Data Distribution

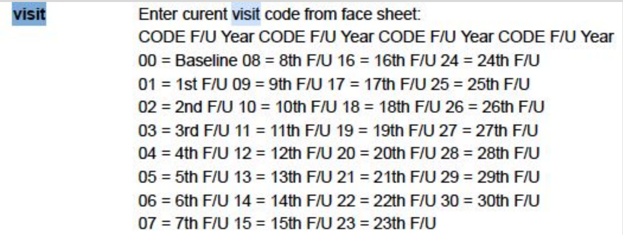
Yunus Shariff

Updates

- Fixed gender for subid 1736 [OHSU]
    - Positive for Mild Cognitive Impairment (+1 for Male in MCI positive subgroup)

- rush_additional.csv and rush_final_20210708.csv:
    - The 6 sub ids that are missing gender in the original mars.csv dataset do not exist in rush_final_20210708.csv (68 subids only)
        - The data is identical for these 68 subjects, with the exception of some additional columns in rush_final_20210708.csv
    - mars_additional.csv file is used to populate the gender and MCI status for 1791 (one of the six).
        - Remaining (5) could not be updated

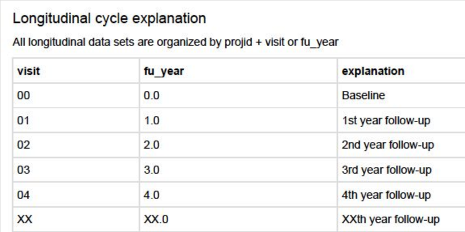
- Updated gender, MCI status for 1791 [RUSH]
    - Normal cognitive status (MCI positive gender distribution is unaltered)
    - Subid is still dropped due to **missing 'enroll_visit'**
        - *mars_additional.csv contains '**fu_year'** (follow up year) - is this the same as 'enroll_visit'?*



visit

Enter curent visit code from face sheet:

CODE F/U Year CODE F/U Year CODE F/U Year CODE F/U Year
00 = Baseline 08 = 8th F/U 16 = 16th F/U 24 = 24th F/U
01 = 1st F/U 09 = 9th F/U 17 = 17th F/U 25 = 25th F/U
02 = 2nd F/U 10 = 10th F/U 18 = 18th F/U 26 = 26th F/U
03 = 3rd F/U 11 = 11th F/U 19 = 19th F/U 27 = 27th F/U
04 = 4th F/U 12 = 12th F/U 20 = 20th F/U 28 = 28th F/U
05 = 5th F/U 13 = 13th F/U 21 = 21th F/U 29 = 29th F/U
06 = 6th F/U 14 = 14th F/U 22 = 22th F/U 30 = 30th F/U
07 = 7th F/U 15 = 15th F/U 23 = 23th F/U



