# Financial & Health Diaries: User Guide

This version: December 18th, 2018

## Project Information

The Financial & Health Diaries project was a joint initiative of the Amsterdam Institute for International Development (AIID – currently AIGHD)[1] and the PharmAccess Foundation (PAF). In Nigeria it was carried out in collaboration with the University of Ilorin Teaching Hospital (UITH). In Kenya, Ipsos Synovate acted as a local research partner.

The principal investigators for the project were Wendy Janssens (AIGHD, Vrije Universiteit Amsterdam) and Berber Kramer (International Food Policy Research Institute – IFPRI). The project was co-initiated by Sicco van Gelder and Annegien Wilms (PAF). Project management was carried out by Marijn van der List (AIGHD) and data management was conducted by David Pap (AIGHD). Mike Murphy (IFPRI) assisted in finalizing the data and created this user guide.

The Financial and Health Diaries data collection was funded by PAF, AIID and the Health Insurance Fund (HIF). Additional funding for research time was provided by the Netherlands Organisation for Scientific Research (NWO), IFPRI, and CGIAR research programs Agriculture for Nutrition & Health (A4NH) and Policies, Institutions & Markets (PIM).

For general project inquiries, please contact Wendy Janssens (*w.janssens@vu.nl*) or Berber Kramer (*b.kramer@cgiar.org*). For data inquiries, or if you are unable to access the data, please contact Mike Murphy (*m.murphy@cgiar.org*).

The dataset should be referenced as:

Janssens, W., B. Kramer, M. van der List, and D. Pap (2018), "Financial and Health Diaries 2012-2013: A year-long weekly panel of farming households in Nigeria and Kenya", published master dataset, Amsterdam Institute for Global Health and Development (AIGHD).

The data are available upon request through AIGHD's data repository upon signing a Data Use Agreement with AIGHD. Interested researchers should send a request through email to *w.janssens@vu.nl* and *b.kramer@cgiar.*org, specifying their research objectives. The email should include:

[1] The Amsterdam Institute for International Development (AIID) merged with the Amsterdam Institute for Global Health and Development (AIGHD) in 2017. All original "AIID"-affiliations of the researchers have been updated to "AIGHD".

- Name:
- Email address:
- Institution:
- Position:
- Objective of the research (minimum 1 paragraph):

If active involvement of AIGHD/IFPRI researchers is desired for data analysis and/or interpretation, the Diaries team is open to discussing co-authorship. Otherwise, researchers are asked to appropriately reference the dataset and acknowledge the AIGHD, IFPRI and PharmAccess International for making available the data.

## Overview

The goal of this guide is to act as a resource for researchers working with the Financial and Health Diaries data, by providing background information on how the data was collected as well as functional examples of how to work with the datasets. The included information is intended to provide sufficient background to begin working with the data, along with some stylized examples that highlight key features of the data.

The guide is divided into four sections. The first provides information on project background, on the selection of participants for the study, and on the data collection instruments that were used. The second section provides a description of the datasets. It also describes how the raw data from participant interviews was processed, and how the variables in the weekly transactions dataset were constructed. The third section provides some examples of how to work with the data using Stata syntax. The fourth section includes an overview of journal articles and working papers using the diaries, both completed and work in progress.

# Contents

# Section 1- Project Background & Data Collection

## 1.1 Motivation

The Financial and Health Diaries project is a unique attempt to capture high-detail high-frequency data on how low-income households in developing countries manage money and make spending choices around healthcare. Since households in developing countries typically pay a large share of health expenditures out-of-pocket, health events can have substantial effects on the financial transfers and economic activity of household members who are coping with healthcare costs. Health insurance may shield households from the most devastating consequences of illness or injuries, but uptake and renewal of insurance in developing countries is typically low – even when premiums are highly subsidized.

At the start of the study, PAF was implementing several health insurances schemes–funded by the Health Insurance Fund (HIF)–in Sub-Saharan Africa. Similar to other insurance programs, these schemes were struggling to keeping enrolment rates sufficiently high to ensure their sustainability. The main motivation behind the Diaries project was therefore to understand why households in the HIF target population would (not) enroll in insurance or renew their policy, and how uninsured households were coping with health shocks.

The primary research objectives were three-fold:

- **Understanding the health characteristics of the population**. What are people's health needs? What determines health-seeking behavior? How much do people spend out-of-pocket on health and how do they finance these expenses?
- **Barriers to insurance uptake and renewal**. Do solvency and/or liquidity constraints prevent households from enrolling in health insurance? Are limited commitments to savings a constraint to enrolment?
- **Exposure to risk and risk management strategies**. To what extent are households exposed to different types of risks? How does this affect their ability to pay for health insurance? How do people cope with (uninsured) risk?

The project aimed to provide an important complement to standard survey data by gathering detailed weekly information on individuals' financial activities, health events and interactions with healthcare providers. Data on household income and expenditures in developing countries typically comes from large-scale annual household surveys that ask respondents to recall information for the preceding year. Lacking detailed financial records, households may have difficulties in providing complete and accurate information. The granular level of the diaries data should substantively reduce these biases.

Other advantages of using the diaries approach instead of annual household survey data are the ability to capture seasonal effects, to measure both major and minor health events as well as foregone care, the likely higher level of trust between respondent and enumerator, and the repeated nature of observations, allowing for the use of panel data models to control for unobserved heterogeneity across households and individuals. Potential disadvantages relate to the substantial requirements in terms of data management, the smaller sample size that limits the number of catastrophic events in the sample, and potential Hawthorne effects, i.e. changes in respondent behavior induced by weekly interviews.

The project took place in two rural low-income target populations of the HIF: in Nandi County, Kenya, and Kwara State, Nigeria. All adult household members in participating households were interviewed on a weekly basis for the duration of a full year and asked to report every financial transaction they had incurred in the previous week as well as every health-related event that had taken place in their household. Men and women were interviewed separately and in private to account for separate incomes and financial responsibilities. To explore how enrolment in health insurance may influence these outcomes, the samples for both countries were split between households with and without health insurance. The next section provides some context on the health insurance plans in each country.

## 1.2 Health Insurance

### Kenya (The Community Health Plan)

At the time of the project, Kenya's national health insurance scheme, the National Hospital Insurance Fund (NHIF), covered in-patient care in public hospitals but not health expenditures in private facilities or expenditures for out-patient procedures. The NHIF was mostly accessibly to formal sector employees. To improve the quality and affordability of healthcare among the farming population in Nandi County, PAF developed the Tanykina Community Health Plan (TCHP). This voluntary insurance scheme was introduced in 2011 in partnership with the Kenyan insurance company AAR and the Tanykina Dairy Plant Ltd., a farmer-owned dairy organization in the area, and funded by the Health Insurance Fund (HIF). The scheme also aimed to improve the quality of healthcare in facilities within its network by implementing quality standards, financing initial facility upgrades, and regularly monitoring quality improvements.

The Kenya sample consisted of dairy farmers located in Nandi County who supplied milk to Tanykina. At baseline, approximately half of the sample was enrolled in health insurance via TCHP. Enrolment was on a family basis. Their monthly insurance premiums were deducted from their milk deliveries to Tanykina – hence enrollment is partly dependent on providing regular milk deliveries to the company. The other half of the sample was uninsured at the time of the baseline survey.

The health insurance plan included both out-patient care and in-patient coverage, including for treatment for chronic diseases, with an actuarially fair premium which varied with household size. Mid-way through the study (in April 2013) TCHP introduced an additional cheaper package, covering only out-patient care. At this stage TCHP also fixed the price per household irrespective of size. The packages were priced at 300 KSh and 1100 Ksh for the 'basic' and 'comprehensive' packages, respectively. Throughout, the plan incorporated a waiting period between sign-up and the policy start date. Households registering from the 1st-25th of the month would be enrolled from the 1st of the following month, those registering after the 25th had to wait until the beginning of the month after that, hence the waiting period ranged from 5-35 days.

### Nigeria (Hygeia Community Healthcare)

Households in the Nigeria sample were residing in the north of Kwara Stata, which is one of the poorest states in Nigeria, with a large share of the population dependent on subsistence farming. In 2007 the Nigerian health maintenance organization Hygeia, together with PAF, introduced the Hygeia Community Health Care (HCHC) plan to farmer households in rural Kwara, which was the first of a number of pilots introduced by HIF to target low income populations in Africa South of the Sahara. In Kwara North, where the study took place, the program upgraded two public health facilities and one private clinic in three major towns. The program also provided subsidized insurance through the Kwara State Health Insurance Program (HIP).

As in Kenya, approximately half of the households included in the diaries study had health insurance coverage at the outset of the study. The insurance scheme was voluntary and based on individual enrolment, which means a household could decide who to enroll and who not to enroll. It was highly subsidized: the insurance premium ('co-premium') was 300 NGN per enrolled family member per year, approximately six percent of the actual premium. The Kwara State government and HIF jointly financed the premium subsidy. The benefit package consisted mainly of primary and secondary healthcare (including HIV/AIDS treatment). There was only one package available with a set price for everyone. Enrolment was on an annual basis and could start on the 1st of each new month. Enrollees needed to renew their contract again each year for 300 NGN per insured household member. The waiting period was the same as in Kenya, with people able to register for the following month from the 1st-25th of the current month.

## 1.3 Sampling methodology

### Sampling: Kenya

In Kenya, the planned sample size was 120 households, selected from households who were members of the Tanykina dairy cooperative in three areas: Salien, Surungai, and Lemook. The goal was to identify 60 households that were insured as part of the Tanykina Community Health Plan, and 60 that were not enrolled in the program.

Administrative data from Tanykina were used to randomly select 60 insured households residing in seven villages in total. In cases where households were no longer Tanykina members, a substitute was drawn from a replacement list. The uninsured sample was selected through a random selection procedure amongst other households in the sample villages. To be eligible for inclusion in the study, households had to be members of the dairy itself (Tanykina), and plan to stay in their village and retain their membership for the upcoming year.

The proportion of insured versus uninsured households sampled within each village was based on the proportion of insured households in a village relative to the total number of insured households in the seven sampled villages. Within sampled households, the Diaries data collection interviewed all financially active adult household members. Financially active adults were defined as all adults in a household who were currently working for income, physically and mentally able to respond, and were present in the household for at least four days a week on average.

In total, 222 eligible households were included in the household listing. Of these, 50 households started but did not complete the baseline survey; 42 households could not be reached for the baseline survey; and 10 households refused to participate in the subsequent Diaries. This left the final sample of 120 households, including 207 adult Diaries respondents and 587 individual household members

### Sampling: Nigeria

The Diaries study was implemented in the surroundings of three major towns in Kwara North: Shonga, Bacita, and Lafiagi. Each district was stratified in one urban cluster (town) and four (in Shonga and Bacita) or five (in Lafiagi) randomly selected rural clusters (villages) within a 15-kilometer radius from town. The goal was to include variation in distance to the nearest upgraded clinic in each of the three towns as part of the insurance scheme, since this would likely affect the benefits of having insurance. In total there were 13 clusters, three urban and ten rural clusters. The total planned sample size was 240 households: approximately 20 households in each cluster, of which half were enrolled in health insurance.

"Uninsured" households were originally defined as households which had never been insured. However, the listing exercise for the baseline survey showed that it was difficult to identify a sufficient number of never insured households in each of the clusters. Therefore, the sampling methodology was adjusted, and the working definition of "never insured" was changed to "uninsured in the past 2 months".

The 240 selected households were randomly assigned to a Diaries group and a Control group. Both took part in the baseline and endline survey, but only the Diaries households also participated in the year-long weekly diaries interviews. Random assignment occurred within insurance-cluster strata, such that the final sample consisted of 120 diaries households (60 insured and 60 uninsured) and 120 control households (60 insured and 60 uninsured). The comparison of the Diaries and the Control group at endline enables us to estimate the extent of any interview effects, for instance due to a Hawthorne effect or increased trust between respondents and enumerators. Due to budget constraints, including a control group without diaries was not possible in Kenya.

In total, 311 adult respondents from the Diaries households participated in the baseline. They represent a total of 829 household members in the Diaries group. Household-level attrition was very low: Only 3 households (2.5%) had dropped out of the Diaries sample by the time of the endline.

## 1.4 Survey Instruments

Two main types of survey instruments were used in the study: general household surveys, and individual financial diaries interviews. Following selection based on the eligibility criteria from the household listing exercise, households participated in a baseline household survey. This was then followed by weekly Financial and Health Diaries interviews over the course of a full year, after which an endline household survey was administered.[2]

In addition to these main data instruments, households completed quarterly stocktaking surveys to track asset ownership; adult respondents participated in behavioral lab-in-the-field experiments that measured risk, time and social preferences; and the enumeration teams collected monthly price data on commodities at local markets. Administrative data on insurance enrolment was collected after the endline survey. Before the start of the baseline surveys, the research team conducted qualitative in-depth interviews in both countries to inform instrument design.

### Household Survey (Baseline)

*Unit of Observation: Individuals*

For both the Kenya and Nigeria samples, detailed household surveys were carried out at baseline which gathered information on all members of the selected households. These included questions relating to the demographic, educational and employment characteristics of household members, and detailed questions on financial status and activities, including past and expected income-generation, asset ownership, planned expenditures, savings, access to credit and other formal or informal financial instruments, and decision-making within the household. The surveys also included modules on individual health status, health expenditures, and health insurance enrolment.

### Financial and Health Diaries (Transactions)

*Unit of Observation: Transactions*

Based on the information collected from the baseline household survey, all financially active adults were invited to participate in weekly private interviews. The diaries interviews followed a more flexible structure than the household survey, with respondents first encouraged to provide an overall description of their financial and health-related activities or events in the preceding week, and then prompted to provide more detail on specific transactions or health incidents.

