Inventory Policies for Pharmaceutical Distribution in Zambia: Improving Availability and Access Equity

Jérémie Gallien
London Business School, London, United Kingdom, jgallien@london.edu

Ngai-Hang Z. Leung
College of Business, City University of Hong Kong, Kowloon, Hong Kong, znhleung@cityu.edu.hk

Prashant Yadav
Technology & Operations Management, INSEAD, Europe Campus, Fontainebleau, France, prashant.yadav@insead.edu
Center for Global Development, Washington, DC
Department of Global Health and Social Medicine, Harvard Medical School, Boston, MA

In many low and middle income countries, including Zambia, stockouts of life-saving medicines threaten the advancement of the Sustainable Development Goals (SDGs); it is therefore vital to reduce stockouts through the use of improved inventory control policies. The associated medicine distribution problem is challenging because it involves seasonality and uncertainty in both demand and lead-times, heterogeneous delivery locations, and lost demand. Besides service level and inventory costs, equity across delivery locations must also be considered. This empirical study is based on an independently-validated simulation model constructed from extensive field data, and addresses the lack of rigorous recommendations of inventory policies in this context. It compares the current base-stock and other policies proposed in the practitioner's literature with an optimization-based policy adapted from research on industrial settings. Although the optimization-based policy may need more implementation efforts, it generally outperforms the other evaluated policies. Our results also suggest that the prevalent proportional inventory rationing rules may lead to substantial service level discrepancies between facilities. Finally, the performance metrics of service level and distribution equity can be at odds, prompting non-trivial design trade-offs and considerations. This work motivated the development of a digital distribution information system involving smartphones with barcode scanners deployed in 60 health centers, posts and district hospitals in Zambia until 2018.

Key words: Inventory Management; Medicine Distribution; Efficiency and Equity; Simulation; Zambia
History: Received: May 2020; Accepted: June 2021 by Subodha Kumar, after 3 revisions.

1. Introduction

Achieving equitable access to health resources is necessary to accomplish Sustainable Development Goal (SDG) 10 ("Reduce inequality within and across countries") and SDG 3 ("Ensure healthy lives and promote well-being for all at all ages"). A key challenge is ensuring access to health products — SDG 3.8 mandates "access to safe, effective, quality and affordable essential medicines and vaccines for all." In many low and middle income countries (LMICs), inadequate pharmaceutical distribution leads to drug stockouts at local health facilities, which contributes to their heavy disease burden. For instance, the average service level of essential medicines such as malaria drugs at public health facilities in sampled LMICs was found as low as 29.4% in some regions (Cameron et al. 2009).

Zambia is representative of the disease burden in sub-Saharan Africa: its under five infant mortality rate is 87 per thousand live births, compared to the sub-Saharan average of 92 per thousand live births (You et al. 2015), with malaria being one of the major causes of under five mortality. The qualitative reviews Yadav (2007) and Yadav et al. (2011) identified lack of reliable transportation, inadequate staffing, and poor facility infrastructure as the main challenges in supplying local health facilities. Subsequent assessments have also reported relatively frequent stockouts at the central warehouse, high rates of stockouts at the health facility level, and substantial variations across regions (Picazo and Zhao 2009, Vledder et al. 2019).

To overcome these challenges, the Zambian Ministry of Health (MoH) and its partners have invested significant resources in the public sector supply chain for essential medicines in recent years, resulting in better distribution system performance (Vledder et al. 2019). However, there is strong evidence that inventory management (i.e., how much inventory should be sent at each time to each location) remains an important improvement opportunity (Leung et al. 2016). Consistent with the SDGs, the equity of access to essential medicines across locations is a key strategic goal formulated by the MoH in its Health Sector Supply Chain Strategy (Zambia Ministry of Health 2015); this contrasts with the utilitarian (or efficiency-driven) objective, which in this context would consist of exclusively maximizing the country-wide service level regardless of any resulting local inequalities (McCoy and Lee 2014). Health access equity is also a sensitive issue in Zambia because its different ethnolinguistic groups are distributed unevenly across the country. Ensuring this equity is made challenging by the prevalence of drug shortages at the national, regional and global levels however (Hedman 2016, Gallien et al. 2017).

Unfortunately, the current peer-reviewed literature does not provide recommendations for how to design an inventory management policy to address the inventory distribution challenge faced by Zambia or similar countries, because the key features of this problem (seasonality and uncertainty in both demand and lead-times, heterogeneous delivery locations and lost sales, frequent central warehouse stockouts, consideration of equity across locations) render mathematical analysis difficult. While some policies have been recommended in the practitioner literature for this problem (e.g., USAID | DELIVER PROJECT (2011a)), their performance in certain circumstances is poor (Leung et al. 2016). In this context, the present empirical study pursues the following objectives:

1. Provide a rigorous description of the challenging medicine distribution problem facing Zambia and other sub-Saharan African countries;
2. Develop and disseminate a rigorous empirical simulation testbed with validated predictive accuracy for this problem;

3. Use this testbed to evaluate various possible policies against inventory distribution policies currently employed in practice, and develop related evidence-based recommendations for practitioners.

There is a substantial existing literature on inventory distribution models, which we discuss in §2. The specific distribution problem considered here adds to that literature because of its above-mentioned features which are particularly challenging from an analytical standpoint, but also because our motivating environment involving essential medicines dictates the consideration of service level equity across delivery locations, a performance dimension which has received little attention in the inventory management literature. We provide a more detailed discussion of this problem in §3, including some additional background on Zambia's public pharmaceutical distribution (in §3.1), a quantitative problem formulation (in §3.2) and a description of the inventory control policies currently used in Zambia and elsewhere to address this challenge (in §3.3). We then present three possible improved policies in §4, specifically two extensions of the current policy recently proposed in USAID | DELIVER PROJECT (2011a) and Watson et al. (2014) (see §4.1 and §4.2), and an optimization-based policy involving an inventory planning model solved on a rolling horizon basis as in Foreman (2008) (see §4.3).

This paper relies on a realistic simulation-based testbed which has validated performance prediction accuracy and enables the empirical evaluation of these policies. We specifically leverage extensive field data to construct a model simulating the inventory of a representative antimalarial drug in 212 delivery locations in Zambia, as described in §5. An important feature of this model is that the accuracy of its stockout predictions can be validated out-of-sample, by comparison with another extensive dataset of historical facility service levels that was collected independently. As discussed in §6, our results suggest that implementing the above optimization-based inventory distribution method would lead to substantial improvements in the availability of drugs at health facilities in Zambia. In addition, these results show that the metric of distribution equity can be at odds with the more traditional performance metrics of system service levels and inventory costs, and that the prevalent proportional inventory rationing rule may lead to substantial inequity in environments with heterogeneous delivery locations. As discussed in §7, this work has motivated the development of a digital distribution information system involving smart phones with barcode scanners, which was used in 60 health centers, posts and district hospitals in Zambia before its deployment was paused in 2018 due to lack of funding. We conclude the paper and discuss opportunities for future related work in §8. The Electronic Companion to this paper contains a summary of mathematical notations (§A), a discussion of policy implementation issues (§B), technical details on the simulation model (§C), including the construction and estimation of its demand (§C.1) and lead-time (§C.2) generation components, and reports on numerical experiments conducted to evaluate the robustness of our study with respect to the equity metric (§E) and information transmission speed (§F). In order to facilitate further research by others on inventory theory and/or the distribution of essential medicines in low-income countries, all datasets as well as the code of the simulation model described in this paper are available in the public repository `https://github.com/zacharyleung/ZappSupplyChainSimulator`.

2. Literature Review

We study inventory policies for a pharmaceutical distribution system in a resource-constrained environment. Accordingly we first discuss the relevant literature on inventory distribution in multi-echelon systems (§2.1), then the relevant literature on global health supply chains (§2.2).

2.1. Multi-Echelon Inventory Distribution Models

Porteus (2002) contain a review of the large stream of literature on inventory policy in multi-echelon systems, and Axsäter et al. (2002) contains a good discussion of the subset of these studies that focus on distribution systems. Within this literature, the problem we consider is distinguished by the assumption of unsatisfied demand being lost (as opposed to backordered), which seems appropriate in a context where life-saving medicines are often distributed to patients who need to walk for several hours to reach care delivery facilities. Remarkably, the optimal replenishment policy is not even known for the single level version of this problem with stationary demand and constant lead-times (Zipkin 2008). Our context is further characterized by heterogeneous, non-stationary and stochastic demand and replenishment lead-times in a distribution system comprising a large number of facilities, and limited inventory available at the supplying warehouse. We are not aware of any policy, let alone an optimal one, described in the peer-reviewed research literature for this problem. The practice-based papers by Foreman et al. (2010) and Caro and Gallien (2010) report heuristic policies adapted to real inventory distribution problems. The optimization-based policy considered here has a structure similar to that described in Foreman et al. (2010), namely an inventory planning model solved on a rolling horizon basis. However, the present paper substantially differs from these studies in the context and details of the distribution system considered.

Our study further adds to the existing literature on inventory distribution because of the consideration of service level equity. Many studies consider the trade-off between efficiency and fairness in the context of both specific applications and theoretical resource allocation problems (see Bertsimas et al. (2011) and references therein). In addition, many studies of operational decisions in health systems recognize the importance of fairness among patients (e.g. McCoy and Lee (2014), Qi (2017)). However, few papers on inventory distribution explicitly capture equity considerations beyond the proportional stock allocation rule that is widely used in practice (e.g., Agrawal and Cohen (2001)). Exceptions include Ernst and Kamrad (1997), who consider the problem of allocating inventory between different retailers with individual service level constraints, and Graves (1996), who introduces a notion of equity based on prioritizing the oldest outstanding orders. The literature also discusses a wide range of possible metrics to evaluate equity (e.g., Marsh and Schilling (1994), Bertsimas et al. (2011), McCoy and Lee (2014)) along with various related theoretical and practical considerations. While the primary metric we use to evaluate distribution equity is the standard deviation of service level across delivery locations, we find that our results continue to hold qualitatively across a range of different equity measures (see §E in the Electronic Companion). Furthermore, the simulation results reported in §6 demonstrate that under realistic and important scenarios some inventory distribution policies can outperform others along the traditional dimensions of average system-wide service level and inventory holding costs while generating substantially lower service level equity across delivery locations. The performance dimension of service level equity thus introduces new and non-trivial design trade-offs and considerations to inventory management, and our study may inform future studies of fairness in inventory systems.