Longitudinal cycle explanation

All longitudinal data sets are organized by projid + visit or fu_year

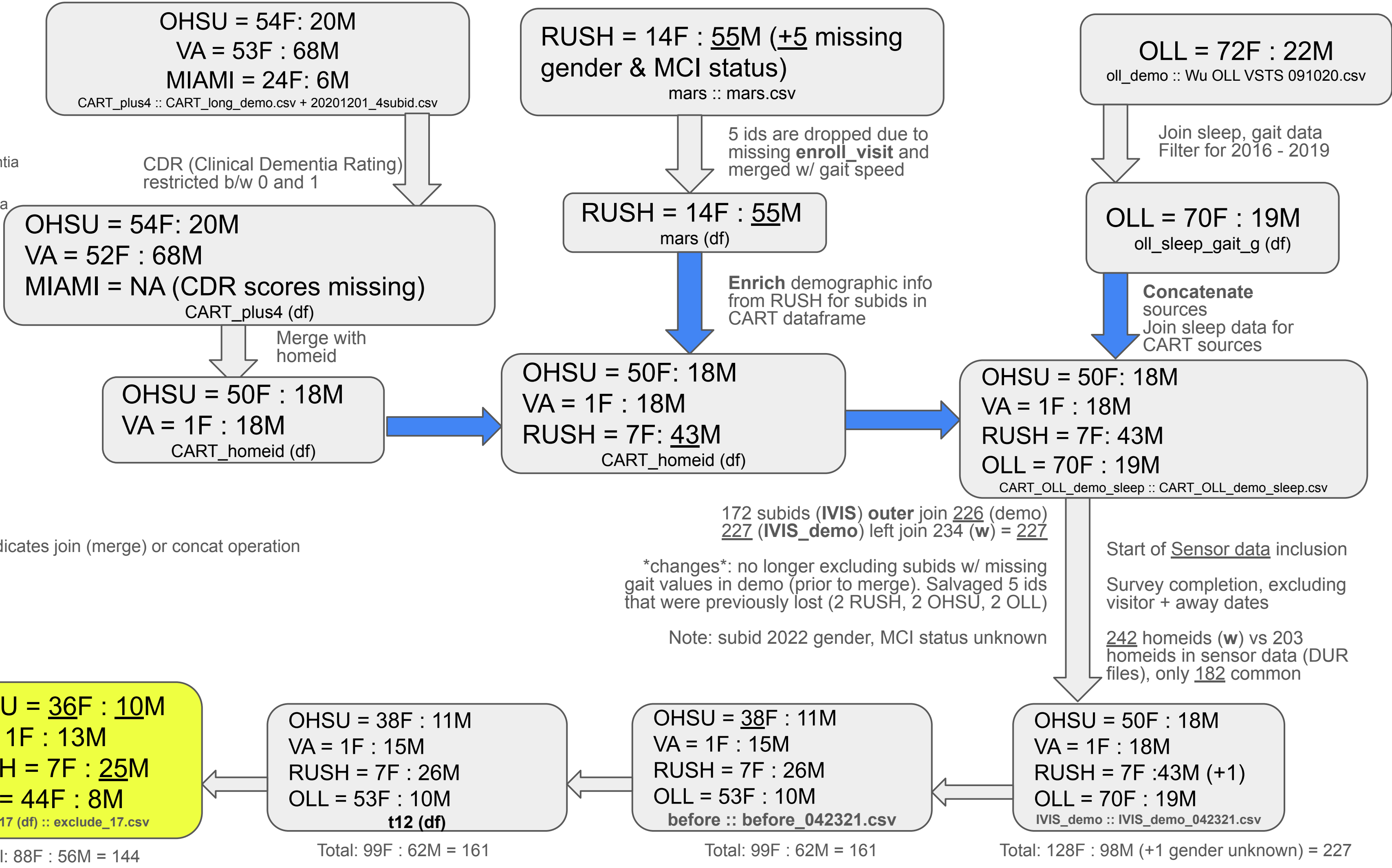
visit	fu_year	explanation
00	0.0	Baseline
01	1.0	1st year follow-up
02	2.0	2nd year follow-up
03	3.0	3rd year follow-up
04	4.0	4th year follow-up
XX	XX.0	XXth year follow-up

- When updating 'gait' for these 69 ids, the gait file for RUSH only has gait recorded for proj id relevant to 67 ids
    - Original set of missing gender (6 + 1) : [1791, 1885, 1886, 2022, 2028, 2179] + [2070]
    - 2 subids, recently recovered subid 1791 along w/ 2070 are missing gait, in RUSH df before merge w/ OHSU
    - Due to the missing gait info, 1791 gets dropped at the process sensor stage

- *Since only common ids (pre-existing) in OHSU source are copied over from RUSH…*
    - *Out of 74 RUSH subids, how many in common w/ OHSU?* ***51***
        - *Why 51? The Single Resident Home ID file only has home info for these ids*
    - *How many of the RUSH ids are updated in RUSH?* ***50*** *(previously 49)*
        - Subid **2022** is the 51st ID
        - excluded due to unknown ['gender', 'mci', **'enroll_visit'** or 'visit' or 'fu_year']

Gender distribution

FYI: CDR range
0 = Normal
0.5 = Very Mild Dementia
1 = Mild Dementia
2 = Moderate Dementia
3 = Severe Dementia



OHSU = 54F: 20M
VA = 53F : 68M
MIAMI = 24F: 6M
CART_plus4 :: CART_long_demo.csv + 20201201_4subid.csv

CDR (Clinical Dementia Rating) restricted b/w 0 and 1

OHSU = 54F: 20M
VA = 52F : 68M
MIAMI = NA (CDR scores missing)
CART_plus4 (df)

Merge with homeid

OHSU = 50F : 18M
VA = 1F : 18M
CART_homeid (df)

RUSH = 14F : 55M (+5 missing gender & MCI status)
mars :: mars.csv

5 ids are dropped due to missing **enroll_visit** and merged w/ gait speed

RUSH = 14F : 55M
mars (df)

Enrich demographic info from RUSH for subids in CART dataframe

OHSU = 50F: 18M
VA = 1F : 18M
RUSH = 7F: 43M
CART_homeid (df)

OLL = 72F : 22M
oll_demo :: Wu OLL VSTS 091020.csv

Join sleep, gait data
Filter for 2016 - 2019

OLL = 70F : 19M
oll_sleep_gait_g (df)

Concatenate sources
Join sleep data for CART sources

OHSU = 50F: 18M
VA = 1F : 18M
RUSH = 7F: 43M
OLL = 70F : 19M
CART_OLL_demo_sleep :: CART_OLL_demo_sleep.csv

172 subids (**IVIS**) **outer** join 226 (demo)
227 (**IVIS_demo**) left join 234 (**w**) = 227

changes: no longer excluding subids w/ missing gait values in demo (prior to merge). Salvaged 5 ids that were previously lost (2 RUSH, 2 OHSU, 2 OLL)

