Final report

International Growth Centre

Impact evaluation of a public health insurance plan in India

Post health event survey pilot

Anup Malani
Cynthia Kinnan
Alessandra Voena

February 2017

When citing this paper, please use the title and the following reference number:
S-35302-INC-1

DIRECTED BY
LSE
UNIVERSITY OF OXFORD

FUNDED BY
UKaid
from the British people

Impact Evaluation of a Public Health Insurance Plan in India: Post Health Event Survey Pilot

FINAL REPORT

Authors: - Anup Malani[1], Cynthia Kinnan[2], Alessandra Voena[3]

[1] University of Chicago

[2] Northwestern University

[3] University of Chicago

Section I: Introduction

Hospitalization (or inpatient care) is required for treating the most severe health shocks. Yet, the cost of hospitalization is a primary driver of financial distress for low---income families. **Hospital insurance tackles both problems by enabling access to inpatient treatment while reducing the financial burden of such treatment**.

The impact of hospital insurance, as with many other public policy innovations, is typically measured using a combination of experimental (or quasi---experimental) methods and household---level data collection. There is a growing interest among academics and policymakers in improving the efficiency of such household---level data collection. For instance, McKenzie (2012) explored the tradeoff between cross---section and time series dimensions of data. Ligon (2016) develops and investigates alternativesto comprehensive, expensive expenditure modules as a measure of welfare. Perhaps closest in spirit to our exercise, Broockman, Kalla, and Sekhon (2016) explore options to design sample frames for online surveys to maximize response rates, thus reducing the number of households who must be recruited.

We address a related – but distinct – set of challenges with measuring the impact of insurance with survey data. Hospitalization is a low frequency event because severe health shocks are fairly infrequent and because needed hospitalization is often foregone due to cost. In India, only 38 of every 1000 persons are hospitalized each year according to theNational Sample Survey Office (NSSO) 2014 'Health in India' report.[1] Consequently, a study of hospital insurance typically requires large, costly samples in order to be powered to detect economically meaningful effects (Cf. Kaboski and Townsend (2011)). Moreover, low frequency (e.g., annual) surveys typically employed in impact evaluations can suffer substantial recall error, especially when measuring low frequency events. Recall error, which increases with the time that has passed since the event(de Nicola & Giné, 2014), in turn reduces power and may introduce bias if recall error is correlated with shock severity. These issues are not unique to hospitalization: other low frequency – but economically important – events include serious crimes, business closures, divorce/desertion/widowhood, and pursuit of advanced degrees.

We proposed a new survey method and instrument, called the Post Health Event Survey (PHES), to mitigate the twin problems of costefficiency and recall error of annual surveys. **First, to address costs**, instead of surveying all households in person, the PHES surveys only a select subsample households in person. The subsample is selected by calling each enrolled household every two months to determine if any household member suffered a serious health event in the last two months. (In the PHES Pilot, given the abbreviated study duration, each household was contacted once.) The PHES then surveys in person only those households that report such an event. (In the PHES Pilot, a small subsample received the in---

[1] This rate varies by the patient's location: the urban hospitalization rate is 44 per 1000 while that of rural India is lower, 35 per 1000 [pg A50]. The rate of patient's reporting an ailment is 89 per 1000 for rural and 118 per 1000 for urban [pg. S1---2].

depth survey by phone.) A person suffering a serious health event is much more likelyto consider and experience hospitalization.[2] Therefore, fewer in---person surveys are required to observe any hospitalization.

Second, to address recall error, the PHES phone survey only asks about health events in the last 2 months and the PHES in---person surveys are administered to each household that reports a serious health event within 14 days of screening. Thus, the maximum period between a health shock and a subject's response to a survey about it is 2.5 months. By contrast, an annual survey of households would have a lag of up to 12 months, or even longer if the survey is begun at the year anniversary and takes several months to complete. Prior research suggests that reducing the time between an event and a survey reduces recall error (de Nicola & Giné, 2014).

The PHES is a module that we are incorporating into an ongoing randomized control trial experiment, the India Health Insurance Experiment (IHIE). The IHIE, launched in 2013, evaluates a live policy question: what are the health and financial benefits of expanding Rashtriya Swasthya Bima Yojana (RSBY), India's largest public health insurance scheme, to economically vulnerable populations of just above poverty line (APL) households? The study takes place in two districts in the state of Karnataka, one district representing central India (Gulbarga) and one representing southern India(Mysore). In its present form, RSBY targets Below Poverty Line (BPL) households, which represent nearly one quarter of India's population; the IHIE has enrolled roughly 11000 APL households that are not currently eligible for RSBY but would be if eligibility were expanded to APL households. A baseline and a post---treatment annual survey are the primary instruments employed to gather data in the HIE. The PHES is an innovation designed to reduce survey costs and recall error in collecting hospitalization data relative to the annual survey.

A pilot of the PHES was launched in December 2015 in order to assess the value of a full---scale rollout of the PHES and to optimize data collection and recall under that rollout. In this report, we assess the lessons of the PHES pilot, including its impact on recall and survey costs, as well the challenges and guidelines for implementation. Section II provides background on the IHIE, the larger context within which we implemented this new survey instrument. Section III describes the methodology of the PHES pilot. Section IV details characteristics of the sample households and the surveys implemented in the pilot through summary statistics, including data on the time (and thus monetary) cost of the survey. Section V discusses the results of the PHES Pilotin terms of reducing recall error. Section V concludes.