2.2. Global Health Supply Chains

Kraiselburd and Yadav (2013) argue that ineffective and poorly designed systems for purchasing and distributing health products are one of the most important access barriers to medicines in low income countries. Consequently, a number of studies discuss interventions on various pharmaceutical supply chain components in resource-constrained environments, including procurement and financing (Sieter 2010), facility location (Raja et al. 2008), information systems (Barrington et al. 2010) and organizational structure (Bossert et al 2007). A particularly relevant study is Vledder et al. (2019), which reports a quasi-randomized experiment (hereafter referred to as the 2009 Zambia supply chain pilot) designed to evaluate two different supply chain structures (intermediary stocking versus cross-docking), and found that the supply chain with fewer tiers results in reduced stock outs at the health facility level. Although Vledder et al. (2019) does not explicitly consider the impact of inventory distribution policies, that study has influenced the design of the current pharmaceutical distribution system in Zambia (see §3.1), and has also generated essential data for this paper (§5).

Turning to inventory control, Leung et al. (2016) presents a single-facility simulation study of the inventory replenishment policy currently used in Zambia. That paper reports evidence that that policy causes regular and predictable stock-outs of drugs with seasonal demand (see §3.3), and more broadly that inventory control constitutes an important opportunity to improve patient access to drugs in Zambia. Following the dissemination of the Leung et al. (2016) study, practitioners have developed additional guidance for changing some parameters of the current inventory policy when replenishing malaria medicines (USAID | DELIVER PROJECT 2011a), and funded another study to develop an extension of the current policy involving look-ahead seasonality indices (Watson et al. 2014).

While Leung et al. (2016) evaluates the policy variants described in USAID | DELIVER PROJECT (2011a), it considers neither the policy proposed later by Watson et al. (2014) nor other policies adapted from the existing literature such as the optimization-based policy considered here. More importantly, while Leung et al. (2016) also leverages data collected in the context of the 2009 Zambia supply chain pilot, that study entirely relies on a model only considering a single facility and assuming unlimited central warehouse inventory. These assumptions are particularly problematic given the study goals stated in §1. This is because central warehouse stockouts are prevalent in Zambia and more broadly sub-Saharan Africa, therefore practitioners routinely face the challenge of allocating inventory between multiple competing locations. Furthermore, developing rigorous knowledge about this allocation challenge seems particularly important in the context of distributing life-saving medicines to the public. Such knowledge may obviously not be developed with a single facility model as that used by Leung et al. (2016) however. In contrast, the present study relies on a realistic evaluation testbed involving a large network of facilities, for which more than 200 heterogeneous demand and access lead-times distributions must be estimated (as opposed to single distributions corresponding to an "typical" facility in Leung et al. (2016)). This network model allows us to evaluate policies for allocating a limited amount of inventory between multiple facilities, and the associated performance dimension of service level equity across delivery locations, which is critical in this setting. Finally, in contrast with Leung et al. (2016) which essentially establishes the inadequacy of current inventory replenishment policies for a single facility but does not present specific alternative proposals, this study also develops a specific and detailed proposal for how practitioners may address the inventory distribution problem in such a network.

3. The Essential Medicines Distribution Problem

3.1. Background on Public Pharmaceutical Distribution in Zambia

The large scale population level distribution of essential medicines and supplies presents a significant challenge due to low population density, poor road and communication infrastructure and flood-related access cutoffs during the rainy season. The three distribution channels for drugs in Zambia include the private sector, faith-based/mission organizations and the public distribution system (Yadav 2007), in which drugs are delivered to patients for free by the government. Because of the limited spending power of Zambia's population (GDP per capita approximately $1305), the public distribution has an important impact on public health. In that system, medicines are initially procured by the government with financial and technical assistance from external donors. Insufficient or delayed funding (Gallien et al. 2017) as well as inaccurate demand forecasts, inefficient procurement and production processes among other factors (Hedman 2016) all combine to create frequent drug shortages at the national level in Zambia, as in most of sub-Saharan Africa (Hwang et al. (2019) and references therein). Once received at the central warehouse in Lusaka, drugs and medical supplies are shipped on a monthly basis to approximately 90 district stores and hospitals (*primary distribution*) by a para-statal agency called Medical Stores Limited (MSL). Upon receiving downstream replenishment orders, MSL carries out picking and packing, and drugs are then shipped to Zambia's 72 districts by a dedicated truck fleet. To smooth transportation and warehouse workload, districts are partitioned into four groups of delivery tours, and MSL's fleet of trucks performs the deliveries to these four groups following the same sequence each month. Picking, packing and loading activities carried out at the central warehouse are also driven by this schedule. Next, *secondary distribution* involves shipment of these drugs from the districts stores to 1500 or so health centers. Secondary distribution is more ad-hoc and unreliable because of chronic shortages of transportation resources and health center staff, and the poorer condition of secondary roads. Following a large-scale pilot experiment (Vledder et al. 2019), the government of Zambia decided in 2010 to transition the entire country from *intermediary stocking* (whereby districts replenish their own unallocated inventory from MSL and prepare shipments for secondary distribution themselves) to *cross-docking* (whereby districts receive from MSL packages already prepared for their final health center destination), which we assume in this paper.

3.2. Problem Statement

We summarize the inventory distribution problem arising for every health product in the context just described as follows: every month, a central warehouse (denoted W) needs to determine shipment quantities x_t^h to every final delivery location h in a set $\mathcal{H}$, where discrete index $t \in \mathcal{T} \triangleq \{1, ..., T\}$ refers to the week in which this shipment is determined (a sensible value of T in practice could cover 6 months to a year, in order to capture key upcoming seasonality features). Delivery locations $\mathcal{H}$ are partitioned into four subsets $(\mathcal{H}_g)_{g \in \{1,2,3,4\}}$ corresponding to each one of the fixed sequential delivery groups used for transportation and warehouse planning purposes (see §3.1). For exposition simplicity we assume that the total duration of each monthly delivery cycle is 4 weeks, and that the shipments to each delivery group are determined at the beginning of the same week each month (in the following we refer to weeks and time periods interchangeably). We define $\mathcal{T}_g \subset \{1, ..., T\}$ as the subset of time periods when the shipments to facilities in shipment group $\mathcal{H}_g$ are being determined, and the shipment schedule constraints can thus be expressed as

$$x_t^h = 0 \text{ for all } (t, h, g) \text{ such that } h \in \mathcal{H}_g \text{ and } t \notin \mathcal{T}_g. \tag{1}$$

That is, in any given week the only shipments being determined are those to the facilities in the delivery group associated with that week.

The information that is relevant to determine the shipments x_t^h includes the on-hand inventory levels I_t^W and I_t^h at the warehouse W and in each location $h \in \mathcal{H}$ at the beginning of week t, and the pipeline inventory to these location. Specifically, the amount shipped in period $t' < t$ but not yet received in location h is denoted $X^h_{t'}$, and the expected supplier delivery to the central warehouse in week $t'' \geq t$ is denoted $R^W_{t''}$. In addition, a distributional forecast $\mathbf{D}^h_{s,t}$ of demand during week $t \geq s$ available at time s is available for each location h, along with the discrete lead-time distribution $\mathbf{L}^h_t$ of the number of weeks necessary for a shipment determined in week t to reach location h. Both demand and lead-time distributions are non-stationary and location-dependent in this context; this may reflect local seasonal patterns such as a demand peak for malaria drugs during the rainy season due to the development of mosquito populations, and interruptions of access to some health centers due to flooding (see facility accessibility estimates from field data).

As discussed in section §C of the Electronic Companion dedicated to the estimation of demand and lead-times from field data, it appears that facility-level demand in this context can be modeled reasonably well by lognormal distributions with a constant coefficient of variation (see §C.1). We thus use the standard multiplicative form of the martingale model of forecast evolution (MMFE) described in Heath and Jackson (1994) in order to model the demand forecast updating process characterizing the distributional forecast $\mathbf{D}^h_{s,t}$. Specifically, distributional forecasts available in period s for the demand in period $t \geq s$ are generated by the process

$$\mathbf{D}^h_{s,t} = \bar{D}^h_t \exp(\epsilon_{t,t} + \epsilon_{t-1,t} + ... + \epsilon_{s,t}) \exp(\boldsymbol{\epsilon}_{s-1,t} + ... + \boldsymbol{\epsilon}_{t-H,t}), \tag{2}$$

where $\bar{D}^h_t = \mathbb{E}[\bar{\mathbf{D}}^h_t]$ is the demand mean obtained from our dataset with the estimation procedure described in section §C.1 of the Electronic Companion, H is the length of the forecast horizon, and $\epsilon_{u,t}$ are normal random variables representing the uncertainty that is revealed in period $t-u$ concerning demand during period t (bold characters are used to differentiate yet unknown random quantities from their known realizations)[1]. In words, the MMFE through equation (2) defines a quantitative process by which the uncertainty affecting demand at a future time is partially resolved during the time period leading to that point. Consistent with (2), the realization of demand in period t is given by $\bar{D}^h_t \exp(\sum_{u=t-H+1}^{t} \epsilon_{u,t})$, which is indeed lognormal. The baseline scenario for forecasting accuracy considered in this paper corresponds to the "statistical method" derived from actual forecasting data in Heath and Jackson (1994), and for which the provided parameters are $H = 3$ months and 44%, 30%, 18% and 7% of demand variability resolved 3, 2, 1 months before sales and during the month of sales, respectively.

The assumed sequence of events at the beginning of each period is that deliveries to the warehouse and the facilities occur first, inventory levels are updated as demand is realized and then decisions are made. Any demand from patients that cannot be satisfied from available inventory in a given week and

[1] If $\epsilon_{u,t} \sim N(\mu_u, \sigma_u^2)$, then parameters μ_u and σ_u^2 are constrained by $\mu_u = -\sigma_u^2/2$ so that $\mathbb{E}[\exp(\epsilon_{u,t})] = 1$. Others constraints stem from the specified moments of demand $D_{t,t}$ and schedule of demand uncertainty resolution with u.

location is assumed to be lost. The relevant objectives include minimizing total expected lost demand and average inventory levels at the delivery locations over the entire network. The importance of the inventory level at the delivery locations, as opposed to the central warehouse, stems from the typically poorer storage conditions in those facilities and their higher resulting holding costs. Another important objective is to minimize the standard deviation of service levels (defined as the ratio of satisfied demand to total demand) across delivery locations, where these quantities are calculated over some specified time horizon T without weighting facilities by population or demand volume. This objective captures the equity of any distribution policy considered across patients living in different locations, which is an important consideration in this setting (see §1). As the standard deviation of service level may seem an ad-hoc choice among many other possible equity metrics (Marsh and Schilling 1994), we report the results obtained with other metrics in the Electronic Companion.

Note that this problem definition ignores potential interactions between different products (e.g., transportation volume constraints). This is aligned with our field observations that in primary distribution sufficient transportation capacity is available and that, given the typical size of individual shipments, in secondary distribution availability of vehicles is a bigger concern than their capacity. Product expiry is also ignored here, because in contrast with other products requiring cold chains (e.g., vaccines), the shelf life of many essential medicines is relatively long (for example, the shelf life of most antimalarials is about 24 months).

