# Optimal Decentralization of Early Infant Diagnosis of HIV in Resource-Limited Settings

**Indian School of Business – Working Paper**

**2014**

**Submitted to *Manufacturing & Service Operations Management***
**manuscript**

# Optimal Decentralization of Early Infant Diagnosis of HIV in Resource-Limited Settings

Sarang Deo
Indian School of Business, Hyderabad, 500032, India, sarang_deo@isb.edu

Milind Sohoni
Indian School of Business, Hyderabad, 500032, India, milind_sohoni@isb.edu

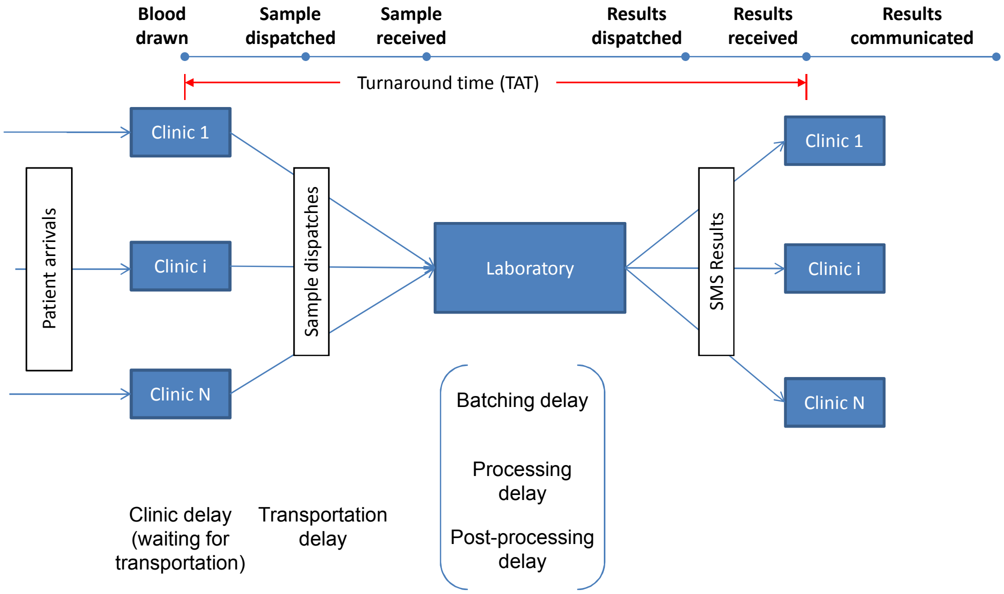
Early infant diagnosis (EID) programs in many resource-limited settings are aimed at diagnosing infants born to HIV positive mothers. Due to the complexity of the diagnostic technology, EID programs are often highly centralized with few laboratories testing blood samples from a large network of health facilities. This leads to long diagnostic delays and consequent failure of patients to collect results in a timely manner. Several point-of-care (POC) devices that provide rapid diagnosis within the health facilities are being developed to mitigate these drawbacks of centralized EID networks. We study the decision of which facilities should receive the POC device (the placement plan) using the EID program in Mozambique as a case-study. We argue that the choice of an appropriate plan is critical to maximizing the public health impact of POC devices in the presence of tight budget constraints. To formalize this argument, we develop a detailed simulation model to evaluate the impact of a placement plan. It comprises two parts: an operational model that quantifies the impact of a POC placement plan on the diagnostic delay and a behavioral part that quantifies the impact of diagnostic delay on the likelihood of result collection by infants' caregivers. We also develop an approximate version of these operational and patient behavior dynamics and embed them in an optimization model to generate candidate POC placement plans. We find that the optimization based plan can result in up to 30% more patients collecting their results compared to rules of thumb that have practical appeal. Finally, we show that the effectiveness of POC devices is much higher than other operational improvements to the EID network such as increased laboratory capacity, reduced transportation delay, and more regularized transport.

*Key words*: health access; resource-limited settings; pediatric HIV; diagnosis delay
*History*: This version: October 15, 2014

## 1. Introduction

Sub-Saharan Africa accounts for about 320,000 newly infected children per year and 18.4 million children living with HIV (WHO 2011). This corresponds to more than 90% of the global pediatric HIV disease burden. However, only 21% of these children receive treatment; a major reason for the low coverage being the lack of access to appropriate diagnostic facilities. All infants, irrespective of their HIV status, inherit HIV antibodies from their

**Figure 1** **Schematic representation of a typical EID network and associated components of delay.**



mothers during gestation. Consequently, antibody tests that are routinely used to diagnose HIV in adults yield a large number of false positives in infants thereby necessitating a more advanced diagnostic method (virologic testing) to confirm the infection (Creek et al. 2007). However, given its technical complexity, virologic testing is typically conducted in a few centralized laboratories in each country. This results in a complex sample transportation and processing network called Early Infant Diagnosis (EID) network as shown in Figure 1.

In this network, blood samples are collected from infants at remote health facilities and dispatched to a centralized laboratory through unorganized means of transport (which we refer to as transportation opportunities). Samples are then batched and processed in the laboratory to economize on the cost of reagents and labor. Finally, results are transmitted back electronically to the health facilities. The centralized structure leads to significant delay (also called turnaround time) between collection of samples and receipt of results at the health facility (Nuwagaba-Biribonwoha et al. 2010, Creek et al. 2008, Khamadi et al. 2008). Long turnaround time (TAT), can adversely impact patient health outcomes because of high mortality rate (Newell et al. 2004) and reduced likelihood of collection of results by mothers (or other caregivers) of infants (Latigo-Mugambi et al. 2013, Nuwagaba-Biribonwoha et al. 2010, Ciaranello et al. 2011, Chatterjee et al. 2011).

Several point-of-care (POC) devices that can rapidly diagnose patients at the health facility and thus mitigate these logistical problems are being developed (Yager et al. 2008, Fiscus 2010, Jani et al. 2010b, Parpia et al. 2010). These devices are designed to be cheaper and more robust to adverse environmental conditions such as heat, dust and lack of uninterrupted electricity compared to the current testing technology. However, they are also likely to be less sensitive, i.e., they are able to detect fewer HIV positive infants compared to virologic testing (Parpia et al. 2010). Because of the high investment required to buy these new devices and train laboratory technicians to use them, it is not feasible to place these devices in all health facilities in countries with limited healthcare budgets. Hence, an important operational decision in this context is designing a POC device allocation plan. In other words, which health facilities should receive the limited number of devices? To address this question, we develop a formal modeling approach and demonstrate its applicability using operational data collected from the EID network in Mozambique.

First, we construct an integrated simulation model consisting of two components to evaluate a given POC device allocation plan. The first component is a discrete event simulation model of the detailed operational dynamics of the EID network. For a given allocation of POC devices, it calculates the turnaround time of each sample as the sum of: (*i*) time spent in the clinic waiting for a transportation opportunity, (*ii*) time spent in transportation from the clinic to the laboratory, and (*iii*) time spent in the laboratory until the result is transmitted back to the clinic. The second component is a Monte Carlo simulation model based on recent empirical evidence regarding the impact of turnaround time on the patients' probability of result collection (Deo et al. 2014). We validate the simulation model using data from a part of the EID network in Mozambique for the year 2011 for a baseline scenario of a completely centralized diagnostic network.

Next, we develop a mixed integer nonlinear optimization model consisting of two parts to generate good POC device allocation plans. The first part approximates the operational dynamics to obtain an analytical characterization of the average turnaround time of samples as a function of the POC device allocation plan. The second part aggregates the behavioral dynamics of individual patients in a clinic to quantify the relationship between the average turnaround time at a clinic with the average fraction of results collected at that clinic.

Finally, we use the validated simulation model to obtain a realistic estimate of the impact of the POC device allocation plan obtained from the optimization model. We find that, for the same device, allocation plan based on the optimization model can increase the number of patients receiving their results by up to 30% in the Mozambique EID program compared to various rules of thumb currently being considered in practice.

The main contribution of our work is to combine two disparate approaches for evaluating and implementing POC devices using operations models and principles. The first approach emphasizes the clinical cost-effectiveness of the device (Shillcutt et al. 2008, Hislop et al. 2010, Rydzak and Goldie 2008, Laurence et al. 2008), which includes a trade-off between the cost of the device and its accuracy. This approach is suitable for comparing different centralized diagnostic technologies. However, it is not suited for evaluating POC devices, whose major benefit is to improve access to health care services by redesigning the health care delivery system. The second approach acknowledges programmatic constraints and espouses scaling up an new medical technology in a hierarchical manner from tertiary hospitals to primary health centers (Girosi et al. 2006, Aledort et al. 2006, Wagner et al. 2010). But it does not usually quantify the impact of various approaches to scale-up on health outcomes.

Our results underline the importance of using implementation plans as the basis for cost-effectiveness analysis of diagnostic devices that fundamentally alter the operational dynamics associated with health care delivery. In effect, our analytical framework allows viewing implementation of POC devices as an operational improvement in a diagnostic network and facilitates its comparison with other operational interventions such as improved sample transportation and increased laboratory capacity.

Our model formulation is related to the extensive literature on facility location problems with stochastic demand and congestion. See Berman and Krass (2002) and Boffey et al. (2007) for detailed reviews of this literature. A subset of papers in this literature (Parker and Srinivasan 1976, Berman and Kaplan 1987, Marianov 2003, Berman et al. 2006, Berman and Drezner 2006, Zhang et al. 2009) explicitly incorporate the impact of delay on the demand in a manner similar to our patient behavior model, e.g., see participation function in Zhang et al. (2009).

However, we cannot directly apply the models and results from this literature because of two key features of our problem context. First, from a system dynamics perspective, the drivers of delay include stochastic transport opportunities at the demand nodes and batching and congestion at the service facility. The former results in arrival stream at the lab comprising batches of random sizes while the latter leads to delay at the lab that is non-monotone function of the utilization. Second, from a patient behavior perspective, the delay primarily affects the objective function (collection of results) and not the operational dynamics (demand for testing itself). To account for and exploit these structural differences, we adopt a different solution approach compared to those adopted in the literature. Specifically, we use a linearization reformulation technique (Forrester et al. 2010) instead of meta-heuristics (e.g., Tabu Search) and complex specialized algorithms (e.g., Lagrangean relaxation) used in the earlier papers.

In the remainder of the paper, we describe our study setting in greater detail in §2 and use its key operational characteristics to formulate the POC device allocation problem in §3. We explain our solution approach in §4 and present the results in §5. Finally, we provide concluding remarks along with potential avenues for future research in §6. Appendix A contains a table of all notation employed in the paper and Appendices B through D contain background results underlying our analytical and empirical approach. Appendix E contains more details about the empirical model of patient behavior and Appendix F contains some extensions of our optimization model.

## 2. EID program in Mozambique

The structure of the early infant diagnosis program (EID) in Mozambique is similar to that in many other countries in the sub-Saharan region. Samples are collected at several hundred health facilities around the country and transported to one of the four laboratories equipped to conduct virologic testing. In the first stage of transportation, samples wait at the health facilities until an unorganized transport opportunity (a healthcare worker or community member traveling to the nearby town on other errands) materializes to transport them to respective provincial headquarters. In the second stage of transportation, a private courier company uses passenger aircrafts to transport samples from the provincial headquarters to the laboratories. At the time of sample collection, mothers are given an initial prophylaxis (protective treatment) and advised to return for a follow-up appointment after one month to collect the results and to seek further medical advice (Creek et al. 2007).

At the laboratory, samples are processed in batches to optimize on the cost of reagents and availability of personnel. All incoming samples are held until a complete batch is formed and a completed batch enters processing once the equipment becomes available after finishing the processing of the previous batch. Because of the physical constraint

**Table 1 Descriptive statistics for major model parameters based on 42 health facilities and the associated centralized laboratory included in our analysis.**

| | Mean | Std. Dev. | Min. | Max. |
|---|---|---|---|---|
| Number of samples | 157.17 | 113.14 | 34.00 | 494.00 |
| Arrival rate per day | 0.46 | 0.33 | 0.10 | 1.44 |
| Fraction of HIV+ samples (positivity rate) | 0.13 | 0.09 | 0.03 | 0.57 |
| Average frequency of dispatch per year | 23.62 | 15.22 | 6.00 | 74.00 |
| Average dispatch batch size | 6.65 | 3.48 | 1.73 | 16.50 |
| Average clinic delay in days | 11.92 | 11.08 | 2.70 | 48.24 |
| Average transportation delay in days | 3.26 | 6.29 | 0.00 | 23.33 |
| Average laboratory delay in days | 12.46 | 9.5 | 0 | 180 |

on equipment capacity, some samples that have arrived in the laboratory may not be accommodated and might have to be left out to wait for the next batch.

In 2011, over 40,000 samples were tested from across Mozambique, of which approximately 13% were HIV positive. The average turnaround time across the country was about 44 days, of which 11 days were spent in the clinic, 8 days in transit and 25 days in the laboratory. Only 32% of the results were delivered to the health facilities before the follow-up appointment at one month. For our analysis, we focus on clinics linked to one of the four central laboratories with the best operational performance among all laboratories: lowest turnaround time (32 days) and the highest fraction of results delivered within one month (58%). Consequently, our estimation of the benefits of optimizing the POC allocation decision is conservative and we expect it to be higher when applied to other parts of the EID network in the country.

The raw data used in our analysis include 7184 patient records from 95 health facilities in the provinces of Tete and Zambezia. Of these, we exclude facilities with less than 30 samples or less than two dispatches in the entire year because they are unlikely to have adequate infrastructure to support a POC device. The resulting data comprise 42 facilities with 6601 samples (∼92% of all observations). The key operational fields include *date of arrival* to the health facility, *date of dispatch* from the health facility, *date of processing* in the laboratory, and *date of result transmission* back to the health facility. We use these data to calculate the three main components of the diagnostic delay–in the health facility, in transportation, and in the laboratory. We also use these data to calculate other operational parameters of interest such as sample arrival rates, dispatch batch sizes, interval between dispatches, and laboratory batch sizes. Table 1 contains a brief summary of these parameters.

## 3. Model Formulation

We consider an EID network with one central laboratory, at a predetermined location, and $n$ clinics indexed by $i \in \{1, 2, \dots, n\}$. Infants arrive for testing at clinic $i$ at a rate $\lambda_i$. For tractability, and due to lack of data, we assume that this arrival rate does not change with the introduction of the POC device. However, our model can be extended to accommodate a constant multiplicative change in the arrival rate (increase or decrease).

We assume that a limited implementation budget permits the decision maker to procure and install POC devices at only $\hat{m}$ facilities and the remaining $n - \hat{m} = m$ facilities remain with the centralized laboratory. A binary decision variable $y_i$ indicates if the facility remains in the centralized laboratory network ($y_i = 1$) or is allocated a POC device ($y_i = 0$). We use $\mathbf{y} = \{y_1, y_2, \dots, y_n\}$ to denote the allocation vector comprising these binary variables.