Section II: Background on the IHIE

There is a pressing need for experimental evaluations to inform the policy questions surrounding RSBY, India's first national health insurance scheme, which was adopted in 2008. The current scheme has already enrolled 150 million below poverty line (BPL) beneficiaries and aims to extend coverage to 300 million BPL persons. Alongside these ambitious goals, policy debates are underway over how RSBY should be reformed to provide more coverage and to more groups in Indian society.**Nonetheless, to**

[2] In the NSSO 2014, whereas the hospitalization rate is 38 per 1000 persons on average, the number of ailments reported during the last 15 days is 101 per 1000.

date, no rigorous evaluation of RSBY, let alone of an expanded RSBY, has been conducted. Nearly all existing studies of RSBY have failed even to include a contemporaneous control group (Hou & Palacios, 2010; Palacios, 2010; Sun, 2010). With one exception (Das & Leino, 2011), all prior studies are non---experimental. The resulting evidence is of limited use because it confounds the true effect of insurance with selection bias; for instance, those who are sicker or wealthier may be more likely to take up insurance. Moreover, non---experimental attempts to address selection bias are ineffective when selection depends on numerous factors, which cannot be measured (Imbens, 2003).

The IHIE seeks to address deficiencies in prior studies. The IHIE examines the impact of expanding RSBY to cover above poverty line (APL) households that are not covered by RSBY or other secondary hospital care insurance plans. It enrolled roughly 11000 APL households (comprising over 50000 individuals) in two districts of Karnataka, one representing central India (Gulbarga) and another representing south India (Mysore).

The IHIE randomizes these households to one of four arms: free insurance through the RSBY program (treatment group 1), an income transfer equal in value to the RSBY premium and the opportunity to buy RSBY (treatment group 2), just the opportunity to buy RSBY (treatment group 3), or no treatment (control group). Each of these arms is intended, by itself or in combination with another arm, to mimic different policy approaches to achieve universal coverage – from free insurance for all, to subsidized premiums, to a public option in the insurance market– or to serve as a concurrent control group. This design will also allow us to separate the causal effects of health insurance from the causal effect of premium subsidies.

Through this rigorous randomized control trial (RCT) design, the IHIE aims to measure a range of key outcomes related to the health and financial impacts of RSBY. Primarily, we aim to measure the impact of health insurance on healthcare usage, health expenditure and health status.The study will examine whether such insurance increases hospital utilization and how it affects healthcare expenditure and health outcomes. In addition, we aim to measure the impact of health insurance on financial status, including the impact on non---healthcare consumption and on financial distress from health shocks. Finally, we also aim to measure willingness to pay for health insurance and the effect of paying for health insurance on the utilization of health insurance.

As initially designed, the IHIE included a baseline survey and a post---treatment annual survey of all enrolled households. The baseline survey was completed in 2014, treatment assignment took place in 2015, and a post---treatment annual survey (midline) is planned to begin in fall 2016 to measure the first year impact of treatment. Both the baseline and midline include health, medical, cognitive, and financial modules. In order to reduce the cost and recall error associated with observing hospitalizations, the key outcome for which the study was powered, we planto implement a full---scale PHES starting in fall 2016 and lasting 1 year to measure the second year impact of treatment.

With the addition of the PHES, the HIE will collate multiple sources and types of data to truly provide a 360---degree perspective on RSBY. Given the design of the study, which varies access to and the price of insurance, the PHES will shed light on how a health insurance scheme designed for a developing country affects healthcare---seeking behavior and the nature of healthcare interactions given the facilities in that country. It will also be able to ascertain how insurance affects household finances, and whether the price of insurance affects utilization. Most significantly, the PHES will help eliminate recall error, particularly long---term biased recall, and thereby improve the statistical precision of the study.

Section III: PHES Methodology

The PHES is intended to be a high---frequency survey instrument that measures how households seek and finance healthcare in the face of a health event soon after the event occurs. Because it focuses on those who have health events, it saves on the costs of surveying households that are very unlikely to seek inpatient care. As stated earlier, as the maximum period between a health shock and a subject's response to a survey is about 2.5 months as opposed to 12 months or longer under an annual survey, data gathered under a PHES about healthcare---seeking behavior and financing will have less recall error.

The PHES has two phases – *screening* and *surveying*. First, in the *screening phase*, we *identify* households that experienced an adverse health event in the past two months to improve targeting of households most likely to have experienced hospitalization. Second, in the *surveying phase*, we *survey* the identified households within a few weeks about their healthcare---seeking and health financing behaviour soon after the event to reduce recall error.

In the scaled up PHES, we will cycle through these two phases – contacting each sample household every two months – throughout an entire year. In the PHES pilot, selected households– every sample household for which we had a phone number– received only 1 screening call during a 5---month period.

Screening Phase

In the pilot**, the *Screening Survey* was conducted by phone over a period offive months** to the 8364 sample households (out of 10879 sample households enrolled in the study) for which we had phone numbers on record.

During this phone call, we ask a total of nine questions to identify a qualifying health event for the in--- depth PHES. We briefly describe three categories of eligible health events (a childbirth in the past two months, an accident that caused the victim to miss at least two days of normal daily activities like work or going to school or doing housework, or a physical functional limitation such as the inability to eat normally, dress oneself, walk with ease or perform basic household tasks) and ask if any individual inthe household suffered such an event. Whether householdswho experienced one of these eventsare "screened in" to a full survey of the eventdepends on the nature of the event. Specifically, our screening algorithm was as follows:

1. **Child Birth:** If anyone had delivered a child in the past 2 months, they were screenedin immediately.
2. **Accident:** Else, if anyone had met with an 'accident'*and* missed at least 2 days of work as a result of this accident, they were screened in.
3. **Functional Limitations:** Else, if any individual in the household was reported as having*three or more* of the functional limitations we asked about, they were screenedin.

If there was more than one individual within a category, we went to the household with all screened---in names and asked the respondent or the head of the household which illness they thought affected the household most defined as the individual or health event with the highest priority for the survey

household. We then proceeded to administer the survey about that person. The screening survey took only four to seven minutes per household.

We do not ask outright about hospitalization or healthcare---seeking behaviour in the Screening Survey as we want to capture what a household does when a health event occurs and whyin the in---depth survey. This encompasses a range of behaviours, including the decision to not seek healthcare, only partially seek healthcare, or choose hospitalization.