The deliveries by suppliers to the central warehouse R_t^W are also considered exogenous. This is justified by the low frequency of supplier deliveries in this setting (typically a couple per year) relative to outbound shipments (monthly); the upstream procurement process also involves different considerations (contractual agreements, production capacity constraints, multiple suppliers), suggesting a hierarchical / decoupled approach. The partition of facilities into delivery routes and the design of these routes are also assumed fixed and exogenous, as in this setting changing the delivery routes on a frequency comparable to how often shipment decisions are made would be challenging from a practical standpoint. Finally, the assumption of lost demand (as opposed to backorders) is justified by the long travel required for many patients to reach a health center, despite the life-saving nature of many of the commodities considered here. These problem features may not all be appropriate in other settings.

3.3. Current Inventory Distribution Policy

The method currently used in Zambia for addressing the medicine inventory distribution challenge just described is sometimes referred to by practitioners as the min/max policy. It corresponds to the base-stock (s, S) policy described in the classical inventory theory literature, however in practice the min (or s) parameter is often set to max $-$ 1 (or $S - 1$), and therefore omitted thereafter. In the following we will only refer to it as the base-stock policy. This method follows the basic guidelines recommended by the influential USAID-funded DELIVER project, and for that reason is used widely throughout sub-Saharan Africa (USAID | DELIVER PROJECT 2011b). This involves the monthly upstream communication by final delivery locations of replenishment order forms called *requests and requisitions forms*, or R&Rs (see Figure 5 in the Electronic Companion for an example). To coordinate the transmission of this information, each location faces a monthly deadline for transmitting its R&R upstream, which is linked to the schedule of deliveries from MSL to the districts (see §3.1).

R&R forms fill two purposes: (i) they document the aggregate impact of inventory transactions (receipt, deliveries, counting adjustments) on local stock, thus providing a monthly snapshot of downstream inventory levels; and (ii) they facilitate the implementation of the base-stock policy

$$O_t^h = (M \times AMI_t^h) - \ IP_t^h, \tag{3}$$

where:

- O_t^h is the replenishment order quantity requested by facility h in week t;
- M is the *maximum stock level*, representing the number of months of recent past consumption to which inventory should be replenished. Our baseline value for M is 4 months, as used in cross-docking facilities during the 2009 supply chain pilot for example (see Figure 5);
- AMI_t^h is a moving average of past monthly quantities issued to patients by the pharmacy in location h over the 3 months preceding week t. Later in this paper (in §4.1), we consider inventory policies involving different periods for calculating this moving average, and expand then that notation to specify explicitly that calculation period in brackets, so that the default method featured in (3) would for example be denoted $AMI_t^h[-3,0]$;
- IP_t^h is the inventory position of facility h at the beginning of week t, that is the sum of inventory on-hand I_t^h and the total quantity ordered in the past by that facility but not yet delivered to date (see column H in Figure 5). Because of practical challenges associated with tracking pipeline inventory, the inventory on hand I_t^h is sometimes used instead of the inventory position IP_t^h in (3).

Once R&Rs from various facilities are received at the central warehouse, an *allocation rule* is followed in situations where the sum of all replenishment orders received in the same week $\sum_{k\in\mathcal{H}} O_t^k$ exceeds the inventory I_t^W available in the warehouse then. A prevalent allocation rule is *first-come-first-serve*, whereby individual shipment requests are filled in the sequence they arrive until no more inventory is available, at which point the current replenishment order being considered may be filled only partially and the subsequent ones not at all. Another prevalent practice is the *proportional* ($PROP$) allocation rule, which involves reducing all the orders received that week by the same proportion so that the sum of shipments equals the inventory available at the warehouse, that is

$$x_t^h = O_t^h \times \min(\frac{I_t^W}{\sum_{k\in\mathcal{H}} O_t^k}, 1). \tag{4}$$

Based on extensive numerical experiments (not reported here) suggesting that the proportional allocation rule outperforms first-come-first-serve, in the rest of this paper we assume the proportional allocation rule $PROP$ defined in (4) is used.

In summary, the family of inventory distribution policies used in Zambia for addressing the problem described in §3.2 can be described (with subscripts and superscripts omitted) by the following symbolic notation:

$$4 \times AMI - \ IP. \tag{5}$$

This mathematical expression refers to the order calculation method used by each facility (e.g., replenish inventory position to 4 months of the average monthly quantity issued, calculated over the last 3 months). In the absence of any ambiguity about the value of specific parameters within the class of policies discussed here, we will hereafter refer to a policy within this class as the *current policy.*

Considering a single typical facility having access to infinite warehouse inventory, Leung et al. (2016) simulate the inventory replenishment policy (3). They find that the policy $4 \times AMI - IP$ may be directly responsible for many of the stockouts of anti-malaria medicines observed in practice. This is because the forecasts of demand and lead-times implicitly associated with that replenishment order calculation ignore the predictable increases of these two quantities during the peak demand period occurring for these medicines in the first quarter of the year (this seasonality is linked to Zambia's rainy season, when mosquito population increases and flooding cuts off access to some facilities). Specifically, when demand starts to increase in the couple of months preceding the demand peak, the quantities ordered then reflect lower past consumption rates associated with the previous 3 months and are therefore insufficient to cover the upcoming demand surge. The safety stock resulting from the maximum stock level M takes some time to deplete however, so that the first stockouts only appear in the second half of the peak demand period. Because of longer lead-times, the high quantities ordered during the first half are only received after the peak demand period is over, resulting in a wasteful accumulation of inventory. This phenomenon has been observed in other inventory systems facing demand seasonality and coined "the landslide effect" (Neale and Willems 2015). Another pattern is that replenishment orders based on past issues (as opposed to demand) create a negative self-perpetrating cycle whereby historical stock-outs are ignored, resulting in insufficient replenishment quantities and increased likelihood that more stock-outs will subsequently occur. Overall that policy is found to satisfy less than 90% of total demand, even though an infinite amount of upstream inventory is assumed and an average on-hand inventory equal to about two months of demand across the year. These performance concerns, which have been communicated to a number of relevant practitioners in both Zambia and the US since 2010, have motivated the development of the enhanced policies described in the next section.

4. Alternative Inventory Distribution Policies

In §3.3, we described the inventory distribution policy which is currently used to address the inventory distribution challenge of Zambia (see §3.2). We now discuss three alternative policies: (1) a simple modification of the current inventory policy proposed in USAID | DELIVER PROJECT (2011a) (§4.1); (2) another modification of the current policy capturing seasonality factors and lead-times recently described in Watson et al. (2014) (§4.2); and (3) a new rolling horizon policy based on an inventory planning optimization model (§4.3). We mostly focus here on the mathematical definition of these policies, and refer the reader to section §B of the Electronic Companion for a discussion of related implementation issues.

4.1. Last Year Policy

The influential USAID DELIVER project has issued several supply chain management guidance documents for practitioners, including USAID | DELIVER PROJECT (2011a) focusing on anti-malarial commodities. It proposes a number of modification of the current policy described in §3.3, specifically considering different calculation periods for the moving average of past issues and using an estimate of demand as opposed to issues for the basis of the calculation, for example by tracking the number of days without stock in any given month and increasing the quantity issued from stock during that month accordingly. The notation we use to refer to such an enhanced policy is directly derived from that defined in §3.3, for example $AMD_t^h[-12,-9]$ denotes average monthly demand calculated over the three months following the week one year before t. In the absence of any ambiguity about parameter values, we will hereafter refer to a policy within the class $M \times AMD[-12,-9] - IP$ as the *last year policy.*

4.2. Lookahead Seasonality Index (LSI) Policy

The USAID DELIVER project has also funded the study by Watson et al. (2014) to design and evaluate an enhancement of order rule (3) explicitly capturing seasonality in demand and access to facilities. The intent of this approach is to address the performance gaps highlighted in Leung et al. (2016) while minimizing the potential implementation challenges and costs associated with changes to the current system. Specific potential benefits of this approach include (i) the continued use of an explicit formula that ideally remains intuitive to all involved; and (ii) the use of processes and information system for inventory management that are similar if not identical to the current ones.

The modified ordering rule proposed by Watson et al. (2014) can be described as

$$\begin{cases} O_t^h = \left(M \times \frac{\tau_2^h(t) - \tau_1^h(t)}{4} \times LSI_t^h \times AMD_t^h[-3,0] \right) - \ I_t^h \\ LSI_t^h = \frac{\text{average}[s_{\tau_1^h(t)-4}, \ldots, s_{\tau_1^h(t)-1}, s_{\tau_1^h(t)}, \ldots, s_{\tau_2^h(t)}, s_{\tau_2^h(t)+1}, \ldots, s_{\tau_2^h(t)+4}]}{\text{average}[s_{t-12}^h, \ldots, s_{t-1}^h]} \end{cases}, \tag{6}$$

where:

- $\tau_1^h(t)$ is the median of the distribution of time at which the shipment triggered by the order quantity determined in week t will be received at facility h, and $\tau_2^h(t)$ is the median of the distribution of time when the following shipment is received at that facility[2]. The time interval $[\tau_1^h(t), \tau_2^h(t)]$ can thus be interpreted as the *shipment consumption period* during which the shipment determined in week t is meant to cover demand. Its duration $\tau_2^h(t) - \tau_1^h(t)$ is divided by 4 in (6) in order to express the terms M and AMD in months as is prevalent in practice, even though the time periods defining our problem dynamics are one week (see §3.2);
- s_t^h is a *seasonality index* quantifying for facility h the average expected change of consumption in period t relative to another period of reference (to simplify notation the superscript is omitted when obvious from context). When enough historical records are available, Watson et al. (2014) propose to compute s_t^h by considering one continuous year of historical consumption data and dividing the record in each period by the consumption in the first period of the year;
- LSI_t^h is the *lookahead seasonality index* meant to correct the implicit demand forecast associated with the term $AMD_t^h[-3,0]$ so that it better reflects the demand to be expected over the shipment consumption period $[\tau_1^h(t), \tau_2^h(t)]$. The inclusion of the seasonality indices $s_{\tau_1^h(t)-4}, \ldots, s_{\tau_1^h(t)-1}$ and $s_{\tau_2^h(t)+1}, \ldots, s_{\tau_2^h(t)+4}$ in the numerator of the fraction defining LSI_t^h is meant to provide some robustness with respect to possible shifts of seasonality (Watson et al. 2014 recommend the inclusion of the months preceding and following the shipment consumption period). The denominator average$[s_{t-12}^h, \ldots, s_{t-1}^h]$ reflects the time period over which the term $AMD_t^h[-3,0]$ is calculated;

In summary, Watson et al. (2014) propose to modify the calculation of the average monthly issues featured in the existing policy (3) to form a prediction of the average monthly demand rate over the anticipated consumption period relevant to the shipment being considered, with some consideration for robustness[3]. In the remainder of the paper, we will refer to the ordering rule described in this section as $M \times LSI \times AMD - I$ or more simply as the *LSI policy* when no ambiguity about the value of specific policy parameters arises.

