GlobalMix Mozambique Aim 2 participant data extraction.

MC Kiti

We use this file to extract participant metadata for further analysis.

Lost sensors

```
# we lost some sensors and we do not have the hwids. let's identify these
# individuals and see their characteristics then delete them from the data.
lost_sensorid <- c("041802E2", "077075CE", "0DD5BD3F", "16B7EA20", "1BF0AB62",
                   "99FF6D3B", "A498A0A5", "B0834C7E", "BD848166", "BFAAAFE9",
                   "E691CC38", "EC52843D", "F554A006", "FC1DE353")

lost_sensors_df <- data %>%
  filter(sensor_id %in% lost_sensorid) %>%
  select(rec_id, study_site, hh_id, isindex, sensor_id, participant_age, participant_sex,
         read_write, occupation, enrolled_school)

summary_lost_sensors <- lost_sensors_df %>%
  tbl_summary(by="study_site",
              percent="column",
              missing="ifany",
              digits = all_categorical() ~ 0) %>%
  bold_labels() %>%
  modify_header(label = "**Variable**") %>%
  modify_caption("")
summary_lost_sensors
```

```
Table printed with `knitr::kable()`, not {gt}. Learn why at
http://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
To suppress this message, include `message = FALSE` in code chunk header.
```

Variable	**Rural**, N = 0	**Urban**, N = 19
rec_id	NA (NA, NA)	903 (630, 1,032)
hh_id		
BL1GG1030	0 (NA%)	1 (5%)
BL1GG1040	0 (NA%)	2 (11%)
BL2CM2003	0 (NA%)	2 (11%)
BLACM1022	0 (NA%)	1 (5%)
QU8NM1002	0 (NA%)	1 (5%)
QUOMU1033	0 (NA%)	1 (5%)
QURMU1007	0 (NA%)	1 (5%)
QUSRT1029	0 (NA%)	1 (5%)
QUTRT1061	0 (NA%)	1 (5%)
QUUGG1003	0 (NA%)	2 (11%)
QUUGG1005	0 (NA%)	1 (5%)
QUVAM2023	0 (NA%)	3 (16%)
QUYPC1144	0 (NA%)	1 (5%)
QUYPC1195	0 (NA%)	1 (5%)
isindex	0 (NA%)	4 (21%)
sensor_id		
041802E2	0 (NA%)	1 (5%)
077075CE	0 (NA%)	1 (5%)
0DD5BD3F	0 (NA%)	2 (11%)
1BF0AB62	0 (NA%)	1 (5%)
99FF6D3B	0 (NA%)	2 (11%)
A498A0A5	0 (NA%)	1 (5%)
B0834C7E	0 (NA%)	1 (5%)
BD848166	0 (NA%)	1 (5%)
BFAAAFE9	0 (NA%)	3 (16%)
E691CC38	0 (NA%)	2 (11%)
EC52843D	0 (NA%)	1 (5%)
F554A006	0 (NA%)	1 (5%)
FC1DE353	0 (NA%)	2 (11%)
Participant age		
<6mo	0 (NA%)	0 (0%)
6-11mo	0 (NA%)	0 (0%)
1-4y	0 (NA%)	6 (32%)
5-9y	0 (NA%)	1 (5%)
10-14y	0 (NA%)	3 (16%)
15-19y	0 (NA%)	2 (11%)
20-29y	0 (NA%)	2 (11%)
30-39y	0 (NA%)	2 (11%)
40-59y	0 (NA%)	3 (16%)