---

[2] In the case of Kenya a full endline was not completed due to budget constraints. Instead, we conducted an abbreviated endline survey on exposure to and coping with shocks, and with a focus on contrasting these shocks with losses from civil conflict around the national election in 2007.

Specifically, respondents provided information about the period since the last interview on:

- Individual level (only for adult Diaries respondents): sales and purchases of goods and services; income from work; gifts, loans or credit given or received; intra-household gifts; and savings deposits or withdrawals.

- Individual level (for all household members: health problems; consultations at formal and informal health providers; and education/working days lost due to health problems or other reasons.

- Household level: sales or consumption of agricultural produce; loss, damage or theft of assets; production, sales and consumption of milk (Kenya only).

Additionally, respondents were asked about any such transactions or events which had occurred in a previous reporting period but that they had forgotten to include in a previous interview. Similarly, if respondents had missed a weekly interview due to absence, they could report transactions and events for the missed period.

### Quarterly Balances (stocktaking surveys)

*Unit of Observation: Households*

In addition to the weekly diaries, quarterly stocktaking surveys were carried out, starting at the baseline survey. The stock-taking surveys consisted of a detailed inventory of a household's physical assets (including livestock and land ownership) as well as Diaries respondents' formal and informal financial assets. These so-called Quarterly Balances help to track changes in asset ownership from the baseline survey and provide an additional data source to compare the changes in asset ownership to the transactions reported in the weekly diaries.

### Household Survey (Endline)

*Unit of Observation: Individuals*

At the conclusion of the project in Nigeria, an endline household survey was conducted which was similar in content to the baseline. In Kenya, an endline survey was carried out that was an abbreviated version from the baseline, except for a new module that focused principally on shocks that had affected the household in the preceding year.

### Monthly price questionnaires

*Unit of Observation: Item*

In addition to the interviews, the research team also gathered monthly price data for a range of locally available goods during the study period. These data were gathered at monthly intervals as the average price for specified goods asked from three sellers at a main local market. They are included as separate datasets for each country.

### Health Insurance

*Unit of Observation: Households (Kenya); Individuals (Nigeria)*

Information on the health insurance status of participants (at the household level in Kenya, and individual level in Nigeria) comes from administrative data from the health insurance providers. This data was matched by the research team to the financial diaries dataset, and included in the weekly diaries datasets (see description below). Note that enrollment status varies by month, while transaction information in the diaries was collected by week.

### Experimental data on risk, time, social preferences and subjective expectations

*Unit of Observation: Individuals*

The research team also undertook a series of experimental games at different periods of the study, which included elicitation of subjective expectations as well as time and risk preferences.

For the Kenya sample, these were randomly asked at the end of the diaries interviews. In each case the questions were the same, with binary choice questions for risk and time preferences, a paired altruism game and the elicitation of expectations of potential financial events for the coming month, all of which were incentivized. For Nigeria, the experiments were carried out following the stocktaking, with respondents participating in incentivized time preference and expectations games.

Datasets from these activities are not included in the current release, for further information please contact Mike Murphy (*m.murphy@cgiar.org*).

## 1.5 Data Collection Timeline

| Month | Activities: Nigeria | Activities: Kenya |
|---|---|---|
| **November 2011** | Begin In-Depth Interviews | |
| **December 2011** | End In-Depth Interviews | |
| **March 2012** | Baseline Survey | |
| **April 2012** | Begin Diaries Surveys (Weekly)<br>Begin Market Surveys (Monthly) | Begin In-Depth Interviews<br>End In-Depth Interviews |
| **June 2012** | Stocktaking 1 | |
| **August 2012** | | Baseline Survey |

| | | |
|---|---|---|
| **September 2012** | Stocktaking 2 | |
| **October 2012** | | Begin Diaries Surveys (Weekly) |
| **November 2012** | | Begin Market Surveys (Monthly) |
| **December 2012** | Stocktaking 3<br>No Market Survey | Stocktaking 1 |
| **March 2013** | Stocktaking 4<br>No Market Survey | Stocktaking 2 |
| **April 2013** | End Market Surveys | |
| **May 2013** | End Diaries Surveys | |
| **June 2013** | Endline Survey | Stocktaking 3 |
| **October 2013** | | Stocktaking 3<br>Endline Survey<br>End Diaries Surveys<br>End Market Surveys |

# Section 2- Data Processing

Following the completion of the interview, the data for each of the survey and diaries datasets was processed in three main stages:

1. Data entry & consolidation
2. Preparation & cleaning of data files
3. Variable construction & calculation of transaction flows

Researchers are invited to contact the team directly for more details on any stage of data processing.

## 2.1 Data Entry & Consolidation

### Paper Questionnaires vs. Computer-Assisted Personal Interviews

The baseline and endline household surveys as well as the monthly price questionnaires were collected through paper-based interviews.

The Diaries data and quarterly Stock-takings for Nigeria were at start collected using paper-based questionnaires as well, which were then manually entered into computers. Beginning in July, the team gradually switched to a Computer Assisted Personal Interview (CAPI) system. By early August, the team fully transitioned to CAPI. For Kenya the CAPI system was used throughout Diaries data collection.

The CAPI tool was developed using Microsoft (MS) Access by the study team. While some validation conditions were used, the CAPI tools were highly flexible in their design, particularly for the Diaries surveys. Enumerators were able to navigate the form freely, switching between models to allow them adding different types of transactions and events as they came up in conversation with the respondent. Checked menu bars ensured that modules were not inadvertently skipped, while a box showing the sum of reported weekly incomes and expenditures enabled the enumerators to ascertain that no large financial transactions went unreported.

### Processing Raw Data

Following data collection, the raw data generated by the MS Access back-end was converted into Stata .dta format. Corrections were applied to ensure that ID variables were correct, duplicate entries were removed and other issues were resolved (notably one laptop was damaged during fieldwork – no data was lost, however additional data recovery steps had to be added into the workflow to incorporate this data). Files from multiple interviewers' computers were consolidated into Stata into a single dataset for each data collection activity (i.e. one dataset for the Kenya household baseline, one for the Nigeria diaries transactions, etc.).

## 2.2 Preparation & Cleaning of Data Files

After processing the raw datasets, there was a secondary stage in which the data was prepared for cleaning. This involved validation of ID variables across datasets, standardization of variable names, correction of data entry errors, consolidation of text responses into categories, removal of personally identifying information and survey metadata (such as GPS points and interviewer comments). This stage of processing was inherently specific to the dataset or variable under consideration. It should be noted that generalized procedures for outlier detection or trimming were not applied across the data.

Following cleaning of the diaries transactions and events data, an additional step was undertaken to integrate 'forgotten' transactions which had not been recorded by a respondent in a given interview week but recalled in a later week – either because it was inadvertently omitted in a previous interview or because the respondent had been absent.

## 2.3 Variable Construction, Account Balances & Flows

Once variables had been cleaned and all forgotten transactions had been integrated, variable construction for the Diaries dataset proceeded in two stages. First, weekly financial flows were calculated for each individual respondent by aggregating transactions by type within a given week, such as loans and borrowing; gifts/remittances given or received; savings deposits or withdrawals; income; expenditures. In addition to these flows by transaction category, we also calculated total aggregate in- and out-flows per respondent per week.

Important is that these flows are calculated for two potential definitions of a week: the week in which the transaction took place (denoted with 'Tr') and the week in which the interview took place (denoted with 'Int'). See next section for more details on the week identifiers.

Next, these flow variables were combined with the data from the quarterly Stock-takings to calculate weekly balances, i.e. the result of these flows on the overall weekly financial situation of the household. These weekly balances were calculated using three constructions: 1) based solely on the transactions information in the diaries (i.e. without reference to the Stock-taking datasets) which are denoted with the prescript 'DO'; 2) in reference to assets owned at the most recent stocktaking *before* the week in question ('Forward', denoted 'FW'); and 3) in reference to the most recent stocktaking *after* the week in question ('Backward', denoted 'BW').

Additional variable construction was carried out to create a cash-flow variable representing the change in an individual's balance associated with a given transaction, and to create variables to capture the timing of the transaction relative to the timing of the interview.

## 2.4 Datasets

The published datasets are organized into five categories: household surveys (baseline and endline); transaction-level diaries data; weekly diaries data; stocktaking data; and price data. The table below shows the available datasets for each category, along with the frequency with which the data was collected, the level of observation, and the variable(s) within each dataset which uniquely identify the observations. The remainder of this section then provides some additional descriptive information on each category, and some more detailed information on the relevant identifying variables.

| Dataset | Files | Frequency | Observation Level | ID Variable |
|---|---|---|---|---|
| **Household Surveys** | baseline_kenya.dta; baseline_nigeria.dta; endline_nigeria.dta | One-time | Individual | *MemberKey* |
| **Financial Diaries (Transaction Files)** | transactions_kenya.dta<br>transactions_nigeria.dta | Weekly | Transaction | *TrKey MemberKey* |
| **Financial Diaries (Weekly Files)** | weekly_kenya.dta<br>weekly_nigeria.dta | Weekly | Week-Individual | *WeekKey MemberKey* |
| **Stocktaking** | stocktaking_kenya.dta<br>stocktaking_nigeria.dta | Quarterly | Round-Individual | *Round MemberKey* |
| **Prices** | prices_kenya.dta<br>prices_nigeria.dta | Monthly | Month-Item | *Month PriceKey* |

### *Household Survey Datasets*

#### Files: baseline_kenya.dta; baseline_nigeria.dta; endline_nigeria.dta

The household survey datasets comprise the baseline household surveys for both countries, and the endline household survey which was carried out for Nigeria. These datasets are at the individual level (ie. one row corresponds to a single individual), with multiple individuals reporting for each household. Households are identified by the *HouseholdKey* variable, to which a two-digit running number is appended to generate the *MemberKey* variable, which uniquely identifies observations (individuals) within the dataset.

The household survey datasets contain information on all household members. This is in contrast to the diaries datasets, which only contain reports by economically active household members. Therefore, the individuals reporting in the diaries datasets are a subset of the individuals in the household survey datasets. Additionally, in the case of Nigeria the household surveys include reports from control households which did not participate in reporting for the diaries datasets.

### *Financial Diaries Datasets*

#### Files: transactions_kenya.dta; transactions_nigeria.dta; weekly_kenya.dta; weekly_nigeria.dta

For each country, the financial diaries data is available in two forms. The first is a transactions-level dataset which includes information on each financial transaction or health event reported by any individual in a given diaries interview. Individuals typically report multiple transactions of different types in a given interview week, hence the transactions dataset will typically contain multiple rows for a given individual (identified by *MemberKey*) in a given week, each of which is identified by a unique transaction identifier: *TrKey*.

The second type of diaries dataset is the weekly datasets. These are an aggregation of all of the financial transactions carried out by a given individual in a given week across a range of categories. For example, consider a week in which Person A makes two purchases and receives one remittance. In the transactions-level dataset, this will be represented as three rows- one per transaction. In the weekly dataset, this will appear as one row, with separate variables indicating the total purchase amount, net remittances received and overall change in cashflow. Observations in the weekly datasets are therefore identified by individual and week, using the variables *MemberKey* and *WeekKey*.

Data on the health insurance enrollment status of participants (at the household level for Kenya, and the individual level for Nigeria) is included in the weekly datasets[3]. For Kenya, insurance status from the administrative dataset is indicated by the variable *InsStat_Admin*. There is also a variable capturing self-reported status: *InsStata_SelfR*. For Nigeria, insurance status is indicated by the variable *Insured*.

### *Stocktaking Datasets*

#### Files: stocktaking_kenya.dta; stocktaking_nigeria.dta

The stocktaking datasets contain observations for each of the stocktaking interviews that were carried out for each household. In Kenya, there were five stocktaking interviews in total, while in Nigeria there were six.  Each dataset contains a variable *Round* which indicates the order of the stocktaking interviews. Since the stocktaking was carried out at the individual level, observations are identified by the individual's ID and the round

[3] As above, note that insurance status varies by month, while transactions were collected by week.

number: *MemberKey* and *Round.* The dataset also contains the *WeekKey* variable, allowing users to determine when the stocktaking visits took place relative to the diaries visits- note however that there may be some variation in the week in which a given round of stocktaking took place across households, so the value of WeekKey for a given round may differ between households.

### *Price Datasets*

#### Files: prices_kenya.dta; prices_nigeria.dta

The price datasets contain information on market prices for a range of items. These were collected for both countries on a monthly basis at local markets. The available items are identified by the variable *PriceKey* which indicates the item type, and *Month* which indicates the calendar month for which the price was recorded.

### *Identifying Variables*

#### HouseholdKey

Identifies households within the dataset. An additional two digits are concatenated to *HouseholdKey* to generate *MemberKey*

#### MemberKey

Identifies individuals within the datasets. Each row in the household datasets has a unique value of *MemberKey*. This can also be matched to other variables within a household to define relations between individuals- for example *FatherKey* takes the value of *MemberKey* of the row within the household corresponding to that person's father.

Important to mention is that *MemberKey* in the baseline and endline household surveys capture all individual household members based on the household roster. In the Diaries, only the financially active adults participated as a respondent. Thus, only a subset of all *MemberKey* values appears in the diaries transactions dataset as a respondent.

#### SubjectKey

Identifies the individual within the Diaries dataset who was the subject of an expenditure (in contrast to the respondent to this question, which is captured in the *MemberKey*). For example, if a mother within a household reports a health expenditure she made for her daughter, the transaction will be recorded under the *MemberKey* value corresponding to the mother in the household survey, while *SubjectKey* will take the value corresponding to the *MemberKey* of the daughter in the household survey.

#### TrKey

Identifies a single transaction. Each row in the transactions dataset has a unique *TrKey*. For a very small number of cases where a transaction was originally recorded as taking two payment forms (ie. cash & mPesa) the transaction is split into two and denoted with a decimal, ie. if original transaction *TrKey* was recorded as “10001000” and split, these would take the values *TrKey* “10001000.1” and “10001000.2”.

