product substitution on retail merchandising. *European Journal of Operational Research*, 135(3):582–601.

Rao, U. S., Swaminathan, J. M., and Zhang, J. (2004). Multi-product inventory planning with downward substitution, stochastic demand and setup costs. *IIE Transactions*, 36(1):59–71.

Sadowski, W. (1959). A few remarks on the assortment problem. *Management Science*, 6(1):13–24.

Shin, H., Park, S., Lee, E., and Benton, W. (2015). A classification of the literature on the planning of substitutable products. *European Journal of Operational Research*, 246(3):686–699.

Slawsky, R. (2015). Kiosks in health care 101. *Network Media Group*.

Taleizadeh, A. A., Akhavan Niaki, S. T., and Hoseini, V. (2009). Optimizing the multi-product, multi-constraint, bi-objective newsboy problem with discount by a hybrid method of goal programming and genetic algorithm. *Engineering Optimization*, 41(5):437–457.

Taleizadeh, A. A., Niaki, S. T. A., and Hosseini, V. (2008). The multi-product multi-constraint newsboy problem with incremental discount and batch order. *Asian Journal of Applied Sciences*, 1(2):110–122.

Tryfos, P. (1985). On the optimal choice of sizes. *Operations Research*, 33(3):678–684.

Turken, N., Tan, Y., Vakharia, A. J., Wang, L., Wang, R., and Yenipazarli, A. (2012). The multi-product newsvendor problem: Review, extensions, and directions for future research. In *Handbook of newsvendor problems*, pages 3–39. Springer.

Vairaktarakis, G. L. (2000). Robust multi-item newsboy models with a budget constraint. *International Journal of Production Economics*, 66(3):213–226.

Waring, A. C. (2012). Risk-averse selective newsvendor problems. *PhD Dissertation*.

Wollmer, R. D. (1992). An airline seat management model for a single leg route when lower fare classes book first. *Operations research*, 40(1):26–37.

Zhang, B., Xu, X., and Hua, Z. (2009). A binary solution method for the multi-product newsboy problem with budget constraint. *International Journal of Production Economics*, 117(1):136–141.