Towards the end of the screening survey period,**we decided to add a question to the screening survey** that asked directly whether someone in the household was hospitalized or not. With this question, we aimed to determine whether there were health events resulting in hospitalization that our screening criteria were failing to capture.

Table 1. Summary of HH surveyed

Calls Made (HH)	Screening Surveys Administered (HH)	HH with Qualifying Health Event
10879	8364	822

Surveying Phase

For the in---depth survey component of the PHES, we experimented with two different mediums of surveying. For a small, random subset of our sample (~20%), we**administered the in---depth PHES over the phone** to test a survey strategy that would result in even greater cost savings. The remaining households received **an in---person survey (PHES I)**. To identify this subset, we chose a village in each district at random and then successively selected the next closest village, toeconomize travel costs during the in---person survey. Among the households that received an in---person survey, 112 households received **a revised in---person survey (PHES II)** between April and May 2016. For the PHES II, the survey was revised to reflect colloquial Kannada (the local language) andwe undertook an improvement in surveyor training, described in more detail below.[3]

Table 2. Summary of HH surveyed with in---depth PHES

Instrument	PHES Phone	PHES I [Dec – Mar 2016]	PHES II Revised [Apr – May 2016]
Gulbarga	73	302	42
Mysore	72	268	70
Total	**145**	**570**	**112**

In the in---depth PHES, we gather information on where and why households choose to seek and finance healthcare in addition to gauging their use of health insurance products available to them.

[3] All subjects receiving the phone survey were surveyed using the PHES I version of the survey instrument.

- **Section A** asks for details regarding the nature of the symptomsobserved by the patient, where they sought healthcare, what sort of treatment they received, to what degree they followed medical instructions, and their experience of availing themselves of medical care.
- **Section B** collects information on the specifics of healthcare expenditure. We ask detailed questions about their use of health insurance, various expenses incurred while accessing healthcare services, and what proportion of these expenses were financed out---of---pocket.
- **Section C** explores the other means through which the family financed the healthcare– whether other family members migrated temporarily or worked extra hours to make up for the loss of income, if jewelry or other household items were sold or pawned to raise money, etc.
- **Section D** attempts to understand the impact of the health event on other related behaviour within the family, including whether food consumption patterns changed for the rest of the household, and whether any major household expenses had to be put off as a consequence of the health shock.

After the first round of the pilot in March 2016, our analysis of the initial data alerted us to some awkward contradictions in respondent answers, leading us to revise both thein---depth PHES and Screening Survey as well as how we trained surveyors.A large subset of respondents answered affirmatively to having RSBY but negatively to having health insurance. We hypothesized that perhaps the questions in the PHES were phrased incorrectly in both English andKannada. Specifically, we hypothesized that perhaps the direct translation for "health insurance" in Kannada was the wrong term to denote the meaning we wanted to convey. Thus, in consultation with a health ethnographer at the University of Pennsylvania, Prof. Vani Kulkarni, we revised the survey instrument to improve the language and to include a photocopy of the RSBY smart card to show to the respondent so that they could identify whether they had one more accurately. On the advice of Prof. Kulkarni, wealso introduced an intervention in the surveyor training. During training for surveyors in Gulbarga, where the second round of the pilot was implemented, the surveyors were made more aware of the PHES pilot's general purpose, given greater context about the experiment and RSBY, and given an explanation of the purpose of each section of the PHES and how the sections related to each other. We launched the second round of the PHES (PHES II) that incorporated all these changes in April 2016. The second round concluded in May 2016.

Survey sample

As noted above, **the Screening Survey was administered to 8364 sample households** for which we had phone numbers on record, out of 10879 sample households enrolled in the study. **For the 2515 households in the sample for which we did not have active phone numbers, we attempted to contact the household through their neighbors.** To update our database, we attempted to trace their current phone numbers by calling other households in the same village and asking them for the contact details of the household with the missing phone number. Sometimes, the other households would have this contact information available with them. Other times, they would ask us to call back at a scheduled time, and would then carry their cellphones over to the relevant household so that we could administer the screening survey and collect an updated phone number. This exercise helped us verify the numbers of 8364 households.

Summary statistics from the Screening Surveys andthe in---depth PHES are provided in the following section.

Section IV: Data Description and Summary Statistics

In this section, we present summary statistics describing the data collected from the PHES screening and in---depth surveys. We first discuss characteristics of the sample households.Then, we present statistics on the surveys themselves (duration, etc.).

Characteristics of sample households

In total, 8364 screening calls were conductedin the Screening Survey: 3977 in Mysore and 4387 in Gulbarga (Table 3). In both districts, the screen---in rate was approximately 10%, consistent with expectations based on NSSO data. (We discuss the comparison with NSSO data in more detail below.)

Table 3. Summary of households (HH) screened in

	Total Screening Calls	Screened---In	Percentage
Mysore	3977	408	10.26%
Gulbarga	4387	414	9.44%
Total	**8364**	**822**	**9.83%**

Of 822 screened---in households, 715 received an in---depth PHES. Of these, 570 were conducted in person, and 145 by phone (Table4). Of the 107 households who were screened in but for whom we do not have in---depth survey data, approximately 20 householdsreceived the in---depth survey, but the data was lost due to software erroron the first data of data collection. The remaining approximately 87 households did not receive an in---depth survey because the survey activities were stopped after the team had completed 715 in---depth surveys due to time constraints.

Table 4. In---person vs. phone PHES I

Mode of Survey	Number of Surveys
In---Field PHES	570
Phone PHES	145
Total	**715**

A key measure of the effectiveness of the PHES is rates of hospitalization among those who were screened in to the in---depth survey (Table 5). **Among those who answered the question, 47.5%of respondents to the in---depth survey reported that they had experienced hospitalization related to the event for which they were screened in.**This indicates that our screening survey is indeed effective at identifying households who are likely to experience hospitalization, since the overall incidence of hospitalization in India is 0.38% perannum (NSSO 2014).