Let $T_i^j$ denote the random turnaround time (TAT) for health facility $i$ that uses diagnostic system $j \in \{L, P\}$, where $L$ represents the laboratory network and $P$ represents the POC device. For the clinics in the centralized laboratory network ($y_i = 1$), TAT $\left(T_i^L\right)$ is a sum of three components: time spent in the clinic before being dispatched to the laboratory ($W_{i,c}$), sojourn time in the laboratory ($W_{i,l}$) and transportation time from the clinic to the laboratory ($W_{i,t}$). Thus,

$$T_i^L(\mathbf{y}) = W_{i,c} + W_{i,l}(\mathbf{y}) + W_{i,t}, \tag{1}$$

where we have emphasized the dependence of $W_{i,l}$ and consequently $T_i^L$ on the allocation vector $\mathbf{y}$. This is because the time spent by samples in the laboratory depends on the total sample load, which in turn, depends on the arrival rate at all clinics associated with that laboratory. For clinics with the POC device ($y_i = 0$), we assume that $T_i^P = 0$ because the results are typically available on the same day.

A fraction $p_i$ of infants arriving at clinic $i$ is infected with HIV. Let $s^j$ denote the sensitivity of the diagnostic system $j$ ($j \in \{L, P\}$), i.e., the fraction of truly HIV positive patients diagnosed correctly using that diagnostic system. In other words, $1 - s^j$ denotes the fraction of false negative results in diagnostic system $j$. Let $\Omega\left(T_i^j\right)$ denote the probability that a randomly chosen infant's caregiver will collect the result, given a random turnaround time $T_i^j$. $\Omega(\cdot)$ captures the combined effect of two factors. First, some results are not collected because the caregivers cannot make repeated costly visits to the facility to check if the results have arrived or not. Second, some results might remain uncollected because of the infant's death; mortality rate among untreated infants can be as high as 30% in the first year (Newell et al. 2004). Thus, the expected number of patients who collect their results at clinic $i$ that uses diagnostic technology $j$ is given by $\lambda_i \mathbb{E}\left[\Omega\left(T_i^j\right)\right]$, where the expectation is with respect to the distribution of the turnaround time $T_i^j$. We assume that all results in facilities with a POC device are collected because of instantaneous delivery of results, i.e., $\Omega(0) = 1$.

We formulate the objective function for the POC device placement decision as the expected number of truly HIV+ infants who receive their results. For facility $i$ that uses diagnostic technology $j$, this number is given by:

$$N_i^j = \lambda_i p_i s^j \mathbb{E}[\Omega\left(T_i^j\right)] \quad i \in \{1, 2, \ldots, n\}, \quad j \in \{L, P\}, \tag{2}$$

which can be reformulated using the binary variables $y_i$ as

$$N_i = y_i \lambda_i p_i \left(s^L \mathbb{E}\left[\Omega\left(T_i^L(\mathbf{y})\right)\right]\right) + (1 - y_i)\lambda_i p_i s^P. \tag{3}$$

Note that this objective function does not include the impact of any false positive results because, as we will describe later, both centralized laboratory and POC device have near perfect specificity in our setting.

Our discussions with the EID program managers confirmed that this measure is a reasonable proxy for health outcomes because infants cannot be initiated on treatment unless they receive their results. While some of the infants whose results are collected do not eventually initiate treatment, we do not include this loss to follow-up while evaluating the impact of POC diagnostic device because it is driven by reasons other than turnaround time.

We assume that the social planner has access to $\hat{m}(= n - m)$ POC devices, all of which need to be installed. This requirement is justified if each POC device has a net positive impact on the overall objective function. We numerically verify this assumption for our setting in §5. Then, the social planner's problem can be formulated by summing (3) over all clinics as:

$$\max_{\mathbf{y}} \sum_{i=1}^{n} y_i \lambda_i \left(B_i \mathbb{E}\left[\Omega\left(T_i^L(\mathbf{y})\right)\right] - A_i\right), \tag{4}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} y_i = m, \tag{5}$$

$$y_i \in \{0, 1\}, \tag{6}$$

where we have substituted $B_i = p_i s^L$, $A_i = p_i s^P$ and dropped the constant term $\sum_i \lambda_i p_i s^P$ since it does not impact the computation of the optimal allocation vector $\mathbf{y}$.

Typically, $s^L > s^P$, i.e., laboratory technology is more accurate in identifying the HIV positive patients than the POC device, which yields $B_i > A_i$. On the other hand, $\mathbb{E}[\Omega(\cdot)] < 1$ indicating that not everyone in the centralized laboratory system collects the results. Thus, the objective function captures the *accuracy vs. access trade-off* associated with decentralization of diagnosis.

Our formulation can accommodate the objective function of maximizing the number of infants initiating treatment if the fraction of patients that does not initiate treatment after receiving their results (say $\alpha_i$) is independent of the TAT. Our field interviews indicate that this is a reasonable assumption. In that case, the objective of maximizing the number of patients initiating treatment can be formalized exactly as in (4), with $A_i = \alpha_i p_i s^P$ and $B_i = \alpha_i p_i s^L$.

In Appendix §F, we discuss two expanded problem formulations that can be valuable when generalizing our approach to other settings. The first formulation accommodates other operational decisions regarding the design of EID network (e.g. lab-clinic linkages) in addition to POC device placement and the second formulation also includes a more detailed account of various costs and health outcomes associated with false negative and false positive cases.

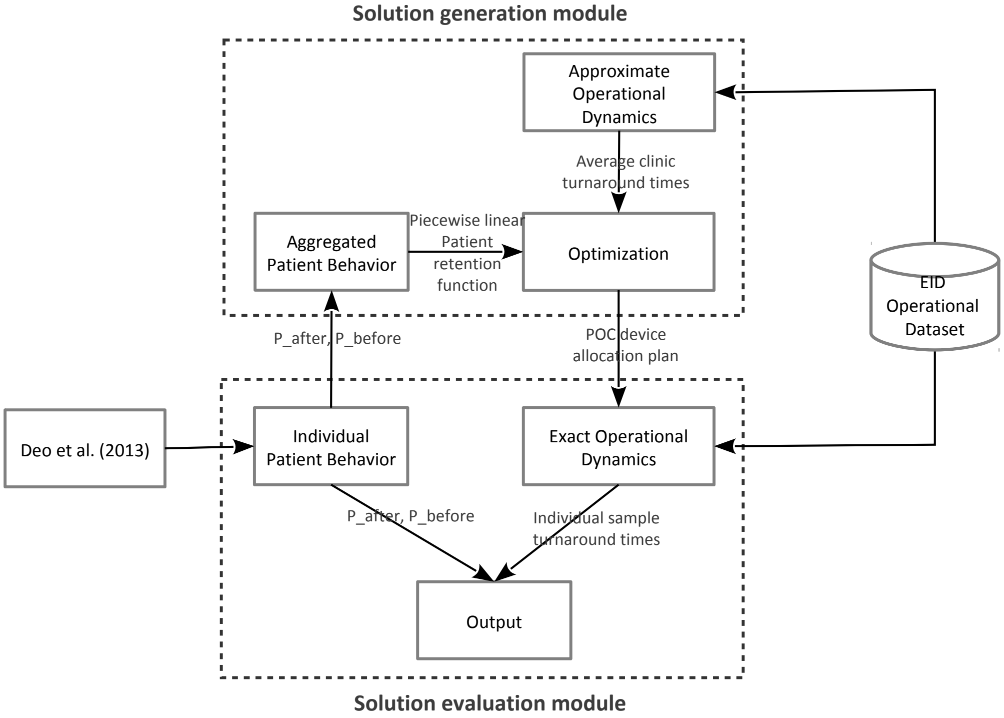
## 4. Solution Methodology

The main obstacle in solving the POC device allocation problem (4)–(6) is that the objective function is non-separable in the allocation decisions $\mathbf{y}$. Also, the functional form of the probability of result collection $\Omega(T_i^j)$ and the distribution of the turnaround time $T_i^L(\mathbf{y})$ have complex structure, which precludes an exact analytic solution approach. Consequently, we adopt a two-pronged approach of *solution generation* and *solution evaluation* as shown in Figure 2. We employ two different (but related) sets of assumptions for these two prongs that are consistent with their respective objectives: analytical tractability in solution generation and fidelity to the application in solution evaluation. We now explain these further in §4.1–§4.3.

### 4.1. Solution Evaluation: Simulation-based approach

We develop a data-driven simulation model that estimates the number of infants receiving their results for any given POC device allocation plan. Our objective here is to incorporate

**Figure 2** Schematic representation of the solution approach described in Section 4


Solution generation module
Approximate Operational Dynamics
Average clinic turnaround times
Aggregated Patient Behavior
Piecewise linear Patient retention function
Optimization
EID Operational Dataset
P_after, P_before
POC device allocation plan
Deo et al. (2013)
Individual Patient Behavior
Exact Operational Dynamics
P_after, P_before
Individual sample turnaround times
Output
Solution evaluation module

detailed operational and patient behavior dynamics, which are informed by EID program data and complemented by our field observations and interviews in Mozambique.

**4.1.1. Operational Module.** We develop a discrete event simulation model (in MATLAB®) to calculate the turnaround time of samples for a given POC device allocation plan. Key inputs of this model include the arrival process of samples and transport opportunities to the clinics, distribution of the transportation delay, and various aspects of laboratory operations (e.g., processing capacity, processing batch sizes, post-processing delay). Next, we describe how we obtain these inputs from the data of the Mozambique EID program data.

*(i)* **Clinic Operations:** In our data, an average health facility did not receive samples on roughly 70% of the days. This is significantly greater than the number of days with zero arrivals predicted by a Poisson process with an equivalent arrival rate. To account for these "excess" zeros in the sample arrival process, we use a Zero Inflated Poisson model (mixture of Bernoulli and Poisson distributions), wherein the probability mass function of the number of sample arrivals per day at clinic $i$ is given by:

$$\mathbb{P}(R_i = r) = \begin{cases} \phi_i + (1-\phi_i)\, e^{-\hat{\lambda}_i}, & r = 0 \\ (1-\phi_i)\, \frac{e^{-\hat{\lambda}_i}\hat{\lambda}_i^r}{r!}, & r = 1, 2, \ldots \end{cases} \tag{7}$$

Each day at clinic $i$, nature draws zero from a Bernoulli distribution with probability $\phi_i$, in which case the number of arrivals is zero at that clinic on that day. Similarly, nature draws one with probability $1-\phi$, in which case the number of arrivals is governed by a Poisson model with rate $\hat{\lambda}_i$. We calculate these parameters by solving the moment conditions for each clinic given by $\mathbb{E}[R_i] = \lambda_i = \hat{\lambda}_i(1-\phi_i)$ and $Var(R_i) = \hat{\lambda}_i(1-\phi_i)(1+\hat{\lambda}_i\phi_i)$ (Cameron and Trivedi 2013). In Section 4.2, we will further approximate the arrival process with a Poisson model with arrival rate $\lambda_i$ for analytical tractability.

Our field observations and interviews reveal that most health facilities do not have a dedicated transportation budget or access to vehicles specifically assigned for sample transportation. They typically rely on informal transportation opportunities such as a health care worker or a community member travelling to the nearby town on a personal errand or other clinic-related tasks. When such an opportunity realizes, all samples collected until then are dispatched. To model this process, we assume that the transportation opportunities arrive at clinic $i$ according to a Poisson process with rate $\eta_i$, independent of the arrival of samples. In our data, opportunities that do not result in a dispatch (due to unavailability of samples at the clinic) are not recorded. To overcome this censoring in the data, we use the memoryless property of the Poisson process to estimate $\eta_i$. Specifically, under these structural assumptions, the average time spent by a sample in the clinic is exactly the inter-arrival time of the transport opportunities:

$$\mathbb{E}\left[W_{i,c}\right] = \frac{1}{\eta_i}. \tag{8}$$

*(ii)* **Transportation:** For transportation delay, at each clinic, we fit a separate empirical distribution to the difference between the date of dispatch from that clinic and date of arrival at the laboratory.

*(iii)* **Laboratory dynamics:** Our data do not contain the identity of the processing batch for each sample at the laboratory. Hence, we assume that all samples with the same processing date comprise a processing batch and derive an empirical distribution of the batch size. However, this is at odds with the stated policy of running batches of fixed size. Our experiments using a fixed batch size, which is equal to the mean of the empirical distribution, produce results that are reasonably close to actual data. Hence, for brevity, we present results corresponding to the fixed batch size policy. We also observe a *post-processing delay* between the processing date and the date of result transmission

to the clinics. This can be attributed to the fact that laboratory staff members undertake multitasking and might not be available to approve and transmit all the results at the exact time instance when the processing is complete. We fit an empirical distribution to these data.

A key parameter of the laboratory operation that is not directly observable to us is its service rate or capacity. Hence, we impute it using actual data and output of the simulation model as follows. We consider the base case of a completely centralized laboratory system without any POC devices. We run the discrete event simulation model (20 replications of 2000 days each with a warm up period of 400 days) for a wide range of effective laboratory capacity. We impute the laboratory capacity as the value (0.26 batches / day) for which the simulated mean laboratory cycle time is not statistically different from that observed in the data (95% confidence interval). This imputation reflects the the effective capacity that accounts for constraints on human resources (e.g., technicians, supervisors, assistants), virologic testing equipment and communication infrastructure (e.g., computer, printers). The discrete event simulation model assumes a continuous operation (24 hours per day and 7 days per week) whereas the actual operation is based on an 8 hour shift. Thus, the imputed capacity is equivalent to an actual capacity of about 1 batch per day for a workweek of 6 days with 8 hours each. This calculation agrees well with the experience of the laboratory supervisors and the data on processing dates.

**4.1.2. Patient behavior module.** Routinely collected operational data in Mozambique do not include information on whether the results were collected by the infants' caregivers or not, which is the key outcome variable in our analysis. Hence, we use the results of a recent empirical study conducted at seven health facilities in Mozambique (Deo et al. 2014) to fit a Logit model to the probability of result collection as follows:

$$Log\left(\frac{\mathbb{P}(C_j=1)}{1-\mathbb{P}(C_j=1)}\right)=\gamma_0+\gamma_1\mathbf{1}_{\{TAT_j>T_a\}}+(CONTROLS)_j+\epsilon_j. \tag{9}$$

$C_j$ is a binary variable denoting whether the result was collected by the caregiver of infant $j$, $T_a$ is the time until the first follow-up appointment (typically one month) and $\text{TAT}_j$ is the turnaround time for the result of infant $j$. The model also includes other patient characteristics such as age and gender of the infant, and fixed effects for clinic and year. More details about the model and the results from Deo et al. (2014) are included in Table A.3 in Appendix E.

Since our primary interest is in understanding the impact of TAT on result collection, we partial out the impact of all other variables and calculate the marginal probability of result collection based on whether the TAT is greater than 30 days and less than 30 days, which we denote by $P_{after}$ and $P_{before}$ (see Figure 2), respectively. We then use these probabilities to simulate whether result of each sample in the operational dataset is collected or not depending on its turnaround time. Finally, we aggregate these outcomes to obtain the number of patients who will collect the results across all facilities, which is the main output of the simulation model.

### 4.2. Solution Generation (I): Optimization-based approach

In this section, we introduce analytical approximations for the operational dynamics ($T_i^L$) and the result collection probability $\Omega(\cdot)$ to lend more structure to the problem formulation (4)–(6) such that it can be solved using commercially available solvers to generate POC device allocation plans.