*TrKey* is distinct from *TrID* which identifies individuals, weeks and transaction types, but does not distinguish between same-type transactions carried out by a given individual in a given week. In practice, *TrID* is unlikely to be required.

### WeekKey

Identifies the interview week. Values are sequential from the first week that a financial diary interview took place for that country and take values of 1-55. Note that this is the week in which the interview took place. The 7-day period for which the individual is reporting may include the preceding week.

# Section 3- Examples

## 3.1 Example 1- Working With the Financial Diaries: Gifts & remittances in Kenya

### Overview

The goal of this first example is to provide a brief introduction to working with the diaries datasets, using the data for Kenya as an example. It will demonstrate how to link household member information from the baseline dataset to the transactions data, and highlight some important considerations in working with the datasets.

### Loading the baseline dataset

To begin, we will load the baseline dataset, which contains information on the 120 households participating in the study. Note that each row in the dataset corresponds to an individual within a household. There are 587 individuals in the dataset. For this example we are going to be using only data from the first two modules of the baseline survey, so for simplicity we will exclude data from Section 3 onward from our dataset.

```
. use "baseline_kenya.dta", clear

. unique HouseholdKey
Number of unique values of HouseholdKey is  120
Number of records is  587

. isid MemberKey

. keep HouseholdKey-Edu_Language        // Drop Section 3 onward
```

### Defining the education variable

For this example, we are going to explore the relationship between respondents' education level and the amount they send and receive in gifts and remittances. To do this, we will construct an indicator variable for whether the individual has completed primary education, which for this example we will define as Primary Class 8 or above. Note that in constructing our indicator variable, we need to take care to ensure missing values in the Edu_Level variable are not encoded as zeroes.

```
. gen CompletedPrimary = .
(587 missing values generated)

. replace CompletedPrimary = 0 if Edu_Ever == 2 | Edu_Level < 8
(332 real changes made)

. replace CompletedPrimary = 1 if Edu_Level >= 8 & Edu_Level != 25 & !missing(Edu
_Leve
> l)
(247 real changes made)
```

```
. lab var CompletedPrimary "Member has compelted Primary Education"

. lab def ind_yesno 0"No" 1"Yes"

. lab val CompletedPrimary ind_yesno
```

### Merging the baseline to the transactions data

We are now ready to merge our dataset to the transactions data. The baseline is uniquely identified by MemberKey, but the transactions dataset can contain many transactions for a given member. Therefore we merge 1-to-many, with our new dataset being identified by the transaction identifier TrKey. When we merge all of the transactions data is matched, but 386 individuals from the baseline dataset are not matched. What happened? Recall that only financially active adult household members were selected to complete the diaries surveys. Of the 386, 301 are children while the remaining 85 are not financially active- hence we can subset out these individuals before proceeding with our analysis.

```
. qui merge 1:m MemberKey using "transactions_kenya.dta"

. tab _m
```

| _merge | Freq. | Percent | Cum. |
|---|---|---|---|
| master only (1) | 386 | 0.22 | 0.22 |
| matched (3) | 172,008 | 99.78 | 100.00 |
| Total | 172,394 | 100.00 | |

```
. count if _m == 1 & Age < 18
  301

. drop if _m == 1
(386 observations deleted)

. drop _m
```

### Exploring the transactions data (1)

Now that we have merged in the transactions data, we can take a look at some of its key features. As we have noted, transactions are identified with the ID variable TrKey. Note there is also a variable TrID which identifies the member, week and type of transaction but is NOT a unique identifier in the dataset- for example if Member A sold farm produce twice to Buyer B in Week 1, the transactions would have the same TrID but a different TrKey. In practice TrID is rarely needed. Two other important Tr* variables are TrType and TrPartner. TrType categorizes the transactions recorded, while TrPartner lists the other party to the transactions

```
. isid TrKey
```

```
. unique TrID
Number of unique values of TrID is  149333
Number of records is  172008

. tab TrType

                 Transaction Type |      Freq.     Percent        Cum.
----------------------------------+-----------------------------------
                Household stories |      2,964        1.72        1.72
                           Health |      1,842        1.07        2.79
     Unable to do daily activities |        577        0.34        3.13
        Assets damaged/stolen/lost |      1,485        0.86        3.99
      Purchases of goods and services |     93,953       54.62       58.61
Gifts, loans, credits, advance, harambe |      6,558        3.81       62.43
                          Savings |      6,283        3.65       66.08
  Harvest/consumption farm products |     34,961       20.33       86.40
   Consultations or health care visit |        769        0.45       86.85
     Income/sales of goods & services |     22,108       12.85       99.70
           Money not given/received |        508        0.30      100.00
----------------------------------+-----------------------------------
                            Total |    172,008      100.00

. inspect TrPartner

TrPartner:  Partner/Provider/Place of Tr        Number of Observations
------------------------------------------------------------------------------
                                         Total      Integers   Nonintegers
|  #                        Negative         -             -             -
|  #                        Zero             4             4             -
|  #                        Positive    32,140        32,140             -
|  #   #                               -------       -------     ---------
|  #   #                    Total       32,144        32,144             -
|  #   #   .   .   .        Missing    139,864
+----------------------                -------
0                     99               172,008
  (36 unique values)

      TrPartner is labeled and all values are documented in the label.
```

### Exploring the transactions data (2)

Other key variables within the transactions data are CF and the Amount variables. CF, for 'Cash-flow', records the amount of the transaction, and takes a positive value if the amount was received by the household and a negative value if the amount was paid out by the household. The amount variables record the value of the transaction, each corresponding to a different form of transfer. Hence AmountCash records the amount paid in cash, AmountMpesa the amount paid with mobile money, etc. Note that CF only takes a value if the transaction occurred in cash, so if we take the absolute value of CF, it will be the same as the AmountCash variable.

```
. gen abs_CF = abs(CF)
(54,902 missing values generated)

. sum CF AmountCash abs_CF

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
          CF |    117,106    12.84447    1556.926     -97000     120000
  AmountCash |    117,106    346.9192    1517.837          0     120000
      abs_CF |    117,106    346.9192    1517.837          0     120000

. sum AmountMpesa CF if !missing(AmountMpesa)

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
 AmountMpesa |        544    455.8162    1804.606          4      26000
          CF |          0
```

### Creating remittance variables

For this example, we are going to look at gift & remittance transactions. These are listed under "Gifts, loans, credits, advance, harambe", which is transaction code 6, and identified by codes '4' and '10' in the loan type variable which further categorizes these transactions. We will subset the dataset to look at only these transactions- note how the sign of the cashflow variable varies with the direction of the transaction. For simplicity, we will create two variables for in- & out-flows of cash gifts & remittances

```
. tab TrType if TrType == 6

                       Transaction Type |      Freq.     Percent        Cum.
----------------------------------------+-----------------------------------
Gifts, loans, credits, advance, harambe |      6,558      100.00      100.00
----------------------------------------+-----------------------------------
                                  Total |      6,558      100.00

. keep if inlist(LoanType,4,10)
(168,263 observations deleted)

. bys LoanType: sum CF

--------------------------------------------------------------------------------
-----
-> LoanType = GIFT/REMITTANCE RECEIVED

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
          CF |      1,823    923.5513    1736.853         40   33333.33

--------------------------------------------------------------------------------
-----
```

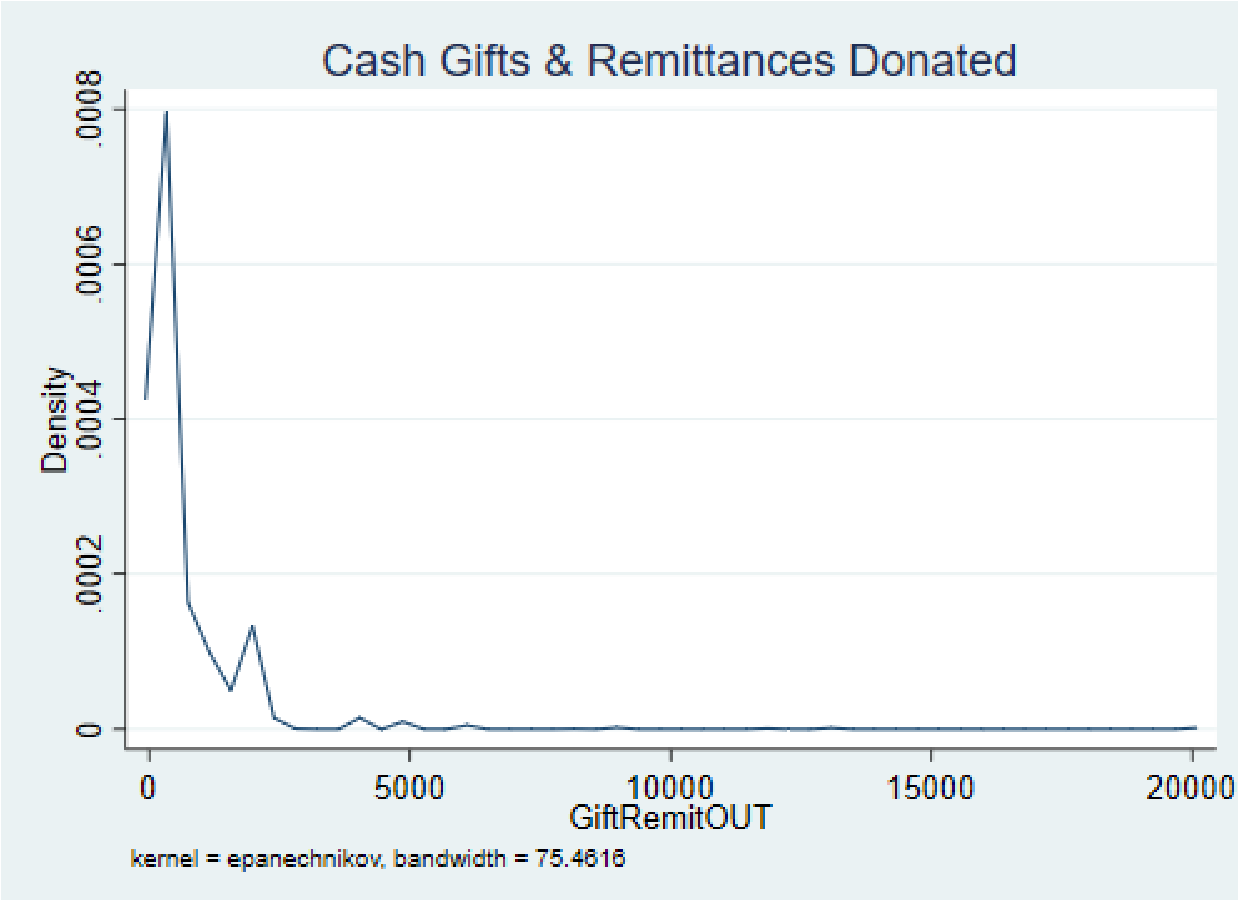
```
-> LoanType = GIFT/REMITTANCE GIVEN

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
          CF |      1,688   -590.1173    1160.943     -20000         -5
```

```
. gen GiftRemitIN = CF if LoanType == 4
(1,922 missing values generated)

. gen GiftRemitOUT = abs(CF) if LoanType == 10
(2,057 missing values generated)
```

### Exploring differences in remittances by education (1)

Now that we have created these remittance variables, we can compare them using the education status variable. One important consideration is that there are some very large transactions within the data- as can be observed from the long right tail in the distribution of GiftRemitOUT- we will ignore this for the purposes of this example, however handling outliers is an important consideration for researchers working with the data. With this caveat in mind, we can carry out a Student's t-test to compare remittance transactions received by education level.

```
. kdensity GiftRemitOUT, title("Cash Gifts & Remittances Donated")

. graph export test.png, width(500) replace
(note: file test.png not found)
(file test.png written in PNG format)
```




```
. ttest GiftRemitOUT, by(CompletedPrimary)

Two-sample t test with equal variances
```