Variable	**Rural**, N = 0	**Urban**, N = 19
60+y	0 (NA%)	0 (0%)
Participant sex		
Female	0 (NA%)	12 (63%)
Male	0 (NA%)	7 (37%)
Can the participant read and write on their own?	0 (NA%)	11 (58%)
What is your occupation?		
Child	0 (NA%)	0 (0%)
Unemployed	0 (NA%)	1 (8%)
Student	0 (NA%)	5 (42%)
Homemaker	0 (NA%)	0 (0%)
Casual laboror	0 (NA%)	0 (0%)
Farmer	0 (NA%)	0 (0%)
Fisherman	0 (NA%)	0 (0%)
Business person	0 (NA%)	1 (8%)
Office worker	0 (NA%)	5 (42%)
Retired	0 (NA%)	0 (0%)
Other	0 (NA%)	0 (0%)
Unknown	0	7
Are you currently enrolled in school?	0 (NA%)	6 (33%)
Unknown	0	1

```
# extract the rec_id associated with the lost sensors. these will be dropped from the main

lost_sensor_recid <- lost_sensors_df %>%
  select(rec_id)

# currently, we have 703 observations (300 urban).
data <- data %>%
  filter(!rec_id %in% lost_sensor_recid$rec_id)
# after dropping, we have 684 records (281 urban), so we have dropped 19 records.

rm(lost_sensor_recid, lost_sensorid)
```

All the 19 individuals who lost sensors are from 14 households in the urban site. Out of these, 5 households had >2 sensors that were lost (1 had 3 missing and 1 had 4 missing). During sensor data cleaning, it will be important to take this into account for these households.

Summary of participant characteristics by site

Table 2 shows the total number of individuals who were recruited into the study and their demographic characteristics by site.

Table 2: Particiant data from Mozambique sites

Variable	**Rural**, N = 302	**Urban**, N = 281
isindex	57 (19%)	61 (22%)
Participant sex		
Female	175 (58%)	166 (59%)
Male	126 (42%)	114 (41%)
Unknown	1	1
Participant age		
<6mo	27 (9%)	22 (8%)
6-11mo	23 (8%)	31 (11%)
1-4y	40 (13%)	37 (13%)
5-9y	50 (17%)	32 (11%)
10-14y	43 (14%)	28 (10%)
15-19y	24 (8%)	19 (7%)
20-29y	44 (15%)	57 (20%)
30-39y	26 (9%)	27 (10%)
40-59y	18 (6%)	20 (7%)
60+y	7 (2%)	8 (3%)
Can the participant read and write on their own?	124 (41%)	154 (55%)
What is your occupation?		
Child	50 (20%)	53 (23%)
Unemployed	66 (26%)	38 (16%)
Student	84 (34%)	68 (29%)
Homemaker	3 (1%)	13 (6%)
Casual laboror	8 (3%)	11 (5%)
Farmer	21 (8%)	0 (0%)
Fisherman	0 (0%)	0 (0%)
Business person	3 (1%)	23 (10%)
Office worker	5 (2%)	25 (11%)
Retired	0 (0%)	1 (0%)
Other	10 (4%)	2 (1%)
Unknown	52	47
Are you currently enrolled in school?	107 (37%)	71 (26%)
Unknown	13	11

Rural sensor metadata

Here, we clean the sensor participant data. Steps are: 1. Check that all the sensor_ids are available. 2. Replace sensor_id written incorrectly. These were identified visually and via code. 3. Drop records with inconsistent sensor_id and share with data managers for checks. 4. Merge with hwid data. This was extracted from the sensors by data managers. 5. Check missing sensor_id and hwid data. Export records for DMs to check.

```
There are 25 sensors that do not have hwid data, used by 67 individuals missing hwids, so we
```

```
  0   1
184  46
```

```
[1] 61
```

```
# summary of updated rural participants
rural_summary_aim2 <- rural_sensors %>%
  select(study_site, isindex, participant_sex, participant_age, read_write,
         occupation, enrolled_school) %>%
  tbl_summary(statistic = list(all_continuous() ~ "{median ({IQR})"),
              percent="column",
              missing="ifany",
              digits = all_categorical() ~ 0) %>%
  bold_labels() %>%
  modify_header(label = "**Variable**") %>%
  modify_caption("")
rural_summary_aim2
```

```
Table printed with `knitr::kable()`, not {gt}. Learn why at
http://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
To suppress this message, include `message = FALSE` in code chunk header.
```

