Submitted to *Operations Research*
manuscript OPRE-2014-10-608.R2

Improving HIV early infant diagnosis supply chains in sub-Saharan Africa: Models and application to Mozambique

Jónas Oddur Jónasson
MIT Sloan School of Management, 30 Memorial Drive, 02142 Cambridge, MA.
joj@mit.edu

Sarang Deo
Indian School of Business, Gachibowli, Hyderabad, India, 50032.
sarang_deo@isb.edu

Jérémie Gallien
London Business School, Regent's Park, London NW1 4SA, UK.
jgallien@london.edu

Early diagnosis of HIV among infants born to HIV infected mothers is critical because roughly 50% of untreated infected infants die before the age of two years. Yet most countries in sub-Saharan Africa experience significant delays in diagnosis due to operational inefficiencies in early infant diagnosis (EID) networks. We develop a two-part modeling framework relying on optimization and simulation to generate operational improvements in the assignment of clinics to laboratories and the allocation of capacity across laboratories, and to evaluate the associated impact on the number of infants initiating treatment. Applying our methodology to EID program data from Mozambique, we validate our simulation model and estimate that optimally re-assigning clinics to labs would decrease the average sample turnaround time (TAT) by 11% and increase the number of infected infants starting treatment by about 4% relative to the current system. Further, consolidating all diagnostic capacity in one centralized lab would decrease average TATs by an estimated 22% and increase the number of infected infants initiating treatment by 7%. Our sensitivity analysis suggests that the consolidation of capacity in a single location would remain near-optimal across a wide range of laboratory utilization levels in Mozambique. However, this full consolidation solution is dominated by configurations with two or more labs for EID networks with average transportation times larger than those currently observed in Mozambique by at least 15%.

Key words: Facility location; queueing networks; supply chain optimization; HIV prevention; early infant diagnosis

History: December 14, 2016

1. Introduction

Approximately 90,000 children died of Acquired Immune Deficiency Syndrome (AIDS) related causes and 150,000 were infected with Human Immunodeficiency Virus (HIV) in sub-Saharan Africa

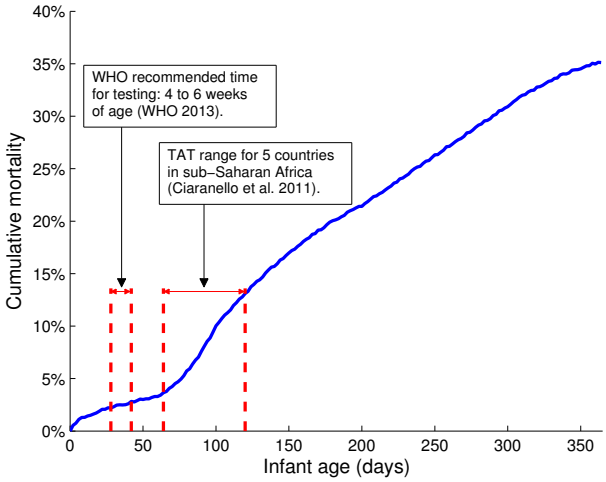
during 2015 (UNAIDS 2016). Without treatment, over a third of the infected infants die before their first birthday and over half of them before their second (Newell et al. 2004, Becquet et al. 2012). While confirmation of infection is a prerequisite for treatment, diagnosing HIV among infants born to HIV positive mothers (HIV exposed infants) is challenging. The most commonly used diagnostic method for adults, HIV antibody blood test, yields a high rate of false-positive results because all infants retain antibodies from their mother for several months after birth. Hence, virological testing must be performed through an enzymatic process called polymerase chain reaction (PCR). However, this method is more expensive and involves more complex technology, so it can only be performed at a few specially equipped laboratories in each country (UNITAID 2015, Creek et al. 2007). As a result, a sample supply-chain is formed where blood samples from HIV exposed infants are obtained at local clinics and sent at the earliest opportunity to labs for analysis. The labs, in turn, send the results back to the clinics where they are, ideally, communicated to caretakers (see §3.1 for a detailed description).

This system of clinics and laboratories is referred to as an *Early Infant Diagnosis (EID) network* (Ciaranello et al. 2011). A key performance measure for such networks is the delay between the drawing of a sample and the results becoming available at the clinic, referred to as *turnaround time* (TAT). Shortening TATs in EID networks is an important and widely shared objective (e.g. Chatterjee et al. 2011), because doing so affects health outcomes in several ways. First, it enables treatment initiation for infants who would otherwise have died before receiving their results (Newell et al. 2004, Becquet et al. 2012) thereby reducing mortality. In particular, Figure 1 shows that even when infants are tested at the recommended age of 4 weeks, in some countries up to 7% of the infected ones are expected to die before their test results become known, due to long TATs. Second, it makes caretakers more likely to collect results (Latigo-Mugambi et al. 2013, Deo et al. 2014). Third, it facilitates early treatment initiation, which has a positive impact on subsequent health outcomes (Violari et al. 2008).

The contribution of our study is twofold. First, we provide a rigorous quantitative framework for the operational design of EID networks in sub-Saharan Africa, where the design decisions considered include the allocation of diagnostic capacity to labs and the assignment of clinics to labs. Second, we illustrate the application of this framework to the EID network in Mozambique. Because this study focuses on operational guidelines potentially affecting the health of vulnerable and underprivileged populations, our methodology includes data-driven model development, rigorous model validation, and predictive accuracy assessment. Its two main components are described below.

First, we develop a discrete event simulation model to predict the health outcome (number of infants initiating treatment) associated with a given EID network configuration. This simulation

Figure 1 First year cumulative mortality for untreated infants infected by HIV at birth



Note. Mortality rate based on clinical trial data from Newell et al. (2004). World Health Organization (WHO) recommendation (WHO 2013), and range of turnaround times for Botswana, Cote d'Ivoire, Kenya, Swaziland, and Tanzania, as reported in Ciaranello et al. (2011).

model itself comprises three sub-models. The operational sub-model includes the detailed dynamics associated with collection of samples at the clinics, their transport to the labs and subsequent batching and congestion at the labs. These dynamics are used to calculate the distribution of TAT at each clinic for a given configuration of the EID network. The clinical sub-model includes information on whether mothers participate in prevention of mother to child transmission (PMTCT) programs, the age at infection and the age at testing. These attributes are used to calculate the probability of an infant being infected and the probability of an infected infant dying before starting treatment. Finally, the behavioral sub-model includes the association between TATs and the likelihood of follow-up by caretakers (Latigo-Mugambi et al. 2013, Deo et al. 2014), and is used to obtain the eventual health outcome of the number of infants initiating treatment.

Second, we develop optimization models to generate alternate configurations for the EID network including improved assignments of clinics to labs for a given distribution of lab capacity (*optimal lab assignment*, or OLA), and improved allocations of lab capacity between several candidate sites (*optimal capacity allocation*, or OCA). The latter problem is more general because it also includes assignment decisions, and constitutes a variant of the classical facility location problem with discrete capacity decisions in each location, stochastic congestion, batch processing and a specific objective associated with health outcomes (see §2.2). While these optimization models also capture the relevant empirical infant survival function (Newell et al. 2004) and behavioral model of caretaker follow-up (Latigo-Mugambi et al. 2013, Deo et al. 2014), in contrast with the simulation model they include approximate dynamics for blood samples at the labs based on a combination

of existing approximations for the $GI^{[X]}/G^{(b,b)}/c$ and $G/G/1$ queueing systems (Hanschke 2006, Sakasegawa 1977).

We apply our methodology to the EID network in Mozambique, relying on extensive field data to estimate model parameters and for out-of-sample validation of the simulation model's prediction accuracy. This model we then use to estimate the potential improvement in health outcomes from alternate design configurations over the current EID network. We find that optimally re-assigning clinics to labs is predicted to reduce average TAT by 11% and increase the proportion of infected infants receiving treatment by 3.6%. Moreover, consolidating all diagnostic capacity in a single location accordingly to the OCA solution is predicted to shorten average TAT by 22% and increase treatment initiation among infected infants by up to 6.9%. In addition we conduct extensive sensitivity analyses, to systematically evaluate the impact of model parameter changes or environment modifications on these results. We find that optimal network design decisions are most important in EID networks with either relatively high or relatively low utilization. We also show that the simple solution of consolidating capacity in a single location is robust to changes in utilization for networks with short to moderate transportation times. However, for networks where average transportation times between clinics and labs are approximately 15% longer than in Mozambique, a decentralized distribution of capacity with at least two labs is predicted to be superior.

To the best of our knowledge, this is the first study to describe an analytical framework for developing operational EID network design recommendations. While Mozambique is our main case study, given the structural similarities in EID networks across sub-Saharan Africa, our methodology produces insights that may be relevant to other countries. In particular, we characterize through extensive sensitivity analysis how the optimal capacity allocation solution changes for a range of laboratory utilization and transportation delays.

In the rest of the paper, we discuss the relevant literature in §2 and provide more detailed descriptions of EID systems and our field dataset in §3. Our simulation and optimization models are discussed in §4 and §5, respectively. Results are reported in §6 and concluding remarks are provided in §7.

2. Literature review

Aside from the EID literature cited in the introduction, our work is related to two distinct streams of literature. First, our work relates to the emerging literature on global health operations management that employs similar model-based approaches to suggest improvements in health systems (§2.1). Second, our methodological approach is related to the extensive literature on stochastic facility location problems (§2.2).

2.1. Global health operations management

Our work contributes to an emerging literature on the management of health care delivery operations in resource limited settings (Kraiselburd and Yadav 2011). This includes studies regarding the impact of uncertain donor funding on inventory management and availability of health products (Rashkova et al. 2016, Natarajan and Swaminathan 2014), the impact of supply uncertainty on the optimal balance between initiation of HIV treatment for new patients and ensuring continuity of treatment for existing ones (Deo et al. 2016), and how to optimally distribute drugs from a central warehouse to clinics (Parvin et al. 2014, Vledder et al. 2013, Leung et al. 2014, Gallien et al. 2014). Furthermore, Taylor and Xiao (2014) analyze whether drug subsidies should be provided at the wholesaler or the retailer level to maximize access and McCoy and Johnson (2014) investigate how a clinic should use limited resources to enroll new patients over. Our work is similar to McCoy and Johnson (2014) and Deo et al. (2016) in that we explicitly capture the impact of operational decisions on health outcomes (as opposed to operational measures). The main differentiating factor of our work is its focus on a diagnostic system as opposed to a drug supply chain or an inventory system. In addition, our modeling of patient mortality and behavior is closely linked with recent empirical studies.

2.2. Facility location problems

Our optimization problem is related to the extensive literature on facility location problems that explicitly capture congestion at the facilities (Baron et al. 2008, Zhang et al. 2009, Marianov and Serra 1998, 2002, Marianov 2003, Marianov et al. 2008, Brimberg et al. 1997, Brimberg and Mehrez 1997 — see Boffey et al. (2007) for a comprehensive survey).

However, existing models cannot be adapted to include several important features of EID networks. First, the operational dynamics in our network are substantially different from those previously considered in the literature. Specifically, our network involves the two echelons of clinics and laboratories and we explicitly model delays at both. The sample dispatches at each clinic are stochastic and batched, and dispatches from different clinics superpose to form the arrival stream at the laboratories. Second, the processing facilities in our model (i.e., the laboratories) are characterized by batched service, which to date have not been considered in the facility location literature. In such systems, delays occur due not only to congestion but also batching. As a result, overall delay in our facilities is a non-monotone function of utilization, which substantially affects the qualitative nature of optimal solutions (e.g., a single moderately loaded service facility with long travel distances may dominate a solution with multiple lightly loaded and highly accessible facilities). Our setting is therefore in strong contrast with that of Chao et al. (2003), whose main determinant of optimal network structure in a stylized model is switching costs (analogous to transportation

times in an EID system). As soon as switching costs increase to a point where any fraction of demand must be processed locally, their optimal solution becomes of the "one large many small" type, whereas in our problem the solution with a single central facility remains optimal for a wide range of transportation times. Similarly the optimal solution in Chao et al. (2003) does not depend on utilization whereas for our setting the structure of the optimal solution varies with utilization.

The work of Deo and Sohoni (2015), which also lies at the intersection of the above three streams of literature, is closest to our paper in its approach. It also develops optimization and simulation models for operational decisions in EID networks but differs substantially in its objective, methodology, decisions considered, and complexity. While their main objective is to establish the qualitative insight that different point-of-care (POC) diagnostic device allocation policies may substantially affect the overall effectiveness of a POC device implementation, our research objective is to generate operational recommendations for actual complex EID networks operating with current technology. In terms of methodology their optimization model considers a simplified model of an EID network involving a single lab serving several clinics and their simulation model is not subjected to a validation study. In contrast, our optimization model considers a full network (requiring a different modeling and solution approach) and the prediction accuracy of our simulation model is validated out-of-sample.

3. Early infant diagnosis systems and the Mozambique data

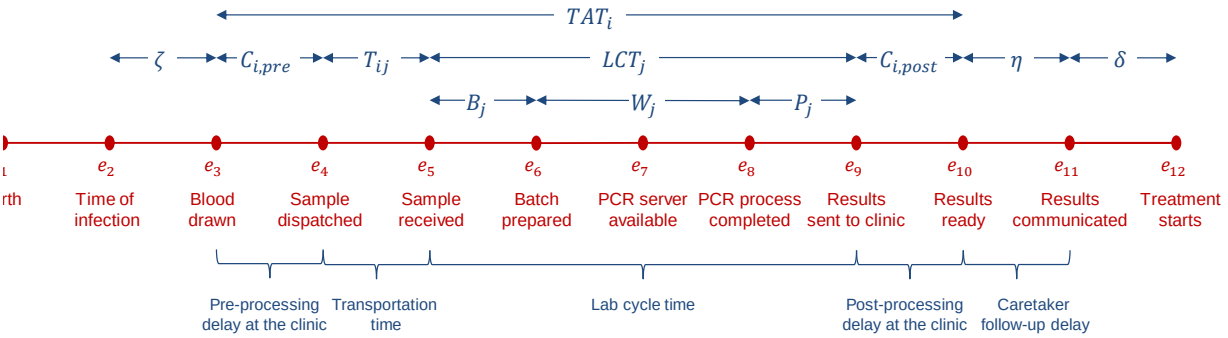
In this section, we provide some background information on EID programs in sub-Saharan Africa (§3.1) and a brief description of our field dataset (§3.2).

3.1. Background on EID systems in sub-Saharan Africa

The risk of HIV transmission from mother to child can be reduced significantly with enrollment in PMTCT programs during pregnancy and delivery (Dabis 1999, Wiktor 1999). Unfortunately, in 2013 only about two-thirds of pregnant women living with HIV in low- and middle-income countries were enrolled in PMTCT programs, due to poor adherence, insufficient access to knowledgeable providers or medicines, and other issues (WHO 2014). As a result, about 17% of the 1.4 million children born to these mothers (also called HIV exposed infants) were infected with HIV. Of these estimated 240,000 newly infected children, about 89% lived in Africa. While all HIV exposed infants are currently recommended to undergo virologic testing between 4 and 6 weeks after birth (WHO 2013), only about 44% of those in low- and middle-income countries received a virological test and less than one quarter was initiated on Antiretroviral Therapy (ART) (WHO 2014).

Figure 2 shows a timeline of various steps involved in testing infants in a typical EID program in sub-Saharan Africa. The process starts when a caretaker brings the HIV exposed infant to a

Figure 2 Timeline for an EID sample



Note. The timeline for an EID sample collected at clinic i, which is assigned to lab j. We include the notation of variables and time epochs for reference in subsequent sections.

clinic for sample collection. At the clinic, a nurse typically collects the blood sample, dries it on filter paper (Smit et al. 2014) and makes a follow-up appointment for the caretaker to visit the clinic again after a month to collect the results. This follow-up interval is chosen to coincide with the provision of other healthcare services to the family such as HIV treatment and postnatal care for the mothers and immunization visits for the infant.

Due to limited resources and infrastructure, clinics do not usually have access to regular and reliable means to transport samples to the laboratory to which they are assigned. Instead, samples accumulate at the clinics until an ad-hoc transportation opportunity (e.g. some local clinic staff or community member traveling to the district capital) is realized. The delay experienced at the clinic before the dispatch of samples is termed *pre-processing clinic delay*. Once the transportation opportunity arises, all accumulated samples are dispatched to the nearest district facility, from where they can be forwarded to the lab by more regular means of transport such as road or air couriers. The delay from dispatch until arrival at the lab is hereafter called *transportation delay*.

Most countries in sub-Saharan Africa have few labs that are capable of performing virologic testing, due to high equipment costs and requirement of trained personnel. The PCR process also requires costly reagents which necessitates conducting the process in batches to minimize the cost of operation. The time for which samples wait in the lab to make fully formed processing batches is called *lab batching delay*. The subsequent wait until a PCR machine becomes available for processing the batch is referred to as the *lab congestion delay*. Upon subsequent completion of the process, a lab supervisor cross-checks the results and types them in a computer leading to additional *post-processing delays*. The sum of lab batching, congestion, processing and post-processing delays for any sample is referred to as the *lab cycle time* (LCT). The communication of test results back to the originating clinic is performed via short message service (SMS) text message, or occasionally via paper courier (using similar transportation methods as the untested samples).

After test results are received at the clinic they are prepared for communication resulting in a short delay termed as the *post-processing clinic delay*. The results are then typically communicated in person to the infant caretakers during their next visit. Finally, additional delays may occur after collection of results until treatment initiation due to other barriers to care, e.g. drug stock-outs or availability of treatment in other facilities (Kieffer et al. 2009).

Longer delays are associated with lower likelihood of collection of results or initiation of treatment, driven by clinical as well as behavioral reasons. First, due to the rapid progression of the disease, several infants may die before receiving their results or initiating treatment (Newell et al. 2004). Second, delays beyond 30 days (coinciding with the follow-up appointment) may reduce the probability of repeated attempts by caretakers to seek results (Deo et al. 2014, Latigo-Mugambi et al. 2013).

3.2. Mozambique EID dataset

The Mozambique EID network[1] is similar in structure to those in most of sub-Saharan Africa. It comprises four labs performing virologic testing for over 400 clinics spread over 11 regions and 128 districts. The labs in Nampula, Quelimane, and Beira have a single automatic PCR machine while that in Maputo has one automatic as well as an older manual one. The structure of the EID program in Mozambique and its practices closely match those described in §3.1 (full details in §EC.2), with the current assignment of clinics to labs based on the regional administrative boundaries.