**4.2.1. Operational dynamics.** Next, we introduce approximations for the mean of each of the three components of turnaround time shown in (1).

*(i)* **Clinic operations:** As alluded to earlier, we approximate the Zero Inflated Poisson model for sample arrivals at clinic $i$ with a corresponding Poisson model with rate $\lambda_i = \hat{\lambda}_i(1-\phi)$ for analytical tractability. Similar to Section 4.1, we assume that transport opportunities arrive according to Poisson process and hence $\frac{1}{\eta_i} = \mathbb{E}[W_{i,c}]$. We further exploit this structure and apply results regarding the superposition of two Poisson processes (sample arrivals and transportation arrivals) to characterize the dispatch process in the following Lemma.

Lemma 1. *Let $X_i$ be the resulting random batch size of samples dispatched from clinic $i$, $I_i^A$ be the inter-arrival time of samples, $I_i^T$ be the inter-arrival time of transport opportunities and $I_i$ be the interval between two successive dispatches. Then, (i) $I_i = I_i^A + I_i^T$, $\mathbb{E}[I_i] = \frac{1}{\eta_i} + \frac{1}{\lambda_i}$, and $Var[I_i] = \frac{1}{\eta_i^2} + \frac{1}{\lambda_i^2}$, (ii) $\mathbb{P}(X_i = n) = \left(\frac{\lambda_i}{\lambda_i+\eta_i}\right)^{n-1}\left(\frac{\eta_i}{\lambda_i+\eta_i}\right)$, $n \geq 1$, and (iii) $\mathbb{E}[X_i] = \frac{\eta_i+\lambda_i}{\eta_i}$ and $Var[X_i] = \frac{\lambda_i(\lambda_i+\eta_i)}{\eta_i^2}$.*

Next, we build on the characterization of dispatch processes at each clinic to characterize the arrival process at the laboratory and the consequent delay therein.

*(ii)* **Laboratory Operations:** The processing of samples in the laboratory can be broken down into three main steps: (i) consolidating dispatch batches from different clinics into

larger batches, (ii) processing these batches in the virologic testing equipment, and (iii) recording, approving and transmitting the results back to the clinics. We assume that the time spent in the third step does not depend on the load and denote it by $\mathbb{E}\left[W_{i,l}^{p}\right]=\delta$.

Next, we model the average time spent in the first two steps using a $\sum_i GI^{[X_i]}/G^{(B_L,B_L)}/1$ queuing system (Bitran and Tirupati 1989, Whitt 1993, Hanschke 2006). The average time for the formation of a processing batch of size $B_L$ is given by:

$$\mathbb{E}\left[W_{i,l}^{b}\right]=\frac{B_L-1}{2\Lambda}, \tag{10}$$

where $\Lambda=\sum_{i=1}^{n}\lambda_i y_i$. The subsequent delay experienced by a processing batch is calculated as the sojourn time of a GI/G/1 queuing system using the Kingman's approximation as:

$$\mathbb{E}\left[W_{i,l}^{c}\right]\approx\frac{1}{\mu}\left(\frac{\rho}{1-\rho}\right)\frac{SCV[I^*]+SCV[S]}{2}+\frac{1}{\mu}, \tag{11}$$

where $S$ denotes the service time, $\mu=\frac{1}{\mathbb{E}[S]}$ is the service rate, and $\rho=\frac{\Lambda}{\mu B_L}$ is the effective utilization. Further, $I^*$ is the random interval between formation of successive batches of size $B_L$, and $SCV[\cdot]$ denotes the squared coefficient of variation of a random variable. Let $X$ and $I$ denote the random variables corresponding to the size and inter-arrival times of laboratory arrivals, respectively. Using Hanschke (2006), we approximate $SCV[I^*]$ by:

$$SCV[I^*]\approx\frac{\mathbb{E}[X]}{B_L}\left(SCV[X]+SCV[I]\right). \tag{12}$$

Note that $I$ is the inter-arrival time of a process that is a superposition of several renewal processes. $I$ approaches exponential distribution when the number of superposing processes is large (Albin 1982, 1984). While this result is not directly applicable in our context because the component processes comprise batched arrivals and not individual samples, we still assume $SCV[I]\approx 1$ because the number of superposing processes is sufficiently large. This considerably improves the analytical and computational tractability of our formulation. In §5.2, we present the actual values of $SCV[I]$ from the simulation model and assess the quality of this approximation. Next, we provide an exact characterization of $X$ under the assumptions made in this section.

Lemma 2. *Recall* $\Lambda(\mathbf{y})=\sum_{i=1}^{n}\lambda_i y_i$. *Define* $SCV[S]=\theta$, $f_i=\frac{1}{\mathbb{E}[I_i]}$, $F(\mathbf{y})=\sum_{i=1}^{n}f_i y_i$, *and* $\tilde{\Lambda}(\mathbf{y})=\sum_{i=1}^{n}\frac{\left(\frac{2\lambda_i^2}{\eta_i}+\lambda_i\right)+\theta\lambda_i}{2}y_i$. *Then,* $\mathbb{P}(X=X_i)=\frac{f_i y_i}{F(\mathbf{y})}$. *Consequently,* $\mathbb{E}[X]=\frac{\Lambda(\mathbf{y})}{F(\mathbf{y})}$, $\mathbb{E}[X^2]=\frac{\tilde{\Lambda}(\mathbf{y})}{F(\mathbf{y})}$.

Substituting the results from Lemma 2 in (12), then (12) in (11), and adding the three components of the laboratory cycle time, we obtain:

$$\mathbb{E}\left[W_{i,l}\left(\mathbf{y}\right)\right] \approx \frac{B_L - 1}{2\Lambda\left(\mathbf{y}\right)} + \frac{\tilde{\Lambda}\left(\mathbf{y}\right)}{\mu B_L\left(\mu B_L - \Lambda\left(\mathbf{y}\right)\right)} + \frac{1}{\mu} + \delta. \tag{13}$$

*(iii)* **Transportation Operations:** We assume that transportation delay ($W_{i,t}$) does not depend on the size of the dispatch batch but only depends on the clinic from which samples are dispatched. The mean transportation delay for health facility $i$ is given by:

$$\mathbb{E}\left[W_{i,t}\right] = \tau_i. \tag{14}$$

Substituting (8), (13) and (14) in (1) gives:

$$\mathbb{E}\left[T_i^L(\mathbf{y})\right] \approx \frac{1}{\eta_i} + \tau_i + \frac{B_L - 1}{2\Lambda\left(\mathbf{y}\right)} + \frac{\tilde{\Lambda}\left(\mathbf{y}\right)}{\mu B_L\left(\mu B_L - \Lambda\left(\mathbf{y}\right)\right)} + \frac{1}{\mu} + \delta. \tag{15}$$

Based on the recent field experience of adult CD4 testing (Jani et al. 2010a), we assume that all patients receive their results without any delay in the POC system. Thus, $\mathbb{E}\left[T_i^P\right] = 0$ .

**4.2.2. Patient retention function.** Using the probability of result collection at the level of individual patient ($\Omega\left(\cdot\right)$), described in §3, directly in the optimization problem is challenging for two reasons. First, functional forms of ($\Omega\left(\cdot\right)$) estimated in the empirical studies (Latigo-Mugambi et al. 2013, Deo et al. 2014) are complex. Second, the calculation of the expected value $\mathbb{E}\left[\Omega\left(\cdot\right)\right]$ requires the knowledge of the distribution of the turnaround time $\left(T_i^L\right)$, which is quite difficult to obtain as the analysis in §4.2.1 highlights.

We overcome these challenges by establishing an approximate relationship between average turnaround time and the average probability of result collection at the clinic level instead of an exact relationship at the patient level. This is achieved in the following steps:

- Calculate the average turnaround time for each clinic directly from observed data.
- Use the simulated probabilities of result collection from §4.1.2 and calculate the mean fraction of results collected for every clinic.
- Fit a piecewise linear curve between the average $TAT$ and the average fraction of results collected at the clinic level, using the usual least squares method.

We call this resulting piecewise linear function as the “patient retention function” and denote it by $l(\mathbb{E}\left[T_i^L\left(\mathbf{y}\right)\right])$. We reiterate that we have not merely interchanged the position of the expectation operator in the probability of an individual patient collecting the result, $\mathbb{E}\left[\Omega\left(T_i^L\right)\right]$. Rather, we have approximated individual behavior with ana ggregate behavior at the clinic level.

We represent the piecewise patient retention function by a set of linear segments $\{\hat{\beta}^p - \beta^p \mathbb{E}\left[T_i^L\right] : p = 1, 2, \ldots, P\}$ that are separated by a set of breakpoints $\{k^p : \ p = 1, 2, \ldots, P+1\}$, where $\hat{\beta}^p \geq \hat{\beta}^{p+1}$ and $\beta^p \geq \beta^{p+1}\ \forall\ p$. In other words, $l(\mathbb{E}\left[T_i^L(\mathbf{y})\right]) = \underset{p}{max}\{\hat{\beta}^p - \beta^p \mathbb{E}\left[T_i^L\right] : p = 1, 2, \ldots, P\}$. However, this characterization of the patient retention function is nonlinear. to improve computational tractability, we linearize it by introducing a set of additional binary variables, one for each segment:

$$\omega_i^p = \begin{cases} 1 \text{ if } k^{p+1} > \mathbb{E}\left[T_i^L\right] \geq k^p, \\ 0 \text{ otherwise.} \end{cases} \tag{16}$$

Then, using these variables, we can write:

$$l\left(\mathbb{E}\left[T_i^L\right]\right) = \hat{\beta} - \beta \mathbb{E}\left[T_i^L\right], \tag{17}$$

where $\hat{\beta} = \sum_{p=1}^{P} \hat{\beta}^p \omega_i^p$ and $\beta = \sum_{p=1}^{P} \beta^p \omega_i^p$ and $\omega_i^p \in \{0,1\}$ such that $\sum_{p=1}^{P} \omega_i^p = 1$. This ensures that only one of the binary variables will be nonzero for each clinic and the intercept and the slope of the line segment corresponding to that variable will determine the patient retention fraction based on (17).

**4.2.3. Linearization and reformulation.** Substituting (15) in (17) and then (17) in (4), we obtain the following reformulation of the POC allocation problem:

$$\max_{\mathbf{y}} \sum_{i=1}^{n} y_i \lambda_i \left( B_i \left( \sum_{p=1}^{P} \hat{\beta}^p \omega_i^p - \sum_{p=1}^{P} \beta^p \omega_i^p \left( \frac{1}{\eta_i} + \tau_i + \frac{B_L - 1}{2\Lambda(\mathbf{y})} + \frac{1}{\mu} + \frac{\tilde{\Lambda}(\mathbf{y})}{\mu B_L \left(\mu B_L - \Lambda(\mathbf{y})\right)} + \delta \right) \right) - A_i \right) \tag{18}$$

$$\text{s.t. } \sum_{i=1}^{n} y_i = m, \tag{19}$$

$$\sum_{p=1}^{P} \omega_i^p = 1, \tag{20}$$

$$y_i, \omega_i^p \in \{0,1\}. \tag{21}$$

Note that even after approximating the operational dynamics and aggregating individual patient behavior at the clinic level, (18–21) is still a nonlinear mixed integer program, which is not guaranteed to be solved to optimality with commercially available solvers. Hence, we linearize this optimization problem by introducing auxiliary variables and additional constraints that enforce the correct values of these variables. This formulation is shown in Appendix B. Evidently, the new formulation is more cumbersome because of several more constraints and variables. However, it can be solved efficiently using commercially available solvers and is amenable to generating operational insights, which are described in the next section.

### 4.3. Solution Generation (II): Heuristic-based approach

We discuss and analyze two rules of thumb for allocating POC devices that have significant practical and/or intuitive appeal for EID program managers. These have also been mentioned by several domain experts during our field visits and interviews.

**Largest volume heuristic (LVH).** New medical technology is often first introduced in tertiary care hospitals and then rolled out to peripheral health centers (Gerlach et al. 2008). Such an allocation scheme is practically appealing for several reasons. It allocates the fixed cost of new devices over a larger sample volume thus yielding a lower average cost per test. Moreover, hospitals are typically staffed by better trained health care workers and have better infrastructure compared to health centers. These factors facilitate a speedy roll out of the new technology. In our context, we model this heuristic as the one that allocates the POC devices to facilities with largest sample volume. Formally, define a set of facilities $\mathcal{S} = \{i : \lambda_i > \lambda_k \ \forall k \notin \mathcal{S}\}$. Then, LVH can be represented in terms of our decision variables as $y_i = 0 \ \forall i \in \mathcal{S}$.

**Minimum turnaround time heuristic (MTH).** During our field visits, we learned that the operational objective pursued by EID program managers is reduction of turnaround time of results across the diagnostic network. This is quite reasonable when the ultimate patient outcomes of interest are highly correlated with TAT. Also operational measures such as TAT are easier to track compared to actual patient outcomes. Saveh-Shemshaki et al. (2012) consider a similar objective when deciding the placement of a novel technology for diagnosing tuberculosis in the Canadian province of British Columbia.

A formal definition of this heuristic can be written in terms of the following optimization problem:

$$\min_{\mathbf{y}} \sum_{i=1}^{n} y_i \lambda_i \left( \frac{1}{\eta_i} + \tau_i + \frac{B_L - 1}{2\Lambda(\mathbf{y})} + \frac{\tilde{\Lambda}(\mathbf{y})}{\mu B_L (\mu B_L - \Lambda(\mathbf{y}))} \right) \tag{22}$$

$$\text{s.t.} \sum_{i=1}^{n} y_i = m, \tag{23}$$

$$y_i \in \{0,1\}, \tag{24}$$

where we have removed the constant terms $\frac{1}{\mu}$ and $\delta$ from the expression of turnaround time because they do not vary across clinics.

**Suboptimality of heuristics.** To better understand the relationship between the two heuristics and the optimal solution, consider a special case of the formulation, where the patient retention function consists of a single segment. Then, the objective function can be written as:

$$\max_{\mathbf{y}} \sum_{i=1}^{n} y_i \lambda_i \left( B_i \hat{\beta} - A_i \right) - \beta \sum_{i=1}^{n} y_i \lambda_i \mathbb{E}\left[ T_i^L \right].$$

Clearly, MTH optimizes the second term of the objective function and will be optimal if $B_i\hat{\beta} = A_i$, which reduces to $s^L\hat{\beta} = s^P$ but is unlikely to be optimal otherwise. Intuitively, one would expect that the performance of this heuristic is closer to the optimal solution for high $\beta$ values for a given value of POC device accuracy as the second term would then be more dominant in the objective function.

On the other hand, we can interpret LVH as the allocation that emphasizes the first term of the objective function. In fact, LVH is optimal if $\beta = 0, p_i = p \ \forall i$ and $s^P > s^L\hat{\beta}$ or $B_i > A_i$. Thus, when patients are not sensitive to delay, there is a threshold accuracy of the POC device beyond which LVH is optimal. Building on this intuition further one would expect that this threshold accuracy could be even lower when patients are more sensitive to delay, i.e., $\beta > 0$.