Variable	**N = 230**
study_site	
Rural	230 (100%)
Urban	0 (0%)
isindex	46 (20%)

Variable	N = 230
participant_sex	
Female	142 (62%)
Male	87 (38%)
Unknown	1
participant_age	
<6mo	22 (10%)
6-11mo	19 (8%)
1-4y	30 (13%)
5-9y	39 (17%)
10-14y	32 (14%)
15-19y	14 (6%)
20-29y	35 (15%)
30-39y	21 (9%)
40-59y	13 (6%)
60+y	5 (2%)
read_write	92 (40%)
occupation	
Child	41 (21%)
Unemployed	52 (27%)
Student	60 (31%)
Homemaker	2 (1%)
Casual laboror	5 (3%)
Farmer	15 (8%)
Fisherman	0 (0%)
Business person	2 (1%)
Office worker	5 (3%)
Retired	0 (0%)
Other	10 (5%)
Unknown	38
enrolled_school	79 (36%)
Unknown	9

Table 4 shows the characteristics of individuals with missing sensor IDS and hwids (n=67).

Table 4: Rural records missing sensor_id and hwid data

Variable	N = 67
isindex	11 (16%)
participant_sex	
Female	31 (46%)

Variable	N = 67
Male	36 (54%)
participant_age	
<6mo	5 (7%)
6-11mo	4 (6%)
1-4y	10 (15%)
5-9y	11 (16%)
10-14y	9 (13%)
15-19y	7 (10%)
20-29y	9 (13%)
30-39y	5 (7%)
40-59y	5 (7%)
60+y	2 (3%)

In total, there are 302 participants who were issued sensors from the rural site. Out of these, we have complete records for 235. This is after dropping records due to missing sensor_id and hwid.

> **Note**
>
> The hwid is an important identifier for the participants and is the primary key for the sensor data. Without the hwid, we cannot link the sensor data to the participant data collected in REDCap.

Checking household IDs

```
cat("Here, we want to find out households thatn are present in all the raw files i.e.
    zipped, unzipped, h5 and REDCap database. This is because I was having a hard time rec
```

```
Here, we want to find out households thatn are present in all the raw files i.e.
    zipped, unzipped, h5 and REDCap database. This is because I was having a hard time recon
```

```
cat("There are", length(unique(rural_sensors$hh_id)), "households in the rural
    area. This means that there are individuals who migrated from one household
    and were now living in a new household, but their perm_id was not updated to
    reflect the move to the new household.")
```

There are 61 households in the rural
    area. This means that there are individuals who migrated from one household
    and were now living in a new household, but their perm_id was not updated to
    reflect the move to the new household.

```
n_rural_household_members <- rural_sensors %>%
  select(rec_id, hh_id) %>%
  group_by(hh_id) %>%
  summarise(n=n())
cat("There are 2 households with only 1 member. This should not be the case.")
```

