Malaria Treatment Distribution in Developing World Health Systems and Application to Malawi

Abstract: Malaria is a major health concern among many developing countries. Developing strategies for efficient distribution of malaria medications, such as Artemesinin Combination Therapies (ACTs), is a key challenge in resource constrained countries. Using Malawi as a test case, this paper develops a solution that integrates strategic-level and tactical-level models to better manage pharmaceutical distribution through a three-tier centralized health system common to Sub-Saharan African countries. At the strategic level, we develop a two-stage stochastic programming approach to address the problem of demand uncertainty. In the first stage, before the malaria season starts, an initial round of shipments is sent to each local clinic from district hospitals, which receive medications from regional warehouses. After the malaria season begins, a recourse action is triggered to avoid shortages in the form of (1) lateral transshipment or (2) delayed shipment. The strategic model yields an optimal policy that suggests small clinic clusters with exclusive transshipment policies. This insight enables us to decompose the problem at the tactical level, solving each clinic cluster independently using a Markov decision process (MDP) approach to determine optimal periodic transshipment policies. We demonstrate the potential to reduce shortage incidents by using our proposed distribution system in a case study of 290 facilities under the control of the Malawi Ministry of Health. Numerical analysis of Malawi’s distribution system indicates that our decomposition method is very near optimal, and that such an approach is robust to challenges of developing countries, including slow paper-based inventory review, uncertain transportation infrastructure, the need for equitable distribution, and seasonal and correlated demand associated with malaria transmission dynamics.

Keywords: Health Care Operations, Humanitarian Logistics, Malaria Treatment Distribution, Stochastic Programming, Markov Decision Processes.

History: TBD.

1 Introduction

Despite decades of elimination and control efforts, malaria remains one of the most common causes of child morbidity and mortality worldwide. According to The World Health Organization (WHO) (2014) there were nearly 207 million suspected malaria cases in 2012. In addition to imposing an immense burden on health and welfare, malaria is a major impediment to the economic development of impoverished nations (see Malaney et al. (2004), Gallup and Sachs (2001)). Thus, for the past decade, malaria control and elimination have been a priority for international and domestic health agencies, non-governmental organizations (NGOs), and health ministries. Malaria is a treatable disease, and prompt administration of medicines for uncomplicated malaria such as Artemesinin Combination Therapies (ACTs) can prevent the most severe outcomes. However, stock outs of essential medications are common in developing countries, particularly those facing disproportionate malaria burdens (see PMI (2014), Sudoi et al. (2012)). Problems in regional supply chains have been noted as a major barrier to timely and efficient distribution of malaria medications to meet local demand (see Daniel et al. (2012), Tetteh (2009), Bateman (2013)).
According to United Nations (2011), Malawi is one of the poorest countries in the world and ranks 153rd out of 177 countries on the Human Development Index (HDI). Due to a combination of intense poverty and environmental and local weather conditions, Malawi suffers from an exceptionally high burden of malaria. Dzinjalamala (2004) indicates that all Malawians live at year round risk for malaria, though incidence peaks during the December-May rainy season. The World Health Organization (WHO) (2014) estimated that at least a third of all medical consultations are malaria related and a recent Malaria Indicator Survey showed that more than a third of all Malawians test positive for recent infections at given any time (see Malawi Ministry of Health (2012)). Malaria spending makes up a major portion of total expenditures on health in Malawi, crowding out spending on other conditions.

1.1 Operational Challenges

Foster (1991) claims that: (1) proper inventory management of medications in Africa can reduce costs by 15-20% and (2) transportation of drugs and medical aid is an especially critical factor in Africa. Our research aims to provide insights into how developing countries with systems of health delivery similar to Malawi’s can most effectively distribute expensive and valuable treatments on a restrictive transportation budget.

According to Claeson and Waldman (2000), the efficacy of delivering health care through such systems has been the subject of debate for decades. For instance, people in developing countries may seek treatment through the private sector, namely from local shops and community members due to real or perceived shortfalls in the public system - see Bustreo et al. (2003). Some research has focused on strategies that circumvent, replace, or radically decentralize public health systems (e.g., Gallien et al. (2012)). In underdeveloped health systems, centralized and hierarchical supply chains in which most of the supplies are distributed from a central authority, sometimes suffer from disorganization and corruption. Information technology is often ineffective making it difficult to track location and quantity of medication supply, hampering forecasting of demand and effective allocation of supplies. Transportation infrastructure is generally poor, fuel shortages complicate matters and roads are often in bad condition, especially during the rainy season when malaria is most prevalent. Cultural issues and regional rivalries prioritize certain areas while ignoring others, leading to inequities in access and supply. This observation is based on one co-author’s on the ground experience working with malaria in Malawi. In contrast, in a decentralized system, distribution of supplies is not necessarily controlled by a central authority. Instead, multiple players (public or private) contribute to the distribution of medications. However, the humanitarian nature of drug supply chains can prevent the development of more sustainable market-based supply chain solutions. Both centralized and decentralized systems have their limitations, but we focus on government sponsored health facilities, which are the most prominent source of medications and healthcare in most developing countries, including Malawi and nearly all sub-Saharan African countries. The methods suggested in this paper, however, can also be...
applied to other medical aid distribution supply chains outside of the public sector, such as those of the NGOs like John Snow Inc.

In this paper we explore transportation schemes that combine both strategic and tactical operations to increase the effectiveness of ACT distribution channels within the public, centralized supply chain of Malawi. At the strategic level, we first develop and solve a large stochastic program capable of optimizing ACT delivery to all 290 hospitals and clinics that treat malaria in Malawi. We then use this model to investigate the impact of transshipment and delayed shipment (where some inventory is held back at the higher echelon) on both transportation cost/feasibility and on ACT shortages. Applying this model to data obtained from the Malawi Ministry of Health, we find that the particular infrastructure in Malawi produces a convenient structure in the optimal solution to the stochastic program, in which the problem can be decomposed into small clusters of clinics with exclusive transshipment policies. This observation then allows us to implement a tactical method for transshipment using a tractable Markov Decision Process model, which could not be solved in the absence of clinic clusters due to the curse of dimensionality. By integrating both models, we are able to analyze unique features of pharmaceutical aid delivery in the developing world, such as poor road conditions, ethical considerations, seasonality of malaria, and the fact that inventories are only reviewed periodically due to hand counting and paper-based inventory systems. Our proposed approach reduces shortage by 40% - 60% compared to the baseline model.

Insights and Contributions in Pharmaceutical Aid Distribution in the Developing World. From the above analyses, we obtain the following managerial insights about pharmaceutical aid distribution in the developing world.

- **Fairness.** Equitable policies are near optimal when clinics are clustered with other clinics nearby
- **Seasonality.** Seasonality is easily handled by transshipment, but random demand variability has a significant impact on cost
- **Periodic Review.** Transshipment is effective even with infrequent inventory review intervals.
- **Transshipment Works with Poor Infrastructure.** The flexibility of transshipment to select alternate routes better meets demand than delayed shipment when roads are worn down or washed out during the rainy season.
- **Delayed Shipment is Best for Shortages.** Delayed shipment, despite being less flexible than transshipment, actually tends to achieve lower shortages, but at a higher transportation cost.
- **Clinic Clusters.** The optimal solution of the strategic model forms clusters of 2-5 clinics that only transship within the cluster. The formation and size of the clusters depends on the tradeoff between shortage penalty and transportation cost.
- **Mid-season Repositioning.** Allowing mid-season repositioning after the first half of the malaria season via a three-stage – instead of a two-stage – stochastic program allows the
model to better utilize inventory to reduce shortages when supply is sufficiently large and when shortage penalties are higher. This is due to targeted repositioning of inventory to distressed areas, which is particularly effective because of the positively correlated demand patterns of malaria.

Beyond these insights, we further contribute to the stream of literature in the following ways: (1) by developing a method for integrating both strategic and tactical phases for very large scale pharmaceutical distribution problems. (2) Applying this approach to two datasets from Malawi that capture the country’s entire ACT distribution system as well as historical demand for ACTs with both geographic and temporal components. (3) Performing extensive numerical analysis that demonstrates the difference between traditional transportation and inventory models and those designed for the developing world. We show that our decomposition method creating a tractable solution to the countrywide problem is in fact very close (0.5% gap on average) to the optimal, fully-integrated solution for smaller instances that are solvable to optimality. (4) Designing a practical and implementable policy based on the insights from the model solutions.

1.2 Malawi’s Existing Health System

Malawi’s public health system is a three tiered network consisting of central warehouses and regional hospitals in the first tier, district hospitals in the second tier, and primary health centers and local community clinics in the third. Each tier receives supplies from and answers to the tier above it with the exception of the central warehouses and regional hospitals which answer directly to the Ministry of Health (see Figure 1). Malaria care is provided at all levels of the Ministry of Health system, though central hospitals principally accept patients on a referral basis only. These patients typically represent the most severe malaria cases. Distribution of pharmaceuticals begins at the Central Medical Stores (CMS) in Lilongwe, Malawi, which allocate drugs to the regional hospitals and central warehouses (first tier). First tier distribution then delivers to district hospitals, which are in turn responsible for supplying primary health centers and local community clinics. The goal of this centralized system is to maintain a constant and equitable supply of drugs to all levels. However, as reported by Lufesi et al. (2007), shortages of essential medications due to logistical problems are common. In this paper we employ stochastic programming and Markov decision models to significantly decrease treatment shortage while keeping transportation costs low and affordable.

2 Literature Review

Literature from a number of areas is relevant to this paper, including (1) disaster preparedness and emergency response with pre-positioned inventories, (2) disease prevention resource allocation, and (3) transshipment and multi-echelon distribution models. In our model we consider a two-stage response in the distribution of medical supplies (as in disaster preparedness) as well as dynamic periodic lateral (bidirectional) transshipment decisions among clinics (the lowest echelon) based on small clusters of nearby clinics that are identified by the higher level two-stage
model. Our context is distinguished by characteristics that include: a centralized distribution system, three echelons and a network of almost 300 stockpoints, non-stationary demand by month, unfilled demand is lost, and heterogeneous shipping cost parameters enabling distances and road conditions to be incorporated in the model.

2.1 Disaster Response and Disease Prevention

Emergency response research tends to focus on broad public health needs that must be addressed in a rapid and targeted manner after a period of prior planning, often involving inventory prepositioning. Published research regarding disaster preparedness and emergency response is extensive and has been well documented by several survey papers, including Altay and Green (2006), Simpson and Hancock (2009), and de la Torre et al. (2011). Readers are encouraged to review these surveys for a more complete understanding of this stream of literature. Particularly relevant to our methodology are emergency response models that employ two-stage stochastic programming. These approaches involve an initial allocation of resources before a disaster and subsequent transportation to affected locations after a large emergency event (see for example Mete and Zabinsky (2010), Salmerón and Apte (2010)). Models of disaster preparedness and emergency response share similarities with our work; however, they typically involve rare events with unknown timing that require a rapid response. Our work targets random and seasonal demand for ACTs over the entire malaria season, served by a large set of stockpoints.

While our focus is on daily treatment distribution, a related area is disease prevention resource allocation, which typically considers public health response and budget allocation to specific disease treatments to mitigate the effects or prevent the spread of a particular disease. For example, Dimitrov and Morton (2009) develop a two-stage approach using MDP to prevent the spread of malaria in Nigeria by considering allocation strategies based on a geographical grid. Due to the large size of our problem (with nearly 300 distribution locations), such approaches are not tractable. Optimal allocation of resources (i.e., vaccines in the case of treating epidemics) is considered in Zaric and Brandeau (2002). Lasry et al. (2011) develops non-linear optimization models to determine the allocation of an HIV treatment and prevention budget over a five-year
horizon. Disease prevention models typically focus on the *effectiveness of treatment strategies*, emphasizing optimal allocation of a limited budget to a number of prevention and treatment actions. However, these models often disregard geographical or transportation components. Our work considers an *operational model* focusing on allocation and transportation of one type of treatment to geographical regions based on the dynamic and random demand over the course of the malaria season.

### 2.2 Transshipment and Multi-echelon Distribution Models

Our work also contributes to the area of transshipment research. Paterson et al. (2011) provides a comprehensive survey of transshipment, identifying areas where additional research is particularly needed. Among multiple areas in need of development, they cite the following three: (1) using transshipment to proactively redistribute/balance the stock with multiple transshipment epochs, (2) further work on larger numbers of locations (rather than the typical two or three), (3) larger networks with 3 or more echelons. As will be seen, our paper addresses these areas of need through a combination of modeling, theoretical analysis, and numerical analysis.

Traditionally, the literature on transshipments has generally addressed problems with only two retailers in analytical approaches for tractability (see Paterson et al. (2011) for references). Much literature assumes infinite capacity for replenishment, though some papers model a finite supply or production capacity. In our setting, the total amount of medication available is restricted, having been donated or sold to the country in large up-front lots – the method preferred by the ministry. Thus replenishment costs are limited to the cost of shipment or transshipment.

In a single echelon setting, a multi-period and multi-location approach is taken in Robinson (1990), and it shares a number of features with our model, such as multiple retailers in a multistage optimization setting with random demand and either backlogging or lost sales. A key feature of this analysis is time-stationarity of the model at each period, which is not an appropriate approximation for our setting. We consider non-stationary demand distributions over time; therefore, the control policies become more complex, in part because the multi-period solution does not reduce to a single-period solution. Further, the Markov Decision Process (MDP) approach taken by this paper would be intractable for our 290 facilities; however, we use the optimal solution of our strategic stochastic program to decompose the network into small clinic transshipment clusters that are solvable by MDP. We use the cluster transshipment policies determined via the MDP to derive operational insights. These include a strong characterization of optimal policies having a threshold structure and performing rebalancing above the threshold.

Herer et al. (2006) extends the multi-period, multi-location work of Robinson (1990), and it differs from our work in ways that include maintaining a stationary model, allowing backordering; it also assumes replenishment from a central supplier in every period. Rosales et al. (2013) provides a model consisting of two retailers and uses simulation to study the impact of model
parameters (e.g., cost, lead-time, and demand uncertainty) on both a transshipment model and an allocation system structure (shipment from a centralized depot). This work also addresses an issue that may be very real in practice: geographical demand correlation (which Herer et al. (2006) also considered). As intuition would suggest, positive correlations in demand across suppliers reduces the benefits of transshipment. Intuition and experience with malaria and its mechanisms suggest that positive correlations can be expected across clinics close to each other, captured in both our stochastic programming and MDP models.

Rottkemper et al. (2012) provides a mixed-integer programming approach to minimize a combination of shortage and operational costs under demand uncertainty in the context of humanitarian operations in a multi-modal transportation setting. They use a limited data set (corresponding to clinics in Kayanza province in Burundi) under eight scenarios to illustrate the effectiveness of their approach. Since the size of our problem is much larger, we construct a more scalable approach. Our paper aligns with Rottkemper et al. (2012) in demonstrating that transshipments can significantly reduce the unsatisfied demand at slightly increased overall cost. In addition, we argue that allowing transshipment actions can result in higher robustness against poor road conditions - an inherent characteristic of distribution problems in the developing world.

A main modeling contribution to the transshipment literature is the integration of both the strategic and tactical levels by combining a stochastic programming approach with a MDP approach. Previous work tends to consider one or the other. This integration is facilitated by the identification and use of the special geographical clinic clustering structure resulting from the optimal solution of the strategic model to decompose the country-wide distribution problem into tractable subproblems that could be solved using MDP.

Other contributions stem from the unique features of our application area: aid distribution in the developing world. First, the situation in aid distribution differs from conventional inventory models which tacitly assume an environment of ongoing production and consumption. However, in very poor countries such as Malawi, pharmaceutical supplies are often donated annually in advance of that year's malaria season with mid-season replenishment being uncommon. This lack of ongoing and predictable supply causes distribution and transshipment to behave differently from traditional contexts. Secondly, we analyze ethical solutions addressing perceived fairness, which is not typically a consideration in traditional transshipment literature. Third, we capture the impact of geographically and temporally correlated demand and seasonality of demand reflecting the characteristics of malaria. Fourth, we explore the impact of transshipment frequency, which is important because many developing world clinics use time consuming paper-based inventory systems and cannot engage in near-continuous review that electronic monitoring systems prevalent in retail and warehousing would allow. Fifth, we explore the impact of poor road conditions along certain transportation routes that are common, particularly during the rainy season (and consequently the peak malaria season) when roads can get washed out.
In the following sections we develop our model and present results based on data from Malawi, compiled from a number of sources. Though these methods could be applied to any product with seasonal demand and geographic variation, this research will address the problems noted in the previous paragraph by integrating strategic and tactical models that can provide workable, on-the-ground solutions to the problem of efficient distribution of anti-malaria medications in Malawi. First, we develop strategic stochastic programming models to address aggregate-level distribution. These models provide both practical results for country-wide distribution of ACTs and insight into an effective management structure, clinic clustering, for implementing a transshipment approach. Clinic clusters enable effective decomposition of the 290 clinic problem into manageable clusters of 3 clinics each on average. The output of the strategic stochastic program is used to parameterize a tactical periodic review model for transshipping between clinics. Clinic cluster decomposition enables tractable solutions to this optimization using a Markov decision process (MDP) approach. From the MDP we also gain valuable insights into the structure of the optimal transshipment policy, which we show to be of threshold type, and perform numerical analyses on several unique features of aid distribution in the developing world. Finally, we propose a simple operational policy that can be implemented even with paper-based records systems.

3 Deterministic Model for Medication Distribution

To demonstrate the effectiveness of incorporating demand uncertainty in distribution decisions we first define a “baseline” model as a surrogate for the current state of ACT distribution in Malawi. Note that this baseline model already represents an optimization with respect to the current practice. However, it is naive in the sense that it ignores the inherent randomness of demand; instead, it distributes medications based on average demand in each clinic by minimizing expected transportation costs and shortage penalties without any real time updates and recourse actions. Notation for all the distribution models that follow is given in Table 1.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{N}$</td>
<td>set of nodes consisting of the central medical storehouse ($m$), regional storehouses ($\mathcal{R}$), district hospitals ($D$), and local clinics ($\mathcal{C}$)</td>
</tr>
<tr>
<td>$\mathcal{A}_R$</td>
<td>subset of arcs connecting the central warehouse to regional warehouses</td>
</tr>
<tr>
<td>$\mathcal{A}_D$</td>
<td>subset of arcs connecting regional warehouses to district hospitals</td>
</tr>
<tr>
<td>$\mathcal{A}_C$</td>
<td>subset of arcs connecting district hospitals to local clinics</td>
</tr>
<tr>
<td>$\mathcal{A}_T$</td>
<td>subset of transshipment arcs connecting local clinics to one another</td>
</tr>
<tr>
<td>$\mathcal{A}$</td>
<td>set of arcs ($\mathcal{A} = \mathcal{A}_R \cup \mathcal{A}_D \cup \mathcal{A}_C \cup \mathcal{A}_T$)</td>
</tr>
<tr>
<td>$\mathcal{S}$</td>
<td>set of demand scenarios</td>
</tr>
<tr>
<td>$p_s$</td>
<td>probability of scenario $s$ where $s \in \mathcal{S}$</td>
</tr>
<tr>
<td>$\pi_i$</td>
<td>penalty of one unit of treatment shortage in clinic $i$</td>
</tr>
<tr>
<td>$c_{ij}$</td>
<td>cost of transporting one unit of treatment on arc $(i, j)$</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>total available supply of treatments</td>
</tr>
<tr>
<td>$d_{is}^s$</td>
<td>demand of local clinic $i$ under scenario $s$ where $i \in \mathcal{C}$ and $s \in \mathcal{S}$</td>
</tr>
</tbody>
</table>

Table 1: Distribution model notation.