Correcting for the fact that only 9.83% of the sample who received the screening call received the in---depth PHES (and assuming no hospitalizations among the screened---out households), **we obtain an**

implied overall hospitalization rate of 4.67% across our sample over a roughly 3 month period. The fact that this is significantly higher than the NSSO rate reflects several factors:1) households in the treatment arms had access to RSBY, which covers inpatient costs; 2) our sample was constructed on the basis of living within 5 km of a hospital and is comprised ofAPL households who are likely better able to afford the costs of hospital care than average among the Indian population;and 3) the shorter recall of our survey likely allowed us tomeasure hospitalizations which might have been under---reported in an annual survey. (We analyze evidence on improved recall/reduced measurement error below in Section V.)

Table 5. Reported hospitalization

Hospitalization Status	Number	Percentage
Yes	311	46.91%
No	344	51.89%
Refuse to Answer	8	1.21%

Table 6 reports on the breakdown of screened---in respondents by category of serious health event. The most common was functional limitations, with 41.1% of all screened---in households. Next was childbirth, with 32.4%. Accidents accounted for the remaining 26.4%.

Table 6. Screen---ins by category

Reason	Number	Percentage
Childbirth	189	32.45%
Accident	232	26.43%
Functional Limitations	294	41.12%

While a majority of households reported only one health event, somehouseholds (12.6% of those called) reported multiple serious health events. Table 7 reports on the breakdown of number of events per household. Though the pilot only collected data on one health event, collectingdetailed information on all events experienced during the lookback period may allow the collection of more complete information. Thus, collecting data on multiple events will be something we consider forthe scaled up PHES.

Table 7. Number of respondents screened In per HH

# Respondents (per HH)	# of HH	Percentage
1	720	87.37%
2	90	10.92%
3	14	1.67%

Characteristics of the surveys

Two sets of characteristics are important to evaluate the efficacy of the PHES relative to standard low--- frequency surveys. One set of characteristics, which we discuss here, is time (and hence monetary)

cost. The other is efficacy in capturing low---frequency events and reducing measurement error, which is addressed in the following section (Section V: Data Analysis).

First, we examine statistics regarding the duration of the screening calls, the in---depth PHES carried out in person, and the in---depth PHES carried out over the phone(Table 8). Screening calls took an average of just under 7 minutes, with a long right tail as evidenced bythe fact that the median duration is roughly 4 and a half minutes and the standard deviation is roughly 8 and a half minutes. These long---duration calls likely reflect households who had difficulty understanding the questions and required more explanation by the surveyor.

The in---person PHES (in---depth survey) took an average of 26 minutes, with a median of just under 20 minutes, again reflecting some right---skewness in the distribution. The standard deviation is just over 15 minutes. Variation in the length of the in---depth survey likely reflects comprehension, as with the screening calls, but also the nature of the survey and its skip patterns (e.g., if no hospitalization was reported, no questions about the nature of the hospitalization will be asked).

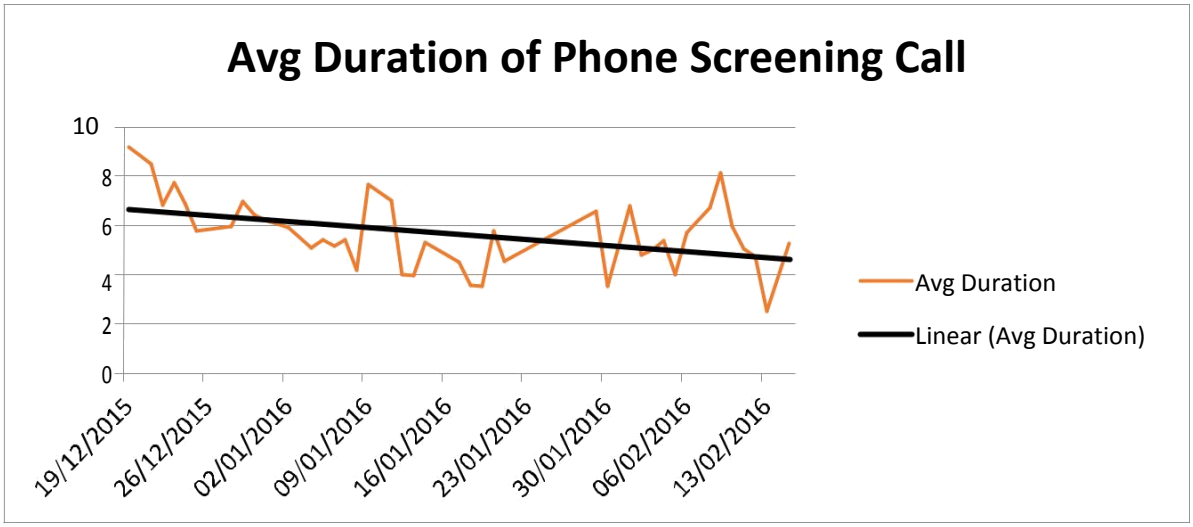
The phone---based PHES took longer on average than the in---field version: an average of 36 minutes and a median of 35 minutes. The lack of right skewness, as well as the slightly lower standard deviation of the phone---based vs. in---field PHES, may reflect that it is somewhat more difficult to provide clarification or ask follow up questions over the phonethan in person. Moreover, the fact that the phone---based survey takes longer overall may dissuade surveyors and/or respondents from engaging in clarifying discussions.