We use routinely collected EID program data from Mozambique to calibrate the simulation and optimization models and to validate the prediction accuracy of the simulation model. The dataset (full details and summary statistics in §EC.3) comprises individual patient records for 34,791 infants, who presented for testing at the 410 EID clinics operating in Mozambique from January to December 2011. The average age at testing was 88 days, 75% of mothers had participated in a PMTCT program, and 13% of infants tested were HIV positive. The average LCT and TAT were 25 and 40 days, respectively. The national dataset does not include information regarding collection of results and initiation of treatment. A small pilot study examining caretaker behavior for about 800 infants found that test results were collected by 39% of caretakers (CHAI 2013).

4. Simulation model and parameter estimation

In this section, we first introduce the operational component of the simulation model (§4.1), followed by the clinical and behavioral components (§4.2), and a description of our validation method (§4.3).

4.1. Operational model

Let $i \in \mathcal{I} = \{1...I\}$ denote a clinic, and $j \in \mathcal{J} = \{1...J\}$ denote a lab. Let the binary variable $z_{ij} \in \{0, 1\}$ denote whether or not clinic i is assigned to lab j. We represent the set of clinics assigned

to lab j by $\mathcal{I}_j = \{i : z_{ij} = 1\}$. For a given lab assignment, the main operational output for clinic i is the turnaround time TAT_i, which consists of the four random components shown in Figure 2,

$$TAT_i = C_{i,pre} + \sum_{j=1}^{J} z_{ij} T_{ij} + \sum_{j=1}^{J} z_{ij} LCT_j + C_{i,post}, \tag{1}$$

where $C_{i,pre}$ denotes the wait for a transportation opportunity at the clinic, T_{ij} is the transportation time from clinic i to lab j, LCT_j is the LCT for lab j, and finally $C_{i,post}$ represents the administrative delay at the clinic once the lab has sent the results back.

4.1.1. Clinic delays. The first component of the TAT, $C_{i,pre}$, is jointly determined by the arrival process of infants (samples) to the clinic and the arrival of transportation opportunities. Because of the preponderance of days with no infant arrivals at each clinic in the data, we use a Zero-Inflated Poisson (ZIP) model for the daily number a_i of arriving samples at clinic i:

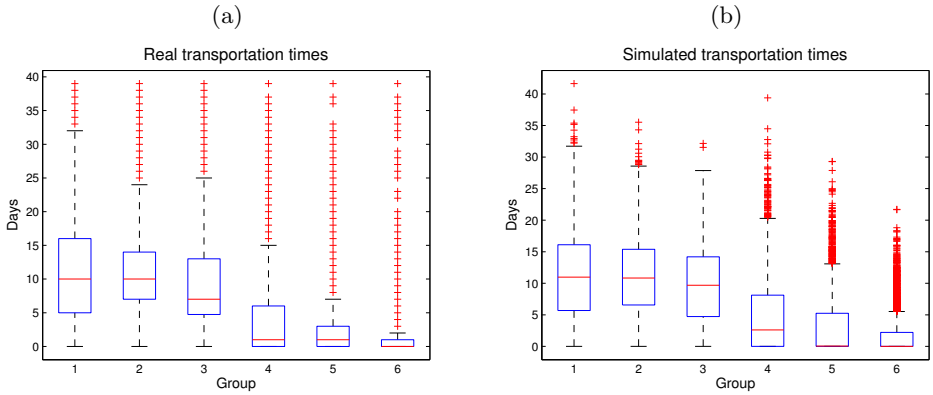
$$\begin{aligned} \mathbb{P}(a_i = 0) &= (1 - \pi_i) + \pi_i e^{-\lambda_i} \\ \mathbb{P}(a_i = k) &= \pi_i \frac{\lambda_i^k e^{-\lambda_i}}{k!}, k > 0. \end{aligned}$$

The ZIP model is a mixture of a Bernoulli random variable (with mean π_i), and a regular Poisson random variable (with rate λ_i). We use the Maximum Likelihood method to estimate the mean and the variance of the ZIP arrival process at clinic i, given by $\mathbb{E}[a_i] = \lambda_i \pi_i$ and $\sigma^2(a_i) = \lambda_i \pi_i + \lambda_i \pi_i (\lambda_i - \lambda_i \pi_i)$, respectively (Cameron and Trivedi 2013).

Based on our discussions with EID field managers, the arrival of transportation opportunities at each clinic i is assumed to be independent of the sample arrival process there, and modeled as a Poisson process (with rate γ_i). The clinic delay before being dispatched thus follows an exponential distribution, and the mean *pre-processing clinic delay* equals the mean inter-arrival time of transportation opportunities. As discussed in §EC.4.1, our dataset of historical clinic delays does provide some support for that assumption. Finally, based on our field work, we assume the *post-processing clinic delay* $C_{i,post}$ to be 1 day for all clinics (Quevedo and Crea 2013).

4.1.2. Transportation delays. Our Mozambique EID dataset includes transportation delays for the current clinic-lab assignment, but not for alternative assignments that might arise during optimization. Hence, we use existing data and the expertise of field managers (Quevedo and Crea 2013) to estimate the transportation delays for all possible alternate lab assignments. Models and details about this analysis can be found in §EC.4.2. In short, the main determinants of transportation delays are the relative location of the clinics and labs. Specifically, whether a clinic is located rurally, in a district capital, or a regional capital and whether it is located in the same region as the

Figure 3 Transportation time comparison



Note. A comparison of the transportation time distributions for each group of clinics. The red line represents the median transportation time for each group, the blue box the interquartile range (IQR), and the whiskers extend to the most extreme data point within 1.5 IQRs from the blue box.

lab or not. Based on this we divide the clinics into 6 groups and empirically estimate the impact of each factor on transportation times.

Figure 3 compares the distribution of transportation times observed in the data with simulated ones (according to the empirical results of §EC.4.2), for each of the six groups. The figure shows that simulated transportation time distributions are comparable to those observed in the field. In particular, the median, the interquartile range and the whiskers match fairly closely between the two figures, even if the right tail is slightly longer in the field data. We observe that there is a substantial difference in transportation delays between groups 1, 2, and 3 (clinic and lab in the same region) and groups 4, 5, and 6 (clinic and lab in different regions). This suggests that samples arriving in a provincial capital without a lab encounter substantial delays before being flown to the region of their assigned lab. Finally, the distributions of transportation times corresponding to clinic-lab assignments different from the current one are simulated using the same models.

To validate our these predicted transportation times, we organized a comprehensive review by a local transportation expert, who confirmed that these predictions were realistic (Quevedo and Crea 2013). In addition, we performed extensive numerical experiments designed to test the sensitivity of our findings to these transportation times (see §EC.8.4).

4.1.3. Laboratory delays. The operational dynamics inside the laboratories are quite complex as they involve several activities (data entry, sample preparation, batch testing, results analysis, etc.) and resources (lab technicians, testing machines, computers, etc.). Unfortunately, we do not have access to data at this granularity, as our dataset only includes the date of arrival, the date of processing, and the date of results transmission for each sample (denoted in Figure 2 by e_5,

e_8 and e_9, respectively). Hence, we estimate the delay of samples LCT_j at each lab j by focusing on the following three main components:

$$LCT_j = B_j + W_j + P_j, \tag{2}$$

where B_j denotes the lab batching delay, W_j denotes the sojourn time of the fully formed batch, and P_j denotes the post-processing delay.

To simulate LCT_j each lab is modeled as a queueing system. In practice the arrival of samples to the lab results from the superposition of arrival processes from the set $\mathcal{I}_j$ of its assigned clinics. Each lab can have multiple (c_j) testing machines working in parallel that process only full batches of $b = 92$ samples. This batch size is determined by the dimensions of prevalent equipment for PCR testing and sample preparation. Hence the lab can be described as a batch arrival/batch processing queueing system $\sum_{i\in\mathcal{I}_j} GI^{[X_i]} \Big/ G_j^{(b,b)} \Big/ c_j$, where X_i denotes the random dispatch count of samples from clinic i, and $G_j^{(b,b)}$ highlights that the service times have general lab-specific distributions with fixed batch size b.

Lab batching. Given the departure streams at each clinic (§4.1.1), the transportation times for each clinic (§4.1.2), the network configuration characterized by $\mathcal{I}_j$, and the processing batch size b, B_j can be simulated directly.

Sojourn time of full batches. The key challenge in estimating W_j is that we do not directly observe service times, i.e. the time required to process each batch. Our approach is to choose a service time distribution for the queueing system that results in the best out-of-sample fit between simulated and observed LCTs. We restrict our attention to Erlang distributions with probability density function

$$f(S_j) = \frac{S_j^{k_j-1} e^{-\frac{S_j}{\mu_j}}}{\mu_j^{k_j}(k_j-1)!}, \tag{3}$$

where $k_j \in \mathbb{N}$ and $\mu_j \in \mathbb{R}^+$ are shape and scale parameters, respectively. This choice of a family of distributions is driven both by tractability considerations and its flexibility to represent a range of distributions, from the exponential to unimodal distributions similar to those found in other service settings (Brown et al. 2005).

We fit the k_j and μ_j parameters using the Mozambique dataset and conduct an out-of-sample validation exercise. Table 1 shows the comparison of simulated and empirical LCT distributions for each lab. §EC.4.2.1 contains a detailed discussion of the fitting methodology and outcomes (particularly for the Maputo lab), but overall, we believe that our fitted queueing model satisfactorily captures lab dynamics in Mozambique for the purpose of this study.

Table 1 Lab cycle time comparison

	Training period			Validation period		
	Field data	Simulation	Difference	Field data	Simulation	Difference
LCT mean (days):						
Lab 1 - Maputo	29.2	27.4	1.8	52.5	38.8	13.7
Lab 2 - Nampula	11.7	11.3	0.4	18.3	19.1	-0.8
Lab 3 - Quelimane	13.9	13.2	0.7	9.2	12.1	-2.9
Lab 4 - Beira	25.4	26.3	-0.9	34.5	37.6	-3.1
LCT std. dev. (days):						
Lab 1 - Maputo	13.1	12.2	0.9	19.0	6.0	13.0
Lab 2 - Nampula	7.7	5.9	1.8	5.8	5.5	0.3
Lab 3 - Quelimane	9.9	7.3	2.6	5.2	5.2	0.0
Lab 4 - Beira	11.6	10.0	1.7	8.4	6.6	1.8

A comparison between the simulated and the real lab cycle times in days. Training and validation periods to the left and right, respectively.

Post-processing. From a survey sent to the labs and conversations with field managers, we determined that P_j is driven by the need to wait for both the lab manager and the lab computer to become available for data entry (Quevedo and Crea 2013). We simulate this delay component using an empirical post-processing delay distribution, which is separately estimated for each lab from data (see §EC.4.2.2 for more details).

4.2. Public health impact model

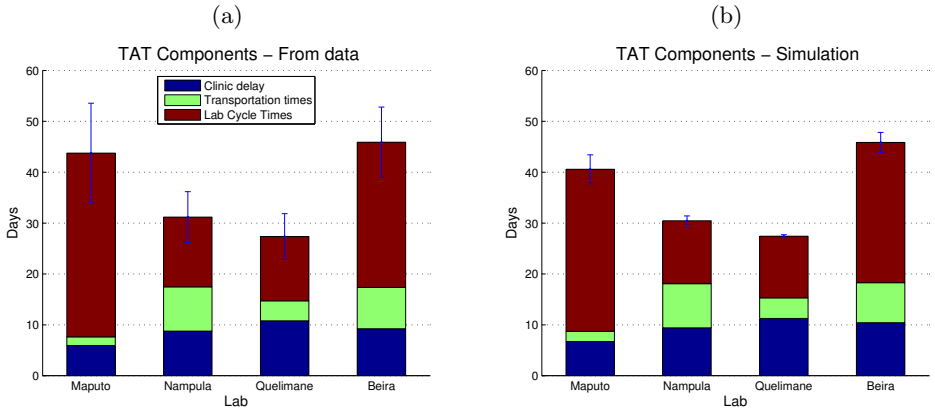
This subsection presents our model for the public health impact of the TAT, which is determined by the transmission of disease from mother to child, the time of infection and the age at testing (§4.2.1), loss to follow-up in collection of results (§4.2.2), and treatment initiation (§4.2.3).

4.2.1. Vertical transmission, age at infection and age at testing. For each infant, we simulate whether its mother participated in the PMTCT program or not based on the participation data for each clinic ($Q_{i,P}$ and $Q_{i,NP}$, respectively). Similarly, we simulate whether an infant was infected with HIV or not based on the empirical mother-to-child transmission rates for mothers who participated in the PMTCT programs (V_P) and those who did not (V_{NP}).

The age of infants at testing is simulated using a country-wide empirical distribution estimated from the Mozambique EID program data. When applicable, the age of infants at the time of infection is simulated based on published estimates (Kourtis et al. 2006). These individual variables are used to calculate the time since infection at the time of testing, which is denoted by ζ.

4.2.2. Results follow-up rates. Using data from 7 clinics in Mozambique, Deo et al. (2014) estimate an 18% reduction in probability of results being picked up by caretakers (the *follow-up rate*) if the TAT is greater than one month as compared to less than one month (44% in the first month and 36% in the second month). We assume that the odds ratio between the first two

Figure 4 Mean TAT validation



Note. a) TAT by lab from field data. b) TAT by lab output from simulation model. The bars denote 95% confidence intervals.

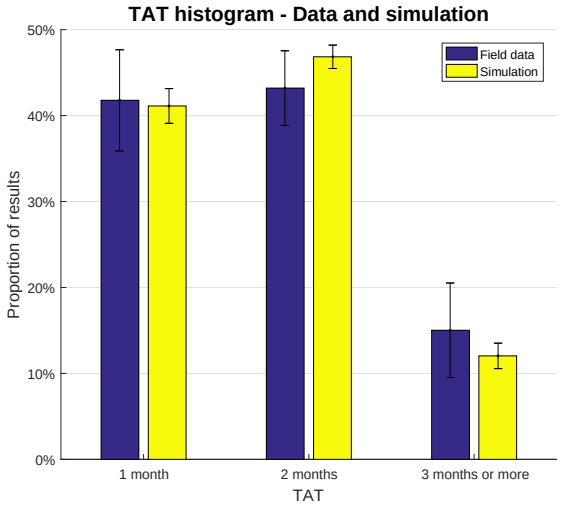
months also holds between months 2 and 3, but conservatively assume a constant follow-up rate for all samples with a TAT of 3 months or more. We then use the results from Latigo-Mugambi et al. (2013) to estimate the odds ratio for collection of results between PMTCT and non-PMTCT caretakers (0.67). These estimates reflect the combined effect of infant mortality as well as lack of follow-up by caretakers of infants who are still alive. In order to separate the two effects, we also simulate the probability of death using the mortality curve estimated by Newell et al. (2004), taking into account the communication delay, η (see Figure 2).

4.2.3. Treatment follow-up rates. Due to infant mortality (Newell et al. 2004, Becquet et al. 2012) or lack of caretaker follow-up (Ciaranello et al. 2011), not all infected infants whose results are collected are eventually initiated on treatment. In the absence of specific data from Mozambique, we assume a constant follow-up rate to treatment of $\upsilon = 75\%$ based on aggregate estimates reported across sub-Saharan Africa (Ciaranello et al. 2011). Finally, when applicable we allow for an additional delay (δ) of 3 weeks between collection of results and initiation of treatment, based on published estimates from other parts of sub-Saharan Africa (Kieffer et al. 2009).

Combining the above elements we obtain the probability density function of follow-up for treatment. This density function is denoted by $F_P(\zeta, TAT_i, \eta, \delta)$ and $F_{NP}(\zeta, TAT_i, \eta, \delta)$ for PMTCT and non-PMTCT participants, respectively. The expected number of infected infants treated as a result of visiting clinic i is thus given by

$$\mathbb{E}[N_i] = \upsilon \pi_i \lambda_i \sum_{k \in \{P,NP\}} \left(Q_{i,k} V_k \mathbb{E}\Big[F_k\big(\zeta, TAT_i, \eta, \delta\big) \Big] \right). \tag{4}$$

Figure 5 TAT distribution validation



4.3. Model validation

To evaluate the predictive accuracy of our simulation model, we compare the simulated output for the current network configuration with the values of three main outcome variables estimated from field data: average TAT at clinic i ($\mathbb{E}[TAT_i]$), proportion of results received within one month ($P(TAT_i < 30)$), and number of infected infants initiated on treatment ($\mathbb{E}[N_i]$).

Figure 4 shows that the simulated average TAT at all four labs is very close to that estimated from field data. Further, the relative contribution of the various components of TAT is also very similar between the simulation output and field data for all four labs. Figure 5 shows that the distribution of TAT, specifically the fraction of results received within one month, also matches across the simulation output and field data. While extensive data on HIV treatment initiation for infants is not currently available in Mozambique, a small pilot study of about 800 infants suggests that the uptake of treatment amongst infected infants is about 39% (CHAI 2013). This matches fairly well with our simulated output of 38%.

The validated simulation model allows us to quantify the impact of a given network configuration on the health outcome of number of infants initiating treatment. But it cannot be used to efficiently search for an improved network configuration. In the next section, we develop optimization models that serve this purpose.

5. Optimization models

In this section we develop optimization models to generate improved assignments of clinics to labs for a given distribution of lab capacity (*optimal lab assignment*, or OLA), and improved allocations of lab capacity between several candidate sites when the assignment decisions are also endogenous (*optimal capacity allocation*, or OCA). As stated in §1, the main objective we pursue is

to maximize the number of infected infants initiating treatment. An important associated challenge is to characterize the impact of these design decisions on LCTs analytically. We first derive a related approximation in §5.1, then use it in §5.2 and §5.3 to formulate tractable optimization models for OLA and OCA, respectively.

5.1. Lab cycle time approximation

Consider lab j, receiving and processing samples sent from clinics $i \in \mathcal{I}_j$. As before, the arrival rate of samples and transportation opportunities at each clinic i are denoted by $\pi_i\lambda_i$ and γ_i, respectively. The random number of samples in a shipment from clinic i is denoted by X_i.

As noted in §4.1.3, lab dynamics before the post-processing phase can be modeled as a $\sum_{i\in\mathcal{I}_j} GI^{[X_i]}/G^{(b,b)}/c$ queueing system. We denote the average total number of samples waiting for testing or being tested in this system, excluding those in post-processing, as $\mathbb{E}[M_j]$. Following Hanschke (2006) we approximate this quantity by considering the two main components of sojourn time for this system, namely (i) aggregation of arriving batches X_i into processing batches (of size $b = 92$), and (ii) a traditional $GI/G/c$ queueing system with fully formed batches as flow units. With this logic we arrive at the following proposition.