The main decision variable in the baseline model, $x_{ij}$, corresponds to the number of malaria treatments transported on arc $(i, j)$. All distribution decisions are made at the same time based
on an historical estimate of demand. An auxiliary variable, $z_s^i$ is introduced to capture the shortage of malaria treatments in clinic $j$ under scenario $s$, in which a demand of $d_s^i$ is realized for clinic $i$. The min-cost flow formulation introduced in (1)-(7) represents the baseline model.

$$\min \sum_{(i,j) \in A} c_{ij}x_{ij} + \sum_{s \in S} \sum_{i \in C} p_s \pi_i z_s^i$$

s.t.

$$\sum_{j: (m,j) \in AR} x_{ij} \leq \sigma$$

$$\sum_{j: (m,j) \in AR} x_{ij} = \sum_{j: (j,i) \in AP} x_{ji} \quad \forall i \in R$$

$$\sum_{j: (i,j) \in AP} x_{ij} = \sum_{j: (j,i) \in AC} x_{ji} \quad \forall i \in D$$

$$\sum_{j: (j,i) \in AC} x_{ji} = d_s^i - z_s^i \quad \forall i \in C, \forall s \in S$$

$$x_{ij} \geq 0 \quad \forall (i,j) \in A$$

$$z_s^i \geq 0 \quad \forall i \in C, \forall s \in S.$$

The objective function (1) minimizes total cost, comprised of transportation costs and shortage penalty and (2) constrains the amount distributed to be at most the available supply ($\sigma$). Constraints (3), (4) represent the flow conservation constraints for regional warehouses to district hospitals and district hospitals to local clinics respectively. The left-hand-side of (5) represents the total flow of ACTs into local clinic $i$ and the right-hand-side corresponds to the total demand of clinic $i$ under scenario $s$ ($d_s^i$) minus the shortage in that clinic under scenario $s$ ($z_s^i$).

### 4 Two-Stage Stochastic Formulation

The models in this section contrast with the baseline model in the sense that additional demand information becomes available and recourse actions are triggered in the second stage. In the first stage, the Malawi Ministry of Health would decide how many ACTs to send to each facility before the malaria season begins. In the second stage, the actual demand is realized and the Ministry can take recourse actions to address the supply and demand mismatch. Here we consider two potential recourse actions: (1) transshipment and (2) delayed shipment.

In the transshipment model (Section 4.1), all the ACTs are distributed among all the facilities (tier 1, 2, and 3) in the first stage. In the second stage, transshipment of ACTs between facilities occurs to adjust inventories in light of new demand information. In the delayed shipment model (Section 4.2), an initial delivery of ACTs is distributed to the clinics, but some is held back at the higher tier. During the malaria season, a better estimate of the demand is realized and a second round of shipments is delivered. There are benefits and drawbacks to both models. Delayed shipment is less cost effective, but from an implementation standpoint it has the political benefit of not needing to take stock from one clinic to give to another. Transshipment, on the other
hand, is more cost-effective but harder to centrally control. Note, however, that according to Kiczek et al. (2009) transshipment already occurs on an ad-hoc basis in Malawi and in a more structured manner in neighboring Zambia (Mtonga (2010)).

Figure 2 illustrates the timeline of events for the two-stage stochastic models. Note that the recourse actions are not necessarily done all at once. Instead, the transshipments or delayed shipments are made throughout the malaria season as needed. Therefore, the recourse decisions considered here are aggregate-level surrogates for the actual periodic adjustments in the inventory level of each facility.

Figure 2: Event timelines for two-stage stochastic models.

4.1 Two-stage Transshipment Model
A necessary assumption for this stochastic programming formulation is that transshipment occurs immediately and instantaneously (as in a continuous review system) in response to every stock out. This approximation is acceptable for the planning stage that the stochastic program represents. To provide insight into how Malawi, Zambia, or similar countries might operationalize transshipment, we develop a periodic review and transshipment model in Sec. 7. We assume a set of scenarios ($S$) where each scenario, $s \in S$ is realized with probability $p_s$. Under scenario $s$, the realized value of demand for clinic $i$ is $d_i^s$. The first-stage problem has objective:

$$\min \sum_{(i,j) \in A} c_{ij} x_{ij} + Q,$$

and constraints (2) - (4) and (6) from the baseline model. The expected recourse function, $Q$, is given by:

$$Q = \min \sum_{s \in S} p_s \left( \sum_{(i,j) \in A^c \cup A^T} c_{ij} y_{ij}^s + \sum_{i \in C} \pi_i z_i^s \right),$$

s.t.
\begin{align}
\sum_{j: (i,j) \in A^C} y^s_{ij} & \leq \sum_{j: (i,j) \in A^D} x_{ij} - \sum_{j: (i,j) \in A^C} x_{ij} & \forall i \in D, \forall s \in S \tag{10} \\
\sum_{j: (i,j) \in A^T \cup A^C} y^s_{ji} - \sum_{j: (i,j) \in A^T \cup A^C} y^s_{ij} + z_i^s & \geq - \sum_{j: (i,j) \in A^C} x_{ji} + d^s_i & \forall i \in C, \forall s \in S \tag{11} \\
y^s_{ij} & \geq 0 & \forall (i,j) \in A^T \cup A^C, \forall s \in S \tag{12} \\
z_i^s & \geq 0 & \forall i \in C, \forall s \in S. \tag{13}
\end{align}

The decision variable \( y^s_{ij} \) corresponds to the aggregate transshipment of ACTs from facility \( i \) to facility \( j \) under scenario \( s \) throughout the malaria season. (9) minimizes the expected cost of the second stage – transshipment cost plus shortage penalty – where \( z_i^s \) represents shortage of medications in clinic \( i \) under scenario \( s \). (10) ensure that the second round of shipments from district hospitals to local clinics \( (\sum_{j: (i,j) \in A^C} y^s_{ij}) \) do not exceed the available ACTs left from the first stage \( (\sum_{j: (i,j) \in A^D} x_{ij} - \sum_{j: (i,j) \in A^C} x_{ij}) \). (11) capture the concept that the net transshipment plus shortages at clinic \( i \) under scenario \( s \) (LHS) should exceed residual demand (demand minus initial allocation of ACTs from stage 1) at clinic \( i \) under scenario \( s \) (RHS). Note that the value of first-stage decisions \( x_{ij} \) is known in the second stage, therefore \( \sum_{j: (i,j) \in A^D} x_{ji} \) is known. Thus, we re-arrange terms in (11) such that decision variables are on the left-hand-side and the known parameters are on the right-hand-side.

### 4.2 Two-stage Delayed Shipment Model

In the delayed shipment model, some ACTs are reserved at a higher tier for shipment after the start of the malaria season. The first-stage problem has an identical formulation to the transshipment model. The expected recourse function, \( Q \), is given by:

\begin{align}
Q = \min_{s \in S} \sum_{s \in S} p_s & \left( \sum_{(i,j) \in A^C} c_{ij} w_{ij}^s + \sum_{i \in C} \pi_i z_i^s \right) \tag{14} \\
\text{s.t.} & \sum_{j: (i,j) \in A^C} w_{ij}^s \leq \sum_{j: (i,j) \in A^D} x_{ij} - \sum_{j: (i,j) \in A^C} x_{ij} & \forall i \in D, \forall s \in S \tag{15} \\
& \sum_{j: (i,j) \in A^C} w_{ji}^s + z_i^s \geq - \sum_{j: (i,j) \in A^C} x_{ji} + d^s_i & \forall i \in C, \forall s \in S \tag{16} \\
w_{ij}^s & \geq 0 & \forall (i,j) \in A^C, \forall s \in S \tag{17} \\
z_i^s & \geq 0 & \forall i \in C, \forall s \in S. \tag{18}
\end{align}

where \( w_{ij}^s \) denotes the amount of ACTs shipped from district hospital \( i \) to clinic \( j \) throughout the malaria season. The objective function (14) minimizes the expected transportation costs and shortage penalties. Equation (15) is essentially similar to Equation (10) from the transshipment model, allowing some inventory to be kept at the district hospital. This means that the transshipment model does have a similar capability of delayed shipment. In most cases, however, the amount of inventory stored at the district hospital in the transshipment model is negligible.

As we will discuss in section 6, under some parameter regimens, especially when the cost of
transshipment arcs exceeds those of delayed shipment arcs, the transshipment model can result in similar outcomes as those generated by the delayed shipment model. Constraints (16) capture the shortage in each clinic \( z^s_i \) after the second round of ACTs is distributed. Constraints (17) and (18) ensure the non-negativity of shortage.

### 4.3 Supply Equity

When the total supply of malaria medication is less than the demand, shortage is inevitable. In such a scenario, it is possible that some clinics may face significantly higher shortages than others. An equitable policy, however, distributes ACTs in a manner that limits the shortage disparity between clinics. Minimizing the sum of absolute differences between the shortage of each clinic and the average shortage among all clinics can limit this disparity. Let \( \bar{z}^s \) be the average shortage of ACTs in all the clinics in scenario \( s \) and \( \tilde{z}^s_i \) be the absolute difference between the shortage in clinic \( i \) and the average shortage. The following constraints capture this concept:

\[
\bar{z}^s = \frac{1}{|C|} \sum_{i \in C} z^s_i \quad \forall s \in S
\]  

\[
\tilde{z}^s_i \geq z^s_i - \bar{z}^s \quad \forall s \in S, \forall i \in C
\]  

\[
\tilde{z}^s_i \geq -z^s_i + \bar{z}^s \quad \forall s \in S, \forall i \in C
\]  

\[
\tilde{z}^s_i \geq 0.
\]

Equations (19) define the average shortage. Equations (20) - (22) linearize the absolute value function. Based on the above definition of equity, we can modify the objective function in each model by adding a new term \( \sum_{s,i} \pi_i \tilde{z}^s_i \). Other approaches can maintain equity without violating linearity as well, such as minimizing the maximum shortage, or minimizing the difference between the minimum and the maximum shortage values.

### 5 Three-Stage Stochastic Formulation

In Section 4, we considered two-stage models in which the demand scenario for each clinic is realized at the beginning of the malaria season. Recourse actions address the disparity between the realized demand and the initial inventory of ACTs at each clinic. One drawback of the two-stage model is that the recourse actions are aggregate-level surrogates for the actual periodic decisions. This assumption allows for a tractable solution at the cost of ignoring the temporal (e.g., bi-monthly, monthly, weekly, etc.) fluctuations in demand. When temporal demand fluctuations are high, the two-stage models may underestimate the actual shortage in each period.

Model accuracy can be improved by increasing the granularity of the recourse actions. For instance, the transshipments or delayed shipments can be delivered periodically so the model can better estimate the actual shortage in each period. However, as the granularity of the model increases, the computation time increases dramatically. Moreover, collecting and processing the demand data at a very detailed level is often not feasible in a developing nation.

In Section 5.1 we explore the benefits of increasing the granularity of the recourse actions by extending the former analysis to a three-stage stochastic program, using a revised timeline.
shown in Figure 3.

5.1 Three-Stage Transshipment Model and Delayed Shipment Models

The first stage of the three-stage problem (initial distribution of medications) is identical to that of the two-stage problem in Sec. 4.1. The second stage represents the initial round of transshipment; therefore it includes an additional term ($Q'$) in the objective function (23) to represent the third-stage problem - the final round of transshipment. We also define a new auxiliary decision variable ($l_i^s$) to represent the number of ACTs left at clinic $i$ under scenario $s$ after the second stage and the term $l_i^s$ is subtracted from the left-hand-side of the flow conservation constraints (29). Note that in the third-stage problem, $l_i^s$ is calculated by the flow conservation constraint of the second-stage problem (25) and thus is considered input data.

$$Q = \min \sum_{s \in S} p_s \left( \sum_{(i,j) \in A^T \cup A^C} c_{ij} y_{ij}^s + \sum_{i \in C} \pi_i z_i^s \right) + Q' \tag{23}$$

s.t.

$$\sum_{j : (j,i) \in A^C} y_{ij}^s \leq \sum_{j : (j,i) \in A^D} x_{ij} - \sum_{j : (j,i) \in A^C} x_{ij} \quad \forall i \in D, \forall s \in S \tag{24}$$

$$\sum_{j : (j,i) \in A^T \cup A^C} y_{ij}^s - \sum_{j : (j,i) \in A^T \cup A^C} y_{ij}^s + z_i^s - l_i^s = - \sum_{j : (j,i) \in A^C} x_{ji} + d_i^s \quad \forall i \in C, \forall s \in S \tag{25}$$

$$y_{ij}^s \geq 0 \quad \forall (i,j) \in A^T \cup A^C, \forall s \in S \tag{26}$$

$$z_i^s \geq 0, l_i^s \geq 0 \quad \forall i \in C, \forall s \in S. \tag{27}$$

For the third-stage problem ($Q'$) represented by (28)-(31), we define $y_{ij}^s$ to be the decision on the number of ACT units transshipped from clinic $i$ to clinic $j$ under scenario $s$ in the third stage. We allow the transshipment cost in the third stage, ($c_{ij}'$), to be different from the second-stage cost ($c_{ij}$).

$$Q' = \min \sum_{s \in S} p_s \left( \sum_{(i,j) \in A^C} c_{ij}' y_{ij}^s + \sum_{i \in C} \pi_i z_i^s \right) \tag{28}$$

s.t.
\[
\sum_{j: (i,j) \in A^T} y^s_{ij} - \sum_{j: (j,i) \in A^T} y^s_{ji} + z^s_i \geq -l^s_i + d^s_i \quad \forall i \in C, \forall s \in S \tag{29}
\]
\[
y^s_{ij} \geq 0 \quad \forall (i,j) \in A^T, \forall s \in S \tag{30}
\]
\[
z^s_i \geq 0 \quad \forall i \in C, \forall s \in S. \tag{31}
\]

The third stage, Eq. 28-31 has the same structure and intuition as the second stage.

**Delayed Shipment.** The form of the three-stage delayed shipment model is analogous to the three-stage transshipment model with the necessary changes illustrated in Sec. 4.2 for the two-stage version. For brevity we do not repeat them here.

### 6 Computational Experiments

In this section, we present the results of computational experiments based on actual locations of health facilities that were mapped in a country-wide survey conducted by the Japanese International Cooperative Agency (JICA) in the year 2000. Facility demand was estimated based on regular malaria case counts as reported by hospitals and clinics to the central Ministry of Health, spanning the years 2003–2008. The data are summarized in annual government reports prepared by Republic of Malawi Ministry of Health (2009). These data were constructed based on case counts reported at the facility level, whether the demand was met or not. As in all developing country contexts where facility level reporting of case counts is often rudimentary, there were some missing and incomplete data at the facility level. Thus, while counts at some facilities were known through government reports, malaria case counts at others were approximated through a combination of district level data and estimated catchments. Using Thiessen polygons and the 1998 Malawian Census facility, catchments were estimated. It was assumed that people used the closest health facility. In reality, this may not always be true, as patients may prefer one facility over another, or transportation (i.e., buses) might facilitate travel to a farther facility. Nonetheless, as in common practice, case counts were assigned proportional to health facilities based on the estimated population catchment and the probability of contracting malaria associated with each geographical region.

Data were averaged over the five years to model a typical malaria season in Malawi in the face of partially missing data at the clinic level. By observing clinics where the data were more complete we noted that, over the five years, there was some variability from year to year but little evidence of overall upward or downward trend. This is further confirmed by the World Malaria Report 2014 (see The World Health Organization (WHO) (2014)), which indicates no significant trend over the time period. Though there was a slight increase in malaria-related hospital admissions over the time frame, the report states that data were insufficiently consistent to assess any trend in Malawi, consistent with observations from our own data. Though we recognize the limitations of the data available, it is still very significant that this effort is data-driven and reflects real temporal and spatial patterns of malaria incidence given current research.

Fig. 4 shows the geographical and seasonal shape of the ACT demand curve from our data.
The darker color (red) in Fig. 4 (a), indicating higher annual demand, tends to appear near populous urban areas like Blantyre and in both urban and rural areas near Lake Malawi where mosquitoes are more prevalent. In our historical demand data, malaria prevalence in each region grew over the malaria season proportional to the infected population. Fig. 4 (b) shows that malaria medication demand basically follows a 6 month seasonal pattern which coincides with the seasonal patterns of rainfall and thus of Anophelene mosquito prevalence.

Based on the demand pattern illustrated in Fig. 4, we divided the year into two (three) periods for the two-stage (three-stage) model respectively. In the two stage model, the first stage is August 1 through November 30, with an average demand of 113,331; the second stage is December 1 through July 31 with an average demand of 1,993,989. In the three-stage model, the first stage is August 1 through November 30; the second stage is December 1 through March 31 with an average demand of 674,702; and the third stage is April 1 through July 31 with an average demand of 1,319,287. In the following sections, we investigate the impact of geographical and seasonal characteristics of the disease on the optimal drug distribution policies.

For each clinic we generated 10 scenarios to populate demand parameters for the second ($d_i^2$) and third stage ($d_i^3$) respectively. We start with 5 main scenarios, assuming each of those scenarios are equally likely to happen. The first three key scenarios were generated by perturbing the original demand from our historical data ($D_i$). To add robustness, two other scenarios were generated using a uniform distribution such that the mean of the uniform distribution equals the average observed demand. For each of those five main scenarios, we generated a less likely variation to capture the potential for rare extreme events. In total, we assumed each main scenario has a probability of 0.19 and each scenario extension has a probability of 0.01. In designing these scenarios, we also capture correlated demand across the different stages as malaria is a transmittable disease whose spread depends on the number of infected persons. That is, high initial demand is more likely to translate into high demand in future stages. Details of these scenarios are described in Table 2. These scenarios were generated based on the historical data and expert opinion from one of our co-authors who has performed extensive field-work in Malawi regarding malaria. The scenarios developed herein were designed and confirmed based on his personal experiences over several years in Malawi. Our data include demand from 290 facilities.
including 3 regional warehouses, 21 district hospitals and 266 local clinics.