However, this rationale needs to be modified further to account for the network externality associated with POC device allocation. Amongst all POC device allocation plans, LVH by definition achieves the largest reduction in the sample load at the laboratory. This induces a large positive externality, i.e., significantly reduces TAT at facilities without a POC device, if congestion is a bigger driver of laboratory delay than batching. In this case, the threshold POC device accuracy would decrease in $\beta$ as expected earlier. However, if batching is a bigger driver of delay than congestion, LVH will induce a large negative externality, i.e., significantly increase TAT at facilities without a POC device. In this case, one would require higher threshold levels of POC device accuracy for the LVH to be optimal. In other words, the threshold POC device accuracy will increase in $\beta$. Further, whether

**Table 2 Parameter specifications for numerical experiments in Section 5**

| | **Simulation Model** | **Optimization Model** |
|---|---|---|
| Sample arrivals | Zero Inflated Poisson (MLE) | Poisson (MLE) |
| Transport arrivals | Poisson (MLE) | Poisson (MLE) |
| Transportation delay | Empirical distribution for each clinic | Mean for each clinic |
| Lab batch size | Empirical distribution / Fixed | Fixed |
| Lab processing time | Deterministic, imputed from data | Deterministic, imputed from data |
| Lab post-processing delay | Empirical distribution | Mean |
| Behavioral model for result collection | Logistic distribution (Eq. 9) | Patient retention function (Eq. 17) |
| POC device sensitivity | Parpia et al. (2010) | Parpia et al. (2010) |

congestion or batching effect is dominant itself depends on the operational parameters $(\mu, B_L, \{\lambda_i\})$. This intuition is formalized in the statement of Proposition A.1 in Appendix C and pictorially represented in Figure A.1.

## 5. Numerical results and policy insights

In this section, we apply the solution approach developed in §4 to the EID program in Mozambique. We assess the internal consistency between the simulation and the optimization models in §5.2, evaluate various POC device allocation plans in §5.3, study the impact of the number of POC devices in 5.4 and compare them with other operational interventions in §5.5.

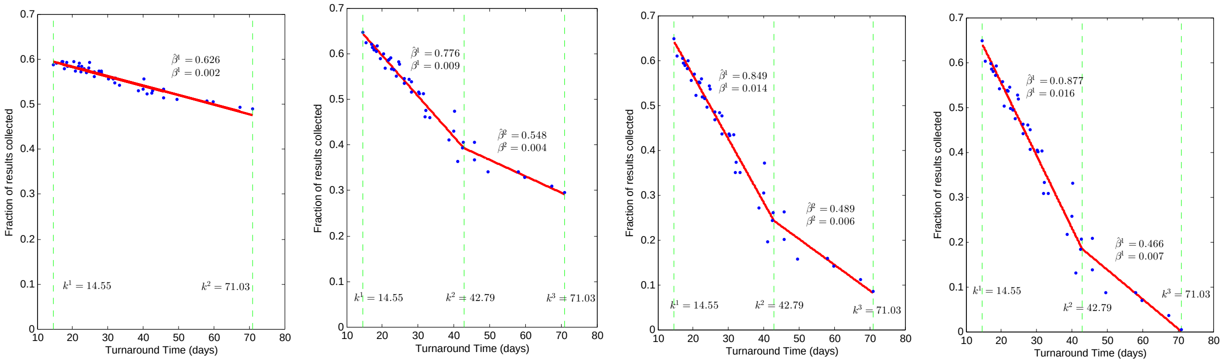
### 5.1. Parameter values

Table 2 provides a summary of methods used to estimate various parameters in the simulation and the optimization models (§4.1 and §4.2, respectively). Note that we make two different related but slightly different sets of assumptions for the simulation and optimization model that are consistent with their respective objectives: fidelity to the context for the former and analytical tractability for the latter.

For the base case, we fix the value of laboratory service time to 0.26/day based on the imputation explained in §4.1 and vary it from 0.26 to 0.32 in the sensitivity analysis. Similarly, we fix the laboratory batch size to the base value of 84 and consider the value of 168 for sensitivity analysis. For the post-processing delay in the laboratory, we use the empirical distribution in our simulation model and its mean (4 days) in the optimization model.

To estimate the patient retention function, we vary the coefficients of the patient-level Logit model in Table A.3 and generate four pairs of probabilities ($P_{after}$ and $P_{before}$) using

**Figure 3 Patient retention functions for varying levels of patient sensitivity to delay based on data from the EID program in Mozambique.**



(a): $P_{before} = 0.60$ and $P_{after} = 0.50$.

(b): $P_{before} = 0.65$ and $P_{after} = 0.294$.

(c): $P_{before} = 0.65$ and $P_{after} = 0.085$.

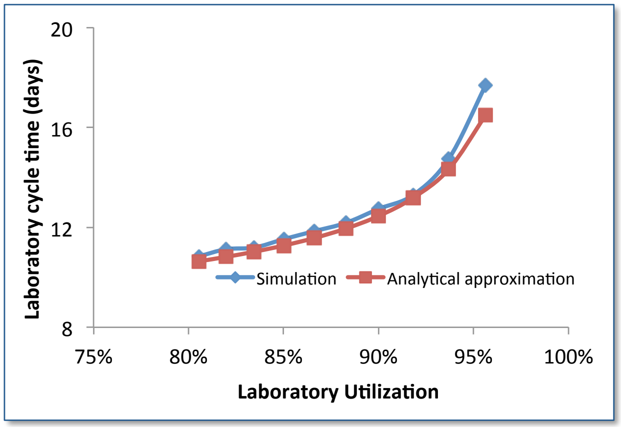
(d): $P_{before} = 0.65$ and $P_{after} = 0.01$.

the method described in §4.1 to denote varying levels of patient sensitivity to delay. For each of these pairs, we simulate patient level outcomes of whether results are collected or not, aggregate them at the clinic level and fit a piecewise linear patient retention function using the method described in Section 4.2.2. These are shown in Figures 3 (a)–(d) in the increasing order of patient sensitivity to delay. In all cases, the resulting $R^2$ was greater than 0.90. Note that the x-coordinate of observations is common across all four figures because they correspond to the actual data whereas y-coordinates are different reflecting different possible patient sensitivity to delay. We use Figure 3 (a) as the base case and use the other three for sensitivity analysis.

We use the characteristics of a POC device under development at a U.S. based university for our computational study. Based on the published results of a preclinical trial (Parpia et al. 2010), we choose the baseline value of sensitivity to 95% and specificity to 100% corresponding to 5% false negatives and no false positives. For sensitivity analysis, we vary the sensitivity of the POC device from 75% to 95%. Based on the evidence from secondary literature, we fix the sensitivity of the existing virologic testing to 100% (Creek et al. 2007).

Finally, the exact amount of budget that would be available for the implementation of POC devices in Mozambique is not yet known. Hence, for illustration, we assume that 10 POC devices are available for allocation corresponding to a coverage of roughly 25% of the facilities in the part of the network that we consider. However, none of our qualitative insights depend on this specific number. In §5.4, we vary the number of POCs to quantify the incremental benefit of a POC device.

**Figure 4** **Comparison of the laboratory cycle time for simulation model and analytical approximation for a completely centralized EID network.**



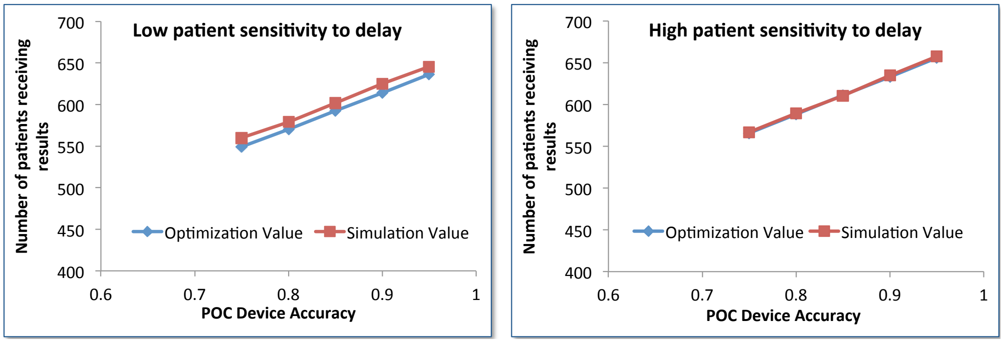
## 5.2. Internal validation of the model

Given the slightly different sets of assumptions made in the optimization and the simulation models (Table 2), we assess the internal consistency of our approach by comparing two outcomes–the laboratory cycle time and the objective value–under the two models.

First, we consider the base case of the centralized laboratory system without any POC device allocation. We generate 10 problem instances by varying the laboratory capacity from the base value of 0.26 batches/day to 0.32 batches/day corresponding to utilization of 80% to 96%. We compare the laboratory cycle time obtained from the simulation model with the value obtained from the analytical approximation in (13). The average gap between the two over these problem instances is within 3% (Figure 4). In addition, we check the appropriateness of the approximation $SCV(I) \approx 1$ used in the optimization model. Since this value is unaffected by laboratory, device or patient behavior characteristics, we run five problem instances corresponding to different number of POC devices to be allocated (0, 5, 10, 15, and 20). The SCV(I) from the simulation model for these problem instances is between 0.94 and 1.00, which indicates that the quality of approximation is very good.

Second, we consider the case of allocating 10 POC devices. We again generate 10 problem instances corresponding to five values of POC device accuracy (from 0.75 to 0.95), each combined with two patient retention functions (Figures 3 (a) and (d)) corresponding to high and low patient sensitivity to delay. For these instances, we compare the objective value obtained directly from the optimization model with that obtained by implementing the optimization solution in the simulation model, which are plotted in Figure 5. The average

**Figure 5** **Comparison of the objective value of the simulation model and the optimization model.**



gap between the two methods over the 10 problem instances is 0.94%. Based on these two sets of results, we expect that the optimal solution to the approximate optimization problem will perform very well when implemented in the (more exact) simulation model.

### 5.3. Comparative performance of different placement plans

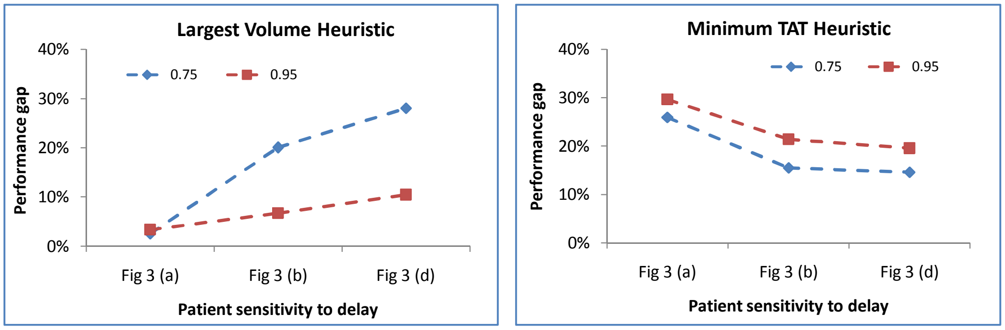
Having established the internal consistency of our solution approach, we now return to our main objective: examining the relative performance of different POC device allocation plans measured in the number of patients collecting results. First, we consider the base case with $\mu = 0.26$, $B_L = 84$, $s^P = 0.95$ and the patient retention function corresponding to Figure 3 (a). For this case, our model predicts that 489 HIV+ infants would receive results in a completely centralized laboratory system. Implementing the optimal POC allocation plan for 10 devices would result in 656 HIV+ infants receiving their results (incremental gain of 167 HIV+ infants). Similarly, LVH and MTH solutions would result in 651 and 607 HIV+ infants receiving results (incremental gain of 162 and 118), respectively. We measure the performance of each heuristic as the percentage gap between the incremental number of HIV+ infants receiving their results under that heuristic compared to that under the optimal policy. Thus, for the base case, the gap for LVH is 3% $(1-162/167)$ and that for MTH is 18% $(1-118/167)$.

Beyond the base case, we calculate the performance gaps over 20 problem instances generated from 5 different values of the device accuracy ($s^P$) and 4 different patient retention functions from Figure 3. Table 3 shows some descriptive statistics related to the performance of the two heuristics. On the whole, LVH performs better (average gap of 12%) compared to MTH (average gap of 20%). Even if the decision maker could choose the best of the two heuristics depending on the problem parameters, the average gap compared to the optimization solution would still be around 10%. Our illustrative numerical examples

**Table 3** **Performance of different heuristics compared to the optimization solution.**

| Solution | Base case | Average | Minimum | Maximum |
|---|---|---|---|---|
| LVH | 3% | 12% | 3% | 28% |
| MTH | 18% | 20% | 14% | 30% |
| Best Performing | 3% | 10% | 3% | 16% |

**Figure 6** **Impact of device accuracy and patients' sensitivity to delay on the performance of heuristics.**



highlight that even with the same number of identical POC devices, health outcomes can be up to 30% lower because of the inability of the simpler rules of thumb to formally incorporate the access vs. accuracy trade-off and the resulting network externality.

Overall, our analysis suggests that although the EID program managers cannot control the design specifications of the POC devices, the device allocation plan provides them with an additional lever to modify their effectiveness.

**5.3.1. Impact of device accuracy and patients' sensitivity to delay.** Table 3 shows that the performance of the two heuristics varies widely over the parameter range of interest. Hence, we analyze the impact of two key parameters, device accuracy and patients' sensitivity to delay, on their performance gaps. We see that the performance of MTH improves as the patients' sensitivity to delay increases. In contrast, the performance of LVH deteriorates with increase in patient sensitivity to delay because of an increase in the magnitude of network externality that is not directly captured in LVH. Moreover, the performance of LVH improves whereas that of MTH deteriorates as the device accuracy increases. Comparison of the two panels in Figure 6 indicates that LVH is a better placement plan compared to MTH under two conditions: (i) high device accuracy, (ii) low device accuracy and low patient sensitivity to delay. On the other hand, MTH seems to be better rule of thumb when the device sensitivity is low and the patient sensitivity to delay is high.

