Stochastic Optimization of Vaccine Vial Replenishment

Zahra Azadi[1], Harsha Gangammanavar[2], Sandra Eksioglu[1]

[1]Industrial Engineering Department of Clemson University

[2] Engineering Management, Information, and Systems Department of Southern Methodist University

{zazadi,seksiog}@clemson.edu

{harsha}@smu.edu

We present a two-stage stochastic linear programming model to manage the inventory of vaccine vials in developing countries. In these countries immunization programs often involve targeted outreach services in remote locations. Organizations managing these programs are in need of tools which identify inventory replenishment and vaccine utilization plans to minimize costs and wastage while achieving greater coverage level. Wastage incurs when doses are discarded from open vials after their safe use time. We propose a model which identifies optimal portfolios of different sized vaccine vials, and corresponding vial-opening decisions in the face of uncertainty in patient arrival. This model is a two-stage stochastic program which we solve using an extension of the stochastic Benders decomposition algorithm (L-shaped method). Computational results on instances built using data from Bangladesh reveal that our solution approach outperforms the standard L-shaped method. Experimental results are used to develop simple and economic vaccine administration policies usable by healthcare administrators.

Key words: Vaccine inventory replenishment; Stochastic programming; L-shaped; Perishable product; Vaccine open vial wastage

1. Introduction

The spread of infectious diseases has significantly reduced in the last century. This can largely be attributed to widespread immunity due to vaccination. Vaccination has helped to completely eradicate diseases such as smallpox, and has severely restricted diseases such as polio, measles, and tetanus. Immunization programs lead by World Health Organization (WHO) and United Nations Children's Emergency Fund (UNICEF) have been instrumental in these global eradication efforts. Global vaccination coverage, which accounts for the fraction of children who receive recommended vaccines, has steadily increased over the past several decades. According to WHO, "Immunization averts nearly 2-3 million deaths every year. An additional 1.5 million deaths could be avoided if global vaccination coverage improves" (WHO 2014a).

Achieving the target immunization levels has been particularly challenging in developing countries. For example, of the 18.7 million children who did not receive the diphtheria-tetanus-pertussis

(DTP3) vaccine worldwide in 2014, 60% live in just ten developing countries (WHO 2014a). Vaccine stock-out is one of the primary reasons for low coverage in these countries. For example, in 2014, 26% of WHO member countries experienced a national level vaccine shortage for at least one month (Subaiya 2015). Another reason for low coverage is vaccine drop-out or wastage, which is higher than 50% of the total vaccines distributed (WHO 2005). Several factors contribute to vaccine drop-outs, such as, exposure to extreme temperature, expiration, and physical damage. This can happen at different levels of supply chain including transportation, storage, and service delivery. According to a document published by theUNICEF (2010), "Highest vaccine wastage occurs at service delivery level as compared to other supply chain levels". The Global Alliance for Vaccines and Immunizations (GAVI) has requested that countries take measures to bring down their vaccine wastage rates (WHO 2005).

In many developing countries, immunization programs are conducted in health centers or through outreach sessions. For example, in Bangladesh 94% of immunization programs are conducted through such services (Guichard et al. 2010). These sessions are organized through programs such as Expanded Programme on Immunization (EPI) and delivered by trained healthcare workers. These health centers are often located in remote areas which lack proper transportation connectivity and amenities such as refrigeration. Due to this, these centers are limited in their vaccine storage capacity and replenishment frequency, and suffer from high wastage. For example, wastage at these centers is 27% for DPT, 61% for bacillus calmette-guérin (BCG), 3.5% for measles, and $< 1\%$ for the others (GAVI 2011, UNICEF 2010). Therefore, vaccine wastage reduction with low stock-outs is critical to deliver stable immunization programs in these developing countries.

Vaccines used for preventable diseases — such as BCG, measles, DTP — are distributed in single and multi-dose vials. In outreach sessions conducted by EPI, a restrictive open vial policy is adopted which requires that opened vials of multi-dose vaccines be discarded at the end of the day or six hours, whichever comes first (WHO 2014b). Discarded doses contribute to what is known as open vial wastage (OVW). While single-dose vials have zero OVW, they are more expensive than multi-dose vials. On the other hand, opening multi-dose vials may lead to wastage during low demand periods. In a recent study in Bangladesh, it was estimated that the OVW for BCG, measles, DTP, and TT were 85%, 71%, 44.2%, and 36.6%, respectively (Guichard et al. 2010). The decision to open a vial of particular size is often made before all the patients it could serve have arrived to the clinic. Since patient arrival is irregular and uncertain, these vial-opening decisions are made in an uncertain environment. In this case, an efficient inventory management plan is necessary which identifies the optimal portfolio of different sized vaccine vials, and the corresponding vial-opening decisions when the patient arrival is uncertain. This plan should aim to achieve maximum coverage while keeping the OVW under an acceptable level in a cost efficient manner.

1.1. Literature Review

The literature on vaccine supply chain management has grown in the recent years. Within this literature, the focus on OVW management started when vaccine vial wastage increased in outreach immunization sessions held under programs such as EPI in developing countries. The work conducted by Drain et al. (2003) was the first that discussed general differences in terms of cost, distribution, coverage, and safety between single- and ten-dose vaccine vials. A detailed survey of vaccine wastage in Bangladesh was provided in Guichard et al. (2010). The authors collected data about the usage and wastage of different vaccine types of varying sized vials in randomly selected outreach and non-outreach clinics. The aim of both these studies was to estimate vaccine wastage, and hence, they did not consider the associated cost. Such costs were considered in the analysis conducted by Parmar et al. (2010). The scope of these studies was restricted to analyzing existing vaccine administration practices and hence, they do not prescribe any plan for efficient inventory management. However, they recognize that "there is an urgent need for more rigorous and systematic wastage monitoring" (Parmar et al. 2010).

The following studies, on the other hand, have used simulation based approaches to assess the impacts of different vial sizes on OVW. The work by Lee et al. (2011) developed a simulation model for Thailand's Trang province to evaluate replacing ten-dose with single-dose vaccines. The results of this study showed that for a fixed arrival rate, the cost associated with wastage due to damage outweighed the cost savings of OVW reductions due to use of single-dose vials. This followed their earlier simulation work where patient arrival was modeled as Poisson distribution (Lee et al. 2010). The authors concluded that the suitable number of doses per vial may vary by region and patient arrival rate. However, the Poisson arrival assumption may not necessarily capture the local circumstances of the clinics, particularly under outreach immunization programs. A similar simulation study was carried out by Yang et al. (2014) to analyze the impact of different patient arrival rates on OVW for five- and ten-dose vaccine vials. In this case the authors used real-life data to estimate parameters for simulation. Since none of these studies consider combination of different vial sizes, they recommend using a particular vial size for a given arrival rate. These recommendations are based on cost estimates obtained using simulation. Once again these works do not prescribe any vaccine inventory management scheme.

In order to prescribe a vaccine managment plan, optimization models are necessary. Such approaches are limited and include deterministic mixed integer programming (MIP) and finite state probabilistic models. In Dhamodharan and Proano (2012) the total cost is minimized via optimal ordering decisions identified by solving a deterministic MIP integrated within a Monte Carlo simulation setup. The authors used the simulation model to generate patient arrivals and compute OVW under a fixed vaccine administration policy for a single vial size. The estimated OVW is then

used in the optimization model to find the optimal ordering decisions. Mofrad et al. (2014) proposed a Markov Decision Process (MDP) model to identify a vial-opening policy which minimizes the expected OVW cost and maximizes vaccine coverage over a finite horizon. The authors used the backward induction algorithm to solve their probabilistic model. This work is further extended in Mofrad et al. (2016) by integrating the MDP model with simulation to analyze the impact of session duration on the optimal policy obtained in their previous work. In both studies the authors determine the clinic closing time or the time when patients should be rejected. However, they do not consider loss in opportunity due to unserved patients. Moreover, they assume that there is enough supply of vaccines which is not always the case, for example in outreach clinics. All these studies address the inventory management of a single vial type and uncertainty in demand is characterized either probabilistically or through simulation. The optimization models focus either on ordering decisions or vial-opening decisions. However, these decisions are interrelated and therefore a model which considers both of them together, as well as uncertainty in patient arrivals will be very effective.

Recall that the vaccines expire after their safe-use time. Managing inventory replenishment decisions for perishable products with fixed shelf-life, which have an expiration date and are not to be consumed after this date, has been extensively studied in the supply chain literature, for example, Nahmias (1982), Raafat (1991), and Bakker et al. (2012). Such problems are referred to as the Economic Lot Sizing (ELS) problems. While some ELS models for perishable products with fixed shelf-life consider demand to be deterministic, others have considered stochastic demand (see Bakker et al. (2012) for a survey). Van Zyl (1963) and Nahmias and Pierskalla (1973) find optimal ordering policies when demand follows continuous and differentiable distributions. This is done by incorporating shortage and ordering costs in the cost function. Nahmias and Pierskalla (1973) also considered wastage (due to expiry) costs in their cost function. An extended m-period dynamic programming model was studied in both Fries (1975) and Nahmias (1975). The ELS for perishable products with fixed shelf-life have been solved using the well-known ordering policies such as (s, S) or (Q, r). A Markov renewal process is used to obtain an ordering policy in Liu and Lian (1999) using a closed-form cost function for an (s, S) policy with back-orders. Weiss (1980) shows the optimality of (s, S) policy when the demand process is compound Poisson. In such studies, the inventory replenishment model is restricted to follow a specific policy which turns out to be optimal only under certain demand distributions and cost functions.

Deriving optimal policies for stochastic ELS models through exact methods is computationally complex. This has resulted the development of number of heuristic policies. For example, Nahmias and Wang developed a myopic-based inventory policy using bounds on wastage cost (Nahmias and Wang 1979). Following this approximation method, two heuristic policies for a joint inventory

control and pricing model were developed in Chen et al. (2014). These policies perform well only when demand is non-stationary. Simulation based methods have also been employed to obtain approximate policies (see Van Donselaar and Broekmeulen (2012)). A combination of dynamic programming and simulation is used to reduce the state space of the stochastic ELS model in Haijema et al. (2009). However, optimization models which provide exact solutions for ELS problem have been limited to deterministic demand case (see Ekşioğlu and Jin (2006)). Stochastic demand is handled only through optimal policies under highly simplifying assumptions or sub-optimal heuristic policies under slightly general settings.

1.2. Our contributions

When demand uncertainties cannot be represented using standard distributions a simple probabilistic model is not effective. In this case, using optimization methods is preferred over obtaining approximate policies through simulation. Moreover, when vaccines are distributed in multiple vial sizes, ordering and vial-opening decisions, the inventory states, and their inter-dependencies become complicated. Such realistic requirements can be captured using stochastic programming (SP) models. Furthermore, SP provides effective tools to design algorithms to deal with such detailed models. In this regard, the contributions of our paper are as follows:

- *Vaccine vial replenishment model:* We propose a two-stage SP model for vaccine vial replenishment which captures (a) the order frequency and quantity for different sized vials, (b) the opening schedule for these vials, and (c) the administration of vaccines to patients. Our model considers the trade-off between OVW and inventory management costs, and accommodates uncertainty in patient arrival. Our model is an extension of stochastic ELS model which captures decisions at different time scales, involves separate inventory for different sized vials, and is independent of demand distribution.
- *Scalable algorithm:* We develop a new solution approach for two-stage stochastic linear programs (2-SLP) with continuous recourse. Our algorithm is motivated by the well-known stochastic Benders method. We use Gomory mixed integer (GMI) and mixed integer rounding (MIR) cuts to address the non-convexity of the first-stage problem. Our approach significantly reduces the computational requirement when compared to the standard method. Such computational enhancements allow our algorithm to be applied to real scale problems such as the vaccine vial replenishment model.
- *Optimization and simulation framework:* We design a computational framework which provides analytical support to immunization programs. We provide a detailed exposition to develop appropriate stochastic models for patient arrival using real-life data. These stochastic models are used to generate scenarios for the optimization model which is then solved using

our extended Benders' algorithm. The experimental results obtained using this framework help us to identify (a) an optimal portfolio of different sized vials for outreach centers located in different regions, (b) the necessary subsidies on purchase costs to achieve high immunization levels, and (c) an appropriate session length to increase coverage level and decrease OVW. These results are validated through an extensive simulation process.