There are 2 households with only 1 member. This should not be the case.

```
# extract the file names in the raw folder.
unzip_path <- c("/Users/mck/Library/CloudStorage/OneDrive-EmoryUniversity/4.Data and Analy
zip_path <- c("/Users/mck/Library/CloudStorage/OneDrive-EmoryUniversity/4.Data and Analysi
h5_path <- c("/Users/mck/Library/CloudStorage/OneDrive-EmoryUniversity/4.Data and Analysis

# library(stringr)
# filterr <- c("SM-rural-supervisors-files-20210727102207",
#              "SM-rural-supervisors-files-20210811091335",
#              "SM-manhica_supervisors-files-20210727102207",
#              "SM-manhica_supervisors-files-20210811091335",
#              "infants_only",
#              "Household sensor ID list comparision.xlsx",
#              "deleted duplicated")
#
rural_unzip_hmid <- as.data.frame(list.files(path = paste(unzip_path))) %>%
  rename(hmid_unzip = names(.[1])) %>%
  # filter(!hmid_unzip %in% unlist(filterr)) %>%
  mutate(hmid_unzip = substr(hmid_unzip, start = 10, stop=nchar(hmid_unzip))) %>%
  # str_replace_all("_", "-") %>%
  na.omit()

rural_zip_hmid <- as.data.frame(list.files(path = paste(zip_path))) %>%
  rename(hmid_zip = names(.[1])) %>%
  # filter(!hmid_zip %in% unlist(filterr)) %>%
  mutate(hmid_zip = substr(hmid_zip, start = 12, stop=nchar(hmid_zip)-25)) %>%
  # gsub("_", "-", .) %>%
  na.omit()
```

```r
rural_h5_hmid <- as.data.frame(list.files(path = paste(h5_path))) %>%
  rename(hmid_h5 = names(.[1])) %>%
  # filter(!hmid_h5 %in% unlist(filterr)) %>%
  mutate(hmid_h5 = substr(hmid_h5, start = 10, stop=nchar(hmid_h5)-8)) %>%
  # str_replace_all("_", "-") %>%
  na.omit()

rural_redcap_hmid <- rural_sensors %>%
  select(hh_id) %>% rename(hmid_redcap = hh_id) %>%
  distinct(hmid_redcap) %>%
  na.omit()

rural_hmid_check <- full_join(rural_zip_hmid , rural_unzip_hmid,
                              by = c("hmid_zip" = "hmid_unzip"),
                              keep = TRUE)
rural_hmid_check <- full_join(rural_hmid_check, rural_h5_hmid,
                              by = c("hmid_zip" = "hmid_h5"),
                              keep = TRUE)
rural_hmid_check <- full_join(rural_hmid_check, rural_redcap_hmid,
                              by = c("hmid_zip" = "hmid_redcap"),
                              keep = TRUE)

# now, extract only the households that appear in all 3 datasets.
rural_complete_hmid <- rural_hmid_check[complete.cases(rural_hmid_check), ]

# now, from the main rural_sensor data, extract households appearing on the list above.
rural_sensors_complete <- rural_sensors %>%
  filter(hh_id %in% rural_complete_hmid$hmid_zip)

cat("After this cleaning up, now there are", length(unique(rural_sensors_complete$hh_id)),
```

After this cleaning up, now there are 47 households in the rural area.

```r
n_rural_household_members_complete <- rural_sensors_complete %>%
  select(rec_id, hh_id) %>%
  group_by(hh_id) %>%
  summarise(n_hhid_complete = n())

rural_hh_size <- n_raw_rural_household_members %>%
  left_join( n_rural_household_members_complete, by = "hh_id")
```

```r
# How many individuals per household?
cat("There are still 2 households with only 1 member.")
```

```
There are still 2 households with only 1 member.
```

```r
# write.xlsx(rural_hh_size,
#            file =  "../data/sensor_metadata/mozambique_hh_size_check.xlsx",
#            sheetName = "rural_hhsize_check",
#            append = FALSE, showNA = FALSE)
#
# write.xlsx(rural_hmid_check,
#            file =  "../data/sensor_metadata/mozambique_hh_size_check.xlsx",
#            sheetName = "rural_hmid_check",
#            append = TRUE, showNA = FALSE)
############


# save rural participant metadata
write.csv(rural_sensors_complete,"../data/sensor_metadata/mozambique_aim2_rural_participan
```

Urban sensor data

```
We have 9 records that have sensor id == 0.
```

```
there now are 273 individuals with sensors
```

```
There are 71 households listed in the
    urban area, and 7 have only 1 member listed.
```

```
visual check reveals 12 missing sensor ids (9)
    or missing hwid (12). So, we drop these.
```

```
So we now have 261 complete records from the urban site.
```

In total, there are 273 participants who were issued sensors from the urban site. Out of these, we have complete records for 261. This is after dropping records due to missing sensor_id and hwid.