[2] This definition generalizes the policy described in Watson et al. (2014), which assumes deterministic lead-times.

[3] Equations (6) actually generalize Watson et al. (2014)'s ordering rule to the case of non-deterministic lead-times. The shipment consumption period $[\tau_1^h(t), \tau_2^h(t)]$ featured in (6) also generalizes and makes explicit their recommendation to increase M for the last shipment before a facility is to get cut off during the rainy season.

4.3. Proposed Optimization-Based Policy

The third main improved policy considered in this paper to address the practical distribution challenge described in §3.2 relies on an inventory planning linear program (LP) that is solved on a rolling horizon basis immediately before any new shipments to facilities are determined. While these high-level features are similar in spirit to the approach described in Foreman et al. (2010), the structure and formulation of the model to be described next are distinct and were developed through extensive and specific numerical experimentation. Its primary decision variables are the quantities of the drug to be sent to each health center as part of each set of weekly shipments scheduled over the planning horizon (see discussion of facility shipping groups in §3.2). These variables thus include both shipments to facilities in the next shipping group and subsequent shipments that are more distant in the future. While these more distant shipments in the time horizon are considered to prevent any myopic behavior (e.g., shipping enough inventory in the short run in anticipation of a possible facility cut-off due to flooding in some future period), upon any model run only computed shipment variables corresponding to the next scheduled delivery are to be implemented. Given the specific seasonality patterns of demand and delivery lead-times in this environment (see §5 for a description of the data), a planning horizon length P of six months to a year seems appropriate (we use $P = 48$ weeks in experiments). The objective retained for this LP is to minimize a weighted combination of expected lost demand and inventory holding costs calculated over the planning horizon for the entire set of delivery locations considered. That expected lost demand function is captured in this LP as a piece-wise linear convex approximation of the original nonlinear lost sales function arising in a model with stochastic demand, using a set of approximating tangents computed from the stochastic primitives of the demand distributions. In the following exact model definition, dependence on the current time period (t_0) is omited when it is clear from context.

Sets and indices:

- $\mathcal{H}$: set of final shipment destinations h (health centers) considered. The central warehouse is denoted W;
- $\mathcal{T}_P = \{t_0, ..., t_0 + P\}$: set of consecutive discrete periods t (weeks) in the planning horizon, where $t_0 \in \{1, ..., T - P\}$ is the first (current) period for which the shipment decisions must be determined;
- $\mathcal{K}_t^h$: set of approximating tangents k for the lost demand function of health center $h \in \mathcal{H}$ in period $t \in \mathcal{T}$.

Input data:

- $D_t^h \triangleq \mathbb{E}[\mathbf{D}_{t_0,t}^h]$: expected demand at health center h during week $t \in \mathcal{T}_P$ estimated in week t_0;

- $L_t^h[\beta]$: β-conservative deterministic equivalent of the lead time from MSL to health center h for a shipment initiated in week t and determined as of t_0, where β is a parameter in $(0,1)$. These β-conservative lead time equivalents are defined as follows:

$$L_t^h[\beta] \triangleq \begin{cases} \min\{n \in \mathbb{N} : \mathbb{P}(\mathbf{L}_t^h = n | \mathbf{L}_t^h > t_0 - t) > 0) \text{ if } t < t_0 \\ \min\{n \in \mathbb{N} : \mathbb{P}(\mathbf{L}_t^h = n) > 0) \text{ if } t = t_0 \\ \min\{n \in \mathbb{N} : \mathbb{P}(\mathbf{L}_t^h \leq n) \geq \beta\} \text{ if } t > t_0 \end{cases}. \tag{7}$$

In words, the β-conservative lead time is equal to the minimum of the lead-time distribution support for the shipment to be determined in the current period t_0 and for any pipeline shipments sent before t_0 but not yet arrived, and it is equal to the β-fractile of the lead time distribution for future shipments. That is, the lead times for past and present shipments is deliberately assumed to be short, while the lead times for all future shipments (most saliently the shipments immediately following the current ones under consideration) is deliberately assumed to be long. The rationale behind this definition is to ensure that the consumption period for the next shipments to be sent, that is the period of time during which these shipments will need to satisfy demand before additional inventory arrives, is conservatively assumed to be long (when β is chosen to be close to 1) by the LP, so it adds appropriate safety stock to avoid lost sales;

- $I_{t_0}^h, I_{t_0}^W$: initial on hand inventory levels at health center h and the central warehouse W, respectively;

- $(X_t^h)_{t<t_0}$: vector of pipeline shipments quantities sent from the warehouse to health center h in a past period $t < t_0$ and which have not yet been received;

- $(R_t^W)_{t>t_0}$: vector of (exogenous) replenishment quantities to be delivered at the warehouse by suppliers in future weeks $t > t_0$ (the assumed sequence of events is such that any supplier delivery in week t_0 is already reflected in the initial warehouse inventory $I_{t_0}^W$);

- A_{tk}^h, B_{tk}^h : slope and intercept of approximating line segment $k \in \mathcal{K}_t^h$ for the lost demand function of health center h during week t (estimated as of week t_0). Since that lost demand $\mathbb{E}[(\mathbf{D}_{t_0,t}^h - i_t^h)^+]$ is a convex function of the starting inventory level i_t^h (see variable definition below), it can be approximated arbitrarily closely by the upper enveloppe of a discrete set $\mathcal{K}_t^h$ of its secants at consecutive points. Their slopes and intercept are calculated using the closed form expressions for the lost demand function that are available under the distributional assumptions for $\mathbf{D}_{t_0,t}^h$ (see §5.2)[4];

- C : weight / cost parameter representing the cost of one unit of lost demand relative to the cost of holding one unit of inventory for one period at a health center.

[4] Omitting subscripts and superscripts t_0, t and h for convenience, for the implementation we define a set of $n+1$ equally spaced fractiles $\{i_0, ..., i_n\}$ as $i_k \triangleq F^{-1}(\frac{k}{n+1})$, where F is the distribution function of $\mathbf{D}$. The approximating line segment with slope A_k and intercept B_k is then obtained as the line joining the points $(i_k, \mathbb{E}[(\mathbf{D} - i_k)^+])$ and $(i_{k+1}, \mathbb{E}[(\mathbf{D} - i_{k+1})^+])$. We use $n = 7$ for the numerical experiments reported in this paper.

Decision variables:

- x_t^h : quantity to be shipped from the warehouse to health center h during week t;
- i_t^h, i_t^W : expected inventory level at the beginning of week t in health center h and the warehouse W, respectively;
- ℓ_t^h : expected shortages (lost demand) at health center h during week t.

Formulation:

$$\min \sum_{h\in\mathcal{H}}\sum_{t\in\mathcal{T}_P}(C\times\ell_t^h + i_t^h) \tag{8}$$

$$\text{subject to:} \quad (1) \text{ and}$$

$$i_{t_0}^W = I_{t_0}^W \tag{9}$$

$$i_{t+1}^W = i_t^W + R_t^W - \sum_{h\in\mathcal{H}} x_t^h \quad \forall t\in\mathcal{T}_P \tag{10}$$

$$i_{t_0}^h = I_{t_0}^h \quad \forall h\in\mathcal{H} \tag{11}$$

$$i_{t+1}^h = i_t^h - D_t^h + \ell_t^h + \sum_{u\in\{u<t_0:u+L_u^h[\beta]=t\}} X_u^h + \sum_{u\in\{u\in\mathcal{T}_P:u+L_u^h[\beta]=t\}} x_u^h \quad \forall h\in\mathcal{H}, t\in\mathcal{T}_P \tag{12}$$

$$\ell_t^h \le D_t^h \quad \forall h\in\mathcal{H}, t\in\mathcal{T} \tag{13}$$

$$\ell_t^h \ge A_{tk}^h i_t^h + B_{tk}^h \quad \forall h\in\mathcal{H}, t\in\mathcal{T}_P, k\in\mathcal{K}_t^h \tag{14}$$

$$x_t^h, i_t^W, i_t^h, \ell_t^h \ge 0 \quad \forall h\in\mathcal{H}, t\in\mathcal{T}_P \tag{15}$$

In this deterministic inventory planning LP, the objective (8) captures the weighted sum of lost demand and inventory holding costs over the planning horizon and the set of health centers considered. Constraint (1) ensures that in any given week only shipments corresponding to the delivery route associated with that week are considered (see §3.2); constraints (9)-(10) and (11)-(12) are the inventory balance equations for the warehouse and all health centers, respectively; constraints (13)-(14) implement the linear piecewise approximation of the lost demand function; and constraint (15) ensures that all decision variables are non-negative, which together which (10) implies that total shipments in any period do not exceed the inventory available at the warehouse then. In the remainder of the paper, we will refer to the policy described in this section as OPT_β^C, or the *optimization policy* when no ambiguity about the value of policy parameters arises.

5. Simulation Model

The simulation model to be described next is used in this paper to evaluate and understand the performance of the proposed enhanced inventory distribution policies described in §4, both in absolute terms and relative to the inventory control policies currently used in Zambia.

Consistent with the problem statement in §3.2, we mainly evaluate policy performance along the following metrics:

System service level: The proportion of total patient demand satisfied from available inventory calculated over all health facilities in the distribution network;

Distribution equity: The standard deviation of service levels calculated independently for each health facility across all sites in the distribution network, reflecting distribution fairness or the degree to which access to drugs may differ for patients depending on their geographic location. Because other equity metrics besides standard deviation also seem meaningful in this context (Marsh and Schilling 1994), we report later in §6 on extensive numerical experiments conducted with other equity metrics;

Average inventory level: The average total on hand inventory level at the health facilities in the distribution network, expressed in weeks of average system demand.

Note that these three metrics allow us to quantify the tradeoffs between inventory cost, system service level, and service equity, as part of this study.

Some questions not considered in this paper are also important in the context of pharmaceutical distribution in low-income countries and/or challenging inventory models (see §8). The facility, demand and lead-time data we rely on, the description of the generation process for that data and the simulation code provided in the public repository `https://github.com/zacharyleung/ZappSupplyChainSimulator` are meant to facilitate future related work by other researchers.

In the remainder of this section, we discuss the scope and structure of our simulation model in §5.1, then the policies and scenarios considered in §5.2. For details on the construction and estimation from data of our demand and lead-time simulation models as well as a description of the work performed to validate and evaluate the predictive accuracy of this simulation model, we refer the reader to section §C in the Electronic Companion.

