

Evaluation of Zipline in Ghana: Evaluation Design and Pre-Analysis Plan

Motivation for PAP

Zipline aims to achieve impact along many indicators of interest (product stocking, product stockouts, treatment behaviour, health outcomes, and response to pandemic), across at ~37 products and at least 13 conditions. Therefore, using all raw variables as primary outcomes is not optimal since it would yield a large number of results which would be difficult to process, present and explain. An additional factor for not including many variables into the analysis is conducting many statistical tests - the more tests are conducted the higher the likelihood of getting a statistically significant result purely by chance. In fact, with 5 independent tests, the probability of a false positive in at least 1 test is roughly 22%¹. Therefore, producing a lot of results and selecting significant ones will not yield valid conclusions. To address the problem, researchers usually (a) shrink the number of variables by creating composite indexes (b) apply multiple hypothesis adjustment correction to adjust for testing multiple outcomes.

This document describes our approach to definition key indicators to estimate Impact of Zipline in Ghana². The pre-analysis plan serves two purposes: (1) critically think through the definition of success of Zipline (b) pre-commit to indicators that would measure the success. IDinsight will conduct additional analysis that is not pre-specified here, however, those analyses will be considered secondary and we will draw overall conclusions from effects on primary outcomes. The additional analyses will provide depth and elucidate mechanisms behind the estimates, but will not be used to tell the impact story.

Program Summary

Zipline is a drone delivery service that provides third-party logistics (3PL) services, holding customers' products in warehouses to minimize the lead time between order and delivery for end recipients. In Ghana, Zipline has partnered with Ghana Health Service (GHS) to provide last mile logistics services and improve access to crucial health products in health facilities. Zipline is establishing four distribution centres (Omenako, Mampong, Walewale, and Sefwi Wiawso) to serve approximately 2,000 health facilities across Ghana. Zipline provides emergency and non-emergency on-demand medical products (such as blood, emergency and routine vaccines, and emergency and essential medicines). This delivery system is intended to improve timely access to medical products, reduce instances and duration of stockouts in health facilities, and ultimately improve health

¹ $1-(1-0.05)^5=0.22$

² Some of the language for this document was adapted from IDinsight's pre-analysis plan for the Evidence Action iron and folic acid evaluation.

outcomes. The process is initiated when a healthcare worker submits an order for a particular product via SMS or a phone call. The order is recorded at the closest distribution center, and the required medical products are packed into a drone. The automatically operated drone drops off the medical products in an allocated location nearby the facility. The medical products are picked up by the facility staff within about 30 minutes.

Theory of Change

The overall theory of change operates as following:

Health workers and health facilities use Zipline to order health care products, which are supplied to Zipline by the Ghana Health Service. Zipline receives the order, then processes and confirms it. The package is then launched and dropped at the health facility. This results in a faster and more convenient delivery system for health facilities. The main outcome is the reduction of stockout instances and duration, improved product availability and a more equitable medical supply system. This can lead to fewer patient referrals and less waiting time before treatment. Ultimately this feeds into the a final impact of a) better quality treatment and improved health outcomes, and b) lower supply chain costs, leading to cost savings for the Ghanaian public health system.

Research Questions

This evaluation will aim to answer the following research questions:

Primary Research Questions:

1. Product Stocking: Did Zipline improve equity of healthcare delivery through improving stocking rates of essential products/vaccines?
2. Did Zipline improve the supply chain? Specifically,
 - a. Product Stockouts: are there reduced instances and duration of stockouts for relevant products?
3. Treatment Behaviour: Did Zipline improve health worker's ability to deliver better care (patients experiencing less referrals due to lack of medical products)?

Secondary Research Questions:

4. Stocking Behaviour: Are healthcare workers more likely to take action to resolve stockouts? Are stockout resolutions more timely?
5. Attitudes and perceptions: What are the views of Zipline by health workers
6. Costs: What effect do Zipline's operations have on supply chain costs? Cost components to be considered include: transport costs; storage and physical infrastructure costs; expiry, waste, and loss of potency; inventory holding cost; and management, overhead, and labour. This piece will be completed by Llamasoft and is outside of the scope for IDInsight
7. Healthcare worker satisfaction: Are healthcare workers more satisfied with their jobs?

Outcomes Metrics

We will use the following indicators to measure the success of the program. The analysis will be based on 18 medical products and 10 vaccines.

Table 1: Primary indicators of interest that define “Success” of the intervention.