Table 8. Duration of Surveys

Mode of Survey	Average Duration	Median Duration	Std. Deviation
Screening Calls	6.78 min	4.4 min	8.47 min
In---Person PHES	26.18 min	19.66 min	15.27 min
Phone PHES	35.98 min	34.95 min	13.91 min

Finally, we examine trends in surveyor efficiency, finding that surveyors are able to more efficiently deliver the PHES over time. Since the PHES (particularly the phone---based screening survey) is a novel instrument, we hypothesized that it is possible surveyors will take time to become familiar with, and efficient at, administering it. Figure 1 analyzes the average duration of phone screening calls over time. There is evidence of a decline in average duration over time, from over 8 minutes early on to less than 6 minutes toward the completion of the screening calls. Thus, efficiency in a full scale up of the PHES pilot (such as we are undertaking beginning in fall 2016) may realize higher efficiency than the pilot since the learning costs can be amortized over more surveys.

Figure 1. Average duration of the screening survey over time



Section V: Data Analysis

Our goal in this section is to examine to what extent the data collection methods used in the PHES reduce recall error, relative to data that would be collected in a traditional survey with longer recall periods, such as a standard annual survey. As noted above, reducing measurement error is the second sets of characteristics important to evaluate the efficacy of the PHES relative to standard low---frequency surveys. Through heaping and dispersion analyses described in detail below, we find that the PHES reduces recall error for monetary values but does not have a clear effect on recall error for time---related measures.

Related literature

In order to contextualize our subsequent analysis of recall error in the PHES pilot data, wediscuss some of the existing literature on recall error that informed the PHES. We are not the first to note that the time elapsed between an event and a survey can have implications for the quality of data collected**.** De Nicola and Gine (2014) examine this issue by comparing administrative records with survey responses regarding income and assets among self---employed households in India. They find thatrecall error increases over time, and that it varies cross---sectionally across households in predictable ways.

While recall error increases over time, there are tradeoffs to surveying too frequently. Arnold et al. (2013) discuss the tradeoff between recall error and precision: a short recall period minimizes recall error but reduces precision because fewer events are captured;the opposite is true for a longer period. They examine multiple datasets containing information on child health from a variety of countries, with differing recall periods, and, in the context of diarrhea, cough, and fever, they find that a 7---day recall period is optimal.

In our context, where serious health events are more infrequent, a two---month recall period was chosen to minimize recall error while maintaining a meaningful sample size. That is, if we only conducted the in---depth survey for health events that had occurred in, say, the last one week, we might have high accuracy but the odds of catching a household in the week after a serious health event is low, so we would capture few events, leading to a small sample. Annual surveys take the opposite approach,

asking about events over a long interval (one year), which leads to capturing more events (a greater sample size) but less precise recall. To our knowledge, we are the first to propose using a short screening survey combined with an in---depth survey among a "qualifying" population in the context of a randomized intervention.

We next turn to examining whether, over the range foundni our data, the recall interval is associated with more error, defining error in a specific sense that we describe in more detail below.

Methods of analysis

In some ways, the ideal comparison to look at the effect of recall interval on error in our contextwould be to compare the PHES data to data from a survey with a longer recall period, covering similar questions. We will be conducting such a survey—the IHIE midline survey—as well as an expanded PHES, which will be heavily informed by this pilot, in fall 2016. When both data sources—midline and PHES—are available, we will compare the data gathered to gain additional insight into this question. At present, only the PHES data are available. Therefore,**we will use (quasi---random) variation in the time elapsed between the occurrence of a health event and administration of the in---depth PHES questionnaire to examine the effect of recall interval on error.**

This variation in recall interval arose primarily due to variation in the length of time elapsed between the health event and the screening call.[4] This time could have been as long as two months or as little as a few days. This variation is quasi---random insofar as households were called at random, and therefore there should be no systematic differences between the characteristics of someone who happens to get a call a few days after an event (say, an illness) vs. 2 full months later.

Measuring recall error: heaping and dispersion

In our analysis, we aim to measure the impact of recall interval, defined asthe total time elapsed between the start of the health event and the in---depth survey, on error in survey responses. Consistent with prior literature and the motive behind the PHES pilot,we expect responses (to questions about days of missed work, total expenditures on the event, etc.)collected when the recall interval is short to exhibit less error than those when the intervalis long. However, we do not know the objective truth. Therefore, **we will use two measures of data qualityto assess error: heaping and dispersion.**

By heaping we mean: is there excess mass of data at"round numbers"? In the PHES, depending on the range of the variable, this could be multiples of 5 or 10 if the variable is days of work missed or multiples of 100s or 1000s of rupees if the variable is monetary. This is a standard measure of data quality(Beegle, Carletto, & Himelein, 2012; Crawford, Weiss, & Suchard, 2015). The intuition behind this measure is that greater heaping indicates that more respondents are providing an inferred or estimated answer—in which case they are more likely to generate a round number—rather than providing the actual answer based on memory. Thus, greater heaping indicates greater recall error.

By dispersion we mean a measure of the variability of the data, such as its standard deviation or coefficient of variation. The intuition behind this measure is that the observed value of a variable is

[4] Additional variation arose due to the time elapsed between the screening call and the in---depth survey, however this variation was typically smaller. As with the main source of variation, there is no systematic correlation between individual characteristics and this time interval.

equal to the truth plus measurement or recall error, assumed to be classical, i.e. uncorrelated with the true value. That is, the measure collected in the survey, is comprised of the "true" value, * plus an "error," . Mathematically, this can be expressed as = * + . If the variance of the error, μ, increases with the recall interval, then the variance/dispersion ofthe measured value will also increase with the recall interval. However, De Nicola and Gine argue that, at long recall periods,respondents "resort to inference rather than memory". Inference could amount to using some variables (call these) to form a "best guess" and reporting the best guess, = + , when the recall interval is long. On the other hand, the household may not resort to inference when the interval is short, so they report as above. If only a fraction of the true variation in * is captured in , the result could be *reduced* dispersion. Thus it is an empirical question whether we more, less, or the sameamount of dispersion under different recall intervals.