5.1. Model Scope and Structure

Our simulation model focuses on the distribution of the anti-malarial medicine Artemether/Lumefantrine (brand name Coartem®, from now on abbreviated as AL), for the following reasons. First, AL is important to global public health as the recommended first-line treatment for malaria in many countries including Zambia. Second, the demand for AL is seasonal because malaria incidence is highly correlated with rainfall patterns due to mosquito population dynamics, and Zambia experiences a marked rainy season between December and March. Third, AL is distributed to all health facilities in Zambia, including locations that are particularly challenging to access through all or part of the year. AL is thus an important product in itself, but is also a meaningful test case from a policy design perspective as the most challenging product, from a distribution standpoint, in a set of several hundred essential medicines that do not involve storage and transportation temperature restrictions or very short shelf lives. Finally, because AL was one of the main tracer drugs considered as part of the 2009 supply chain pilot, we can use for model validation purposes the availability evaluation data collected independently for this product then (see §6.1). While AL products come in four different pack sizes, demand and inventory for these pack sizes are fully substitutable, so that in all experiments except those reported in §6.1 we aggregate them and consider these pack sizes together as a single product[5].

Written in the Java programming language, our model uses a discrete-event structure and weekly time period and predicts on-hand inventory dynamics of all combined AL products in each location of a network comprising the central warehouse, 12 district health offices and 212 health facilities. This geographic coverage amounts to approximately 17% of Zambia's facilities and corresponds to the districts for which historical demand and lead time data could be collected by leveraging the presence of a commodity planner during the 2010 supply chain pilot (see §3.1). We note that these districts were selected as a subset that is representative of the entirety of Zambia as part of the experimental design of that pilot (Vledder et al. 2019). In addition, these facilities cover three of the four sequential distribution groups discussed in §3.2, allowing investigation of the impact of this partition.

The sequence of events simulated by this model in each period is the following:

1. Planned receipts are added to the on-hand inventory of each location;

2. Demand in each health center is generated according to the nonstationary stochastic demand model described in §C.1 of the Electronic Companion. This demand is debited from the local on-hand inventory and lost demand is recorded in case demand exceeds available inventory;

3. Shipments from the central warehouse to the set of facilities on the shipment schedule that week are computed according to the inventory policy being simulated (see §5.2) and debited from the central warehouse inventory. Reflecting the actual pre-determined monthly primary distribution schedule of fixed truck routes covering each a subset of districts (see §3.2), the simulated districts are evenly partitioned into subsets associated with single monthly shipment opportunities. Lead times for these shipments are generated according to the nonstationary stochastic lead time model described in §C.2 of the Electronic Companion, and the corresponding planned receipts are added to a list of future events.

[5] AL comes in four different pack sizes (6, 12, 18 and 24), with the numbers indicating the quantity of pills included in a tablet constituting a single treatment dose for a patient. Any box of AL contains 30 individual tablets regardless of pack size. The treatment dose for a patient is dependent on his/her body weight, with the pack size of 24 intended for adults and smaller ones for children with smaller body weights. The pills provided in each tablet are rigorously identical across pack sizes however, so that two tablets of AL 6 or half of a tablet of AL 24 would be be provided to a patient requiring a dose of 12 pills whenever a stockout of AL 12 would occur, etc. Our aggregation of these four pack sizes accounts for the different number of pills in each tablet.

This simulation model is designed to capture some key features of Zambia's distribution system that include the predictable and unpredictable variability associated with both demand and shipment lead times, the monthly order and shipment schedule at the central warehouse and the scarcity of central inventory. Its assumptions reflect the discussion in §3.2 and include the independence of different health products, the exogenous supplier deliveries to the central warehouse and the exogenous schedule of delivery routes to the districts. Instead of using actual data of inbound deliveries by suppliers to the central warehouse, we will assume regular shipments of equal quantities every quarter of the year (as is frequently observed in practice) and will vary the quantities of these shipments in relation to total demand in the distribution network in order to examine different levels of supply scarcity.

5.2. Policy and Scenario Parameters

The families of inventory distribution policies we evaluate here are as described in §3.3 and §4.1-§4.3, respectively. Within each family, a given policy is characterized by family-specific inventory control parameters such as the maximum stock level M, the lost demand penalty C and the lead time fractile β.

The key simulation scenario variable that we consider is the scarcity of inventory available at the central warehouse relative to network-wide demand at the health facilities. Specifically, we assume that the upstream procurement process is characterized by the delivery of R^W units of inventory to the central warehouse four times per year (52 weeks), and vary R^W to achieve different values of the *suppy/demand ratio* (hereafter denoted S/D) defined as

$$S/D \triangleq \frac{4R^W}{\sum_{t=1}^{52}\sum_{h\in\mathcal{H}}\bar{D}_t^h},$$

where the denominator represents the sum over all health facilities of average simulated demand through one year.

In order to generalize our results (to medicines besides AL and countries besides Zambia), we ran a set of sensitivity analysis experiments, where we considered different demand and facility lead time scenarios. For the experiments reported in §6.3.1 we derive different demand datasets from our original Zambia dataset by varying the relative level of demand seasonality using a *demand seasonality parameter* $\phi_D \in [0,1]$. Specifically, the mean demand in each period t and location h in these modified dataset is obtained by replacing the analogous quantity $\bar{D}_t^h$ of the original dataset by

$$\bar{D}_t^h[\phi_D] \triangleq \bar{D}_t^h + (1-\phi_D)\left(\dot{D}^h - \bar{D}_t^h\right),$$

where $\dot{D}^h \triangleq \frac{1}{T}\sum_{t=1}^{T}\bar{D}_t^h$ is the average demand mean for facility h throughout the year. As a result, $\phi_D = 1$ recovers our original malaria drug demand dataset, while at the other extreme $\phi_D = 0$ generates a dataset with constant expected demand in each location (that expected demand may still vary across locations however).

For the experiments reported in §6.3.2, we similarly derive different facility lead time datasets from our original Zambia dataset by varying the relative level of facility access challenges using a *facility access challenge parameter* $\phi_L \in [0,1]$. Specifically, the accessibility probability for facility h at time t in these modified datasets (see §C.2) is obtained by replacing the analogous quantity a_t^h estimated in our original dataset by

$$a_t^h[\phi_L] \triangleq \phi_L a_t^h + (1-\phi_L).$$

As a result, $\phi_L = 1$ recovers our original Zambia lead time dataset, while at the other extreme $\phi_L = 0$ generates a dataset where every facility is always accessible and has a constant lead-time through the year (lead-times may still vary across locations however).

6. Numerical Simulation Results

Our first set of simulation experiments reported in §6.1 focuses on validating and establishing the predictive accuracy of the simulation model just defined. We then discuss our baseline policy performance evaluation experiments in §6.2, and numerically explore in §6.3 the robustness of these policies with respect to various parameters and environment features. We finally provide an interpretation and summary of our results in §6.4. In addition, we report our empirical findings that our results are not qualitatively affected by the choice of the equity metric in Electronic Companion section §E. For each simulation experiment we report the maximum relative margin of error over all simulation estimates, defined as the half-width of the 95% confidence interval divided by the estimate.

6.1. Simulation Model Validation

Our validation relies on a survey of facility performance over the 4th quarter of 2009 that was conducted by a private contractor as part of the 2009 public sector supply chain pilot evaluation (see §3.1 for background and (Vledder et al. 2019) for a more detailed discussion). Importantly, the data source just mentioned is completely distinct / independent from all other data collection activities described earlier in this paper, including those used to develop our simulation model and estimate its parameters.

Specifically, among other data that contractor collected then from the locally-maintained, paper-based stock records of 192 health centers the number of days of stockout in Q4 2009 (with a maximum possible value of 92 days) for all drugs in a tracer list that included the four pack sizes of AL. In addition, the intersection of the set of pilot facilities surveyed then and the set of health centers captured by the simulation model described earlier in this section includes 51 health centers that were supplied by cross-docking districts, and 34 health centers that were supplied by intermediate stocking districts during the

Table 1 Model Validation Results

Intervention	Pack Size	Mean Actual Stockout Days	Mean Simulated Stockout Days	Simulated Fractile of Mean Actual Stockout Days
Intermediate Stocking	6	12.62	13.29	0.412 (0.382, 0.443)
	12	4.85	5.76	0.349 (0.320, 0.379)
	18	7.56	5.86	0.819 (0.794, 0.842)
	24	7.85	11.26	0.075 (0.060, 0.093)
Cross-docking	6	4.18	3.32	0.786 (0.760, 0.810)
	12	0.00	0.32	0.521 (0.490, 0.552)
	18	1.39	0.33	0.932 (0.915, 0.946)
	24	1.80	1.66	0.622 (0.592, 0.652)

Note. Simulated results were obtained with 1000 replications. For the simulated fractile of mean actual stockout days, 95% confidence intervals calculated with the Wilson formula are reported.

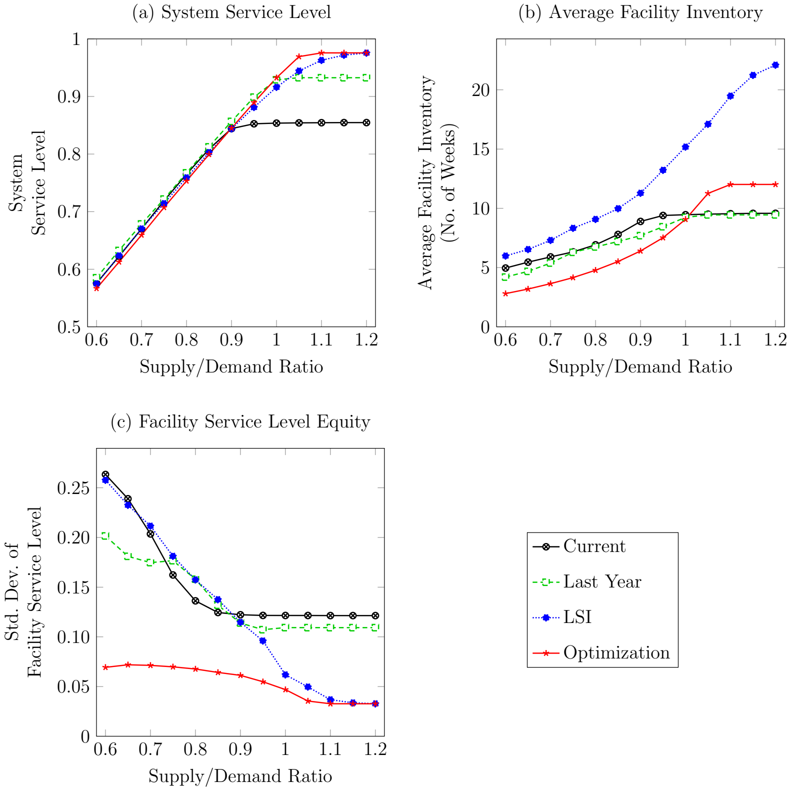
pilot (see §3.1). These facilities are hereafter denoted *validation facilities*. This situation thus offers an opportunity to compare for these facilities the number of stockout days measured empirically as part of this independent evaluation with the simulated number of stockout days predicted by our simulation model when assuming the same inventory distribution policies as those used during the pilot. Because it was not clear how to meaningfully aggregate stockout days across different pack sizes from the field observations, as explained in more details in section §D of the Electronic Companion we set up the model to separately simulate the inventory of all four pack sizes. We otherwise used the same discrete-event simulation dynamics, demand model and lead time model that were used for all other simulation experiments reported in this paper (see §5.1 and sections §§C.1 and C.2 of the Electronic Companion, respectively). The simulated number of stockout days was then recorded for each replication between weeks 36 and 48 of that simulated time period, corresponding to the fourth quarter of 2009 for which the same data was obtained from the field.