Table 5 shows the characteristics of individuals with missing sensor IDS and hwids (n=67).

Table 5: Urban records missing sensor_id and hwid data

Variable	N = 12
isindex	2 (17%)
participant_sex	
Female	7 (58%)
Male	5 (42%)
participant_age	
<6mo	0 (0%)
6-11mo	1 (8%)
1-4y	2 (17%)
5-9y	0 (0%)
10-14y	0 (0%)
15-19y	0 (0%)
20-29y	8 (67%)
30-39y	0 (0%)
40-59y	1 (8%)
60+y	0 (0%)

Checking household IDs in urban site

```
cat("There are", length(unique(urban_sensors$hh_id)), "households in the urban
    area. This means that there are individuals who migrated from one household
    and were now living in a new household, but their perm_id was not updated to
    reflect the move to the new household.")
```

```
There are 70 households in the urban
    area. This means that there are individuals who migrated from one household
    and were now living in a new household, but their perm_id was not updated to
    reflect the move to the new household.
```

```
n_urban_household_members <- urban_sensors %>%
  select(rec_id, hh_id) %>%
  group_by(hh_id) %>%
  summarise(n=n())
urban_hhsize_2 <- n_urban_household_members %>%
  filter(n < 2)
```

```r
cat("There are", nrow(urban_hhsize_2), "households with 1 member. There should
    not be households with only 1 member (11).")
```

```
There are 9 households with 1 member. There should
    not be households with only 1 member (11).
```

```r
# extract the file names in the raw folder.
unzip_path <- c("/Users/mck/Library/CloudStorage/OneDrive-EmoryUniversity/4.Data and Analy
zip_path <- c("/Users/mck/Library/CloudStorage/OneDrive-EmoryUniversity/4.Data and Analysi
h5_path <- c("/Users/mck/Library/CloudStorage/OneDrive-EmoryUniversity/4.Data and Analysis


urban_unzip_hmid <- as.data.frame(list.files(path = paste(unzip_path))) %>%
  rename(hmid_unzip = names(.[1])) %>%
  # filter(!hmid_unzip %in% unlist(filterr)) %>%
  mutate(hmid_unzip = substr(hmid_unzip, start = 10, stop=nchar(hmid_unzip))) %>%
  # str_replace_all("_", "-") %>%
  na.omit()

urban_zip_hmid <- as.data.frame(list.files(path = paste(zip_path))) %>%
  rename(hmid_zip = names(.[1])) %>%
  # filter(!hmid_zip %in% unlist(filterr)) %>%
  mutate(hmid_zip = substr(hmid_zip, start = 10, stop=nchar(hmid_zip)-25)) %>%
  # gsub("_", "-", .) %>%
  na.omit()

urban_h5_hmid <- as.data.frame(list.files(path = paste(h5_path))) %>%
  rename(hmid_h5 = names(.[1])) %>%
  # filter(!hmid_h5 %in% unlist(filterr)) %>%
  mutate(hmid_h5 = substr(hmid_h5, start = 10, stop=nchar(hmid_h5)-8)) %>%
  # str_replace_all("_", "-") %>%
  na.omit()

urban_redcap_hmid <- urban_sensors %>%
  select(hh_id) %>% rename(hmid_redcap = hh_id) %>%
  distinct(hmid_redcap) %>%
  na.omit()

urban_hmid_check <- full_join(urban_zip_hmid , urban_unzip_hmid,
                              by = c("hmid_zip" = "hmid_unzip"),
                              keep = TRUE)
```

```r
urban_hmid_check <- full_join(urban_hmid_check, urban_h5_hmid,
                              by = c("hmid_zip" = "hmid_h5"),
                              keep = TRUE)
urban_hmid_check <- full_join(urban_hmid_check, urban_redcap_hmid,
                              by = c("hmid_zip" = "hmid_redcap"),
                              keep = TRUE)

# now, extract only the households that appear in all 3 datasets.
urban_complete_hmid <- urban_hmid_check[complete.cases(urban_hmid_check), ]

# now, from the main urban_sensor data, extract households appearing on the list above.
urban_sensors_complete <- urban_sensors %>%
  filter(hh_id %in% urban_complete_hmid$hmid_zip)

cat("After this cleaning up, now there are", length(unique(urban_sensors_complete$hh_id)),
```

```
After this cleaning up, now there are 49 households in the urban area.
```

```r
n_urban_household_members_complete <- urban_sensors_complete %>%
  select(rec_id, hh_id) %>%
  group_by(hh_id) %>%
  summarise(n_hhid_complete = n())

urban_hh_size <- n_raw_urban_household_members %>%
  left_join( n_urban_household_members_complete, by = "hh_id")

# How many individuals per household?
cat("There are 6 households with 1 member.")
```

```
There are 6 households with 1 member.
```

```r
write.xlsx(urban_hh_size,
           file =  "../data/sensor_metadata/mozambique_urban_hh_size_check.xlsx",
           sheetName = "urban_hhsize_check",
           append = FALSE, showNA = FALSE)

write.xlsx(urban_hmid_check,
           file =  "../data/sensor_metadata/mozambique_urban_hh_size_check.xlsx",
           sheetName = "urban_hmid_check",
```

```
            append = TRUE, showNA = FALSE)
############


# save final urban household file
# # save urban participant metadata
write.csv(urban_sensors_complete,"../data/sensor_metadata/mozambique_aim2_urban_participan
```