Proposition 1. *Assuming a coefficient of variation of 1 for the interarrival times of sample shipments arriving to each lab and a constant average clinic batch size ($\frac{\pi_i\lambda_i}{\gamma_i} = E$), Sakasegawa's (1977) approximation, Hanschke's (2006) approximation, and Little's Law yield the following approximation of $\mathbb{E}[LCT_j]$, in EID lab j:*

$$\mathbb{E}[LCT_j] \approx \frac{(b-1)\mathbb{E}[S_j]}{2bc_j\rho_j} + \frac{\mathbb{E}[S_j]\rho_j^{\sqrt{2(c_j+1)}-1}}{bc_j(1-\rho_j)}\left(E + \frac{1}{2} + \frac{b_j SCV[S_j]}{2}\right) + \mathbb{E}[S_j] + \mathbb{E}[P_j]. \tag{5}$$

For a detailed discussion of assumptions and a full derivation of the approximation, see §EC.5. In the above approximation, the four terms on the right-hand side represent lab batching, congestion, processing, and post-processing delays, respectively. Note that (5) is now a convex function of ρ_j, first decreasing and then increasing for $0 \le \rho_j \le 1$. As discussed later, this feature has important implications for network design as both low and high utilization levels can increase TAT due to long lab batching and congestion delays, respectively.

5.2. Lab assignment

Using notation from previous sections and (4), the problem of finding the assignment of clinics to labs z_{ij} that maximizes the expected number of infants initiating treatment can be stated as:

$$\underset{z_{ij},\rho_j}{\text{maximize}}\ \upsilon \sum_{i=1}^{I} \pi_i\lambda_i \left(\sum_{k\in\{P,NP\}} Q_{i,k}V_k \mathbb{E}\Big[F_k\big(\zeta, TAT_i, \eta, \delta\big)\Big] \right), \tag{6}$$

subject to:

$$\sum_{j=1}^{J} z_{ij} = 1 \qquad \forall i \in \mathcal{I}, \tag{7}$$

$$\rho_j = \frac{\sum_{i=1}^{I} z_{ij}\pi_i\lambda_i\mathbb{E}\left[S_j\right]}{b_j c_j} \qquad \forall j \in \mathcal{J}, \tag{8}$$

$$\rho_j \leq 1-\epsilon \qquad \forall j \in \mathcal{J}, \tag{9}$$

$$LCT_j = B_j + W_j + P_j \qquad \forall j \in \mathcal{J}, \tag{10}$$

$$TAT_i = C_{i,pre} + \sum_{j=1}^{J} z_{ij}T_{ij} + \sum_{j=1}^{J} z_{ij}LCT_j + C_{i,post} \qquad \forall i \in \mathcal{I}, \tag{11}$$

$$z_{ij} \in \{0,1\}; \rho_j \in \mathbb{R}_+. \tag{12}$$

Constraint (7) ensures that each clinic is assigned to a single lab. Constraint (8) calculates the utilization resulting from the assignment variables at each lab, and (9) restricts this utilization to be less than one in order to ensure system stability (we use $\epsilon = 0.0001$ in our numerical experiments). Constraint (10) defines LCT for each lab whereas constraint (11) defines TAT for each clinic. Note that both (10) and (11) include random variables, themselves nonlinear functions of the decision variables z_{ij}. Furthermore, both B_j and W_j depend non-linearly on utilization. Finally, the objective function contains an expectation of a function $F_k(\cdot)$ of random variable TAT_i.

In §EC.6.1 we derive and interpret the following linear reformulation of problem (6)-(12) which is computationally tractable despite these challenges:

$$\textbf{OLA}: \underset{z_{ijd},\tau_{jd},\rho_j}{\text{maximize}} \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{d=1}^{D} z_{ijd}\left(\sum_{m=1}^{M}\phi_{dm}\omega_{im}\right), \tag{13}$$

subject to (9) and:

$$\sum_{j=1}^{J}\sum_{d=1}^{D} z_{ijd} = 1 \qquad \forall i \in \mathcal{I}, \tag{14}$$

$$\rho_j = \frac{\sum_{i=1}^{I}\sum_{d=1}^{D} z_{ijd}\pi_i\lambda_i\mathbb{E}\left[S_j\right]}{b_j c_j} \qquad \forall j \in \mathcal{J}, \tag{15}$$

$$\sum_{d=1}^{D} \tau_{jd} d \geq \alpha_{ju}\rho_j + \beta_{ju} \qquad \forall j \in \mathcal{J},\ u \in \mathcal{U}, \tag{16}$$

$$\sum_{d=1}^{D} \tau_{jd} = 1 \qquad \forall j \in \mathcal{J}, \tag{17}$$

$$z_{ijd} \leq 0 \qquad \forall i \in \mathcal{I},\ j \in \mathcal{J},\ 0 \leq d \leq \Delta_{ij}, \tag{18}$$

$$z_{ijd} \leq \tau_{j(d-\Delta_{ij})} \qquad \forall i \in \mathcal{I}\,,\ j \in \mathcal{J},\ \Delta_{ij} < d \leq D, \tag{19}$$

$$z_{ijd}, \tau_{jd} \in \{0,1\}; \rho_j \in \mathbb{R}_+. \tag{20}$$

In formulation (13)-(20), indices $d \in \{1,2,..,D\}$ and $m \in \{1,2,..,M\}$ respectively represent the days and months passed since sample collection (D and M represent some sensible upper bound on

the possible length of TATs). Parameter ω_{im} denotes, for clinic i, the expected number of infants starting treatment if the results were to become available in month m, and parameter ϕ_{dm} denotes the proportion of results that are expected to become available in month m if the expected TAT is d days. Both can be estimated from data, as discussed in §EC.6.1 and §EC.6.2. Decision variables z_{ijd} expand the definition of the previous assignment decision variable z_{ij} as

$$z_{ijd} = \begin{cases} 1 & \text{if } z_{ij} = 1 \text{ and } \lfloor \mathbb{E}\left[TAT_i\right] \rfloor = d, \\ 0 & \text{otherwise.} \end{cases}.$$

Decision variables τ_{jd} are binary indicators for whether the mean LCT of lab j is d days, and parameters α_{ju} and β_{ju} denote the slopes and intercepts of a discrete set $\mathcal{U}$ of tangents approximating $\mathbb{E}[LCT_j]$ as a function of ρ_j on $(0,1)$ (see §EC.6.3). Finally, Δ_{ij} represent all the delays associated with clinic i being served by lab j, except the time spent at the lab, i.e. $\Delta_{ij} = \lceil \mathbb{E}(C_{i,pre}) + \mathbb{E}(T_{ij}) + \mathbb{E}(C_{i,post}) \rceil$.

The MIP formulation (13)-(20) is still challenging to solve for realistically sized problems. In EC.6.4 we introduce cuts to improve computational performance. These result in substantial reductions of computational times.

5.3. Capacity allocation

We now provide a tractable formulation for the more general problem of optimally distributing the PCR testing machines available nationally (their number is henceforth denoted by K) to a number of given potential lab locations, where the resulting assignment of clinics to these locations is also a decision (OCA). To model the capacity allocation decisions, we use a generalized definition for the index set $\mathcal{J}$, now defined as the product of the set of available lab locations (noted $\mathcal{L}$) with the set $\{0,1,2,...,K\}$ of possible capacity decisions in each site. Any element $j \in \mathcal{J}$ can thus be represented as $j = (\ell_j, c_j)$ with $\ell_j \in \mathcal{L}$ and $c_j \in \{0,1,2,...,K\}$, and corresponds to the option of installing c_j testing machines in lab location ℓ_j. The capacity allocation variables can thus be represented by binary variables y_j indexed over $\mathcal{J}$. Although the notation $\mathcal{J}$ is now overloaded because it both refers to the set of lab locations in the context of OLA and to the product set just defined in the context of OCA, this notational choice is justified by the much simpler following statement of optimization problem OCA:

OCA : maximize (13),

subject to:

(14), (15), (9), (18), (19), (EC.15), (EC.16), (EC.17),

$$\sum_{d=1}^{D} \tau'_{jd} \geq \alpha_{ju}\rho_j + \beta_{ju}y_j \qquad \forall j \in \mathcal{J},\, u \in \mathcal{U}, \tag{21}$$

$$\sum_{i=1}^{I}\sum_{d=1}^{D} z_{ijd}\pi_i\lambda_i \leq \frac{\mathbb{E}[S_j]}{b_j c_j} y_j \qquad \forall j \in \mathcal{J}, \tag{22}$$

$$\sum_{j=1}^{J} y_j c_j = K, \tag{23}$$

$$y_j \in \{0,1\}. \tag{24}$$

Note that constraint (16) is replaced by constraint (21) with right hand side $\alpha_{ju}\rho_j + \beta_{ju}y_j$, which ensures that lab configurations receiving no samples have LCTs set to zero. Constraint (22) ensures that no clinics can be assigned to a lab configuration unless it is part of the solution being considered, and (23) ensures that the total number of testing machines allocated to all different locations does not exceed national capacity.

6. Results

In this section, we report simulated[2] estimates of the impact of the mild intervention of OLA (§6.1) and the more drastic intervention of OCA (§6.2) as compared to the status quo (SQ hereafter). In §6.3 we use the case of Mozambique as a base case to generate more general managerial insights for EID network design in sub-Saharan Africa.

6.1. Optimal lab assignment (OLA)

First, consider the predicted impact of optimally assigning clinics to labs. Table 2 demonstrates how the relatively mild intervention of OLA results in a significant impact on health outcomes, summarized in the observation below.

Observation 1 (OLA impact). *The OLA solution is predicted to reduce average TATs by 11% (from 37 to 33 days) and increase the number of results becoming available within a month by 26%, compared to SQ. This, in turn, increases the number of infected infants starting treatment by 4%.*

This increase in infected infants starting treatment is a joint effect of 12% fewer children dying before receiving the results and 3% fewer children being lost due to caretakers not following up and amounts to roughly 50 infants every year based on the Mozambique arrival data.

The operational mechanisms driving the above result can be observed in Table 3. First, we note that the utilization of the Maputo lab has been lowered, resulting in a substantial increase in the number of infected infants treated at the clinics assigned to that lab. Effectively, this reduction is achieved through load-balancing, which becomes possible in OLA as samples from multiple regions are allowed to be processed at each lab. In fact, each lab receives samples from an average of 7 regions as opposed to less than 3 in SQ, in which the lab assignment is determined by administrative boundaries.

Table 2 Main outcomes of OLA and OCA compared to SQ

	SQ	OLA	OCA	OLA impact	OCA impact
Operational performance:					
Mean TAT	37.4	33.3	29.1	−11%	−22%
Results available within a month	41.1%	51.7%	61.9%	+26%	+50%
Public health performance:					
Infants treated	40.9%	42.4%	43.7%	+ 4%	+ 7%
Infants lost to death	7.9%	7.0%	5.8%	−12%	−27%
Infants lost due to non-follow-up	35.9%	35.0%	34.2%	− 3%	− 5%

The public health metrics are reported in expected percentages of infected infants per year. All improvements are significant at $p < 0.001$.

Second, we note the different roles played by the Nampula and Beira labs in OLA. For Nampula, while the average LCTs are practically unchanged in OLA compared to SQ, the average TAT is shortened by a significant 3 days. At Beira, LCTs are shortened by 5 days on average, but with no discernible impact on average TATs. This is because in OLA Nampula serves clinics with low clinic delays — for which a short LCT means that a high fraction of results becomes available within a month. Beira, on the other hand, serves the clinics of all regions (hence the longer transportation times) which have long clinic delays — indicating low probability of getting results back within a month, regardless of lab performance. Since marginally longer TATs have less of an effect on the clinics served by Beira than those served by Nampula, it is optimal to operate Beira at a slightly higher utilization, effectively pooling the negative congestion externality of clinics with long exogenous delays in a single lab. We summarize these insights in the following observation.

Observation 2 (OLA operational mechanisms.) *The main operational drivers for improvement in the OLA solutions are: (i) Load-balancing: In the OLA solution the distribution of sample load is more even across labs. (ii) Clinic-prioritization: Clinics with long exogenous delays are assigned to a specific lab, allowing clinics with a high probability of receiving results within a month to benefit from less congested labs.*

6.2. Optimal capacity allocation

We next consider the impact of optimally re-allocating diagnostic capacity across the four labs of Mozambique. Table 3 shows that the OCA solution is to consolidate all diagnostic capacity, while Table 2 demonstrates the significant performance benefit of this configuration over SQ.

Observation 3 (OCA impact). *It is optimal to consolidate all available testing machines in a single lab located in Nampula. This results in a 22% decrease in average TATs, a 50% increase in the number of results becoming available within one month and a 7% increase the number of infected infants starting treatment.*

Table 3 Operational comparison of SQ, OLA, and OCA

	PCR machines c_j			Utilization ρ_j			Clinic delay $\mathbb{E}[C_{i,pre}+C_{i,post}]$			Transport delay $\mathbb{E}[T_{ij}]$			Lab cycle time $\mathbb{E}[LCT_j]$			Total TAT $\mathbb{E}[TAT_i]$		
Lab	SQ	OLA	OCA	SQ	OLA	OCA	SQ	OLA	OCA	SQ	OLA	OCA	SQ	OLA	OCA	SQ	OLA	OCA
L1 - Maputo	2	2		100%	92%		7.7	7.4		2.0	2.1		31.9	19.7		41.6	29.2	
L2 - Nampula	1	1	5	89%	89%	90%	10.4	8.5	10.2	8.7	7.6	10.7	12.4	12.2	8.1	31.5	28.2	29.1
L3 - Quelimane	1	1		73%	89%		12.3	9.9		4.1	5.2		12.1	16.2		28.4	31.3	
L4 - Beira	1	1		98%	92%		11.4	16.0		7.9	8.5		27.6	22.6		46.9	47.2	
National	5	5	5	90%	90%	90%	10.2	10.2	10.2	5.7	5.8	10.7	21.5	17.3	8.1	37.4	33.3	29.1

TAT components are expressed in days.

This estimated public health impact is a combined effect of 27% fewer children dying before receiving the results and 5% fewer children being lost due to caretakers not following up, corresponding to 74 and 57 infants per year, respectively.

Furthermore, Table 2 is useful for comparing the impact of OCA and OLA. Clearly the OCA intervention is different from OLA in that it is more drastic and requires more changes to the system. Nevertheless, we observe that its benefits over OLA are substantial. Specifically, on every operational (e.g. TATs and results becoming available within a month) and health outcome (e.g. infants treated, infant mortality, caretaker non-follow-up) metric the positive impact of OCA is nearly twice that of OLA.

Based on Table 3, we note that the operational drivers for the improvement in OCA are substantially different from those in OLA. As a result of consolidating capacity in one lab we predictably observe an increase in transportation times (from 6 days to 11 days). However, this increase is outweighed by a dramatic reduction in LCTs (from 21 to 8 days). This total reduction is a combination of batching and congestion delays being reduced from 2 days and 9 days in SQ to 0.5 and 1 day on average in OCA, respectively. In addition there is a decrease of 4 days in post-processing delays due to the more efficient post-processing practices at Nampula, as observed from the data. In fact, in EC.8 we demonstrate that if post-processing delays in Maputo were as short as those in Nampula, the OCA is to consolidate all capacity in the capital of Maputo, taking advantage of its better transportation links. We summarize our observations of the main operational drivers for OCA performance improvements as follows.

Observation 4 (OCA operational mechanisms). *The OCA performance impact is driven by: (i) Consolidation trade-off: For the Mozambique EID network, the increase in average transportation times associated with capacity consolidation is far outweighed by the decrease in average LCTs. (ii) Congestion and batch pooling: Traditional capacity pooling leads to reduced congestion delays while batch formation pooling leads to shortened pre-processing delays.*

Finally, in EC.7 we show that despite the optimization formulations derived in §5 not explicitly capturing the important notion of equity of access across patients or geographic locations, the

improvements achieved by OLA and OCA do not significantly impact standard measures of equity. Further, in EC.9 we conduct extensive sensitivity analyses which among other results show that: (i) the positive impact of OLA and OCA above SQ is statistically significant for a range of reasonable parameters for caretaker follow-up and can be as large as 8.5% and 14.4%, respectively, (ii) that for the baseline parameters there is not a significant benefit from segmenting samples based on caretaker participation in PMTCT programs, and (iii) that by removing the behavioral and clinical factors from the objective function and simply minimizing the operational outcome of TAT leads to significantly worse solutions in some reasonable instances.

6.3. Managerial considerations for EID network design

In this section we summarize three main observations for EID network design, beyond the case of Mozambique. Full details of this analysis are included in EC.8. First, in many settings it might be desirable to transport the samples from all labs in the same district or region to the same lab for administrative convenience. By adding a constraint ensuring each clinic from the same district (or region) is assigned to the same lab and resolving the OLA problem we arrive at the following observation.

Observation 5 (Administrative considerations for OLA). *Restricting all clinics within the same district to be assigned to the same lab achieves statistically the same benefits as OLA, whereas optimizing at the regional level results in significantly worse performance.*

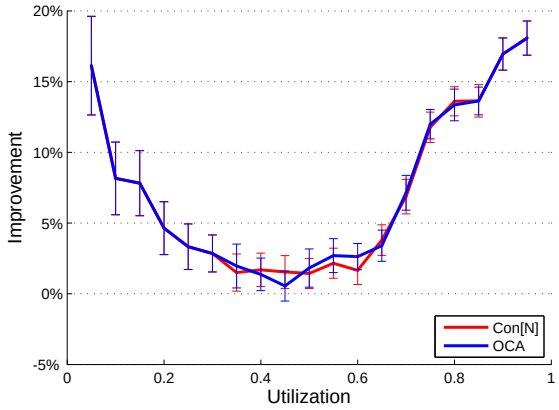
Second, we turn to the OCA solution of consolidating all capacity in Nampula. To generalize our findings to EID networks in other countries, we adjust all sample arrival rates at each clinic by a common multiplicative factor so that system utilization is varied from 5% to 95%, re-solve the OCA problem and simulate the solution for each instance. Figure 6 shows the performance improvement of Con[N] and OCA over SQ for each experiment.

Observation 6 (OCA robustness to utilization changes). *Capacity consolidation is optimal or near-optimal for any level of utilization.*

This result is important from a practical perspective, because it suggests that during the scale-up phase of an EID program when managers initially have ample capacity compared to workload, appropriate allocation of this capacity to labs still has a significant impact on system performance.