**Figure 7 Impact of laboratory capacity on the performance of the heuristics for two different batch sizes.**

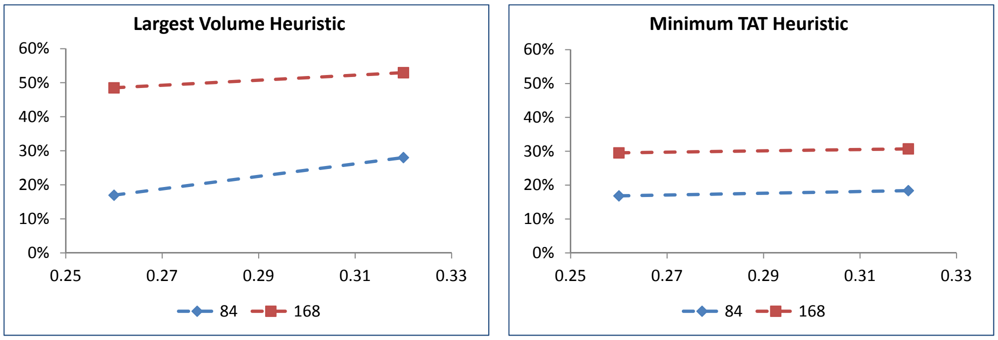
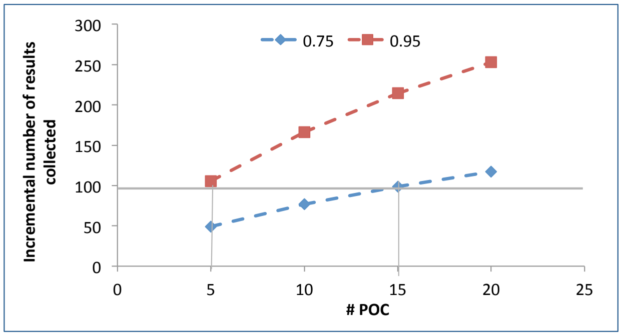


**Figure 8 Diminishing marginal benefit of POC devices.**



**5.3.2. Impact of laboratory capacity and batch size.** Figure 7 describes the performance of the two heuristics for different levels of laboratory capacity and two batch sizes (84 and 168). We see that performance of LVH worsens for larger values of laboratory capacity and batch size. Under these conditions, batching is the main driver of delay in the laboratory network and hence removing high volume clinics from the centralized laboratory network imposes a greater negative externality on other clinics. Interestingly, performance of MTH is relatively insensitive to changes in laboratory capacity and batch size because of the ability of the MTH solution to adapt these changing operational conditions. Consequently, MTH outperforms LVH when either the laboratory batch size or the laboratory capacity is relatively large.

**5.3.3. Structure of the optimization solution.** Here, we attempt to obtain some insights regarding the structure of the optimization solution and how it differs from the other heuristics. These insights can guide decision makers toward more effective placement plans even if they do not have access to the optimization model.

For the base case (95% POC sensitivity, patient retention corresponding to Figure 3 (a), $\mu = 0.26$ and $B_L = 84$), we find that the optimization solution coincides with a greedy one that allocates POCs to clinics with highest number of HIV+ infants. This is because, for these parameter values, the direct benefit accrued at the clinics with POC devices through reduced patient loss to follow-up outweighs any potential network externality imposed on other clinics without POC devices. For this case, LVH overlaps substantially with the optimization solution (8 out of 10 clinics), which explains its excellent performance for the base case. On the other hand, MTH allocates POC devices preferably to clinics that have high arrival rate and higher clinic and transportation delays irrespective of the number of HIV+ infants. Consequently, the overlap between MTH and the optimal solution is less (6 out of 10 clinics) and the performance of MTH is poor.

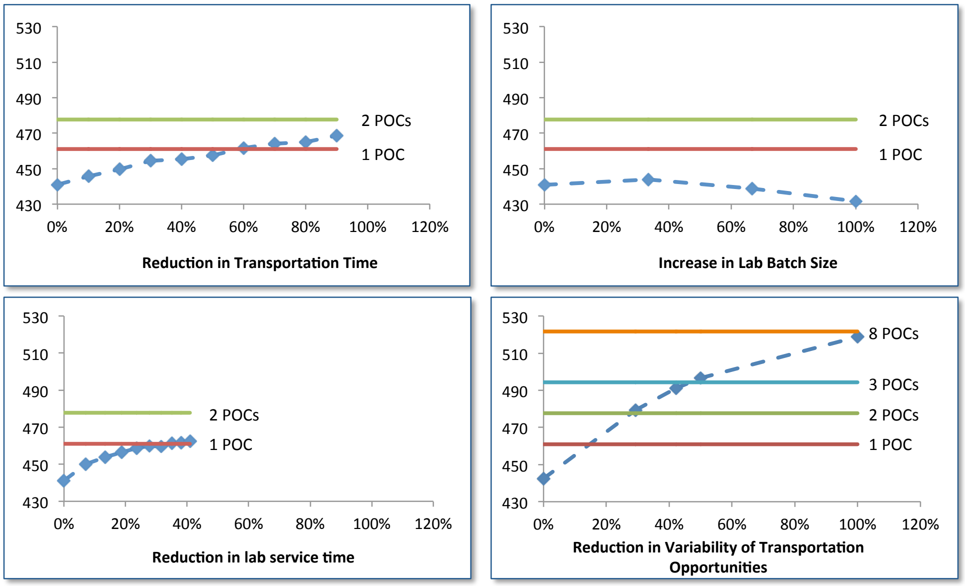
However, as patient sensitivity to delay increases (Figure 3 (a) to (d)), clinics with fewer HIV+ patients and shorter clinic and transportation delay from among those in the optimization solution are excluded. They are replaced with clinics that have even fewer HIV+ patients but substantially greater clinic and transportation delay than the excluded clinics. This helps to reduce the overall delay in the clinics that do not have POC devices and improve patient retention. This benefit more than compensates for the slight reduction in HIV+ infants having direct access to POC devices. Since the LVH solution remains unchanged across all cases (it depends only on the arrival rate of samples), the overlap between the optimization solution and LVH reduces with increased patient sensitivity to delay. For the case corresponding to Figure 3 (d), only 6 out of 10 clinics are common to both solutions.

Keeping the patient sensitivity to delay at levels of Figure 3 (d), if the POC device accuracy is reduced from 0.95 to 0.75, the optimal solution further includes clinics that have fewer HIV positive patients but higher clinic and transportation delay. As a result, for the extreme case with $s^P = 0.75$ and patient retention function as in Figure 3 (d), only half of the clinics are common between the optimal solution and LVH, which translates into poor performance of LVH.

### 5.4. Number of POC devices.

We analyze the incremental benefit of installing POC devices by solving the optimization problem for different number of POC devices (5, 10, 15, and 20) for two values of POC device accuracy (0.75, 0.95) and fix the patient retention function to that in Figure 3 (a). The objective value for these different problem instances is shown in Figure 8. It shows a reduction in the incremental benefit, which can be attributed to of the increase in

**Figure 9 Comparison of the effectiveness of POC devices with other operational interventions.**



batching time due to reduced sample load at the laboratory. This leads to the question of whether the marginal benefit of a POC device can be negative. To answer this question, we solved the optimization problem with an inequality constraint ($\sum_i y_i \geq m$) instead of the equality in (5) but found that it was binding for all problem instances corresponding to the parameter regimes of interest thus confirming that the marginal benefit of a POC device is not negative for parameter values of practical relevance. Consequently, the optimal number of POCs in this environment is simply the one that exhausts the entire budget.

We also see that the incremental benefit is lower and declines faster for a less accurate device. Thus, in order to achieve the same level of incremental benefit, the program manager would have to install a greater number of less accurate devices. For instance, Figure 8 shows that to achieve the collection of 100 additional results, one would require 5 POC devices of 95% accuracy but 15 devices of 75% accuracy. These results can help EID program managers negotiate appropriate prices with the device manufacturers depending on the accuracy of the device and/or to choose the appropriate device depending on the accuracy and price.

### 5.5. Comparison with other operational interventions

One of the advantages of our approach is that it provides a common platform to allow us to compare the effectiveness of POC devices with that of other operational improvements such as the following: (*i*) reduction in the transportation delay, (*ii*) increase in the laboratory batch size, (*iii*) reduction in the laboratory processing time, and (*iv*) reduction in the variability of inter-arrival times for transport opportunities. Our estimates of their comparative effectiveness, when combined with the cost of implementation, can help the EID program managers evaluate their relative attractiveness compared to the allocation of POC devices.

Figure 9 shows the numerical results for the case, where $s^P = 0.75$ and the patient retention function corresponds to Figure 3(b). Each horizontal line corresponds to the effectiveness of a different number of POC devices, which is identified alongside that line. Note that the first three interventions are not very effective compared to the allocation of POC devices. Even if one were to ensure almost instantaneous transportation of samples to the laboratory (100% reduction in transportation time), the resulting improvement would be less than that obtained from an optimal allocation of just two POC devices. Similarly, maximum improvement obtained by increasing the laboratory batch size is less than that obtained from one POC device. Reducing the laboratory batch size from its current value is not a feasible option as it would reduce the laboratory capacity below the sample arrival rate making the laboratory an unstable system. Further, a 40% reduction in laboratory processing time is equivalent to optimal allocation of one POC device. The second and third interventions improve the effective laboratory capacity as measured in samples processed per unit time. But the second intervention also has the negative effect of increasing the average batching time, which reduces its effectiveness compared to reducing the laboratory processing time. Reduction in the variability of transportation opportunities is the most effective of these interventions. A 50% reduction in the coefficient of variation is equivalent to installing 4 POC devices whereas making the transportation opportunities entirely deterministic (without changing the rate of their arrivals) is equivalent to optimal allocation of eight POC devices.

To capture a best-case scenario, we also analyzed a “super intervention” comprising these individual components at their respective optimal levels (100% reduction in transportaion delay, 33% reduction in laboratory batch size, 40% reduction in laboratory service time, and 100% reduction in variability of inter-arrival times of transportation opportunities) and found it to be equivalent to implementation of 12 POC devices.

# 6. Discussion

## 6.1. Summary of main insights

In this paper, we use the context of the EID program in Mozambique to analyze an important public health challenge currently faced by many developing countries: How to allocate new POC devices to health facilities in an effective manner? We develop an integrated approach to evaluate the impact of the POC device placement plan (operational decision) on the number of patient receiving their results (a proxy for health outcomes). We achieve this by combining an operational model of the EID network with a patient behavior model of result collection. Our main finding is that the same device can yield very different outcomes depending on the placement plan. This underlines the importance of using the placement plan (as against the device itself) as the unit of analysis. Our simulation model presents a unifying framework to compare the effectiveness of various operational interventions in improving health outcomes. It can thus form the backbone of a decision support tool that can be used to prioritize various interventions on the ground.

## 6.2. Limitations

Given our objective of developing and demonstrating a new approach to evaluate POC devices, we have maintained a relatively narrow scope of our model. Our objective function, number of infants receiving results, is not an exact measure of population health outcomes. It does not include the loss to follow-up of patients between result collection and treatment initiation. It also does not directly include the benefit of early initiation of treatment (Violari et al. 2008), which will further enhance the incremental benefit of POCs. It only indirectly accounts for the budget constraints by restricting the number of POC devices to be allocated but does not include a detailed account of various cost components associated with the centralized lab system or the POC devices. Next, we discuss some model extensions that can address these limitations

## 6.3. Future work

**Joint optimization with other operational decisions.** Our analysis is useful when EID managers are interested in comparing the performance of deployment of POC devices with that of adjusting other operational aspects of the centralized system such as capacity and batch size at the laboratory and sample dispatch policies at the health facilities. However, in some situations, there might be an opportunity to jointly optimize these decisions along with the POC allocation decision if appropriate data are available and policy environment is favorable. Appendix F.1 presents one such formulation, where the POC allocation decision and the clinic-lab assignment decision is jointly optimized.

The problem formulation can be further enriched by directly incorporating the health benefits and implementation costs in the objective function and the constraints respectively. See Appendix F.2 for an illustrative formulation. Here, the objective function accounts for the health impact of various diagnosis and treatment outcomes including false positives and false negatives. Similarly, the constraint explicitly captures various operational cost components associated with the centralized lab system and POC devices and imposes a budget constraint on them.

**Detailed clinical models.** It can be argued that instead of prescribing a follow-up visit after one month, this follow-up interval should depend on the likelihood of the turnaround time being less than that interval. The main trade-off involved here is that a longer duration will allow more results to be available at the clinic but could lead to further delay in result collection, which has negative clinical implications. A careful quantification of this trade-off would require a more comprehensive disease progression model to quantify the impact of diagnostic delays on health outcomes. It would also require obtaining a deeper understanding of patient behavior so as to generate estimates of the probability of result collection under modified follow-up appointment policies that are not observed in retrospective data. We discuss this issue next.

**Understanding patient behavior.** Our model demonstrates the significance of the patient retention function in linking operational decisions with health outcomes. More comprehensive data collection and more elaborate empirical analysis would help to uncover various aspects of caretaker behavior such as the interval between successive attempts to collect results and the probability of returning after each unsuccessful attempt.

**Strategic models.** Another interesting avenue for future work is investigating strategic issues related to the design of POC devices. For instance, how should the POC devices be priced and what should be their target accuracy? Answering these questions would require abstracting away from the operational details described in this paper and explicitly modeling the incentives of the manufacturer of POC devices. It would also be interesting to study the response of the incumbent firms with laboratory technologies to the entry of POC devices in the diagnostics market and the impact of this competition on social welfare.

## Acknowledgments

The authors are grateful to Jorge Quevedo, Trevor Peter, Lindy Crea, (Clinton Health Access Initiative) for several insightful discussions about EID programs in Mozambique and sub-Saharan Africa in general. The authors are indebted to Ilesh Jani (Instituto Nacional De Saude, Mozambique) for allowing the use of the EID program data for this paper and for his insights into the operational aspects of new POC devices.The authors also acknowledge the inputs of Kara Palamountain and David Kelso (Northwestern University) on the technical and operational aspects of POC devices for EID programs.

## References

S. Albin. On Poisson approximations for superposition arrival processes in queues. *Management Science*, 28 (2):126–137, 1982. ISSN 0025-1909.

S. Albin. Approximating a point process by a renewal process, II: Superposition arrival processes to queues. *Operations Research*, 32(5):1133–1162, 1984. ISSN 0030-364X.

J. Aledort, A. Ronald, S. Le Blancq, R. Ridzon, A. Landay, M. Rafael, M. Shea, J. Safrit, R. Peeling, N. Hellmann, et al. Reducing the burden of HIV/AIDS in infants: The contribution of improved diagnostics. *Nature*, 444, 2006.

O. Berman and Z. Drezner. Location of congested capacitated facilities with distance-sensitive demand. *IIE Transactions*, 38(3):213–221, 2006.

O. Berman and E. Kaplan. Facility location and capacity planning with delay-dependent demand. *International Journal of Production Research*, 25(12):1773–1780, 1987.

O. Berman and D. Krass. *Facility location problems with stochastic demands and congestion*, pages 329–371. Springer Verlag, 2002.

O. Berman, D. Krass, and J. Wang. Locating service facilities to reduce lost demand. *IIE Transactions*, 38 (11):933–946, 2006.

G. Bitran and D. Tirupati. Approximations for product departures from a single-server station with batch processing in multi-product queues. *Management science*, 35(7):851–878, 1989. ISSN 0025-1909.

B. Boffey, R. Galvao, and L. Espejo. A review of congestion models in the location of facilities with immobile servers. *European Journal of Operational Research*, 178(3):643–662, 2007.

A. C. Cameron and P. Trivedi. *Regression analysis of count data.* Cambridge University Press, 2013.

A. Chatterjee, S. Tripathi, R. Gass, N. Hamunime, S. Panha, C. Kiyaga, A. Wade, M. Barnhart, C. Luo, and R. Ekpini. Implementing services for Early Infant Diagnosis (EID) of HIV: A comparative descriptive analysis of national programs in four countries. *BMC Public Health*, 11(1):553, 2011.

A. Ciaranello, J. Park, L. Ramirez-Avila, K. Freedberg, R. Walensky, and V. Leroy. Early infant hiv-1 diagnosis programs in resource limited settings: opportunities for improved outcomes and more cost-effective interventions. *BMC Medicine*, 9(1):59, 2011.