Final dataset

```
final_participants <- rbind(rural_sensors_complete, urban_sensors_complete)

summary_aim2_final <- final_participants %>%
  select(study_site, isindex, participant_sex, participant_age, read_write,
         occupation, enrolled_school) %>%
  tbl_summary(by="study_site",
              statistic = list(all_continuous() ~ "{median ({IQR})"),
              percent="column",
              missing="ifany",
              digits = all_categorical() ~ 0) %>%
  bold_labels() %>%
  modify_header(label = "**Variable**") %>%
  modify_caption("")
summary_aim2_final
```

```
Table printed with `knitr::kable()`, not {gt}. Learn why at
http://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
To suppress this message, include `message = FALSE` in code chunk header.
```

Variable	**Rural**, N = 168	**Urban**, N = 190
isindex	35 (21%)	43 (23%)
participant_sex		
Female	106 (63%)	114 (60%)
Male	61 (37%)	75 (40%)
Unknown	1	1
participant_age		
<6mo	18 (11%)	15 (8%)
6-11mo	14 (8%)	20 (11%)
1-4y	21 (12%)	26 (14%)
5-9y	29 (17%)	21 (11%)

Variable	**Rural**, N = 168	**Urban**, N = 190
10-14y	24 (14%)	21 (11%)
15-19y	11 (7%)	13 (7%)
20-29y	23 (14%)	34 (18%)
30-39y	16 (10%)	20 (11%)
40-59y	8 (5%)	13 (7%)
60+y	4 (2%)	7 (4%)
read_write	68 (40%)	105 (55%)
occupation		
Child	32 (23%)	35 (22%)
Unemployed	37 (26%)	24 (15%)
Student	45 (32%)	48 (30%)
Homemaker	2 (1%)	9 (6%)
Casual laboror	3 (2%)	8 (5%)
Farmer	9 (6%)	0 (0%)
Fisherman	0 (0%)	0 (0%)
Business person	2 (1%)	17 (11%)
Office worker	4 (3%)	16 (10%)
Retired	0 (0%)	1 (1%)
Other	7 (5%)	0 (0%)
Unknown	27	32
enrolled_school	58 (36%)	49 (27%)
Unknown	7	7