Finally, while the main OCA recommendation for capacity consolidation in Nampula is optimal for the EID network in Mozambique, which has average transportation times of 11 days, we now examine the robustness of this solution for networks with longer transportation times. Specifically,

Figure 6 The treatment initiation improvement of OCA and Con[N] over SQ as a function of system utilization.



Note. The blue and red curves represent the improvement of the OCA and Con[N] upon the current lab assignment in the number of infected infants starting treatment. Confidence intervals represent the 95% level.

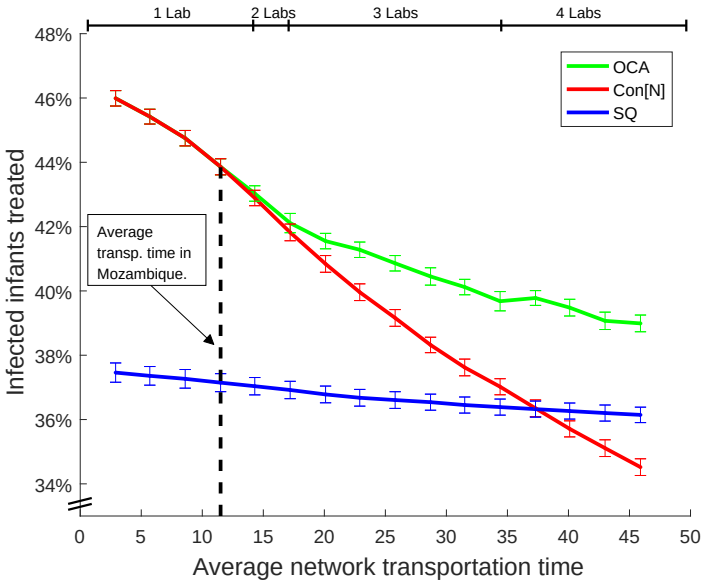
we consider a range of EID networks obtained by scaling the transportation times estimated for Mozambique from 3 days to 36 days, on average. For the resulting continuum from dense to sparser networks, Figure 7 shows the simulated performance of the SQ, the baseline OCA of capacity consolidation in Nampula (labelled $Con[N]$), and the re-optimized OCA for each instance.

Observation 7 (OCA for increasing transportation times). *As EID network transportation times are scaled up we note that: (i) Con[N] is optimal for transportation times up to 15% longer than empirically observed. (ii) Con[N] outperforms SQ for transportation times up to 300% longer than empirically observed. (iii) For very high transportation times, resulting in the same capacity allocation in OCA as in SQ, OCA still significantly outperforms SQ due to its improved lab assignment.*

7. Conclusion

In this paper, we develop a two-part modeling framework (simulation and optimization) capturing operational, clinical, and behavioral features in order to evaluate the impact of EID network design on sample TATs and the number of infants initiated on treatment. The optimization model generates alternate configurations of the EID network, and the simulation model evaluates the health impact of these configurations. Applying this framework to the EID network in Mozambique, we are able to validate the predictive accuracy of our simulation model for that country and find that the relatively simple intervention of optimally re-assigning clinics to labs should increase the number of infected infants starting treatment by almost 4%. This potential impact is driven improving the likelihood of results being received before the next monthly visit of caretakers. The more drastic intervention of consolidating all current diagnostic capacity in a centralized lab located in Nampula

Figure 7 EID performance as a function of network transportation times.



Note. The top bar indicates the number of active lab locations in each OCA solution instance.

is furthermore predicted to increase the number of infected infants starting treatment by 7%. These results are driven by strong capacity pooling benefits that are specific to EID networks, because of batching delays at labs. They are also consistent with the recent adoption by Uganda of a single centralized lab for its EID network (Kiyaga et al. 2013).

With the goal of informing more general EID network design guidelines, we perform extensive sensitivity analysis to test the robustness of our recommendations. Our results indicate that the heuristic of consolidating capacity in a centralized lab remains optimal or near-optimal across all system utilization levels. In addition, we find that a centralized lab system is optimal for transportation times up to about 15% longer than currently observed in practice.

A limitation highlighted in §EC.9.1 is that our estimates of impact on the number of infected infants receiving treatment are quite sensitive to the behavioral sub-model assumed, and the field data available to estimate this sub-model is limited. While our predictions of impact on the number of treated infants should thus be qualified, a related observation is that our focus on this particular metric ignores the important health benefits associated with earlier treatment initiation (Violari et al. 2008), and therefore results in a conservative assessment of the health benefits associated with shorter TATs. Capturing the benefits of earlier treatment initiation would thus improve our understanding of EID network design decisions and likely increase their perceived importance, but we do not attempt this here. This is in part because the relevant models of disease progression and relationship with treatment timing are not yet as developed for infants as they are for adults.

From an operational perspective two opposing trends are currently observed for most HIV testing

modalities (UNITAID 2015). On the one hand, POC devices are being developed, enabling a more decentralized EID network structure, and the other, high-volume testing equipment is becoming increasingly efficient (UNITAID 2015) and cheap (UNAIDS 2015), making centralized labs even more cost-effective. In the long run, each country will most likely adopt a combination of efficient high-volume centralized labs and selectively deployed POC devices for EID testing. Therefore various aspects of EID systems warrant further analysis. At the lab level, it would be interesting to analyze the relative costs and benefits of alternative batching policies, possibly leveraging existing results on the optimal control of batch service queues (Deb and Serfozo 1973). At the network level, it would be interesting to combine the decision models from this paper with that of Deo and Sohoni (2015) to jointly optimize the existing laboratory network and POC device placement. We believe that the quantitative framework presented here may constitute a useful tool when considering these research questions.

Endnotes

1. The EID system in Mozambique has experienced minor changes since this dataset was acquired. In particular, the overall annual sample volume has increased moderately and a fifth laboratory was opened. As of writing, limited field deployment of point-of-care testing devices has recently begun. However, in the short term the proportion of demand handled by these devices is anticipated to be limited and in the long term at most 50%, so lab-based PCR services continue to be an integral part of the EID system (Crea 2016, as well as discussion in §7). Furthermore, long TATs continue to be a challenge. Overall our study dataset therefore still appears to be relevant and representative of the key qualitative features of EID in Mozambique.

2. All optimization problems were solved to optimality using IBM® ILOG® CPLEX® Optimization Studio 12.5.1. Simulation experiments were performed using MathWorks® MATLAB® R2013a. P-values and 95% confidence intervals were calculated by paired t-tests based on 50 replications (Law 2007).

Acknowledgments

We thank the National Institute of Health in Mozambique and the Clinton Health Access Initiative for providing data and information regarding the problem setting, in particular Dr. Ilesh Jani, Jorge Quevedo, Lindy Crea, Trevor Peter, Marcel Andela, and Lara Vojnov. Nicos Savva and the participants of the Collaborative Academic/Practitioner Workshop on Operational Innovation (London, 2013), the MSOM Annual Meetings (New York 2012, Fontainebleau 2013), the INFORMS Healthcare Conference (Chicago 2013) and the Health Systems Optimization Workshop (Chicago 2014) provided many insightful comments. We are also grateful to the Department Editor Pinar Keskinocak, the Associate Editor and three Referees for many useful suggestions.

References

Albin, Susan L. 1982. On poisson approximations for superposition arrival processes in queues. *Management Science* **28**(2) 126–137.

Albin, Susan L. 1984. Approximating a point process by a renewal process, II: Superposition arrival processes to queues. *Operations Research* **32**(5) 1133–1162.

Baron, O., O. Berman, D. Krass. 2008. Facility location with stochastic demand and constraints on waiting time. *Manufacturing and Service Operations Management* **10**(3) 484–505.

Becquet, Renaud, Milly Marston, François Dabis, Lawrence H Moulton, Glenda Gray, Hoosen M Coovadia, Max Essex, Didier K Ekouevi, Debra Jackson, Anna Coutsoudis, et al. 2012. Children who acquire hiv infection perinatally are at higher risk of early death than those acquiring infection through breastmilk: a meta-analysis. *PloS one* **7**(2) e28510.

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

Brimberg, J., A. Mehrez. 1997. A note on the allocation of queuing facilities using a minisum criterion. *Journal of the Operational Research Society* 195–201.

Brimberg, J., A. Mehrez, G. Wesolowsky. 1997. Allocation of queuing facilities using a minimax criterion. *Location Science* **5**(2) 89–101.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, L. Zhao. 2005. Statistical analysis of a telephone call center. *Journal of the American Statistical Association* **100**(469) 36–50.

Cameron, A. Colin, Pravin K. Trivedi. 2013. *Regression analysis of count data*. 53, Cambridge university press.

CHAI. 2013. 3 PCR Pilot Cohort Data. Obtained from the Clinton Health Access Initiative (CHAI) via personal communication with Jorge Quevedo.

Chao, Xiuli, Liming Liu, Shaohui Zheng. 2003. Resource allocation in multisite service systems with intersite customer flows. *Management Science* **49**(12) 1739–1752.

Chatterjee, A., S. Tripathi, R. Gass, N. Hamunime, S. Panha, C. Kiyaga, A. Wade, M. Barnhart, C. Luo, R. Ekpini. 2011. 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.

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

Codato, Gianni, Matteo Fischetti. 2006. Combinatorial Benders' cuts for mixed-integer linear programming. *Operations Research* **54**(4) 756–766.

Crea, Lindy. 2016. Personal communication. Project meeting on December 9th.

Creek, T.L., G.G. Sherman, J. Nkengasong, L. Lu, T. Finkbeiner, M.G. Fowler, E. Rivadeneira, N. Shaffer. 2007. Infant human immunodeficiency virus diagnosis in resource-limited settings: Issues, technologies, and country experiences. *American journal of obstetrics and gynecology* **197**(3) S64–S71.

Dabis, Francois et al. 1999. 6-month efficacy, tolerance, and acceptability of a short regimen of oral zidovudine to reduce vertical transmission of HIV in breastfed childre in Cote D'Ivoire and Burkina Faso: A double-blind placebo-controlled multicentre trial. *The Lancet* **353** 786–792.

Deb, Rajat K, Richard F Serfozo. 1973. Optimal control of batch service queues. *Advances in Applied Probability* 340–361.

Deo, Sarang, Charles J Corbett, S Mehta. 2016. Dynamic allocation of scarce resources under supply uncertainty. *Working paper, available at SSRN 1619408* .

Deo, Sarang, Jorge Quevado, Jonathan Lehe, Trevor Peter, Ilesh Jani. 2014. Expedited results delivery systems using SMS technology significantly reduce Early Infant Diagnosis test turnaround times. *Journal of Acquired Immune Deficiency Syndrome (Forthcoming)* .

Deo, Sarang, Milind Sohoni. 2015. Optimal decentralization of early infant diagnosis of hiv in resource-limited settings. *Manufacturing & Service Operations Management* **17**(2) 191–207.

Gallien, J., Z. Leung, P. Yadav. 2014. Rationality and Transparency in the Distribution of Essential Drugs in Sub-Saharn Africa: Analysis and Design of an Inventory Control System for Zambia. Working paper.

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

Kieffer, M., C. Chouraya, B. Lukhele, M. Adler. 2009. Implementing early antiretroviral treatment for infants in Swaziland. In conference: The 2009 HIV/AIDS Implementers' Meeting, Windhoek, Namibia.

Kiyaga, Charles, Hakim Sendagire, Eleanor Joseph, Ian McConnell, Jeff Grosz, Vijay Narayan, Godfrey Esiru, Peter Elyanu, Zainab Akol, Wilford Kirungi, et al. 2013. Uganda's new national laboratory sample transport system: A successful model for improving access to diagnostic services for early infant hiv diagnosis and other programs. *PloS one* **8**(11) e78609.

Kourtis, Athena P., Francis K. Lee, Elaine J. Abrams, Denise J. Jamieson, Marc Bylterys. 2006. Mother-to-child transmission of HIV-1: Timing and implications for prevention. *The Lancet Infectious Diseases* **6** 726–732.

Kraiselburd, S., P. Yadav. 2011. Supply chains and global health: An imperative for bringing operations management scholarship into action. *Production and Operations Management* .

Latigo-Mugambi, Melissa, Sarang Deo, Addeodata Kekiitinwa, Charles Kiyaga, Mendel Singer. 2013. Do diagnosis delays impact receipt of test results? Evidence from the HIV early infant diagnosis program in Uganda. *PLoS One (Forthcoming)* .

Law, Averill M. 2007. *Simulation modeling and analysis - International edition*. McGraw Hill.

Leung, Z., A. Chen, P. Yadav, J. Gallien. 2014. The impact of inventory management on stock-outs of essential drugs in Sub-Saharan Africa: Secondary analysis of a field experiment in Zambia. Working paper.

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

Marianov, V., T.B. Boffey, R.D. Galvão. 2008. Optimal location of multi-server congestible facilities operating as $M/Er/m/N$ queues. *Journal of the Operational Research Society* **60**(5) 674–684.

Marianov, V., D. Serra. 1998. Probabilistic, maximal covering location-allocation models for congested systems. *Journal of Regional Science* **38**(3) 401–424.

Marianov, V., D. Serra. 2002. Location - allocation of multiple-server service centers with constrained queues or waiting times. *Annals of Operations Research* **111**(1) 35–50.

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

McCoy, Jessica H, M Eric Johnson. 2014. Clinic capacity management: Planning treatment programs that incorporate adherence. *Production and Operations Management* **23**(1) 1–18.

Natarajan, Karthik V, Jayashankar M Swaminathan. 2014. Inventory management in humanitarian operations: Impact of amount, schedule, and uncertainty in funding. *Manufacturing & Service Operations Management* **16**(4) 595–603.

National AIDS Council. 2008. *Progress report for the United Nations General Assembly Special Session on HIV and AIDS - for the period 2006-2007*. Republic of Mozambique - National AIDS Council, Mozambique.

National AIDS Council. 2010. *Progress report for the United Nations General Assembly Special Session on HIV and AIDS - for the period 2006-2007*. Republic of Mozambique - National AIDS Council, Mozambique.

National AIDS Council. 2012. *Global AIDS response progress report for the period 2010-2011*. Republic of Mozambique - National AIDS Council, Mozambique.

Newell, Marie-Louise, Coovadia Noosen, Marjo Cortina-Borja, Nigel Rollins, Philippe Gaillard, Francois Dabis. 2004. Mortality of infected and uninfected infants born to HIV-infected mothers in Africa: a pooled analysis. *The Lancet* **364** 1236–1243.

Parvin, Hoda, Shervin Ahmad-Beygi, Jonathan Helm, Mark Van Oyen, Peter Larson. 2014. Malaria Treatment Distribution in developing world health systems and application to Malawi. Working paper obtained from Jonathan Helm.

Quevedo, Jorge, Lindy Crea. 2013. Personal communication. Project meeting on May 9th.

Rashkova, Iva, Yadav Prashant, Rifat Atun Atun, Jérémie Gallien. 2016. National drug stockout risks in Africa: The global fund disbursement process, 2002-2013. *Production and Operations Management* Forthcoming.

Sakasegawa, H. 1977. An approximation formula $l_q \simeq \alpha \cdot \rho^{\beta}/(1-\rho)$. *Annals of the Institute of Statistical Mathematics* **29**(1) 67–75.

Smit, Pieter W, Kimberly A Sollis, Susan Fiscus, Nathan Ford, Marco Vitoria, Shaffiq Essajee, David Barnett, Ben Cheng, Suzanne M Crowe, Thomas Denny, et al. 2014. Systematic review of the use of dried blood spots for monitoring hiv viral load and for early infant diagnosis. *PLoS One* **9**(3) e86461.

Taylor, T, Wenqiang Xiao. 2014. Subsidizing the distribution channel: Donor funding to improve the availability of malaria drugs. *Management Science* Forthcoming.

UNAIDS. 2015. Breakthrough global agreement sharply lowers price of early infant diagnosis of hiv. *www.unaids.org* URL `http://www.unaids.org/en/resources/presscentre/pressreleaseandstatementarchive/2015/july/20150719_eid_pressrelease`.

UNAIDS. 2016. *AIDS by the numbers*. Joint United Nations Programme on HIV/AIDS (UNAIDS).

UNAIDS, United Nations. 2013. Country report - Mozambique. URL `http://www.unaids.org/en/regionscountries/countries/mozambique/`.

UNITAID. 2015. *HIV/AIDS Diagnostics Technology Landscape, 5th edition*. World Health Organization.

Violari, A., M.F. Cotton, D.M. Gibb, A.G. Babiker, J. Steyn, S.A. Madhi, P. Jean-Philippe, J.A. McIntyre. 2008. Early antiretroviral therapy and mortality among hiv-infected infants. *New England Journal of Medicine* **359**(21) 2233–2244.

Vledder, M., P. Yadav, Sjoblom M., Brown T. 2013. Optimal supply chain structure for distributing essential drugs in low income countries: Results from a randomized experiment. Working Paper, The World Bank, Washington, DC.

WHO. 2013. *Consolidated guidelines on the use of antiretroviral drugs for treating and preventing HIV infection: Recommendations for a public health approach*. World Health Organization, Geneva.

WHO. 2014. *Global update on health sector response to HIV, 2014*. World Health Organization, Geneva.

Wiktor, Stefan Z. et al. 1999. Shourt-course oral zidovudine for prevention of mother-to-child transmission of HIV-1 in Abidjan, Cote d'Ivoire: A randomised trial. *The Lancet* **353** 781–785.

Wooldridge, Jeffrey M. 2002. *Econometric analysis of cross section and panel data*. The MIT press.

Yang, XQ, CJ Goh. 1997. A method for convex curve approximation. *European Journal of Operational Research* **97**(1) 205–212.

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

E-companion

Improving HIV early infant diagnosis supply chains in sub-Saharan Africa: Models and application to Mozambique

Jónas Oddur Jónasson
MIT Sloan School of Management, 30 Memorial Drive, 02142 Cambridge, MA.
joj@mit.edu
Sarang Deo
Indian School of Business, Gachibowli, Hyderabad, India, 50032.
sarang_deo@isb.edu
Jérémie Gallien
London Business School, Regents Park, London NW1 4SA, UK.
jgallien@london.edu

EC.1. Glossary of parameters

Table EC.1 lists all the main variables of our models along with their description.