<table>
<thead>
<tr>
<th>No</th>
<th>Name</th>
<th>Description</th>
<th>Probability</th>
<th>Demand in Stage 2 ($d_i^s$)</th>
<th>Demand in Stage 3 ($d_i^{s'}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>LOW1</td>
<td>Low total demand, variation 1</td>
<td>0.19</td>
<td>$\frac{1}{7}D_i$</td>
<td>$\frac{2}{5}D_i$</td>
</tr>
<tr>
<td>2</td>
<td>LOW2</td>
<td>Low total demand, variation 2</td>
<td>0.01</td>
<td>$\frac{1}{5}D_i$</td>
<td>$\frac{3}{5}D_i$</td>
</tr>
<tr>
<td>3</td>
<td>MED1</td>
<td>Medium total demand, variation 1</td>
<td>0.19</td>
<td>$\frac{1}{7}D_i$</td>
<td>$\frac{2}{5}D_i$</td>
</tr>
<tr>
<td>4</td>
<td>MED2</td>
<td>Medium total demand, variation 2</td>
<td>0.01</td>
<td>$\frac{1}{5}D_i$</td>
<td>$\frac{3}{5}D_i$</td>
</tr>
<tr>
<td>5</td>
<td>HIGH1</td>
<td>High total demand, variation 1</td>
<td>0.19</td>
<td>$\frac{3}{5}D_i$</td>
<td>$\frac{2}{5}D_i$</td>
</tr>
<tr>
<td>6</td>
<td>HIGH2</td>
<td>High total demand, variation 2</td>
<td>0.01</td>
<td>$\frac{3}{5}D_i$</td>
<td>$\frac{2}{5}D_i$</td>
</tr>
<tr>
<td>7</td>
<td>CONS1</td>
<td>Uniform total demand, variation 1</td>
<td>0.19</td>
<td>$U[0.9D_i^{d}, 1.1D_i^{d}]$</td>
<td>$2d_i$</td>
</tr>
<tr>
<td>8</td>
<td>CONS2</td>
<td>Uniform total demand, variation 2</td>
<td>0.01</td>
<td>$U[0.9D_i^{d}, 1.1D_i^{d}]$</td>
<td>$\frac{10}{9}d_i$</td>
</tr>
<tr>
<td>9</td>
<td>VAR1</td>
<td>Uniform demand variation 3</td>
<td>0.19</td>
<td>$U[0.2D_i^{d}, 1.8D_i^{d}]$</td>
<td>$\frac{3}{4}d_i$</td>
</tr>
<tr>
<td>10</td>
<td>VAR2</td>
<td>Uniform demand variation 4</td>
<td>0.01</td>
<td>$U[0.2D_i^{d}, 1.8D_i^{d}]$</td>
<td>$d_i$</td>
</tr>
</tbody>
</table>

Table 2: Ten demand scenarios where generated based on the observed historical demand for each clinic ($D_i$).

To capture the transportation cost ($c_{ij}$) we calculated the distance between each facility pair in kilometers. For the purposes of this research, we used Euclidean distance. A road map for Malawi was available, and more accurate measures of road distance based on a map could have been produced, but it was found that the quality of the map varied by geographic area. It was found that Euclidean distances accurately reflect road-based measures regionally in Malawi and other research has confirmed that this measure is satisfactory compared to more sophisticated methods (see Nesbitt et al. (2014)). To account for cartographic shortfalls on published maps and thereby maintain consistency over the region of interest, we use straight line distance.

Road quality in Malawi varies widely and the system is mostly underdeveloped so the unit transportation cost can vary by route, though specific data on road quality are not available. Thus, we initially use the average transportation cost per kilometer in Malawi reported by Lall et al. (2009) which is about 4 cents (or 228.4 kwacha, the Malawian currency), but vary the costs to account for road conditions in our sensitivity analyses.

### 6.1 Sensitivity Analysis on the Value of Shortage Penalty

Estimating the shortage penalty for malaria medications is non-trivial. Factors such as loss of income and productivity (for patients and relatives) during the course of infection, and health care expenditures should be taken into account in order to obtain a correct estimate. It should be noted that given the type of parasite, the symptoms and their severity vary dramatically. Some people may have already developed immunity while for others (especially children) the disease can be deadly. Furthermore, malaria can have a higher indirect impact on children by hampering their physical and intellectual growth. Accounting for all these factors and monetizing their impact is key to determining the actual value of the shortage penalty and is beyond the scope of this paper. However, note that in the proposed approach, the rationale for including a shortage penalty is simply to ensure that medications are distributed according to the demand. In absence of such penalty, the model simply stops distributing medications in order to save transportation cost. Due to a lack of reliable data regarding health care expenditures, we performed a sensitivity analysis...
analysis on the value of the shortage penalty. We considered Malawi’s national income per capita, $810 reported by The World Health Organization (WHO) (2014) as a basis. According to UNICEF (2004), malaria can slow the economic growth in sub-Saharan Africa by 1.3% annually. Based on these statistics, one can estimate the economic impact of malaria in Malawi at the individual level to be about $10.5 in lost economic growth. Thus we ran the models for a range of potential shortage penalty values between $10 and $100. For those experiments, we set the available supply of ACT to 1.5 million units. Based on our initial computation, even a low number (such as $20) is high enough to trigger effective distribution of medications. As the value of the shortage penalty is increased, reducing shortages becomes more critical relative to transportation cost.

Figure 5 clearly depicts the changes in the models’ behavior as we change the value of the shortage penalty. Under lower shortage penalty values, models tend not to satisfy all the demand to save on transportation cost. For extremely low shortage penalty values such as $10, we observe little difference between the actions of the stochastic models and that of the baseline deterministic model. As the value of shortage penalty increases, the models ship more medications to decrease shortage volume and thus incur a higher transportation cost. This aligns with the results reported by Rottkemper et al. (2012).

The three-stage delayed shipment model tends to be more effective at addressing shortage, though at a higher transportation cost. As the shortage penalty increases, however, we observe that the gap between the three-stage delayed shipment model and the three-stage transshipment model shrinks. Also note that under high shortage penalty values, the two-stage delayed shipment model results in higher shortage than the two-stage transshipment model. This occurs because once the actual demand is realized, the delayed shipment model can only send additional shipments of medications from the district hospitals to the local clinics to address shortage. The transshipment model, on the other hand, has a broader base of facilities from which to satisfy demand. In some sense, this confirms the results of Rosales et al. (2013) that transshipment models outperform generalized allocation mechanisms under most parameters.

Managerial Insights. Here we summarize some key takeaways from this analysis. First, the three stage models outperform the two stage models on shortages while maintaining a similar
transportation cost. This is because the added control to be able to reposition inventory once part of the malaria season has been observed allows for more targeted efforts at distribution in distressed areas. This is especially important because of the demand correlation from period to period. Second, the delayed shipment model actually has lower shortages than transshipment in most cases. At first, this may seem counterintuitive because there is more flexibility in the transshipment model. However, this flexibility actually causes the first stage to distribute all the inventory out to the clinics to save on transportation cost, shipping direct instead of prepositioning a large stock at the district hospitals. This creates greater dispersion of the inventory across the country, making it more likely that sufficient inventory is not nearby the point of need and transshipment may have to come from a clinic far away and therefore is not worth the transportation cost. Whereas in the delayed shipment model, the district hospitals tend to be centrally located with many clinics around them. Since there is a larger stock stored at these hospitals initially (by design), there will be more incentive to take the recourse shipping action in stages two and three due to sufficient inventory and proximity. This also explains why shipping cost is higher for delayed shipment; rather than shipping direct, much of the product must follow first a route from the main dispensary to the district hospital, followed by a second route from the hospital to the clinics. Hence, the key insight is that if the government has sufficient transportation budget and wants to avoid any shortages, then delayed shipment may well be a better transportation structure.

6.2 Sensitivity Analysis on Supply Availability

Supply availability is a major challenge in distributing malaria medications in holoendemic areas. As reported by Natarajan and Swaminathan (2014) the process of procuring humanitarian supplies can be subject to delays and uncertainty. To better assess the effectiveness of our proposed stochastic models, we compare their results for a range of possible supply values, between 500,000 to 2,000,000 units while fixing the shortage penalty at $100. Fig. 6 (b), shows that the stochastic models can better utilize the additional supply of medications to address shortage compared to the baseline model. Among the stochastic models, three-stage models tend to be better at utilizing additional supply than two-stage models. This insight is similar to the key takeaway from Sec. 6.1; specifically, the three-stage model better utilizes the supply of ACTs through targeted repositioning in stages two and three. Obviously reducing shortage will result in higher transportation costs; but as seen in Fig. 6 (a), the difference in transportation cost for different models is relatively small.

6.3 Two-Stage vs. Three-Stage Models

Through computational experiments, we also analyze the marginal benefit of adding another recourse stage to the stochastic model. To focus on this impact given different uncertainty profiles, we performed separate experiments focusing on two scenario subsets: (1) scenarios where the majority of demand is realized in the second stage (LOW2, MED2, HIGH1, CONS2, and VAR1); (2) scenarios where the majority of the demand is realized in the third stage (LOW1,
Figure 6: Sensitivity analysis on supply availability (shortage penalty = $100).

MED1, HIGH2, CONS1, and VAR2). The three-stage model provides more opportunities to react to demand uncertainty, but adding more stages makes the problem harder to solve.

As observed in Figures 5, 6, and 7, three-stage models outperform two-stage and the difference is accentuated as supply increases. From Figure 7, the marginal benefit of moving from the baseline model to the two-stage model is higher in transshipment models. On the other hand, the three-stage delayed shipment model is more effective at reducing shortage compared to the three-stage transshipment model. Depending on parameters such as supply, shortage penalty, and demand uncertainty profile, the magnitude of these marginal benefits can vary.

Figure 7: Stochastic models compared to the baseline.

6.4 Transshipment Cost Sensitivity Analysis

In this section we analyze the sensitivity of the transshipment model to the transshipment cost and compare it with the delayed shipment model. In particular we explore the cases where (1) transshipment is cheaper and (2) more expensive than shipment along the main channels from the regional and district hospitals. It may be possible that shipments are cheaper because smaller and more frequent shipments may be transported with smaller and cheaper transportation methods, such as a motorcycle or small vehicle that can more easily pass difficult terrain or roads that are damaged by heavy rains. In these cases, a large truck may have difficulty navigating certain routes and therefore be more costly relative to transshipment. On the other hand, it may also be possible that frequent transshipment loses economies of scale, making clinic-to-clinic shipping costs more expensive. Thus we analyze how the models react in both cases. To do so, we modify the cost of clinic-to-clinic transshipment to be X% of the standard cost of shipment, where X
ranges from 0 - 150%. Costs for the other routes remain unchanged.

In Figure 8(a) and (b) the dashed lines represent the expected transportation cost and expected shortage volume respectively for the delayed shipment models. These are constant across all scenarios because delayed shipment does not use the clinic-to-clinic routes. The solid lines represent the transportation and expected shortage costs for the transshipment model. When transshipment is less expensive, all the inventory is initially allocated to one clinic in the cluster, which then transships to the other clinics due to the lower cost of transshipment relative to the cost of the initial distribution. The transportation costs initially increase, though more slowly, as clinic-to-clinic routes become more expensive. The model behaves this way because of the trade-off between shortage penalties and the transshipment recourse action. When transshipment is 60% as expensive as the main routes, the transportation cost begins to decrease while the shortage penalty increases more sharply. This inflection point occurs because at this level, clinic-to-clinic transportation becomes expensive enough that the model will stop transshipping along certain routes altogether, preferring some shortages rather than incurring high shipping costs, e.g., shipping across the country to fulfill a small amount of demand. Note that as previously mentioned in Sec. 4, the similarity between Equations (10) and (15) enables the transshipment model to store some inventory at district hospitals and delay shipments if necessary. Eventually both costs approach those of the delayed shipment model as transshipment becomes so expensive that the model chooses to avoid clinic to clinic shipments almost entirely and uses primarily the non-clinic-to-clinic routes that the delayed shipment model is restricted to.

Figure 8: Transshipment cost sensitivity analysis using the two-stage model

While the shape of the curve is driven by the particular network structure as well as the shortage penalty and original transportation cost values, changing these values would likely shift the inflection point while maintaining the overall shape. A key insight is that, once transshipment reaches a scenario-specific cost threshold (e.g., 50-60% in Fig. 8), the transportation cost will remain relatively stable, as the optimization becomes more conservative as to how far one would be willing to ship medications to satisfy unmet demand in a different region. That is, as clinic-to-clinic transportation becomes more expensive, transshipment eliminates routes from consideration due to high cost thus becoming less effective in satisfying all demand. Routes will be eliminated starting with the longest routes, keeping only the shorter routes available for trans-
shipment and thereby localizing transshipping needs around increasingly proximate geographical clusters. As an extension of this line of reasoning, the higher the penalty cost for unmet demand, the longer the model resists localizing transshipment efforts in favor of more regional/national transshipment. Thus, depending on the strategic goals and constraints of the distributor, the shipping network may be more localized (prioritizing transportation cost due to a constrained budget) or more national (sufficient budget shifts the focus to shortages).

6.5 Road Condition Analysis
As mentioned in Sec. 1.1, poor road conditions can make certain transportation routes difficult and therefore more costly, requiring for instance special vehicles or delayed travel during especially poor weather periods. To analyze this feature of supply distribution in the developing world, we design a scenario in which a proportion of the roads in our supply network is made more costly to travel due to poor road conditions. Since we are not aware of any data on the actual road conditions of the thousands of potential supply routes, we test the model’s sensitivity to poor road conditions by varying the proportion of total routes that are considered to be in poor condition from 10% up to 50%, with the poor routes being selected at random and the additional cost of shipping along the given route also generated randomly. In the scenario where 10% of roads are considered in poor condition, we assigned a Bernoulli indicator to each road where the road is considered in poor condition with \( p = 0.1 \) and standard condition with \( 1 - p = 0.9 \). If the road was found to be poor, we multiplied the transportation cost by \( 1 + U(0, 0.5) \), where \( U(\cdot) \) is a uniform random variable. We modify the cost per km rather than adding a random quantity to each route because when traveling on a poor road for a longer distance, the cost should increase more than when traveling on a poor road for a shorter distance. We then generated five outcome samples for the entire set of routes. For the 20% scenario, we started with the same bad roads as the 10% scenario and then modified the remaining roads that had not been touched in the previous scenario using a Bernoulli probability that guaranteed that 20% of the total routes would be modified (in this case \( p = 0.111 \)). This yields a coherent comparison between the different scenarios. The rest of the scenarios (30%-50%) were generated in the same manner.

Fig. 9 shows the results for the different percentages of roads in poor condition in terms of transportation cost (Fig. 9 (a)) and the expected shortage (Fig. 9 (b)). The X’s represent the solution of the transshipment model for the five different random scenarios we generated at each percentage of poor roads. The dashed line represents the average of the five scenarios for transshipment. Likewise the plus symbol and solid line represent the corresponding outcomes for the delayed shipment model.

First note that the transportation cost remains relatively stable as the percentage of bad roads increases. This is because the transportation cost is high enough that it becomes more beneficial to keep medications locally rather than ship across routes with poor road conditions for a small reduction in shortages. Delayed shipment costs trend downward because there are fewer viable
options when a key route becomes affected by poor road conditions so the model chooses to accept more shortages. The transshipment model, on the other hand, has more flexibility because there are many more options when clinic-to-clinic routes are added. When one route becomes more expensive, the model is able to find other viable routes to transship product. The transshipment model’s transportation cost demonstrates a slight upward trend as the transshipment model seeks alternative routes that allow for more movement of medications at a slightly higher price, however not so high that the routes become untenable.

The cost of storing more medications locally can be measured in terms of increased shortages. As seen in Fig. 9 (b), the slope of the increase in shortage is steeper for the delayed shipment model than the transshipment model, which implies that the transshipment model is better at meeting demand as road conditions worsen.

![Graphs showing expected transportation cost and shortage volume](image)

Figure 9: Analysis of road conditions using the two-stage model.

A key takeaway from this analysis is that poor road conditions lead to an increase in shortages and less movement of product around the network. However, the flexibility of the transshipment model to choose alternate routes enables more demand to be satisfied relative to delayed shipment, albeit at increased transportation cost due to using more expensive alternate routes.

### 6.6 A New Distribution Structure: Establishing Clinic Clusters

One of the key insights gained from the computational experiments on the strategic-level stochastic program is the appearance of what we call **clinic clusters**. That is, the transshipment stochastic models group clinics together into clusters such that transshipment often occurs *within* clusters only, and very rarely *between* different clusters. In our transshipment model experiments, for example, 15% to 25% of clinics send ACTs to other clinics in the recourse stage. These *sender* clinics transship their excess inventory to between 2 and 5 proximal *receiver* clinics in the vicinity of the sender clinic. Figure 10 illustrates five representative clinic clusters in the northern area of Malawi. This idea of clusters can be used to decompose the nationwide problem into tractable cluster-level problems that can be solved independently at the tactical/operational level. This is the key to integrating the strategic models with the operational models that we develop in Section 7.

This decomposition has further benefits, addressing some of the distribution issues that make developing country drug distribution different. The geographic proximity of clinics within a
cluster (1) increases the likelihood that the clinics would be willing to work together toward a transshipment program; this partially mitigates challenges associated with regional rivalries identified in Sec. 1.1; (2) it makes transshipment over the short distances between clinics more feasible by motorcycle or small vehicle that are less likely to get stuck due to poor road conditions; and (3) it increases ease of communication, which represents one of the major issues in breakdown of supply.

7 A Tactical/Operational Model for Transshipment in Clusters

For the strategic planning models of Sec. 4 and 5, it was necessarily assumed that recourse occurred under continuous review. In reality, transshipments would likely occur periodically at regularly scheduled intervals after stock reviews at each facility. In this section, we use MDP to develop a mechanism to operationalize the transshipment concept at a cluster level based on the strategic plan developed in the previous sections.

First, we formulate a MDP model to analyze the dynamics of a periodic review system for the clinic clusters. Second, we analyze the structure of the optimal policy under a reasonable demand assumption for clinics that are within close geographic proximity. We are able to show that balancing the load evenly across the clinic cluster is optimal, but re-balancing occurs only in a cluster-level “sweet spot” where there is not too much or too little inventory in the cluster as a whole. Third, we parameterize the model with historical demand data (as described at the beginning of Sec. 6) and solve the MDP numerically to illustrate the behavior of the model and explore unique features of aid distribution in the developing countries.