T. Creek, G. Sherman, J. Nkengasong, L. Lu, T. Finkbeiner, M. Fowler, E. Rivadeneira, and N. Shaffer. Infant human immunodeficiency virus diagnosis in resource-limited settings: Issues, technologies, and country experiences. *American Journal of Obstetrics and Gynecology*, 197(3S):64–71, 2007.

T. Creek, A. Tanuri, M. Smith, K. Seipone, M. Smit, K. Legwaila, C. Motswere, M. Maruping, T. Nkoane, R. Ntumy, et al. Early diagnosis of human immunodeficiency virus in infants using polymerase chain reaction on dried blood spots in botswana's national program for prevention of mother-to-child transmission. *The Pediatric Infectious Disease Journal*, 27(1):22, 2008.

S. Deo, E. S. Gudo, A. Vubil, L. Crea, J. Quevedo, J. Lehe, O. Tobaiwa, L. Vojnov, T. Peter, and I. Jani. Reduced diagnostic delays is associated with increased collection of results: Findings from HIV Early Infant Diagnosis (EID) program in Mozambique. Technical report, 2014.

S. A. Fiscus. Early Infant Diagnosis: Current Tools and Prospects of Point of Care Technology. In *XVIII International AIDS Conference*, 2010.

R. Forrester, W. Adams, and P. Hadavas. Concise RLT forms of binary programs: A computational study of the quadratic knapsack problem. *Naval Research Logistics*, 57(1):1–12, 2010. ISSN 1520-6750.

J. Gerlach, D. Boyle, G. Domingo, B. Weigl, and M. Free. Increased Access to Diagnostic Tests for HIV Case Management. *AIDS2031 Working Papers Series*, 2008.

F. Girosi, S. Olmsted, E. Keeler, B. Hay, Y. Lim, J. Aledort, M. Rafael, K. Ricci, R. Boer, L. Hilborne, et al. Developing and interpreting models to improve diagnostics in developing countries. *Nature*, 444: 3, 2006.

T. Hanschke. Approximations for the mean queue length of the $GI^X/G^{(b,b)}/c$ queue. *Operations Research Letters*, 34(2):205–213, 2006. ISSN 0167-6377.

J. Hislop, Z. Quayyum, G. Flett, C. Boachie, C. Fraser, and G. Mowatt. Systematic review of the clinical effectiveness and cost-effectiveness of rapid point-of-care tests for the detection of genital chlamydia infection in women and men. *Health Technology Assessment*, 14:29, 2010.

I. V. Jani, N. Sitoe, E. Alfai, P. Chongo, J. Lehe, B. Rocha, J. Quevedo, and T. Peter. Point-Of-Care CD4 Improves Patient Retention and Time-To-Initiation for ART in Mozambique. In *XVIII International AIDS Conference*, 2010a.

I. V. Jani, N. Sitoe, E. Alfai, P. Chongo, J. Lehe, B. Rocha, J. Quevedo, and T. Peter. Evaluation of Point-Of-Care CD4 and Toxicity Monitoring for Resource-Limited ART Clinic Settings in Mozambique. In *XVIII International AIDS Conference*, 2010b.

S. Khamadi, V. Okoth, R. Lihana, J. Nabwera, J. Hungu, F. Okoth, K. Lubano, and M. Mwau. Rapid identification of infants for antiretroviral therapy in a resource poor setting: the kenya experience. *Journal of Tropical Pediatrics*, 54(6):370, 2008.

M. Latigo-Mugambi, S. Deo, A. Kekitiinwa, C. Kiyaga, and M. E. Singer. Do diagnosis delays impact receipt of test results? evidence from the hiv early infant diagnosis program in uganda. *PloS One*, 8(11):e78891, 2013.

C. Laurence, A. Gialamas, L. Yelland, T. Bubner, P. Ryan, K. Willson, B. Glastonbury, J. Gill, M. Shephard, J. Beilby, et al. A pragmatic cluster randomised controlled trial to evaluate the safety, clinical effectiveness, cost effectiveness and satisfaction with point of care testing in a general practice setting–rationale, design and baseline characteristics. *Trials*, 9(1):50, 2008. ISSN 1745-6215.

V. Marianov. Location of multiple-server congestible facilities for maximizing expected demand, when services are non-essential. *Annals of Operations Research*, 123(1):125–141, 2003.

M. Newell, H. Coovadia, M. Cortina-Borja, N. Rollins, P. Gaillard, and F. Dabis. Mortality of infected and uninfected infants born to HIV-infected mothers in Africa: A pooled analysis. *The Lancet*, 364(9441): 1236–1243, 2004.

H. Nuwagaba-Biribonwoha, B. Werq-Semo, A. Abdallah, A. Cunningham, J. Gamaliel, S. Mtunga, V. Nankabirwa, I. Malisa, L. Gonzalez, C. Massambu, et al. Introducing a multi-site program for early diagnosis of HIV infection among HIV-exposed infants in Tanzania. *BMC Pediatrics*, 10(1):44, 2010.

B. Parker and V. Srinivasan. A consumer preference approach to the planning of rural primary health-care facilities. *Operations Research*, 24(5):991–1025, 1976.

Z. Parpia, R. Elghanian, A. Nabatiyan, D. Hardie, and D. Kelso. p24 antigen rapid test for diagnosis of acute pediatric hiv infection. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 55(4):413, 2010.

C. Rydzak and S. Goldie. Cost-effectiveness of rapid point-of-care prenatal syphilis screening in sub-Saharan Africa. *Sexually Transmitted Diseases*, 35(9):775, 2008.

F. Saveh-Shemshaki, S. Shechter, P. Tang, and J. Isaac-Renton. Setting sites for faster results: Optimizing locations and capacities of new tuberculosis testing laboratories. *IIE Transactions in Health Systems Engineering*, To Appear, 2012.

S. Shillcutt, C. Morel, C. Goodman, P. Coleman, D. Bell, C. Whitty, and A. Mills. Cost-effectiveness of malaria diagnostic methods in sub-Saharan Africa in an era of combination therapy. *Bulletin of the World Health Organization*, 86:101–110, 2008.

A. Violari, M. Cotton, D. Gibb, A. Babiker, J. Steyn, S. Madhi, P. Jean-Philippe, and J. McIntyre. Early antiretroviral therapy and mortality among HIV-infected infants. *New England Journal of Medicine*, 359(21):2233, 2008.

M. J. Wagner, P. Yadav, and S. Finkelstein. The role of point-of-care CD4 diagnostics in reducing the burden of HIV/AIDS in developing countries. In *XVIII International AIDS Conference*, 2010.

W. Whitt. Approximations for the GI/G/m queue. *Production and Operations Management*, 2(2):114–161, 1993.

WHO. Global HIV-AIDS Resopnse: Epidemic Update and Health Sector Progress Towards Universal Access, 2011. Accessed 24th June 2013 at http://www.who.int/hiv/pub/progress_report2011/en/index.html.

P. Yager, G. Domingo, and J. Gerdes. Point-of-care diagnostics for global health. *Annual Reviews*, 2008.

Y. Zhang, O. Berman, and V. Verter. Incorporating congestion in preventive healthcare facility network design. *European Journal of Operational Research*, 198(3):922–935, 2009.

# Appendix A: Notation

**Table A.1 List of notations.**

| Symbol | Description |
|---|---|
| $n$ | : Number clinics, |
| $\hat{m}$ | : Number of POC devices to be allocated in the network, |
| $m = n - \hat{m}$ | : Number of clinics with the lab after allocation of POC devices, |
| $\mathbf{y} = \{y_1, \ldots, y_n\}$ | : A binary POC allocation decision vector, |
| $\lambda_i$ | : Arrival rate at clinic $i$, |
| $p_i$ | : Fraction of infants arriving into clinic $i$ infected with HIV, |
| $s^P$ | : Sensitivity of the POC diagnostic device, |
| $s^L$ | : Sensitivity of the equipment at the lab, |
| $N_i^L$ | : Number of patients receiving their results at clinic $i$ with the lab, |
| $N_i^P$ | : Number of patients receiving their results at clinic $i$ with a POC device, |
| $N_i$ | : The total number of patients receiving their results at clinic $i$ under POC allocation $\mathbf{y}$, |
| $T_i^L(\mathbf{y})$ | : Random turnaround time depending on POC allocation $\mathbf{y}$ for clinic $i$ with the : lab ($L$), |
| $T_i^P$ | : Random turnaround time for clinic $i$ with the POC device ($P$), |
| $W_{i,c}$ | : Random time in clinic before sample is dispatched, |
| $W_{i,l}(\mathbf{y})$ | : Random sojourn time in lab as function of the POC allocation decision, |
| $W_{i,l}^b$ | : Random batch build-up time in the lab, |
| $W_{i,l}^c$ | : Random congestion related delay in the lab, |
| $\delta = W_{i,l}^p$ | : Average post-processing delay in the lab, |
| $W_{i,t}$ | : Random transportation time from clinic to lab, |
| $\Omega(T_i^L)$ | : An individual's probability of collecting results at clinic $i$ when TAT is $T_i^L$, |
| $B_i = p_i s^L$ | : The first objective coefficient in (4), |
| $A_i = p_i s^P$ | : The second objective coefficient in (4), |
| $C_j$ | : Binary variable in (9) denoting whether the result was collected or not, |
| $P_{before}$ | : Marginal probability of result collection based on whether TAT is : less than the appointment date, |
| $P_{after}$ | : Marginal probability of result collection based on whether TAT is : greater than the appointment date, |
| $\eta_i$ | : Average availability of the transportation opportunity at clinic $i$, |
| $X_i$ | : Random batch size dispatched from clinic $i$, |
| $\Lambda(\mathbf{y}) = \sum_{i=1}^n \lambda_i y_i$ | : Aggregate arrival rate under POC allocation $\mathbf{y}$ in Lemma 2, |
| $\tilde{\Lambda}(\mathbf{y})$ | : Adjusted aggregate arrival rate defined in Lemma 2 under POC allocation $\mathbf{y}$, |
| $I_i^A$ | : Random inter-arrival time of samples at clinic $i$, |
| $I_i^T$ | : Random inter-arrival time of transport opportunities at clinic $i$, |
| $B_L$ | : Average batch size at the lab, |
| $\mu$ | : Average service rate at the lab, |
| $S$ | : Average service time at the lab, |
| $\rho$ | : Effective utilization of the lab equipment, |
| $l(\mathbb{E}[T_i^L(\mathbf{y})])$ | : The aggregate patient retention function (17) based on average TAT : under POC allocation $\mathbf{y}$, |
| $k^p$ | : The $p^{\text{th}}$ break-point of the piece-wise linear approximation of $l(\mathbb{E}[T_i^L(\mathbf{y})])$ in (17), |
| $\hat{\beta}^p$ | : Intercept of the $p^{\text{th}}$ segment of the piece-wise linear approximation of : $l(\mathbb{E}[T_i^L(\mathbf{y})])$ in (17), |
| $\beta^p$ | : Slope of the $p^{\text{th}}$ segment of the piece-wise linear approximation of : $l(\mathbb{E}[T_i^L(\mathbf{y})])$ in (17). |

## Appendix B: Mixed Integer Formulation

First, for the operational component, we define additional variables:

$$x_i^b = \frac{(B_L - 1)\, y_i}{2\Lambda(\mathbf{y})}, \tag{A.1}$$

$$x_i^c = \frac{\tilde{\Lambda}(\mathbf{y})}{\mu B_L \left(\mu B_L - \Lambda(\mathbf{y})\right)}, \tag{A.2}$$

where $x_i^b$ is the average batching delay and $x_i^c$ is the average congestion delay experienced by the samples of clinic $i$.

Further, to linearize (A.1) and (A.2), we introduce additional variables $v_{ik}^b = x_i^b\ y_k$, $v_{ik}^c = x_i^c\ y_k$ and $t_{ik} = y_i y_k$ and include the following equality constraints:

$$(B_L - 1)\, y_i - 2\sum_{k=1}^{n} \lambda_k v_{ik}^b = 0, \tag{A.3}$$

$$(\mu B_L)^2\, x_i^c - \mu B_L \sum_{k=1}^{n} \lambda_k v_{ik}^c - \frac{1}{2}\sum_{k=1}^{n} \tilde{\lambda}_k t_{ik} = 0, \tag{A.4}$$

where $\tilde{\lambda}_k = \frac{2\lambda_k^2}{\eta_k} + \lambda_k$. We introduce additional set of linear constraints to ensure that the correct values of composite variables $v_{ik}^b, v_{ik}^c$, and $t_{ik}$ are enforced. These are shown in the detailed formulation (CMBL) in Appendix B.

Second, for the patient retention function, to enforce that $\omega_i^p = 1$ if and only if $k^p \le \mathbb{E}[T_i^L] < k^{p+1}$, we define:

$$\omega_i^p = \underline{\omega}_i^p\ \overline{\omega}_i^p, \tag{A.5}$$

where

$$\underline{\omega}_i^p = \begin{cases} 1 \text{ if } \mathbb{E}[T_i^L] \ge k^p,\ p > 1, \\ 1 \text{ if } p = 1\ \forall\ \mathbb{E}[T_i^L], \\ 0 \text{ otherwise,} \end{cases} \tag{A.6}$$

$$\overline{\omega}_i^p = \begin{cases} 1 \text{ if } \mathbb{E}[T_i^L] \le k^p,\ p < P+1, \\ 1 \text{ if } p = P+1\ \forall\ \mathbb{E}[T_i^L], \\ 0 \text{ otherwise.} \end{cases} \tag{A.7}$$

We include additional constraints to enforce the right values of $\underline{\omega}_i^p$ and $\overline{\omega}_i^p$ as shown in (A.6) and (A.7), which are shown in formulation CMBL in Appendix B. Figure 3 shows the values of these variables for an illustrative example with three linear segments and four breakpoints (including the endpoints). Note that $k^2 < \mathbb{E}[T_i^L] < k^3$ and consequently $\omega_i^p = \underline{\omega}_i^2 = \overline{\omega}_i^2 = 1$.

Finally, we define additional composite variables $\xi_{ip} = y_i \omega_i^p$, $m_{ip}^b = \omega_i^p x_i^b$, and $m_{ip}^c = \omega_i^p x_i^c$ to completely linearize the objective function. We develop additional linear constraints to enforce appropriate values of these variables depending on which linear segment $p$ is chosen (i.e., $\omega_i^p = 1$) and the values of $y_i$, $x_i^b$, and $x_i^c$, respectively. Below, we formally state the equivalence between the linearized reformulation and the original formulation.