- *Vaccine administration policies:* In addition to the base policy obtained using our vaccine vial replenishment model, we propose simple and economic vaccine administration policies for outreach clinics. These policies are motivated by our experimental results with the base policy and are effective for different regional settings. We establish their comparative effectiveness through a replicated simulation process.

In what follows, we present our stochastic programming model in §2 and discuss our solution approach in §3. We describe our data and the computational results in §4. Ideas regarding extensions of the current model and the solution approach are presented in §5.

2. Problem Formulation

In this section we present models that aid the design and management of immunization programs conducted through outreach clinics in developing countries. These outreach sessions target the infant population in rural areas and are held over a short period of time, usually lasting a few weeks. Due to the nature of these outreach clinics it is critical to identify replenishment schedules as well as vaccine administration policies. These should be accomplished by considering the uncertainty associated with patient arrivals to such clinics as well as the associated costs which comprise of purchasing, inventory holding, and OVW penalty costs. Vaccines are distributed in vials of different sizes with purchase costs proportional to vial sizes. Therefore, clinics maintain inventories of vials of different sizes and inventories of doses from opened vials. We propose an extension of the stochastic ELS model to capture the trade-offs that exist between purchasing and wastage cost, and between OVW and vaccination coverage in an uncertain environment. The schematic representation of this problem is presented in Figure 1.

2.1. Vaccine vial replenishment model

In our model we consider two major decisions: (a) inventory replenishment/ordering decisions, and (b) vial-opening/consumption decisions. The ordering decisions help us to identify the replenishment schedule and order size. These decisions are made on a weekly basis. On the other hand, the consumption decisions are a function of the vaccine administration policy which depends on patient arrival at a much finer timescale. We will use $\mathcal{T} = 1, \ldots, T$ to denote the ordering decision epochs, and each ordering period consists of N consumption decision epochs. Therefore, $\mathcal{N} = 0, \ldots, NT$ will denote all the consumption decision epochs over the planning horizon. We consider vials of

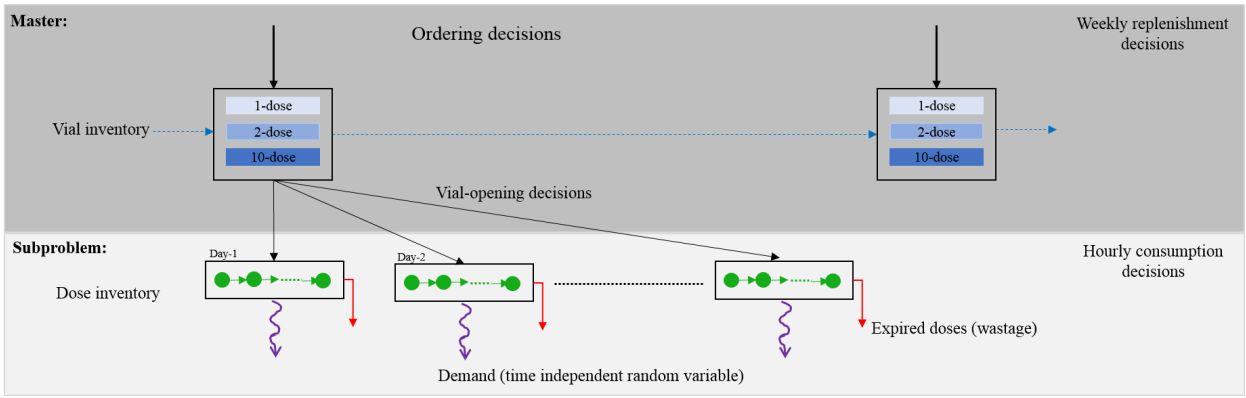


Figure 1 **Inventory dynamics in vaccine vial replenishment model.**

different sizes which are denoted by the set $\mathcal{V}$. Once the vials are opened they have to be consumed within their safe use time, we will use τ to denote this limit.

At $t \in \mathcal{T}$, ordering decisions z_t are made at a fixed cost of f_t. If an order is placed ($z_t = 1$), then the replenishment quantity for different sized vials is to be determined. We will denote these decisions by $r_{\nu t}$ and the corresponding variable purchase cost by $c_{\nu t}$ $\forall \nu \in \mathcal{V}$. At each consumption decision epoch $n \in \mathcal{N}$, vial-opening decisions $u_{\nu n}$ are made. Replenishment quantity decisions $r_{\nu t}$ and vial-opening decisions $u_{\nu n}$ together determine the state of vial inventory which we denote as $s_{\nu n}$ $\forall \nu \in \mathcal{V}$. The evolution of inventory for each vial size $\nu \in \mathcal{V}$ is captured by the following flow-balance equations:

$$s_{\nu Nt} = s_{\nu(Nt-1)} + r_{\nu t-1} - u_{\nu Nt} \qquad \forall t \in \mathcal{T}, \tag{1a}$$

$$s_{\nu n} = s_{\nu n-1} - u_{\nu n} \qquad \forall n \in \mathcal{N} \setminus \{N, 2N, \ldots, TN\}, \tag{1b}$$

where the initial inventory $s_{\nu 0}$ is assumed to be known. While (1a) captures the arrival of new orders $r_{\nu t}$ at a replenishment decision epoch, (1b) ensures that the flow is balanced at the remaining consumption epochs. The vials are held in the inventory at a holding cost of $d_{\nu t}$. Recall that the order quantity for each vial size is determined only if an order is placed in replenishment decision epoch $t \in \mathcal{T}$. This is ensured by the following inequality:

$$\sum_{\nu \in \mathcal{V}} r_{\nu t} \leq M_t z_t \qquad \forall t \in \mathcal{T}, \tag{2}$$

where M_t is the limit on the number of vials ordered and varies over time period t.

We will use a single decision vector $x \in \mathcal{X}^0 \subset \mathbb{Z}_+^{n_1}$ to collectively denote the ordering decision z_t, the replenishment quantity $r_{\nu t}$, vial-opening decisions $u_{\nu n}$, and the state variables $s_{\nu n}$, i.e., $x = (z, r, u, s)$. Note that, the decision vector x consists of binary (z_t) as well as pure-integer decisions. Moreover, all these decisions are required to be made before the realization of patient arrivals at the clinic.

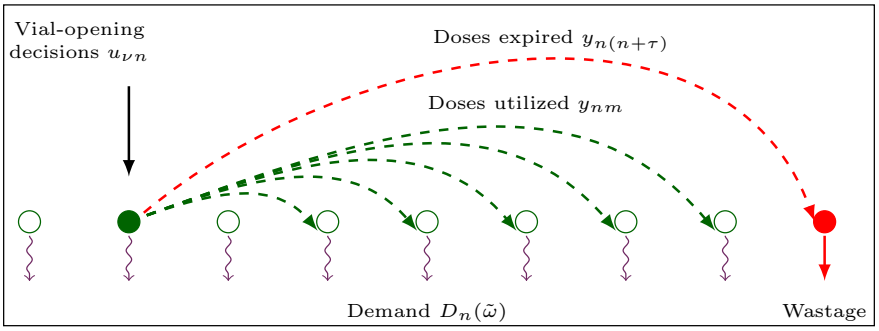


Figure 2 **Dose utilization for $n=2$ with $\tau=6$.**

The vials opened in consumption period n can be utilized in period $m < n+\tau$, or before the end of a session, whichever occurs first. We will use y_{nm} to denote the number of doses obtained from vials opened in period n, and used in period m. Using this notation, $y_{n(n+\tau)}$ represents the number of doses which are past their safe use time, and hence contribute to OVW. The relationship between opened vials of different sizes and the number of doses available is captured by the following equation:

$$\sum_{m=n}^{n+\tau-1} y_{nm} + y_{n(n+\tau)} = \sum_{\nu\in\mathcal{V}} q_\nu u_{\nu n} \qquad \forall n \in \mathcal{N}, \tag{3}$$

where q_ν is the number of doses in vial ν. The right-hand side of the above expression is the total number of doses obtained by opening vials in period n, while the left-hand side is the sum of utilized doses and expired doses. The doses are used to meet the demand over all consumption periods $n \in \mathcal{N}$. In this model we consider patient arrival to be a random variable denoted by $D_n(\tilde{\omega})$. If the total number of patient arrivals exceeds the available doses, then the system is penalized (denoted by ℓ_n) at a cost of p for each unserved patient. Each of the expired doses, on the other hand, is disposed at a fixed cost g. With these, the dose inventory must satisfy the following flow-balance equations:

$$\sum_{m=n-\tau+1}^{n} y_{mn} + \ell_n = D_n(\omega) \qquad \forall n \in \mathcal{N}. \tag{4}$$

Here ω is a realization of the random variable $\tilde{\omega}$. Note that the right-hand sides of equations (3) and (4) take only discrete values. The first term on the left-hand side is the total number of unexpired doses, and hence can be viewed as the state of the inventory of doses. Figure 2 demonstrates how the doses obtained from opened vials in period $n=2$ are utilized over $\tau=6$ consumption periods. The expired doses are represented by the arc in red.

The objective function denoted by $F(x)$ includes the fixed and variable purchasing costs, the inventory holding costs, and the expected value of wastage and penalty costs. This objective function is given by:

$$F(x) = \sum_{t\in\mathcal{T}}(f_t z_t + \sum_{\nu\in\mathcal{V}} c_\nu r_{\nu t}) + \sum_{\nu\in\mathcal{V}}\sum_{n\in\mathcal{N}} d_\nu s_{\nu n} + \mathbb{E}\{h(x,\tilde{\omega})\}, \tag{5}$$

where $h(x,\tilde{\omega})$ captures the penalty due to unserved patients and disposal cost for expired doses. Note that, this cost depends on random patient arrivals ($D_n(\tilde{\omega})$) and is bounded from below by zero.

2.2. Formulation of the two-stage stochastic linear programming model

The vaccine vial replenishment problem can be written as a 2-SLP to capture the impact of the stochastic nature of patient arrival on vaccine inventory replenishment and vaccine dose consumption decisions. Since the decisions in x have to be made prior to realization of patient arrivals., they are non-anticipative in nature (Birge and Louveaux 2011). These form the first-stage decisions which construct the vaccine and dose inventory. On the other hand, the consumption decisions y_{nm} depend on patient arrivals, and therefore they are made in an adaptive manner. This allows us to write the replenishment model as:

$$\begin{aligned} &\min\ F(x) \\ &\quad s.t.\ (1),(2),\ \text{and}\ z_t\in\{0,1\},\ r_{\nu t}, u_{\nu n}, s_{\nu n},\in\mathbb{Z}^+ \qquad \forall n\in\mathcal{N}, t\in\mathcal{T}, \nu\in\mathcal{V}, \end{aligned} \tag{6}$$

where the recourse function $h(x,\omega)$ corresponding to the first-stage decisions x and realization ω of random variable $\tilde{\omega}$ is given by:

$$\begin{aligned} h(x,\omega) = \min\ &\sum_{n\in\mathcal{N}}(g y_{n(n+\tau)} + p\ell_n) \\ &s.t.\ (3),(4),\ \text{and}\ y_{nm}, \ell_n\in\mathbb{Z}^+ \qquad \forall n\in\mathcal{N}, m\in\{n,\ldots,n+\tau\}. \end{aligned} \tag{7}$$

In the literature (6) is referred to as the master problem and (7) as the subproblem. Notice that, the decisions in both the master problem (6) and the subproblem (7) are discrete variables. Also note that, uncertainty affects only the right-hand sides of equation (4) (patient arrival realization). The recourse matrix characterized by the left-hand sides in equations (3) and (4) and the transfer matrix characterized by the right-hand side of equation (3) are independent of randomness. Therefore, the above 2-SLP has a fixed recourse (Birge and Louveaux 2011).