EC.2. Mozambique EID system in more detail

HIV/AIDS is a significant public health problem in Mozambique. In 2012, the HIV prevalence in adults was estimated at about 11.1%, 1.6 million people were living with HIV, and 77,000 people lost their lives due to AIDS (UNAIDS 2013). The EID program was initiated in 2006 by the Ministry of Health (National AIDS Council 2008) and gradually scaled up to test about 25,000 infants in 2010 and 35,000 infants in 2011 (National AIDS Council 2012). Initially only two laboratories were equipped to perform PCR testing, one in the city of Maputo (region of Maputo) and another in the northern city of Nampula (region of Nampula). Two additional laboratories were added in the cities of Beira (Sofala region) and Quelimane (Zambezia region) in 2010 (National AIDS Council 2010).

As a result the Mozambique EID network currently comprises four labs performing virologic testing for over 400 clinics spread over 11 regions and 128 districts. The labs in Nampula, Quelimane and Beira have a single automatic machine each, while that in Maputo has one automatic machine as well as an older manual one. Both machine types use the same standard batch size, but it is theoretically possible to process half batches with a manual machine.

Table EC.1 List of parameters

Operational model development

Parameter	*Description*
$\mathcal{I} = \{1...I\}$	Set of clinics, which are indexed by i.
$\mathcal{J} = \{1...J\}$	Set of labs, which are indexed by j.
$z_{ij} \in \{0,1\}$	Main decision variable, indicating the assignment of clinic i to lab j.
$\mathcal{I}_j = \{i : z_{ij} = 1\}$	Set of clinics assigned to lab j.
TAT_i	Random turnaround time (see Figure 2) for clinic i.
$C_{i,pre}$	Random delay due to the wait for a transportation opportunity at clinic i.
T_{ij}	Random transportation time from clinic i to lab j.
LCT_j	Random lab-cycle-time for lab j.
$C_{i,post}$	Deterministic administrative delay at the clinic once results are back i.
π_i	Mean parameter for the Bernoulli part of the zero-inflated Poisson (ZIP) distributed arrival rate.
λ_i	Rate parameter for the Poisson part of the zero-inflated Poisson (ZIP) distributed arrival rate.
B_j	Batch-formation delay at lab j.
W_j	Sojourn time of a fully formed batch at lab j.
P_j	Post-processing delay at lab j.
b	Fixed batch size at all labs.
X_i	Random dispatch batch-size of samples from clinic i.
$f(S_j)$	Erlang distribution of service times at lab j.
k_j	Shape parameter for $f(S_j)$.
μ_j	Scale parameter for $f(S_j)$.

Public health modeling

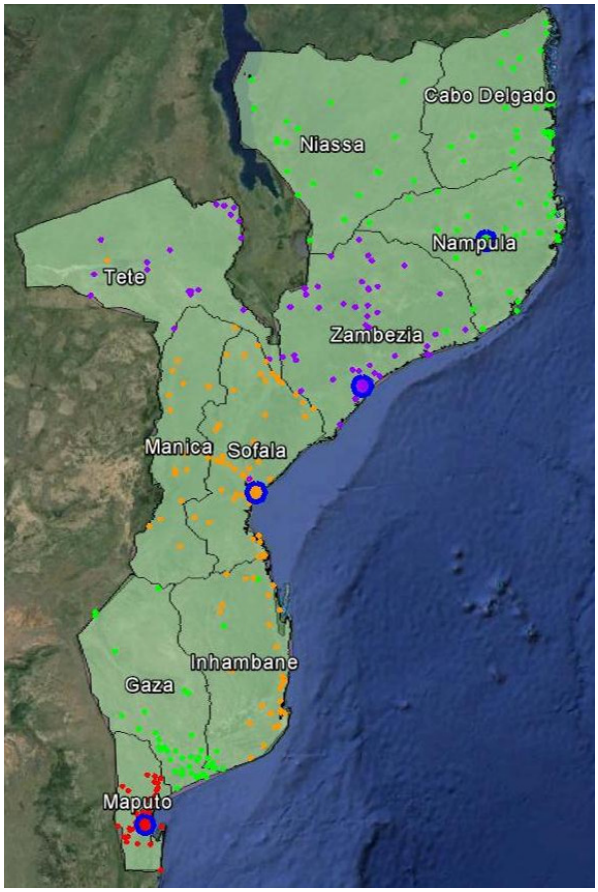
Parameter	*Description*
$Q_{i,P}$	PMTCT participation rate at clinic i.
$Q_{i,NP}$	PMTCT non-participation rate at clinic i.
V_P	Vertical HIV transmission rate for PMTCT participants.
V_{NP}	Vertical HIV transmission rate for PMTCT non-participants.
ζ	Time since infection, at the time of testing.
η	Communication delay (wait for caretakers).
υ	Follow-up rate from result receipt to treatment initiation.
δ	Delay from collection of results to treatment initiation.
$F_P(\zeta, TAT_i, \eta, \delta)$	Density of treatment initiation for PMTCT participants, as a function of delays.
$F_{NP}(\zeta, TAT_i, \eta, \delta)$	Density of treatment initiation for PMTCT non-participants, as a function of delays.
$\mathbb{E}[N_i]$	Expected number of infected infants treated as a result of visiting clinic i.

Additional variables for optimization models

Parameter	*Description*
$\mathbb{E}[M_j]$	Average total number of samples waiting for testing or being processed in an EID system.
ρ_j	Utilization of lab j.
c_j	Number of PCR servers/machines at lab j. (Parameter in OLA, decision variable in OCA.)
ϵ	Stability parameter for optimization models.
$d \in \{1, 2, .., D\}$	Days passed since sample collection (where D is some sensible upper bound).
$m \in \{1, 2, .., M\}$	Months passed since sample collection (where M is some sensible upper bound).
ω_{im}	Expected number of infants starting treatment if results become available in month m at clinic i.
ϕ_{dm}	Proportion of results expected to become available in month m if $\lceil \mathbb{E}(TAT) \rceil$ is d.
$z_{ijd} \in \{0,1\}$	Expanded definition of decision variable z_{ij}, indicating an expected TAT of d days.
$\tau_{jd} \in \{0,1\}$	Indicator for $\lceil \mathbb{E}(LCT) \rceil$ of lab j being d days.
α_{ju}	Slopes of a set $\mathcal{U}$ of tangents approximating $\mathbb{E}[LCT_j]$ (see §EC.6.3).
β_{ju}	Intercepts of a set $\mathcal{U}$ of tangents approximating $\mathbb{E}[LCT_j]$ (see §EC.6.3).
Δ_{ij}	Integer approximation of all the non-lab related delays associated with clinic i being served by lab j.
K	Total number of PCR testing machines available nationally.
$\mathcal{L}$	Set of available lab locations.
$y_j \in \{0,1\}$	Capacity allocation decision variable for OCA problem.

The structure of the EID program in Mozambique (Figure 2) and its practices closely match those described in §3.1. The assignment of clinics to labs is based on the boundaries of Mozambique's administrative regions as shown in Figure EC.1. Thus, the Maputo lab serves its own region; the Beira lab serves the Sofala, Inhambane and Manica regions; the Qelimane lab serves the Zambezia

Figure EC.1 The Mozambique EID system



Note. The eleven regions of Mozambique, the clinics (smaller circles), and the laboratories (larger circles). Each colored circle is an EID clinic which collects blood samples. The four labs are marked as a larger shapes surrounded by a blue line. The color of each clinic corresponds to the lab to which its blood samples are sent for diagnosis. (Map: Google Earth.)

and Tete regions; and the Nampula lab serves the three northernmost regions as well as the Gaza region. Generally samples are driven by car from clinics to district and regional headquarters, and from there all collected samples are flown by an air courier company to the city of their respective lab. The only exception is Gaza, located in the southern part of the country. Specifically, all samples from Gaza are first driven to Maputo, then flown to Nampula for processing. This is necessitated by lack of capacity in the Maputo lab, but availability of excess capacity in the Nampula lab, and further facilitated by availability of reliable air connection between Maputo and Nampula.

EC.3. Mozambique dataset

The dataset includes key patient characteristics such as age, gender, HIV status and whether the mother is a PMTCT program participant. It also includes time stamps for six of the twelve epochs shown in Figure 2, specifically the date of birth (e_1), the date of sample collection (e_3), the date

Table EC.2 Summary statistics for the Mozambique EID clinics

	Mean	Std. Dev.	Min	Max
Daily arrival rate of infants to clinic	0.24	0.32	0.003	2.37
Average pre-processing clinic delay (days)	2.2	2.6	0.012	20.5
Average number of samples per dispatch to lab	3.5	2.2	1.0	15.0
Average transportation delay from clinic to lab (days)	7.2	5.6	0.0	26.7

Table EC.3 Summary statistics for the Mozambique EID labs

Lab	Daily arrival rate of blood samples		Average LCT (days)		Average TAT (days)	
	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.
Lab 1 - Maputo	28	30	39	29	48	32
Lab 2 - Nampula	28	50	14	8	32	19
Lab 3 - Quelimane	17	25	13	10	28	19
Lab 4 - Beira	23	38	31	19	50	28
National average	24	37	25	22	40	27

of dispatch to the lab (e_4), the date of arrival at the lab (e_5), the date of processing (e_8), and the date of results being sent back to the clinic via the SMS printer (e_9).

Summary statistics for the clinics and labs are included in Tables EC.2 and EC.3, respectively.

EC.4. Additional analyses for simulation model parameter-fitting

This section describes in more detail the parameter fitting methodology introduced in §4.

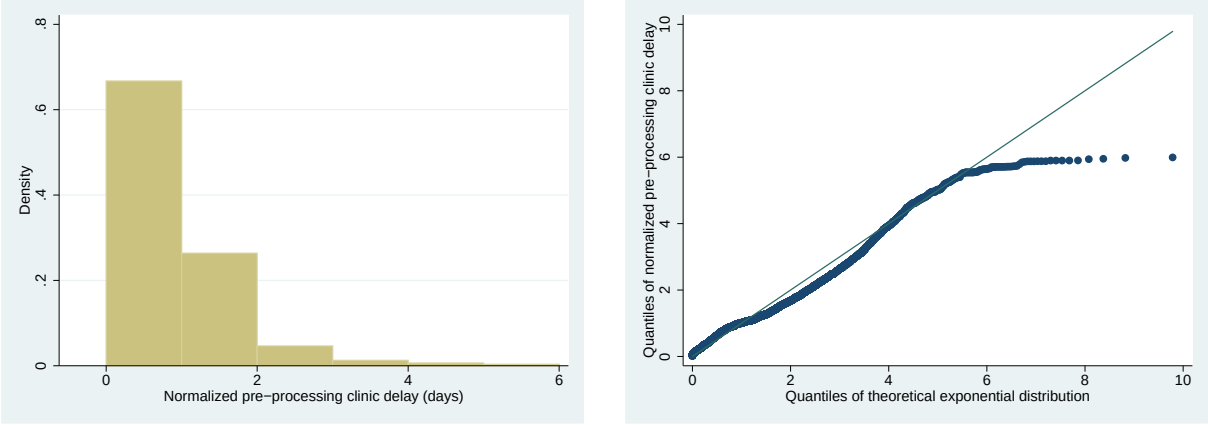
EC.4.1. Distribution of pre-processing clinic delays

In subsection 4.1.1 we discuss the modeling of pre-processing clinic delays. In this subsection we investigate the model assumption that pre-processing clinic delays follow an exponential distribution. To that end we normalize the historical clinic delays at each one of the 410 clinics by dividing the original data for each clinic by its mean, remove the 1% longest delays as outliers, and consider the resulting aggregated dataset across all clinics. Figure EC.2 contains a histogram of the resulting empirical distribution, and a Q-Q plot comparing that empirical distribution to an exponential distribution with the same mean.

The histogram seen in Figure EC.2 shows that the distribution of empirical data does share some important features with the exponential distribution, in particular a decreasing density function. In addition, the Q-Q plot reveals that over a large range (lower and medium values of the support), the fractiles of the empirical distribution fit the exponential quite well. However, the exponential has slightly more weight in the right tail, which we may interpret as possible specific actions by the clinic staff to generate a transportation opportunity faster when samples have been waiting for more than a certain time threshold larger than the mean of the distribution. We note that

Figure EC.2 Histogram of the empirical distribution of normalized pre-processing clinic delays and Q-Q plot of that empirical distribution against the exponential distribution with the same mean.

(a) (b)



goodness-of-fit tests, such as Chi Squared or Kolmogorov –Smirnov, reject the null hypothesis that the data distribution is equal to the exponential distribution. This is because the former is very sensitive to large sample sizes, and the latter calculates the largest difference between the CDFs and is thus affected by the differences in the tail probabilities. This is partly why we performed an extensive out-of-sample validation of our simulation model in order to assess its predictive accuracy with respect to the key performance/outcome measure of sample TATs (see §4.3). This exercise and its relatively positive results provided further support for this model assumption.

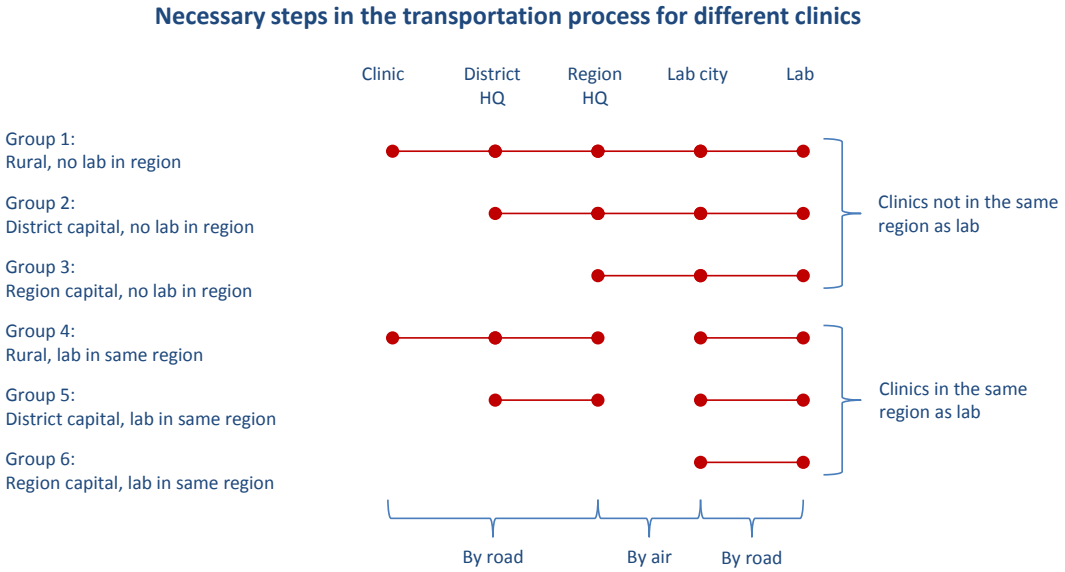
EC.4.2. Transportation model regression output

In order to estimate transportation times we divide clinics into 6 groups based on their location (rural area, district capital or a provincial capital) and whether they are in the same region as the lab they are assigned to. This is a result of our extensive conversations with Quevedo and Crea (2013) who pointed out that, as shown in Figure EC.3, the number of steps required to transport a sample from a clinic to a lab varies depending on the type of clinic. As an example, clinics in Group 6 require only one transportation step, from the clinic directly to the lab. On the other hand, clinics in Group 1 require 4 transportation steps, from the clinic to the district headquarters, from there to the regional headquarters, from which they are flown to the lab city and driven to the lab. As a result we expect transportation delays for clinics in Group 6 to be shorter than those for clinics in Group 1.

We estimate the following Tobit model (Wooldridge 2002) using data on transportation times corresponding to the current EID network configuration:

$$T_{ij} = max(0, \hat{T}_{ij}), \text{ where} \tag{EC.1}$$

Figure EC.3 The transportation steps necessary for each clinic group



Note. The dots represent the locations where samples are gathered on their way to the lab. The links indicate the necessity of transporting the samples between the given nodes, for each group.

$$\hat{T}_{ij} = \beta_0 + \sum_g \beta_{1,g} G_g + \sum_i \beta_{2,i} H_i + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2), \tag{EC.2}$$

where G_g denotes a binary indicator for whether a sample belongs to a clinic classified in group $g \in \{1, 2, ..., 6\}$, and H_i is a binary indicator associated with each clinic $i \in \mathcal{I}$. The H_i variables are designed to capture clinic fixed effects, or unobservable factors associated with individual clinics that might affect transportation delays, e.g. some clinics may have more motivated staff or the support of an NGO for sample transport. The Tobit model is adapted to the fact that transportation times are strictly positive with a probability mass on zero in our dataset, and our choice of variables reflects the input of field managers. The resulting parameter estimates (calculated using the *tobit* function in Stata/IC® 12.1) for (EC.1) and (EC.2) are listed in table EC.4.

The groups in the table correspond to those discussed above and represented in Figure EC.3. The reason the regression includes 9 groups as opposed to the 6 in the figure is that we allow the coefficients for groups 1, 2, 3 in the region of Gaza to differ from the rest. This is due to the Gaza samples not being transported by air, as is generally the rule, but by car to Maputo, from where they are flown to their destination. By conducting joint F tests on the $\beta_{1,g}$ variables on the one hand and the $\beta_{2,i}$ variables on the other we conclude that each set is jointly significant at the 0.1%.

Finally, the distributions of transportation times corresponding to clinic-lab assignments differing from the current one are then simulated using (EC.1) and (EC.2). We use the parameters β_0, $\beta_{1,g}$ and $\beta_{2,i}$ estimated in-sample from the current assignment, but update the clinic fixed effect and transportation group variables to reflect any new assignment considered.

Table EC.4 Estimates for regression (EC.2)

Independent variables	*Dependent variable:* Transportation time
Group	
2	4.166
	(3.173)
3	.750
	(3.312)
4	−14.945***
	(3.252)
5	−4.620
	(3.382)
6	−12.929***
	(3.173)
7 (Gaza)	−.667
	(3.536)
8 (Gaza)	−.389
	(3.166)
9 (Gaza)	−.994
	(3.169)
Clinic fixed effects	*Included*
Observations	32969
Pseudo R^2	0.1253

Group 1 is the benchmark group.
∗, ∗∗, and ∗ ∗ ∗ denote significance at the 5%, 1%, and 0.1% level, respectively.