To identify clinic clusters, we solve the strategic-level transshipment model to optimality and calculate the optimal values of transshipment between two clinics under all scenarios, i.e., $y^s_{ij}$. Then we take the maximum of transshipment values across all scenarios defined as $y^\text{max}_{ij} = \max_{s \in S} y^s_{ij}$. Note here we are not using an exact clustering method (e.g., k-means). Instead, we are implying cluster structures from the results of the transshipment model. The downside to this method of developing clusters is that the results may not always be clear-cut. Depending on the actual parameter values, cluster boundaries may become a bit blurry. For instance as the cost of transshipment increases, the model tends to hold on to some inventory at the district hospitals and ship them to clinics with a delay. In another extreme case when transshipment is very inexpensive, a clinic may receive a small shipment from another clinic that is not in the same cluster. Therefore using the results of the stochastic program will not always guarantee...
that we obtain mutually exclusive clusters. However, in all those cases we observed that by eliminating shipments that were lower than a predefined threshold, we could obtain mutually exclusive clusters. To implement this idea, if there has been a “significant” transshipment between those two clinics, i.e., $y_{ij}^{\text{max}}$ exceeds a pre-determined threshold, we assume those two clinics are in the same cluster. To conduct computational experiments, we set this threshold to the average monthly demand of the receiving clinic (i.e., $j$) across all scenarios; i.e., $\sum_{s \in S} p_s d_j$.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Xi$</td>
<td>$n$-dimensional vector for the amount of inventory at each clinic at the beginning of the period</td>
</tr>
<tr>
<td>$\xi_j$</td>
<td>the $j$th component of $\Xi$, indicating how much inventory is at clinic $j$</td>
</tr>
<tr>
<td>$U_\Xi$</td>
<td>$n$-dimensional integer vector space where $u \in U$ is defined in (33), which enforces flow conservation</td>
</tr>
<tr>
<td>$u_j$</td>
<td>the $j$th component of $u \in U$, which is the action describing how much to increase or decrease clinic $j$’s inventory level via transshipment</td>
</tr>
<tr>
<td>$\Pi$</td>
<td>$n$-dimensional vector of shortage penalty</td>
</tr>
<tr>
<td>$c$</td>
<td>unit cost of transshipment between clinics</td>
</tr>
<tr>
<td>$d_n$</td>
<td>random variable for pharmaceutical demand in period $n$</td>
</tr>
<tr>
<td>$\Phi$</td>
<td>set of clinics in the cluster, a subset of $\mathcal{C}$</td>
</tr>
</tbody>
</table>

Table 3: Clinic transshipment model dynamic program notation.

In our operational model, each clinic cluster is modeled separately with a periodic review cycle in which each clinic’s inventory is surveyed and then a decision is made as to how much to transfer to other clinics within the cluster. At the beginning of each period, each clinic incurs a shortage penalty for unmet demand from the prior stage – indicated by a negative inventory value. Next, a decision is made regarding how much product to ship between clinics. Finally, demand arrives to each clinic within the cluster according to a distribution for epoch $n$ of $d_n \sim F_n$ and the state is updated for the next decision epoch. The finite-horizon MDP formulation is given in (32) with notation in Table 3. Equation (33) limits the action space to allow transshipment only if inventory is available.

$$
    f_n(\Xi) = \Pi^T(-\Xi)^+ + \min_{u \in U_\Xi} \left\{ c \sum_{j \in \Phi} (u_j)^+ + E\{f_{n-1}((\Xi)^+ + u - d_n)\} \right\},
$$

(32)

where the action space is given by:

$$
    U_\Xi = \{ u = (u_1, ..., u_s) : u_j \leq \xi_j \text{ and } \sum_{j \in \Phi} u_j = 0 \}.
$$

In the tactical model, we only focus on the clinics in one cluster, which means they are all in close proximity. This makes the distance between each clinic in the cluster approximately the same, which allows us to safely approximate $c_{ij} = c$, where $i$ and $j$ are in the same cluster. However, an advantage of the MDP approach is that it easily accommodates nonlinear cost functions and transport capacity limits if needed. The expected cost-to-go is based on the positive part of $\Xi$, because in malaria treatment, the dynamics behave as “lost sales,” not backorders.

### 7.1 Structural Properties and Insights for Clinic Cluster Transshipment

In this section we analyze several structural properties of our MDP model to gain insight into the optimal transshipment policy for clinic clusters. We specifically show that (1) the entire cluster...
is better off when any clinic in the cluster increases its initial supply, (2) balancing the inventory among clinics is optimal, (3) the optimal transshipment policy is of threshold nature.

**Individual supply benefits the group.** Theorem 7.1 shows that the entire cluster is always better off if any one of its clinics receives more supplies. This theorem supports the need for a strategic planning model that initially allocates ACTs to clusters in an effective and equitable manner.

**Theorem 7.1.** $f_n(\Xi)$ is non-increasing in $\xi_j$ for all $n$ and $j$.

**Proof:** We prove this theorem by induction.

**Base Case:** $f_0(\Xi) = 0$ for all $\Xi$ and therefore is trivially non-increasing.

**Induction Step:** Assume $f_{n-1}(\Xi)$ is non-increasing in $\Xi$.

\[
\begin{align*}
    f_n(\Xi) - f_n(\Xi - e_j) &= \Pi^T(-\Xi)^+ + \min_{u \in U_{\Xi}} \{ c \sum_{j \in \Phi} (u_j)^+ + E\{f_{n-1}((\Xi)^+ + u - d_n)\} \} - \\
    &\quad - \Pi^T((-\Xi - e_j)^+) + \min_{u \in U_{\Xi - e_j}} \{ c \sum_{j \in \Phi} (u_j)^+ + E\{f_{n-1}((-\Xi - e_j)^+ + u - d_n)\} \}.
\end{align*}
\]

$e_j$ is the unit vector with 1 in the $j^{th}$ dimension and 0’s elsewhere. We compare the Equation term by term. First, the instantaneous cost is clearly greater in the system with less inventory:

\[
\Pi^T((-\Xi)^+ - (-\Xi - e_j)^+) \leq 0. \quad (34)
\]

Next we compare the minimization term. If the optimal action, $u^*$, is the same in both $f_n(\Xi)$ and $f_n(\Xi - e_j)$, it follows from the induction hypothesis that:

\[
E\{f_{n-1}((\Xi)^+ + u^* - d_n)\} - E\{f_{n-1}((-\Xi - e_j)^+ + u^* - d_n)\} \leq 0.
\]

If, on the other hand, the optimal actions for $f_n(\Xi)$ and $f_n(\Xi - e_j)$ are different, without loss of generality we assume that optimal action in state $(\Xi - e_j)$ is $u^0$. We then have:

\[
\begin{align*}
    \min_{u \in U_{\Xi}} \{ c \sum_{j \in \Phi} (u_j)^+ + E\{f_{n-1}(\Xi + u - d_n)\} \} - c \sum_{j \in \Phi} (u_j^0)^+ - E\{f_{n-1}((-\Xi - e_j)^+ + u^0 - d_n)\} \leq \\
    c \sum_{j \in \Phi} (u_j^0)^+ + E\{f_{n-1}((\Xi)^+ + u^0 - d_n)\} - c \sum_{j \in \Phi} (u_j^0)^+ - E\{f_{n-1}((-\Xi - e_j)^+ + u^0 - d_n)\} \leq 0. \quad (35)
\end{align*}
\]

Inequality (34) follows because the minimizing action at $\Xi$ is clearly at least as small as action $u^0$. Inequality (35) follows directly from the induction hypothesis. This completes the proof. □

**Optimality of Inventory Balancing within a Cluster.** In this section we develop a model for a cluster consisting of two clinics and show that the optimal policy balances the inventory between the two clinics. In Section 7.2, we extend the insights regarding cluster balancing from the analytical model to show numerically that the same structure holds more generally by applying historical data to larger clinic clusters.
We begin by defining what it means for a function to be balanced. Next, we show that the balanced property is preserved by the expectation operator in Lemma 7.1. This lemma supports development of further operational insights including the key result (Theorem 7.2 and Corollary 7.1) that the optimal transshipment policy is of threshold nature; and depending on the cluster-wide inventory levels and the disparity between the clinics the optimal action will either (1) re-balance the inventory across the cluster so that each clinic has the same inventory level or (2) do nothing. This result is supported by deriving ancillary insights that show the optimal states for a clinic cluster possess the property that all clinics have “roughly equal” inventory levels (Lemma 7.3), and transshipment only occurs from clinics with higher inventory to clinics with lower inventory (Lemma 7.2). These last two Lemma’s also guarantee that our decision support has the appealing property of being perceived as fair by implementing clinics: no clinic with less inventory will ship to one with higher inventory, and the goal of the algorithm is to achieve inventory balance among the clinics.

As a precursor to model analysis, we begin by describing a reasonable assumption about the demand for ACTs induced by tightly clustered clinic groupings. The clinic clusters resulting from the strategic stochastic programming model are composed of clinics that are geographically very close to one another. Given the close proximity of the modeled clinics, it would be reasonable to assume that the severity of malaria outbreak will follow a similar pattern among the clinics of the same cluster. Mathematically, we take a similar pattern to mean that it is equally likely to see malaria incidence of x in clinic A and y in clinic B as it is to see incidence y in clinic A and x in clinic B. We call this a symmetric demand distribution. Here we argue that, in the presence of this assumption, we can prove certain properties of the optimal transshipment policy. Note that the MDP model we propose is capable of handling asymmetrical demand as well. However, the properties argued in this section will not hold under asymmetric demand.

We now begin our analysis with a definition of a balanced function and then proceed to show that the MDP value function is balanced, which guarantees the optimality of balancing inventory levels across the clinics within a given cluster.

**Definition 7.1.** We call a function \( f : \mathbb{R}^2 \rightarrow \mathbb{R} \) balanced if given \( \Xi \) and \( \Xi' \), such that \( \xi_1 + \xi_2 = \xi'_1 + \xi'_2 \), if \( |\xi_1 - \xi_2| \leq |\xi'_1 - \xi'_2| \) then \( f(\Xi) \leq f(\Xi') \).

In the following lemma we show for symmetric demand distributions that the expected cost-to-go function of the MDP preserves the balanced property. We then use this lemma (with proof in Appendix A) to prove structural properties of our periodic review with transshipment MDP and build up insights supporting the optimality of balanced inventory levels across the cluster.

**Lemma 7.1.** If the two clinics in a cluster have a symmetric demand distribution and the function \( f_n(\Xi) \) is balanced for all \( n \), then \( g(\Xi) = \mathbb{E}[f_n(\Xi - d_n)] \) is also balanced.

Building upon the preceding lemma, we are able to show that if transshipment occurs, it will
only occur with clinics that have more inventory shipping to those with less inventory (Lemma 7.2 stated below and proved in Appendix A). This property makes practical sense, since any solution that makes clinics with less inventory ship to clinics with more inventory would almost certainly be met with skepticism. From a mathematical standpoint, this lemma serves to simplify the action spaces and enables us to prove further structural properties of the value function. From a computational perspective, the reduced action space further enables much faster numerical solutions for the MDP presented in Section 7.2.

**Lemma 7.2.** If function \( f_n(\Xi) \) is balanced for all \( n \), the optimal action will never transship from the clinic with lower inventory to a clinic with higher inventory.

This result, combined with the following lemma, reduces the action space significantly since the optimal action \( u^* \) must be an element of \( \{0, \ldots, \lfloor 0.5 \cdot (\max\{\xi_1, \xi_2\} - \min\{\xi_1, \xi_2\}) \rfloor \} \). \( u^* \) will be the amount of medication shipped from the clinic with higher inventory to the one with lower inventory. The result of this lemma will be employed to show our key result, Theorem 7.2, which leads to the operational insight that clinics will either transship to balance inventory levels or do nothing. The next lemma (proved in Appendix A) demonstrates that a completely balanced inventory distribution is the lowest cost state for a clinic cluster. Hence each clinic cluster will desire to move toward a cluster-wide balanced inventory as long as the cost of achieving the balance is not too great (which is shown by Theorem 7.2 and Corollary 7.1).

**Lemma 7.3.** If \( f_n(\Xi) \) is a balanced function for all \( n \), then for any total inventory level \( \xi_1 + \xi_2 \), the value function is minimized where \( \xi_1^* = \xi_2^* \) if \( \xi_1 + \xi_2 \) is even and \( |\xi_1^* - \xi_2^*| = 1 \) if \( \xi_1 + \xi_2 \) is odd.

Now we are ready for the main result, which is that the optimal transshipment policy follows a threshold in which the clinics will either (1) re-balance the inventory across the cluster so that each clinic has the same inventory level or (2) do nothing. We first show that the MDP value function is balanced in Theorem 7.2. This means that the optimal solution of our MDP has the properties of Lemma 7.2 and Lemma 7.3. This property also leads directly to the main result which provides an operational structure for our inventory management system: the optimality of balancing inventory levels via a threshold policy in Corollary 7.1.

**Theorem 7.2.** When the demand vector has a symmetric distribution, the value function in (32) is balanced.

**Proof:** Without loss of generality, we assume \( \xi_1 \leq \xi_2 \) and \( \xi_1' \leq \xi_2' \). This ordering along with our assumption that \( \xi_1 + \xi_2 = \xi_1' + \xi_2' \) and \( \Delta\Xi = |\xi_1 - \xi_2| \leq |\xi_1' - \xi_2'| = \Delta\Xi' \), implies that \( \xi_1' \leq \xi_1 \leq \xi_2 \leq \xi_2' \). We proceed to prove this theorem by using induction.

**Base Case:** Since \( f_0(\Xi) = 0 \) for all \( \Xi \). As a result, the induction hypothesis holds.

**Induction Step:** We assume that the induction hypothesis holds for stage \( n - 1 \). In order to show
$f_n(\Xi)$ is less than or equal to $f_n(\Xi')$, we first write the expressions for both cases:

$$f_n(\Xi) = \Pi^T(-\Xi)^+ + \min_{u \in U_\Xi} \{c \sum_{j \in \Phi} (u_j)^+ + E[f_{n-1}(\Xi') + u - d_n]\},$$

$$f_n(\Xi') = \Pi^T(-\Xi')^+ + \min_{u \in U_{\Xi'}} \{c \sum_{j \in \Phi} (u_j)^+ + E[f_{n-1}((\Xi')^+ + u - d_n)]\}.$$

Let $\bar{\xi} = [\xi_1 + \xi_2]$. Applying the inductive hypothesis that $f_{n-1}$ is a balanced function, and Lemma 7.1 and Lemma 7.3, at stage $n - 1$ the state $[\bar{\xi}, \xi_1 + \xi_2 - \bar{\xi}] = [\xi_1, \bar{\xi}]$ achieves the minimum value for the expectation of the function $f_{n-1}(\cdot)$. By the induction hypothesis we have $f_{n-1}(\Xi') \geq f_{n-1}(\Xi)$. Since $\xi_1 + \xi_2 = \xi_1' + \xi_2'$, we can conjecture that $f_{n-1}(\Xi)$ is the state which reaches the minimum possible cost.

$$f_n(\Xi') - f_n(\Xi) = \Pi^T(-\Xi')^+ + \min_{u \in U_{\Xi'}} \{c \sum_{j \in \Phi} (u_j)^+ + E\{f_{n-1}((\Xi')^+ + u - d_n)\}\} -$$

$$\Pi^T(-\Xi)^+ - \min_{u \in U_\Xi} \{c \sum_{j \in \Phi} (u_j)^+ + E\{f_{n-1}((\Xi)^+ + u - d_n)\}\}.$$

First we compare the instant cost associated with the shortage penalties. As we mentioned before, without loss of generality, we consider the cases where $\xi_1 \leq \xi_2$ and $\xi_1' \leq \xi_2'$. Other cases can be investigated similarly. There are the following cases:

1. $\xi_1', \xi_2', \xi_1, \xi_2 \geq 0$: In this case both $f_n(\Xi)$ and $f_n(\Xi')$ incur zero shortage penalties. As a result $\Pi^T(-\Xi')^+ - \Pi^T(-\Xi)^+ = 0$.

2. $\xi_1' \leq 0$ and $\xi_2, \xi_1, \xi_2' \geq 0$: In this case $f_n(\Xi')$ has a positive shortage cost while $f_n(\Xi)$ incurs zero shortage penalty. Therefore $\Pi^T(-\Xi')^+ - \Pi^T(-\Xi)^+ \geq 0$.

3. $\xi_1', \xi_1 \leq 0$ and $\xi_2, \xi_1', \xi_2' \geq 0$: In this case $f_n(\Xi')$ and $f_n(\Xi)$ both have positive shortage cost but since $\xi_1' \leq \xi_1$, $f_n(\Xi)$ has a greater shortage penalty. As a result $\Pi^T(-\Xi')^+ - \Pi^T(-\Xi)^+ \geq 0$.

4. $\xi_1', \xi_1, \xi_2 \leq 0$ and $\xi_2', \xi_2 \geq 0$: We conclude that having $\xi_2' \geq 0$, will result in $\xi_1' \leq \xi_1 + \xi_2$ as a direct result of the assumption, since $\xi_1 + \xi_2 = \xi_1' + \xi_2'$. Therefore $\Pi^T(-\Xi')^+ - \Pi^T(-\Xi)^+ \geq 0$.

5. $\xi_1', \xi_1, \xi_2, \xi_2' \leq 0$: this implies that $\Pi^T(-\Xi')^+ = \Pi^T(-\Xi)^+ \geq 0$.

The next step is to investigate the possible actions and compare the cost-to-go terms for both $f_n(\Xi)$ and $f_n(\Xi')$. Let assume the optimal action in state $\Xi'$ is $u^*$. Having $\xi_1' \leq \xi_2'$ and based on the result of Lemma 7.2, $u^* \in \{0, ..., u_\bar{\xi}\} = U_{\Xi'}$ and $u^* = (u^*, -u^*)$ , the optimal action either will be to ship from clinic 2 to clinic 1 or do nothing (result of Lemma 7.2). We also know that $\xi_2' - \xi_1' \geq \xi_2 - \xi_1$, as a result, the optimal action $u^*$ of the state $\Xi'$ will be a member of $\{0, ..., u_\bar{\xi}\}$. We have:

$$U_{\Xi} \subset U_{\Xi'}$$

There are two possible scenarios, either $u^* \in U_{\Xi}$ or $u^* \in U_{\Xi'} \setminus U_{\Xi}$. 