Note: subid 2022 gender, MCI status unknown

Start of Sensor data inclusion

Survey completion, excluding visitor + away dates

242 homeids (**w**) vs 203 homeids in sensor data (DUR files), only 182 common

OHSU = 50F : 18M
VA = 1F : 18M
RUSH = 7F :43M (+1)
OLL = 70F : 19M
IVIS_demo :: IVIS_demo_042321.csv

Total: 128F : 98M (+1 gender unknown) = 227

OHSU = 38F : 11M
VA = 1F : 15M
RUSH = 7F : 26M
OLL = 53F : 10M
before :: before_042321.csv

Total: 99F : 62M = 161

OHSU = 38F : 11M
VA = 1F : 15M
RUSH = 7F : 26M
OLL = 53F : 10M
t12 (df)

Total: 99F : 62M = 161

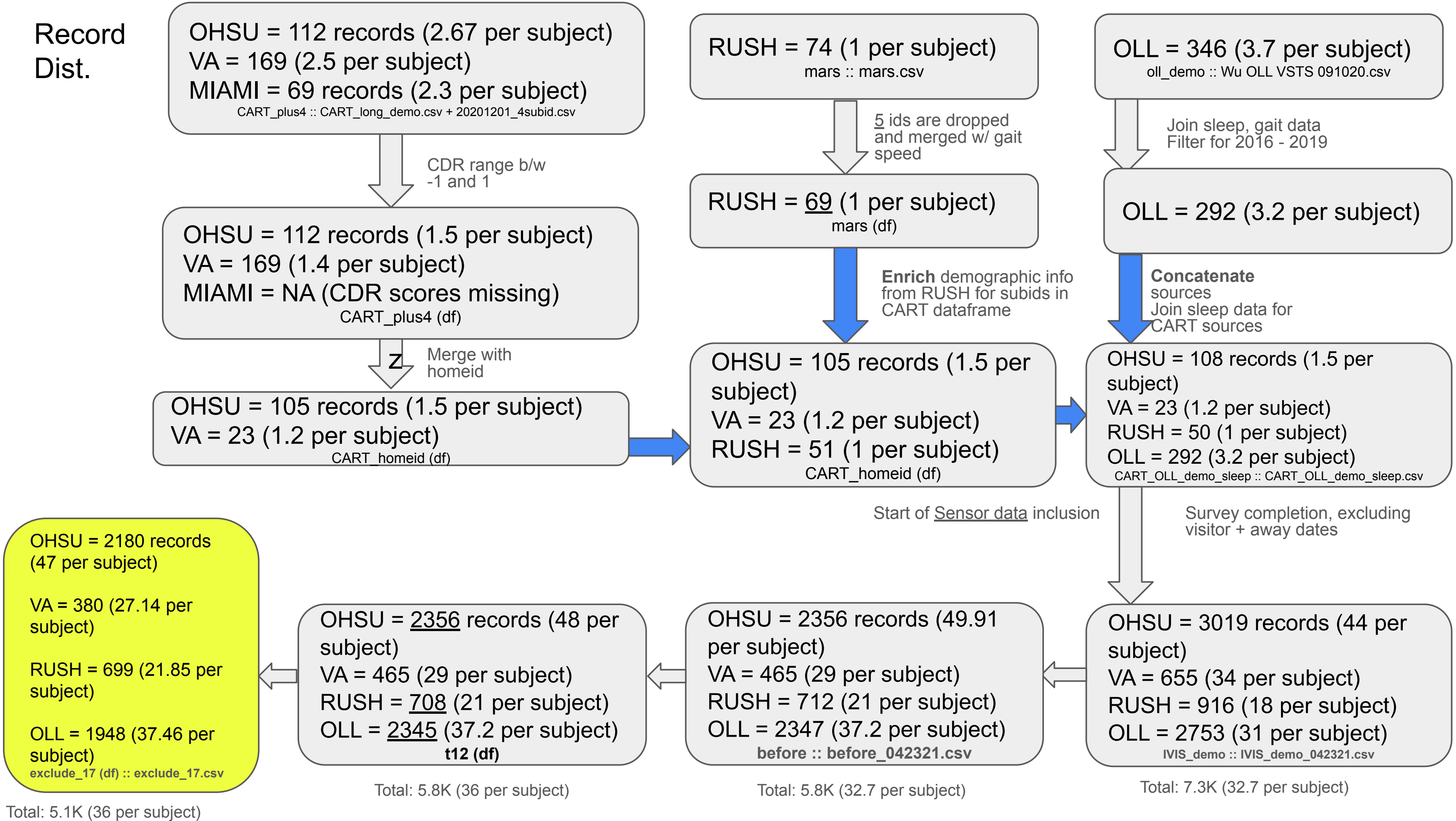
OHSU = 36F : 10M
VA = 1F : 13M
RUSH = 7F : 25M
OLL = 44F : 8M
exclude_17 (df) :: exclude_17.csv

Total: 88F : 56M = 144

Note:

➡ - indicates join (merge) or concat operation

Record Dist.