Results: Heaping

We will examine heaping for several variables: total out of pocket expenditure, total medicine out of pocket expenditure, days in hospital, and days missed work. These were chosen because they are non--- binary (binary variables cannot exhibit classical measurement error) and important for understanding the economic consequences of serious health shocks. For each variable, we will examine whether instances with above---median recall intervals exhibit differential heaping, relative to those with below--- median recall intervals. While some heaping will be present in the true data if prices of, .eg., medical procedures or medicines tend to be round numbers, greater heaping should only be seen fordata with longer recalls if it is a consequence of recall error.

We present the results in the form of two---way tabulations, structured as follows. The columns divide observations into cases with a long recall (short recall=0) vs short recall (short recall=1). The rows divide observations into cases that are not heaped (round=0) vs those that are heaped (round=1). The unit used to define heaping varies depending on the variable, as explained below. We conclude that there is more heaping for long recall periods if the short recall=0 column has a higher percentage of heaped observations than the short recall=1 column. We also report p---values for the difference in the share of heaped observations; these are computed by regressing the "round" variable on the "short recall" variable and a constant term (regressions not reported).

The first variable we examine is the total out of pocket expenditure (OOP) reported by the respondent as a consequence of the health event, in Indian rupees (INR). We consider a responseto be "heaped" if it is a multiple of INR 500 (roughly 10 USD). The results are presented below, in Table9. Consistent with recall error increasing with the recall interval, when the interval is longer than median (short recall=0), 90 of 317 observations, or 28%, exhibit heaping. When the interval is shorter than median, only 76 of 346, or 22%, of observations exhibit heaping. This difference is significant at the 10% level (p=.057).

Table 9. *Total out---of---pocket expenditures (OOP)*

		short recall		
round		0	1	**Total**
(500 Rs)	0	227	270	**497**
	1	90	76	**166**
	Total	**317**	**346**	**663**

Repeating the procedure with OOP expenditure on medicines yields a similar pattern, with even more evidence of heaping: 78% of long---interval observations display heaping, vs. only 56% of short---interval observations (Table 10). This difference is significant at the 5% level (p---value .013).

Table 10. Total medicine out---of---pocket expenditures (OOP)

		short recall		
		0	1	**Total**
round (500 Rs)	0	13	20	**33**
	1	47	25	**72**
	Total	**60**	**45**	**105**

Next, we examine the number of days spent in the hospital (including zeros if no hospitalization occurred). Heaping is now defined as the response being a multiple of 5 days.The same pattern is seen, but is not statistically significant (p=value .343): 28% of observations are heaped when the interval is short, vs. 23% when long (Table 11).

Table 11. Days in hospital

		short recall		
round		0	1	**Total**
(5 days)	0	108	125	**233**
	1	41	37	**78**
	Total	**149**	**162**	**311**

Finally, we examine the number of days of work orschool missed due to the health event. Here there is essentially no relationship between heaping and recall length: heaping is observed for 68.7% of short--- interval observations and 69.4% of long---interval observations (Table 12), and the small difference is not statistically significant (p---value 0.914).

Table 12. Days of work missed

		short recall		
round		0	1	**Total**
(5 days)	0	21	37	**58**
	1	46	84	**130**
	Total	**67**	**121**	**188**

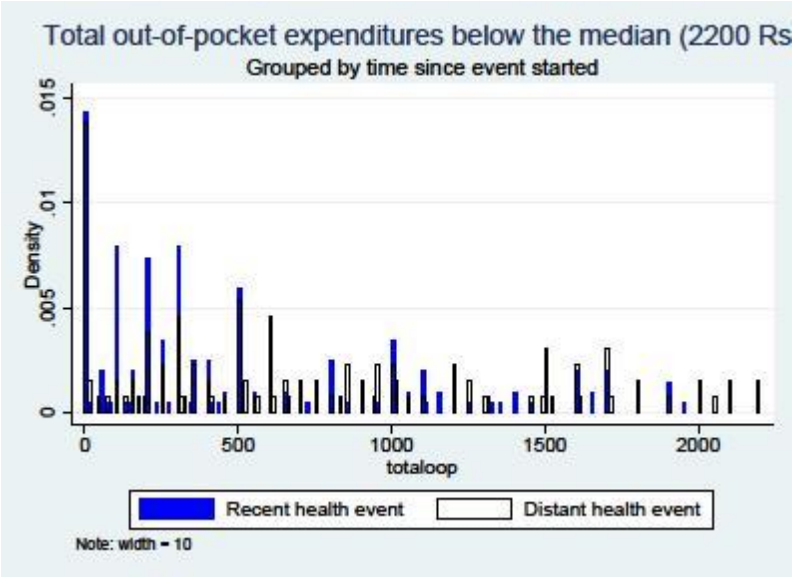
In sum, for both expenditure variables (OOP and medicine OOP), there is evidence of greater heaping when the elapsed time between event and survey is longer, and the association is statistically significant at conventional levels. For the days variables (days of work missed and days in hospital), the association is insignificant and of inconsistent sign, underscoring the need for more data. The difference between the expenditure vs. days variablesis a potentially intriguing finding that we will explore further when data from the scaled---up PHES becomes available.

Results: Dispersion

Finally, we will examine the amount of dispersion seen in the data. We present this analysis pictorially, in the form of histograms.

