STATISTICAL ANALYSIS PLAN

Incentivising fruit and vegetable consumption in urbanising India

London School of Hygiene and Tropical Medicine, UK,

National Institute of Nutrition, India,

Indian Institute of Public Health Hyderabad, &

Indian School of Business

Prepared by: Poppy Mallinson

Approved by: Sanjay Kinra (PI)

Sanjay Kinra

Signature Date: 29th April 2021

Contents

1. Introduction

This Statistical Analysis Plan (SAP) describes the planned quantitative analyses and reporting of results of the cluster randomised trial component of the study "Incentivising fruit and vegetable consumption in urbanising India".

Full description of background and methods of the study are provided in the study Protocol (Version 2.0, 22-03-2021), which should be read alongside this SAP. Reporting of results of the process/qualitative evaluation components of the study are not covered in this SAP (refer to the Protocol for a description of these).

2. Objectives

The aim of this study is to develop and evaluate a sustainable business model involving financial incentives to increase consumption of fruit and vegetables (FV) amongst peri-urban communities in India.

The study objectives are:

1. To implement a sustainable financial incentive scheme in 3 villages (with another 3 villages as a control group) for 3-months and conduct a process and outcome evaluation.
    1.1. To conduct a survey of community members to assess whether the intervention results in a short-term change in fruit and vegetables (FV) purchase.
    1.2. To assess stakeholder satisfaction with intervention (local community, vendors, incentive partners) and evidence of any unintended or perverse effects (e.g., misuse of incentives, price fluctuations or supply constraints).
    1.3. To review the findings with stakeholders and policy makers to establish the need for a full-scale trial or a pragmatic evaluation of the scheme in an implementation mode.

This SAP pertains to the statistical analysis for objective 1.1. The primary hypothesis for this objective was that an incentive-based intervention for increasing FV purchase would be associated with greater quantity of FV purchased by households (primary outcome), and greater amount spent on FV by households, quantity of FV obtained from all sources by households, and quantity/value of FV sold by village vendors (secondary outcomes).

3. Study design

For objective 1.1, the study design is a cluster randomised trial nested within the 29 villages participating in the APCAPS study (Ranga Reddy district, Telangana). Six comparable villages were selected based on pre-defined criteria, and randomly allocated to receive a community-wide fruit and vegetable (FV) incentives intervention (intervention villages) or no intervention (control villages). The intervention is a coupon-based incentive scheme implemented in partnership with village FV vendors. The intervention will be implemented for 3 months.

Baseline (prior to intervention) and end-line (after intervention) telephone surveys will be conducted over a week on a random sample of households from the study villages (N=1500) to assess households' FV purchasing over the study period. Baseline, mid-line (x2) and end-line telephone surveys will also be conducted with all FV vendors in the study villages (N=38) to assess FV sales over the study period.

4. Statistical analyses

4.1. General principles

All statistical analyses are pre-specified. All data management, cleaning and analysis will be blinded (i.e., the analyst does not know which trial arm is which). Prior to analysis, data will be visually inspected for any implausible values (extreme outliers), which will be investigated by cross-checking with paper questionnaires and contacting participants where possible. Any remaining implausible values will be assumed to be data errors and set to missing.

For the community survey, all descriptions and analyses will be done at the household level. For the vendor survey, all descriptions and analyses will be done at the vendor level. Evaluation of the intervention effects will be performed on the principle of 'intention to treat' unless otherwise specified (i.e., community members will be assigned to intervention/control according to the village they reside in, regardless of whether they personally participated in the incentive scheme).

All statistical tests will be two-sided. Analyses will be performed using STATA Version 16.

4.2. Baseline description and comparison

For descriptive analyses, continuous variables will be presented as means and standard deviations, or medians and inter-quartile ranges if data are skewed. Categorical variables will be presented as frequencies and percentages.

4.2.1. Community survey

Number of households attempted to be contacted, successfully contacted, consented to participate in the study, and completing and partially completing the telephone surveys at baseline and end-line will be tabulated by intervention group, and presented in a flow diagram.

Baseline demographic variables and information on FV purchasing in participating households will be summarised by intervention group. Chi-squared tests (proportions of categorical variables) or t-tests (means of continuous variables) for difference between intervention and control group will be presented. Characteristics which will be described include:

- Number of household members
- Proportion of household members <18 years old
- Household dependency ratio (ratio total n members: n economically active members)
- Highest education level in household
- Household asset score (baseline survey items 3.1-3.14 combined using PCA)
- Whether household produces any FV for selling
- Whether household produces any FV for own consumption
- Whether village market is a main source of FV for the household
- Weekly quantity of FV purchased (for consumption) in kg (total and per capita)
- Weekly cost of FV purchased in rupees (total and per capita)
- Weekly quantity of FV obtained from other sources (e.g. grown, foraged) in kg (total and per capita)

4.2.2. Vendor survey

Number of FV vendors attempted to be contacted, number and % successfully contacted, number and % consented to participate in vendor surveys, number and % completing and partially

completing vendor surveys at baseline, midline 1, midline 2 and end-line, will be tabulated by intervention group.