OHSU = 112 records (2.67 per subject)
VA = 169 (2.5 per subject)
MIAMI = 69 records (2.3 per subject)
CART_plus4 :: CART_long_demo.csv + 20201201_4subid.csv

CDR range b/w -1 and 1

OHSU = 112 records (1.5 per subject)
VA = 169 (1.4 per subject)
MIAMI = NA (CDR scores missing)
CART_plus4 (df)

Z Merge with homeid

OHSU = 105 records (1.5 per subject)
VA = 23 (1.2 per subject)
CART_homeid (df)

RUSH = 74 (1 per subject)
mars :: mars.csv

5 ids are dropped and merged w/ gait speed

RUSH = 69 (1 per subject)
mars (df)

Enrich demographic info from RUSH for subids in CART dataframe

OHSU = 105 records (1.5 per subject)
VA = 23 (1.2 per subject)
RUSH = 51 (1 per subject)
CART_homeid (df)

OLL = 346 (3.7 per subject)
oll_demo :: Wu OLL VSTS 091020.csv

Join sleep, gait data
Filter for 2016 - 2019

OLL = 292 (3.2 per subject)

Concatenate sources
Join sleep data for CART sources

OHSU = 108 records (1.5 per subject)
VA = 23 (1.2 per subject)
RUSH = 50 (1 per subject)
OLL = 292 (3.2 per subject)
CART_OLL_demo_sleep :: CART_OLL_demo_sleep.csv

Start of Sensor data inclusion

Survey completion, excluding visitor + away dates

OHSU = 3019 records (44 per subject)
VA = 655 (34 per subject)
RUSH = 916 (18 per subject)
OLL = 2753 (31 per subject)
IVIS_demo :: IVIS_demo_042321.csv

Total: 7.3K (32.7 per subject)

OHSU = 2356 records (49.91 per subject)
VA = 465 (29 per subject)
RUSH = 712 (21 per subject)
OLL = 2347 (37.2 per subject)
before :: before_042321.csv

Total: 5.8K (32.7 per subject)

OHSU = 2356 records (48 per subject)
VA = 465 (29 per subject)
RUSH = 708 (21 per subject)
OLL = 2345 (37.2 per subject)
t12 (df)

Total: 5.8K (36 per subject)

OHSU = 2180 records (47 per subject)

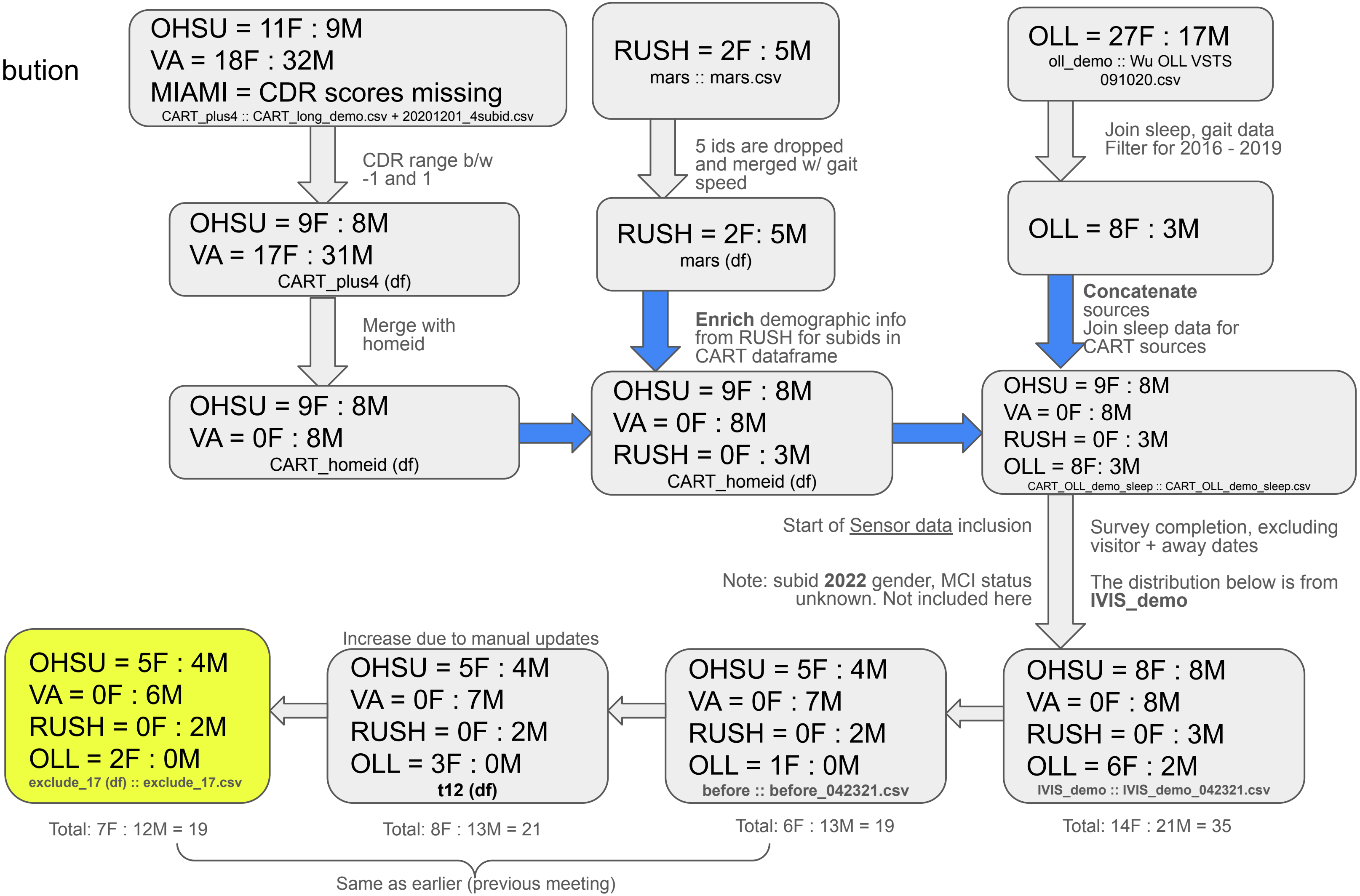
VA = 380 (27.14 per subject)