The results of these validation experiments are shown in Table 1, which specifically reports for every pack size and intervention the estimated fractiles of the actual number of stockout days measured empirically relative to the simulated distribution of stockout days predicted by the model (last column). In other words, these fractiles provide an indication of how likely it is that the actual observed values of stockout days could have been generated by the simulation model, with moderate values between 0.05 and 0.95 indicating for example that the hypothesis that the simulation model would generate the same stockout days as observed in the field cannot be rejected at the 10% confidence level. These computed fractiles indeed have moderate values from a statistical testing standpoint, with the exception of their minimum (0.075) and maximum (0.932). Table 1 thus indicates that the actual measurements of stockout days are all fairly likely under the mathematical environment assumed for our simulation model. While the actual measurements of stockout days for pack size 24 in the intermediate stocking districts and for pack size 18 in the cross-docking districts seem slightly less likely than all other values, we observe that the corresponding difference between actual and simulated mean estimates remains relatively small in absolute terms. On this basis we conclude that our simulation model seemingly offers a suitably realistic prediction of the service level associated with a given inventory distribution policy in this setting, which lends support to the conclusions derived from the numerical performance evaluation experiments presented in §6.

6.2. Baseline Performance Evaluation

Our baseline simulation experiments investigate the relative performance of policies $4 \times AMI - IP$, $4 \times AMD[-12, 9] - IP$, $4 \times LSI \times AMD - I$ and $OPT^{16}_{0.99}$ when the supply/demand ratio S/D varies between 0.6 and 1.2. The specific parameters characterizing these policies ($M = 4$, $\beta = 0.99$, $C = 16$) were selected through extensive numerical experimentation as the best performing variants of the three enhanced policy families discussed in §§4.1-4.3. In the remainder of this section these four policies will be referred to as *current, last year*, *LSI* and *optimization*, respectively. Our baseline experiments results are summarized in Figure 1.

Figure 1 Performance of the three enhanced distribution policies against the current distribution system in Zambia for different supply/demand ratios.



(a) System Service Level

(b) Average Facility Inventory

(c) Facility Service Level Equity

Note. Displayed simulation results have a relative margin of error lower than 8.5%.

Figure 1 shows that the system service levels achieved by all policies are similar and close to the maximum achievable for values of S/D lower than 0.85 – when the overall inventory available to cover demand is grossly insufficient, distribution decisions resulting from even simple policies have little impact on system-wide service level as the probability of fulfilling some demand with each available unit of inventory is similarly high regardless of where it is shipped. For S/D larger than 0.85, all three proposed enhanced policies substantially outperform the current one, which does not ever achieve a system service level greater than 85% and keeps facilities stocked at about 7 weeks of inventory on average despite the increasing availability of central inventory. The last year policy performs particularly well in terms of both service level and facility inventory for S/D smaller than 1, which is remarkable given its simplicity (see §B). For S/D values above 1 however, its performance also flattens at a service level of approximately 93% and 9 weeks of average facility inventory. In contrast, the optimization policy achieves a service level of approximately 98% with an average facility inventory of 12 weeks when S/D reaches 1.1 and beyond. Finally, the LSI policy only achieves the same service level of 98% when S/D reaches 1.2, and its performance is substantially worse than all other policies in terms of facility inventory (a 98% service level requires facilities to carry about 22 weeks of inventory on average).

Panel (c) in Figure 1 shows the performance of these four policies along the dimension of geographic equity in service level. We find that the optimization policy significantly outperforms all other considered policies for all S/D values, with the exception of the LSI policy for the highest S/D value considered. Of note, the equity performance of the current and last year policy remains flat and substantially worse than that of the LSI and optimization policies even for high S/D values when ample inventory is made available to them.

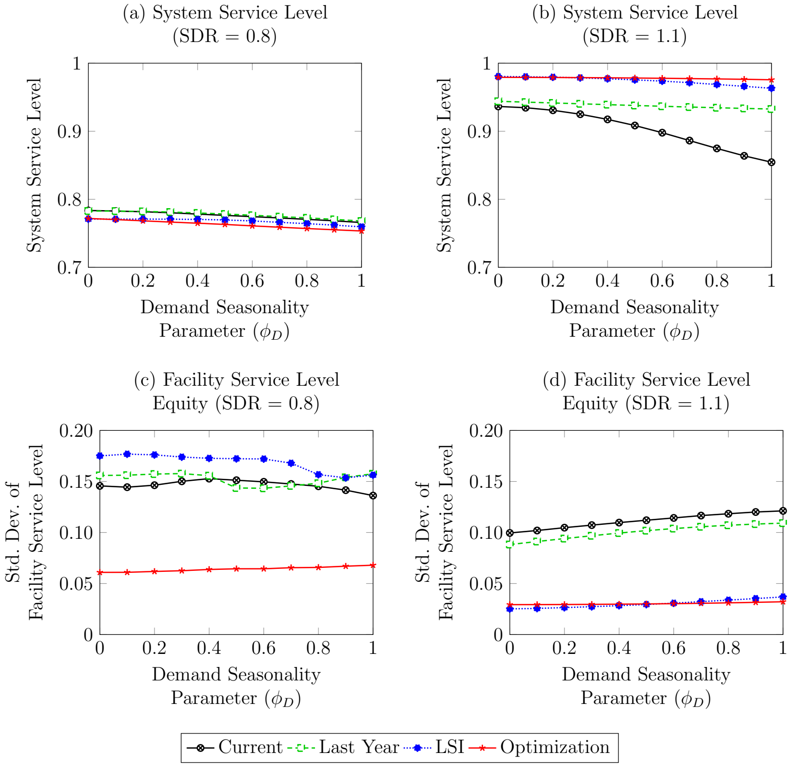
6.3. Sensitivity Analysis

We now report the results of sensitivity analysis experiments conducted to evaluate the impact on policy performance of demand seasonality (in §6.3.1) and facility lead time and access challenges (in §6.3.2). In section §F of the Electronic Companion we also consider the impact of information transmission speed.

6.3.1. Demand Seasonality Figure 2 shows the simulated performance of the four considered policies for different levels of demand seasonality ϕ_D and inventory scarcity S/D (see §5.2).

As seen in panels (a) and (c) of Figure 2 ($S/D = 0.8$), the performance of all policies considered appears relatively insensitive to the level of demand seasonality when inventory is scarce. When more inventory becomes available however, the service level performance of the current policy as well as the equity performance of the current and last year policies, and to a lesser extent the LSI policy, degrade substantially as demand seasonality increases (panels (b) and (d)). In contrast, the performance of the optimization policy on all dimensions considered appears relatively robust to different demand seasonality levels.

Figure 2 Performance of the current, last year, LSI and optimization policies for different levels of demand seasonality and inventory scarcity.

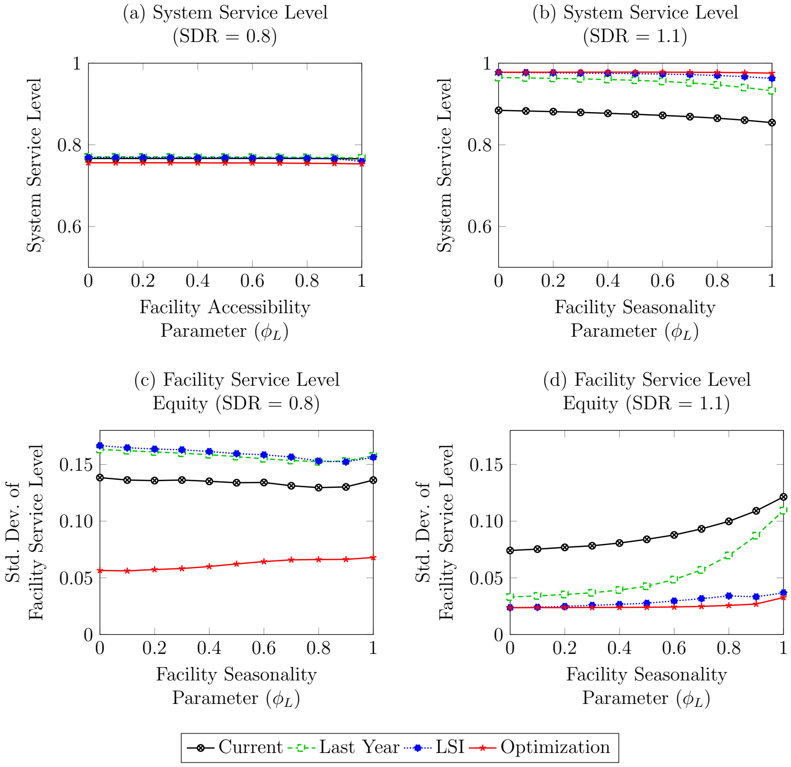


Note. $\phi_D = 1$ corresponds to the original Zambia demand dataset, $\phi_D = 0$ to stationary demand in each facility. Relative margin or error lower than 8.8% on all displayed estimates.

6.3.2. Facility Access Challenges Figure 3 shows the simulated performance of the four considered policies for different levels of facility accessibility ϕ_L and inventory scarcity S/D (see §5.2).

As seen in panels (a) and (c) of Figure 3 ($S/D = 0.8$), when inventory is scarce the performance of all policies considered appears mostly insensitive to the onset of facility accessibility challenges, with the possible exception of the optimization policy, which sees its equity performance slightly degrade as some facilities become harder to reach in some parts of the year. Although these resuts are not reported in Figure 3, we also observe that the LSI policy sees a substantial increase of its average facility inventory levels as facility access probabilities become close to their values in the original Zambia facility dataset. When more inventory is available relative to demand ($S/D = 1.1$) however, as accessibility challenges appear the service level performance of the current, last year and LSI policies slightly degrade (panel b) while the facility inventory levels of the LSI and optimization policies increase somewhat. But the most salient variation is the substantial deterioration of the current and last year policies' equity performance as accessibility challenges appear, even as the performance of the LSI and optimization policies along this dimension remains relatively stable (panel d).