Proposition 1 *The linear programming solutions of subproblem* (7) *are integers.*

Proof: Subproblem (7) is an uncapacitated minimum cost network flow model. Therefore, the recourse matrix satisfies the total unimodularity property (Ahuja et al. 1988). In other words, since the recourse matrix is integer, every basic feasible solution to the linear relaxation of subproblem (7) is integer for any integer vector $(\sum_{\nu\in\mathcal{V}} q_\nu u_{\nu n}, D_n(\omega)), \forall n \in \mathcal{N}$. which appears on the right-hand side. ■

To summarize, our proposed model is a 2-SLP where some decisions are made before the realization of uncertain parameters, and certain recourse actions can be taken in response to their actual realization. These decision variables are divided into two groups (See figure (1)): the non-anticipative decisions belong to the first-stage (master problem) and adaptive decisions belong to the second-stage (subproblem). In our model both the first- and second-stage variables are required to take integer values. However, the subproblem satisfies total unimodularity, and therefore solutions obtained by solving its linear relaxation satisfy the integrality requirement. Moreover, the recourse matrix is not affected by randomness and hence our 2-SLP model has fixed recourse.

3. Solution Approach: An extended L-shaped method

Since its conception in Beale (1955) and Dantzig (1955), 2-SLP has been used to model many applications in the fields of financial planning, capacity expansion, manufacturing, resource allocation, etc. Over the past several decades different algorithms have been proposed to tackle these problems. To achieve computational tractability, many of these methods represent uncertainty through a finite number of realizations (scenarios). If $\Omega = \{\omega_1, \omega_2, \ldots, \omega_S\}$ represents this finite set of scenarios with respective probabilities $p_i,\ i = 1, \ldots, S$, then the expectation function in (6) can be written as:

$$\mathbb{E}\{h(x, \tilde{\omega})\} = \sum_{i\in\Omega} p_i h(x, \omega_i). \tag{8}$$

In this case, we can always formulate the full deterministic equivalent linear program (LP) which is often termed as the extensive scenario formulation (ESF). Even with a moderate number of realizations the ESF can become quite large. Recall that our model is a network flow model with fixed charge cost. It is known that solving these problems is NP-Hard (Hochbaum and Segev 1989), and therefore solving the ESF using standard LP methods is beyond the current computational capabilities.

Decomposition based methods, notably the stochastic Benders decomposition (Van Slyke and Wets 1969), Dantzig-Wolfe decomposition (Dantzig and Wolfe 1960) and progressive hedging (Rockafellar and Wets 1991), have proven effective in addressing this concern. These methods iteratively build approximations to the expected recourse function. Dantzig-Wolfe decomposition is

an inner linearization method in which the dual of the master problem is solved. However, this method is not applicable to the class of problems with discrete first stage variables, which is the case in our model. Progressive hedging is a primal-dual method where in each iteration a penalty is associated with a deviation from a feasible solution. However, using this method requires selecting an appropriate proximal parameter which is instance dependent and hard to discern (Watson and Woodruff 2011). Therefore, we base our solution approach on stochastic Benders decomposition which is also known as the L-shaped method.

Suppose that we can generate S realizations of random vector $\tilde{\omega}$ using Monte Carlo simulation techniques. We can view the set Ω to comprise of these simulated vectors, each having the same probability $p_i = 1/S$. With this simulated set we can write the sample average approximation (SAA) problem as follows:

$$\min_{x \in \mathcal{X}} \ \hat{F}_S(x) = c^\top x + \frac{1}{S} \sum_{i \in \Omega} h(x, \omega_i). \tag{9}$$

The second term in the objective function above is an unbiased estimator of the expectation function in (8). The SAA problem is a SP with discrete distribution and hence can be solved by the classical L-shaped method with a branch and cut procedure to recover the integrality of the master problem.

L-shaped method is an iterative method used to solve SP problems with continuous recourse. It is well known that in these problems the expected recourse function is piece-wise linear and convex (Birge and Louveaux 2011). Therefore, the subproblem dual solutions are used to build a lower bound/outer approximation of this expected recourse function in each iteration of L-shaped method. When subproblem variables are discrete valued then additional steps are necessary to achieve convexity (see Sen (2011) for details). Due to total unimodularity of our subproblem, employing such procedures is not necessary. In order to solve problems with binary constraints in the first stage, Laporte and Louveaux proposed the integer L-shaped method in 1993 (Laporte and Louveaux 1993). Later, improved optimality cuts were introduced by Hjorring and Holt (1999) to strengthen the method. This method was further accelerated by using local branching techniques by Rei et al. (2009). These methods work for problems with only first-stage binary variables and hence are not employable to our problem. In order to improve computational time of solving large mixed integer 2-SLPs with continuous recourse, valid integer cuts were used by Bodur et al. (2014) within L-shaped method. The authors evaluate the use of these cuts within the LP relaxation of Benders reformulation. Their computational results show an improvement in the performance of the L-shaped algorithm. The solution algorithm we develop is also an extension of the L-shaped method to solve SAA problems which include mixed integer variables in the first-stage. We address

the integrality by introducing valid integer cuts, particularly, we use GMI and MIR cuts which strengthen the Benders' cuts. In this regard, our approach resembles the work of Bodur et al. (2014) in which cuts are being added to the ESF.

Standard L-shaped method: The SAA problem of the form described in (9) can be solved using the standard L-shaped method. When the feasible region $\mathcal{X}^k$, which comprises of original set of constraints and a set of affine functions, is accessible, then, an approximation to the expected recourse function is represented by the maximum of the set of affine functions in $\mathcal{X}^k$. This can be accomplished using an auxiliary variable η. Notice that at the beginning of the algorithm the feasible region comprises of only the original constraints, that is $\mathcal{X}^0 := \{x \mid Ax = b\} \subset \mathbb{Z}^n_{1+} \times \mathbb{R}_+$. In iteration k the algorithm begins by identifying the solution x^k to the following mixed-integer LP:

$$\min \ \{c^\top x + \eta \mid (x, \eta) \in \mathcal{X}^k\}. \tag{M^k_{mip}}$$

Using this solution and a realization $\omega_i \in \Omega$ we solve the subproblem $h(x^k, \omega_i)$ to obtain the optimal dual solution. Let π^k_{in} and θ^k_{in} denote the optimal dual solutions to constraints (3) and (4) respectively. This procedure is enumerated for every $\omega_i \in \Omega$. Using the dual solutions, a lower bounding optimality cut is computed with the following affine form:

$$l^k_{opt}(x, \eta) := \frac{1}{S} \sum_{i \in \Omega} \left[\sum_{n \in \mathcal{N}} \left(\theta^k_{in} D_n(\omega_i) + \sum_{\nu \in \mathcal{V}} \pi^k_{in} q_\nu u_{\nu n} \right) \right] - \eta \leq 0. \tag{10}$$

Using this the feasible region is updated as:

$$\mathcal{X}^{k+1}(x) = \mathcal{X}^k(x) \cup (l^k_{opt}(x, \eta) \leq 0). \tag{11}$$

This completes an iteration of the L-shaped algorithm. Recall that the subproblem in (7) satisfies the relatively complete recourse assumption, that is, the subproblem is feasible for all $\omega_i \in \Omega$ and $x \in \mathcal{X}^0$. Finally, this algorithm is known to converge in a finite number of iterations if ω has a finite support (Laporte and Louveaux 1993).

Integer cuts: In the standard L-shaped method for problems with discrete first-stage variables, every iteration involves solving (M^k_{mip}) which is a mixed integer program. As the algorithm evolves, the number of affine functions used in the approximation increases and hence the size of the problem enlarges. This results in significantly high computational time even for a moderately sized problem. Therefore, exploring algorithms which address this concern are required to achieve scalability.

Adding integer cuts to the LP-relaxation of deterministic MILP has been proven to be effective (Higle and Sen 1999). These integer cuts are valid inequalities to the original problem, thus, they cut off some feasible points of LP-relaxation but not integer points. There are several families of valid inequalities to mixed integer LPs such as lift-and-project cuts, GMI cuts, MIR cuts, and split cuts. These cuts are briefly discussed here:

1. *Split cuts*: This set of cuts are a special case of disjunctive cuts. They are basically derived using split disjunctions (Balas 1998). Lift-and-project, GMI, and MIR cuts are special cases of split cuts (Cornuéjols 2008).
2. *Lift-and-project cuts*: These are a set of inequalities that are obtained from using the lift-and-project approach and are valid for binary programs. The original lift-and-project method is proposed in (Balas et al. 1993). This method uses a higher dimensional representation to formulate an LP which results in some valid cuts (Balas et al. 1993).
3. *GMI cuts*: This set of cuts are generated using integer rounding on a given row of the LP tableau for a basic integer variable with non- integer value (Gomory 1963). GMI closures are equivalent to split cuts (See the proof in Balas et al. (1993)). The numerical experiments in the literature indicate that a split closure is tighter than a GMI closure. However, optimization over a split disjunction is NP-hard (Cornuéjols 2008).
4. *MIR cuts*: These cuts are generated using integer rounding on the coefficients of integer variables and the right-hand side of a constraint (Wolsey and Nemhauser 2014). These are equivalent to GMI cuts if generated from simplex tableau rows (Wolsey and Nemhauser 2014).

These integer cuts are computed based on the solutions obtained by solving the corresponding LP-relaxation, and are shown to improve the convergence rate of the L-shaped method (Laporte et al. 2002).

In iteration k, the master problem in (M^k_{mip}) can be viewed as a deterministic MIP. Furthermore, any integer cut computed for this master problem remains a valid inequality even when the approximation is updated using (10).

Extended L-shaped:

As in the case of standard L-shaped method, we assume that we have access to the current feasible region denoted by $\overline{\mathcal{X}}^k$ which includes the original set of constraints and a set of affine functions. Unlike the standard method, in our algorithm we solve the LP-relaxation of the master problem. Therefore, in addition to the lower bounding approximations, the set $\mathcal{X}^k$ includes inequalities to enforce integrality. The master problem is given by:

$$\min \ \{c^\top x + \eta \mid (x,\eta) \in \overline{\mathcal{X}}^k\}. \qquad (\overline{M}^k_{mip})$$

Let (x^k, η^k) and v^k_S denote the optimal solution and objective function value obtained by solving the LP-relaxation of the above optimization problem (denoted as $\overline{M}^k_{lp}$), respectively. Using the current solution x^k, an affine lower bounding function $l^k_{opt}(x,\eta)$ is computed using (10).

In each iteration, a fractional portion of $\mathcal{X}$ is cut off, thereby resulting in a tightened feasible region. This is accomplished by adding GMI and MIR cuts which are constructed based on x^k.

Note that, these integer cuts also effect the lower bounding approximation. Notice that, the set $\overline{\mathcal{X}}^k$ consists of linear inequalities which can be modified to include slack variables, and are represented in a general form as: $\{(x,\eta) \in \mathbb{Z}_+^{n_1} \times \mathbb{R}_+ : \sum_{j=1}^{n} a_j x_j + g\eta = b\}$. We will use this general constraint form to briefly present the construction of these cuts and refer the reader to Cornuéjols (2008) for a detailed exposition.