RQ	#	Indicator	Calculation
Product Stocking	1	Number of unique products stocked	Total number of products of interest
<u>Product Stockouts - addressing continuity of supply</u>	2	Percentage of days for which a facility experienced a stockout (across relevant products that the facility stores ³)	$\frac{\text{Total number of days of stockouts}}{\text{Relevant period in days}}$ A value of zero is assigned for products which did not experience a stockout. If the product has not been stocked at the facility, the fraction will be set to missing. A fraction will first be created for every relevant product at the facility, and the an average will be taken across all relevant products
	3	Percentage of days for which a facility experienced a stockout (vaccines) ⁴	$\frac{\text{Total number of days of stockouts}}{\text{Relevant period in days}}$ A value of zero is assigned for products which did not experience a stockout. If the product has not been stocked at the facility, the fraction will be missing. A fraction will first be created for every relevant product at the facility, and the an average will be taken across all relevant products
	4	Average time to resolve the last stockout across relevant medical products	The average first be constructed within the facility across relevant products that the facility stores.
<u>Treatment Behaviour - reduction in referrals & ability to provide “on-demand-supply”</u>	5	At least 1 patient didn’t get vaccinated in time because of stockout (past 3 months) ⁵	=1 if the number of patients that didn’t get any vaccine in time because of stockouts, =0, otherwise
	6	Number of patients didn’t get vaccinated in time because of stockouts (past 3 months)	Either a raw variable or a natural logarithm will be taken of the measure to address outliers, depending on the presence of outliers. Zeros will be replaced with small non-zero values before taking logs.
	7	At least 1 patient treated by a healthcare worker was referred because there was no medical product (Self-reported referral rates for all patients , general healthcare worker)	=1 if the number of patients referred due not having medical products is non-zero, =0, otherwise
	8	Number total patients that were referred because there was no medication	Either a raw variable or a natural logarithm will be taken of the measure to address outliers, depending on the presence of outliers. Zeros will be replaced with small non-zero values before taking logs.

³ The recall period may differ depending on the round

⁴ Only started recording in August 2020

⁵ The recall period will be consistent across survey rounds

	9	At least 1 critical patient treated by a healthcare worker was referred because there was no medical product	=1 if the number of patients referred due not having medical products is non-zero, =0, otherwise
	10	Number of critical patients treated by a healthcare worker referred because there was no stock of medicine ⁶ (Self reported treatment for critical care patients , general healthcare worker)	Either a raw variable or a natural logarithm will be taken of the measure to address outliers, depending on the presence of outliers. Zeros will be replaced with small non-zero values before taking logs. Only relevant to facilities which had at least 1 patient in a critical state
	11	Referral rates for only selected conditions, using administrative data (Endline only) ⁷	$\frac{\text{Total Number of patients referred for a particular condition in the past 6 months}}{\text{Total number of patients with the particular conditions}}$ Referral rate will be calculated separately for each condition. Average referral rate will be reported across all relevant conditions. Missing values due to not having patients with a particular condition will not be replaced with zeros. Uncomplicated malaria, Inpatient cases of malaria, Dog bites/rabies, Snake bites, HIV, TB, Post partum haemorrhage, Eclampsia, Pre-eclampsia

To understand the broader Zipline story, we will also look at

- 1) Treatment of the last patient with certain conditions for general and maternity worker
- 2) Procedures to resolve and consequences of the last stockout of certain medical products and vaccines

We are including indicators of interest for each type of survey that will be reported in the analysis. Primary indicators of interest are highlighted in yellow.

Sample Selection

IDinsight will use a matching (with ANCOVA) identification strategy to estimate the causal impact of Zipline services on the outcomes of interest. The estimation strategy relies on the assumption that trends in treatment facilities are comparable to similar facilities in the control group.

The treatment sample is selected from all facilities located in Zipline service zones. These facilities are situated within 80km range of the distribution centers and can be reached by drones. Facilities outside of the 80km range of the distribution centers cannot be reached by Zipline drones. The control sample is selected from facilities located outside of the Zipline service zones.

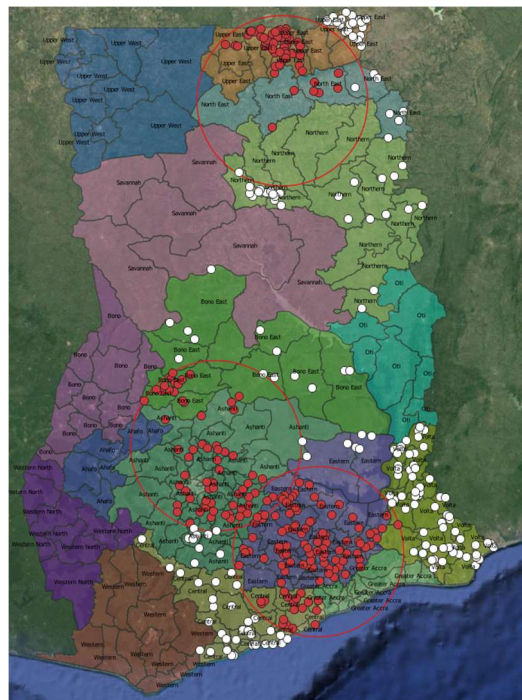
⁶ Almost all facilities have at least 1 critical patients who was referred, so the binary indicator is not a meaningful measure of impact.