28
1. Scenario 1: \( u^* \in \mathcal{U}_\Xi \):

\[
\begin{align*}
&Q^1 \geq 0 \\
&f_n(\Xi') - f_n(\Xi) \geq \Pi^T(-\Xi')^+ - \Pi^T(-\Xi)^+ + cu^* + E\{f_{n-1}(\Xi')^+ + u^* - d_n)\} - \min_{u \in \mathcal{U}_\Xi}(c \sum_{j \in \Phi}(u_j)^+ + E\{f_{n-1}(\Xi)^+ + u - d_n)) \geq \\
&Q^1 + cu^* + E\{f_{n-1}(\Xi')^+ + u^* - d_n)\} - cu^* + E\{f_{n-1}(\Xi)^+ + u^* - d_n)\} = \\
&E\{f_{n-1}(\Xi')^+ + u^* - d_n)\} - E\{f_{n-1}(\Xi)^+ + u^* - d_n)\} \geq 0. \text{ by induction hypothesis}
\end{align*}
\]

2. Scenario 2: \( u^* \in \mathcal{U}_\Xi \setminus \mathcal{U}_\Xi \):

\[
\begin{align*}
&Q^2 \geq 0 \\
&f_n(\Xi') - f_n(\Xi) \geq \Pi^T(-\Xi')^+ - \Pi^T(-\Xi)^+ + cu^* + E\{f_{n-1}(\Xi')^+ + u^* - d_n)\} - \min_{u \in \mathcal{U}_\Xi}(c \sum_{j \in \Phi}(u_j)^+ + E\{f_{n-1}(\Xi)^+ + u - d_n)) \geq \\
&Q^2 + cu^* - cu_{\xi} + E\{f_{n-1}(\Xi')^+ + u_{\xi}^* - d_n)\} - E\{f_{n-1}(\Xi)^+ + u_{\xi} - d_n)\} \geq 0. \text{ by induction hypothesis}
\end{align*}
\]

We should note that \( B \geq 0 \) since the total units shipped from clinic 2 to clinic 1 in scenario 2 is more than \( u_{\xi} \). Also after transshipping \( u_{\xi} \), \( f_{n-1}(\Xi) \) is reaching its minimum (as a direct result of Lemma 7.3), therefore \( C \geq 0 \). This ends the proof. \( \square \)

**An Optimal Threshold Policy for Inventory Balancing.** The lowest cost state for the clinic cluster is a balanced inventory level. However, to achieve a balanced state in each epoch requires paying a transshipment cost, so it may not be optimal to re-balance the cluster in every epoch. This section provides the key insight that the clinics should follow a threshold policy that re-balances inventory when the difference between inventory levels is above a certain threshold, but will not re-balance if both clinics have either too little inventory or a surplus of inventory. Figure 11 provides a typical example of the optimal transshpement areas. In Area 1 there is not enough inventory within the cluster (shortages being likely at both clinics) and in Area 3 there is sufficient inventory in the cluster (shortages being unlikely at either clinic); hence no transshipment occurs. In Area 5, Clinic 1 has surplus inventory while Clinic 2 does not have enough, with the reverse occurring in Area 4, and so re-balancing occurs in both Area 4 and Area 5. The structure demonstrated in Figure 11 is guaranteed by the following Corollary, which follows directly from Theorem 7.2.

**Corollary 7.1.** Under a non-decreasing shipping cost, the optimal policy is of threshold nature with stage-dependent thresholds. Depending on the shipping cost, the optimal action will perform the minimal amount of transshipment necessary to balance the inventory (in the sense of Definition 7.1) or do nothing.
As an example of Corollary 7.1, consider the case where there is a fixed cost per shipment. The optimal policy balances the inventories between the two clinics in the following way: $u = 0$ if $c > E\{f_{n-1}((\Xi)^+ - d_n)\}$; otherwise $u$ is the optimal action that brings the inventory levels of the clinics to $\lfloor \frac{\xi_1 + \xi_2}{2} \rfloor$ and $\xi_1 + \xi_2 - \lfloor \frac{\xi_1 + \xi_2}{2} \rfloor$. Any analytical proof regarding the structure of an optimal policy for clusters consisting of more than two clinics can be complex.

### 7.2 Illustrative Example of the Optimal Area-based Transshipment Policy

In this section, a numerical example is used to gain insight into state-specific optimal actions. For the purpose of exposition, we begin with an example of a cluster consisting of two clinics. We also scale the units of demand and supply to obtain the following restricted state space:

$$\Xi = \left\{ (\xi_1, \xi_2) \in \mathbb{R}^2 : -5 \leq \xi_i \leq 9, \forall i \in \{1, 2\} \text{ and } \xi_1 + \xi_2 \leq 9 \right\}.$$

We solve the two-dimensional MDP under three different parameter settings where the ratio of shortage penalty to unit transportation cost was either: (1) low, (2) moderate, or (3) high and solve it for six stages (one stage for each month of the malaria season). Figure 11 illustrates the optimal actions at stage $5$ (i.e., $n - 1$) for each state for cases (1), (2) and (3). The optimal actions in Figure 11 are identified by five areas, 1 through 5, described in more detail in Table 4.

<table>
<thead>
<tr>
<th>Area</th>
<th>Description</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Clinics 1 and 2 face shortage</td>
<td>No action is possible</td>
</tr>
<tr>
<td>2</td>
<td>Clinics have low inventories</td>
<td>No action is recommended</td>
</tr>
<tr>
<td>3</td>
<td>Both clinics have high surplus</td>
<td>No action is required</td>
</tr>
<tr>
<td>4</td>
<td>Clinic 1 is facing shortage while clinic 2 has surplus</td>
<td>Clinic 2 transships to clinic 1</td>
</tr>
<tr>
<td>5</td>
<td>Clinic 2 is facing shortage while clinic 1 has surplus</td>
<td>Clinic 1 transships to clinic 1</td>
</tr>
</tbody>
</table>

Figure 11: Optimal actions in period 5 for three parameter settings.

Table 4: Five areas in the two-dimensional illustrative example.

Fig. 11 shows that the ratio of shortage penalty to transportation cost ($\pi/c$) plays an important role. As this ratio increases, there is more transshipment between clinics; transshipment Areas 4 and 5 becomes larger while no action Areas 2 and 3 shrink. As the $\pi/c$ ratio decreases we observe less transshipment, which has the opposite affect on the areas. This behavior demonstrates the importance of low cost and accessible shipping options for short distance transport as this leads to more effective transshipment policies.
In Section 7.1 we found the structure of an optimal transshipment policy with two-clinic clusters and symmetric demand. While the setup considered was reasonable both in the demand assumption (as argued previously) and size (a number of clusters from the strategic model contained only two clinics), we can further extend the analytical results to clusters containing 3 clinics through numerical analysis. The insights are summarized below:

**Insight 7.1.** Corollary 7.1 extends to clusters of size greater than two. When the demands of all the clinics in a cluster are symmetrically distributed, the optimal action is to balance the inventory between the clinics in the cluster.

**Insight 7.2.** As the ratio of the shortage penalty to the transshipment cost increases, it is optimal to ship more units between the clinics; increasing the effectiveness of transshipment in preventing ACT shortage.

Similar to the previous example, we scale the units of demand and supply to obtain the following restricted state space:

$$\Xi = \left\{ (\xi_1, \xi_2, \xi_3) \in \mathbb{R}^3 : -5 \leq \xi_i \leq 9, \forall i \in \{1, 2, 3\} \text{ and } \xi_1 + \xi_2 \leq 9 \right\}.$$

The system state has three dimensions, so we only illustrate the optimal actions for three inventory levels at clinic 3: 0, 2, and 4. For ease of comparison, we chose a moderate ratio of shortage penalty to transportation cost, i.e., $\pi/c = 10$. The optimal actions in Figure 12 are identified by five areas, detailed in Table 5. Figure 12, shows that the optimal policy balances the inventory between the three clinics when demand is symmetric. As the inventory level of clinic 3 increases, that clinic ships more pharmaceutical units to clinics 1 and 2 if needed.

![Figure 12: Optimal actions in period 5 for a cluster consisting of three clinics.](image)

<table>
<thead>
<tr>
<th>Area</th>
<th>Description</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Clinics 1 &amp; 2 face shortage</td>
<td>If 3 has surplus, transships to 1 and 2, no action otherwise</td>
</tr>
<tr>
<td>2</td>
<td>Clinics 1 &amp; 2 have low inventories</td>
<td>If 3 has surplus, transships to 1 or 2, no action otherwise</td>
</tr>
<tr>
<td>3</td>
<td>Clinics 1 &amp; 2 have high surplus</td>
<td>1 &amp; 2 transship to 3 if needed</td>
</tr>
<tr>
<td>4</td>
<td>Clinic 2 has surplus</td>
<td>Clinic 2 transships to clinics 1 and/or 3 if needed</td>
</tr>
<tr>
<td>5</td>
<td>Clinic 1 has surplus</td>
<td>Clinic 1 transships to clinics 1 and/or 3 if needed</td>
</tr>
</tbody>
</table>

Table 5: Five areas in the three-dimensional illustrative example.
7.3 Clinic Clustering vs. Fully Integrated Optimization

In Sec. 6.6 we introduced a method for integrating the strategic and tactical levels of supply allocation through the decomposition of the strategic model into clinic clusters based on the structure of the solution. While this approach has significant computational advantages (the tactical MDP is intractable at the level of the full-scale problem), a question remains regarding loss of optimality stemming from the cluster-based decomposition. In this section, we address this issue by studying four larger groups of clinics for which the strategic model recommended cluster decomposition. Through numerical analysis, we compare for each group (1) the fully integrated solution, (2) the cluster decomposition solution suggested in Sec. 6.6, (3) and the solution of the baseline (i.e., naive) model from Sec. 3.

For the fully integrated solutions, we solve a full-scale MDP that allows transshipment between any pair of clinics within the group of clinics studied and then selects the optimal initial inventory to minimize the total cost using an exhaustive search. For the cluster decomposition solution, we first run the strategic optimization to identify the optimal clusters within the larger group and initial allocation of inventory within each cluster (as described in Sec. 6.6). Next we solve an MDP for each cluster separately, only allowing transshipment between clinics within the same cluster. For the baseline model, we simply solve the baseline optimization and allocate inventory accordingly. In solving the MDP, we used the actual demand patterns (scaled down for tractability) at each clinic to incorporate (1) non-stationary demand by month during the 6 month malaria season, and (2) demand variability by year by including very high, high, medium, low, and very low years. The same approach is used in the other analyses of this paper.

In identifying the four larger groups of clinics to study, we first identified groups of four and five clinics that (1) were all in close proximity to one another and (2) were split into two clusters based on the global strategic solution (containing all 290 clinics). We capped the group size at five clinics because analyzing any larger group of clinics causes the fully integrated solution to be intractable due to the curse of dimensionality. Further, including larger groups usually entails groups of clinics with significant distance between clusters, in which case the optimal solution would almost never use that shipping route (as seen in the strategic solution). Hence, this analysis should be sufficient to capture the key comparison of full integration versus the decomposition approach. The following table shows distances in kilometers between clinics in each of the four master groups.

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
<th>Group 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>139</td>
<td>142</td>
<td>143</td>
<td>274</td>
</tr>
<tr>
<td>139</td>
<td>142</td>
<td>143</td>
<td>274</td>
</tr>
<tr>
<td>139</td>
<td>142</td>
<td>143</td>
<td>274</td>
</tr>
<tr>
<td>139</td>
<td>142</td>
<td>143</td>
<td>274</td>
</tr>
</tbody>
</table>

Table 6: Intraclinic distances for four larger groups of clinics that can be decomposed into clinic clusters. Data in matrix form with clinic number as the row and column headers.
The clusters derived from the strategic solutions for each group were as follows. Group 1: Cluster 1 = Clinic 274 and 285; Cluster 2 = Clinic 139, 142, and 143. Group 2: Cluster 1 = Clinic 249 and 252; Cluster 2 = Clinic 1, 2, and 263. Group 3: Cluster 1 = Clinic 12 and 13, Cluster 2 = Clinic 256 and 265. Group 4: Cluster 1 = Clinic 3 and 4; Cluster 2 = Clinic 225 and 252.

We now present the optimality gap of both the cluster-based decomposition strategy and the baseline modeling approach compared to the fully integrated optimal solution. We analyze different starting levels of total inventory for the entire group. Table 7 presents the results for the four groups for very low initial inventory levels (10) up to high initial inventory levels (32). Beyond this size of initial inventory the fully integrated model for 5 clinic groups (group 1 and 2) became intractable. There are several key observations from the table. First is that the percent optimality gap for the proposed decomposition heuristic is very small for both 4 and 5 clinic groups, typically between 0-1%, with the average gap being 0.5% and the maximum only reaching 3.2%. Second, the gap for the decomposition heuristic remains stable as the initial inventory increases, whereas the gap for the baseline model grows at an increasing rate. When the initial inventory is low, all models can use nearly all of the initially allocated demand. However, as the inventory increases, the less flexible baseline model is at an increasing disadvantage and ends up wasting inventory. The decomposition heuristic, on the other hand, continues to track the fully integrated model closely because the clinic clusters created by the decomposition heuristic have the property that cross-cluster shipping is generally undesirable so the fully integrated model rarely uses these shipping lanes. Hence, the fully integrated model behaves like the clustered model in most cases, which leads to the small optimality gap.

Table 7: Optimality gap (%) for Decomposition Heuristic and Baseline Optimization versus the Fully-Integrated Optimization for clinic groups 1-4

<table>
<thead>
<tr>
<th>Initial Inventory In Group</th>
<th>10</th>
<th>12</th>
<th>14</th>
<th>16</th>
<th>18</th>
<th>20</th>
<th>22</th>
<th>24</th>
<th>26</th>
<th>28</th>
<th>30</th>
<th>32</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>Decomp</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>0.4</td>
<td>1.1</td>
<td>1.1</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>3.5</td>
<td>4.8</td>
<td>7.4</td>
<td>10.6</td>
<td>14.4</td>
<td>19.0</td>
<td>26.3</td>
<td>37.7</td>
<td>52.1</td>
<td>70.6</td>
<td>95.1</td>
</tr>
<tr>
<td>Group 2</td>
<td>Decomp</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>0.2</td>
<td>0.2</td>
<td>0.3</td>
<td>0.5</td>
<td>0.8</td>
<td>1.9</td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>1.8</td>
<td>3.9</td>
<td>6.5</td>
<td>9.7</td>
<td>13.6</td>
<td>18.5</td>
<td>26.5</td>
<td>34.3</td>
<td>49.7</td>
<td>69.9</td>
<td>96.2</td>
</tr>
<tr>
<td>Group 3</td>
<td>Decomp</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>0.1</td>
<td>0.3</td>
<td>0.4</td>
<td>0.5</td>
<td>2.0</td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>1.7</td>
<td>2.8</td>
<td>5.1</td>
<td>7.9</td>
<td>11.3</td>
<td>15.5</td>
<td>22.4</td>
<td>28.9</td>
<td>41.8</td>
<td>58.4</td>
<td>79.3</td>
</tr>
<tr>
<td>Group 4</td>
<td>Decomp</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>0.1</td>
<td>0.5</td>
<td>0.3</td>
<td>0.7</td>
<td>1.4</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>1.7</td>
<td>2.8</td>
<td>5.2</td>
<td>8.0</td>
<td>12.8</td>
<td>17.3</td>
<td>22.8</td>
<td>31.7</td>
<td>45.0</td>
<td>62.3</td>
<td>85.1</td>
</tr>
</tbody>
</table>

7.4 Operational Considerations and Insights for Pharmaceutical Distribution in Developing Countries

As mentioned earlier in Section 1.2, drug distribution and transshipment between clinics within a centrally controlled (government) distribution network in the developing world has several distinguishing features that contrast this environment with that studied in the traditional transshipment literature. These features include ethical considerations not present in typical supply chains, periodic review with lengthy intervals due to time consuming paper-based inventory
calculation methods, geographically and temporally correlated and variable demand including seasonality due to characteristics of malaria. To capture the correlation, we modeled the joint distribution of the demand at two clinics that formed a cluster. For these two clinics, we observed in our data the phenomenon that a high (low) demand at one clinic was more likely to coincide with a high (low) demand at the other. In this section we provide insights into how these features impact the system cost of a transshipment scheme by performing extensive numerical experiments on a two clinic cluster formed in the optimal solution of the stochastic program from Sec. 6.

The two clinics (Clinic 124 and Clinic 133 in our dataset) in the chosen cluster are from the eastern portion of central Malawi to the northeast of the capital city of Lilongwe. As with previous numerical experiments, we calculate the transshipment cost at $0.45 per unit (determined by distance and transportation cost per km) and the shortage cost to be $20. The ratio is about 50:1, which is in keeping with the middle range ratio of our previous computational experiments. Demand was taken from monthly case counts at the two facilities, which had similar levels of demand. We ran the experiments for a three month period, since it is unlikely that the clinics will deplete their supplies during the first few months of the malaria season and transshipment only has a major effect when supplies run low at the clinics. When analyzing the system cost of each solution we present an average and a maximum cost over all reasonable starting inventory levels at Clinic 124 and 133.

From these experiments we have the following findings: (1) Ethical policies are very close to optimal when clinics are clustered with other clinics nearby, (2) Seasonality (predictable variability) is easily handled by transshipment, but random demand variability has a significant impact on system cost, (3) transshipment is effective even with infrequent intervals under periodic review. We also propose a simple heuristic that can be easily implemented and appears to generate near optimal results.

**Ethical Transshipment.** In delivering pharmaceuticals to patients in need, a prime consideration is fair and ethical distribution. We consider two different ethical strategies based on how quickly transshipment can occur. The “chase strategy” assumes that if there is shortage in one clinic in a given period, the other clinic is able to (and will) immediately transship available supply to satisfy that demand in the same period that it occurs. In the second scenario we assume (as we do in the previously described formulation of the MDP) that demand at any clinic that cannot be satisfied immediately is lost and a shortage penalty is incurred. The “balanced” strategy will transship in each period to rebalance the supply at the clinics in the cluster to observe the ratio of expected demand at each clinic. For example, if both clinics have the same expected demand over future periods, then after transshipment in each period both clinics will have the same supply (+/- 1). This strategy is based on the structural results from Thm. 7.2 and is also equitable since ensures both clinics receive equal supply (relative to their expected
demand) in each period.

The result of these experiments shows that the ethical solution is very close to optimal. The cost gap between the ethical balanced (chase) solution and the optimal solution was $12.3 ($7.7), or 0.6% (0.4%) of the total cost, over a three month time period on average, which is the cost equivalent of 0.4 shortages over three months. The maximum value was $40.6 ($39.9), or 2.1% (2.1%) of the total cost, which is equivalent to 2 shortages over three months. One reason why the gap between the ethical and optimal solutions is so small is because the cost of a shortage is high relative to the cost of transshipping. Thus the optimal policy generally strives to satisfy all demand. The two policies diverge when inventory at the transshipping clinic was also low, and it would be likely to experience high demand at that clinic in the near future. In this case, it actually makes sense to hold inventory back at the primary clinic because the inventory will be used anyway, but without the additional cost of transshipping.

Fig. 13 shows the gap between the ethical and optimal solutions for both (a) the chase strategy and (b) the inventory balancing strategy as a function of initial inventory levels at clinic 1 and 2. In both strategies there is almost no difference when the inventory is initially distributed evenly between both clinics, with the highest gap occurring when the initial inventory level is highly imbalanced. This occurs because both ethical strategies attempt to restore the balance and will often incur unnecessary shipping cost rather than anticipating future demands. Note that with the balanced strategy, the cost gap rises again if initial inventories are very high. This is because a completely balanced inventory is not needed to satisfy all the future demand, so the ethical solution ends up overshipping supply that will not be needed.

**Demand Variability and Seasonality.** Because malaria is seasonal, demand for ACTs follows a seasonal pattern with variability from season to season as well as geographically correlated demand. We model these features through numerical experiments and investigate the impact of predictable variation (seasonality) and random variation (year to year) in the presence of geographically correlated demand. To capture seasonality, we take the actual demand pattern for Clinics 124 and 133 from the monthly case count data. We compare this with a pattern that has the same mean demand in each time period (and the same total demand as the seasonal
pattern over the planning horizon). To capture random variation, (1) we model demand as a Normal random variable with the case counts representing the mean demand, and (2) we vary the standard deviation to achieve different coefficients of variation ($CV = \sigma/\mu$): low ($CV=0.5$), medium ($CV=1$), and high ($CV=2$). We then took discrete points on a grid of +/- 0, 0.25, and 0.5 standard deviations to generate 5 different demand scenarios for each period in the MDP. The intervals were chosen to provide significant dispersion while avoiding negative demand scenarios. We also consider the interaction between random and seasonal variation.