The GMI cuts are generated for fractional basic solutions to $(\overline{M}_{lp}^k)$. Therefore, multiple GMI cuts can be generated in iterations with a fractional basic solution. A GMI cut corresponding to the fractional basic solution x_j^k can then be constructed as:

$$l_{gmi}^{k,j}(x,\eta) := 1 - \sum_{j:f_j \le f_0} \frac{f_j}{f_0} x_j - \sum_{j:f_j > f_0} \frac{1-f_j}{1-f_0} x_j - \sum_{g>0} \frac{g}{f_0}\eta + \sum_{g<0} \frac{g}{1-f_0}\eta \le 0, \tag{12}$$

where $f_0 = b - \lfloor b \rfloor$ and $f_j = a_j - \lfloor a_j \rfloor$. Let $\mathcal{L}_{gmi}^k$ denote the set of all GMI cuts generated in iteration k.

On the other hand, MIR cuts are generated only for constraints that have mixed integer variables. In our model the original constraints ((1), (2)) have only integer variables. The affine function added in (10), however, introduces the continuous variable η. Therefore, MIR cuts are applicable only for these constraints. Moreover, the observation that the recourse function $h(\cdot)$ is lower bounded by zero, almost surely, plays critical role in deriving MIR inequalities. The MIR cut generated for affine function $j = 1, \ldots, k$ in iteration k has the following form:

$$l_{mir}^{k,j}(x,\eta) := \sum_{j=1}^{n} (\lfloor a_j \rfloor + \frac{(f_j - f_0)^+}{1-f_0} x_j + \frac{1}{1-f_0} \sum_{g<0} g\eta - \lfloor b \rfloor \le 0, \tag{13}$$

where f_0 and f_j are introduced earlier, and $(f)^+ = \max(0, f)$. The set of MIR inequalities added in iteration k is denoted by $\mathcal{L}_{mir}^k$.

The lower-bounding affine function l_{opt}^k, a set of GMI cuts $\mathcal{L}_{gmi}^k$ and a set of MIR cuts $\mathcal{L}_{mir}^k$ constructed in iteration k are added to the master problem to obtain the updated feasible region:

$$\overline{\mathcal{X}}^k = \overline{\mathcal{X}}^{k-1} \cap \{l_{opt}^k(x,\eta) \le 0\} \cap \{l_{gmi}^{k,j}(x,\eta) \le 0\}_{j \in \mathcal{L}_{gmi}^k} \cap \{l_{mir}^{k,j}(x,\eta) \le 0\}_{j \in \mathcal{L}_{mir}^k}. \tag{14}$$

This concludes one iteration of our algorithm.

Recall that v_S^k is the optimal objective function value to $(\overline{M}_{lp}^k)$ and therefore provides a lower bound (LB) to the original problem at x^k. On the other hand, $\hat{F}_S(x^k)$ computed from (9) is an unbiased estimate of the objective function value, and hence is the upper bound (UB) to the problem at x^k. Let, K denote the first iteration when the relative gap between $\hat{F}_S$ and v_S^k falls below a predefined error ϵ. In this iteration, the optimal solution to $(\overline{M}_{lp}^k)$ is given by x^K. If $x^K \in \mathbb{Z}_+^{n_1}$ then the algorithm is terminated and x^K is declared as the optimal solution, i.e., $\hat{x}_S = x^K$ with the

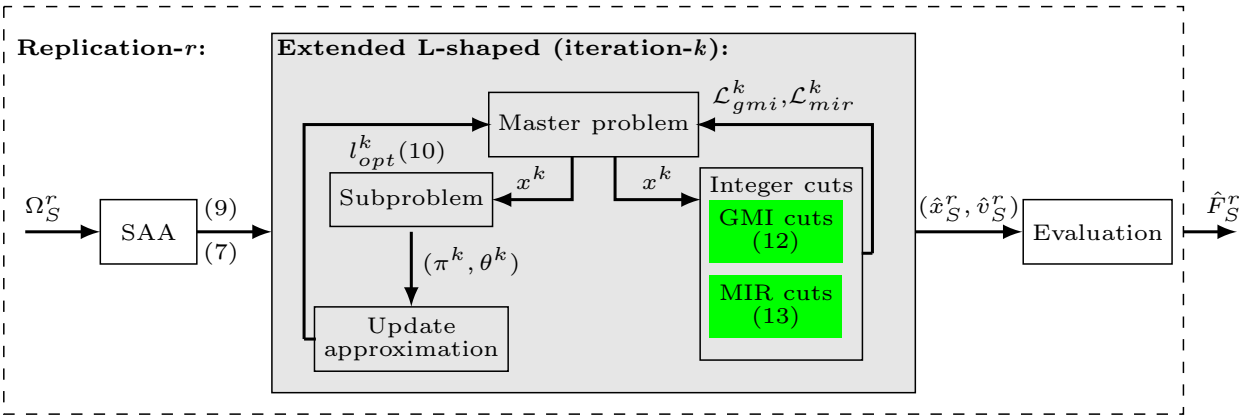


Figure 3 **Schematic representation of extended L-shaped algorithm.**

corresponding objective function value $\hat{v}_S = v^K_S$. If $x^K \notin \mathbb{Z}^{n_1}_+$ we solve $(\overline{M}^k_{mip})$ to obtain an integer optimal solution $\hat{x}_S$ and value $\hat{v}_S$.

Note that $(\hat{x}_S, \hat{v}_S)$ is the optimal solution-value pair to the SAA problem in (9). However, the sample size S required to ensure that $\hat{x}_S$ is an optimal solution to the true problem is problem specific. One can increase this probability of success by replicating the process with different sets of scenarios. The idea of replication has been used in the SP literature to provide statistical performance guarantees (Mak et al. 1999, Kleywegt et al. 2002). For a given selection of S, let Ω^r_S denote the set of scenarios and let $(\hat{x}^r_S, \hat{v}^r_S)$ be the optimal pair obtained using the extended L-shaped algorithm in replication $r = 1, \ldots, R$. In this regard, an estimator of the optimal objective function value of the true problem is given by $\bar{v}^R_S = \frac{1}{R}\sum_{r=1}^R \hat{v}^r_S$. Note that $\hat{v}^r_S$ is a LB and therefore $\bar{v}^R_S$ is a biased estimator of the objective function value.

Furthermore, in each replication we evaluate the quality of the solution $\hat{x}^r_S$ by simulating the subproblem using scenarios different from the ones in Ω^r_S. We term this process as out-of-sample evaluation. This evaluation is continued until when $(1-\alpha)$ confidence interval is built on $\hat{F}^r_{S'}$ where $(S' \neq S)$ is the number of scenarios used in this evaluation process. Consider the estimator $\Delta_r = \hat{F}^r_S(\hat{x}^r_S) - \hat{v}^r_S$ of the optimality gap. The mean and variance of this estimator are given by:

$$\overline{\Delta}^R = \frac{1}{R}\sum_{r=1}^R \Delta^r, \quad \sigma^2_\Delta = \frac{1}{R(R-1)}\sum_{r=1}^R (\Delta^r - \overline{\Delta}^R)^2. \tag{15}$$

The above estimator overestimates the optimality gap and therefore, can be viewed as a pessimistic gap. We will report this gap as well as the distribution of $\hat{F}^r_{S'}$ over replications in our computational study.

To summarize, our algorithm solves a linear relaxation of the master problem in every iteration as opposed to MILP in the standard approach. This significantly reduces the computational time, as we will see in the next section. GMI and/or MIR cuts are sequentially added to the master

Table 1 Cost and weight of different vaccine vial sizes of Pentavalent.

Vaccine size	Purchase cost per dose ($)			Vaccine weight per dose (gr)		
	Lower limit	Mean	Upper limit	Lower limit	Mean	Upper limit
1 dose	3.24	3.60	3.96	1.542	1.713	1.885
2 dose	3.25	3.50	3.75	1.723	1.914	2.106
10 dose	2.00	1.80	2.20	3.169	3.522	3.874

problem to enforce integrality. Our computational setup also includes a replicated out-of-sample procedure to obtain an estimate of optimiality gap. The schematic representation of our algorithm is illustrated in Figure 3.

4. Computational Experiments: A case study for Bangladesh

In this section we describe our computational experience of solving the vaccine replenishment model. We built instances of this model using real data available for different regions in Bangladesh. These instances are built to highlight the impact of data (deterministic as well as stochastic) on model decisions. Such analyses helps us draw insights on the vaccine vial replenishment problem and design effective policies for outreach clinics in developing countries. Analysis of large SPs require an effective algorithm such as the one we presented in §3. The policy obtained from our extended L-shaped algorithm is used as a benchmark in our computational study. Therefore, we first provide an analysis of our approach in terms of solution quality and the computational time compared to the standard methods. This section is organized into four categories: first, we explain the data input used for the instances; second, we present the setup used for the experiments; third, we analyze the vaccine vial replenishment problem via different sets of experiments; and finally, we discuss different heuristic policies and provide a comparison between them.

4.1. Data input model

We used two different data sources in our case study to develop instances for an outreach clinic conducted by EPI in Bangladesh. Our model requires deterministic data which is known *a priori* to the decision makers, as well as stochastic data which is uncertain and beyond the control of decision makers. The first set of data includes purchase and inventory holding costs. This data was complied using 2012 WHO vaccine volume calculator (WHO 2012). The second set of data is based on the Demographic and Health Survey (DHS) of Bangladesh. This is a population-based survey executed by National Institute for Population Research and Training in 2011 (NIPORT 2011) and includes individual household level DHS data. The details of deterministic and stochastic parameters are described next.

4.1.1. Deterministic parameters: The EPI services in rural areas are scheduled on a weekly basis (GAVI 2011). Therefore, in our experiments we consider a clinic which operates over a 2 weeks horizon and 5 working days per week. We allow the operating hours to vary between 2 to 8 hours a day. While the ordering decisions are made at the beginning of each week, the vials are opened hourly. We consider the Pentavalent vaccine which is distributed in single-dose, two-dose, and ten-dose vials. However, our setup can be used for other pediatric vaccines as well. The Pentavalent vaccines has a 6 hours safe use time (i.e. $\tau = 6$). The 2012 WHO vaccine volume calculator (WHO 2012) provides the data for vial purchase costs and corresponding weights. These are summarized for Pentavalent vaccine in Table (1). The inventory holding cost was estimated based on the weight of vaccines. In our experiments, we use the mean of purchase and inventory holding costs. The ordering cost is assumed to be constant over the horizon and is set to \$10. Recall that in our model there is a penalty associated with unserved patients, we use a constant penalty cost of \$100. This penalty reduces the number of patients not served and thereby ensures high vaccine coverage level. Finally, OVW penalty cost is set to a value greater than the highest purchase cost for all experiments.

4.1.2. Stochastic parameters: Recall that the only stochastic element in our model is patient arrival which follows a stochastic process $\{D_n\}_{n\in\mathcal{N}}$. This stochastic process was characterized using the DHS data set. This data set includes both demographic information as well as interviews with families from each region. The interviews provide information about the history of vaccinations for children under 5 years. We used each child vaccination date to count the number of observations in each day for one year time interval. This data was used to get an estimate of the distribution associated with the random variable D_n, the number of patients arriving at the clinic every day. We conducted goodness-of fit tests to find the best distribution for our data. We assume that the outreach clinics are open only on weekdays and therefore, only weekday data is considered for this purpose. The results of goodness-of fit tests are summarized in Table (2). Our model is targeted for outreach clinics which are held temporarily by EPI where vaccines are administered over a short period of time. In order to obtain an estimate of patient arrival to such clinics, we consider demand aggregated over a certain period of time which depends on the frequency at which these programs are held. We assume that these clinics are held annually and hence we aggregate the target infant population. The number of vaccines depends on the number of births (NB), the population size (PS), and the number of healthcare facilities (HF). In order to forecast the annual vaccine requirement (AVR) we use a linear regression model with NB, PS, and HF as independent variables. We used the data from 2008 to 2015 for each independent variable. The coefficients and intercept of this regression model were estimated using the least square method.

Table 2 Data used to predict the daily demand distribution of Pentavalent through different regions in Bangladesh.