| Group | Obs | Mean | Std. Err. | Std. Dev. | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| No | 419 | 826.0382 | 99.56669 | 2038.079 | 630.3244 | 1021.752 |
| Yes | 1,269 | 512.2206 | 17.7405 | 631.9703 | 477.4167 | 547.0246 |
| combined | 1,688 | 590.1173 | 28.2569 | 1160.943 | 534.695 | 645.5396 |
| diff | | 313.8175 | 64.98377 | | 186.3602 | 441.2749 |

```
    diff = mean(No) - mean(Yes)                                   t =   4.8292
Ho: diff = 0                                     degrees of freedom =     1686

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

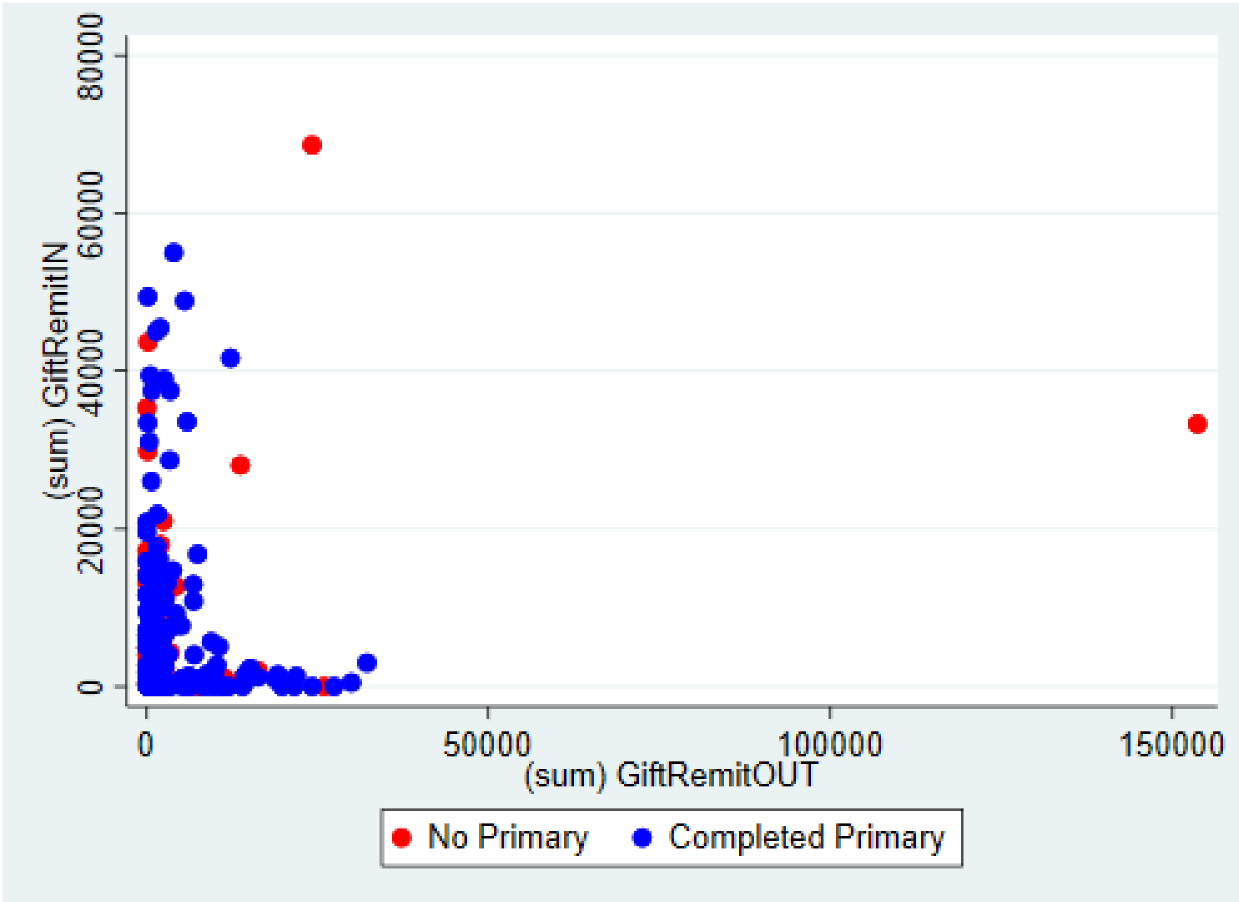
### Exploring differences in remittances by education (2)

Aside from our concern about outliers, we need to be careful in interpreting this comparison. Recall that the dataset is identified by transactions, so the difference we see is comparing the average transaction size, rather than the total quantity of transactions across the year. To look at total amounts, we need to collapse the dataset. Colapsing also allows us to compare multiple transcations for a given unit of observation, in this case hosuehold members- in this case we can plot the total amount given and received as gifts and remittances, by members' education.

```
. collapse(sum) GiftRemitIN GiftRemitOUT, by(MemberKey CompletedPrimary)

. ttest GiftRemitOUT, by(CompletedPrimary)

Two-sample t test with equal variances
```

| Group | Obs | Mean | Std. Err. | Std. Dev. | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| No | 60 | 5768.5 | 2607.432 | 20197.08 | 551.0412 | 10985.96 |
| Yes | 131 | 4961.893 | 585.6342 | 6702.889 | 3803.286 | 6120.5 |
| combined | 191 | 5215.277 | 908.2292 | 12551.98 | 3423.77 | 7006.785 |
| diff | | 806.6069 | 1960.962 | | -3061.577 | 4674.791 |

```
    diff = mean(No) - mean(Yes)                                   t =   0.4113
Ho: diff = 0                                     degrees of freedom =      189

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.6594         Pr(|T| > |t|) = 0.6813          Pr(T > t) = 0.3406

. twoway (scatter GiftRemitIN GiftRemitOUT if CompletedPrimary == 0, mcolor("red"
)) //
> /
>        (scatter GiftRemitIN GiftRemitOUT if CompletedPrimary == 1, mcolor("blue
")),
> ///
>        legend(order(1 "No Primary" 2 "Completed Primary"))

. graph export test2.png, width(500) replace
(note: file test2.png not found)
(file test2.png written in PNG format)
```




## Conclusion

In this example, we have focused on some key features of working with the household survey and diaries transaction datasets. We have reviewed how to merge the household data to the transactions dataset and noted some important issues relating to identifying variables and how to interpret data with these concerns in mind. In the next example, we will look at the weekly dataset which uses the transactions & stocktaking data to construct a variety of variables measuring weekly household financial flows.

## 3.2 Example 2- Working with the Health Diaries: Malaria in Nigeria

### Overview

In the previous example, we looked at how to use the diaries to look at financial data from Kenya. In this example, we will focus on working with the health components, using the Nigeria data. In particular, we will look at reported cases of malaria and use these to explore some key features of the data. Note that for ease of explanation, we will make some simplifying assumptions about the data. These are noted in the text below, but users are recommended to explore these issues in more detail in their own analysis.

### Loading the transactions data & formatting the date variable

To start, we will load the transactions dataset for Nigeria and take a look at two important time variables: Week_Int, which lists the week in which the interview took place, and TrWhen which records the date of the transaction. As we see, Week_Int takes values from 1-55 indicating the running count of the survey week. TrWhen is a little more complicated- it is a string which lists all of the dates associated with a transaction separated by slashes. For simplicity, we will take the last date in the list (the earliest) and convert this to a date variable, which we will call EventDate

```
. clear all

. use "transactions_nigeria.dta"

. sum Week_Int
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| Week_Int | 218,914 | 27.16861 | 15.77727 | 1 | 55 |

```
. tab TrWhen if _n < 30 & TrWhen != "."
```

| Date of Transaction or Event | Freq. | Percent | Cum. |
|---|---|---|---|
| 20120423 | 1 | 3.45 | 3.45 |
| 20120515 | 1 | 3.45 | 6.90 |
| 20120627 | 1 | 3.45 | 10.34 |
| 20120720 | 1 | 3.45 | 13.79 |
| 20120728 | 1 | 3.45 | 17.24 |
| 20120731 | 1 | 3.45 | 20.69 |
| 20120806 | 1 | 3.45 | 24.14 |
| 20120814 | 1 | 3.45 | 27.59 |
| 20120821 | 1 | 3.45 | 31.03 |
| 20120824 | 1 | 3.45 | 34.48 |
| 20120902 | 1 | 3.45 | 37.93 |
| 20120905 | 1 | 3.45 | 41.38 |
| 20120907 | 1 | 3.45 | 44.83 |
| 20120913 | 1 | 3.45 | 48.28 |
| 20121008 | 1 | 3.45 | 51.72 |

```
     20121017\20121016\20121015 |          1        3.45       55.17
     20121028\20121027\20121026 |          1        3.45       58.62
                       20121029 |          1        3.45       62.07
                       20121126 |          1        3.45       65.52
                       20121217 |          1        3.45       68.97
                       20121226 |          1        3.45       72.41
                       20130110 |          1        3.45       75.86
                       20130130 |          1        3.45       79.31
                       20130210 |          1        3.45       82.76
                       20130214 |          1        3.45       86.21
                       20130331 |          1        3.45       89.66
                       20130402 |          1        3.45       93.10
                       20130413 |          1        3.45       96.55
                       20130501 |          1        3.45      100.00
--------------------------------+-----------------------------------
                          Total |         29      100.00

. tempvar date_temp

. gen date_temp = substr(TrWhen,-8,.)
(33,541 missing values generated)

. replace date_temp = "" if inlist(date_temp,"98","99")
(0 real changes made)

. gen EventDate = date(date_temp,"YMD")
(33,541 missing values generated)

. format EventDate %td
```

### Subsetting transactions data into Health Events & Consultations

Now as we saw in the previous example, the variable TrType identifies the type of the transaction recorded in the diaries data. In the case of health, there are two categories we are interested in- health events (TrType == 2) & health consultations (TrType == 12). Health events refer to incidence of sickness or other health-related issues which occurred in the household that week, while health consultations refer to health services. To work with this data, we will subset it by these categories into two temporary datasets. An important variable here is the SubjectKey variable which identifies the household member who is the *subject* of the event, ie. the person who had the illness or received the consultation. Here for simplicity, we are going to restrict the datasets to only include the first listed transaction for a given subject in a given week for each category. First we create a dataset of health events, where the recorded symptom is “Fever/Malaria”, next we create a dataset of consultations where the reported reason is “Illness/Injury”

Health Events (TrType == 2)

```
. preserve

.     keep if TrType == 2 & HealthSympt == 1
(217,847 observations deleted)

.     keep SubjectKey Week_Int EventDate HealthConsult

.     sort SubjectKey Week_Int EventDate

.     by SubjectKey Week_Int: keep if _n == 1
(206 observations deleted)

.     tempfile malaria_events

.     save "`malaria_events'"
file C:\Users\MMURPHY\AppData\Local\Temp\ST_1b14_000002.tmp saved

. restore
```

Medical Consultations (TrType == 12)

```
. keep if TrType == 12 & ConsultWhy == 1
(216,897 observations deleted)

. keep SubjectKey Week_Int EventDate

. ren EventDate ConsultDate

. sort SubjectKey Week_Int ConsultDate

. by SubjectKey Week_Int: keep if _n == 1
(163 observations deleted)

. tempfile health_consultations

. save "`health_consultations'"
file C:\Users\MMURPHY\AppData\Local\Temp\ST_1b14_000003.tmp saved
```

## Merging the survey dataset to the health events dataset

Suppose we were interested in seeing the incidence of malaria in our sample. To do this, we need some individual data from our household surveys. We will the endline data for this example to ensure we include members who were born after the baseline. To get the characteristics of those affected we need to merge on the SubjectKey variable, so that we don't end up matching to the person who reported the transaction. Note that we also need to subset the dataset to exclude control households who did not take part in completing the diaries.

```
. use "endline_nigeria.dta"

. keep if typequest == 1
(864 observations deleted)

. clonevar SubjectKey = MemberKey

. drop MemberKey

. keep SubjectKey Age Gender

. tempfile endline_demographics

. save "`endline_demographics'"
file C:\Users\MMURPHY\AppData\Local\Temp\ST_1b14_000004.tmp saved

. merge 1:m SubjectKey using "`malaria_events'", keep(match master)

    Result                           # of obs.
    -----------------------------------------
    not matched                           504
        from master                       504  (_merge==1)
        from using                          0  (_merge==2)

    matched                               860  (_merge==3)
    -----------------------------------------
```

### Creating & summarizing an indicator for a malaria health event

Now we can use the merge status to determine whether or not someone was reported as having malaria/fever in the last year. Let's look at the incidence, which is high: around 63%. But wait- if we look at the number of observations we have 1364, but we only have 919 individual IDs. What's going on? Note that we merged 1-to-many so our dataset is now identified by SubjectKey AND Week_Int. To get the true incidence for the period, we need to collapse the sample so that a row is identified by a SubjectKey. When we do that, we find the incidence is actually somewhat lower: approximately 45%

```
. gen HadMalaria = 0        if _m == 1
(860 missing values generated)

. replace HadMalaria = 1    if _m == 3
(860 real changes made)

. drop _m

. sum HadMalaria

    Variable |        Obs        Mean    Std. Dev.       Min        Max
```

```
  HadMalaria |      1,364    .6304985    .4828468          0          1

. unique SubjectKey
Number of unique values of SubjectKey is  919
Number of records is  1364

. preserve

.     collapse(max) HadMalaria, by(SubjectKey)

.     sum HadMalaria

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
  HadMalaria |        919    .4515778    .4979207          0          1

. restore
```

### Matching Health Events to Consultations

Now we have matched demographic characteristics to incidents of malaria, we can also use the data on consultations to see how long it took people to receive treatment. Here we will look at reported consultations in the same week as the report of the health event, and assume that if an individual was the subject of the event and a consultation, that the consultation was for malaria, though in pratice we would likely want to use additional variables in the dataset to ensure that this was the case. On this basis, we find that in approximately 90% of cases is there a reported consultation for a subject with malaria. We also find a similar percentage looking at the self-report variable HealthConsult.

```
. merge 1:1 SubjectKey Week_Int using "`health_consultations'", keep(match master
)

    Result                           # of obs.
    -----------------------------------------
    not matched                           587
        from master                       587  (_merge==1)
        from using                          0  (_merge==2)

    matched                               777  (_merge==3)
    -----------------------------------------

. gen ReceivedConsultation = 0  if _m == 1
(777 missing values generated)

. replace ReceivedConsultation = 1 if _m == 3
(777 real changes made)
```