RUSH = 699 (21.85 per subject)

OLL = 1948 (37.46 per subject)
exclude_17 (df) :: exclude_17.csv

Total: 5.1K (36 per subject)

MCI distribution



OHSU = 11F : 9M
VA = 18F : 32M
MIAMI = CDR scores missing
CART_plus4 :: CART_long_demo.csv + 20201201_4subid.csv

CDR range b/w -1 and 1

OHSU = 9F : 8M
VA = 17F : 31M
CART_plus4 (df)

Merge with homeid

OHSU = 9F : 8M
VA = 0F : 8M
CART_homeid (df)

RUSH = 2F : 5M
mars :: mars.csv

5 ids are dropped and merged w/ gait speed

RUSH = 2F: 5M
mars (df)

Enrich demographic info from RUSH for subids in CART dataframe

OHSU = 9F : 8M
VA = 0F : 8M
RUSH = 0F : 3M
CART_homeid (df)

OLL = 27F : 17M
oll_demo :: Wu OLL VSTS 091020.csv

Join sleep, gait data
Filter for 2016 - 2019

OLL = 8F : 3M

Concatenate sources
Join sleep data for CART sources

OHSU = 9F : 8M
VA = 0F : 8M
RUSH = 0F : 3M
OLL = 8F: 3M
CART_OLL_demo_sleep :: CART_OLL_demo_sleep.csv

Start of Sensor data inclusion

Note: subid **2022** gender, MCI status unknown. Not included here

Survey completion, excluding visitor + away dates

The distribution below is from **IVIS_demo**

OHSU = 8F : 8M
VA = 0F : 8M
RUSH = 0F : 3M
OLL = 6F : 2M
IVIS_demo :: IVIS_demo_042321.csv

Total: 14F : 21M = 35

OHSU = 5F : 4M
VA = 0F : 7M
RUSH = 0F : 2M
OLL = 1F : 0M
before :: before_042321.csv

Total: 6F : 13M = 19

Increase due to manual updates

OHSU = 5F : 4M
VA = 0F : 7M
RUSH = 0F : 2M
OLL = 3F : 0M
t12 (df)

Total: 8F : 13M = 21

OHSU = 5F : 4M
VA = 0F : 6M
RUSH = 0F : 2M
OLL = 2F : 0M
exclude_17 (df) :: exclude_17.csv

Total: 7F : 12M = 19

Same as earlier (previous meeting)

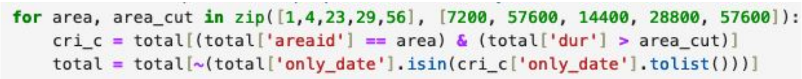
Clarifications pertaining to ids with missing gait in slides *

- Demo Clean Script:
    - 2022 is excluded from RUSH set due to missing enroll_visit
        - 2022: gender, MCI status unknown
    - 1791: gender, MCI status extracted from rush_additional but gait is still missing
        - rush gait file does not have ‘gait’ for this subject’s projid
    - These subids are copied over into CART source after joining with home id file
        - same fields are missing [‘mci’,’gender’]
- Process Sensor Script:
    - These subids get filtered out due to ‘livewhere’ being missing in subsequent steps i.e., ‘before’ df processing

Sensor Data inclusion/cleansing steps

- **w (df):** Filter for subjects:
  - Include who have completed surveys
  - Exclude dates when they had visitors or reported away to avoid external influences in routine/activities
  - Day of return/departure to also be excluded