Region	Number of facilities	Infant population	Daily fitted distribution	Hourly estimated distribution
Barisal	41	8,147,000	NB(0.65,0.43)	NB(5,0.43)
Chittagong	141	288,079,000	NB(0.53, 0.81)	NB(286, 0.81)
Dhaka	171	46,729,000	NB(2.08, 0.16)	NB(2, 0.16)
Khulna	70	15,563,000	NB(0.95, 0.33)	NB(4, 0.33)
Rajshahi	182	18,329,000	NB(0.84, 0.31)	NB(2, 0.31)
Rangpur	43	15,665,000	NB(5, 0.93)	NB(327, 0.93)
Sylhet	62	9,807,000	NB(0.78,0.38)	NB(5,0.38)

The $p-$value for each independent variable coefficient revealed that only the intercept, number of births, and population size are significant at the 5% significance level. The regression model is of the form: $AVR \sim -529,720 + 0.0044PS + 0.9408NB$. The AVR is forecasted using this model and is proportionally distributed to each region based on regional infant population size described in Table (2). We assume that the infant population is uniformly distributed over the outreach clinics in that region. Furthermore, we also assume that the arrival rate at any clinic is uniform over its operational horizon. The last column of Table (2) shows the estimated arrival rates at each pediatric clinic.

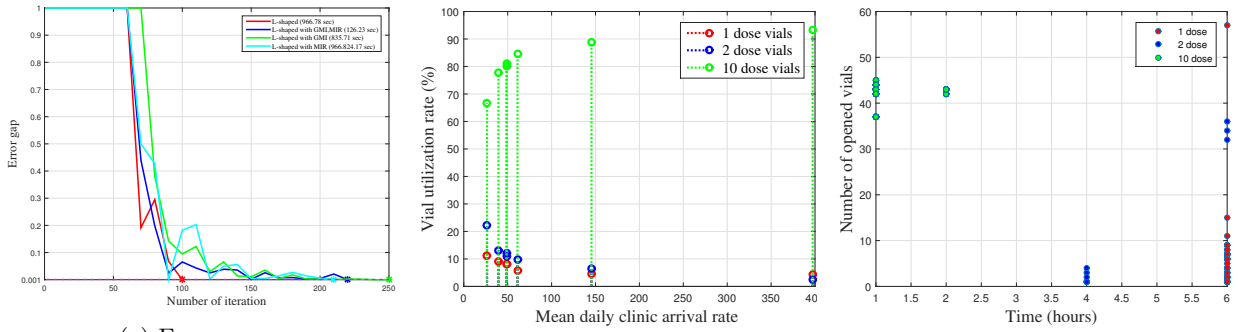
Scenario selection: Selecting an appropriate number of scenarios in the SAA formulation is critical. A very large number of scenarios results in a computationally unwieldy instance. On the other hand, a very small number of scenarios may not appropriately represent the uncertainty. In order to determine the appropriate number of scenarios we implement two different sets of analysis which assess the difference of the optimal solutions ($\hat{x}^S$ and $\hat{x}^{S'}$) obtained using S and S' scenarios. The first set of analysis captures the impact of the optimal solution on the recourse cost. This is done by fixing the optimal solution and simulating the recourse function over the two sets of scenarios and performing the paired t-test on the function values. The $p-$values for each region are presented in Table (3). A $p-$value greater than 0.05 implies that the two solution $\hat{x}^S$ and $\hat{x}^{S'}$ have a statistically indistinguishable impact on the recourse cost. The second set of analysis is based on the Euclidean norm of the two solutions. The results are shown in Table (3). A high value indicates that the solutions are significantly different. As can be seen in Table (3), $\hat{x}^{1000}$ and $\hat{x}^{2000}$ have acceptable $p-$values and relative Euclidean distances. For all regions considered in this study, increasing the number of scenario from 1000 to 2000 is not necessary and therefore, the number of scenarios considered in building our instances is set to $S = 1000$.

4.2. Experimental setup

Our solution algorithm was implemented in C programming language on a 64 bit Intel Core i5 processor @ 1.9 GHz with 8 GB RAM. All linear and mixed integer programs were solved using

Table 3 Paired t-test between different number of scenarios.

	$p-$value (95% CI)							$\|\hat{\mathbf{x}}^{\mathbf{S}} - \hat{\mathbf{x}}^{\mathbf{S}'}\| / \|\hat{\mathbf{x}}^{\mathbf{S}}\|$						
S; S′	Rajshahi	Barisal	Khuluna	Sylhet	Dhaka	Rangpur	Chittagong	Rajshahi	Barisal	Khuluna	Sylhet	Dhaka	Rangpur	Chittagong
100; 200	0.99	0.00	0.00	0.00	0.14	0.00	0.44	0.30	0.07	0.04	0.06	0.02	0.02	0.05
200; 500	0.99	0.84	0.29	0.80	0.03	0.08	0.01	0.33	0.05	0.01	0.05	0.03	0.02	0.05
500; 1000	0.99	0.67	0.89	0.35	0.48	0.00	0.96	0.22	0.05	0.03	0.05	0.02	0.02	0.02
1000; 2000	0.99	0.74	0.66	0.99	0.58	0.39	0.36	0.00	0.00	0.00	0.00	0.00	0.00	0.00



(a) Error gaps. (b) Vial portfolio. (c) Daily consumption schedule.

Figure 4 Base policy: analysis of solution algorithm and parameters.

Table 4 Comparison of solution quality.

Algorithm	LB ($)	UB ($)	CPU time (s)	No. of iter.	No. of LP	No. of MILP
LS	$21,571.7 \pm 12.1$	$22,047.9 \pm 33.8$	859.7 ± 245.9	233.8 ± 11.2	0	232.7 ± 11.2
LSM	$21,574.9 \pm 11.8$	$22,048.9 \pm 30.7$	350.4 ± 86.1	836.3 ± 83.5	756.4 ± 79.3	78.9 ± 11.8
LSG	$21,571.9 \pm 12.1$	$22,047.1 \pm 33.6$	412.2 ± 160.1	837.9 ± 93.8	756.4 ± 88.5	80.5 ± 12.7
LSMG	$21,571.7 \pm 12.1$	$22,047.9 \pm 33.9$	137.2 ± 47.3	431.1 ± 26.3	335.6 ± 16.6	94.5 ± 12.4

CPLEX callable subroutines. Our experiments begin by first finding an optimal or near optimal first stage solution (gap < 0.01) using the extended L-shaped method. In the posterior analysis the quality of these first stage solutions is evaluated by fixing them and simulating the subproblems (7) or (17). The scenarios used for optimization as well as the scenarios used for evaluation are generated from the same distribution. This evaluation is terminated when the standard deviation of the recourse function values is within a tolerable limit. Recall that this process is replicated using independent samples, all our experiments are replicated over $R = 30$. We present all our results in terms of the empirical distribution over replications. Heuristic policies are also treated in a similar manner.

4.3. Base policy

We begin by verifying the performance of our solution algorithm and the model adequacy. We term the policy identified by solving the vaccine replenishment model using our extended L-shaped method proposed in §3 as the base policy. We provide results for our base policy and compare them with those obtained using the standard L-shaped method.

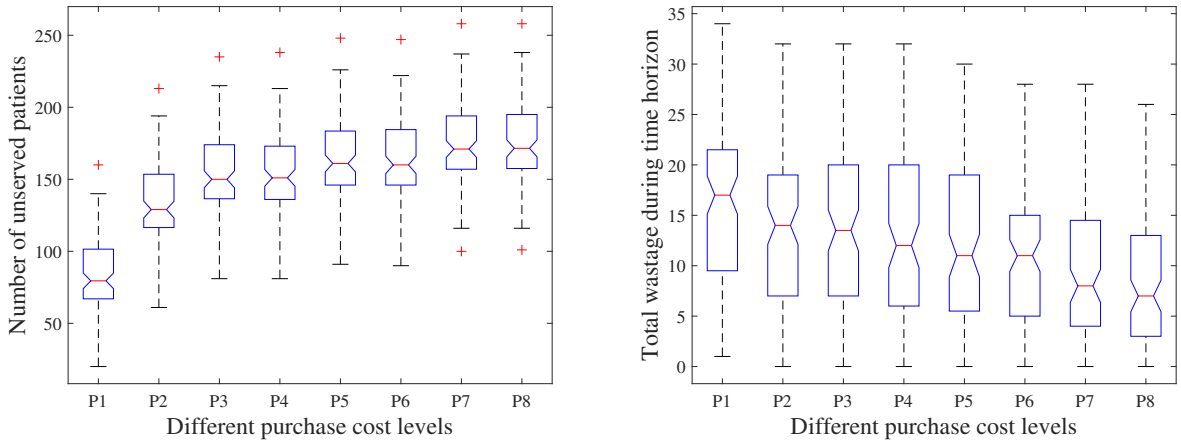
4.3.1. Performance of solution algorithms This set of experiments is implemented using Chittagong instance with one vial size (10 dose Pentavalent). The goal is to evaluate the impact of MIR and GMI cuts on the performance of the extended L-shaped method. We compared the quality of solutions obtained under the following conditions: (a) generating only MIR cuts (LSM), (b) generating only GMI cuts (LSG), and (c) generating both MIR and GMI cuts (LSMG). We compared these solutions with the standard L-shaped method (LS). In Table (4) we present the solution quality of each algorithm in terms of the CPU time, the number of iterations, LB, UB, the number of LPs solved, and the number of MILPs solved over 30 replications. Note that the values of LB and UB for all algorithms are similar and the LS method uses the least number of iterations. However, in every iteration of LS an MILP is solved and therefore the CPU time is significantly high. In algorithms which require integer cuts, MILPs are solved only when the error gap is below a threshold $\epsilon = 0.1\%$ and $x \notin \mathbb{Z}_+^{n_1}$. Therefore, the master problem is solved as a LP in most iterations (particularly in the beginning). This results in a higher number of iterations but smaller total computational time.

Furthermore, in our experiments we observed that using MIR cuts resulted in a greater decrease in computational time as compared to using GMI cuts. The greatest saving in computational time (by a factor 6.2) was observed when both types of integer cuts were incorporated. Figure (4a) illustrates the error gaps over algorithm iterations, the total number of iterations, and the CPU time for one of the replications. For algorithms with integer cuts the figure also shows the iterations where MILPs are solved (when $\epsilon \leq 0.001$).

This analysis establishes that the LSMG algorithm provides solutions which are similar to those obtained from LS method in significantly less CPU time. Such computational enhancements are necessary to satisfy the need to replicate SPs. This also allows our algorithm to be applied to real scale instances such as the one considered in our study.

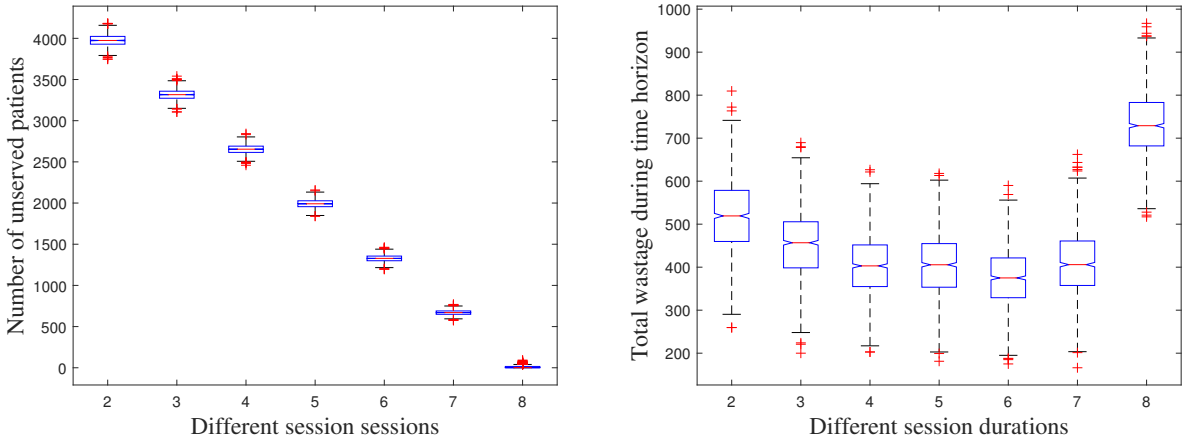
4.3.2. Analysis of parameters The solutions to vaccine vial replenishment model are sensitive to parameters such as patient arrival, purchase cost, and duration of a session. Analysis of these parameters help healthcare policy makers to understand the effectiveness of such outreach clinics and design appropriate policies. In the first set of experiments we evaluate the impact of different patient arrivals on vial portfolio selection. The second experiment is implemented to study the effect of varying purchase cost on OVW and number of unserved patients. Finally, we analyze the impact of the session duration on OVW and number of unserved patients.