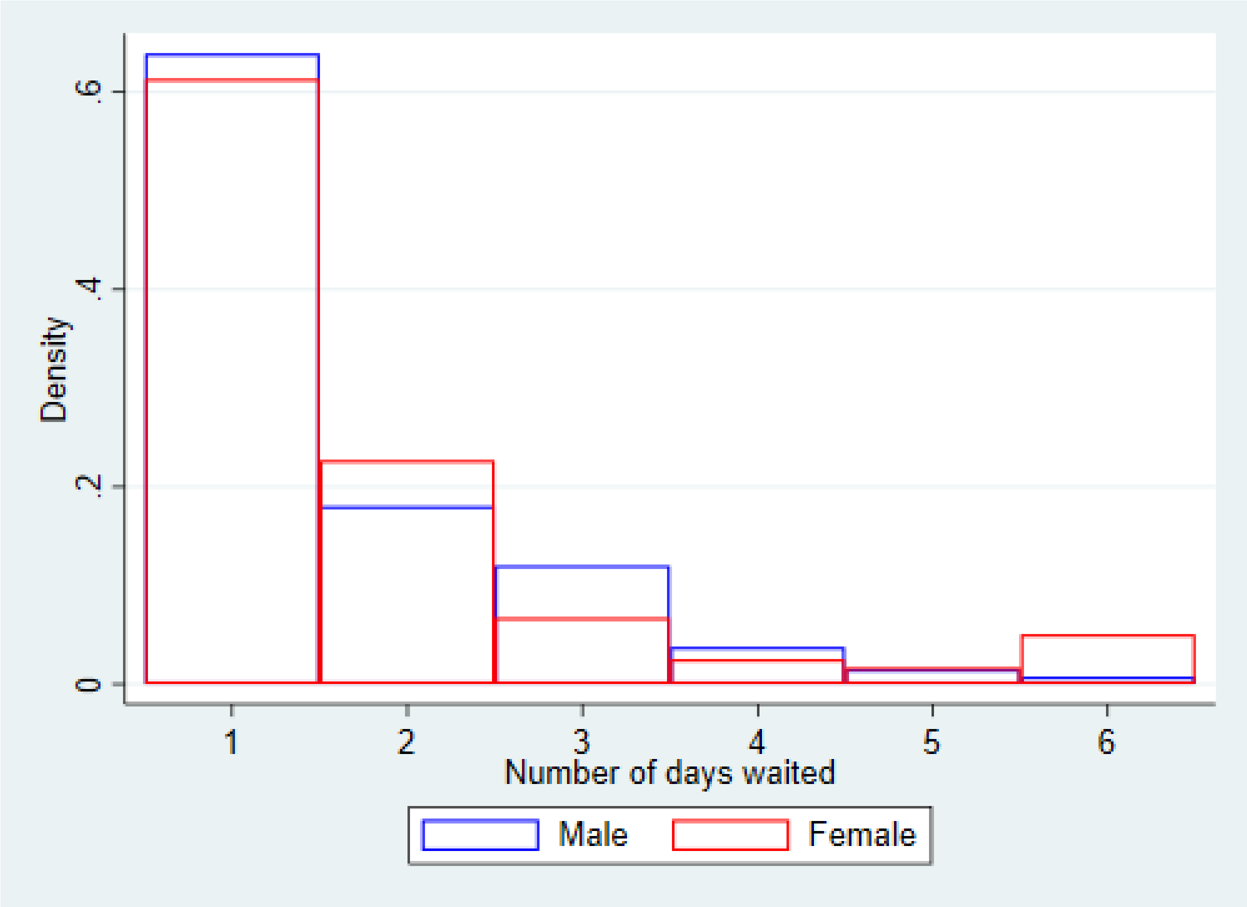
```
. drop _m

. sum ReceivedConsultation if HadMalaria == 1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
ReceivedCo~n |        860    .9034884    .2954634          0          1

. tab HealthConsult if HadMalaria == 1

     Anyone |
  Consulted |
 for Health |
    Problem |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         72        8.37        8.37
          1 |        788       91.63      100.00
------------+-----------------------------------
      Total |        860      100.00
```

### Calculating number of days between event and consultation

Combining the event and consultation data in this way also allows us to estimate how many days passed between when the subject reported being sick, and when they received a consultation. For example, we can see if there is any difference by gender in the average number of days between a child getting malaria or a fever and when they received a consultation, assuming it was not the same day.

```
. gen DaysWaited = ConsultDate - EventDate if ReceivedConsultation == 1
(620 missing values generated)

. sum DaysWaited

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
  DaysWaited |        744    .8400538    1.287671         -4          7

. replace DaysWaited = . if DaysWaited < 1
(372 real changes made, 372 to missing)

. bys Gender: sum DaysWaited if Age < 18

-------------------------------------------------------------------------------------
-----
-> Gender = MALE

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
```

```
  DaysWaited |        133    1.631579    1.018568          1          6

--------------------------------------------------------------------------------
-> Gender = FEMALE

    Variable |        Obs        Mean    Std. Dev.        Min        Max
-------------+---------------------------------------------------------
  DaysWaited |        119    1.756303     1.30167          1          6

--------------------------------------------------------------------------------
-> Gender = .

    Variable |        Obs        Mean    Std. Dev.        Min        Max
-------------+---------------------------------------------------------
  DaysWaited |          0


. twoway hist DaysWaited if Gender == 1 & Age < 18, color(blue) fcolor(none) fint
ensit
> y(inten30) discrete ///
>       || hist DaysWaited if Gender == 2 & Age < 18, color(red) fintensity(inten3
0) fc
> olor(none) discrete ///
>       xtitle("Number of days waited") legend(order(1 "Male" 2 "Female"))

. graph export example2_1.png, width(500) replace
(note: file example2_1.png not found)
(file example2_1.png written in PNG format)
```




### Exploring data by week

Suppose now that we wanted to look at the incidence by week, to see if there were particular times of the year when the incidence was higher. Our transactions data shows us the cases where health incidents were reported, but we need to also ensure that we account for weeks in which there was no reported health event. To do this, we can re-load the endline data, and use *expand* to create a row in the dataset corresponding to every week of the study, giving us 919 individuals x 55 weeks = 50455 rows.

```
. use "`endline_demographics'", clear

. expand 55
(49,626 observations created)

. sort SubjectKey

. by SubjectKey: gen Week_Int = _n

. inspect Week_Int

Week_Int:                                    Number of Observations
___________                             _______________________________________
```

```
                                              Total      Integers   Nonintegers
|  #   #   #   #   #       Negative              -             -             -
|  #   #   #   #   #       Zero                  -             -             -
|  #   #   #   #   #       Positive         50,545        50,545             -
|  #   #   #   #   #                    ----------    ----------    ----------
|  #   #   #   #   #       Total            50,545        50,545             -
|  #   #   #   #   #       Missing               -
+----------------------                 ----------
1                     55                    50,545
  (55 unique values)
```

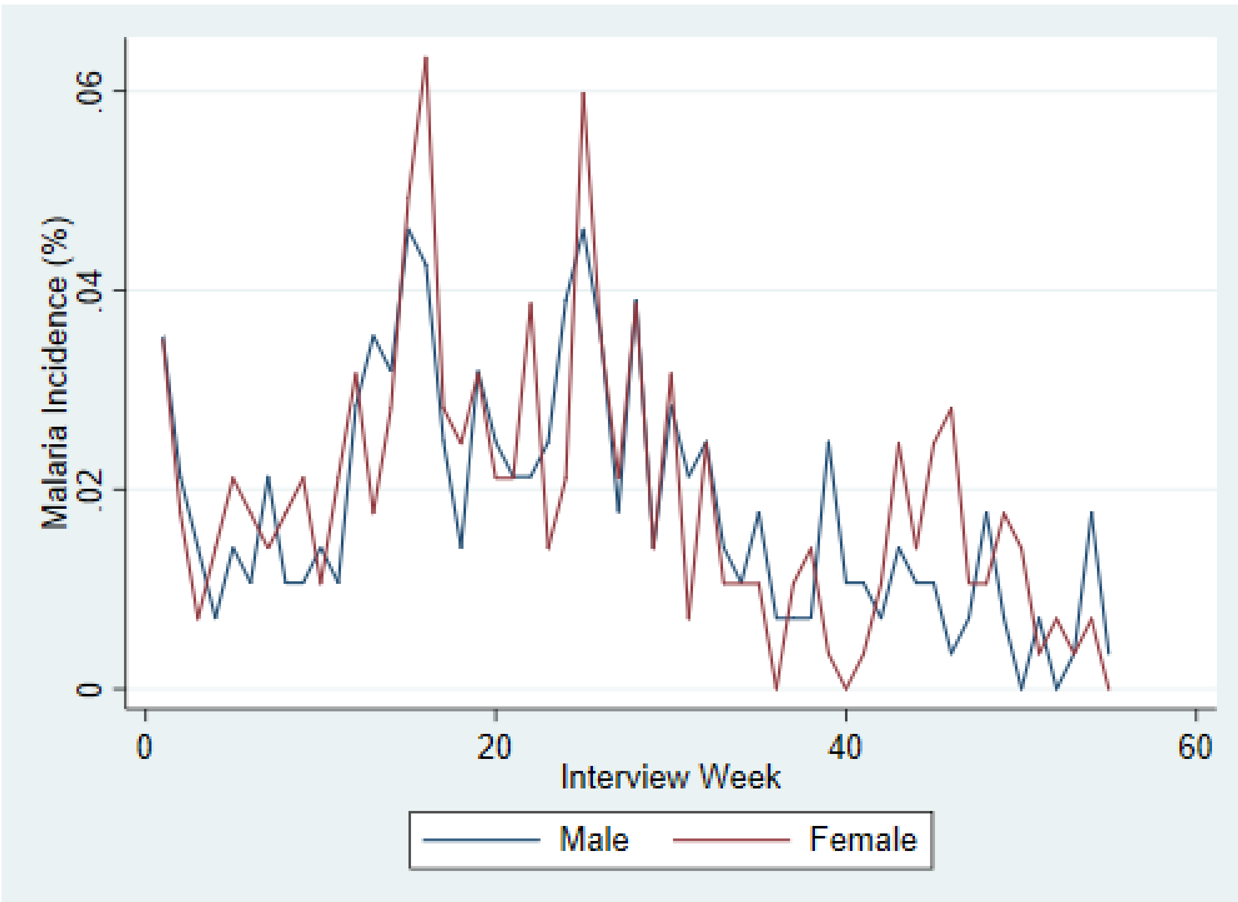
Now we can do a 1-to-1 merge to the weekly malaria data we created, and look at how the incidence varies by week. Note that for this example we are assuming complete reporting, ie. that there were no health events which were not recorded in the transactions dataset. Keep in mind that since we know the diaries interviews were not completed every week for every household, this is not likely to be a realistic assumption!

```
. merge 1:1 SubjectKey Week_Int using "`malaria_events'", keep(match master)

    Result                           # of obs.
    -----------------------------------------
    not matched                        49,685
        from master                    49,685  (_merge==1)
        from using                          0  (_merge==2)

    matched                               860  (_merge==3)
    -----------------------------------------

. gen HadMalaria = 0        if _m == 1
(860 missing values generated)

. replace HadMalaria = 1    if _m == 3
(860 real changes made)

. drop _m
```

Having merged the data and constructed our variable for malaria incidence, we can start to explore whether incidence varies by time or individual characteristics. For example here we create a line graph comparing malaria incidence by gender among children aged less than 18.

```
. preserve

.     keep if Age < 18
(19,415 observations deleted)

.     collapse(mean) HadMalaria, by(Week_Int Gender)

.     twoway line HadMalaria Week_Int if Gender == 1 ||  line HadMalaria Week_Int
if G
> ender == 2, ///
```

```
>         legend(order(1 "Male" 2 "Female")) xtitle("Interview Week") ytitle("Mal
aria
> Incidence (%)")

.     graph export example2_2.png, width(500) replace
(note: file example2_2.png not found)
(file example2_2.png written in PNG format)

. restore

. tempfile malaria_status_by_week

. save "`malaria_status_by_week'"
file C:\Users\MMURPHY\AppData\Local\Temp\ST_1b14_000007.tmp saved
```




### Using the weekly diaries dataset

We can also explore whether malaria status in a given week affects financial outcomes. Here we could use the transactions dataset to construct a series of aggregates to merge to our dataset, but fortunately most of the constructed variables we are likely to need have already been created in the weekly transactions dataset. Hence we simply need to merge to this dataset (note that we need to rename SubjectKey so that we correctly merge the subject of the health event to the same individual's reported transactions). In addition, the weekly dataset contains the variable Insured which provides the current health insurance status of an individual for that week, allowing us to compare insured and uninsured individuals. Let's merge in this data, and create an indicator for whether any member of a household had malaria in a given week

```
. ren SubjectKey MemberKey

. ren Week_Int WeekKey

. merge 1:1 MemberKey WeekKey using "weekly_nigeria.dta", keep(match) nogen

    Result                           # of obs.
    -----------------------------------------
    not matched                             0
    matched                            49,775
    -----------------------------------------

. keep if Result == 1 & Age > 18 & !missing(Age)
(36,195 observations deleted)

. bys WeekKey HouseholdKey: egen HouseholdMalaria = max(HadMalaria)
```

## Combining health events and financial flows

Now using this data, we can look at how a househol's malaria status may be related to members' financial behaviour. For example, we see that in weeks in which at least one of their household members has malaria, on average individuals make much larger withdrawals from their savings than individuals in households where no member reports malaria, and that on average this difference is greater for uninsured households relative to insured households.