Baseline data on purchase and sales of FV by the vendors will be summarised by intervention group, for the following variables:

- Quantity of FV purchased (for sale) in last week in kg
- Cost of FV purchased in last week in rupees
- Quantity of FV sold within the village in last week in kg
- Cost of FV sold within the village in last week in rupees.
- Number of days operating within the village in last week.

4.3. Primary outcome analysis

The primary outcome is the quantity of FV purchased (in kg) by the households in the past week. Quantity purchased will be treated as a continuous outcome. Appropriate continuous transformation of data will be performed if normality of residuals assumption is violated.

Difference between mean outcome in intervention vs. control participants will be assessed using multilevel linear regression (95% confidence intervals and Wald test p-values). A random intercept term at the village level will be used to account for clustering in outcomes within villages, and the Kenward-Roger correction will be used to account for smaller number of clusters[1].

Main models will be adjusted for baseline quantity of FV purchased by the household, baseline village-level mean FV purchase (to improve precision)[1], household size, proportion of household members who are <18 years of age, household dependency ratio, highest education level in household, household asset score, whether village market is household's main source of FV, and weekly quantity of FV obtained from other sources by the household (excluded any sparse or colinear variables). Unadjusted models will also be presented, along with the village-level clustering coefficients for each model.

As a robustness check, the analysis will be repeated using regular linear regression with the outcome of change in quantity of FV purchased between baseline and end-line, a fixed effect for trial arm, and with standard errors clustered at the village level.

4.4. Secondary outcomes analyses

For the amount spent on FV (in rupees) by households in a week, analysis will be conducted as for the primary outcome.

For the quantity of FV obtained (in kg) from all sources by households in a week (i.e. sum of purchase, own production, gifts, wild harvest and others), analysis will be conducted as for the primary outcome (except without adjustment for quantity obtained from other source).

For the quantity (in kg) and value (in rupees) of FV sold by vendors in the village in the past week, the unit of analysis will be the follow-up observations of each individual vendor (i.e. 3 per vendor). Three-level multilevel linear regression models will be used, with observations clustered within vendors (level 1) clustered within villages (level 2). Difference between mean outcomes in intervention vs control villages will be compared (using 95% confidence intervals and Wald test p-values with Kenward Roger correction as previously). Main models will be adjusted for baseline levels of the outcome (individual vendors and cluster means) and total population of village. An interaction between intervention and follow-up number (midline 1, midline 2 or end-line) will be

examined (by Wald test) to test the additional hypothesis the effect of intervention on vendors' FV sales increased with duration of time since the intervention began. In the situation that the statistical models cannot be estimated (e.g., due to limited sample size), emphasis will be placed on the descriptive visualisation of the data accompanied with appropriate non-parametric tests of difference between study arms.

All analyses will also be repeated for the above primary and secondary outcomes considering fruits and vegetables separately (rather than a combined total).

4.5. Sub-group analyses

Pre-specified sub-group analyses will be conducted as for the primary outcome, with inclusion of an interaction term between intervention group and the following:

i) Socioeconomic position: whether household has above or below median asset index score (hypothesis: households with low socioeconomic position will respond better to the incentive scheme as price of FV is a more significant barrier to purchase for this group)
ii) Baseline FV purchase: whether household was in the bottom quartile of weekly FV purchasing (in rupees) at baseline (in rupees) (hypothesis: households which spent very little on FV at baseline will be less able to benefit the incentive scheme).

For each sub-group analysis, we will present stratum-specific intervention effect estimates and Wald test p-value for the significance of the interaction.

We will also conduct an exploratory per-protocol analyses in which the exposure will have three categories: control households, intervention households who reported never availing the incentive scheme (non-responders), and intervention households who reported some use of the incentive scheme (responders) (hypothesis: households which reported availing the incentive scheme will have higher baseline-adjusted mean FV purchase).

4.6. Missing data handling

Potential missing data can be categorised into 3 types:

a) Non-response – some households decline to participate in the study from the beginning.
b) Incomplete weekly reporting data – some participants are not reachable for every call during the week-long FV purchasing survey, resulting in partial information for that week.
c) Loss to follow-up – participants who completed baseline surveys do not participate in subsequent surveys.

For a), we will exclude all non-responders from the study. We will present descriptions of the differences between responders and non-responders by household size and socioeconomic position (asset index) based on a previous household survey we conducted in 2014.

For b), we will examine and identify patterns of FV purchasing over the week among those with complete data (separately for each village as they have different market days). If clear patterns emerge (e.g. households purchase the majority (e.g. >80%) of FV on village market day only), we will impute "0" FV purchase on non-market days, but will not impute values on market days (and instead consider the whole week to be missing for that participant). We will do this manually as we expect the proportion of partial-completers to be very low.

For c), we will present descriptions of the baseline differences between those followed-up and those not followed-up (including all baseline characteristics in 4.2.1). We will use attrition weights to re-weight the sample according to its baseline socio-demographics and FV purchasing if i) loss to follow-up is high (>20%) and ii) baseline comparisons (see 4.2.1) show significant differences ($p<0.05$) in characteristics between those followed-up and those not followed-up. If these conditions are not met, we will conduct a complete case analysis as the primary analysis.

5. References

1 Hooper R, Forbes A, Hemming K, Takeda A, Beresford L. Analysis of cluster randomised trials with an assessment of outcome at baseline. *BMJ* 2018; **360**: k1121.