1. ***Patient arrival***: This experiment was conducted using regions with different patient arrival rates. For example, the patient arrival rate in Rajshahi is lowest since this is the least populated region, and the patient arrival rate of Chittagong is highest since this is the most populated



(a) Unserved patients (b) Wastage

Figure 5 Effect of varying purchase cost on dose utilization.



(a) Unserved patients in Chittagong (b) Wastage in Chittagong

Figure 6 Impact of session duration on OVW and number of unserved patients in Chittagong

region. Figure (4b) depicts the relationship between the number of vial sizes ordered and the patient arrival rate. Going from Rajshahi to Chittagong, the arrival rate increases and so does the number of 10-dose vials. This is attributed to the low cost per dose of multi-dose vials. At high arrival rates, the single-dose vials are ordered to supplement the multi-dose vials, particularly to meet end of the day arrivals. Based on our results in Figure 4b it is economically efficient to procure larger size vials for moderately and highly populated regions. One could expect that the usage of single and two dose vials increases in clinics with small patient arrival rates, such as, outreach centers in remote locations.

2. ***Purchase cost***: In order to conduct these experiments we developed 8 problem instances, P1-P8. Each instance has the same patient arrival rate as Rajashahi, but, different purchase cost. We used the lower, mean, and upper limit of purchase costs presented in Table (1) in order

to build these instances. Instance P1 has the lowest purchase costs. These costs increase from P1 to P8. This set of experiment is replicated once. Figure 5a and 5b present the unserved patients and OVW observed during the evaluation step of one of the replications. As purchase cost increases, the vial-opening decisions become conservative in order to reduce wastage. This in turn increases the demand lost. This analysis provides a tool for policy makers to identify the necessary subsidies on vaccine purchase costs to achieve a certain level of immunization. For example, when costs were set to their lower limit in instance P1 ($c_1 = 1.6, c_2 = 1.5, c_{10} = 1$) the total number of unserved patients was less than 100 for 99% of scenarios tested (as indicated by the top whisker), while it increased to 175 for P3-P4, and was as high as 190 for P8.

3. ***Session duration***: The choice of session duration affects the number of patients who can attend these sessions. Since vaccines expire after their safe use time the choice of session duration will help reduce the OVW. In order to evaluate the impact of session duration on OVW, we created instances for different regions with session duration varying between 2 to 8 hours. When the session duration is less than 8 hours, we assume that the patients, who cannot attend the session during operational hours, are lost or unserved. As expected, a shorter session results in an increase in the number of unserved patients (see Figure 6a). This is because we lose more patients during the time interval between the intended session length and 8 hours. The figure also indicates that the OVW is minimum when the session duration is equal to the safe use time ($\tau = 6$ hours). When the session duration is short, even unexpired doses contribute to OVW. However, more vials will be open when the session duration is longer than 6 hours resulting in additional unused, unexpired doses. While increasing the session duration decreases the number of unserved patients, the need to open new vials to serve patients arriving in later hours impacts OVW. Note that, increasing the session duration also increases the operational costs which were not considered in our model.

Based on our analysis of the base policy, the usage of multi-dose vials with complementary single-dose vials for use in last operating hours (when $t > \tau$), is recommended for highly populated and well connected regions such as Chittagong. Demographic data about Chittagong indicates that the majority of the population is employed in jobs which are less flexible and therefore longer sessions are more beneficial. Furthermore, due to better connectivity, orders can be placed frequently, thus, the high OVW due to longer sessions is less critical. On the other hand, Rajshahi region is mostly rural and a majority of the population is employed in agricultural sector (Islam and Hassan 2011). This provides greater flexibility and hence having a 6 hours session duration is appropriate to achieve high coverage and low OVW. Our model and subsequent analysis is targeted for developing countries and hence the purchase cost for these vaccines is a critical parameter. In addition to reducing cost, the purchase cost also impacts the vaccine coverage and OVW. Therefore, healthcare

policy makers should negotiate necessary subsidies to achieve vaccination target in a cost efficient manner. Our setup provides not only an optimal policy but also a systematic tool to conduct such analyses.

4.4. Heuristic policies

Solving the vaccine vial replenishment model to optimality and obtaining a base policy as described in §(4.3) requires extensive engineering tools. These tools are commercially expensive and are rarely accessible to healthcare administrators in developing countries. In order to address this, we developed and evaluated four simple vaccine administration policies using our understanding of the base policy. These policies are: (a) first multi-dose, last single-dose (FMLS) policy, (b) shorter duration session (SDS) policy, (c) single vial size policy, and (d) wait to open (WO) policy. These policies impact both, order replenishment and administration of vaccines. Thus the master problem in (6) and the subproblem in (7) are appropriately modified to accommodate these policies. These policies are described below.

a *FMLS*: The principle of this policy follows the observation made in Figure 4c regarding optimal consumption schedule. This policy suggests that healthcare practitioners open multi-dose vials in the early hours of the session and single-dose vials as session comes to an end. To implement this, we divide the session length into three equal intervals such that for $t \in \mathcal{T}$ (a) 10 dose vials are opened in hours $\mathcal{N}_1 = \{Nt-5, Nt-4\}$, (b) 2-dose vials are open in hours $\mathcal{N}_2 = \{Nt-3, Nt-2\}$, and (c) 1-dose vials are open in the last two hours $\mathcal{N}_3 = \{Nt-1, Nt\}$. Mathematically, this corresponds to setting $u_{\nu n} = 0$ for $n \in \mathcal{N} \setminus \mathcal{N}_i$ when $q_\nu = i$ in (6). For example, for time periods $n \in \mathcal{N}_3$ when only 10-dose vials are opened, $u_{\nu n} = 0$ if $q_\nu = 1$ or 2.

b *SDS*: In Figure 6 we observed that 6 hours session length has the minimum OVW and longer sessions (8 hours) reduce the number of unserved patients. Therefore, if we encourage people arriving in hours 7 and 8 to attend the session in the first 6 hours, we can decrease the number of unserved patients while keeping OVW under control. The SDS policy achieves low OVW by trimming duration of the immunization session from 8 to 6 hours and moving certain percentage of patients arriving in hours $7-8$ uniformly to earlier hours. This effectively increases the arrival rate in hours 1-6 which is accommodated by modifying the data generation process used in setting up the subproblem (7).

c *Single vial size*: Clinics often use vials of a single size which reduces the flexibility in vial-opening and thereby increases the overall cost. Such single vial size policies have received significant attention in literature (Dhamodharan and Proano 2012, Guichard et al. 2010). We will analyze these policies to highlight the advantage of using vials of different sizes. We implement the single vial policy using vials of size 1, 2, or 10, by adjusting the set $\mathcal{V}$ in the master problem and subproblem.

d *WO*: Certain clinics do not start vaccine administration until a sufficient number of patients have arrive to the clinic (Bosu et al. 1997). This provides flexibility in vaccine administration and therefore has the potential of reducing OVW. However, if the waiting time is too long, a large number of patients may leave without vaccination. This adversely effects vaccine coverage. In order to study such a policy, we present an extension to our vaccine vial replenishment model. This model captures the impact of patient waiting time on vaccine vial utilization decisions while considering trade-offs among OVW, unserved patients, and waiting time.

The inclusion of patient waiting time affects only the subproblem in (7). In order to capture this, we introduce a new variable to the formulation. Let w_{jn} be the number of patients who arrived in period j and were served in period n. The patients arriving in period j can be served in period $j \leq n < j + L$ where L is the maximum waiting time in hours. Patients not served within this time window cost the system. These costs are represented by a non-linear function $v(j-n)$ which depend on the waiting time $j-n$, and unit cost v. Figure 7 illustrates how vaccines are administered to patients arriving in time period n. Mathematically, the flow-balance equation (4) will be changed to the following constraints:

$$\sum_{m=n-\tau+1}^{n} y_{mn} = \sum_{j=n-L+1}^{n} w_{jn} \qquad \forall n \in \mathcal{N}, \tag{16a}$$

$$\sum_{j=n}^{n+L-1} w_{nj} + \ell_n = D_n(\omega) \qquad \forall n \in \mathcal{N}. \tag{16b}$$

The left-hand side of (16a) represents the inventory on hand and the right-hand side represents the total number of patients who have arrived within the last L periods. The inventory consists of doses from vials opened within the last τ periods. For example, when $L = 2$ the doses can be used for patients who arrived in periods $(n-2)$, $(n-1)$, and n as shown in Figure (7). The patients arriving in current time period $D_n(\omega)$ can be served only over the next L hours. Patients waiting longer than L hours are considered to be lost. These are captured by the terms on the left-hand side of equation (16b) and the incoming arcs (see Figure 7). We can now rewrite the recourse function $h(x, \omega)$ as follows:

$$\begin{aligned} h(x,\omega) = \min \; & \sum_{n\in\mathcal{N}} (g y_{n(n+\tau)} + p\ell_n + \sum_{j=n+1}^{n+L} v(j-n) w_{nj}) \\ s.t. \; & (3), (16) \\ & y_{nm}, \ell_n, w_{nj} \in \mathbb{Z}^+ \qquad \forall n \in \mathcal{N}, m \in \{n, \ldots, n+\tau+1\}, j \in \{n, \ldots, n+L\}. \end{aligned} \tag{17}$$

In our experiment the maximum waiting time was set to $L = 2$ hours. Also, the waiting time was penalized at $v = 0.24\$$ per time period which is the average hourly wage in Bangladesh.