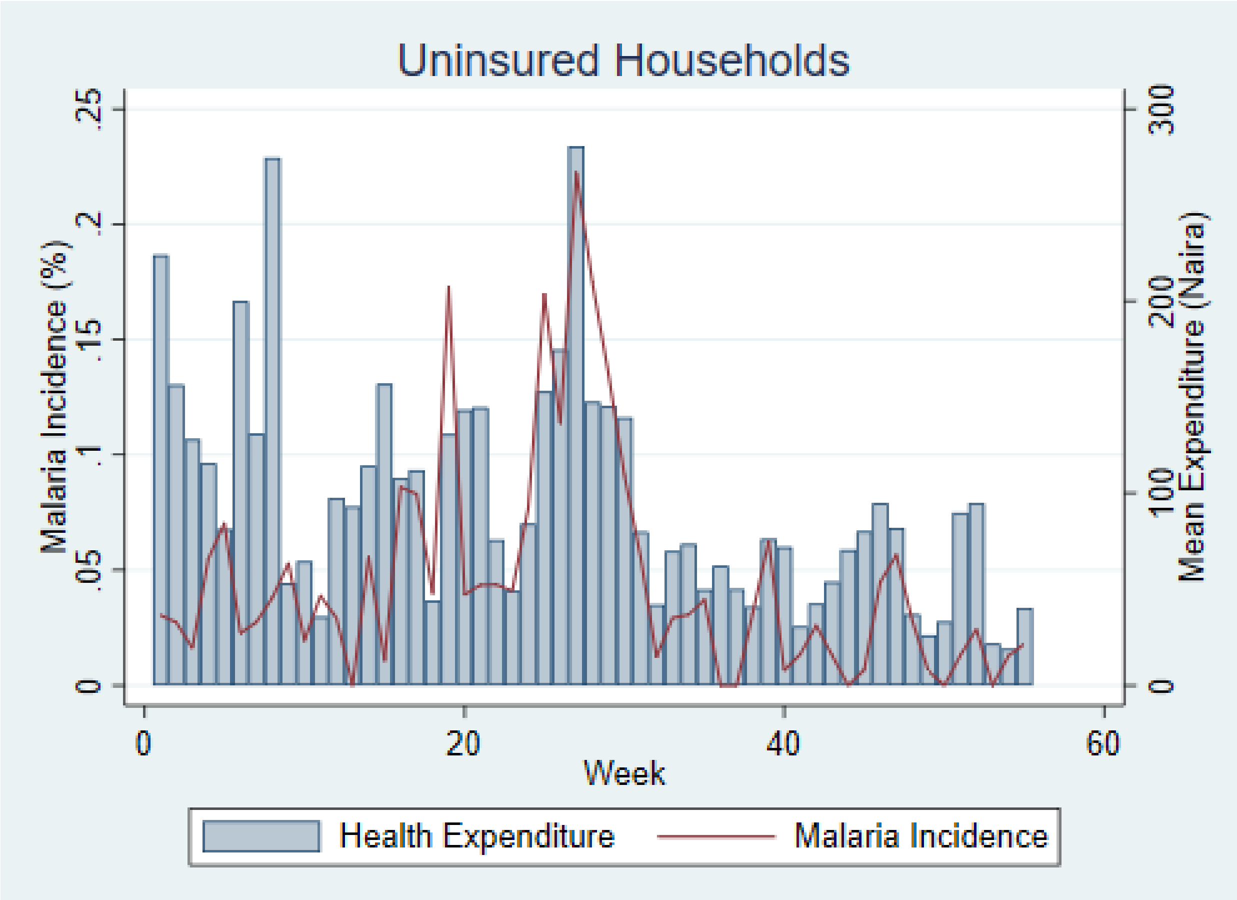
```
. bys HouseholdMalaria: sum withdrawal_Int

-------------------------------------------------------------------------------
-----
-> HouseholdMalaria = 0

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
withd~al_Int |     12,954    2748.034    13555.31          0     758050

-------------------------------------------------------------------------------
-----
-> HouseholdMalaria = 1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
```

```
-------------+---------------------------------------------------------
withd~al_Int |        626    4787.004    24144.34          0     500000


. bys Insured: sum withdrawal_Int if HouseholdMalaria == 1

-----------------------------------------------------------------------------------
-----
-> Insured = No

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
withd~al_Int |        421    5172.913    27069.17          0     500000

-----------------------------------------------------------------------------------
-----
-> Insured = Yes

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
withd~al_Int |        204    3997.027    16668.39          0     132500

-----------------------------------------------------------------------------------
-----
-> Insured = .

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
withd~al_Int |          1        3475           .       3475       3475
```

We can also compare weekly healthcare expenditures in uninsured households and see how this compares to the incidence of malaria in those households.

```
. replace health_Int = abs(health_Int)
(1,443 real changes made)

. collapse(mean) HouseholdMalaria health_Int, by(WeekKey Insured)

. twoway bar health_Int WeekKey if Insured == 0, yaxis(2) fintensity(inten30) yti
tle("
> Mean Expenditure (Naira)", axis(2)) || ///
>        line HouseholdMalaria WeekKey if Insured == 0, yaxis(1) yscale(alt) ysca
le(al
> t axis(2)) ///
>        legend(order(1 "Health Expenditure" 2 "Malaria Incidence")) xtitle("Week
") //
> /
>        ytitle("Malaria Incidence (%)") title("Uninsured Households")
```

```
. graph export example2_3.png, width(500) replace
(note: file example2_3.png not found)
(file example2_3.png written in PNG format)
```




## Conclusion

In this example we have explored how to work with some of the key variables around health using the Nigeria datasets. We have looked at matching health events to consultations, and seen how to use the diaries to look at time series data for both healthcare events and expenditures. In the next example, we will look in more detail at some potential ways to combine the health and financial data within the diaries.

## 3.3 Example 3- Working with the combined Financial & Health Diaries

### Overview

In the first two examples we looked at using financial data with the Kenya sample, and health data with the Nigeria sample. In this final example, we will look at how to combine data for both countries to explore linkages between financial and health outcomes in a combined sample. To do so, we will use data from the transactions datasets to identify individuals who purchased goods with known effects on health- alcohol and tobacco- and then combine the list of individuals we generate with other datasets to explore health and financial outcomes for users of these products.

### Subsetting country data of alcohol & tobacco purchases

First we are going to load the transactions data for Kenya, and subset it to include only transactions listed as purchases or sales (*TrType* == 5) and create a string variable to indicate the country. Then we do the same with the Nigeria data and append it.

We also include a couple of checks, first prior to merging we confirm that the item codes identifying the alcohol and tobacco products we are interested in match across the two samples. While major variables such as *TrType* and *PurchSaleItem* have been created such that their encoding matches across both datasets, this is not always the case. For example, both datasets contain an *Edu_Level* variable, but the values labels of this variable are specific to the country context and are not directly comparable across datasets. Users should always compare the encoding of categorical variables prior to merging country datasets.

```
. use "transactions_kenya.dta", clear

. keep if TrType == 5
(78,055 observations deleted)

. keep MemberKey TrType PurchSaleCategory PurchSaleItem PurchSaleItemS

. tab PurchSaleItem if inlist(PurchSaleItem,42,115,487,501,533)
```

| Good/Service Sold or Purchased | Freq. | Percent | Cum. |
|---:|---:|---:|---:|
| BEER | 83 | 35.47 | 35.47 |
| CIGARETTE | 137 | 58.55 | 94.02 |
| SNUFF | 8 | 3.42 | 97.44 |
| SPIRITS | 4 | 1.71 | 99.15 |
| TOBACCO | 2 | 0.85 | 100.00 |
| Total | 234 | 100.00 | |

```
. gen country = "kenya"

. tempfile purchases_kenya
```

```
. save "`purchases_kenya'"
file C:\Users\MMURPHY\AppData\Local\Temp\ST_1b14_000008.tmp saved

. use "transactions_nigeria.dta"

. keep if TrType == 5
(100,622 observations deleted)

. keep MemberKey TrType PurchSaleCategory PurchSaleItem PurchSaleItemS

. tab PurchSaleItem if inlist(PurchSaleItem,42,115,487,501,533)

         Good/Service Sold or Purchased |      Freq.     Percent        Cum.
----------------------------------------+-----------------------------------
                                   BEER |         30        5.69        5.69
                              CIGARETTE |        314       59.58       65.28
                                  SNUFF |        178       33.78       99.05
                                TOBACCO |          5        0.95      100.00
----------------------------------------+-----------------------------------
                                  Total |        527      100.00

. gen country = "nigeria"
```

### Appending purchase data & reviewing identifying variables

After subsetting and reviewing these categorical variables, we append our purchases dataset for Kenya to that for Nigeria. Now that the datasets are combined, it's a good idea to check that there are no common values of our MemberKey variable across the two datasets. We see there are not, which should help prevent problems later when we merge to other datasets. Note that for the Kenya dataset MemberKey values are 6 digits, while for Nigeria they are eight digits- so these IDs cannot overlap. However, other identifying variables, such as TrKey may be the same length, so care should be taken with these variables as they may indeed overlap between the two country datasets. If in doubt, adding a variable to identify the country can help ensure valid merges.

```
. append using "`purchases_kenya'", nolab

. unique MemberKey if country == "kenya"
Number of unique values of MemberKey is  199
Number of records is  93953

. unique MemberKey if country == "nigeria"
Number of unique values of MemberKey is  308
Number of records is  118292

. unique MemberKey
```

```
Number of unique values of MemberKey is  507
Number of records is  212245
```

### Creating a list of alchohol & tobacco users

Now that we have our combined purchases dataset, we can find members who report using alcohol or tobacco. To do this, we can use the *PurchSaleItem*, as well as some encoding some 'other' responses recorded under *PurchSaleItemS*. Then we can collapse the dataset to have a list of users of both products

```
. gen tobacco = 1 if inlist(PurchSaleItem,115,487,533)
(211,601 missing values generated)

. gen alcohol = 1 if inlist(PurchSaleItem,42,501)
(212,128 missing values generated)

. replace alcohol = 1 if strpos(lower(PurchSaleItemS),"cohol")
(14 real changes made)

. replace alcohol = 1 if strpos(lower(PurchSaleItemS),"brew")
(47 real changes made)

. collapse(max) tobacco alcohol, by(MemberKey)

. tempfile tobacco_alcohol_users

. save "`tobacco_alcohol_users'"
file C:\Users\MMURPHY\AppData\Local\Temp\ST_1b14_000009.tmp saved
```

### Matching the users list to household datasets

Now that we have identified these individuals, we can match them to the demographic characteristics in the household datasets to be able to better characterise them. Start by creating a simple combined dataset. In this case, we will retain variables for age and gender, and for some health outcomes: number of hospital visits, number of doctor visits, and insurance status at baseline.

```
. use "baseline_kenya.dta", clear

. keep MemberKey Age Gender q03_01-q03_04 status

. tempfile bl_kenya

. save "`bl_kenya'"
file C:\Users\MMURPHY\AppData\Local\Temp\ST_1b14_00000a.tmp saved
```

```
. use "baseline_nigeria.dta", clear

. keep MemberKey Age Gender q04_01-q04_04 q04_10

. gen country = "nigeria"

. append using "`bl_kenya'"
(note: variable Age was byte, now int to accommodate using data's values)

. replace country = "kenya" if country == ""
(587 real changes made)
```

After creating this dataset, we can merge in the indicators we created for tobacco and alcohol use.

```
. merge 1:1 MemberKey using "`tobacco_alcohol_users'", keep(match master) nogen

    Result                           # of obs.
    -----------------------------------------
    not matched                         1,722
        from master                     1,722
        from using                          0

    matched                               505
    -----------------------------------------

. replace tobacco = 0 if tobacco == .
(2,184 real changes made)

. replace alcohol = 0 if alcohol == .
(2,212 real changes made)

. gen alcohol_and_tobacco = 0 if tobacco == 0 & alcohol == 0
(53 missing values generated)

. replace alcohol_and_tobacco = 1 if tobacco == 1 | alcohol == 1
(53 real changes made)
```

### Using the household datasets to construct baseline healthcare usage

Now we can construct some health variables. Let's do one for the total number of visits to any type of health provider reported in the previous year. Note that the required variables we are using have different names in each dataset since they were collected using different surveys. They also include some differences in content. For example, *q04_03* from the Nigeria dataset asks about visits to patent medicine vendors, for which there is not a corresponding question in the Kenya dataset. For our example, we are going to include these in our constructed variable for number of health visits, but note that this aggregation may not be appropriate in other contexts. We will also create a common variable for insurance status at baseline- to make the example simpler we will use self-reported insurance status for Nigeria, and subset out individuals for whom this is missing.

```
. egen visits_kenya = rowtotal(q03_01-q03_04)

. egen visits_nigeria = rowtotal(q04_01-q04_04)

. gen HealthVisits_BL = visits_kenya + visits_nigeria

. gen AnyHealthVisits_BL = 0 if HealthVisits_BL == 0
(1,872 missing values generated)

. replace AnyHealthVisits_BL = 1 if HealthVisits_BL > 0 & !missing(HealthVisits_B
L)
(1,872 real changes made)

. drop visits_*

. gen IsInsured_BL = 0 if status == 2 | q04_10 == 2
(1,407 missing values generated)

. replace IsInsured_BL = 1 if status == 1 | q04_10 == 1
(631 real changes made)

. drop if IsInsured_BL == .
(776 observations deleted)

. drop status q*

. lab val IsInsured_BL yesno
```

### Exploring demographic and healthcare status of users at baseline

Now that we've constructed these variables, let's take a look at the characteristics of the alcohol and tobacco users in the combined data. We're going to first restrict the dataset to adults, since hopefully only adults are using these products (which we actually see in the data), then generate some simple descriptive statistics. We see that alcohol and tobacco users are somewhat older and much more likely to be male than adults who do not use these products. They are also more likely to have visited a provider of medical services, and somewhat less likely to be insured than non-users.

```
. keep if Age >= 18
(763 observations deleted)

. gen isFemale = Gender - 1

. bys alcohol_and_tobacco: sum Age isFemale AnyHealthVisits_BL IsInsured_BL
```

```
-> alcohol_and_tobacco = 0

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         Age |        644    41.01087    16.50821         18        104
    isFemale |        644    .6102484    .4880729          0          1
AnyHealthV~L |        644    .7919255     .406246          0          1
IsInsured_BL |        644    .4642857    .4991105          0          1

-> alcohol_and_tobacco = 1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         Age |         44    49.86364     15.3768         24         85
    isFemale |         44    .2272727    .4239151          0          1
AnyHealthV~L |         44    .8409091    .3699894          0          1
IsInsured_BL |         44    .4090909    .4973503          0          1


. tempfile baseline_with_status

. save "`baseline_with_status'"
file C:\Users\MMURPHY\AppData\Local\Temp\ST_1b14_00000b.tmp saved
```

### Creating a list of healthcare visits in the diaries data

We can also look at whether users of these products reported more health visits in the course of the diaries study. To do this, we can create a combined dataset of health consultations, collapsing to get the number of visits and the total expenditure on those visits. For each dataset we create an indicator for cases where the transaction is a health visit that was for an illness or injury, we then collapse those by *SubjectKey* to calculate the number of cases for each person.