⁷ This may not be feasible through phone surveys and the outcome may be dropped

The research team conducted a two stage matching process to create a comparable treatment and control sample of facilities.

The first stage of matching was completed before the baseline data collection using District Health Information Management System (DHIMs). The dataset contained basic information about facilities and health records for selected outcomes. Researchers have used coarsened exact matching (Blackwell et. al. 2009) to create a more comparable treatment and control sample based on observable characteristics. The process created a sample of 452 facilities across the two groups which were interviewed during baseline visits. The baseline sample consisted of three level facilities: district hospitals, healthcare centers, and Health Service Community-Based Health Planning and Services (CHPS). During the data collection, the following staff members were interviewed to comprehensively document pre-treatment outcomes of interest: heads of facilities, staff members in charge of all dispensaries/stock rooms, health workers and patients. The main outcomes of interest included stocking/stockout levels, treatment behaviour; and patients to understand their experiences with the healthcare provision at the facility. Non-eligible facilities (non-government, private facilities, and healthcare centers that fell outside of the three main categories) and those that no longer existed were excluded from the sample. The final sample for baseline consisted of 422 health facilities (238 and 184 within and outside of Zipline zones, respectively).

Figure 1: 422 facilities in the baseline sample.



The second stage of coarsened matching is completed using information collected during the baseline and the order data. Both of those datasets will be used to further narrow down to the sample of the most comparable facilities and facilities that actually use Zipline services. The final study sample will be finalized before endline. Prior to matching, we will exclude the facilities that were dropped from Zipline operations completely due to external factors and not facility demand.

We will either use 1:1 matching or 1:many depending on the balance and the number of observations in the final sample.⁸ The impact estimation will be made on the sub-sample of those facilities, with an exception of hospitals where all hospitals in the current sample will be included.

We will validate the sample with Zipline prior to results finalization using the following considerations:

- 1) Balance of baseline indicators of interest to ensure both IDinsight and Zipline are satisfied with the balance
 - a) All available primary indicators of interest (or close proxies)
 - b) Fixed facility-level characteristics
- 2) Location of the facilities to ensure that the selected facilities are fairly evenly distributed
- 3) Composition of the sample by facility type and location (north vs south)

The estimation of impacts will be run on sample of CHPS, Healthcare centers and hospitals. Due to the fact that hospitals are extremely different from the rest and constitute a very small fraction of the sample, we will run estimation on CHPS and Healthcare centers separately.

Estimation

Quantitative impacts:

The impacts on outcomes of interest will be measured using data phone surveys. To quantify impacts of Zipline service we will use ANCOVA specifications^{9 10 11}:

$$Y_{ft} = Y_{ft0} + \beta_1 * Zipline_{ft} + \alpha_f + \gamma_t + \beta X_f + \epsilon_{ft}$$

where

- Y_{ft} is a facility-level outcome for facility f in survey round t , if applicable
- Y_{ft0} is a facility-level outcome at baseline (or a close proxy), making the estimation strategy ANCOVA
- α_f are facility type fixed effects (CHPS maternal, CHPS, HC centers and hospitals) these fixed effects are necessary to control for differences between facility types.
- γ_t are survey round fixed effects (if there are more than 1 data point per facility) to control for idiosyncratic time shocks

⁸ If we use 1:many matching, we will apply weights generated for each stratum

⁹ The sample will either consist of all facility types or only HC/CHPS facilities, depending on how well the hospital sample is balanced and how much noise it adds to the estimate. The hospital sample is very different from a non-hospital sample, and adding them to a pooled estimate may create noise.

¹⁰ We prefer ANCOVA specification over difference in differences due to 2 reasons: (1) it has more power when estimating outcomes that are not strongly autocorrelated and (2) there are differences in measurements between surveying rounds. For example, at baseline we measured outcomes with 6 month recall, while during high frequency checks the recall period is 1.5 months.

¹¹ We will also consider running treatment on the treated specification, where “takeup” will be defined as ordering at least 1 product of interest.