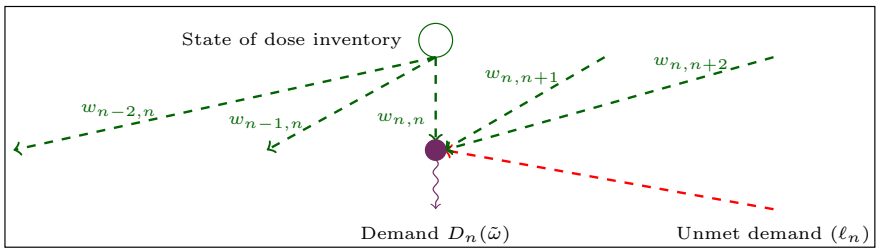
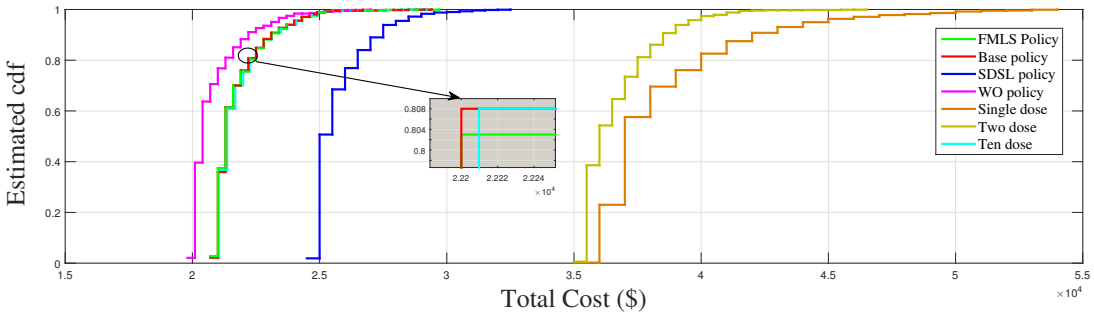
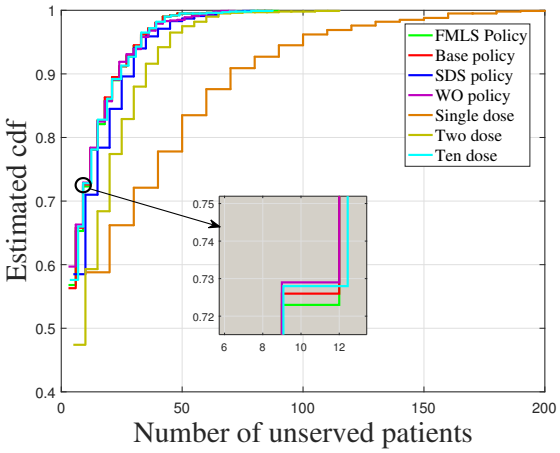


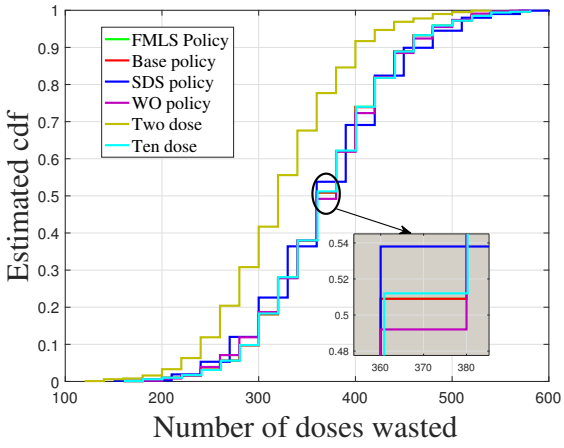
Figure 7 **Dose utilization and demand waiting for $n = 2$ with $L = 2$.**



(a) Total cost.



(b) Unserved patients.

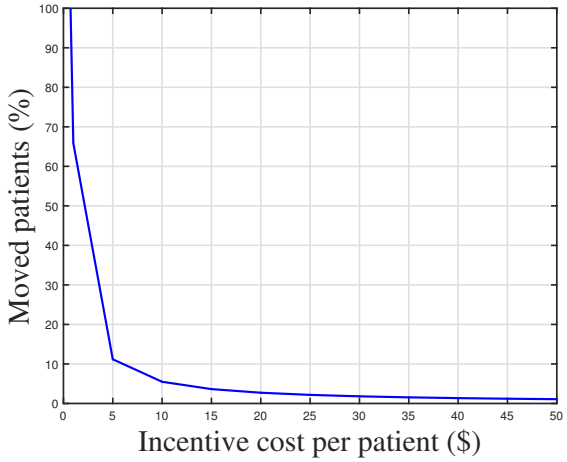


(c) OVW.

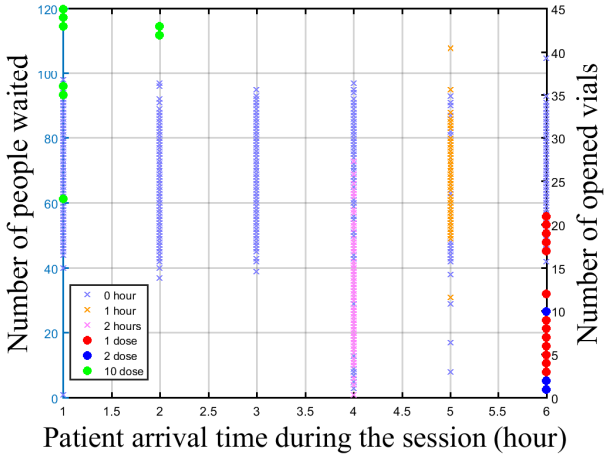
Figure 8 **Comparison of heuristic policies.**

4.4.1. Analysis and comparison In order to evaluate the performance of the policies presented above we compare their total costs, OVW, and number of unserved patients with the base policy. Figure 8 illustrates the cumulative distribution function (CMF) of the total cost, the number of unserved patients, and the OVW for the average of 30 replications.

1. *FMLS*: Figure 8a shows that the probability of achieving a certain cost target under FMLS policy is only slightly lower than the base policy. For example, the probability that the total cost is less than $2.22E4$ is 0.81 for base policy and 0.8 for FMLS policy. Moreover, the results in figures (8b) and (8c) indicate that the number of unserved patients and OVW are comparable



(a) Incentive cost versus the percent of people who attend.



(b) The hourly consumption schedule and waited people.

Figure 9 SDS and WO analysis results.

to the base policy. This can be attributed to the similarity in vial-opening decisions under base policy and the vial-opening windows set under FMLS policy (see Figure 4c). However, the FMLS policy mandates using single-dose vials even when multiple patients arrive in the last time window when using multi-dose vials is more economical. In any case FMLS policy is easy to implement and fairly economical.

2. *SDS*: The SDS policy increases the effective arrival rate in the 6 hours time window by allowing a fraction of patients, originally scheduled to attend in hours $7-8$, to attend in the first 6 hours. This increased arrival rate increases the vaccine requirement, but it has a marginal impact on OVW and the fraction of arrived patients left unserved. The greater vaccine requirement and the number of unserved patients results in a higher cost, which can be seen in Figure 8a. The results for SDS policy in Figure 8 were obtained for instances with incentive cost set to 1\$ and 65% of patients rescheduled from hours $7-8$. An increase in the fraction of patients rescheduled improves the vaccination coverage, however this may require some incentives to motivate patients to reschedule. It makes economical sense if the total amount of incentives paid and the increased purchase cost is lower than the penalty associated with unserved patients. Figure 9a shows the fraction of patients who should be encouraged to reschedule at different unit incentive costs. For example, if we pay an incentive of \$4 to each patient arriving during the $7th$ and $8th$ hours, the maximum percent of people that can be rescheduled economically is equal to 20%. Note that, as the incentive cost increases, only a small fraction of patients can be rescheduled while ensuring that the clinic does not lose money.
3. *Single vial size*: When each region is required to choose a single vial size, then, the choice of vial size would depend on its patient arrival rate. For regions with higher patient arrival rates,

it is beneficial to use 10-dose vials as it is more cost efficient. Furthermore, the use of smaller sized vials increases the number of unserved patients, as it becomes more economical to reject patients when compared to purchasing new vials. On the other hand, for regions with lower rate, 1- and 2-dose vials are economical as 10-dose vials result in higher unutilized doses, and hence contribute to OVW. These observations were captured in our experimental analyses. Figure 8 summarizes these observations for a region with high arrival rate (Chittagong). The figure shows that, with probability 0.95, the number of unserved patients is as high as 45 for 2-dose policy, while it increases to 100 for 1-dose policy, however, the use of such lower-dose policies decreases OVW (it is in fact 0 for 1-dose policy). Since the penalty on unserved patients is higher than OVW penalty, the total cost for 10-dose policy is lower than 2- and 1-dose policies. In fact, the probability of having a cost less than $2.2E4$ is 0.7 for 10-dose policy, while this probability is 0.81 for the optimal policy. This indicates that the 10-dose policy is near optimal for this region.

4. *WO*: This policy provides flexibility in how vaccines are administered by allowing patients to wait. While under the base policy patients are turned away when the opened vials are exhausted, under WO policy, patients can be served over a longer (waiting) window. This reduces the number of unserved patients significantly (see Figure 8b). Furthermore, this also allows for utilizing larger sized vials without exposing the clinics to OVW, as seen in Figure 8c. Due to reduction in the number of unserved patients as well as OVW the total cost of WO policy also reduces as shown in Figure 8a, when compared to the base policy. Moreover, as can be seen in Figure 9b, WO policy suggests that the healthcare practitioners use 10-dose vials for patients arriving in the early hours of the session while not making them wait, and open smaller sized vials at the end of the session while making them wait.

In order to compare the above policies with our base policy, we conducted the two-sample t-test on results obtained over 30 replications. These results are summarized in Table (5). The second column shows the estimated difference in total cost between the base and heuristic policies. In this table, the negative numbers indicate that the base policy has a lower total cost. The table also shows the 95% CI for difference in objective function values and the corresponding $p-$values. A $p-$value greater than 0.05 indicates that the null hypothesis, which states that the objective function values are statistically indistinguishable from one another, cannot be rejected at 95% confidence level. As can bee seen, the WO policy is the only policy that has a lower total cost when compared to the base policy. Moreover, the $p-$value in the last column indicates that the total costs of FMLS policy is statistically indifferent from the total cost of the base policy. This is further corroborated by the fact that 95% CI for estimated difference includes zero. However, the

Table 5 Two-sample t-test for differences between the base and heuristic policies in Chittagong.

Two-sample t-test	Estimated difference	Minimum difference	Maximum difference	95% CI for difference	$p-$value
Base, FMLS	-10.42	-82.19	-0.05	$[-26.45, 0]$	0.198
Base, SDS	-3918.80	-4111.46	-3813.67	$[-3946.00, -3891.60]$	0.000
Base, WO	42.97	2.52	62.91	$[26.99, 58.96]$	0.000
Base, 1 dose	$-15,840.10$	$-15,942.90$	$-15,786.20$	$[-15,861.20, -15,819.00]$	0.000
Base, 2 dose	$-14,855.10$	$-14,953.80$	$-14,802.80$	$[-14,875.90, -14,834.30]$	0.000
Base, 10 dose	-26.5	-99.12	-0.16	$[-46.9, -6.0]$	0.012

null hypothesis is rejected for other policies establishing that they are significantly different from the base policy.

Based on our analysis, we recommend that clinics implement the WO policy. It is clear that this policy achieves the highest coverage at the least cost. However, implementation of this policy must take into consideration the availability of waiting space in clinics and demographic preferences (e.g. willingness to wait). The success of FMLS policy depends on the selection of appropriate windows for opening vials of different sizes, which requires extensive experiments. When such experiments can be conducted, the ease of implementing FMLS policy can be garnered in a combination of FMLS and WO policies. The SDS policy has the most favorable impact on vaccine coverage which can be harnessed by designing appropriate incentive programs. The additional cost incurred due to these incentives are justified when tangible benefits of eradication are taken into consideration. Finally, in regions where vaccines are distributed in single vial sizes, choosing a supplier who distributes vaccine in vial size proportion to regional arrival rate is critical.

5. Conclusion and Future Research

In this paper we presented a 2-SLP model for vaccine vial replenishment and usage for targeted immunization outreach clinics. To the best of our knowledge this is the first stochastic optimization model which simultaneously captures replenishment order and vial-opening/consumption decisions. This model minimizes replenishment costs and wastage due to unused doses at the end of the immunization session. Different from related works in the literature, we model the performance of mixed vial-sized vaccine replenishment decisions; and compare these decisions to single vial-sized replenishment decisions and other replenishment policies which are simple to implement. We extended the L-shaped method by incorporating GMI and MRI cuts in the master problem. Via an extensive numerical study we show that the proposed algorithm is scalable; and outperforms the L-shaped method by providing high quality solutions in much smaller CPU time. We developed a case study using real-life data from Bangladesh. Through these experiments, we observed how different parameters affect the system's behavior. We summarize our observations as:

1. Population size impacts decisions about the mix of vials used in a region. For example, use of multi-dose with complementary single-dose vials is recommended in highly populated regions.