We find that, while random variation has a significant impact on the objective, the demand seasonality brought about by the characteristics of malaria is handled well by transshipment. The average (maximum) cost gap between the seasonal and flat demand patterns was $14 ($58) over a three month time period on average, or 1.2% (4.9%) when compared with the baseline cost, which is the cost equivalent of 0.7 (2.9) shortages over three months. Fig. 14 (a) shows that seasonality has little effect if there are either high or low levels of initial inventory in the cluster, with the largest impact occurring if there are moderate levels of inventory. Fig. 14 (c) demonstrates that random variability has a much larger impact on total cost than seasonality. Also observe that as random variation increases, so does the impact of seasonality (black solid line versus dashed gray line).

The structure of the cost gap for low ($CV=0.5$) versus high ($CV = 2$) random variation relative to initial starting inventory at Clinics 124 and 133 is nearly identical to the seasonality gap, but the magnitude is much higher. The average (maximum) cost gap between the low and high variation demand patterns was $146 ($686), or 12% (57%) of the baseline cost, over a three month time period on average, which is the cost equivalent of 7.3 (34.3) shortages over three months. Note for both seasonal and random variation, the initial inventory imbalance has little effect on the cost gap.

**Periodic Review Interval.** In the developing world, pharmaceutical inventories at distributing clinics are often taken using a paper-based system rather than an electronic inventory. This process can be time consuming, so for practicality it is better to consider periodic system review rather than continuous review. We study the impact of the transshipment frequency by varying the length of the periodic review interval over a three month span and comparing with a daily transshipment policy (similar to continuous review).
Fig. 15 (a) shows the cost gap going from daily transshipment to once every ten days. There is little impact if there is too little inventory or an abundance of inventory and initial inventory imbalance does not play a big role for the same reasons described in the previous scenario. The average (maximum) cost gap between transshipping daily versus once every ten days was $71 ($267), or 3% (11%) of the baseline cost, over a three month time period on average, which is the cost equivalent of 3.6 (13.4) shortages over three months. Fig. 15 (b) shows the cost gap between various transshipment intervals and the daily transshipment interval for high, medium, and low levels of initial supply. The frequency of transshipment does have an impact, but not nearly as large as that of random variability with an average gap equivalent to only one extra shortage each month. This is encouraging for developing countries where continuous or frequent review enabled by electronic tracking is not likely to be available.

**Insights and Policy Design.** The results of our numerical experiments yield some interesting insights useful in designing an effective distribution strategy. First, the largest cost impact comes from year to year variability in demand. Better data collection and forecasting can help reduce some of this uncertainty. Second, transshipment frequency did not have a large impact on cost. Third, an ethical strategy that balances inventory between the clinics in a cluster at each transshipment period is very close to optimal. Further, if the initial inventories start out balanced, then there is almost no difference between the ethical inventory balancing policy and the optimal policy. As an implementable policy, inventory balancing also has many desirable attributes. First, it is simple and easy to implement. Second, it is likely to be perceived as fair by all parties involved. Finally, it is very close to the optimal policy. For these reasons, we posit that an inventory balancing (relative to average demand at each clinic) scheme could be an effective method for implementing transshipment within clusters.

### 7.5 Methods for Implementing Transshipment Policies within a Clinic Cluster

From the analysis of the MDP model presented in this section we now propose a practical operational policy for managing ACT distribution within a clinic cluster. This policy would set up transshipment areas based on the amount of inventory at each clinic in the cluster that
would dictate how many units to transship. These areas would be based on a number of factors. First, the areas would consider the total number of ACTs allocated to the cluster from the strategic planning model. The cost of transshipment can be adjusted to include both the actual shipping cost and the willingness of clinics to transship. If clinics are opposed to transshipping, this cost would increase and the transshipment areas would shrink. However, clinics are likely to understand that if they are unwilling to transship, they also will not receive transshipments when needed due to the cluster level policy. This should encourage their willingness to participate in a transshipment scheme.

In support of transshipment as a valid operational policy, neighboring Zambia already has clinic-to-clinic transshipment policies in place (see Mtonga (2010)). According to Zambia’s Ministry of Health Standard Operating Procedure (Mtonga (2010)) and interviews of a person familiar with the drug distribution system in Zambia, ministry of health clinics use a paper-based system to keep a monthly record of 3 months average consumption and stock levels. This is used to calculate restocking levels. At any point in time, most clinics are in short supply of some drugs and have a surplus of others. The district level health office, which is in charge of 20-30 clinics, distributes medications from the clinics that have surplus to the clinics with shortages. Where all clinics have a low inventory, the district level health office will typically try to balance the shortages across clinics. Notably, transshipment (or transferring drugs between clinics) is already being practiced on an ad-hoc basis in Malawi itself as confirmed by conversations with local professionals and Kiczek et al. (2009).

There is an opportunity to leverage this type of stock count communication scheme. After the transshipment areas are designed and agreed upon, the district level office can use a decision support system to guide periodic transshipment of ACTs. The district office would periodically calculate stock levels for each clinic within a cluster. Determining the zones of the cluster would then guide the district managers as to whether or not transshipment should occur and how much product to transship between clinics. This decision support system could replace or augment the heuristic approach currently being employed in Malawi and Zambia to better serve the populations of each clinic cluster. Further, because the clinic clusters identified in Sec. 6.6 are generally small, the incentive to share between them is greater because the populations served by these clusters are all neighbors. It may be possible to establish authority over each cluster to manage the within-cluster transshipment as a method to increase adoption.

8 Conclusion and Future Research

This paper addresses the challenging problem of distributing pharmaceutical products with seasonal demand through a centralized public health delivery system common to developing countries. Specific challenges to distributing pharmaceuticals in countries such as Malawi include: under-developed transportation infrastructure, spatially and temporally uncertain demand, and limited financial resources. This paper develops an analytical approach to effective distribution
of malaria drugs, integrating strategic level planning (where the planning horizon spans through the malaria season) with a tactical (periodic) level transshipment optimization. We do so by decomposing the national problem into localized clinic clusters. This enables a tractable solution to the periodic review tactical MDP. Through analysis of the MDP’s structural properties, we show that a simple area-based transshipment policy could easily support transshipment decisions even in a paper-based inventory management environment. This decomposition heuristic was shown to be very near the optimal when compared with a fully integrated optimization that would be intractable at the national scale. Further, the clinic clusters identified in the optimal solution of our strategic model can be a novel mechanism to overcoming political concerns while taking full advantage of the transshipment approach.

Using the strategic and tactical models, we explore several unique features of aid distribution in the developing world through a set of computational experiments using compiled estimates of facility level malaria counts for Malawi. Our initial results show that strategic planning can reduce expected ACT shortages by at least 16% while controlling transportation costs. We further show that the optimal transshipment solution cost converges to the delayed shipment solution cost as the clinic-to-clinic transshipping becomes more expensive. However, transshipment is more effective in dealing with poor infrastructure and bad road conditions prevalent in the developing world. Investigating other features of the medical aid distribution in the developing world, we found that ethical strategies are near optimal for geographically proximate clinic clusters, transshipment is fairly robust to the length of the periodic review interval that may be imposed by paper-based inventory systems in the developing world, transshipment solutions are robust to seasonality (as in the case of malaria), and year to year variation has the largest impact of any of the factors mentioned previously.

A challenge regarding the successful application of the methods offered is in the implementation itself. Successful implementation would require proactive efforts from the Ministry of Health of Malawi to improve the current system. The problems of inadequate communication and transportation infrastructure could hinder full implementation. A low-cost, cell phone based reporting system such as SMS For Life in Tanzania could at least overcome the former. Given the small and easy to carry nature of malaria medications and the relatively short distance between facilities, the latter might be overcome simply by sending goods through public transport such as minibuses or by taking medications directly to facilities by bicycle. It is difficult to foresee how these methods, which rely on a connected and cooperative system of facilities, could be utilized in the private sector in Malawi, which represents a patchwork of small sole-proprietorships which might be hesitant to transship goods for free and even more hesitant to allow other shops to take away potential customers. It is possible that innovative mechanisms enabling compensated transshipment might be developed, though that topic is outside the scope of this paper.

In conclusion, the result of our integrated strategic and tactical models is a workable decision
support system approach that can guide government policy in driving better health outcomes at a lower cost. Our methods could be extended to fit pharmaceutical products for other diseases such as diarrhea, dengue fever, and influenza. Further, the methods offered here could be applied not only to public sector supply chains but also to NGO or private sector supply chains for products which have seasonal and geographic demand heterogeneities.

References


## Appendices

### A Proofs.

**Lemma 7.1.** *Proof:* Without loss of generality let $d_n$ be distributed as $q_{i,j}$ for $i, j = 1, \ldots, n$ where $q_{i,j} = q_{j,i}$ is the probability of observing $i$ units of demand in clinic 1 and $j$ units of demand in clinic 2 and vice versa. For notational convenience, for any state $\Xi = (\xi_1, \xi_2)$, define $\Delta\Xi = |\xi_1 - \xi_2|$ as the absolute difference between the inventory at the two clinics. Consider two different states, $\Xi$ and $\Xi'$, such that $\xi_1 + \xi_2 = \xi'_1 + \xi'_2$ and $\Delta\Xi \leq \Delta\Xi'$. We now show that
\[ E[f_n(\Xi') - d_n)] - E[f_n(\Xi - d_n)] \geq 0. \]

\[ E[f_n(\Xi' - d_n)] - E[f_n(\Xi - d_n)] = \sum_{i=1}^{n} \sum_{j=1}^{n} q_{i,j} f_n(\Xi' - ie_1 - je_2) - \sum_{i=1}^{n} \sum_{j=1}^{n} q_{i,j} f_n(\Xi - ie_1 - je_2) \]

\[ = \sum_{i=1}^{n} \sum_{j=i}^{n} \left( q_{i,j} [f_n(\Xi' - ie_1 - je_2) - f_n(\Xi - ie_1 - je_2)] + q_{j,i} [f_n(\Xi' - je_1 - ie_2) - f_n(\Xi - je_1 - ie_2)] \right). \]  

(36)

We now perform a term by term comparison of (36). Without loss of generality assume that \( \xi_1 < \xi_2 \) and \( \xi'_1 < \xi'_2 \). First note that if \( j - i \leq \Delta \Xi \) then both terms within the sum are positive. Otherwise it is possible that \( f_n(\xi'_1 - i, \xi'_2 - j) - f_n(\xi_1 - i, \xi_2 - j) \) is negative, while \( f_n(\xi'_1 - j, \xi'_2 - i) - f_n(\xi_1 - j, \xi_2 - i) \) remains positive. What we show is that the magnitude of the negative portion is smaller than the magnitude of the positive portion, which implies that the sum of the negative and positive portions will be non-negative. To do so we consider two cases:

**Case 1: \( \Delta \Xi < j - i < \Delta \Xi' \).**

First, when \( j \) is not too much larger than \( i \) we show that \( f_n(\xi'_1 - i, \xi'_2 - j) - f_n(\xi_1 - i, \xi_2 - j) \geq 0 \) so the sum of all 4 terms will be positive. In the cases where \( f_n(\xi'_1 - i, \xi'_2 - j) - f_n(\xi_1 - i, \xi_2 - j) < 0 \), we show that the imbalance between states \( (\xi'_1 - i, \xi'_2 - j) \) and \( (\xi_1 - i, \xi_2 - j) \) is smaller than the imbalance between the states of the positive terms: \( (\xi'_1 - j, \xi'_2 - i) \) and \( (\xi_1 - j, \xi_2 - i) \). This directly implies, by the fact that \( f_n \) is balanced, that \( |f_n(\Xi' - ie_1 - je_2) - f_n(\Xi - ie_1 - je_2)| < |f_n(\Xi' - je_1 - ie_2) - f_n(\Xi - je_1 - ie_2)| \) and therefore:

\[ f_n(\Xi' - ie_1 - je_2) - f_n(\Xi - ie_1 - je_2) + f_n(\Xi' - je_1 - ie_2) - f_n(\Xi - je_1 - ie_2) \geq 0. \]

The imbalance for each term of the sum in (36) is given below.

\[ \Delta(\Xi' - ie_1 - je_2) = \xi'_2 - j - \xi'_1 + i = \Delta \Xi' - (j - i), \]  

(37)

\[ \Delta(\Xi - ie_1 - je_2) = \xi_1 - i - \xi_2 + j = -\Delta \Xi + (j - i), \]  

(38)

\[ \Delta(\Xi' - je_1 - ie_2) = \xi'_2 - i - \xi'_1 + j = \Delta \Xi' + (j - i), \]  

(39)

\[ \Delta(\Xi - je_1 - ie_2) = \xi_2 - i - \xi_1 + j = \Delta \Xi + (j - i). \]  

(40)

In Equations (37) and (38), if \( -\Delta \Xi + (j - i) \leq \Delta \Xi' - (j - i) \), then \( |\xi_2 - j - (\xi_1 - i)| < |\xi'_2 - j - (\xi'_1 - i)| \) and since \( f_n \) is balanced we have that \( f_n(\xi'_1 - i, \xi'_2 - j) - f_n(\xi_1 - i, \xi_2 - j) \geq 0 \), so that term of the sum in (36) will be positive. If, however, the opposite is true, then the amount of imbalance for the negative term – which directly correlates with the magnitude – is given by subtracting (37) from (38). In this situation, the state \( (\xi'_1 - i, \xi'_2 - j) \) actually becomes more balanced than the state \( (\xi_1 - i, \xi_2 - j) \). Therefore the difference in the amount of imbalance of the negative term is given by:

\[ 0 \leq -\Delta \Xi + (j - i) - \Delta \Xi' - (j - i) < -\Delta \Xi + (j - i) < \Delta \Xi' - \Delta \Xi. \]  

(41)

The first inequality holds by the assumption that \( -\Delta \Xi + (j - i) \geq (\Delta \Xi' - (j - i)) \). Then second inequality holds because we have \( j - i < \Delta \Xi' \Rightarrow \Delta \Xi' - (j - i) > 0 \). The final inequality follows from the fact that \( (j - i) < \Xi' \).
Likewise we know that the difference in imbalance for the positive term, \( f_n(\xi_1' - j, \xi_2' - i) - f_n(\xi_1 - j, \xi_2 - i) \), is at least as large as the difference in imbalance for the negative term by subtracting (40) from (39).

\[
0 \leq \Delta \Xi' + (j - i) - (\Delta \Xi + (j - i)) = \Delta \Xi' - \Delta \Xi.
\] (42)

Where the inequality follows from the fact that the \( \Xi' \) term is more imbalanced than the \( \Xi \) term and the equality follows directly. Clearly the negative term has less difference in imbalance between its components than the positive term, and therefore \(|f_n(\Xi' - ie_1 - je_2) - f_n(\Xi - ie_1 - je_2)| < |f_n(\Xi' - je_1 - ie_2) - f_n(\Xi - je_1 - ie_2)|\), which implies that \( f_n(\Xi' - ie_1 - je_2) - f_n(\Xi - ie_1 - je_2) + f_n(\Xi' - je_1 - ie_2) - f_n(\Xi - je_1 - ie_2) \geq 0 \).

**Case 2:** \( \Delta \Xi' \leq j - i \).

This case is straightforward, because we now have that for the pair of terms for the negative term, \( f_n(\xi_1' - i, \xi_2' - j) - f_n(\xi_1 - i, \xi_2 - j) \), both \( \xi_2 - j < \xi_1 - i \) and \( \xi_2' - j < \xi_1' - i \). Therefore the imbalance for each component is now given by:

\[
\Delta(\Xi' - ie_1 - je_2) = \xi_1' - i - \xi_2' + j = -\Delta \Xi' + (j - i)
\] (43)

\[
\Delta(\Xi - ie_1 - je_2) = \xi_1 - i - \xi_2 + j = -\Delta \Xi + (j - i)
\] (44)

\[
\Delta(\Xi' - je_1 - ie_2) = \xi_2' - i - \xi_1' + j = \Delta \Xi' + (j - i)
\] (45)

\[
\Delta(\Xi - je_1 - ie_2) = \xi_2 - i - \xi_1 + j = \Delta \Xi + (j - i)
\] (46)

For the negative term from (36), \( f_n(\Xi' - ie_1 - je_2) - f_n(\Xi - ie_1 - je_2) \), the \( \Xi' \) component is more balanced, (43), than the \( \Xi \) component, (44). The difference in imbalance between the two terms is given by:

\[
-\Delta \Xi + (j - i) - (\Delta \Xi' + (j - i)) = \Delta \Xi' - \Delta \Xi.
\] (47)

For the positive term, the difference in imbalance remains the same:

\[
\Delta \Xi' + (j - i) - (\Delta \Xi + (j - i)) = \Delta \Xi' - \Delta \Xi.
\] (48)

Therefore in Case 2, the difference in imbalance between the components of the negative term and the difference in the imbalance between the components of the positive term are equal and thus the subtraction will be 0. □

**Lemma 7.2.** *Proof:* To prove this, we show that the action of “do nothing” will result in less cost than shipping from the clinic with a lower inventory level to the one with a higher stock of medication. Let \( \xi_1 \leq \xi_2 \). If the action is to do nothing, the cost function will be \( J_n^0 = \Pi^T(-\Xi)^+ + 0 + \mathbb{E}\{f_{n-1}((\Xi)^+ - d_n)\} \), otherwise the amount of \( \hat{u} \) medication is moved from clinic 1 to clinic 2 (\( \hat{u} = (-\hat{u}, \hat{u}) \)), we have \( J_n^\hat{u} = \Pi^T(-\Xi)^+ + c\hat{u} + \mathbb{E}\{f_{n-1}((\Xi)^+ - \hat{u} - d_n)\} \).