To simplify notation, for each clinic $i$, we define the following:

$$\tilde{\lambda}_i = \frac{2\lambda_i^2}{\eta_i} + \lambda_i. \tag{A.8}$$

Additionally, for the entire health network, recollect the following definitions:

$$\Lambda(\mathbf{y}) = \sum_{i=1}^{n} \lambda_i y_i, \tag{A.9}$$

$$\tilde{\Lambda}(\mathbf{y}) = \sum_{i=1}^{n} \tilde{\lambda}_i y_i. \tag{A.10}$$

The linear mixed integer formulation of (18)–(21) can now be written as follows:

$$\begin{aligned}\text{CMBL:}\ \max & \sum_{i=1}^{n} B_i \lambda_i \left[ \sum_{p=1}^{P} \left( \hat{\beta}^p - \left( \frac{1}{\eta_i} + \tau_i + \frac{1}{\mu} + \delta \right) \beta^p \right) \xi_{ip} - \sum_{p=1}^{P} \beta^p \left( m_{ip}^b + m_{ip}^c \right) \right] \\ & - \sum_{i=1}^{n} A_i \lambda_i y_i, \end{aligned} \tag{A.11}$$

s.t.

$$\text{Clinics with the lab: } \sum_{i=1}^{n} y_i = m, \tag{A.12}$$

$$\text{Piecewise linear constraints:} \left\{ \begin{array}{l} \sum_{p=1}^{P} \omega_i^p = y_i \quad \forall\ i, \\ \omega_i^p \ge \underline{\omega}_i^p + \overline{\omega}_i^p - 1 \quad \forall\ p,\ i, \\ \omega_i^p \le \underline{\omega}_i^p \quad \forall\ p,\ i, \\ \omega_i^p \le \overline{\omega}_i^p \quad \forall\ p,\ i, \\ t_i^L - k^p \le M \underline{\omega}_i^p \quad \forall\ p,\ i, \\ k^p - t_i^L \le M \left( 1 - \underline{\omega}_i^p \right) \quad \forall\ p,\ i, \\ t_i^L - k^{p+1} \le M \left( 1 - \overline{\omega}_i^p \right) \quad \forall\ p,\ i, \\ k^{p+1} - t_i^L \le M \overline{\omega}_i^p \quad \forall\ p,\ i, \end{array} \right. \tag{A.13}$$

$$\text{Turn around time: } t_i^L = x_i^b + x_i^c + y_i \left( \frac{1}{\eta_i} + \tau_i + \delta + \frac{1}{\mu} \right) \quad \forall\ i, \tag{A.14}$$

$$\text{Modeling } m_{ip} = \omega_i^p x_i^u,\ u \in \{b, c\}\text{:} \left\{ \begin{array}{l} m_{ip}^u \le x_i^v \quad \forall\ i, p \quad u \in \{b, c\}, \\ m_{ip}^u \le M \omega_i^p \quad \forall\ i, p \quad u \in \{b, c\}, \\ m_{ip}^u \ge M \left( \xi_{ip} - 1 \right) + x_i^u \quad \forall\ i, p \quad u \in \{b, c\}, \end{array} \right. \tag{A.15}$$

$$\text{Modeling } \xi_{ip} = \omega_i^p y_i\text{:} \left\{ \begin{array}{l} \xi_{ip} \le y_i \quad \forall\ i, p, \\ \xi_{ip} \le \omega_i^p \quad \forall\ i, p \\ \xi_{ip} \ge y_i + \omega_i^p - 1 \quad \forall\ i, p, \end{array} \right. \tag{A.16}$$

$$\text{Lab batching:} \qquad (B_L - 1)\, y_i - 2\sum_{k=1}^{n} \lambda_k v_{ik}^b = 0, \tag{A.17}$$

$$\text{Lab congestion:} \qquad (\mu B_L)^2\, x_i^c - \mu B_L \sum_{k=1}^{n} \lambda_k v_{ik}^c - \frac{1}{2}\sum_{k=1}^{n} \tilde{\lambda}_k t_{ik} = 0, \tag{A.18}$$

$$\text{If } y_i = 0 \text{ then } x_i^u = 0\text{: } x_i^u \le M y_i \quad \forall\ i, \quad u \in \{b, c\}, \tag{A.19}$$

$$\text{Modeling } v_{ik}^u = x_i^u y_k,\ u \in \{b,c\} \begin{cases} v_{ik}^u = v_{ki}^u \quad \forall\ i,k, \quad u \in \{b,c\}, \\ v_{ik}^u \le x_i^u \quad \forall\ i,k, \quad u \in \{b,c\}, \\ v_{ik}^u \le M y_k \quad \forall\ i,k, \quad u \in \{b,c\}, \\ x_i^u - v_{ik}^u \le M(1 - t_{ik}) \quad \forall\ i,k, \quad u \in \{b,c\}, \end{cases} \tag{A.20}$$

$$\text{Modeling } t_{ik} = y_i y_k\text{:} \begin{cases} t_{ik} \le y_i \quad \forall\ i,k, \\ t_{ik} \ge y_i + y_k - 1 \quad \forall\ i,k, \\ t_{ik} = t_{ki} \quad \forall\ i,k, \end{cases} \tag{A.21}$$

$$x_i^u, v_{ik}^u, m_{ip}^u, \xi_{ip} \ge 0 \quad \forall\ i,k,p \quad u \in \{b,c\}, \tag{A.22}$$

$$t_{ik} \in \{0,1\} \quad \forall\ i,k, \tag{A.23}$$

$$y_i \in \{0,1\} \quad \forall\ i, \tag{A.24}$$

$$\omega_i^p, \overline{\omega}_i^p, \underline{\omega}_i^p \in \{0,1\} \quad \forall\ i,p. \tag{A.25}$$

THEOREM A.1. *The linear mixed integer program (CMBL) in* (A.11)–(A.25) *is equivalent to the nonlinear mixed integer program in* (18)–(21).

## Appendix C: Optimality of LVH

PROPOSITION A.1. *Consider the optimization problem in* (18)–(21) *with* $P = 1$ *and* $p_i = p$, $\frac{\lambda_i}{\eta_i} = \theta\ \forall\ i$. *Let* $\underline{\lambda} = \min_i\{\lambda_i\}$ *and* $\overline{\lambda} = \max_i\{\lambda\}$. *Define* $\overline{s}^P = \hat{\beta} - \frac{\beta B\theta m}{\mu B_L}\left(\frac{2\underline{\lambda}\mu B_L - m\overline{\lambda}^2}{\left(\mu B_L - m\overline{\lambda}\right)^2}\right)$ *and* $\underline{s}^P = \hat{\beta} - \frac{\beta B\theta m}{\mu B_L}\left(\frac{2\overline{\lambda}\mu B_L - m\underline{\lambda}^2}{(\mu B_L - m\underline{\lambda})^2}\right)$.

*(i) The largest volume heuristic is optimal if* $s^P > \overline{s}^P(\beta)$ *and is not optimal if* $s^P < \underline{s}^P(\beta)$.

*(ii)* $\overline{s}^P(\beta)$ *is increasing in* $\beta$ *if and only if* $\underline{\lambda}/\overline{\lambda} > \frac{m\overline{\lambda}}{2\mu B_L}$

*(iii)* $\overline{s}^P(\beta) > \underline{s}^P(\beta)$

Figure A.1 provides a pictorial representation of the proposition. Thus, for a given patient sensitivity to delay and operational parameters, LVH is optimal if the POC device accuracy is high enough and is not optimal if the accuracy is low enough. This rule can be used to check whether a given POC device should be implemented using LVH allocation rule or not.

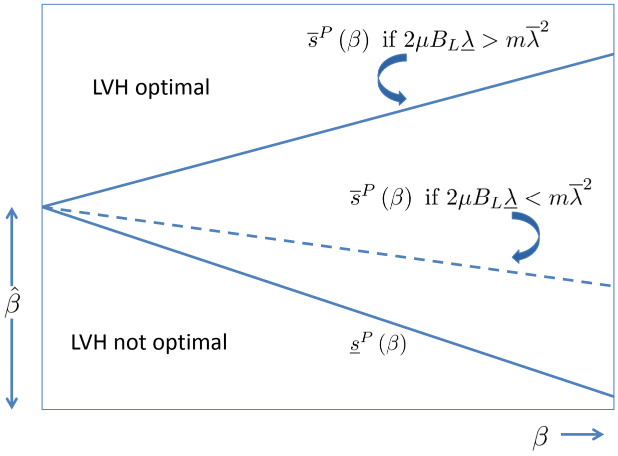

$\bar{s}^P(\beta)$ if $2\mu B_L\underline{\lambda} > m\overline{\lambda}^2$
LVH optimal
$\bar{s}^P(\beta)$ if $2\mu B_L\underline{\lambda} < m\overline{\lambda}^2$
$\hat{\beta}$
LVH not optimal
$\underline{s}^P(\beta)$
$\beta$

**Figure A.1 Regions of optimality for LVH.**

## Appendix D: Proofs of theoretical results

**Proof of Lemma 1** (i) Set $t = 0$ at the time of a dispatch. For a transport opportunity to become a dispatch, one needs at least one sample arrival. The time until the first sample arrival is given by the inter-arrival time $I_A$. The time until the next transport opportunity is again $I_T$ because of the memoryless nature of the process. Thus, the time between two dispatches is $I_A + I_T$. The mean and the variance results follow directly because of the independence of the two random variables.

(ii) Let $X_i^A(t)$ be the number of samples arriving at clinic $i$ in time t. Let $X_i^T$ denote the number of samples arriving after a dispatch and before the next transport opportunity. Since both samples and transport opportunities arrive according to Poisson processes,

$$\mathbb{P}\left(X_i^T = n\right) = \left(\frac{\lambda_i}{\lambda_i + \eta_i}\right)^n \left(\frac{\eta_i}{\lambda_i + \eta_i}\right), \; n \geq 0 \tag{A.26}$$

Since a transport opportunity will actually become a dispatch only if there are one or more sample arrivals before the arrival of the transportation opportunity,

$$\begin{aligned}
\mathbb{P}\left(X_i = n\right) &= \mathbb{P}\left(X_i^T = n | X_i^T > 0\right) &\text{(A.27)}\\
&= \frac{\mathbb{P}\left(X_i^T = n\right)}{\mathbb{P}\left(X_i^T > 0\right)} &\text{(A.28)}\\
&= \frac{\mathbb{P}\left(X_i^T = n\right)}{1 - \mathbb{P}\left(X_i^T = 0\right)} &\text{(A.29)}\\
&= \left(\frac{\lambda_i}{\lambda_i + \eta_i}\right)^{n-1} \left(\frac{\eta_i}{\lambda_i + \eta_i}\right) &\text{(A.30)}
\end{aligned}$$

■

**Proof of Lemma 2** From Lemma 1, we know that the mean time interval between dispatches is $f_i = \frac{1}{\eta_i} + \frac{1}{\lambda_i} = \frac{\eta_i \lambda_i}{\eta_i + \lambda_i}$. Then, it is evident that $\mathbb{P}(X = X_i) = \frac{f_i y_i}{\sum_k f_k y_k}$. From the result on mixture of distributions,

$$\mathbb{E}[X] = \sum_i \frac{f_i y_i}{\sum_k f_k y_k} \mathbb{E}[X_i] \tag{A.31}$$

$$\mathbb{E}[X^2] = \sum_i \frac{f_i y_i}{\sum_k f_k y_k} \left(\mathbb{E}[X_i]^2 + Var[X_i]\right) \tag{A.32}$$

Substituting $\mathbb{E}[X_i]$ and $f_i$ from Lemma 1, we obtain:

$$\begin{align} \mathbb{E}[X] &= \frac{\sum_i \frac{\eta_i+\lambda_i}{\eta_i} \frac{\eta_i\lambda_i}{\eta_i+\lambda_i} y_i}{\sum_k f_k y_k} \tag{A.33} \\ &= \frac{\sum_k \lambda_k y_k}{\sum_k f_k y_k} \tag{A.34} \end{align}$$

Similarly, substituting $Var[X_i]$ and $f_i$ from Lemma 1, we get:

$$\begin{align} \mathbb{E}[X^2] &= \frac{\sum_i \frac{\eta_i\lambda_i}{\eta_i+\lambda_i} \left(\left(\frac{\eta_i+\lambda_i}{\eta_i}\right)^2 + \frac{\lambda_i(\lambda_i+\eta_i)}{\eta_i^2}\right) y_i}{\sum_k f_k y_k} \tag{A.35} \\ &= \frac{\sum_i \left(\frac{\lambda_i(\lambda_i+\eta_i)}{\eta_i} + \frac{\lambda_i^2}{\eta_i}\right) y_i}{\sum_k f_k y_k} \tag{A.36} \\ &= \frac{\sum_i \left(\lambda_i + 2\frac{\lambda_i^2}{\eta_i}\right) y_i}{\sum_k f_k y_k} \tag{A.37} \end{align}$$

■

**Proof of Theorem A.1:** To show equivalence, we first show that any feasible solution to the original problem has a unique feasible solution under the linear reformulation.

To do so, first consider a feasible solution to the original problem, which comprises of the allocation vector $\mathbf{y}$ and resultant $t_i^L \; \forall \; i$ as given by (15). Now, using $\mathbf{y}$ define two sets $Y_1 = \{i : y_i = 1\}$ and $Y_0 = \{i : y_i = 0\}$. Thus, from equations (A.1) and (A.2) we have

$$x_i^b = \begin{cases} \frac{(B_L - 1)}{2\Lambda(\mathbf{y})}; & i \in Y_1, \\ 0; & i \in Y_0, \end{cases} \tag{A.38}$$

$$x_i^c = \begin{cases} \frac{\left(\frac{1}{2}\tilde{\Lambda}(\mathbf{y})\right)}{\mu B_L(\mu B_L - \Lambda(\mathbf{y}))}; & i \in Y_1, \\ 0; & i \in Y_0. \end{cases} \tag{A.39}$$

First, we verify that for any given value of $t_i^L(\mathbf{y})$, constraint block (A.13) selects the appropriate linear segment by enforcing the correct values of $\overline{\omega}_i^p$, $\underline{\omega}_i^p$ and consequently $\omega_i^p$. This is because the constraints guarantee that $\underline{\omega}_i^p = 1$ when $k^p \leq t_i^L$ and 0 otherwise. Similarly, the constraints also ensure that $\overline{\omega}_i^p = 1$ when $k^p \geq t_i^L$ and 0 otherwise. Finally, the constraints ensure that $\omega_i^p = 1$ if and only if $\overline{\omega}_i^p = \underline{\omega}_i^p = 1$, i.e., when $t_i^L \in [k^p, k^{p+1})$. Consequently, the values of $\beta_i$ (slope) and $\hat{\beta}_i$ (intercept) are set appropriately (for the corresponding segment $p$) through these constraints. Second, since $\omega_i^p$ takes a value of 1 for the appropriate segment $p$ and 0 for all other segments, constraint block (A.16) sets the value of $\xi_{ip}$ to 1 for the corresponding segment $p$ if and only if $y_i$ and $\omega_i^p$ are both equal to 1. Third, constraint block (A.15) ensures that $m_{ip}^b$ and $m_{ip}^c$ are set to $x_i^b$ and $x_i^c$, respectively for the appropriate segment $p$ for which $\omega_i^p$ takes on a value of 1.

Now consider various combinations of values taken by pairs of variables $y_i$ and $y_k$. Further, suppose $\omega_i^{\tilde{p}} = 1$ for some $\tilde{p}$ (corresponding to every clinic $i$) due to constraint block (A.13). Then, it can be verified that the following table forms a unique feasible solution for all the remaining variables in the linear integer reformulation, i.e., these are the only values that satisfy all the constraints.