Note the use of *SubjectKey*, since we are concerned about the number of cases in which the individual was the patient (if we used *MemberKey* we would calculate the number of cases that the individual reported). We then combine the two so that we have a simple dataset with IDs, number of visits & total expenditure. Hence when we think about the total visit cost, we are creating the total amount spent ON that individual, not the amount spent BY that individual on health care.

First we load the transactions data for Nigeria and use collapse to construct the healthcare outcomes by *SubjectKey*

```
. use "transactions_nigeria.dta", clear

. gen HealthVisits_Diaries = 0

. replace HealthVisits_Diaries = 1 if TrType == 12 & ConsultWhy == 1
(2,017 real changes made)

. egen HealthVisitCost_Diaries = rowtotal(Amount*) if TrType == 12 & ConsultWhy =
= 1
(216897 missing values generated)

. collapse(sum) HealthVisits_Diaries HealthVisitCost_Diaries, by(SubjectKey)

. drop if missing(SubjectKey)
(1 observation deleted)

. gen country = "nigeria"

. tempfile consultations_nigeria

. save "`consultations_nigeria'"
file C:\Users\MMURPHY\AppData\Local\Temp\ST_1b14_00000c.tmp saved
```

Then we do the same for Kenya

```
. use "transactions_kenya.dta", clear

. gen HealthVisits_Diaries = 0

. replace HealthVisits_Diaries = 1 if TrType == 12 & ConsultWhy == 1
(546 real changes made)

. egen HealthVisitCost_Diaries = rowtotal(Amount*) if TrType == 12 & ConsultWhy =
= 1
(171462 missing values generated)

. collapse(sum) HealthVisits_Diaries HealthVisitCost_Diaries, by(SubjectKey)

. drop if missing(SubjectKey)
(1 observation deleted)

. gen country = "kenya"
```

And combine the two

```
. append using "`consultations_nigeria'"
(note: variable country was str5, now str7 to accommodate using data's values)

. tempfile diaries_consultations
```

```
. save "`diaries_consultations'"
file C:\Users\MMURPHY\AppData\Local\Temp\ST_1b14_00000d.tmp saved
```

### Matching healthcare visits from the diaries data to our alcohol/tobacco user list

Using the consultation data that we created, we can merge this to our baseline dataset. We're going to restrict the dataset to only include individuals who participated in the diaries (ie. financially active adults, excluding those from the Nigeria sample that were in the control group). To do this we combine the transasctions datasets, and generate a list of all member IDs. We can then merge this with the baseline file we already created to add demographic characteristics. We then create a copy of the MemberKey variable called SubjectKey, so that we can match to the subjects of the health consultations in the dataset we just created.

Create a subset of all the individuals who participated in the diaries

```
. use "transactions_kenya"

. qui append using "transactions_nigeria", force

. bys MemberKey: keep if _n == 1
(390,410 observations deleted)

. keep MemberKey

. tempfile diaries_participants

. save "`diaries_participants'"
file C:\Users\MMURPHY\AppData\Local\Temp\ST_1b14_00000e.tmp saved
```

And merge to the baseline & consultations datasets we created

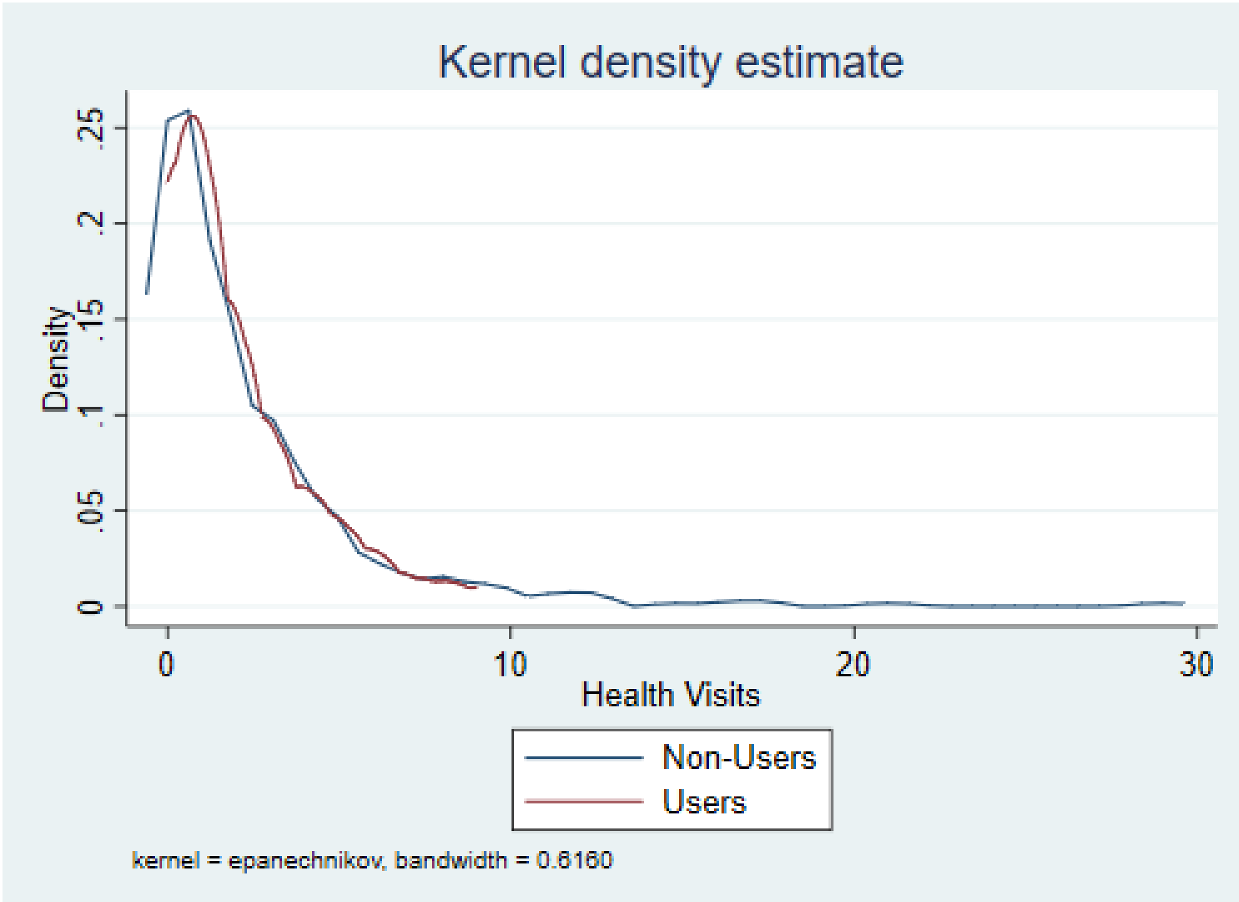
```
. merge 1:1 MemberKey using "`baseline_with_status'", keep(match) nogen

    Result                           # of obs.
    -----------------------------------------
    not matched                             0
    matched                               406
    -----------------------------------------

. clonevar SubjectKey = MemberKey

. merge 1:1 SubjectKey using "`diaries_consultations'", keep(match master)

    Result                           # of obs.
    -----------------------------------------
    not matched                            35
        from master                        35  (_merge==1)
        from using                          0  (_merge==2)
```

```
    matched                                   371  (_merge==3)
    -----------------------------------------

. replace HealthVisits_Diaries = 0 if _m == 1
(35 real changes made)

. replace HealthVisitCost_Diaries = 0 if _m == 1
(35 real changes made)

. drop _m
```

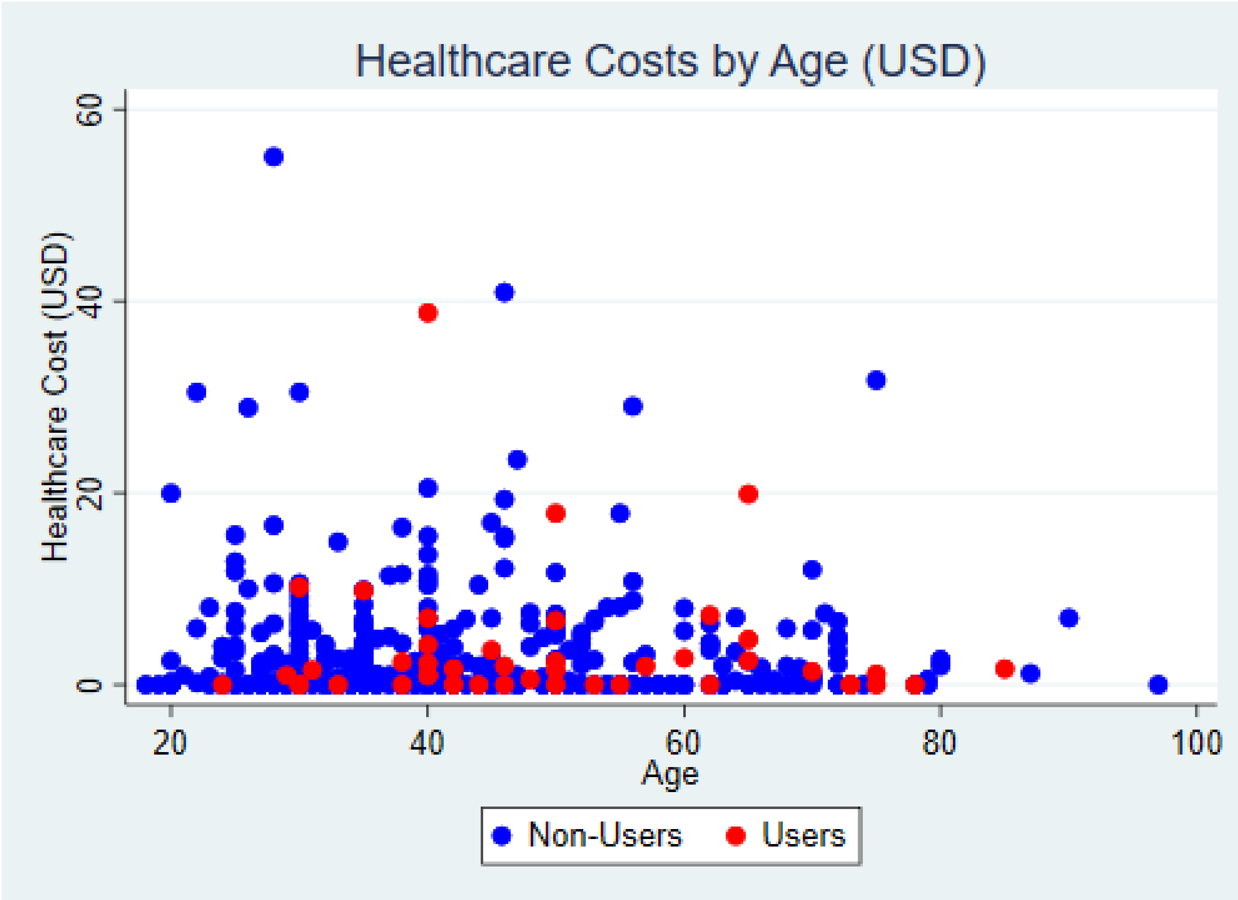
## Plotting healthcare visits of users & non-users by age

Now that we have our completed dataset which matches alcohol/tobacco status to health incidents, we can use this to do some initial analysis. In this case we overlay the distributions of the number of visits by users/non-users of these products, and then generate a scatter plot showing the relationship between health visits costs and age for users & non-users. To make the cost variables comparable we convert the expenditure variables into USD (for this example we simply use today's exchange rate)

```
. replace HealthVisitCost_Diaries = HealthVisitCost_Diaries / 102.384 if country
== "k
> enya"
(95 real changes made)

. replace HealthVisitCost_Diaries = HealthVisitCost_Diaries / 360.844 if country
== "n
> igeria"
(158 real changes made)

. kdensity HealthVisits_Diaries if alcohol_and_tobacco == 0, addplot(kdensity Hea
lthVi
> sits_Diaries if alcohol_and_tobacco ==1) ///
>     legend(order(1 "Non-Users" 2 "Users")) xtitle("Health Visits")

. graph export example3_1.png, width(500) replace
(note: file example3_1.png not found)
(file example3_1.png written in PNG format)
```




```
. twoway scatter HealthVisitCost_Diaries Age if alcohol_and_tobacco == 0 & Health
Visit
> Cost_Diaries < 100, color("blue") || ///
>         scatter HealthVisitCost_Diaries Age if alcohol_and_tobacco == 1, color(
"red"
> ) ///
>         legend(order(1 "Non-Users" 2 "Users")) xtitle("Age") ytitle("Healthcare
Cost
>  (USD)") title("Healthcare Costs by Age (USD)")

. graph export example3_2.png, width(500) replace
(note: file example3_2.png not found)
(file example3_2.png written in PNG format)
```




### Matching user list to weekly transactions datasets

Similarly we could merge the dataset we created with alcohol/tobacco status to the weekly transactions dataset to compare seasonal trends for these individuals relative to non-users. To do this we again create a master list of diaries participants with alcohol/tobacco status and combine this with the weekly datasets, just as we did for the transactions dataset. We are also going to restrict to completed interviews.