- **IVIS:** Considering data involving 4 area ids and movement is detected
  - Area ids are: 1,4,23,29 for bedroom, bathroom, kitchen and living room
  - Study is only interested in transitions between these 4 areas and duration of transitions is >= 20 seconds
  - Exclude records where with <=10 transitions per day and excessive time is spent detected within these areas



```
for area, area_cut in zip([1,4,23,29,56], [7200, 57600, 14400, 28800, 57600]):
    cri_c = total[(total['areaid'] == area) & (total['dur'] > area_cut)]
    total = total[~(total['only_date'].isin(cri_c['only_date'].tolist()))]
```

- **IVIS_demo:** Merge this with demographic and sleep data (demo), excluding any with missing gait measurements and survey data (w)

- **before:**
  - records where IS is between 0 and 1 &
  - Only precovid data (before March 2020) &
  - Number of transitions is within 3 std &
  - Consider who did not require any assistance in grooming, medication mgmt (as indicated in surveys) &
  - Living situation is retirement community and home/apt (people who can manage daily independently) &
  - Not belonging to homeids: 411, 1468, 1527, 1619, 1913 due to missing scores and/or living situation &
  - Subjects with atleast 2 weekly surveys

- **t12:**
  - Manual updates (age, race, out_time_week,mci status)
  - Left join with demo based on subid and date to enrich with sleep and demographic data (currently only contains test scores, surveys and activities)
  - Drop duplicates: if surveys from subsequent weeks are filled 1 day apart, discard them as they would be identical

- **exclude_17:**
  - Discard **17** subids excluded due to incorrect sensor installation, door left open all night and other poor data quality issues
  - OHSU = 3 [2F:1M], VA = 2[M], RUSH = 1[M], OLL = 11 [9F:2M]; VA = 1805 +ve for MCI, OLL = 737[F] MCI

Step and sleep + demographic

Daily

- Step: 195 OHSU/VA + 70 RUSH = 265
- Sleep: 193 OHSU/VA + 70 = 263
    - [1652 (VA), 2199 (OHSU)] diff wrt step data

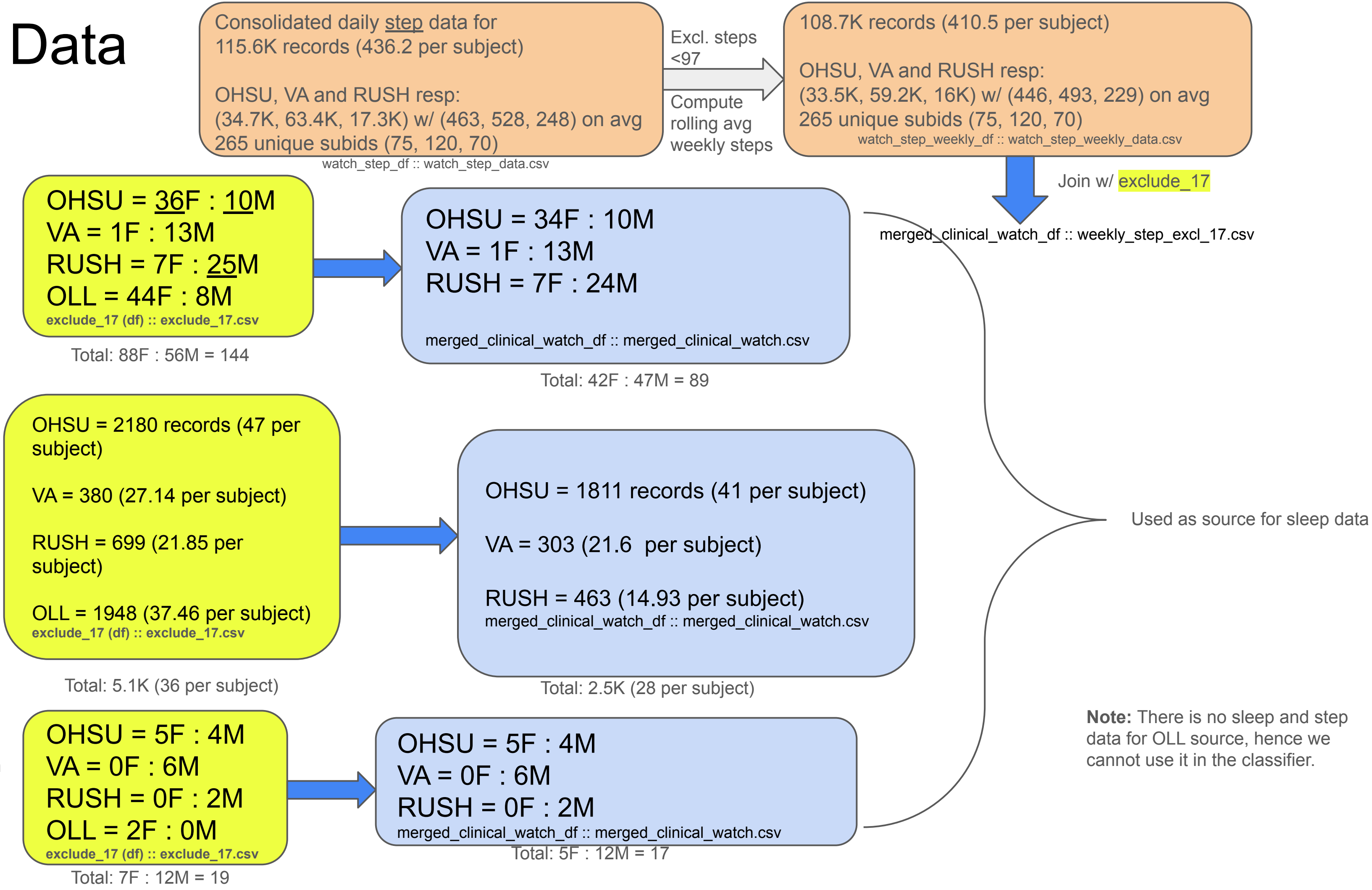
- Source is missing [2253, 2254] from VA
    - Step: 263 can be enriched w/ demographic
    - Sleep: 261 can be enriched w/ demographic
    - Step & Sleep: 261