Comparing the two term by term:

\[
c\hat{u} \geq 0,
\] (49)

\[
\mathbb{E}\{f_{n-1}((\Xi - \hat{u}e_1 + \hat{u}e_2)^+ - d_n)\} \geq \mathbb{E}\{f_{n-1}((\Xi)^+ - d_n)\}
\] (50)
Equation (50) follows from the fact that \( |\xi_2 - \xi_1| \leq |\xi_2 + \hat{u} - \xi_1 + \hat{u}| \) and that \( f_n \) is balanced, which carries through to the expectation via Lemma 7.1. From (49) and (50), \( f_n^*(\Xi) \geq f_n^0(\Xi) \) follows directly. \( \square \)

**Lemma 7.3.** *Proof:* Consider \( \Xi^* \) where \( \xi_1^* = \xi_2^* \) versus \( \Xi \) where \( \xi_1 \neq \xi_2 \).

\[
f_n(\Xi) = \Pi^T(-\Xi)^+ + \min_{u \in U}(c \sum_{j \in \Phi} (u_j)^+ + E[f_{n-1}((-\Xi)^+ + u - d_n)]),
\]

\[
f_n(\Xi^*) = \Pi^T(-\Xi^*)^+ + \min_{u \in U}(c \sum_{j \in \Phi} (u_j)^+ + E[f_{n-1}((-\Xi^*)^+ + u - d_n)]).
\]

We know that \( f_n(\cdot) \) is balanced, thus by Lemma 7.1 for all \( n \), \( E\{f_{n-1}((-\Xi)^+ + u - d_n)\} \) is also balanced. Therefore the optimal action at \( \Xi^* \) is \( u^* = 0 \), which achieves the lowest possible value for the expectation. Thus we have:

\[
E\{f_{n-1}((-\Xi^*)^+ - d_n)\} \leq E\{f_{n-1}((-\Xi)^+ + u - d_n)\}, \tag{51}
\]

and because the optimal policy at \( \Xi^* \) has no penalty because \( u^* = 0 \) it is clear that:

\[
\min_{u \in U}(c \sum_{j \in \Phi} (u_j)^+ + E[f_{n-1}((-\Xi^*)^+ + u - d_n)]) \leq \min_{u \in U}(c \sum_{j \in \Phi} (u_j)^+ + E[f_{n-1}((-\Xi)^+ + u - d_n)]).
\tag{52}
\]

Finally, it can quickly be verified that the instantaneous cost is lower for the balanced inventory \( (\Xi^*) \):

\[
\Pi^T(-\Xi^*)^+ \leq \Pi^T(-\Xi)^+ \tag{53}
\]

Thus we have shown that \( f_n(\Xi^*) \leq f_n(\Xi) \). \( \square \)
Malaria treatment distribution in developing world health systems and application to Malawi: Review Response

1. General Response to the Senior Editor and both Reviewers
   We appreciate the thorough review of our paper. The comments from the senior editor and the two reviewers have prompted us to deepen our analysis and explore new directions that have significantly improved the contribution of the work. We wish to also thank the senior editor and the reviewers for the opportunity to review, revise, and resubmit our paper. The following is a summary of the main revisions that we have undertaken during this round.

   - **Cluster Decomposition vs. Full Integration:** We have added Section 7.3 along with extensive new numerical analysis employing parallel computing that has enabled us to (1) solve smaller instances of the problem using a fully integrated strategic and tactical optimization, (2) compare the fully integrated optimization with the cluster decomposition we propose for tractability purposes. The primary result (in Table 7) indicates that the decomposition approach is in fact very near the true optimal solution, with and average (maximum) percent optimality gap of 0.5% (3.2%) over a six month period across four different groups of clinics tested (groups of four and five clinics as larger problems were not tractable even after employing parallel computing with 48 cores) and 12 different starting inventory levels (from low to high for each scenario). This is an exciting result, with implications for tractably and accurately solving large-scale, countrywide problems such as the distribution of ACTs in Malawi.

   - **Realism and Practical Grounding:** We have carefully gone through the referees’ comments and the paper itself to better support the practical grounding of the work. This includes more citations to support our claims and an explanation that some of the practical grounding also comes from one of our co-author’s experiences and domain knowledge acquired during extensive malaria field work on the ground in Malawi. We have also provided more detailed description of the demand estimation approach and other model parameter estimates. All of these are addressed in greater detail in the individual referee reports.

   - **Managerial take-aways:** We have now clearly highlighted some interesting managerial takeaways, and added new ones based on the referees’ comments. These are now summarized in the introduction and further explained in the appropriate sections throughout the paper.

   - **Introduction:** We have added a high-level description of the model, approach, results, and contributions to the Introduction.

   - **Clustering:** We have added more details about our cluster-based decomposition approach that we hope clarifies this new method for the reader.

   - **Clarity and Flow:** We have carefully gone through the paper to ensure that flow and continuity are preserved throughout and that our descriptions are clear. We also scoured the paper to improve the exposition and correct typos.

2. Senior Editor
   Below, we will first repeat your comment, then follow with our response.
This revision was significant in addressing the concerns of the referees about the first version of the paper. The paper does a better job of modeling and studying several variations on the supply (with potential for resupply/transshipment) for malaria treatments in Malawi. Some MDP results are derived, and an informative numerical study was provided. I think this is a good example of usage of operations research in solving an important problem.

Thank you for the positive encouragement. We hope that this revision has taken further steps toward providing a strong contribution in this area of research.

The referees have provided good feedback on the paper. Here are some main themes from their feedback.

- Some additional discussion about the context and justification of claims would be appropriate. In the report with 4 major issues, an example would be major issue 1: regarding inventory allocation in clusters, and transshipment. The referee with 6 main points identifies such issues in his/her points 1, 3 and 4 well.

  We have spent significant effort in addressing these points to provide a deeper discussion of context and justification of claims. In particular, we discuss our logic behind estimating the ACT demand, estimating model parameters and designing the case study, and adding citations/references to support our claims. We also add a more thorough explanation of the clinic clustering and decomposition method. The details are provided in the section with our response to that reviewer.

- Points 2, 5 at 6 of the referee with 6 main points identify issues of distilling out key take-aways more crisply and cleanly (from managerial and from academic perspectives).

  In this revision we have taken great care to highlight the key take-aways, which are in fact plentiful, due to the suggestions of the review team and the unique nature of the problem domain. First, we have, as suggested by the referees, summarized these takeaways in the introduction. We have also added some summarizing paragraphs regarding take-aways at the end of all major analysis sections, as suggested by Referee 1. Finally, we have performed an extensive numerical analysis to address point 2 and were very pleased with the results, which indicate that our decomposition method is very near the true optimal of the fully integrated model.

- The other main points of the referee with 4 major issues also go to supporting the reader’s walk through the paper.

  We have focused in this revision on providing clearer and deeper explanations to improve the reader’s walk through the paper, with special emphasis on addressing Referee 2’s remaining points. In particular, explaining why delayed shipment often yields lower shortages and exploring scenarios when the two recourse models behave very similarly, a deeper discussion of our cluster-based decomposition approach, and improving the former Section 7.3 (now Sec. 7.4) to give the reader a sense of the magnitude of the percent optimality gap relative to the objective.
One referee has suggested a major revision, the other considered both major and minor in private communication. In the sense that points 2 and 3 of the report with 6 points, for example, might require additional numerical work, a major revision would seem a more appropriate characterization of what is being asked to be done. Given the change between the original submission and the current submission, it would appear that this can be done in a next revision, and I would invite that next revision to be resubmitted to the journal. Ultimately, both referees will need to be satisfied with the replies to their concerns.

Thank you for your positive feedback. In this revision, we addressed all of the comments received from the referees. We agree with you that the additional discussions and justifications about some of the claims in the paper improve the research’s merit and paper’s readability significantly. Our goal is to have a paper with a logical flow where our academic contributions and analysis can be read, understood, and utilized. We spent significant amount of time to improve the flow of the paper. We also revised the presentation of the results and the research ideas and takeaways to state our contribution more clearly.

3. Referee 1
The authors respond to a number of concerns raised in the previous round most notably contribution to the literature on transshipment. Their response regarding the connection between strategic and tactical level models is still not very satisfactory. In addition, their method of estimating demand and the description of how it incorporates various contextual details needs to be strengthened considerably. I recommend a major revision so that the authors can take care of these concerns.

a. Thank you so much for your detailed review and positive encouragement. We considered all of your comments very carefully and have attempted to address each one. We believe these changes have strengthened our manuscript significantly.

1. Characterization of demand: I am glad that the authors provide more information on the method of generating scenarios for their stochastic optimization model and some field data on the number of cases reported. However, it raises a number of further questions that need to be addressed in the next revision.

1. How can we assess the validity for the demand model? In other words, what is the validity of assigning 19% probability to each of the “main” scenarios and 0.01 to each of the “variation” scenarios? Further what defines what is main and what is a variation?

a. Thank you for the comment. The information technology infrastructure in Malawi is underdeveloped and thus there is no reliable demand forecast available. However, one of the co-authors spent three to four months each year for several years in Malawi doing field work and worked on this problem on an operational level. We used the data he had collected in conjunction with his domain knowledge (expert opinion) regarding potential demand variations to create the demand scenarios. We started with 5 main scenarios, each representing a general pattern based on what had been observed in the past. Due to the lack of any long-term demand data, we assumed each main scenario is equally likely to occur and thus assigned equal probabilities (0.2) to each scenario. Then for each of those 5 main scenarios, we generated a less likely variation to capture the potential to capture extreme (but rare) scenarios. Thus we assigned the probability of 0.01 to each variation and 0.19 to each main scenario. We cannot claim that the demand scenarios we developed are a true
indicator of the actual demand, however they were informed by both actual historical ACT data and the long-term experience of our co-author in Malawi working on malaria, which is the subject of his research in the area of public health. We have added this information to the beginning of Section 6, where we describe the data effort:

“These scenarios were generated based on the historical data and expert opinion from one of our co-authors who has performed extensive field-work in Malawi regarding malaria. The scenarios developed herein were designed and confirmed based on his personal experiences over several years in Malawi.”

2. Figure 4(b) clearly shows a trend, i.e., some part of the variability in demand is predictable. If so, the demand during the second and third period should be correlated with the demand in the first period (before the recourse). Both the model and the estimation procedure should take this into account.

a. Thank you for the careful reading of the paper. We agree that Figure 4(b) indeed shows a trend and we agree that the demand in the second and third stages should be correlated with the demand in the first period. We included this seasonality and demand correlation when we created the 10 demand scenarios. So both the stochastic programming model and the estimation procedure do take this correlation into account. We have made this more explicit in the paper in Section 6 as well, where we describe the scenarios:

“In designing these scenarios, we also allow for correlated demand across the different stages as malaria is a transmittable disease whose spread depends on the number of infected persons. That is, high initial demand is more likely to translate into high demand in future stages.”

3. Can the implementers on the ground replicate this method for generating demand scenarios and do they have required data to do so?

a. This is a great point. We have spent a significant amount of time discussing this with our co-author who has direct experience in Malawi to ensure that the model we developed can be replicated and implemented in Malawi and generate practical results. We agree that extensive data collection and forecasting of the demand as well as calibration of the model parameters (such as transportation costs and shortage penalties) is important for obtaining high quality results. However, we are confident that even with limited data the model would (1) perform better than the current heuristic approach, and (2) would be fairly robust to errors in estimation. According to our co-author who works with the Malawi Ministry of Health, they do have historical data to replicate our estimation method (since our data comes from the Ministry) and the experts on the ground to provide input on the scenario perturbations of the raw data. Further, it is likely that Malawi Ministry of Health has more data than what we were provided, enabling more accurate estimates of demand than we were able to perform.

2. Differentiation from / contribution to the literature: I like the authors' attempt at distinguishing their paper from the transshipment literature on the basis of the link between strategic and tactical levels. However, they do not explicitly quantify the benefits of doing so. In other words, what would be the increase in cost if the tactical level transshipment decisions were optimized without accounting for the strategic level decisions and vice versa? Ideally, to do so, one would need three types of models: (i) a
benchmark formulation that truly integrates the decisions at the two levels, (ii) a pair of models at the two levels, similar to the ones in the paper, but with a demonstration of how they could be derived from the integrated model, (iii) a naive benchmark of making two levels of decisions in a completely uncoordinated fashion. An extensive numerical analysis of these three variants would be necessary to quantify the benefit of the authors' modeling approach. See also managerial insights below.

a. This is an important point and we thank you for prompting us to think about this issue. To address this comment we developed additional models and conducted extensive computational experiments. We have added a new section (Sec. 7.3) to the paper that performs extensive analysis of the tradeoffs between computational efficiency of our decomposition method versus the loss of optimality. Interestingly, we find the percent optimality gap of our decomposition heuristic (versus a fully-integrated strategic and tactical model) to be very small, with an average (maximum) percent gap of 0.5% (3.2%) across four different groups of clinics. Further, the percent optimality gap for our decomposition heuristic remains stable across all levels of initial inventory across the group, whereas the naive model (which we take to be our baseline model from the stochastic programming section) experiences a convex growth in percent optimality gap as the starting level of inventory increases. This has significant implications for our decomposition heuristic as a tractable method for solving large scale integrated strategic and tactical transshipment problems with near-optimal results.

3. **Description of data and contextual information:** As highlighted in the point about demand characterization, I am still not satisfied with the description and preprocessing of data to obtain the inputs to the model. The explanation at the beginning of Section 6 needs to be elaborated further and supplemented with intermediate data / quantification. For instance,

1. What were the demand patterns over the five years? How did the authors conclude that the demand did not vary much over the course of five years? Is this demand or is the demand that was met since Figure 4(b) mentions units of ACTs?

   a. Thank you for prompting us to clarify our approach to demand estimation. To address the two questions we have now added to the paper the following description:

   “By observing clinics where the data was more complete we noted that, over the five years, there was little evidence of overall upward or downward trend. This is corroborated by the World Malaria Report 2014 (see World Health Organization (WHO) (2014)), which indicates no significant trend over the time period. Though there was a slight increase malaria-related hospital admissions over the timeframe, the report states that data was insufficiently consistent to assess any trend in Malawi, which is consistent with observations from our own data.”

   Regarding the final question, we have added the following to the paper to be clearer:

   “This data was based on case counts reported at the facility level, whether the demand was met or not”

2. Similarly, how is $20 estimation of shortage cost linked with the per capita income of $810 is not clear. Ideally, it should be linked to the duration of the infection, loss in wages and productivity, health care expenditure etc. But the authors do not explain any of these steps nor do they provide any external reference or validation for this number.
a. Thanks for the comment. We have updated the manuscript (section 6.1, paragraph 1) to be clearer as follows:

Estimating the shortage penalty for malaria medications is non-trivial. Factors such as loss of income and productivity (for patients and relatives) during the course of infection, and health care expenditures should be taken into account in order to obtain a correct estimate. It should be noted that given the type of parasite, the symptoms and their severity can be quite different. Some people may have already developed immunity while in others (especially children) the disease can be deadly. Furthermore, malaria can have a higher indirect impact on children by hampering their physical and intellectual growth. Accounting for all these factors and monetizing their impact is key to determining the actual value of shortage penalty and is beyond this paper’s contribution. However, note that in the proposed approach, the rationale for including a shortage penalty is simply to ensure medications are being distributed according to the demand. In absence of such penalty, the model simply stops distributing medications in order to save transportation cost. In the absence of reliable data regarding health care expenditures, we performed a sensitivity analysis on the value of shortage penalty. We considered Malawi’s national income per capita, $810 reported by World Health Organization (WHO) (2014) as a basis. According to UNICEF’s Malaria Fact Sheet (UNICEF) (2004), malaria can slow the economic growth in sub-Saharan Africa by 1.3% annually. Based on these statistics, one can estimate the economic impact of malaria in Malawi at the individual level to be about $10.5 in lost economic growth. Thus we ran the models for a range of potential shortage penalty values between $10 and $100. For those experiments, we set the available supply of ACT to 1.5 million units. Based on our initial computation, even a low number (such as $20) is high enough to trigger effective distribution of medications. As the value of the shortage penalty is increased, reducing shortages becomes more critical in the trade-off between shortage penalty and transportation cost.

Reference: http://www.unicef.org/media/media_20475.html

3. How was the number of 1.5 million doses of ACT as the amount of available supply arrived at? I do see that the authors have done a sensitivity analysis around this number but still some justification is required for the assumption of the base case of the parameter value.

a. As the reviewer mentioned, we did not assume a fixed value for the total supply. Based on our co-author’s experience in Malawi, the available supply can vary. He suggested that we start with a total supply of 1.5 million units and perform sensitivity analysis around that value. Further, a report from the World Health Organization indicated that in 2008 ACTS were delivered in approximately 85% of reported malaria cases in Malawi. Given that our data indicated approximately 2.1 million cases, this would mean approximately 1.8 million units. However, to perform sensitivity analysis we tested values from 500,000 to 2 million units, which covers a broader range of possibilities. Further, the shortages between 1.5 million and 2 million units delivered were not significantly different.

4. A related but slightly different issue is that of definitions of various periods and matching those with calendar months in the Malaria season. Can the authors clarify what would be the length of the first and second periods in two-stage and three-stage models respectively? Accordingly, how much demand would be observed according to Figure 4(b) in these stages? In other words, Figures 2, 3 and 4 need to be internally consistent.
a. Thank you for the comment. In our model, we divided the year into three periods, each consisting of four months. The first period, August 1st through November 30th is considered the pre-malaria season and had an average demand of 113,331 over the timeframe. December 1st through July 31st is considered the malaria season. In the two stage model, the first stage is August 1st through November 30th and the second stage is December 1st through July 31st. In the three-stage model, the first stage is August 1st through November 30th, the second stage is December 1st through March 31st, with an average demand of 674,702, and the third stage is April 1st through July 31st with an average demand of 1,319,287. We updated the manuscript (section 6) accordingly.

4. Appropriate citations: For a paper that is focused on a real application, the authors should be more careful in citing all their sources. For instance, at several places, it is mentioned that the entire supply of medications is received at the beginning of the year but there is no citation for this. Other evidence from the global health supply chain literature suggests that supplies might be received periodically depending on the disbursements of funds by multinational agencies and donors, a process that is associated with a lot of delays and uncertainty (Natarajan and Swaminathan 2014). Similarly, at several places, authors mention terms like “cultural issues”, and “regional rivalries” without any proper citation of whether these issues exist in reality or not. There also needs to be a definition of what is meant by the terms centralized and decentralized before using them to describe supply chains in developing countries.

a. Supply uncertainty: Thank you for bringing this to our attention. We have updated the manuscript accordingly.