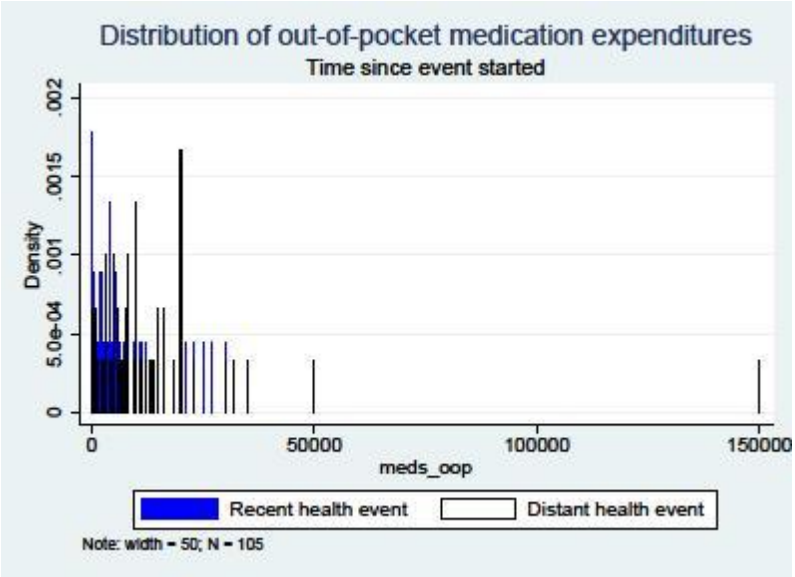
For total OOP, because the presence of a long right tail of high expenditures makes the data difficult to read, we focus on observations below the median (INR 2200). In this sample there is little clear---cut evidence of differential dispersion across the two groups(Figure 2):

Figure 2. Total out---of---pocket expenditures (OOP) below the median (2200 Rs.)



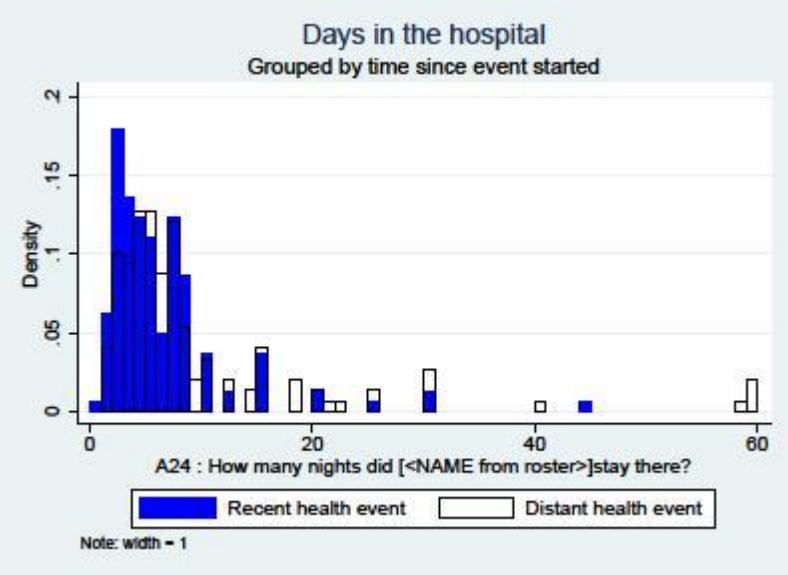
When we turn to OOP expenditures on medicine, however, there does appear to be greater dispersion among those with a longer recall interval, reflected in a longer right tail(Figure 3):

Figure 3. Total medicine out---of---pocket expenditures (medicine OOP)



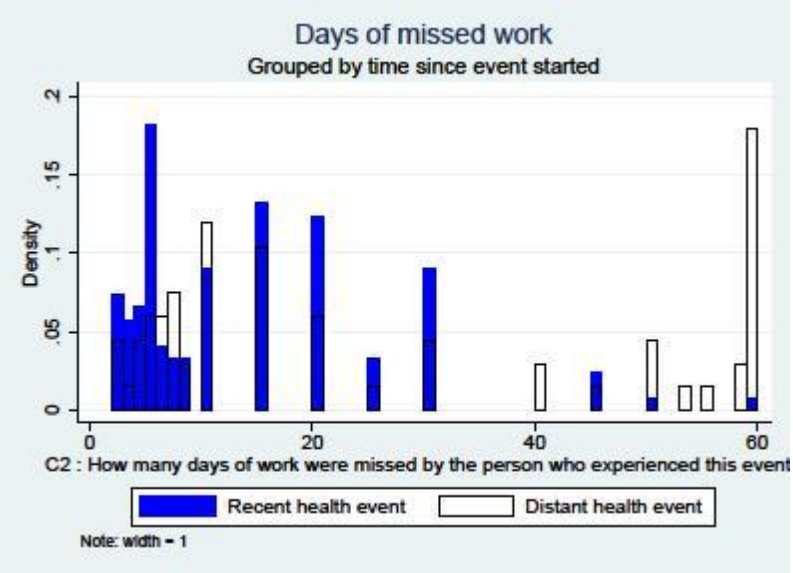
Turning to days spent in hospital, there is again evidence of more dispersion when the recall interval is longer (Figure 4):

Figure 4. Days in the hospital



Finally, we examine days of work or school missed.Again, there is suggestive evidence of more dispersion when the recall interval is long (Figure 5):

Figure 5. Days of work missed



In sum, there appears to be greater dispersion when the recall interval is long for three of the variables examined: medicine OOP, days in hospital, and days of work missed, supporting our hypothesis that responses collected when the recall interval is longer exhibit more error. However, there is little evidence of greater dispersion for longer recall in thetotal OOP variable for observations below the median. This variable simply has so much dispersion in general that our method of dispersion

analysis is unable to pick out differences according to the recall interval. Recall, however, that our heaping analysis (above) did find greater evidence of heaping for this variable when the recall interval is long, which is an indicator that recall error is increased by long recall times.

Section VI: Discussion and Conclusion

In this section we discuss what we have learned from the PHES pilot and the ways in which it will inform our scaled---up PHES. Several conclusions have emerged from our analysis, including that recall error does increase meaningfully with time elapsed, while the exercise of implementing the pilot PHES led to valuable insights in terms of ensuring that the survey is understood by respondent and quantifying the cost savings of the PHES.

Recall interval and data quality

We find evidence of statistically significant increases in measurement error, measured as increases in data heaping, at longer recall interval for key monetary variables such as total out of pocket expenditure (OOP) reported by the respondent as a consequence of the health event andOOP expenditure on medicines. Interestingly, we do not find evidence of greater heaping at longer recall intervals for measures of the time spent in hospital or days or work or school missed as a consequence of the health event. Thus it appears that monetary---denominated variables are harder to recall with precision after time elapses, while variables measured in days are easier to remember. It may be that individuals can mentally benchmark time---related measures but find it difficult to do so where quantities of money are involved.