Weekly (only weekly data is exported to csv)

- Step: 195 OHSU/VA + 70 RUSH = 265
- Sleep: 193 OHSU/VA + 70 = 255 (discard sleep duration > 3*std)
    - [1642, 1652, 1782, 1812, 1879, 1910, 2015, 2199, 2242, 2259] diff wrt step data

- Source is missing [2253, 2254] from VA
    - Step: 263 can be enriched w/ demographic
    - Sleep: 253 can be enriched w/ demographic
    - Step & Sleep: 253
- Left join is used so file will still contain 265 and 255 unique subids respectively
    - Step + sleep + demo csv file contains also has 265 subids

Step Data



Consolidated daily step data for
115.6K records (436.2 per subject)

OHSU, VA and RUSH resp:
(34.7K, 63.4K, 17.3K) w/ (463, 528, 248) on avg
265 unique subids (75, 120, 70)

watch_step_df :: watch_step_data.csv

Excl. steps <97

Compute rolling avg weekly steps

108.7K records (410.5 per subject)

OHSU, VA and RUSH resp:
(33.5K, 59.2K, 16K) w/ (446, 493, 229) on avg
265 unique subids (75, 120, 70)

watch_step_weekly_df :: watch_step_weekly_data.csv

Join w/ exclude_17

merged_clinical_watch_df :: weekly_step_excl_17.csv

Gender distribution

OHSU = 36F : 10M
VA = 1F : 13M
RUSH = 7F : 25M
OLL = 44F : 8M

exclude_17 (df) :: exclude_17.csv

Total: 88F : 56M = 144

OHSU = 34F : 10M
VA = 1F : 13M
RUSH = 7F : 24M

merged_clinical_watch_df :: merged_clinical_watch.csv

Total: 42F : 47M = 89

Record distribution

OHSU = 2180 records (47 per subject)

VA = 380 (27.14 per subject)

RUSH = 699 (21.85 per subject)

OLL = 1948 (37.46 per subject)

exclude_17 (df) :: exclude_17.csv

Total: 5.1K (36 per subject)

OHSU = 1811 records (41 per subject)

VA = 303 (21.6 per subject)

RUSH = 463 (14.93 per subject)

merged_clinical_watch_df :: merged_clinical_watch.csv

Total: 2.5K (28 per subject)