Section 6.2:

Supply availability is a major challenge in distributing malaria treatments in holoendemic areas. As reported by Natarajan and Swaminathan (2014) the process of procuring humanitarian supplies can be subject to delays and uncertainty. To better assess the effectiveness of our proposed stochastic models, we compare their results for a range of possible supply values, between 500,000 to 2,000,000 units while fixing the shortage penalty at $100. Fig. 6 (b), shows that the stochastic models can better utilize the additional supply of medications to address shortage compared to the baseline model. Among the stochastic models, three-stage models tend to be better at utilizing additional supply than two-stage models. Obviously reducing shortage results in higher transportation costs; but as seen in Fig. 6 (a), the difference in transportation cost for different models is relatively small.

b. Cultural issues and regional rivalries: We understand the reviewers concern and appreciate their thorough comments. As mentioned before, one of the co-authors has spent a considerable amount of time in Malawi working on this problem at the operational level. Through personal experience interacting with the ministry of health, local clinic personnel, and patients, he concluded that regional and cultural issues can often add to the complexity of the problem. We have added that this is based on personal observation and experience working in Malawi as follows:

“Cultural issues and regional rivalries prioritize certain areas while ignoring others, leading to inequities in access and supply. This observation is based on one co-author’s on the ground experience working with malaria in Malawi.”
c. Centralized vs. decentralized system: Thank you for this comment. We modified the manuscript (section 1.1) as follows to address your comment:

“According to Claeson and Waldman (2000), the efficacy of delivering health care through such systems has been the subject of debate for decades. For instance, people in developing countries may seek treatment through the private sector, namely from local shops and community members due to real or perceived shortfalls in the public system - see Bustreo et al. (2003). Some research has focused on strategies that circumvent, replace, or radically decentralize public health systems (e.g., Gallien et al. (2012)). In underdeveloped health systems, centralized and hierarchical supply chains in which most of the supplies are distributed from a central authority, sometimes suffer from disorganization and corruption. Information technology is often ineffective making it difficult to track location and quantity of medication supply, hampering forecasting of demand and effective allocation of supplies. Transportation infrastructure is generally poor, fuel shortages complicate matters and roads are often in bad condition, especially during the rainy season when malaria is most prevalent. Cultural issues and regional rivalries prioritize certain areas while ignoring others, leading to inequities in access and supply. This observation is based on one co-author's on the ground experience working with malaria in Malawi. In contrast, in a decentralized system, distribution of supplies is not necessarily controlled by a central authority. Instead, multiple players (public or private) contribute to the distribution of medications. However, the humanitarian nature of drug supply chains can prevent the development of more sustainable market-based supply chain solutions. Both centralized and decentralized systems have their limitations, but we focus on government sponsored health facilities, which are the most prominent source of medications and healthcare in most developing countries, including Malawi and nearly all sub-Saharan African countries. In addition, the methods suggested in this paper can be applied to other medical aid distribution supply chains outside of the public sector, such as those of the NGOs like John Snow Inc.”

5. Managerial take-aways: The authors need to distill down the key managerial take-aways and include them in the abstract nor the introduction. Section 6 is currently more than 7 pages long and can definitely use some breaks in the flow of the text by highlighting of a few key insights. For instance, they could be about the comparison of (i) two-stage vs. three-stage models, (ii) delayed shipment vs. transshipment models. The authors could also investigate the structure of the facility clusters within which transshipment is undertaken and explain how it depends on various model parameters.

a. Thank you for the suggestion on improving the flow and insights in the paper. At the end of section 6.1 we have added the following managerial insights (per your suggestion):

“Managerial Insights. Here we summarize some key takeaways from this analysis. First, the three stage models outperform the two stage models on shortages while maintaining a similar transportation cost. This is because the added control to be able to reposition inventory once part of the malaria season has been observed allows for more targeted efforts at distribution in distressed areas. This is especially important because of the demand correlation from period to period. Second, the delayed shipment model actually has lower shortages than transshipment in most cases. At first, this may seem counterintuitive because there is more flexibility in the transshipment model. However, this flexibility actually causes the first stage to distribute all the inventory out to the clinics to save on transportation cost, shipping direct instead of prepositioning a large stock at the district
hospitals. This creates greater dispersion of the inventory across the country, making it more likely that sufficient inventory is not nearby the point of need and transshipment may have to come from a clinic far away and therefore not worth the transportation cost. Whereas in the delayed shipment model, the district hospitals tend to be centrally located with many clinics around them. Since there is a larger stock stored at these hospitals initially (by design), there will be more incentive to take the recourse shipping action in stages two and three due to sufficient inventory and proximity. This also explains why shipping cost is higher for delayed shipment, because rather than shipping direct, much of the product must follow first a route from the main dispensary to the district hospital and then a second route from the hospital to the clinics. Hence, the key insight here is that if the government has sufficient transportation budget and cares more about avoiding shortage, then delayed shipment may be a better structure.”

In Sec. 6.2, we added the following:

“Among the stochastic models, three-stage models tend to be better at utilizing additional supply than two-stage models. This insight is similar to the key takeaway from Sec. 6.1 in that the three-stage model benefits from targeted repositioning in stages two and three after observing some of the malaria season’s demand and accounting for correlation from one period to the next.”

In Sec. 6.4 we summarize the key factors in driving the creation and characteristics (size, distances) of clinic clusters (per the suggestion above):

“A key insight is that, once transshipment reaches a scenario-specific cost threshold (e.g. 50-60% in Fig. 8) the transportation cost will remain relatively stable by becoming more conservative as to how far one would be willing to ship medications to satisfy unmet demand in a different region. That is, as clinic-to-clinic transportation becomes more expensive, transshipment eliminates routes from consideration due to high cost thus becoming less effective in satisfying all demand. Routes will be eliminated starting with the longest routes, keeping only the shorter routes available for transshipment and thereby localizing transshipping needs around increasingly proximate geographical clusters. As an extension of this line of reasoning, the higher the penalty cost for unmet demand, the longer the model resists localizing transshipment efforts in favor of more regional/national transshipment. Thus, depending on the strategic goals and constraints of the distributor, the shipping network may be more localized (prioritizing transportation cost due to a constrained budget) or more national (sufficient budget shifts the focus to shortages).”

At the end of Sec. 6.5 we provide the following:

“A key takeaway from this analysis is that poor road conditions lead to an increase in shortages and less movement of product around the network. However, the flexibility of the transshipment model to choose alternate routes enables more demand to be satisfied relative to delayed shipment, albeit at increased transportation cost due to using more expensive alternate routes.”

We also present the following insights at the beginning of Sec. 7

“From these experiments we have the following findings: (1) Ethical policies are very close to optimal when clinics are clustered with other clinics nearby, (2) Seasonality (predictable variability) is easily handled by transshipment, but random demand variability has a significant impact on system cost, (3) transshipment is effective even with infrequent intervals under periodic review. We
also propose a simple heuristic that can be easily implemented and appears to generate near optimal results.”

And the following are presented to conclude section 7:

“Insights and Policy Design. The results of our numerical experiments yield some interesting insights that can help design an effective distribution strategy. First, the largest cost impact comes from year to year variability in demand. Better data collection and forecasting could help reduce some of this uncertainty. Second, transshipment frequency did not have a large impact on cost. Third, an ethical strategy that balances inventory between the clinics in a cluster at each transshipment period is very close to optimal. Further, if the initial inventories start out balanced there is almost no difference between the ethical inventory balancing policy and the optimal policy. An implementable policy, inventory balancing also has many desirable attributes. First, it is simple and easy to implement. Second, it is likely to be perceived as fair by the parties involved. Finally, it is very close to the optimal policy. For these reasons, we posit that an inventory balancing (relative to average demand at each clinic) scheme could be an effective method for implementing transshipment within clusters.”

6. Writing: At places, the writing of the paper can be improved. The entire introduction does not contain any description of the problem being addressed in the paper, its main solution approach, results and contributions. Perhaps, the theoretical development in Sections 6 and 7 can be combined and similarly the numerical experiments in the two sections can be combined in a separate section to ensure better flow and continuity.

   a. Thank you for the suggestions. We have now expanded the introduction (primarily Sec. 1.1) to better explain the problem, the solution approach, results, and contributions.

   We further attempted to find a way to combine the theory and the computational sections, but the fact that we need the results from the stochastic programming computational section (i.e. the establishment of clinic clusters) to motivate Section 7 made this approach difficult. We have, however, gone through the paper thoroughly to try to smooth out the flow and create better transitions between topics.


4. Referee 2
The authors addressed all major and minor comments provided in the first revision. I only have a few additional comments outlined below.

   a. Thank you for your careful review and the positive encouragement. We found your comments very helpful and relevant. We believe the updated manuscript (after addressing all your comments) is much stronger.

Major Issues
1. In Page 9 the authors present the two-stage transshipment model. If I understand correctly equation (10) implies that district hospitals keep some inventory that is later allocated to clinics. If this is true (please confirm), a few concerns regarding this are:

   a. The transshipment model allows for both delayed allocation as well as transshipment. While there is nothing wrong with this, it seems to imply that the transshipment model has the flexibility of both delaying the allocation of inventory as well as transshipment between clinics. This flexibility will give the transshipment model a very marked advantage over the delayed allocation model. It would be interesting to know how much inventory is kept at the district hospitals under the transshipment and delayed models. Inventory literature suggests that the amount of inventory kept at the district hospitals should be zero or relatively small given the option to transship. Nevertheless the fact that transshipment is carried out only within clusters may provide motivation to keep some stock at the district hospitals.

   Thank you for your careful reading. This is true. In our approach we allow the transshipment model to keep some inventory in a similar fashion to the delayed shipment model. Interestingly, as you mentioned, our findings are consistent with the existing literature – the amount of inventory kept at the district hospitals is negligible. This is mostly due to the fact that transshipment is generally less expensive than delayed shipment. However, as transshipment becomes more expensive, the transshipment model results become more similar to those of the delayed shipment – see Figure 8.

   We updated section 4.2 to clarify the point that (according to equation 10) the transshipment model does have the capability to keep some inventory at the district hospitals. Here is the updated text:

   “Equation (15) is essentially similar to Equation (10) from the transshipment model – allowing some inventory to be kept at the district hospital. This means that the transshipment model does have a similar capability of delayed shipment. In most cases, however, the amount of inventory stored at the district hospital in the transshipment model is negligible. As we will discuss in section 6, under some parameter regimens, especially when the cost of transshipment arcs exceeds those of delayed shipment arcs, the transshipment model can result in similar outcomes as those generated by the delayed shipment model.”

   b. A more detailed discussion on how the district hospital’s inventory is allocated to the different clinic clusters would be valuable. It would be interesting to know if only one clinic within a cluster receives inventory from the district hospital in the second or third stages.

   This is a very interesting point. We observe that when transshipment is less expensive, only one clinic in each cluster receives the initial shipment. As transshipment becomes more expensive, the transshipment model behaves more similarly to the delayed shipment and thus, clusters become less relevant.

   To better clarify this point we updated section 6.4 as the following:

   “In Figure 8(a) and (b) the dashed lines represent the expected transportation cost and expected shortage volume respectively for the delayed shipment models. These are constant across all scenarios because delayed shipment does not use the clinic-to-clinic routes. The solid lines represent the transportation and expected shortage costs for the transshipment model. When transshipment is less expensive, all the inventory is initially allocated to one clinic in the cluster,
which then transships to the other clinics due to the lower cost of transshipment relative to the cost of the initial distribution. We see that the transportation costs initially increase as clinic-to-clinic routes become more expensive as does the expected shortage, though more slowly. This is because the model is considering the trade-off between shortage penalties and the transshipment recourse action. At 60% the transportation cost then begins to decrease while the shortage penalty increases more sharply. This inflection point occurs because at this level, clinic-to-clinic transportation becomes expensive enough that the model will stop transshipping along certain routes altogether, preferring to accept some shortages rather than incur high shipping costs, e.g. shipping across the country to fulfill a small amount of demand. Note that as previously mentioned in section 4, the similarity between equations (10) and (15) enables the transshipment model to store some inventory at district hospitals and delay shipments if necessary. Eventually both costs approach those of the delayed shipment model as transshipment becomes so expensive that the model chooses to avoid clinic to clinic shipments almost entirely and uses primarily the non-clinic-to-clinic routes that the delayed shipment model is restricted to.”

2. The results obtained with the three-stage model in Section 6 are interesting. It is surprising to see in Figure 7(b) that under a three-stage model the delayed allocation model provides a greater reduction in expected shortages than the transshipment model compared to the baseline. While this would imply lower shortage costs for the delayed shipment model, Figure 7(a) also shows that transportation costs are higher for the delayed model and as a result this model may still result in higher costs. Nevertheless given the application under study in this paper it would seem that reducing shortages may have a higher priority than cost reduction. In section 6.4 the authors provide a nice sensitivity analysis for the transshipment model including identifying “cost thresholds” that affect the performance of the transshipment model. A similar analysis exploring parameter values that result in the delayed model performing similarly to the transshipment model would be insightful and could result in some general guidelines for the use of either delayed shipment or transshipment in practice.

Thank you for bringing this to our attention. This is certainly something that we found quite interesting as well. As such, we have added a discussion at the end of Sec. 6.1 where the phenomenon of lower shortage in delayed shipment first appears (excerpted here):

“The delayed shipment model actually has lower shortages than transshipment in most cases. At first, this may seem counterintuitive because there is more flexibility in the transshipment model. However, this flexibility actually causes the first stage to distribute all the inventory out to the clinics to save on transportation cost, shipping direct instead of prepositioning a large stock at the district hospitals. This creates greater dispersion of the inventory across the country, making it more likely that sufficient inventory is not nearby the point of need and transshipment may have to come from a clinic far away and therefore not worth the transportation cost. Whereas in the delayed shipment model, the district hospitals tend to be centrally located with many clinics around them. Since there is a larger stock stored at these hospitals initially (by design), there will be more incentive to take the recourse shipping action in stages two and three due to sufficient inventory and proximity. This also explains why shipping cost is higher for delayed shipment, because rather than shipping direct, much of the product must follow first a route from the main dispensary to the district hospital and then a second route from the hospital to the clinics. Hence, the key insight here is that if the
government has sufficient transportation budget and cares more about avoiding shortage, then delayed shipment may be a better structure.”

Regarding the sensitivity analysis to investigate parameters that cause the transshipment model and the delayed shipment model to behave similarly, we present the following result – which shows that as the penalty term increases, the transshipment model begins to behave more like the delayed shipment model:

“The three-stage delayed shipment model tends to be more effective at addressing shortage, though at a higher transportation cost. As the shortage penalty increases, however, we observe that the gap between the three-stage delayed shipment model and the three-stage transshipment model shrinks.”

3. Page 20, line 20 the authors describe how they determine which clinics belong to a cluster by comparing the maximum transshipment values obtained against a predetermined threshold. Please explain how this threshold is obtained. Given that the clusters are obtained as a result of the structure of the optimal solution to the strategic stochastic program, it is important to understand why is a threshold needed, and how it is obtained.

a. Thanks for the comment. Note that in this paper we are not using exact clustering methods (e.g. k-means). Instead, we imply cluster structures from the results of the transshipment model. The downside to this method of developing clusters is that the results may not always be clear-cut. Depending on the actual parameter values, cluster boundaries may get a bit blurry. For instance as described in figure 8, as the cost of transshipment increases, the model tends to hold on to some inventory and ship them to clinics with a delay. In other extremes when transshipment is very inexpensive, a clinic may receive a small shipment from another clinic that is not in the same cluster. This is why using the results of the stochastic program will not always guarantee that we obtain mutually exclusive clusters. However, in all those cases we observed that by eliminating shipments that were lower than a predefined threshold, we can obtain mutually exclusive clusters. For the computational experiments, we set this threshold value to the average demand of the receiving clinic (i.e. clinic j) across all scenarios (i.e. $\sum_{s \in S} p_s d_j^s$).

We updated the manuscript (section 7) to clarify this issue:

To identify clinic clusters, we solve the strategic-level transshipment model to optimality and calculate the optimal values of transshipment between two clinics under all scenarios, i.e. $y_{ij}^s$. Then we take the maximum of transshipment values across all scenarios defined as $y_{ij}^{max} = \max_{s \in S} y_{ij}^s$. Note here we are not using an exact clustering method (e.g. k-means). Instead, we are implying cluster structures from the results of the transshipment model. The downside to this method of developing clusters is that the results may not always be clear-cut. Depending on the actual parameter values, cluster boundaries may get a bit blurry. For instance as the cost of transshipment increases, the model tends to hold on to some inventory at the district hospitals and ship them to clinics with a delay. In another extreme case when transshipment is very inexpensive, a clinic may receive a small shipment from another clinic that is not in the same cluster. Therefore using the results of the stochastic program will not always guarantee that we obtain mutually exclusive clusters. However, in all those cases we observed that by eliminating shipments that were lower than a predefined threshold, we can obtain mutually exclusive
clusters. To implement this idea, if there has been a “significant” transshipment between those two clusters, i.e. $y_{ij}^{max}$ exceeds a pre-determined threshold, we assume those two clinics are in the same cluster. To conduct computational experiments, we set this threshold to the average monthly demand of the receiving clinic (i.e. $j$) across all scenarios; i.e. $\sum_{S\in S} p_S d_j^S$.

5. The new Section 7.3 is a good addition to the paper as well. It provides great managerial insights and a more clear perspective of the uniqueness of the problem under study. In order to improve this section it would be great if the authors could provide the results not only in $\$ amounts but also in percentages relative to the corresponding benchmark. For instance, when analyzing the impact of demand variability and seasonality it would nice to know if the $14 (\$58)$ difference between the seasonal and flat demand patterns represent a small or a large percentage of the total cost obtained with the flat demand pattern. Please do the same for all $\$ measures included in this section.

   a. Thank you for the suggestion. We agree that it provides a better understanding to present the percentage of the baseline cost. This has now been updated for all dollar figures in this section.

Minor Issues

a. In general while the paper is well written there are quite a few errors throughout the paper. Please proofread the paper thoroughly.

   a. We have proof-read the paper carefully, with special attention to typos and errors. Thank you also for your careful reading and suggestions for improvement, which we have addressed in this revision.

b. Page 2, line 4 - erase extra “related”.

   Done.

c. Page 6, line 9 - the comment: “...to behave slightly different from traditional contexts” requires at least a brief discussion to clarify its meaning to the reader.

   Done.

d. Page 7, line 3 - the text: “Unlike conventional inventory .... with mid-season restock being uncommon” was already mentioned in the paper previously. The authors may consider eliminating repetitions throughout the paper.

   Done.

e. Page 11, Section 5.1 – at the end of the section title erase the extra “Transshippment”

   Done.

f. Page 15, line 7 – change “confirm” to “confirms”.

   Done.

g. The different lines shown in Figure 6(a) are hard to differentiate particularly when printed in black/white. Please change the scale or make the Figure larger.
We tried our best to make it easier for the readers to differentiate between the lines in Fig. 6(a). However, in that particular experiment, all the five models result in very similar transportation costs.

h. Results shown in Figures 8 and 9 should indicate if they were obtained using the two or three-stage models.

Done.

i. Page 20, line 20 – at the end of the line, the word “cluster” should be “clinics”?

Done.

j. Page 20, Table 3 – please specify the “timing” of variables such as the amount of inventory at each clinic. Is this the inventory at the beginning of the time interval or at the end? The amount of inventory at each clinic is it defined as the inventory at the beginning of the time interval or at the end?

This is the starting inventory at the beginning of the period. We have now updated table 3 to be more clear about this

k. Page 26, first paragraph:

i. Should “zones” refer to “areas” in Figure 11? Note that “areas” is used both in the Figure as well as later in the same section, for consistency the authors may consider using “areas” instead of “zones”.

Done.

ii. There seems to be some errors in the description of Figure 11. In line 6, should “Zone 2” be “Zone 3”? Also in line 8, should “Zone 3” be “Zone 5”? In line 9, should “Zone 3” be “Zone 5”? Please correct or clarify.

The reviewers are right. There has been a typo. It is fixed now.

l. Page 27, line 4 – change “zone” for “areas”.

Done.

m. Page 30 – The “Demand Variability and Seasonality” discussion appears twice.

Done.

n. Page 30, line erase “a” in “malaria is a seasonal,..”.

Done.

o. Figure 14(c) is wrongly labeled as 14(b).

Done.