Figure 3 **Performance of the current, last year, LSI and optimization policies for different levels of facility accessibility and inventory scarcity.**



Note. $\phi_L = 1$ corresponds to the original Zambia lead times dataset, $\phi_L = 0$ to year-round accessibility and stationary lead times for each facility. Relative margin or error lower than 8.5% on all displayed estimates.

6.4. Results Interpretation and Summary

Supporting our interpretation of the results just presented, Figure 4 provides X-Y plots of every facility along the dimensions of average service level and mean access lead-time for each one of the four policies just discussed when S/D is set to 0.8; the delivery groups of each facility (see §3.2) and the coefficient of variation of the average weekly demand series it faces are also highlighted for analysis purposes.

We now highlight and explain three salient observations from the experimental results just presented.

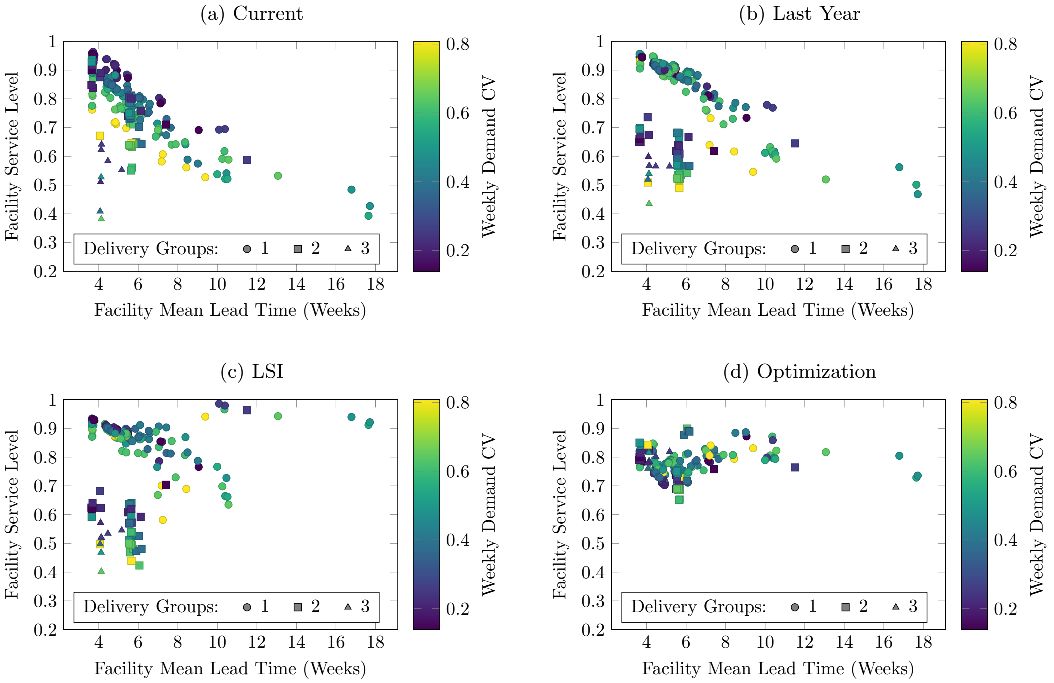
Observation 1: When inventory is scarce relative to demand the choice of the inventory distribution policy has limited impact on the utilitarian objective of system-wide service level.

This observation is clear from panel (a) of Figure 1, and arises because when the S/D ratio is low the probability of fulfilling some demand with each available unit of inventory is similarly high regardless of where it is shipped.

Observation 2: The consideration of equity across facilities has salient policy implications.

In contrast with the previous observation, Figure 4 and panel (c) of Figure 1 show that even in situations of inventory scarcity where different policies have similar average service levels, their equity performance can be very different. In particular a recommendation to use the last year policy, which

Figure 4 **Average service levels of individual facilities under the current, last year, LSI and optimization policies.**



Note. Assumed supply/demand ratio is 0.8. Relative margin of error lower than 6.5% on service level estimates and lower than 2.8% on lead time estimates. Weekly Demand CV is the coefficient of variation of the average weekly demand series over one year for each facility.

could be easily justified in light of its simplicity and relatively good performance for service level and facility inventory level (panels (a) and (b) of Figure 1), should arguably be reconsidered when equity must also be taken into account (panel (c) of Figure 1, panel (b) of Figure 4). Likewise, Figure 1 shows that the overall performance of the LSI policy is quite appealing in situations with high enough supply to demand ratio and when average facility inventory levels are not too much of a concern. Whenever inventory becomes scarce however, its poor equity performance (panel (c) of Figure 1, panel (c) of Figure 4) seems to preclude recommendation.

Observation 3: The set of policies adapted to a specific distribution environment depends on its demand seasonality, facility access heterogeneity and inventory scarcity.

Panel (a) of Figure 4 shows the high correlation between facility demand seasonality (the color of each point) and service level for the current policy, and panel (b) of Figure 2 shows the substantial degradation of both system-wide service level and equity generated by that policy when demand seasonality increases. This is explained by that policy's lack of any anticipation for upcoming predictable seasonal changes in demand, which is clear from its defining equation (3). As a result, it generates substantial stockouts during the rainy season when demand for antimalarial medicines is highest, regardless of how much central inventory is available then, a well-known phenomenon which has been previously coined "the landslide effect" (Neale and Willems 2015). These results contrast with the lack of apparent negative correlation between demand seasonality and performance for the LSI and optimization policies seen in panels (c) and (d) of Figure 4, confirmed by the performance robustness of these two policies against demand seasonality seen in Figure 2. This is consistent with the mathematical definitions of the LSI and optimization policies, which both involve more sophisticated mechanisms for capturing demand seasonality (see (6) and (2)/(12), respectively).

The negative correlation between facility service level and mean access lead time seen in panels (a) and (b) of Figure 4 for the current and last year policies, along with their performance sensitivity seen in panels (b) and (d) of Figure 3, show that access lead times constitute another driver of inequity between facilities for these policies. This contrasts with the LSI and optimization policies, for which no such correlation can be observed in Figure 4, and is confirmed by the results of the broader set of experiments reported in Figure 3. These observations are consistent with the lack of any component capturing lead times explicitly in the equations characterizing the current and last year policies (see §3.3 and §4.1), and the relative sophistication with which the lead times of different facilities are accounted for in the mathematics defining the LSI and optimization policies (see (6) and (7)). This feature of the LSI and optimization policies explain both the higher system service level and higher facility inventory levels they achieve for high S/D values. Specifically, whenever sufficient central inventory is available these policies send larger shipments to facilities with access challenges before their cutoff periods, which the current and last year policies fail to do.

Finally panels (a), (b) and (c) of Figure 4 show that facilities from shipment groups 2 and 3 tend to have lower service levels than facilities in group 1 with the current, last year and LSI policies. This is consistent with they myopic nature, that is their restricted consideration of only demand from the shipment group for which shipments are being determined. In contrast, panel (d) in Figure 4, panel (c) of Figure 1 and panels (c)-(d) of Figures 2 and 3 show that the optimization policy does not generate a consistent pattern of inequity across facilities with different access lead-times, demand seasonality and/or delivery groups. Although equity is not explicitly captured in the objective (8) of the LP formulation used by this policy, the convexity of this objective with respect to the inventory levels resulting from the shipments being determined (which impact both terms i_t^h and ℓ_t^h of (8), see §4.3) still ensures a relatively balanced allocation of inventory across facilities. In contrast with the last year and LSI policies, the optimization policy is not myopic as it considers through the summation over t in its objective (8) the implications of current decisions on service levels of all facilities in future periods. As a result, the optimization policy does not penalize facilities in delivery groups 2 and 3 relative to those in delivery group 1 to the same extent that the current, last year and LSI policies do.

Table 2 Qualitative Policy Robustness Results

	Robustness with respect to:		
Policy	Demand Seasonality	Heterogeneous Facility Access	Dynamic Rationing of Limited Inventory
Current	No	No	No
Last Year	Mixed	No	No
LSI	Yes	Yes	No
Optimization	Yes	Yes	Yes

Summarizing these observations, Table 2 provides a qualitative summary of our findings on the robustness of the inventory policies evaluated in this paper with respect to several important environmental features, which is relevant to the potential application of our results across a range of different drugs and countries. Implementation issues are further discussed in the next section.

7. Implementation Discussion

While our study focuses on inventory distribution as opposed to routing, the framework introduced by Vries and Wassenhove (2020) for analyzing the cost-effectiveness of decision support systems for humanitarian logistics remains relevant. Specifically, in times when sufficient central inventory is available, the LSI policy and (for drugs and countries with little demand seasonality and/or facility access challenges) the last year and the current policies are relatively appealing from an overall cost-effectiveness standpoint. The current and last year policies, and to a lesser extent the LSI policy, also offer some benefits in terms of organizational culture because their rationale is more easily explained relative to the "black box" nature of the optimization policy. In any situation when central inventory becomes scarce however, these decentralized policies have lower performance in terms of service level and/or equity relative to the centralized optimization policy considered here, with potentially severe negative health consequences.

The optimization policy does involve implementation challenges and costs linked to its higher informational and computational requirements however. In particular, this policy requires the capabilities of lead times and demand forecasting for each product and facility in the network. It is therefore synergistic with a digital distribution information system. As seen in Figure 11, such a system would only have a mild impact on the main performance metrics considered in this paper. But its most important benefits would arguably include increasing system transparency and accountability, reducing non-value added and manual work linked to inventory management in chroniquely understaffed patient-facing facilities, and enabling a shift of demand forecasting for procurement from notoriously unreliable epidemiology-based "quantification" exercises (Management Sciences for Health 1997) to more reliable bottom-up and data-driven methods.

A preliminary version of the results described in this paper was presented in 2010 to Zambia's Ministry of Health, MSL and other partners including the World Bank, DFID, UNDP and Crown Agents. Motivated by these results and the benefits of supply chain digitization, these institutions formed in 2010 a partnership known as eZICS (enhanced inventory control system for Zambia). Partnering with IBM, this alliance proceeded to develop a system comprising connected smart phones with a bar code scanner at all inventory storage locations; a forecasting component with a user-friendly interface; a shipment optimization component interfacing with the central warehouse management software used by MSL, and a web-based transaction and performance reporting system (IBM 2014). Its field deployment started in early 2016 (The World Bank 2016), and by 2018 it was used in 60 health centers, posts and hospitals located in 8 different districts where it was used to manage the flow of products on a routine basis. While the feedback received from many stakeholders including health center staff was very positive, deployment was paused in 2018 before a field-based quantitative performance evaluation could be completed. This interruption was due in part to funding constraints, reflecting the well-documented challenges faced by low-income countries seeking to improve their health systems and infrastructure (so called "horizontal" investments, as opposed to disease-specific "vertical" programs). It may also reflect the challenges of managing different objectives and constraints within partnerships that involve public and private stakeholders, and the challenges of implementing disruptive technological changes in the highly political environment of international development assistance for health. From the broader perspective of using optimization-based planning systems in humanitarian contexts that is discussed in Vries and Wassenhove (2020), this experience also suggests that the implementation of such systems remains challenging even when their cost-effectiveness is high.