EC.4.2.1. Lab processing distribution To fit k_j and μ_j we split the Mozambique dataset into a training dataset (the first 9 months of 2011) and a validation dataset (final 3 months of 2011). Using the actual arrival batches to each of the labs, as well as their respective post-processing delay distributions, we simulate LCT_j for a range of values of the two parameters. We conduct 20 replications using each parameter combination with a warm-up period of 3 months. We compare the means and standard deviations of the simulated LCT_j values to those in the data, in terms of mean squared errors (MSEs). For each lab we choose the 10 best parameter combinations and simulate the system again for the validation period. We finally choose the parameter combination that best fits the validation dataset (see table EC.5). While the historical capacity utilization resulting from this estimation procedure was found to be strictly less than 1 for the labs in Nampula, Quelimane and Beira, the corresponding estimate for Maputo was approximately equal to 1 (1.00067). We note that some qualitative features of the historical data for that lab (occasionally long lab cycle times during the year, slight accumulation of unprocessed samples at the end of the year) were consistent with that estimate. Furthermore, by design our simulation experiments consider transient system behaviour in the course of a year and do not attempt to reproduce or measure steady state. Finally, while our optimization models do rely on the assumption that all processing facilities considered are stable because their formulations involve steady-state queueing formulas, we only use that model for generating improved clinic assignment and capacity allocation solutions, as opposed to estimating the performance of either existing or proposed solutions (for which we use the simulation model). Indeed, for all the scenarios we consider our optimization models do indeed identify more

desirable solutions where the clinic assignment is such that all the labs are stable.

Table 1 in the main text shows that the fitted model predicts cycle times in the validation period fairly accurately for Nampula, Quelimanne and Beira, even though they are very different from those in the training period. The higher discrepancy for the Maputo lab is explained by a specific operational reason. In particular, the annual EID program report for the year 2011 reveals that this particular laboratory was short staffed for a brief period towards the end of 2011 due to exceptional staff turnover, which resulted in temporarily longer TATs. Since this was a non-systematic and isolated effect, we decided to not change our model as a result of this discrepancy.

Table EC.5 Properties of the service time distributions for each lab

Lab	Shape	Scale	Mean	Variance	SCV
Lab 1 - Maputo	5	1.3	6.5 days	8.5	0.2
Lab 2 - Nampula	2	1.4	2.8 days	3.9	0.5
Lab 3 - Quelimane	3	1.3	3.9 days	5.1	0.3
Lab 4 - Beira	3	1.3	3.9 days	5.1	0.3

EC.4.2.2. Post-processing at lab Implicit to the discussion of lab post-processing times in §4.1.3 is that we assume post-processing delay distribution to be independent from the sample load assigned to the lab, i.e., a function of general lab practices as opposed to decision variables. This modeling choice is partly justified by the lab survey findings, which surfaced the key role of the lab manager and the lab computer with respect to post-processing delays. Because we perceive these resources to be less affected by the current sample load of the lab than the pre-processing and processing delays, this lead to the decision of modeling post-processing delays as exogenous random variables, meaning that while they are indeed variable and random in our model as described above, they are not affected by the optimization decisions (capacity allocation and assignment of clinics to labs). In contrast, the pre-processing delays in our model are endogenous, in the sense that they are affected by the network design decisions considered, through their impact on lab sample workload and the dynamics of the queueing model described in section §4.1.3. Therefore, even though the service time itself is assumed to be exogenous, the sum of pre-processing delays and service time $e_8 - e_5$ observed as part of the data is considered endogenous.

These assumptions regarding which lab delay components depend on the decision variables (endogenous components) and which ones do not (exogenous components) can actually be somewhat validated a posteriori from data. We have examined the historical correlation between lab load (defined as the number of samples already at the lab at the time of arrival of a given sample) and the two components of lab delays experienced by each sample and available in our dataset, namely the sum of the pre-processing delay and service time ($e_8 - e_5$) and post-processing delays

$(e_9 - e_8)$. For this analysis we excluded the first month of the year as we do not know what the sample spill over was from the previous year, and we removed outliers by deleting the longest 2% of delays. In addition, we also calculated the correlation between the daily and weekly averages of the random components considered. The results we obtained for each lab are shown in Table EC.6.

Table EC.6 Correlation between load and lab delays.

	Pre-processing & processing times ($e_5 \to e_8$)	Post-processing times ($e_8 \to e_9$)
National		
Each sample	0.57	0.22
Daily averages	0.61	0.24
Weekly averages	0.66	0.26
Lab 1 - Maputo		
Each sample	0.26	-0.15
Daily averages	0.29	-0.21
Weekly averages	0.30	-0.30
Lab 2 - Nampula		
Each sample	0.49	0.11
Daily averages	0.49	0.19
Weekly averages	0.57	0.22
Lab 3 - Quelimane		
Each sample	0.47	0.27
Daily averages	0.54	0.33
Weekly averages	0.67	0.32
Lab 4 - Beira		
Each sample	0.34	-0.04
Daily averages	0.31	-0.13
Weekly averages	0.42	-0.16

Notes: Lab load is defined as the total number of samples already at the lab at the time of arrival for a given sample. The epochs e_5, e_8, and e_9 refer to the timeline of Figure 2.

The data demonstrates that the pre-processing and processing times are more highly correlated with the number of samples at the lab than the post-processing delays (regardless of whether one aggregates at the weekly or daily level, or not at all). Specifically, the absolute values of the estimates of correlation with lab load are systematically higher in each lab for the pre-processing and service times than for the post-processing delays. In the labs of Maputo and Beira, we even observe slightly negative correlations with the post-processing delays, suggesting that these delays may actually decrease with load, perhaps because a high load of sample may attract some additional attention from the lab manager in charge of post-processing. This effect is not consistent across labs however. In contrast, we see stronger and positive correlation between load and pre-processing and

service times, which seems to partly justify the dependence of this delay on the decision variables in our model.

EC.5. Proof of Proposition 1

EC.5.1. Proof outline

As mentioned in 5.1 we denote the average total number of samples waiting for testing or being tested at an EID lab, excluding those in post-processing, by

$$\mathbb{E}[M_j] = \mathbb{E}[Z_j^{\infty}] + b(\mathbb{E}[Q]_{GI/G/c_j} + c_j\rho_j), \tag{EC.3}$$

where $\mathbb{E}[Z_j^{\infty}]$ represents the average number of samples not yet a part of a fully formed batch, $\mathbb{E}[Q]_{GI/G/c_j}$ denotes the average number of fully formed batches waiting to be processed, ρ_j denotes utilization, and $c_j\rho_j$ denotes the average number of batches being processed.

We first evaluate the following approximation (eq. 14 in Hanschke (2006)) for the squared coefficient of variation of the interarrival time between fully formed batches, $SCV[Y_j^*]$, for the case of an EID lab;

$$SCV[Y_j^*] \approx \frac{\mathbb{E}[X_j]}{b}(SCV[X_j] + SCV[Y_j]). \tag{EC.4}$$

Second, we substitute this approximation into Sakasegawa's approximation for $\mathbb{E}[Q]_{GI/G/c_j}$, the average queue length in a queue with general arrivals, general service times, and c_j servers;

$$\mathbb{E}[Q]_{GI/G/c_j}(\rho_j, SCV[Y_j^*], SCV[S_j]) \approx \frac{\rho_j^{\sqrt{2(c_j+1)}}}{1-\rho_j} \times \frac{SCV[Y_j^*] + SCV[S_j]}{2}, \tag{EC.5}$$

where S_j is the distribution of processing time at each machine (see §4.1.3).

Third, we incorporate this approximation in Hanschke's approximation for the average number of customers in a $\sum_{i\in\mathcal{I}_j} GI^{[X_i]}/G^{(b,b)}/c_j$ queueing system (eq. 15 in Hanschke (2006));

$$\mathbb{E}[M_j] \approx \mathbb{E}[Z_j^{\infty}] + b\mathbb{E}[Q]_{GI/G/c_j}(\rho_j, SCV[Y_j^*], SCV[S_j]) + bc_j\rho_j. \tag{EC.6}$$

Finally, we apply Little's law to approximate the sojourn times in such a system, and add the expected post-processing delay.

$$\mathbb{E}[LCT_j] = \frac{\mathbb{E}[M_j]}{\sum_{i\in\mathcal{I}_j}\pi_i\lambda_i} + \mathbb{E}[P_j]. \tag{EC.7}$$

In deriving the approximation we make two further assumptions for analytical tractability. First, we assume that the coefficient of variation of the interarrival times of sample shipments arriving to each lab is 1, which seems reasonable since sample arrivals at the lab result from a superposition of a large number of arrival processes, each with exponential inter-arrival times (Albin 1982, 1984). Second, we assume that the average clinic batch size $\frac{\pi_i\lambda_i}{\gamma_i}$ is constant (noted E). This implies that the frequency of transportation opportunities at a given clinic is proportional to the arrival rate of infants, which seems reasonable based on EID program data.

EC.5.2. Proof details

(i) Denote the random number of samples for a given shipment from clinic i by X_i. At any given time the probability of a transportation opportunity arriving before the next sample is denoted by $\frac{\gamma_i}{\lambda_i\pi_i+\gamma_i}$. Hence, the batch size can be modeled as $geom(\frac{\gamma_i}{\lambda_i\pi_i+\gamma_i})$, i.e. the number of trials before a success, where success is the next arrival being a transportation opportunity. Then the expectation and variance of X_i can be written as;

$$\begin{aligned}\mathbb{E}[X_i] &= \frac{1-\frac{\gamma_i}{\gamma_i+\pi_i\lambda_i}}{\frac{\gamma_i}{\gamma_i+\pi_i\lambda_i}} \\ &= \frac{\pi_i\lambda_i}{\gamma_i}, \text{ and}\end{aligned}$$

$$\begin{aligned}\sigma^2[X_i] &= \frac{1-\frac{\gamma_i}{\gamma_i+\pi_i\lambda_i}}{\left(\frac{\gamma_i}{\gamma_i+\pi_i\lambda_i}\right)^2} \\ &= \frac{\pi_i\lambda_i(\gamma_i+\pi_i\lambda_i)}{\gamma_i^2}.\end{aligned}$$

(ii) Denote the random size of the sample shipments arriving at lab j by X_j. Since the lab receives sample shipments from a number of clinics, the size of the arriving shipments depends partly on the frequency of transportation opportunities from each clinic. That is, the distribution of X_j given by the distributional mixture $X_j \sim X_i$ with probability $\frac{\gamma_i}{\sum_{k\in\mathcal{I}_j}\gamma_k}$. Hence, at the lab level the properties of the arrival batch distribution are as follows:

$$\begin{aligned}\mathbb{E}[X_j] &= \sum_{i\in\mathcal{I}_j}\frac{\gamma_i}{\sum_{k\in\mathcal{I}_j}\gamma_k}\mathbb{E}[X_i] \\ &= \sum_{i\in\mathcal{I}_j}\frac{\gamma_i}{\sum_{k\in\mathcal{I}_j}\gamma_k}\frac{\pi_i\lambda_i}{\gamma_i} \\ &= \frac{\sum_{i\in\mathcal{I}_j}\pi_i\lambda_i}{\sum_{i\in\mathcal{I}_j}\gamma_i},\end{aligned}$$

$$\begin{aligned}\sigma^2[X_j] &= \mathbb{E}[X_j^2]-\mathbb{E}[X_j]^2 \\ &= \sum_{i\in\mathcal{I}_j}\frac{\gamma_i}{\sum_{k\in\mathcal{I}_j}\gamma_k}\mathbb{E}[X_i^2]-\mathbb{E}[X_j]^2 \\ &= \sum_{i\in\mathcal{I}_j}\frac{\gamma_i}{\sum_{k\in\mathcal{I}_j}\gamma_k}\left(\mathbb{E}[X_i]^2+\sigma^2[X_i]\right)-\left(\frac{\sum_{i\in\mathcal{I}_j}\pi_i\lambda_i}{\sum_{i\in\mathcal{I}_j}\gamma_i}\right)^2 \\ &= \frac{\sum_{i\in\mathcal{I}_j}\frac{\pi_i\lambda_i(2\pi_i\lambda_i+\gamma_i)}{\gamma_i}}{\sum_{i\in\mathcal{I}_j}\gamma_i}-\left(\frac{\sum_{i\in\mathcal{I}_j}\pi_i\lambda_i}{\sum_{i\in\mathcal{I}_j}\gamma_i}\right)^2.\end{aligned}$$

(iii) By assumption, the interarrival times of clinic shipments arriving at the lab, Y_j, is assumed to have the property that $SCV[Y_j] = 1$.

(iv) Hence the approximation given in (EC.4) becomes:

$$\begin{aligned} SCV[Y_j^*] &\approx \frac{\mathbb{E}[X_j]}{b}(SCV[X_j] + SCV[Y_j]) \\ &= \frac{\sum_{i\in\mathcal{I}_j}\pi_i\lambda_i}{b\sum_{i\in\mathcal{I}_j}\gamma_i}\left(\frac{\sigma^2[X]}{\mathbb{E}[X]^2}+1\right) \\ &= \frac{\sum_{i\in\mathcal{I}_j}\pi_i\lambda_i}{b\sum_{i\in\mathcal{I}_j}\gamma_i}\left(\frac{\left(\sum_{i\in\mathcal{I}_j}\gamma_i\right)\left(\sum_{i\in\mathcal{I}_j}\frac{\pi_i\lambda_i(2\pi_i\lambda_i+\gamma_i)}{\gamma_i}\right)}{\left(\sum_{i\in\mathcal{I}_j}\pi_i\lambda_i\right)^2}\right) \\ &= \frac{\sum_{i\in\mathcal{I}_j}\frac{\pi_i\lambda_i(2\pi_i\lambda_i+\gamma_i)}{\gamma_i}}{b\sum_{i\in\mathcal{I}_j}\pi_i\lambda_i} \end{aligned}$$

(v) Thus, substituting into the approximation for $\mathbb{E}[Q]_{GI/G/c_j}$, from (EC.5) we get:

$$\begin{aligned} \mathbb{E}[Q]_{GI/G/c_j}(\rho_j, SCV[Y_j^*], SCV[S_j]) &\approx \frac{\rho_j^{\sqrt{2(c_j+1)}}}{1-\rho_j}\times\frac{SCV[Y_j^*]+SCV[S_j]}{2} \\ &= \frac{\rho_j^{\sqrt{2(c_j+1)}}}{1-\rho_j}\left(\frac{\sum_{i\in\mathcal{I}_j}\frac{\pi_i\lambda_i(2\pi_i\lambda_i+\gamma_i)}{\gamma_i}}{2b\sum_{i\in\mathcal{I}_j}\pi_i\lambda_i}+\frac{SCV[S_j]}{2}\right). \quad \text{(EC.8)} \end{aligned}$$

(vi) Now consider the first term of (EC.6), $\mathbb{E}[Z_j^\infty]$, representing the average number of samples not yet a part of a fully formed batch. If we denote the number of samples waiting to form a processing batch at lab j and time t by $Z_j(t)$, we get (e.g., Hanschke (2006)):

$$\mathbb{E}[Z_j^\infty] = \lim_{t\to\infty}\mathbb{E}\big[Z_j(t)\big] = \frac{b-1}{2}. \quad \text{(EC.9)}$$

(vii) Substituting the results from (EC.8) and (EC.9) into (EC.6) yields:

$$\mathbb{E}[M_j] \approx \frac{b-1}{2}+\frac{\rho_j^{\sqrt{2(c_j+1)}}}{1-\rho_j}\left(\frac{\sum_{i\in\mathcal{I}_j}\frac{\pi_i\lambda_i(2\pi_i\lambda_i+\gamma_i)}{\gamma_i}}{2\sum_{i\in\mathcal{I}_j}\pi_i\lambda_i}+\frac{bSCV[S_j]}{2}\right)+bc_j\rho_j$$

$$= \frac{b-1}{2} + \frac{\rho_j^{\sqrt{2(c_j+1)}}}{1-\rho_j}\left(\frac{\sum_{i\in\mathcal{I}_j}\frac{(\pi_i\lambda_i)^2}{\gamma_i}}{\sum_{i\in\mathcal{I}_j}\pi_i\lambda_i} + \frac{1}{2} + \frac{bSCV[S_j]}{2}\right) + bc_j\rho_j$$

(viii) With the additional assumption that transportation resources γ_i are distributed so that the average clinic batch size $\frac{\pi_i\lambda_i}{\gamma_i} = E$ is about constant for all i, then the expression above becomes:

$$\mathbb{E}[M_j] \approx \frac{b-1}{2} + \frac{\rho_j^{\sqrt{2(c_j+1)}}}{1-\rho_j}\left(E + \frac{1}{2} + \frac{bSCV[S_j]}{2}\right) + bc_j\rho_j. \tag{EC.10}$$

(ix) Finally, applying Little's law and adding the expected post processing delay yields an approximation of the LCT as a univariate function of utilization:

$$\mathbb{E}[LCT_j](\rho_j) \approx \frac{(b-1)\mathbb{E}[S_j]}{2bc_j\rho_j} + \frac{\mathbb{E}[S_j]\rho_j^{\sqrt{2(c_j+1)}-1}}{bc_j(1-\rho_j)}\left(E + \frac{1}{2} + \frac{b_jSCV[S_j]}{2}\right) + \mathbb{E}[S_j] + \mathbb{E}[P_j]. \tag{EC.11}$$

□

EC.6. Additional analysis for optimization models

EC.6.1. Derivation of linear reformulation (13)-(20)

First, consider the objective (6). Let indices $d \in \{1,2,..,D\}$ and $m \in \{1,2,..,M\}$ respectively represent the days and months passed since the time of sample collection (D and M represent some sensible upper bound on the possible length of TATs). Recall from §3.1 that caretakers are given follow-up appointments at monthly intervals. Hence, we define the expected follow-up rate conditional on results becoming available in the m-th month following sample collection as

$$\mathbb{E}[F_k^m] = \mathbb{E}[F_k(\zeta, TAT, \eta, \delta) \,|\, \lceil TAT/30\rceil = m] \qquad \forall k \in \{P, NP\}, m \in \{1,...,M\}. \tag{EC.12}$$

Denoting by ω_{im} the expected number of infants starting treatment if the results were to become available at clinic i in month m, this quantity can then be expressed as

$$\omega_{im} = \upsilon\pi_i\lambda_i \sum_{k\in\{P,NP\}} \left(Q_{i,k}V_k\mathbb{E}[F_k^m]\right). \tag{EC.13}$$

Further, denote by $\phi_{dm} = P[\lceil TAT\rceil = m | \lfloor\mathbb{E}[TAT]\rfloor = d]$ the proportion of results that are expected to become available in month m if the expected TAT is d days. The distributional family and parameters characterizing ϕ_{dm} can be estimated off-line from data (see e-companion EC.6.2), and

incorporated in the model as coefficients. We can also expand the definition of the main assignment decision variable z_{ij} as

$$z_{ijd} = \begin{cases} 1 & \text{if } z_{ij} = 1 \text{ and } \lfloor \mathbb{E}[TAT_i] \rfloor = d, \\ 0 & \text{otherwise.} \end{cases}.$$

Using the new definitions z_{ijd}, ϕ_{dm} and ω_{im}, the objective (6) of maximizing the number of infected infants receiving their results can be expressed as (13). Similarly, constraints (7) and (8) can be rewritten as (14) and (15).