- *Zipline* is an indicator for whether a facility is serviced by Zipline at the time of the survey. β_1 estimates the causal effect of interest.
- X_i is a vector of covariates from the baseline dataset. Covariate data is crucial for the causal estimators described below. It allows us to control for imbalances between treatment and control facilities on relevant observable characteristics, such as differences in stockouts of products of interest, distances to road and other important characteristics that are correlated with the outcomes of interest. Controlling for these variables also improves the precision of our estimators. We will use the following as control variables¹²
 - Staff/Capacity:
 - Number of healthcare providers per bed, average daily working hours
 - Infrastructure: number of coolers available for medicine and blood storage, source of water (2 most common), has access to electricity, uses electronic system to manage patients, number of flush toilets
 - Self reported time traveled to get to the road using typical transport
 - Whether facility is in north/south region
 - Patients: Average number of patients per week
 - Typical order volume from regional management store
 - Location: Proximity to the closest paved road and distance to regional medical stores
- ϵ_{ist} is an error term. We will use Conley adjustment¹³ to the standard errors to account for serial and spatial autocorrelation of the error terms. We will use distance cutoff of 10 km (assuming that errors are correlated at most within 10 km radius¹⁴), and time cutoff of infinity (which produces standard errors equivalent to cluster command in Stata at the facility level).

We will present both ITT (intention to treat) and TOT (treatment on the treated estimates). Treatment on the treated estimates will be derived using a sub-sample of facilities that used Zipline service regularly (defined as using at least once in the past 6 months). A comparable matching group will be selected using coarsened exact matching, and balance will be presented for both the full sample and the user subgroups.

Qualitative analysis:

Due to a low number of hospitals in the sample, qualitative analysis will be used to understand differences in behavior between treatment and control facilities when it comes to products/conditions only relevant to the hospital sample. Qualitative insights will only produce

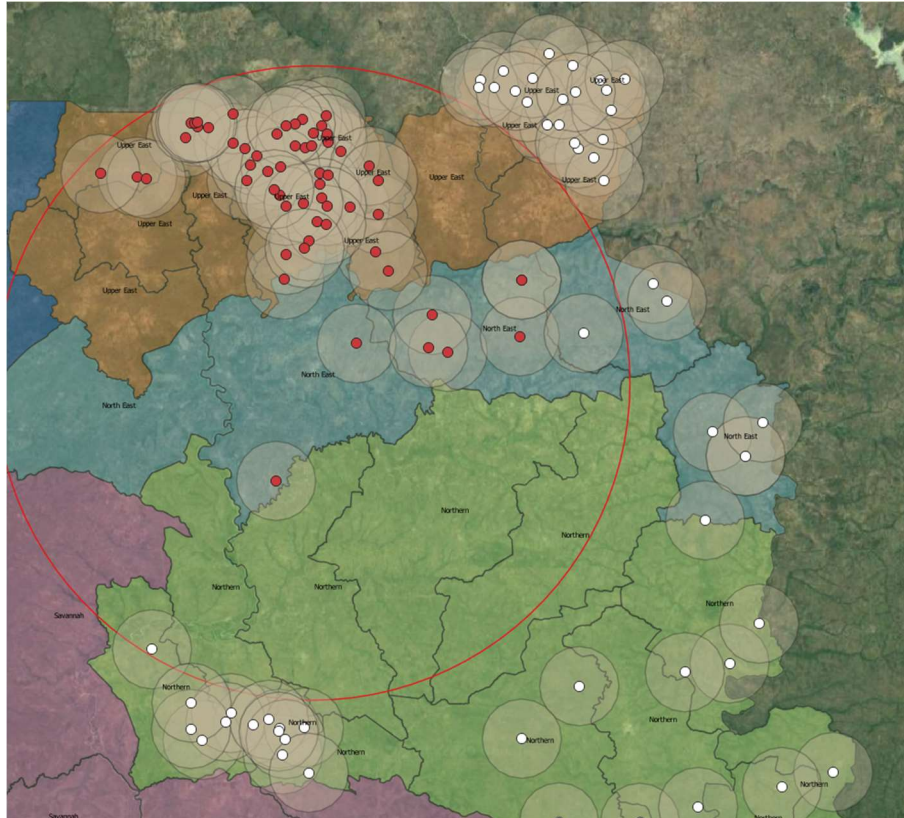
¹² We will consider adding variables that are not balanced in the final sample to control for potential underlying differences.