8. Conclusion

Improved inventory control policies can improve patient access to drugs in Zambia, as demonstrated by the empirical results reported in §6. Our results have broader relevance beyond Zambia. Zambia's current inventory policy is the specific base-stock policy recommended by the multi-country USAID-funded DELIVER project. Similar policies are widely used throughout sub-Saharan Africa (USAID | DELIVER PROJECT 2011b) and LMICs in other regions. Specifically, we find that the current inventory policy as well as the recently proposed enhancements to that policy (so called last year and LSI policy in the present paper) exhibit a relatively poor performance in terms of equity whenever inventory in the central warehouse is scarce, a situation prevalent in Zambia and many LMICs. In contrast, the optimization policy described in this paper appears relatively robust with respect to inventory scarcity. These results appear to be robust across a variety of demand seasonality and facility access challenge levels, as well as various equity metrics. This suggests that the dimension of the service level equity across facilities or regions should be considered in the design of inventory management policies for countries' health systems. Equity metrics should be evaluated in the context of evaluating global progress towards the SDGs, in particular SDG 3.8.

Our empirical results also suggest broader observations on distribution equity. Firstly, the proportional inventory rationing rule, which is prevalent in practice, may lead to substantial service level discrepancies between facilities whenever there is any heterogeneity between them in terms of access lead-times or timing. Furthermore, the performance dimension of distribution equity seems frequently at odds with the more traditional performance metrics of both system service level and inventory costs, and this finding appears to be robust across a range of different equity metrics. The consideration of distribution equity, which seems important in any health-related distribution system, therefore gives rise to non-trivial design considerations and trade-offs. In particular, the existence of three relevant metrics (inventory cost, system service level, service level equity) suggests theoretical considerations extending beyond the classical trade-off of efficiency versus equity highlighted in the literature dedicated to other contexts or more generic resource allocation problems (e.g. Bertsimas et al. (2011)), which would add to the existing fairness literature.

Future research could also investigate the specific problem of distributing products with strict storage and transportation temperature restrictions such as vaccines, which require both different physical assets and management policies compared to the essential medicines that are covered by the present study. Separate studies could also consider other important distribution system components beyond inventory control, such as facility location and delivery route design as well as incentive aspects and outsourcing of some distribution activities to private providers. We hope that the validated simulation model and related datasets made public as part of the present study will facilitate these endeavours.

References

Agrawal, Narandra, Morris A. Cohen. 2001. Optimal material control in an assembly system with component commonality. *Naval Research Logistics* **48**(5) 409–429.

Axsäter, S, J Marklund, E A Silver. 2002. Heuristic methods for centralized control of one-warehouse, n-retailer inventory systems. *Manufacturing Service Oper. Management* **4**(1) 75–97.

Barrington, J, Olympia Wereko-Brobby, Peter Ward, Winfred Mwafongo, Seif Kungulwe. 2010. SMS for life: a pilot project to improve anti-malarial drug supply management in rural Tanzania using standard technology. *Malaria Journal* **9** 298.

Bertsimas, Dimitris, Vivek Farias, Nikolaos Trichakis. 2011. The price of fairness. *Operations Research* **59**(1) 17–31. Doi 10.1287/opre.1100.0865.

Cameron, A, M Ewen, D Ross-Degnan, D Ball, R Laing. 2009. Medicine prices, availability, and affordability in 36 developing and middle-income countries: a secondary analysis. *The Lancet* **373**(9659) 240–249.

Caro, Felipe, Jérémie Gallien. 2010. Inventory management of a fast-fashion retail network. *Operations Research* **58**(2) 257–273.

Daff, D. M., C. Seck, H. Belkhayat, P. Sutton. 2014. Informed push distribution of contraceptives in senegal reduces stockouts and improves quality of family planning services. *Global Health Science Practice* **2**(2) 245–252. Doi: 10.9745/GHSP-D-13-00171.

Ernst, Ricardo, Bardia Kamrad. 1997. Allocation of warehouse inventory with electronic data interchange and fixed order intervals. *European Journal of Operational Research* **103** 117–128.

Foreman, John. 2008. Optimized supply routing at Dell under non-stationary demand. Master's thesis, Massachusetts Institute of Technology.

Foreman, John, Jérémie Gallien, Julie Alspaugh, Fernando Lopez, Rohit Bhatnagar, Chee Chong Teo, Charles Dubois. 2010. Implementing supply routing optimization in a make-to-order manufacturing network. *Manufacturing & Service Operations Management* **10**(4) 547–568.

Gallien, Jeremie, Iva Rashkova, Rifat Atun, Prashant Yadav. 2017. National drug stockout risks and the global fund disbursement process for procurement. *Production and Operations Management* **26**(6) 997–1014.

Graves, Stephen C. 1996. A multiechelon inventory model with fixed replenishment intervals. *Management Science* **42**(1) 1–18.

Heath, D C, P L Jackson. 1994. Modeling the evolution of demand forecasts with application to safety stock analysis in production/distribution systems. *IIE Transactions* **26**(3) 17–30.

Hedman, Lisa. 2016. Medicines shortages: Global approaches to addressing shortages of essential medicines in health systems. *WHO Drug Information* **30**(2) 180–185.

Hwang, Bella, Amir Shroufi, Tinne Gils, Sarah Jane Steele, Anna Grimsrud, Andrew Boulle, Anele Yawa, Sasha Stevenson, Lauren Jankelowitz, Marije Versteeg-Mojanaga, Indira Govender, John Stephens, Julia Hill, Kristal Duncan, Gilles van Cutsem. 2019. Stock-outs of antiretroviral and tuberculosis medicines in South Africa: A national cross-sectional survey. *PLOS ONE* **14**(3). Https://doi.org/10.1371/journal.pone.0212405.

IBM. 2014. Zambian government and ibm provide improved access to life saving drugs. https://www-03.ibm.com/press/us/en/pressrelease/43960.wss.

Kraiselburd, Santiago, Prashant Yadav. 2013. Supply chains and global health: an imperative for bringing operations management scholarship into action. *Production and operations management* **22**(2) 377–381.

Leung, Ngai-Hang Z., Ana Chen, Prashant Yadav, Jérémie Gallien. 2016. The impact of inventory management on stock-outs of essential drugs in sub-saharan africa: Secondary analysis of a field experiment in zambia. *PLOS One* **11**(5). E0156026. doi:10.1371/journal.pone.0156026.

Makridakis, S, S Wheelwright, R Hyndman. 1998. *Forecasting: methods and applications*. 3rd ed. Wiley.

Management Sciences for Health. 1997. *Managing Drug Supply: The Selection, Procurement, Distribution, and Use of Pharmaceuticals*. 2nd ed. Kumarian Press, West Hartford, CT.

Marsh, Michael T., David A. Schilling. 1994. Equity measurement in facility location analysis: A review and framework. *European Journal of Operations Research* **74** 1–17.

McCoy, Jessica H., Hau L. Lee. 2014. Using fairness models to improve equity in health delivery fleet management. *Production and Operations Management* **23**(6) 965–977.

Neale, John, Sean Willems. 2015. The failure of practical intuition: How forward-coverage inventory targets cause the landslide effect. *Production and Operations Management* **24**(4) 535–546.

Picazo, O F, F Zhao. 2009. Zambia health sector public expenditure review: accounting for resources to improve effective service coverage. World Bank Publications.

Porteus, E L. 2002. *Foundations of Stochastic Inventory Theory*. Stanford Business Books.

Qi, Jin. 2017. Mitigating delays and unfairness in appointment systems. *Management Science* **63**(2) 566–583.

Quick, Jonathan D., James D. Rankin, Richard O. Laing, Ronald W. O'Connor, Hans V. Hogerzeil, M.N.G. Dukes, Andrew Garnett. 1997. *Managing Drug Supply*. 2nd ed. Management Sciences for Health in Collaboration with the World Health Organization, Kumarian Press, West Hartford, Connecticut.

Sieter, A. 2010. *A practical approach to pharmaceutical policy*. World Bank Publications.

The World Bank. 2016. Health services improvement project (p145335) implementation status & results report. Tech. rep. Accessed from `https://documents1.worldbank.org/curated/en/531311467315639833/pdf/ISR-Disclosable-P145335-06-30-2016-1467315625397.pdf` on November 2, 2017.

USAID | DELIVER PROJECT. 2011a. *Guidelines for Managing the Malaria Supply Chain: A Companion to the Logistics Handbook*. USAID | DELIVER PROJECT, Task Order 3, Arlington, Virginia. Accessed from http://apps.who.int/medicinedocs/documents/s21981en/s21981en.pdf on November 2, 2017.

USAID | DELIVER PROJECT. 2011b. *The Logistics Handbook: A Practical Guide for the Supply Chain Management of Health Commodities*. Second edition ed. USAID | DELIVER PROJECT, Task Order 1, Arlington, Virginia. Accessed from `http://apps.who.int/medicinedocs/documents/s20211en/s20211en.pdf` on November 2, 2017.

Vledder, Monique, Jed Friedman, Mirja Sjoblom, Thomas Brown, Prashant Yadav. 2019. Improving supply chain for essential drugs in low-income countries: Results from a large scale randomized experiment in zambia. *Health Systems and Reform* **5**(2) 158–177.

Vries, Harwin De, Luk Van Wassenhove. 2020. Do optimization models for humanitarian operations need a paradigm shift? *Production and Operations Management* **29**(1) 55–61.

Watson, Noel, Loren Bausell, Andrew Ingles, Naomi Printz. 2014. Malaria seasonality and calculating resupply: Applications of the look-ahead seasonality in Zambia, Burkina Faso and Zimbabwe. Tech. rep., USAID — DELIVER PROJECT, Task Order 7, Arlington, VA, USA.

Yadav, P, H L Tata, M Babaley. 2011. Supply chain management for essential medicines. *World Medicines*. World Health Organization.

Yadav, Prashant. 2007. Analysis of the public, private and mission sector supply chains for essential drugs in Zambia. Tech. rep., MIT-Zaragoza International Logistics Program.

You, Danzhen, Lucia Hug, Simon Ejdemyr, Priscila Idele, Daniel Hogan, Colin Mathers, Patrick Gerland, Jin Rou New, Leontine Alkema, et al. 2015. Global, regional, and national levels and trends in under-5 mortality between 1990 and 2015, with scenario-based projections to 2030: a systematic analysis by the UN Inter-agency Group for Child Mortality Estimation. *The Lancet* **386**(10010) 2275–2286.

Zambia Ministry of Health. 2015. Health sector supply chain strategy and implementation plan.

Zipkin, P. 2008. Old and new methods for lost-sales inventory systems. *Operations Research* **56** 1256–1263.