2. Vaccine purchasing cost impacts the achievable immunization level given a budget limit. Thus, the models presented here aid policy makers to negotiate the necessary subsidies to achieve the target vaccine coverage levels.
3. Session length impacts replenishment costs, the number of patients served and OVW. Session lengths which equal vaccine safe use time do minimize the total cost. Short sessions do minimize costs and OVW in sparsely populated regions.

These observations motivated us to design vaccine administration policies that are simple and economic. The results demonstrated that WO policy has the least total costs and therefore is highly recommended. For high populated and well-connected regions, FMLS policy works well since it provides high vaccine coverage level at a lower cost. Moreover, for the regions where there is only access to a single vial size, one should select a size which is proportional to the regional patient arrival rate (see Table (5)).

We plan to extend this research in the following ways. First, since there is no clear guideline for determining the number of scenarios used, investigating applications of sequential sampling algorithms, such as two-stage stochastic decomposition (SD) method (Higle and Sen 1991) is necessary. The SD method was originally designed for 2-SLPs and does not require *a priori* selection of scenarios. This algorithm assumes that the master problem is linear. However, our model is an MILP, thus, we plan to extend this method to accommodate discrete decision variables in the first stage. Second, our proposed model identifies the inventory replenishment decisions of a single clinic. As part of our future work, we plan to extend this model to consider multiple clinics within a region. We expect that the clinics will coordinate their decisions about inventory and operating hours in order to minimize costs and OVW. Third, we plan to develop an extension of the proposed model to aid replenishment decisions in clinics which handle vaccines of different types that have different safe use times, such as, liquid with 28 days and lyophilized with 6 hours.

References

2005. Monitoring vaccine wastage at country level: guidelines for programme managers .

2012. Vaccine volume calculator. Tech. rep., World Health Organization and others.

2014a. Global immunization data. Tech. rep., WHO world Health Organization and others.

Ahuja, Ravindra K, Thomas L Magnanti, James B Orlin. 1988. Network flows. Tech. rep., DTIC Document.

Bakker, Monique, Jan Riezebos, Ruud H Teunter. 2012. Review of inventory systems with deterioration since 2001. *European Journal of Operational Research* **221**(2) 275–284.

Balas, Egon. 1998. Disjunctive programming: Properties of the convex hull of feasible points. *Discrete Applied Mathematics* **89**(1) 3–44.

Balas, Egon, Sebastián Ceria, Gérard Cornuéjols. 1993. A lift-and-project cutting plane algorithm for mixed 0–1 programs. *Mathematical programming* **58**(1-3) 295–324.

Beale, Evelyn ML. 1955. On minimizing a convex function subject to linear inequalities. *Journal of the Royal Statistical Society. Series B (Methodological)* 173–184.

Birge, John R, Francois Louveaux. 2011. *Introduction to stochastic programming*. Springer Science & Business Media.

Bodur, Merve, Sanjeeb Dash, O Günlück, James Luedtke. 2014. Strengthened benders cuts for stochastic integer programs with continuous recourse. Tech. rep., Technical Report. Optimization Online 2014-03-4263.

Bosu, William K, Doris Ahelegbe, Emmanuel Edum-Fotwe, Kobina A Bainson, Paa Kobina Turkson. 1997. Factors influencing attendance to immunization sessions for children in a rural district of ghana. *Acta tropica* **68**(3) 259–267.

Chen, Xin, Zhan Pang, Limeng Pan. 2014. Coordinating inventory control and pricing strategies for perishable products. *Operations Research* **62**(2) 284–300.

Cornuéjols, Gérard. 2008. Valid inequalities for mixed integer linear programs. *Mathematical Programming* **112**(1) 3–44.

Dantzig, George B. 1955. Linear programming under uncertainty. *Management science* **1**(3-4) 197–206.

Dantzig, George B, Philip Wolfe. 1960. Decomposition principle for linear programs. *Operations Research* **8**(1) 101–111.

Dhamodharan, Aswin, Ruben A Proano. 2012. Determining the optimal vaccine vial size in developing countries: a monte carlo simulation approach. *Healthcare Management Science* **15**(3) 188–196.

Drain, Paul K, Carib M Nelson, John S Lloyd. 2003. Single-dose versus multi-dose vaccine vials for immunization programmes in developing countries. *Bulletin of the World Health Organization* **81**(10) 726–731.

Ekşioğlu, Sandra Duni, Mingzhou Jin. 2006. Cross-facility production and transportation planning problem with perishable inventory. *Computational Science and Its Applications-ICCSA 2006*. Springer, 708–717.

Fries, Brant E. 1975. Optimal ordering policy for a perishable commodity with fixed lifetime. *Operations Research* **23**(1) 46–61.

GAVI. 2011. Comprehensive multi-year plan of the national immunization program of bangladesh 2011 2016. Tech. rep., GAVI.

Gomory, Ralph E. 1963. An algorithm for integer solutions to linear programs. *Recent Advances in Mathematical Programming* **64** 260–302.

Guichard, Stephane, Karen Hymbaugh, Brent Burkholder, Serguei Diorditsa, Christine Navarro, Selina Ahmed, Mohd Mahbubur Rahman. 2010. Vaccine wastage in bangladesh. *Vaccine* **28**(3) 858–863.

Haijema, René, Nico van Dijk, Jan van der Wal, Cees Smit Sibinga. 2009. Blood platelet production with breaks: optimization by sdp and simulation. *International Journal of Production Economics* **121**(2) 464–473.

Higle, Julia L, Suvrajeet Sen. 1991. Stochastic decomposition: An algorithm for two-stage linear programs with recourse. *Mathematics of Operations Research* **16**(3) 650–669.

Higle, Julia L, Suvrajeet Sen. 1999. *Stochastic decomposition: a statistical method for large scale stochastic linear programming*, vol. 1. John Wiley & Sons, Inc.

Hjorring, Curt, John Holt. 1999. New optimality cuts for a single-vehicle stochastic routing problem. *Annals of Operations Research* **86** 569–584.

Hochbaum, Dorit S, Arie Segev. 1989. Analysis of a flow problem with fixed charges. *Networks* **19**(3) 291–312.

Islam, Md Rahedul, Md Zahidul Hassan. 2011. Land use changing pattern and challenges for agricultural land: A study on rajshahi district. *Journal of Life and Earth Science* **6** 69–74.

Kleywegt, Anton J, Alexander Shapiro, Tito Homem-de Mello. 2002. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization* **12**(2) 479–502.

Laporte, Gilbert, François V Louveaux. 1993. The integer l-shaped method for stochastic integer programs with complete recourse. *Operations Research Letters* **13**(3) 133–142.

Laporte, Gilbert, François V Louveaux, Luc Van Hamme. 2002. An integer l-shaped algorithm for the capacitated vehicle routing problem with stochastic demands. *Operations Research* **50**(3) 415–423.

Lee, Bruce Y, Tina-Marie Assi, Korngamon Rookkapan, Diana L Connor, Jayant Rajgopal, Vorasith Sornsrivichai, Shawn T Brown, Joel S Welling, Bryan A Norman, Sheng-I Chen, et al. 2011. Replacing the measles ten-dose vaccine presentation with the single-dose presentation in thailand. *Vaccine* **29**(21) 3811–3817.

Lee, Bruce Y, Bryan A Norman, Tina-Marie Assi, Sheng-I Chen, Rachel R Bailey, Jayant Rajgopal, Shawn T Brown, Ann E Wiringa, Donald S Burke. 2010. Single versus multi-dose vaccine vials: an economic computational model. *Vaccine* **28**(32) 5292–5300.

Liu, Liming, Zhaotong Lian. 1999. (s, s) continuous review models for products with fixed lifetimes. *Operations Research* **47**(1) 150–158.

Mak, Wai-Kei, David P Morton, R Kevin Wood. 1999. Monte carlo bounding techniques for determining solution quality in stochastic programs. *Operations Research Letters* **24**(1) 47–56.

Mofrad, Maryam H, Gian-Gabriel P Garcia, Lisa M Maillart, Bryan A Norman, Jayant Rajgopal. 2016. Customizing immunization clinic operations to minimize open vial waste. *Socio-Economic Planning Sciences* **54** 1–17.

Mofrad, Maryam H, Lisa M Maillart, Bryan A Norman, Jayant Rajgopal. 2014. Dynamically optimizing the administration of vaccines from multi-dose vials. *IIE Transactions* **46**(7) 623–635.

Nahmias, Steven. 1975. Optimal ordering policies for perishable inventoryii. *Operations Research* **23**(4) 735–749.

Nahmias, Steven. 1982. Perishable inventory theory: A review. *Operations research* **30**(4) 680–708.

Nahmias, Steven, William P Pierskalla. 1973. Optimal ordering policies for a product that perishes in two periods subject to stochastic demand. *Naval Research Logistics Quarterly* **20**(2) 207–229.

Nahmias, Steven, Shan Shan Wang. 1979. A heuristic lot size reorder point model for decaying inventories. *Management Science* **25**(1) 90–97.

NIPORT. 2011. Bangladesh demographic and health survey 2011. Tech. rep., NIPORT National Institute of Population Research and Training, Associates for Community and Population Research (ACPR), and ICF International.

Parmar, Divya, Elaine M Baruwa, Patrick Zuber, Souleymane Kone. 2010. Impact of wastage on single and multi-dose vaccine vials: Implications for introducing pneumococcal vaccines in developing countries. *Human Vaccines* **6**(3) 270–278.

Raafat, Fred. 1991. Survey of literature on continuously deteriorating inventory models. *Journal of the Operational Research society* 27–37.

Rei, Walter, Jean-François Cordeau, Michel Gendreau, Patrick Soriano. 2009. Accelerating benders decomposition by local branching. *INFORMS Journal on Computing* **21**(2) 333–345.

Rockafellar, R Tyrrell, Roger J-B Wets. 1991. Scenarios and policy aggregation in optimization under uncertainty. *Mathematics of Operations Research* **16**(1) 119–147.

Sen, Suvrajeet. 2011. Stochastic mixed-integer programming algorithms: Beyond benders' decomposition. *Wiley Encyclopedia of Operations Research and Management Science* .

Subaiya, Patrick Lydon Marta Gacic-Dobo Rudolf Eggers Laura Conklin, Laure Dumolard. 2015. Global routine vaccination coverage, 2014. Tech. rep., Centers for Disease Control and Prevention.

UNICEF. 2010. Vaccine wastage assessment. *National Rural Health Mission* .

Van Donselaar, Karel H, Rob ACM Broekmeulen. 2012. Approximations for the relative outdating of perishable products by combining stochastic modeling, simulation and regression modeling. *International Journal of Production Economics* **140**(2) 660–669.

Van Slyke, Richard M, Roger Wets. 1969. L-shaped linear programs with applications to optimal control and stochastic programming. *SIAM Journal on Applied Mathematics* **17**(4) 638–663.

Van Zyl, Gideon Johannes Jacobus. 1963. Inventory control for perishable commodities. Ph.D. thesis, University of North Carolina at Chapel Hill.

Watson, Jean-Paul, David L Woodruff. 2011. Progressive hedging innovations for a class of stochastic mixed-integer resource allocation problems. *Computational Management Science* **8**(4) 355–370.

Weiss, Howard J. 1980. Optimal ordering policies for continuous review perishable inventory models. *Operations Research* **28**(2) 365–374.

WHO, World Health Organization. 2014b. Who policy statement: Multi-dose vial policy (mdvp)-revision 2014. geneva, 2014. Tech. rep., WHO/IVB/14.07.

Wolsey, Laurence A, George L Nemhauser. 2014. *Integer and Combinatorial Optimization*. John Wiley & Sons.

Yang, Wanfei, Monika Parisi, Betsy J Lahue, Md Jasim Uddin, David Bishai. 2014. The budget impact of controlling wastage with smaller vials: A data driven model of session sizes in bangladesh, india (uttar pradesh), mozambique, and uganda. *Vaccine* **32**(49) 6643–6648.