In terms of future practice, this suggests that measures of expenditure are particularly sensitive to the recall interval (the total time elapsed between the start of the health event and being asked about it). Thus, high---frequency, short recall surveys should prioritize collecting this information, especially since measures of the financial consequences of a health event are a key input to understanding the consequences for household welfare. Questions about time---use consequences of health events, on the other hand, appear well suited to lower frequency surveys with longer recall intervals, such as annual surveys.

Other types of data, such as information on symptoms experienced or treatments sought, were notwell suited to our heaping and dispersion analysis. Thus, **we will examine these variables when we have data from both the PHES and the annual survey.**

Using the vernacular to define health insurance

Another very important conclusion from the PHES pilotwas that the survey instrument must refer to health insurance in a way that is colloquially intelligible. In the final waves of pilot data, a large subset of respondents answered affirmatively to having RSBY but negatively to having health insurance.This alerted us that we were not defining health insurance in an intuitive way.In consultation with a health ethnographer at the University of Pennsylvania, Prof. Vani Kulkarni, we revised the survey instrument to improve the language and to include pictures of health cards associated with different types of health insurance. This updated language and methodology has been incorporated into both the PHES scale up and the midline annual survey.

Cost---benefit of PHES

The PHES is significantly more cost---effective than a standard annual survey. We have confirmed, through a competitive bid process, that the data gathering market in India estimates the cost of a full---

scale PHES to be around $250,000 versus the cost of an annualmidline at $570,000. To put it another way, the full---scale PHES will cost ~45% of the annual survey.Even after equalizing the survey duration and the equipment costs between the two surveys, the full---scale PHES will cost ~75% of the annual survey exercise. The cost savings are attributed to the screening exercise that helpsreduce the number of households that need to receive an in---person survey and also make possible a leaner field management team structure to achieve, arguably, a higher quality of data.

Summing up

Measuring a complex, multi---dimensional phenomenon such as a household's response to serious health events, and the threat of such events happening in the future, is not easy: "the devil is in the details." **The IGC---funded PHES pilot has provided invaluable insights about the best ways to undertake this task. Those insights are now being put into practice informing the scaled up PHES and an annual survey that will soon be launched.**

Works Cited

Arnold, Benjamin F., et al. "Optimal recall period for caregiver---reported illness in risk factor and intervention studies: a multicountry study." *American Journal of Epidemiology* 177.4 (2013): 361---370.

Beegle, K., Carletto, C., & Himelein, K. (2012). Reliability of recall in agricultural data.*Journal of Development Economics, 98*(1), 34---41. doi:http://dx.doi.org/10.1016/j.jdeveco.2011.09.005

Broockman, D. E., Kalla, J., & Sekhon, J. S. (2016). Testing Theories ofAttitude Change with Online Panel Field Experiments. *Available at SSRN*.

Crawford, F. W., Weiss, R. E., & Suchard, M.A. (2015). Sex, Lies, and Self---Reported Counts: Bayesian Mixture Models for Heaping in Longitudinal Count Data Via Birth---Death Processes. *The Annals of Applied Statistics*, *(9)*2, 572---596. doi:http://doi.org/10.1214/15---AOAS809

Das, J., & Leino, J. (2011). Evaluating the RSBY: Lessons from an Experimental Information Campaign. *Economic & Political Weekly, 46*(32), 85.

de Nicola, F., & Giné, X.(2014). How accurate are recall data? Evidence from coastal India. *Journal of Development Economics, 106*, 52---65. doi:http://dx.doi.org/10.1016/j.jdeveco.2013.08.008

Hou, X., & Palacios, R. (2010). *Hospitalization patterns in RSBY: preliminary evidence from the MIS*. RSBY Working Paper #6. Retrieved fromhttp://www.rsby.gov.in/Documents.aspx?ID=14

Kaboski, J. P., & Townsend, R. M. (2011). A Structural Evaluation of a Large---Scale Quasi---Experimental Microfinance Initiative. *Econometrica, 79*(5), 1357---1406. doi:10.3982/ECTA7079

Ligon, Ethan. (2016) "Estimating household neediness from disaggregate expenditures.", *Department of Agriculture and Resource Economics,* University of California---Berkeley. doi: http://escholarship.org/uc/item/5gc4h1fm --- page---2

McKenzie, D. (2012). Beyond baseline and follow---up: The case for more T in experiments.*Journal of Development Economics, 99*(2), 210---221. doi:http://dx.doi.org/10.1016/j.jdeveco.2012.01.002

Palacios, R. (2010). *A new approach to providing health insurance to the poor in India: The early experience of Rashtriya Swasthya Bima Yojna*. RSBY Working Paper #1. Retrieved from http://www.rsby.gov.in/Documents.aspx?ID=14

Sun, C. (2010). *An analysis of RSBY enrolment patterns: Preliminary evidence and lessons from the early experience*. RSBy Working Paper #2. Retrieved from http://www.rsby.gov.in/Documents.aspx?ID=14

The International Growth Centre (IGC) aims to promote sustainable growth in developing countries by providing demand-led policy advice based on frontier research.

Find out more about
our work on our website
www.theigc.org

For media or communications
enquiries, please contact
mail@theigc.org

Subscribe to our newsletter
and topic updates
www.theigc.org/newsletter

Follow us on Twitter
@the_igc

Contact us
International Growth Centre,
London School of Economic
and Political Science,
Houghton Street,
London WC2A 2AE

IGC
International
Growth Centre

DIRECTED BY
LSE
UNIVERSITY OF OXFORD

FUNDED BY
UKaid
from the British people

Designed by soapbox.co.uk