Second, we reformulate constraint (10) by replacing the random sum in its right-hand-side by approximation (5). As this approximation of $\mathbb{E}[LCT_j]$ is a convex function of ρ_j, we can approximate it arbitrarily closely on $(0,1)$ by the upper envelope of a discrete set $\mathcal{U}$ of tangents $(\alpha_{ju}, \beta_{ju})_{u \in \mathcal{U}}$ with slopes α_{ju} and intercepts β_{ju} (see e-companion EC.6.3). Defining binary variables τ_{jd} to indicate whether the mean LCT of lab j is d days, constraint (10) can finally be replaced with (16) and (17).

Third, constraint (11) can now be replaced by arguments ensuring that $z_{ijd} = 0$ for all $d \leq \mathbb{E}[TAT_i]$, if clinic i were to be assigned to lab j. We let Δ_{ij} represent all the delays associated with clinic i being served by lab j, except the time spent at the lab, i.e. $\Delta_{ij} = \lceil \mathbb{E}(C_{i,pre}) + \mathbb{E}(T_{ij}) + \mathbb{E}(C_{i,post}) \rceil$. The required condition is then ensured by constraints (18) and (19). Specifically, (18) states that no assignment between clinic i and lab j can have a shorter expected TAT than Δ_{ij} days. Similarly (19) ensures that if a lab assignment results in a LCT of d days at lab j, then $\tau_{jd} = 1$ and the expected TAT for any clinic i assigned to lab j will be exactly $d + \Delta_{ij}$ days. This slightly unconventional approach of using index d in (18) and (19) to add LCT_j and Δ_{ij} improves computational tractability relative to a formulation with big M constraints, which tend to be computationally inefficient (Codato and Fischetti 2006).

With the transformation of the decision variables to z_{ijd} and the introduction of ω_{im} and ϕ_{dm}, the objective function (13) depends now only on $\mathbb{E}[TAT_i]$ instead of the random variable TAT_i. Thus, by calculating the expected LCT in (16) and (17) and adding it to the expected TAT in (18) and (19), we obtain the approximate linear reformulation of optimization problem (6)–(12) stated in (13)-(20).

EC.6.2. Calculating ϕ_{tb}

To incorporate the ϕ_{tb} function in the optimization formulation we must first assume a distributional family. We use maximum likelihood to fit 9 different two-parameter families of probability distributions to the Mozambique data for the TAT_i data for each clinic. The log-normal distribution has the best fit in terms of log likelihood numbers and is therefore chosen as the distributional family for TATs in the MIP formulation.

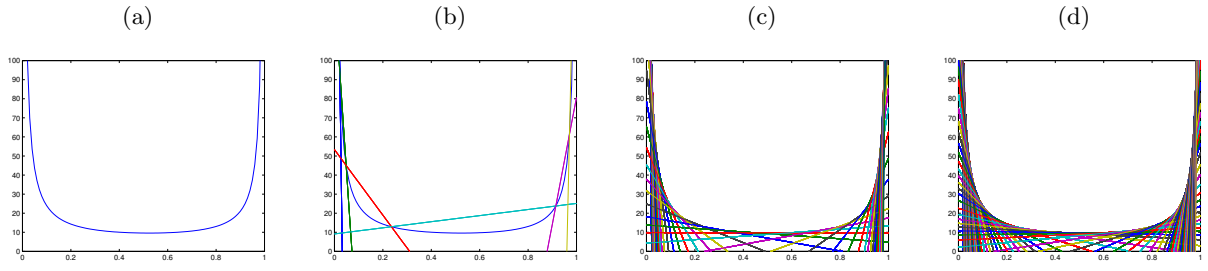
To be able to calculate the ϕ_{tb} function we now only require an approximation for the second moment of the distribution, as the first is determined in constraints of the OLA optimization problem in §5.2. The fitted distributions have a fairly constant second distributional parameter (SCV=0.11) and so we parsimoniously assume the second parameter to be independent of the first and constant. Note that the methodology can easily incorporate other approximations for the second moment. We considered both empirical and theoretical approaches but found this to be the simplest justifiable assumption.

EC.6.3. Calculating α_{ju} and β_{ju}

As discussed in §5.2, the estimate of the total LCT needs to be approximated by linear functions in order to formulate the problem as a MIP. In the optimization models derived in §5.2 and §5.3, the optimal solutions are achieved by strategically reducing LCT. In light of this a conservative approximation of the convex LCT function is one that approximates it using line segments between its functional values giving an upper bound, as opposed to using tangents (found by using the derivative of the function) which would give a lower bound.

The approximation algorithm we employ is based on the functional-values algorithm presented in Yang and Goh (1997). Figure EC.4 shows how increasing the number of iterations improves the approximation for typical EID lab with one server.

Figure EC.4 The lab cycle time approximation



Note. (a) the approximate waiting time function. (b) the linear segments after 5 iterations. (c) the linear segments after 43 iterations. (d) the linear segments after 96 iterations.

EC.6.4. Computational cuts

The MIP formulation (13)-(20) is still challenging to solve for realistically sized problems. In our computational experiments, for instance, we observe that continuous problem relaxations can have an optimal solution with some weight on both τ_{j1} and τ_{jD} (meaning that some samples will have expected LCTs of 1 day the rest will have expected LCTs of D days). Because such solutions are not

helpful in finding an optimal integer solution, we exclude them by introducing an additional binary variable τ'_{jd}, redefining τ_{jd} to be continuous between zero and one, and reformulating constraints (16) and (17) as

$$\sum_{d=1}^{D} \tau'_{jd} \geq \alpha_{ju}\rho_j + \beta_{ju} \qquad \forall j \in \mathcal{J},\, u \in \mathcal{U}, \tag{EC.14}$$

$$\tau'_{jd} \geq \tau'_{j(d+1)} \qquad \forall j \in \mathcal{J},\, d \in \{1...(D-1)\}, \tag{EC.15}$$

$$\tau_{jd} \leq \tau'_{jd} - \tau'_{j(d+1)} \qquad \forall j \in \mathcal{J},\, d \in \{1...(D-1)\}, \tag{EC.16}$$

$$\sum_{d=1}^{D} \tau_{jd} = 1 \qquad \forall j \in \mathcal{J}, \tag{EC.17}$$

$$\tau'_{jd} \in \{0,1\} \qquad \forall j \in \mathcal{J},\, d \in \{1...D\}. \tag{EC.18}$$

Here, (EC.15) ensures that each τ'_{jd} can only have a positive weight if $\tau'_{j(d-1)}$ also has a positive weight, and an integer solution will have $\tau'_{jd} = 1$ for all d lesser than or equal to the expected LCT of lab j. As a result, the expected LCT will be captured by τ_{jd} as before, due to constraint (EC.16). In our numerical experiments we indeed observe substantial reductions of computational time when replacing constraints (16) and (17) with (EC.14)–(EC.18) in the OLA formulation.

EC.7. Equity in OLA and OCA solutions

Because the optimization formulations derived in §5 do not explicitly capture the important notion of equity of access across patients or geographic locations, we have also conducted numerical experiments to evaluate the OLA and OCA solutions on this dimension. Following the framework described in Marsh and Schilling (1994), we specifically report in Table EC.7 the variance and Gini coefficient of the simulated TAT and proportion of infected infants treated across different levels of geographic aggregation, for both the OLA and OCA solutions and the current EID network configuration in Mozambique.

As seen in Table EC.7, these statistics suggest that the OLA solution either maintains or slightly increases the heterogeneity of TATs relative to the existing network, however this increase (at most two percentage points in the Gini ratio) does not seem significant. Likewise, the OLA solution is found to leave the heterogeneity in the proportion of infected infants treated across clinics, districts and regions almost unchanged relative to the status quo configuration.

As noted in the discussion of the OCA results (§6.2) the consolidation of capacity results in increased average transportation times for some clinics but decreased LCTs for all clinics. This observation invites equity concerns. Using the same methodology as before we compare outcomes between different geographic areas of Mozambique to find that the OCA solution tends to reduce inequities in both TATs and proportions of infected infants treated across nearly all geographic

Table EC.7 Simulated equity results for baseline network configurations.

	Status Quo		OLA		OCA	
	Var	Gini	Var	Gini	Var	Gini
TAT						
Patient	414	0.29	363	0.29	259	0.27
Clinic	178	0.18	243	0.20	165	0.19
District	143	0.17	176	0.18	105	0.17
Regional	83	0.12	83	0.13	39	0.10
Proportion of infected infants treated						
Clinic	0.055	0.37	0.055	0.37	0.055	0.36
District	0.031	0.24	0.027	0.24	0.040	0.27
Regional	0.002	0.06	0.003	0.07	0.002	0.05

Variance and Gini coefficient computed for simulated TAT per patient and proportion of infected infants treated per clinic. For higher level aggregations variance and Gini coefficients are computed for mean performance outcomes associated with each clinic, district or region.

aggregation levels. This is consistent with the discussion that the decrease in average LCTs associated with the OCA solution is far outweighed by the increase in transportation times resulting from centralizing capacity in a single location. Indeed, in a centralized solution the decrease in average LCTs promotes equity because it equally affects all originating patients and geographic locations, while the increase in transportation times promotes inequity because it affects to a greater extent the patients and locations that are far away from the single lab in Nampula.

EC.8. Managerial considerations for EID network design details

In this section we provide the full details of the analyses summarized in §6.3. First, we investigate how the improvements associated with OLA and OCA would be affected by the addition of certain constraints motivated by administrative considerations in §EC.8.1 and §EC.8.2, respectively. Second, we provide a detailed analysis of the optimal network structure for networks with increasing transportation times (combined with different utilization levels) in §EC.8.4.

EC.8.1. Administrative considerations for OLA

A potential drawback of the OLA solution from an implementation standpoint is that each lab serves clinics located in multiple administrative regions (3–10), in contrast with current practice wherein each lab serves only one or two regions (see §EC.2). Likewise, assigning clinics in the same district to different labs may create implementation challenges. This is because some planning and coordination activities related to the transportation of EID samples are currently conducted at both regional and district levels. Hence we conduct sensitivity analysis involving additional restrictions on the feasible set of lab assignments so that they are more aligned with current administrative boundaries. Specifically, we add the constraint that all clinics in a district must be assigned to the same lab (*optimal district-level lab assignment*, or ODLA) and the constraint that all clinics

in a region must be assigned to the same lab (*optimal region-level lab assignment*, or ORLA). We find that the simulated performance of ODLA is not statistically different from that of OLA, however the performance of ORLA is significantly worse, with only a 2.5% increase in infected infants initiating treatment against 4% for ODLA and OLA.

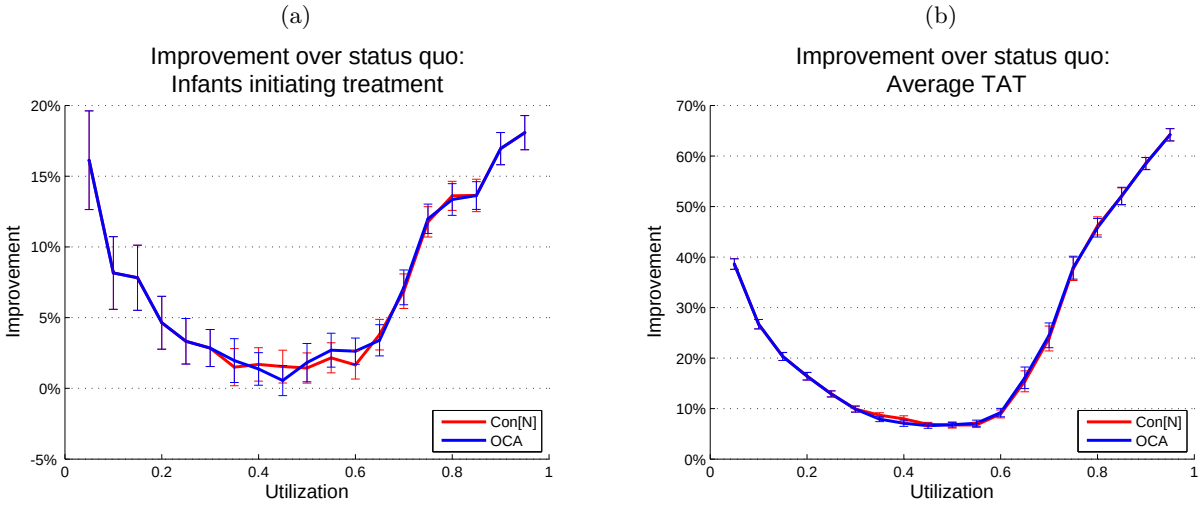
EC.8.2. Administrative considerations for OCA

Turning now to the baseline OCA solution consisting of pooling all available lab capacity in the city of Nampula (see §6.2), a potential implementation difficulty is that this location is not the main administrative center in Mozambique. Instead, a single centralized lab in the capital city of Maputo could be more desirable from an administrative standpoint. Under baseline parameters however, the solution consisting of locating all available testing machines in Maputo performs substantially worse: it only increases the number of infants initiated on treatment by 4.3% over the current solution, as against 6.9% with a single lab in Nampula. The main driver for this result is the larger average post-processing delay estimated for Maputo (7 days) relative to Nampula (2 days), which outweigh the slight advantage of Maputo over Nampula in terms of average sample transportation times (11 days versus 12 days). Indeed, under a modified scenario where all lab locations are assumed to have the same post-processing delay distribution, the optimal solution of the OCA model becomes a single lab located in Maputo. In conclusion, consolidating all testing capacity in Maputo could be justified by shorter transportation delays and greater administrative convenience, but only if accompanied by effective interventions to reduce current post-processing delays in that lab.

EC.8.3. Utilization

The sample arrival rates and average server processing times estimated from the Mozambique dataset result in an overall system utilization of 90.4%. To generalize our findings, we adjust all sample arrival rates by a common multiplicative factor so that system utilization is varied from 5% to 95% in increments of 5%, re-solve the OCA problem and simulate the resulting solution in each instance. In the following we will refer to the configuration where all capacity is consolidated in Nampula as Con[N]. Figure EC.5 shows the performance improvement of Con[N] and the OCA solution over the status quo for each experiment. It predictably illustrates that the benefits of optimal capacity allocation decisions and the performance differences between all solutions considered are greatest for high utilization levels, due to the nonlinear aspect of congestion that is also observed in many other queueing systems. Because of the batch processing policy at the labs however, we also find that the potential for performance improvement over the status quo solution (and

Figure EC.5 The improvement of OCA and Con[N] over the status quo for different utilization levels.



Note. In (a), (b), and (c) the blue and red curves represent the improvement of the OCA and Con[N] upon the current lab assignment on KPIs 1, 2, and 3, respectively. Confidence intervals represent the 95% level. Note the different ranges of the y-axes in (a) and (b).

thus the importance of capacity allocation decisions) is also substantial for low utilization levels. This result is important from a practical perspective, because it suggests that during the scale-up phase of an EID program when managers initially have ample capacity compared to workload, appropriate allocation of this capacity to labs still has a significant impact on system performance.

For levels of utilization at the low and high end of the range (above 85% or below 35%), the OCA solution coincides with Con[N] and consists of consolidating all available capacity in a centralized lab, hence the identical values of the two curves in Figure EC.5. For intermediate utilization levels, the OCA solution involves the allocation of capacity to one or two additional labs (see Table EC.8) but the marginal benefit of doing so relative to the consolidated solution is not significant, as seen in Figure EC.5. This indicates that for countries with average sample transportation times to labs comparable to those in Mozambique, consolidation is a robust recommendation across different system loads, including those in initial phases of scale-up and those that are more established with high utilization.

EC.8.4. Transportation times

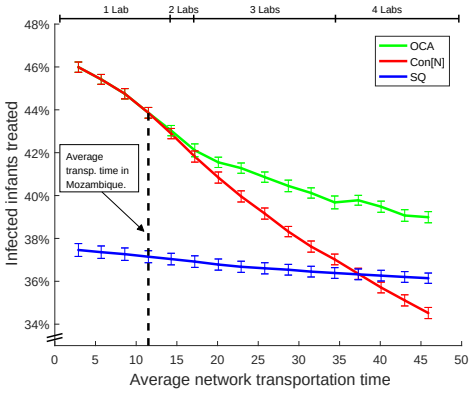
We now examine the impact of sample transportation times and network density on capacity allocation decisions. Specifically, we consider the range of EID networks obtained by scaling the transportation times from each clinic to each lab in our dataset by a common multiplicative factor between 0.25 and 4, which corresponds to average network transportation times ranging from 3 to

Table EC.8 OCA as a function of utilization.

Utilization	Number of servers				Mean LCT				Mean transp. time				Mean expected TAT				Proportion of results in 1st month				Mean transp. time	Mean LCT	Mean TAT
	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4			
5%		5				15				11				38				39%			10.7	14.6	37.8
10%		5				10				11				33				53%			10.8	9.7	32.8
15%		5				8				11				31				58%			10.7	7.9	31.0
20%		5				7				11				30				60%			10.7	7.0	30.1
25%		5				6				11				30				62%			10.7	6.5	29.6
30%		5				6				11				29				63%			10.7	6.1	29.3
35%	2	3			11	6			2	10			28	30			68%	61%			8.6	7.3	29.3
40%	2	3			9	6			2	11			27	30			72%	60%			8.2	7.1	29.1
45%	1	3	1		9	6	11		2	10	3		26	30	28		73%	59%	67%		7.1	7.7	28.8
50%	1	3	1		9	6	11		2	10	3		25	30	29		73%	61%	68%		7.2	7.4	28.5
55%	1	3	1		9	6	10		2	10	3		25	30	27		73%	62%	70%		7.3	7.2	28.3
60%	1	3	1		9	6	10		2	10	5		25	28	31		74%	64%	63%		7.4	7.1	28.3
65%	1	3	1		9	6	10		2	10	4		25	29	30		74%	63%	67%		7.4	7.0	28.1
70%	1	3	1		9	6	9		2	10	3		25	29	27		74%	62%	71%		7.5	6.9	28.0
75%	1	3	1		9	6	9		2	10	5		26	28	30		72%	63%	65%		7.6	7.1	28.3
80%	1	3	1		9	6	10		2	10	6		26	28	33		71%	64%	58%		7.4	7.4	28.6
85%	1	4			9	6			2	10			26	29			72%	62%			8.8	6.4	28.5
90%		5				6				11				29				63%			10.7	5.9	29.0
95%		5				7				11				30				58%			10.8	7.1	30.3

Each row shows the results of the optimization model for various levels of utilization, shown in the first column. The next 5 column groups show the results by lab, while the final 3 columns contain national averages.