Load participant list and merge to baseline with alcohol/tobacco status

```
. use "`diaries_participants'", clear

. merge 1:1 MemberKey using "`baseline_with_status'", keep(match) nogen

    Result                           # of obs.
    -----------------------------------------
    not matched                             0
    matched                               406
    -----------------------------------------
```

Then merge to Kenya weekly data

```
. preserve

.     keep if country == "kenya"
(205 observations deleted)

.     merge 1:m MemberKey using "weekly_kenya.dta", keep(match) nogen

    Result                           # of obs.
    -----------------------------------------
    not matched                             0
    matched                            11,055
    -----------------------------------------

.     tempfile flows_kenya

.     save "`flows_kenya'"
file C:\Users\MMURPHY\AppData\Local\Temp\ST_1b14_00000g.tmp saved

. restore
```

And to Nigeria weekly data, and combine.

```
. keep if country == "nigeria"
(201 observations deleted)

. merge 1:m MemberKey using "weekly_nigeria.dta", keep(match) nogen

    Result                           # of obs.
    -----------------------------------------
    not matched                             0
    matched                            11,275
    -----------------------------------------

. tostring InterviewDate, replace
InterviewDate was long now str8

. append using "`flows_kenya'"
(note: variable BA_POS_income_Int was long, now double to accommodate using data'
s
       values)
(note: variable repaymentmade_Int was long, now double to accommodate using data'
s
       values)
(note: variable BA_POS_income_Tr was long, now double to accommodate using data's
       values)
(note: variable repaymentmade_Tr was long, now double to accommodate using data's
       values)
(note: variable AD_NEG_purch_Tr was long, now double to accommodate using data's
       values)
(label Result already defined)
(label Q00_02 already defined)
```

```
. keep if Result == 1
(5,144 observations deleted)
```

### Creating a weekly date identifier across country datasets

Now that we have a combined dataset, we can look at flows across the two countries. One issue however is that the WeekKey variable which identifies the interview week in the data (and takes the values 1-55), refers to different time periods for the two different countries. To deal with this for this example, we are going to use the most common interview date reported for a given WeekKey to assign a fixed date to each week. First we need to convert the interview date into a Stata date var. Then assign the date for each week by country (in cases of mulitple modes we use the earliest date)

```
. gen FormattedDate = date(InterviewDate,"YMD")

. format FormattedDate %td

. bys country WeekKey: egen WeekDate = mode(FormattedDate), minmode

. format WeekDate %td
```

### Plotting weekly cash purchases by country, alchohol/tobacco status

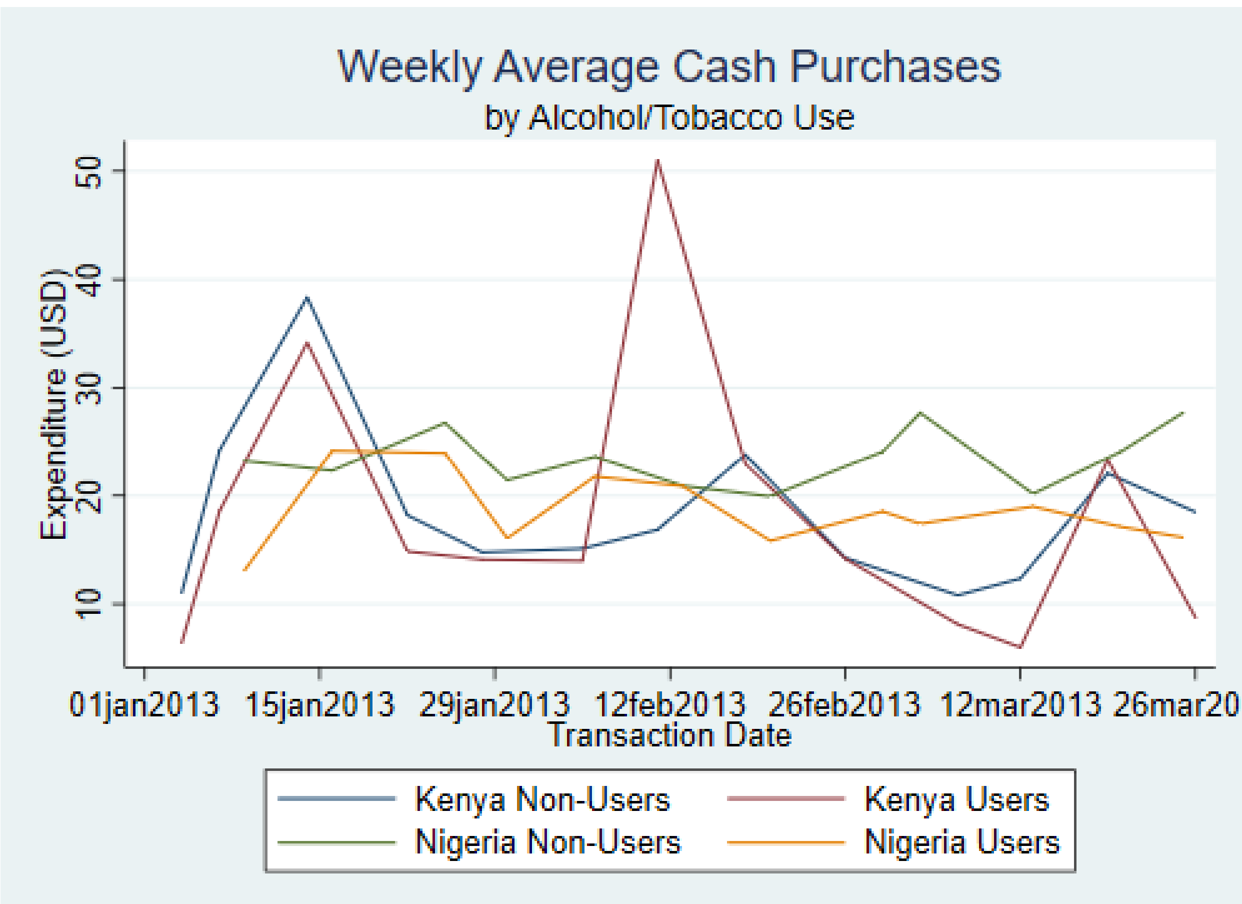
Now that we have a date associated with each week, we can create meaningful visualizations of flows across periods. For example, here we explore patterns of purchases made in cash for both groups for the first 3 months of 2013. We start by converting the variable to a positive value (since purchases are an outflow) and then convert to USD so that the two countries are comparable for the example.

Construct mean weekly purchases variable in USD

```
. replace purchase_Int = abs(purchase_Int)
(17,050 real changes made)

. replace purchase_Int = purchase_Int / 102.384 if country == "kenya"
(7,633 real changes made)

. replace purchase_Int = purchase_Int / 360.844 if country == "nigeria"
(9,417 real changes made)

. collapse(mean) purchase_Int, by(country WeekDate alcohol_and_tobacco)

. gen Q1_2013 = 1 if WeekDate>=mdy(1,1,2013) & WeekDate<mdy(3,31,2013)
(168 missing values generated)
```

And plot by interview week

```
. twoway line purchase_Int WeekDate if Q1_2013 == 1 & country == "kenya" & alcoho
l_and
> _tobacco == 0 || ///
> line purchase_Int WeekDate if Q1_2013 == 1 & country == "kenya" & alcohol_and_t
obacc
> o == 1 || ///
> line purchase_Int WeekDate if Q1_2013 == 1 & country == "nigeria" & alcohol_and
_toba
> cco == 0 || ///
> line purchase_Int WeekDate if Q1_2013 == 1 & country == "nigeria" & alcohol_and
_toba
> cco == 1, ///
> legend(order(1 "Kenya Non-Users" 2 "Kenya Users" 3 "Nigeria Non-Users" 4 "Niger
ia Us
> ers")) ///
> xlabel(#6) xtitle("Transaction Date") ytitle("Expenditure (USD)") ///
> title("Weekly Average Cash Purchases") subtitle("by Alcohol/Tobacco Use")

. graph export example3_3.png, width(500) replace
(note: file example3_3.png not found)
(file example3_3.png written in PNG format)
```



### Conclusion

In this final example, we have explored how we can use data from the transactions datasets to identify individuals purchasing goods with known health outcomes. We have demonstrated how to link this to both household survey data and other transaction data to explore how other financial and health outcomes may vary for users of these goods. We have also worked with combining the Kenya and Nigeria datasets and noted some areas of concern when working with the combined data.

## Acknowledgement

Embedding of Stata code and output for these examples generated using the excellent markstat command, created by Germán Rodríguez. For more details and documentation, go to http://data.princeton.edu/stata/markdown/.

## Section 4- Overview of publications based on the Financial and Health Diaries datasets

### Articles in international peer-reviewed journals


Geng, X., W. Janssens, B. Kramer and M. van der List (2018), "Health Insurance, a Friend in Need? Evidence from Financial and Health Diaries in Kenya", *World Development,* Vol. 111, November 2018, pp. 196-210*.* *doi:10.1016/j.worlddev.2018.07.004*

Barr, A., M. Dekker, W. Janssens, B. Kebede and B. Kramer (Forthcoming), "Cooperation in Polygynous Households", *American Economic Journal: Applied Economics*

Janssens, W, B. Kramer and L. Swart (2017), "Be patient when measuring hyperbolic discounting: Stationarity, time consistency and time invariance in a field experiment", *Journal of Development Economics*, Vol. 126, May 2017, pp. 77-90. *doi: 10.1016/j.jdeveco.2016.12.011*.


### Working papers *(under review)*


Geng, X., W. Janssens and B. Kramer (2017), "Liquid Milk: Effects of Cash Constraints on Collective Marketing in the Kenyan Dairy Sector", IFPRI discussion paper 01602, (January 2017).


### Blogs


Kramer, B. (June 1, 2017). World Milk Day: How Kenyan dairy farmers manage their money. Posted on the IFPRI Blog, *http://www.ifpri.org/blog/world-milk-day-how-kenyan-dairy-farmers-manage-their-money*

Gustafson, S. and Kramer, B. (July 27, 2017). Be patient: Avoid "false positives" in measuring commitments to save. Posted on the IFPRO Blog, *http://www.ifpri.org/blog/be-patient-avoid-"false-positives"-measuring-commitments-save*

Kramer, B. (February 8, 2013). What's Next? Connecting Finance and Health. Posted on the Financial Access Initiative (FAI) NYU Wagner Blog, *https://www.financialaccess.org/blog/2015/7/30/whats-next-connecting-finance-and-health*

## Policy briefs

Janssens, W., B. Kramer, T.M. Akande, G.K. Osagbemi, H. Ameen, M. van der List and A. Langedijk-Wilms (2017), "Using diaries to improve a health insurance program to better meet health needs in rural Nigeria", PharmAccess Health Analytics Brief, No. 4, April 2017.

Janssens, W., B. Kramer, T.M. Akande, G.K. Osagbemi, H. Ameen, M. van der List and A. Langedijk-Wilms (2017), "What do the Financial and Health Diaries tell us about decisions to renew in health insurance?", PharmAccess Health Analytics Brief, No. 3, April 2017.

Janssens, W., B. Kramer, M. van der List, C. Lohr and A. Langedijk-Wilms (2016), "What factors influence the decision to enroll in health insurance in rural Kenya?", PharmAccess Learning & Analysis Brief, No. 1 March 2016.

Janssens, W., B. Kramer, M. van der List, C. Lohr and A. Langedijk-Wilms (2015), "An Overview of the Financial and Health Diaries Study in Nigeria and Kenya", PharmAccess Learning & Analysis Brief, No. 13 December 2015.

## MSc theses

Jager, A. de (2016), "Economic Impacts of Illness, Coping Strategies, Health Insurance and Crowding-Out Effects: Dynamic Effects from a Financial Diary Survey in Nigeria", University of Amsterdam, MSc thesis.

Hardy, M. (2014), "Illness, bargaining and health care: Health care decisions within households in Kwara, Nigeria", Vrije Universiteit Amsterdam, Department of Economics, MSc thesis.

Hofstee, S.N. (2014), "Informal Firm Owners and Shocks in Developing Countries: An Empirical Investigation of a Household-Firm Model in Nigeria", Vrije Universiteit Amsterdam, Department of Economics, MSc thesis.

Hermans, I. (2014), "Measurement of Income Patterns Based on Subjective Expectations", Vrije Universiteit Amsterdam, Department of Economics, MSc thesis.

Ohrnberger, J. (2014), "Of Farmers and Prospect Theory – A Behavioral Economic Analysis of Agricultural Investment in Nigeria", Vrije Universiteit Amsterdam, Department of Economics, MSc thesis.

Ide, V.N. (2014), "Shocks, Remittances, Insurance and M-Pesa: Evidence from Western Kenya", Vrije Universiteit Amsterdam, Department of Economics, MSc thesis.

Swart, L. (2012), "Irrational Financial Decision-Making? Evidence from a Field Experiment in Rural Nigeria", Tinbergen Institute MPhil thesis.

## Other work-in-progress

Dekker, M., W. Janssens, B. Kramer and P. Rossi, “Intra-household risk-sharing and polygamy”.

Nelissen, H.E., D. Brals, W. Janssens, B. Kramer, M. van der List, T.M. Akande, A.H. van ’t Hoog, “Healthcare provider choices in case of a health event in the context of a health insurance scheme in rural Nigeria”.

Janssens, W. and B. Kramer, “Wealthy, Healthy and Insured? Cash Constraints and Health-Seeking Behavior in Financial and Health Diaries Data”.

Okunogbe, A., B. Kramer., M. Pradhan and W. Janssens, “The impact of health insurance on informal risk-sharing and consumption smoothing in Nigeria”.

Janssens, W., B. Kramer and M. Murphy, “Gender differences in health risk management: Evidence from financial diaries on the vulnerability and resilience of rural women in Kenya and Nigeria”

Janssens, W., B. Kramer and L. Swart, “Why do people revise their plans?  Combining a time preference experiment with financial diaries in Nigeria”.

Lehmann-Uschner, K. and B. Kramer (2018), “Impacts of health shocks on nutrition: Evidence from financial diaries”.