**Table A.2 Unique solutions for possible values of $y_i$ and $y_k$.**

| | $y_i$ | $y_k$ | $v_{ik}^b$ | $v_{ik}^c$ | $t_{ik}$ | $\xi_{i\tilde{p}}$ | $m_{i\tilde{p}}^b$ | $m_{i\tilde{p}}^c$ |
|---|---|---|---|---|---|---|---|---|
| Constraint set → | – | – | (A.20) | (A.20) | (A.21) | (A.16) | (A.15) | (A.15) |
| | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 1 | 0 | 0 | 0 | 0 | 1 | $\frac{(B_L-1)}{2\Lambda(\mathbf{y})}$ | $\frac{\left(\frac{1}{2}\tilde{\Lambda}(\mathbf{y})\right)}{\mu B_L(\mu B_L-\Lambda(\mathbf{y}))}$ |
| | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 1 | 1 | $\frac{(B_L-1)}{2\Lambda(\mathbf{y})}$ | $\frac{\left(\frac{1}{2}\tilde{\Lambda}(\mathbf{y})\right)}{\mu B_L(\mu B_L-\Lambda(\mathbf{y}))}$ | 1 | 1 | $\frac{(B_L-1)}{2\Lambda(\mathbf{y})}$ | $\frac{\left(\frac{1}{2}\tilde{\Lambda}(\mathbf{y})\right)}{\mu B_L(\mu B_L-\Lambda(\mathbf{y}))}$ |

Thus, the value of the objective function (A.11) is computed correctly in the linear reformulation.

To show that any feasible solution to the linear reformulation is a feasible solution to the original problem is trivial because the linear reformulation subsumes the constraints of the original formulation and only adds new ones. In essence, we have increased the dimensionality of the problem to model the non-linear objective function, without affecting the feasible space of $\mathbf{y}$. ■

**Proof of Proposition A.1:** Consider $S^* = \{i : y_i = 1, \lambda_i < \lambda_k \ \forall \ k \notin S^*\}$. In other words, $S^*$ is that set of clinics with the laboratory whose arrival rate is less than the arrival rate of all clinics who are with the POC device. Thus, this represents the LVH solution. Consider another solution $S^{**} = \{S^* \backslash j\} \cup k$, where $k \in (S^*)^C$. Denote $\overline{S} = S^* \cap S^{**}$. Further, define $A_i = p_i \left(s^L \hat{\beta} - s^P\right)$ and $B_i = p_i s^L$. Then, the objective value corresponding to the solution $S^*$ is given by:

$$F(S^*) = \sum_{i \in S^*} \lambda_i A_i - \beta \sum_{i \in S^*} \lambda_i B_i \left( \frac{B_L - 1}{2 \sum_{l \in S^*}} + \frac{\sum_{l \in S^*} \tilde{\lambda}_l}{\mu B_L \left( \mu B_L - \sum_{l \in S^*} \lambda_l \right)} \right) \tag{A.40}$$

Note that $p_i = p \ \forall \ i \Rightarrow A_i = A, B_i = B \ \forall \ i$ and that $\frac{\lambda_i}{\eta_i} = \kappa \ \forall \ i$. Also, define $\overline{\Lambda} = \sum_{i \in \overline{S}} \lambda_i, \overline{\overline{\Lambda}} = \mu B_L - \overline{\Lambda}, \Lambda^* = \sum_{i \in S^*} \lambda_i$. Substituting these in A.40, we obtain:

$$F(S^*) = A\overline{\Lambda} + A\lambda_j - \beta B \left[ \frac{\Lambda \left( B_L - 1 \right)}{2 \left( \Lambda + \lambda_j \right)} \right] + \frac{\kappa \left( \overline{\Lambda} + \lambda_j \right)}{\mu B_L \left( \overline{\overline{\Lambda}} - \lambda_j \right)} \tag{A.41}$$

Writing a similar expression for $F(S^{**})$, calculating the difference and simplifying, we obtain:

$$F(S^{*})-F(S^{**})=\left(A-\frac{\beta\kappa B}{\mu B_L}\left(\frac{\overline{\Lambda}^2+2\overline{\Lambda}\,\overline{\overline{\Lambda}}+\overline{\overline{\Lambda}}\left(\lambda_j+\lambda_k\right)-\lambda_j\lambda_k}{\left(\overline{\overline{\Lambda}}-\lambda_j\right)\left(\overline{\overline{\Lambda}}-\lambda_k\right)}\right)\right)\left(\lambda_j-\lambda_k\right) \tag{A.42}$$

Since $(\lambda_j-\lambda_k)<0$ by construction, for LVH to be optimal, we need $F(S^{*})-F(S^{**})>0$, which is equivalent to:

$$A\leq\frac{\beta\kappa B}{\mu B_L}\left(\frac{\overline{\Lambda}^2+2\overline{\Lambda}\,\overline{\overline{\Lambda}}+\overline{\overline{\Lambda}}\left(\lambda_j+\lambda_k\right)-\lambda_j\lambda_k}{\left(\overline{\overline{\Lambda}}-\lambda_j\right)\left(\overline{\overline{\Lambda}}-\lambda_k\right)}\right)\ \forall\ j,k \tag{A.43}$$

Note that $\overline{\overline{\Lambda}}+\overline{\Lambda}=\mu B_L$. Also the definitions of $\underline{\lambda}$ and $\overline{\lambda}$ imply that $\left(\overline{\overline{\Lambda}}-\lambda_k\right)\leq\left(\mu B_L-m\overline{\lambda}\right)$, $\overline{\Lambda}\geq(m-1)\underline{\lambda}$, $\overline{\overline{\Lambda}}\geq\mu B_L-(m-1)\overline{\lambda}$. Using these relationships, we obtain a lower bound on the right hand side of A.43 and the following sufficient condition:

$$A\leq\frac{\beta\kappa B}{\mu B_L}\frac{\left(2m\underline{\lambda}\mu B_L-m^2\overline{\lambda}^2\right)}{\left(\mu B_L-m\overline{\lambda}\right)} \tag{A.44}$$

Substituting the definitions of $A$ and $B$ yields result $(i)$.

The proof of result $(ii)$ proceeds similarly by noting that for LVH to not be optimal, we need $F(S^{*})-F(S^{**})>0$, which is equivalent to:

$$A\geq\frac{\beta\kappa B}{\mu B_L}\left(\frac{\overline{\Lambda}^2+2\overline{\Lambda}\,\overline{\overline{\Lambda}}+\overline{\overline{\Lambda}}\left(\lambda_j+\lambda_k\right)-\lambda_j\lambda_k}{\left(\overline{\overline{\Lambda}}-\lambda_j\right)\left(\overline{\overline{\Lambda}}-\lambda_k\right)}\right)\ \forall\ j,k \tag{A.45}$$

Now, use $\left(\overline{\overline{\Lambda}}-\lambda_k\right)\geq(\mu B_L-m\underline{\lambda})$, $\overline{\Lambda}\leq(m-1)\overline{\lambda}$, $\overline{\overline{\Lambda}}\leq\mu B_L-(m-1)\underline{\lambda}$ to obtain an upper bound on the right hand side of A.45 and the following sufficient condition:

$$A\geq\frac{\beta\kappa B}{\mu B_L}\frac{(2m\overline{\lambda}\mu B_L-m^2\underline{\lambda}^2)}{(\mu B_L-m\underline{\lambda})} \tag{A.46}$$

Substituting the definitions of $A$ and $B$ yields result $(ii)$.

## Appendix E: Empirical estimation of probability of result collection

In this section, we summarize the findings from **?** that estimates the impact of turnaround time on the probability of result collection. The study collected data on 1679 samples from 7 health facilities in Mozambique between 2009 and 2010. The outcome variable was whether the result was collected by the infant caregiver. The main predictor variable was an indicator variable $late_30$ that captured whether the turnaround time of a particular sample was greater than 30 days. To avoid any confounding, we also included the gender and age of the infant to account for higher mortality rate (and hence lower collection probability) in older infants and any possible gender bias. Similarly, we included fixed-effects for each clinic and each calendar year to account for unobserved heterogeneity at the level of these units of analysis. These include possibilities that some clinics might be more effective in following up with patients compared to other clinics and the EID program, in aggregate, might have improved or worsened over time. Finally, we also included an indicator variable to capture whether ERS (expedited results system that entails delivering mobile text messages from the laboratory that can be remotely printed at the health facilities) was implemented. The main intention was to tease out the effect of reduced TAT from other unobserved effects of the ERS program such as improved awareness among the caregivers thereby leading to higher result collection. The results of this model are included below for easy reference.

**Table A.3 Estimating the impact of turnaround time on the probability of result collection (?)**

| Variable | Coefficient(Std. Err.) |
|---|---|
| Intercept | $0.898^{**}(0.170)$ |
| late_30 | $-0.399^{*}(0.171)$ |
| age | $-0.116^{**}(0.025)$ |
| gender | -0.140(0.148) |
| 1b.clinic_code | 0.000(0.000) |
| 2.clinic_code | $-0.481^{**}(0.079)$ |
| 3.clinic_code | $-1.059^{**}(0.062)$ |
| 4.clinic_code | $-0.469^{**}(0.033)$ |
| 5.clinic_code | $0.176^{**}(0.039)$ |
| 6.clinic_code | $-0.499^{**}(0.080)$ |
| 10.clinic_code | 0.120(0.091) |
| 2010.year | $-1.347^{*}(0.604)$ |
| ERS | -0.235(0.286) |
| N | 1679 |
| Log-likelihood | -1016.07 |

These results support our hypothesis that, results delivered later than 30 days have significantly lower probability of collection. We also find that older infants are less likely to collect results (potentially due to mortality). Similarly, we find significant differences across clinics and calendar years. Finally, we fail to find any additional impact of ERS on result collection apart from the one that operates via reduced TAT.

## Appendix F: Expanded model formulations

In this section, we develop two models that expand our main formulation to include: (i) multiple labs with optimization over clinic-lab allocation decision, and (ii) objective of maximization of health outcomes subject to an overall budget constraint.

### F.1. Clinic-Lab assignment decision

Let $l$ index over the set of centralized laboratories $\mathcal{L}$ within the health network. Further, let $y_{il}$ represent the binary decision of assigning clinic $i$ lab $l$ or allocating it a POC device, i.e.

$$y_{il} = \begin{cases} 1 & : i\,\text{assigned to lab}\,l, \\ 0 & : i\,\text{assigned a POC.} \end{cases}$$

Then the number of HIV+ infants receiving their results at clinic $i$ is given by

$$N_i = \lambda_i \left( \sum_l \left( B_{il} y_{il} \left( \mathbb{E}\left[ \Omega\left( T_i^l\left(\mathbf{y}\right)\right)\right]\right)\right) + A_i \left( 1 - \sum_l y_{il} \right) \right),$$

where $\mathbf{y} = \left\{ y_{11}, \ldots, y_{n|\mathcal{L}|} \right\}$ represents the POC allocation vector across $n$ clinics and $|\mathcal{L}|$ laboratories in the health network, $B_{il} = p_i s^l$ and $A_i = p_i s^P$. The social planner's problem in this setting can be formulated as

$$\begin{aligned} \max_{\mathbf{y}} \quad & \sum_i N_i & & \text{(A.47)} \\ \text{s.t.} \quad & \sum_l y_{il} & \leq 1 \ \forall i, & \quad \text{(A.48)} \\ & \sum_i y_{il} & \leq W_l \ \forall l, & \quad \text{(A.49)} \\ & \sum_i \sum_l y_{il} & = m, & \quad \text{(A.50)} \\ & y_{il} \in \{0,1\} \ \forall i, l, & & \text{(A.51)} \end{aligned}$$

where $W_l$ represents the maximum number of clinics that can be associated with laboratory $l \in \mathcal{L}$. Note that this formulation includes two more constraints compared to our original formulation. The first one reflects the fact that every clinic is at most assigned to one lab but it might not be associated with any lab ($\sum_i y_i = 0$), in which case it was allocated a POC device. The second constraint allows for a maximum number of clinics being assigned to each lab (potentially for administrative purposes). Finally, the constraint on the total number of POC devices is appropriately modified by summing up the clinic-lab assignment variables over all clinics and all labs.

### F.2. Maximization of health outcomes

Let $u^l$ and $u^P$ represent the specificity of the diagnostic equipment at lab $l$ and the POC device respectively. Let $N_i^{++}$, $N_i^{-+}$, and $N_i^{+-}$ denote the number of truly positive, false positive, and false negative results reported at clinic $i$. Then

$$N_i^{++} = \lambda_i p_i \left( \sum_l \left( s^l y_{il} \left( \mathbb{E}\left[ \Omega\left( T_i^l\left(\mathbf{y}\right)\right)\right]\right)\right) + s^P \left( 1 - \sum_l y_{il} \right) \right),$$

$$N_i^{-+} = \lambda_i \left(1 - p_i\right) \left( \sum_l \left( \left(1 - u^l\right) y_{il} \left( \mathbb{E}\left[ \Omega\left(T_i^l\left(\mathbf{y}\right)\right)\right]\right)\right) + \left(1 - u^P\right) \left(1 - \sum_l y_{il}\right)\right),$$
$$N_i^{+-} = \lambda_i p_i \left( \sum_l \left( \left(1 - s^l\right) y_{il} \left( \mathbb{E}\left[ \Omega\left(T_i^l\left(\mathbf{y}\right)\right)\right]\right)\right) + \left(1 - s^P\right) \left(1 - \sum_l y_{il}\right)\right).$$

Let $\mathcal{B}$ denote the total budget available to improve the health outcomes in the network by deploying POC devices. Further, let $c_{il}$ denote the cost of assigning clinic $i$ to laboratory $l$ (which may include transportation costs), $c^P$ denote the average per usage charge of the POC device, $c^T$ denote the average societal cost of starting detected (truly positive and false positive) patients on treatment. $I_i^P$ denote the amortized fixed cost of installing a POC device at a clinic.

Similarly, let $q^{++}$denote the average quality adjusted life years gained per truly positive infants detected and results received, $q^{-+}$ denote the average quality adjusted life years gained from initiating healthy infants on unnecessary treatment (which is potentially negative) and $q^{+-}$ denote the average quality adjusted life years gained by not detecting HIV+ infants (which might reflect the enhancd mortality of such infants). Then, we can express the optimization problem of maximizing overall health outcomes as follows:

$$\begin{aligned}
\max_{\mathbf{y}} \quad & \sum_i \left(q^{++} N_i^{++} + q^{-+} N^{-+} + q^{+-} N^{+-}\right) \\
\text{s.t.} \quad & \sum_l y_{il} \leq 1 \ \forall i, \\
& \sum_i y_{il} \leq W_l \ \forall l, \\
& \sum_i \left( \sum_l \lambda_i y_{il} c_{il} + \lambda_i (1 - \sum_l y_{il}) c^P + \left(N_i^{++} + N_i^{-+}\right) c^T - I_i^P \left(1 - \sum_l y_{il}\right)\right) \leq \mathcal{B}, \\
& y_{il} \in \{0, 1\} \, \forall i, \, l.
\end{aligned}$$