Figure EC.6 EID performance as a function of network transportation times (copy of Figure 7).



Note. The y-axis has a range from 34% to 48%.

46 days. For the resulting continuum from dense to sparser networks, Figure 7 (copied in Figure EC.6) shows the simulated proportion of infected infants treated that is associated with the status quo, OCA, and Con[N] solutions to the capacity allocation problem.

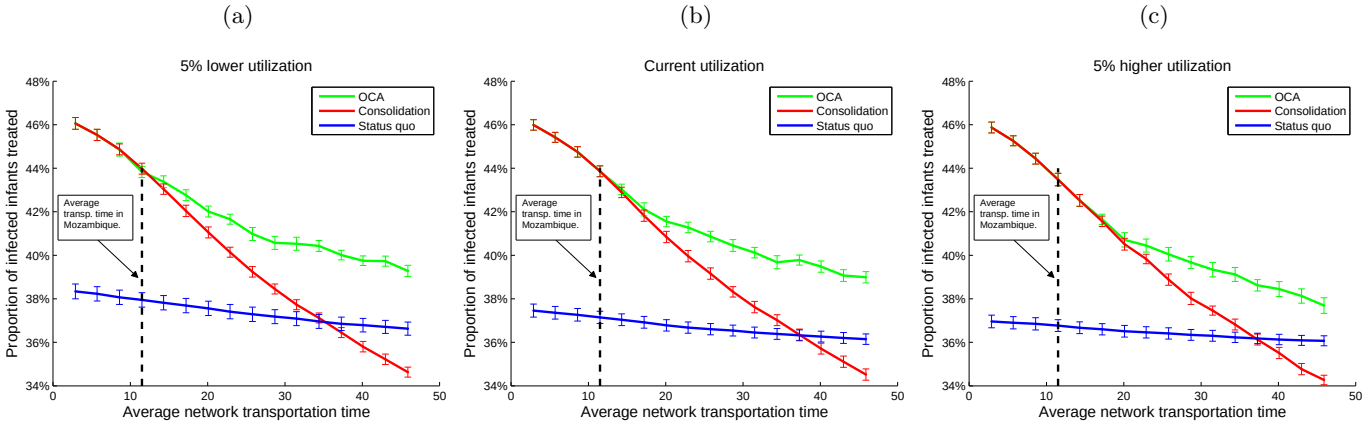
Note that the OCA and Con[N] solutions coincide for average transportation time shorter than 12.8 days (15% longer than that in Mozambique). Beyond that the suboptimality of Con[N] increases with the sparsity of the network, but the consolidation solution continues to outperform the status quo for average transportation times up to 35 days. Interestingly, for average transportation times equal to 33.5 days and above (scaling factor of 3), the location of testing machines in the OCA solution is the same as in the current status quo in Mozambique. There is still a substantial

Table EC.9 OCA as a function of network transportation times.

Scaling factor	Number of servers				Mean LCT				Mean transp. time				Mean expected TAT				Proportion of results in 1st month				Mean transp. time	Mean LCT	Mean TAT
	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4			
0.25		5				6				3				25				84%			3	6	25
0.50		5				6				6				28				77%			6	6	28
0.75		5				6				9				31				70%			9	6	31
1.00		5				6				12				34				63%			12	6	34
1.05		5				6				12				35				61%			12	6	35
1.10		5				6				13				35				60%			13	6	35
1.15		5				6				14				36				59%			14	6	36
1.20	1	4			11	7			2	14			31	37			70%	55%			12	7	36
1.25	1	4			11	7			3	14			31	38			69%	54%			13	8	37
1.50	1	3	1		12	8	12		3	14	15		32	36	50		67%	47%	54%		13	10	40
1.75	2	2	1		9	9	13		19	13	12		48	35	44		56%	46%	50%		15	10	42
2.00	2	2	1		9	9	13		23	14	10		51	37	40		55%	41%	49%		17	10	44
2.25	2	2	1		9	9	13		26	15	13		52	38	45		54%	37%	48%		19	10	45
2.50	2	2	1		9	9	13		29	17	16		52	40	48		54%	34%	48%		21	10	46
2.75	2	2	1		9	9	13		33	19	14		54	42	45		54%	32%	47%		24	10	47
3.00	2	1	1	1	9	13	13	15	15	13	10	36	44	39	40	68	55%	41%	47%	14%	21	12	50
3.25	2	1	1	1	9	13	13	15	18	13	12	39	46	39	43	68	54%	39%	46%	14%	22	12	51
3.50	2	1	1	1	9	13	13	15	26	13	14	40	50	40	44	68	54%	38%	46%	14%	24	12	51
3.75	2	1	1	1	9	13	13	15	28	14	15	42	50	40	46	68	53%	37%	45%	14%	26	12	52
4.00	2	1	1	1	9	13	13	15	31	15	13	44	50	42	43	69	53%	37%	45%	13%	28	12	53

For each row in the table we scale the transportation times of the Mozambique EID network using the scaling factor in the first column. We then re-optimize each instance, and simulate the system using the corresponding OCA. The next 5 column groups show the results by lab, where L1 to L4 refers to Maputo, Nampula, Quelimane, and Beira, respectively. The final 3 columns contain national averages for each instance.

Figure EC.7 EID performance as a function of network transportation times, for three levels of utilization.



performance gap between these two solutions however, due to different assignments of clinics to labs. This suggests that the current distribution of capacity in Mozambique would only be justified if transportation times were much longer than currently observed in that country, and highlights the importance of clinic assignments over and above that of lab location decisions.

The structure and detailed performance of the OCA solutions obtained over the range of scaled networks is reported in Table EC.9. As expected, the available testing capacity is distributed over an increasing number of locations as average transportation times and network sparsity increase. This reflects the trade-off between LCTs (which decrease when capacity is centralized due to pooling

Table EC.10 OCA as a function of network transportation times - 5% higher utilization.

Scaling factor	Number of servers				Mean LCT				Mean transp. time				Mean expected TAT				Proportion of results in 1st month				Mean transp. time	Mean LCT	Mean TAT
	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4			
0.25		5				8				3				27				79%			3	8	27
0.50		5				8				6				30				72%			6	8	30
0.75		5				8				9				33				65%			9	8	33
1.00		5				8				12				36				57%			12	8	36
1.25		5				8				15				39				51%			15	8	39
1.50	1	4			15	9			3	17			35	43			59%	42%			15	10	42
1.75	1	4			15	9			3	20			35	45			58%	38%			18	10	44
2.00*	2		3		12		10		24		16		55		42		48%		35%		19	11	48
2.25	2		3		12		10		27		18		57		44		48%		32%		22	11	49
2.50	2	2	1		12	14	17		30	18	12		56	46	47		47%	26%	39%		22	14	50
2.75*	2	2	1		12	14	17		33	20	26		56	48	48		47%	24%	38%		24	14	51
3.00	2	2	1		12	14	17		36	21	11		57	50	46		47%	23%	38%		26	14	52
3.25*	2	2	1		12	14	17		36	22	26		56	51	54		47%	21%	37%		28	14	53
3.50	2	2	1		12	14	17		36	26	29		54	54	55		47%	20%	37%		30	14	54
3.75	2	2	1		12	14	17		34	27	38		53	55	60		47%	19%	37%		32	14	56
4.00*	2	2	1		12	14	17		39	28	37		55	55	59		47%	19%	36%		34	14	56

Note: For the starred scaling factors the solution described is the result of running the optimization from 6 hours. The optimality gaps for these cases are 0.08%, 0.4%, 0.34%, and 0.15%, respectively.

effects) and transportation times (which decrease when capacity is distributed geographically). Through this exercise we are able to quantify the specific network densities and scaling factors for which the minimal and maximal capacity distribution solutions are optimal (average transportation times lower than 1.15 times and higher than 3 times those observed in Mozambique, respectively). These can act as guide posts when designing EID networks in different countries.

Finally, note from Table EC.9 that once the transportation times become so long that a sufficiently large number of clinics have a negligible probability of receiving results within the first month (average transportation times equal to 33.5 days and above, more than 3 times those found in Mozambique), the last lab is opened to specifically serve those clinics. This lab has a higher utilization than other labs in the optimal solution because further increasing the delay at the clinics they serve will not worsen the outcome significantly. This allows other labs to have lower utilization and higher probability of delivering results within a month.

We also examine the sensitivity of these results with regard to utilization. Figure EC.7 is the equivalent of Figure EC.6, comparing the performance of OCA, Con[N], and the status quo as transportation times are gradually increased, for three levels of utilization. Similarly, in Tables EC.11 and EC.10 we provide sensitivity analysis of Table EC.9. These tables describe the nature of the OCA as the network diameter increases for a 5% decreased utilization and a 5% increased utilization, respectively. First, note that as transportation times gradually increase, additional labs are opened later when utilization is high. This is due to the pooling benefits of consolidation being more substantial for higher utilization, and thus outweighing the increased transportation times.

Second, Tables EC.11 and EC.10 demonstrate that even if the number of operational labs remains constant for increasing transportation times their location and capacity can vary. For instance in

Table EC.11 OCA as a function of network transportation times - 5% lower utilization.

Scaling factor	Number of servers				Mean LCT				Mean transp. time				Mean expected TAT				Proportion of results in 1st month				Mean transp. time	Mean LCT	Mean TAT
	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4			
0.25		5				6				3				25				84%			3	6	25
0.50		5				6				6				28				77%			6	6	28
0.75		4	1			6	10			9	4			29	38			71%	68%		8	16	31
1.00	1	4			10	6			2	11			29	34			74%	62%			10	16	33
1.25	1	3	1		11	7	10		2	13	10		31	34	44		70%	55%	63%		11	8	36
1.50	1	3	1		11	7	10		3	15	13		31	37	46		69%	49%	62%		13	8	38
1.75	2	2	1		8	8	10		19	13	10		48	34	38		57%	47%	60%		15	8	41
2.00	2	2	1		8	8	10		23	14	9		51	36	37		56%	43%	59%		17	8	42
2.25	2	2	1		8	8	10		26	16	11		51	37	41		56%	39%	58%		19	8	44
2.50	2	1	1	1	8	11	11	11	14	9	8	30	41	33	36	63	57%	48%	55%	20%	17	10	46
2.75	2	1	1	1	8	11	11	11	15	10	11	33	42	33	39	63	56%	47%	53%	20%	19	10	47
3.00	2	1	1	1	8	11	11	11	16	12	11	36	43	36	40	64	56%	45%	53%	19%	21	10	48
3.25	2	1	1	1	8	11	11	11	19	12	11	38	46	37	40	64	55%	43%	51%	19%	22	10	48
3.50	2	1	1	1	8	11	11	11	26	13	11	40	49	39	38	64	55%	42%	51%	18%	25	10	49
3.75	2	1	1	1	8	11	11	11	28	13	14	43	49	38	43	64	54%	40%	49%	18%	26	10	50
4.00	2	1	1	1	8	11	11	11	37	14	12	43	53	39	41	63	54%	40%	49%	18%	29	10	50

Table EC.11, when scaling from 0.50 to 0.75 a new lab (Lab 3) is opened to relieve the load of the main lab (Lab 2). This benefits the clinics assigned to Lab 2, as clinics that don't benefit much from shorter LCTs and transportation times are assigned to the new lab. In contrast, when scaling from 0.75 to 1.00 the second lab (Lab 1) is opened for the benefit of the clinics assigned there. A similar effect can be seen when scaling from 1.75 to 2.00 in Table EC.10.

Third, note that the lab assignment is being jointly optimized with the capacity allocation. For instance, consider scaling from 1.75 to 2.00 in Table EC.10. The capacity allocation is constant but some of the other performance parameters change. This is due to changes in the lab assignment. Specifically 21 clinics change assignments, 10 moving from Lab 3 to Lab 1 and 11 doing the opposite. Those clinics moving from Lab 3 to Lab 1 incur slightly longer exogenous delays after the move whereas those clinics that move from Lab 1 to Lab 3 incur little or no increase in exogenous delays, i.e. are being favored. This indicates that those clinics are critical for the particular scaling factor of 2.25 as they are being moved to the Lab which has a shorter LCT.

EC.9. Sensitivity analysis

In this section we provide full discussion of the sensitivity analyses conducted. We first examine how predicted improvements associated with the OLA and OCA solutions are affected by changes in the caretaker behavior model (§EC.9.1). Second, we analyze the potential impact of segmenting samples for processing based on caretaker participation in PMTCT programs (§EC.9.2). Third, we evaluate the robustness of the OCA solution (consolidation of all lab capacity in one location) for networks with different loads (§EC.8.3). Finally, we examine how the optimal solutions are affected by optimizing for operational outcomes (minimize expected TAT) as opposed to the public health outcome of treatment initiation (§EC.9.3).

EC.9.1. Caretaker sensitivity to delays

The most uncertain input parameters in our model may be those related to follow-up behavior of caretakers, as discussed in §4.2. Hence, we use the estimates in Deo et al. (2014) to generate values for caretaker sensitivity to delay that are two standard deviations above and below the mean estimate, and recalculate the impact of the OCA and OLA solutions on the number of infants initiating treatment. In the scenario where follow-up by caretakers is least sensitive to delay, the OLA and OCA solutions are predicted to increase the number of infants initiated on treatment by 1.3% and 1.7%, respectively. However, these estimates increase to 8.5% and 14.4% in the scenario where caretaker follow-up is most sensitive to delay.

EC.9.2. PMTCT segmentation

Infants born to HIV-infected mothers who have not participated in a PMTCT program are more likely to be infected by HIV, and individual PMTCT participation status can be determined at the time of sample collection. Hence we study versions of our network design models wherein the flows of PMTCT and non-PMTCT samples from each clinic can be differentiated. This allows processing non-PMTCT samples faster by sending them to closer or more responsive labs. Hence, we define expansions of our OLA and OCA optimization models where the original assignment variables from each clinic to each lab were split into two variables corresponding to PMTCT and non-PMTCT samples from each clinic, respectively. These new models are denoted *optimal segmented lab assignment* (OSLA) and *optimal segmented capacity allocation* (OSCA).

Unfortunately however, our simulation results show that outcomes associated with the OSLA and OSCA solutions under baseline parameter values are not statistically different from those of the OLA and OCA solutions. This can be explained by the fact that the higher positive rate of non-PMTCT samples is mitigated by the lower rates of follow-up for result collection by non-PMTCT caretakers in our behavioral model (see §4.2.2). When our behavioral model is modified so that PMTCT participation status no longer has an impact on the probability of results collection (conditional on infants being still alive), the OSLA solution is indeed estimated to increase the number of infants initiated on treatment by 3% relative to the OCA solution. The OSLA solution then involves shipments of samples to different labs based on PMTCT status for 138 out of 410 clinics, Nampula becomes a priority lab for non-PMTCT samples (with a much higher proportion of these samples than all other labs and lower overall utilization), and across the entire network the average TAT of non-PMTCT samples is shorter than that of PMTCT samples by 7 days. Even with this modified behavioral model however, the OSCA and OCA solutions remain identical and still consist of consolidating all lab capacity in Nampula. In other words, segmentation offers no additional benefits over and above those derived by consolidating lab capacity.

EC.9.3. Operational optimization problem objective

An alternative to maximizing the (approximate) expected number of infants initiating treatment (13) as part of formulations OLA and OCA is to minimize instead the average expected TAT $\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{d=1}^{D}\lambda_i z_{ijd} d$. We denote by THLA and THCA (*minimum TAT heuristic for lab assignment/capacity allocation*) the formulations obtained by substituting this last expression for the objective (13) in the OLA and OCA formulations, respectively. Instantiating THLA and THCA is easier than OLA and OCA, because the TAT objective does not require estimates of parameters ω_{im} and ϕ_{dm} capturing caretaker follow-up behavior and the conditional distribution of TATs, respectively (see §5.2). The simpler computational set-up associated with THLA and THCA is therefore appealing from an implementation standpoint.

Numerical experiments reveal that for our baseline problem data, the simulated performance of THLA (resp. THCA) is statistically identical to that of OLA (resp. OCA) at the 10% level. While the lab assignments generated by THLA and OLA differ slightly, THCA and OCA both generate the exact same solution (consisting of a single central lab in Nampula). Finally, for our baseline problem there are no significant differences in computation times required to solve these formulations to optimality (17min 16s for THLA vs. 17min 48s for OLA, 1h 0min 58s for THCA vs. 1h 1min 44s for OCA). These observations do not generalize across problem parameters however. When caretaker sensitivity to delay is increased by two standard deviations (as in §EC.9.1) and network utilization is increased by 5% for example, the simulated average number of infected infants receiving treatment with OLA is 1.8% larger than with THLA, and that difference is statistically significant at the 10% level (the average TATs obtained with both formulations are not statistically different however). In addition, the computation time required to solve OLA to optimality is about 40% less than THLA (1h 44min 58s for THLA vs. 42min 31s for OLA). These results are consistent with the main relative weakness of THLA, which is that it ignores the differences in public health impact that the same absolute reduction in TATs may have across different clinics, which is captured by parameters ω_{im} and ϕ_{dm} in OLA. For example, THLA would not differentiate between reductions of average TAT for a given clinic from 15 to 5 days and from 30 days to 20 days, even though the latter would have a greater impact on the number of infants initiating treatment because it would bring more samples below the TAT threshold of 30 days which is sensitive for caretaker follow-up. Because they reflect this and other relevant behavioral features, in scenarios with high caretaker sensitivity to delays parameters ω_{im} and ϕ_{dm} also effectively reduce the range of solutions to be evaluated by the optimization algorithm, which explains the differences in computation times. Likewise, some differences between the THCA and OCA solutions are observed away from baseline parameter values. When transportation times are scaled up by a factor of 3 for example (as in §EC.8.4), OCA allocates some capacity to all four lab locations whereas THCA recommends only three.

This qualitative difference of the THCA solution results in a statistically significant reduction by 1.61% of the simulated number of results becoming available in the first month following sample collection, relative to OCA.