¹³ Hsiang, Solomon M. "Temperatures and cyclones strongly associated with economic production in the Caribbean and Central America." *Proceedings of the National Academy of sciences* 107, no. 35 (2010): 15367-15372. <http://www.fight-entropy.com/2010/06/standard-error-adjustment-ols-for.html>

¹⁴ 10 kilometer roughly reflects the spatial distribution of facilities in the data. Baseline data doesn't suggest that spatial correlation is significant for this sample and applying adjustment doesn't significantly change the standard error of a few sampled estimates.

indicative results and will be presented in a narrative form comparing behaviour/outcomes across treatment/control facilities.

Figure 2: A zoomed in sample of facilities with the 10km radius around each



Heterogeneity

- We plan to do a pooled analysis across all facility types, however, we recognize that hospitals are extremely different from all other health centers, so we will look at CHPS and HC separately if the estimates for the full sample are too noisy, or the hospital sample is imbalanced.
- We will also look at heterogeneity by
 - Remoteness which will be defined as self-reported travel time to the nearest nearest RMS since distance to the medical store is correlated with orders
 - Baseline availability of cold storage for storing medicine across two groups (below and above median)
 - North vs south to look at regional differences
- Zipline has requested to disaggregate impacts by model of delivery: sole vs supplementary model. Due to a small sample size (at most 11 of sole providers in the study pool), we will not be quantifying causal effects on this sub-group. IDInsight can provide some qualitative evidence but the insights will be extremely limited given such a low sample size.

- Zipline has also requested to look at impacts on high vs low users: The above caveats apply to our ability to complete this analysis. If we are able to predict which facilities in the control group would be high vs low users, we will disaggregate treatment effects by usage. However, we are concerned that the sample size will not be high enough to detect impacts.

Multiple Hypothesis Adjustment

Each of the 12 indicators of interest and the rest of the estimates will be tested against the null hypothesis that there is no effect. In frequentist statistics, the more tests are performed the larger the chance of detecting a false positive (i.e. type 1 error - the tests comes back as statistically significant, however, it is driven by chance). The magnitude of the problem is dramatic and with 5 independent tests, the probability of a false positive in at least 1 test is roughly 22%¹⁵. To address the problem, researchers usually (a) shrink the number of variables by creating composite indices to make an adjustment on a smaller number of variables (b) apply multiple hypothesis adjustment correction to adjust for testing multiple outcomes. We aim to apply false discovery rate (FDR) adjustment to all estimates following Benjamini et al. (2006)¹⁶. Under this approach, in expectation, less than 5% of null hypotheses will be incorrectly rejected. This approach is less conservative than others, however, the losses to statistical power are relatively minor. This adjustment will be applied to 2 batches separately: the first batch will consist of only success indicators and the second batch will consist of all other secondary analysis we will perform. The second batch will include the indicators from the first batch.

Appendix

List of robustness checks

1. Running regressions on products of interest that Zipline did not supply during the study period, if those exist. Detecting treatment effects on those products would indicate the effects are driven by differences between facilities rather than Zipline effects.
2. Varying distance of assumed auto correlation between units from 5 to 15 and observing the sensitivity of estimates to choice of distance
3. Replicating the analysis after excluding facilities that are within 10 km of the zip border to check for spillover effects.

¹⁵ $1-(1-0.05)^5=0.22$

¹⁶ Benjamini, Yoav, Abba M. Krieger, and Daniel Yekutieli. "Adaptive linear step-up procedures that control the false discovery rate." *Biometrika* 93, no. 3 (2006): 491-507.

Data

The following data collection is conducted throughout the project.

Table 1: Summary of topics covered in each survey

Survey	Topic	Baseline	Midline Surveys	Endline
Maternity health worker	This survey tracks maternal health outcomes and stocking of maternal products.	X	X	X
General health worker	This survey tracks the treatment behaviour of general health workers, and health worker satisfaction.	X	X	X
Vaccine stocking	This survey tracks vaccine stocking (vaccine availability).	X	X	X
Medicine stocking	This survey tracks medical product stocking (medicine availability).	X	X	X
Blood stocking	This survey tracks blood stocking (blood availability).	X	X	
Facility head	This survey tracks facilities' equipment, infrastructure, logistics (including referrals) and staff.	X		X
Stockroom attendant	This survey tracks medicine stocking (product availability) and stocking behaviour, such as restocking actions and timelines for restoring out of stock products.	X		
DHIMS	This survey tracks patient health outcomes for specific conditions using administrative data records.	X		X
Patient exit interview	This survey tracks treatment received by patients in cases where medicine is in stock and out of stock. It also tracks patients' treatment seeking behaviour.	X		