Used as source for sleep data

MCI distribution

OHSU = 5F : 4M
VA = 0F : 6M
RUSH = 0F : 2M
OLL = 2F : 0M

exclude_17 (df) :: exclude_17.csv

Total: 7F : 12M = 19

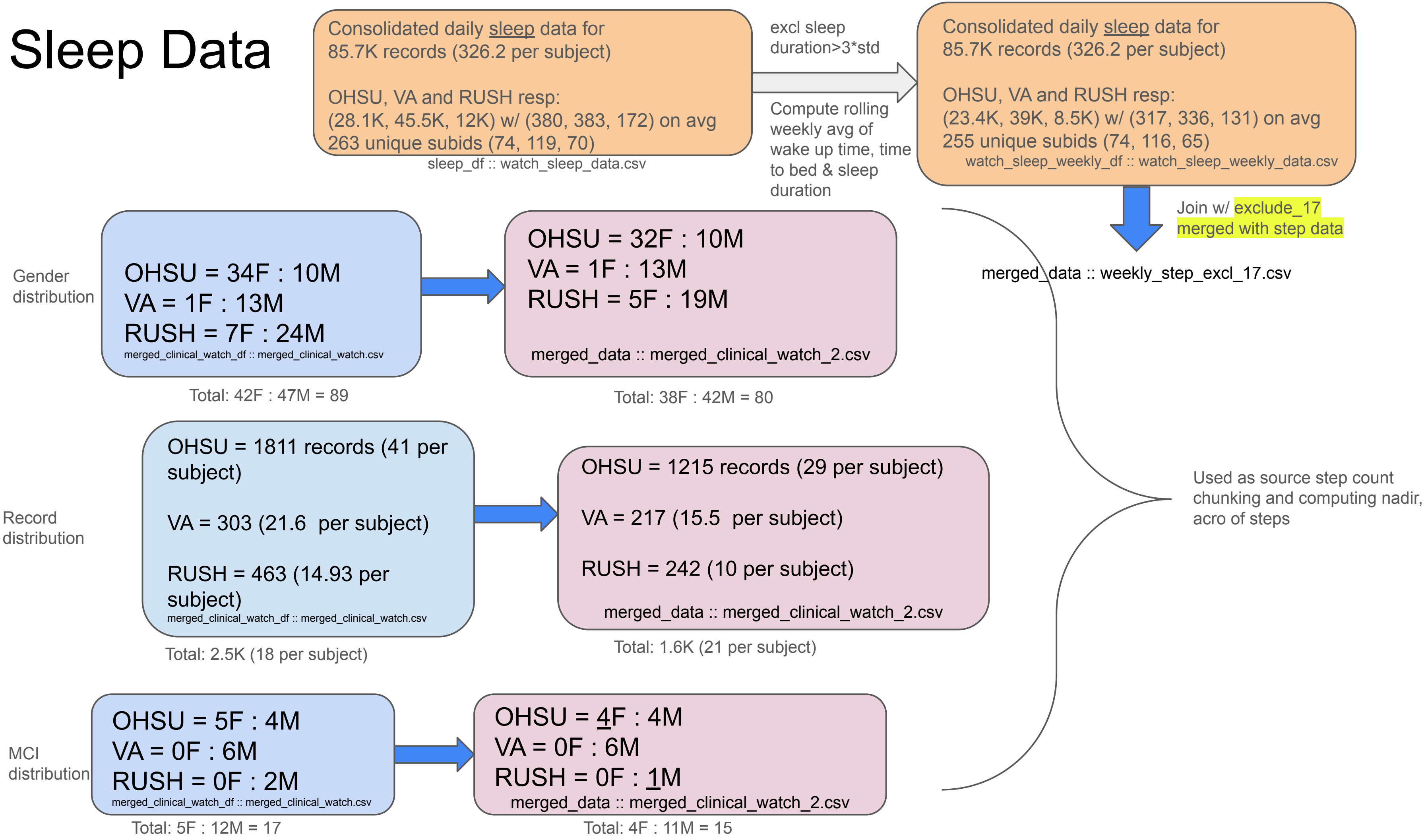
OHSU = 5F : 4M
VA = 0F : 6M
RUSH = 0F : 2M

merged_clinical_watch_df :: merged_clinical_watch.csv

Total: 5F : 12M = 17

Note: There is no sleep and step data for OLL source, hence we cannot use it in the classifier.

Sleep Data



Consolidated daily sleep data for
85.7K records (326.2 per subject)

OHSU, VA and RUSH resp:
(28.1K, 45.5K, 12K) w/ (380, 383, 172) on avg
263 unique subids (74, 119, 70)

sleep_df :: watch_sleep_data.csv

excl sleep
duration>3*std

Compute rolling
weekly avg of
wake up time, time
to bed & sleep
duration

Consolidated daily sleep data for
85.7K records (326.2 per subject)

OHSU, VA and RUSH resp:
(23.4K, 39K, 8.5K) w/ (317, 336, 131) on avg
255 unique subids (74, 116, 65)
watch_sleep_weekly_df :: watch_sleep_weekly_data.csv

Join w/ exclude_17
merged with step data

merged_data :: weekly_step_excl_17.csv

Gender
distribution

OHSU = 34F : 10M
VA = 1F : 13M
RUSH = 7F : 24M
merged_clinical_watch_df :: merged_clinical_watch.csv

Total: 42F : 47M = 89

OHSU = 32F : 10M
VA = 1F : 13M
RUSH = 5F : 19M

merged_data :: merged_clinical_watch_2.csv

Total: 38F : 42M = 80

Record
distribution

OHSU = 1811 records (41 per subject)

VA = 303 (21.6 per subject)

RUSH = 463 (14.93 per subject)
merged_clinical_watch_df :: merged_clinical_watch.csv

Total: 2.5K (18 per subject)

OHSU = 1215 records (29 per subject)

VA = 217 (15.5 per subject)

RUSH = 242 (10 per subject)

merged_data :: merged_clinical_watch_2.csv

Total: 1.6K (21 per subject)

MCI
distribution

OHSU = 5F : 4M
VA = 0F : 6M
RUSH = 0F : 2M
merged_clinical_watch_df :: merged_clinical_watch.csv

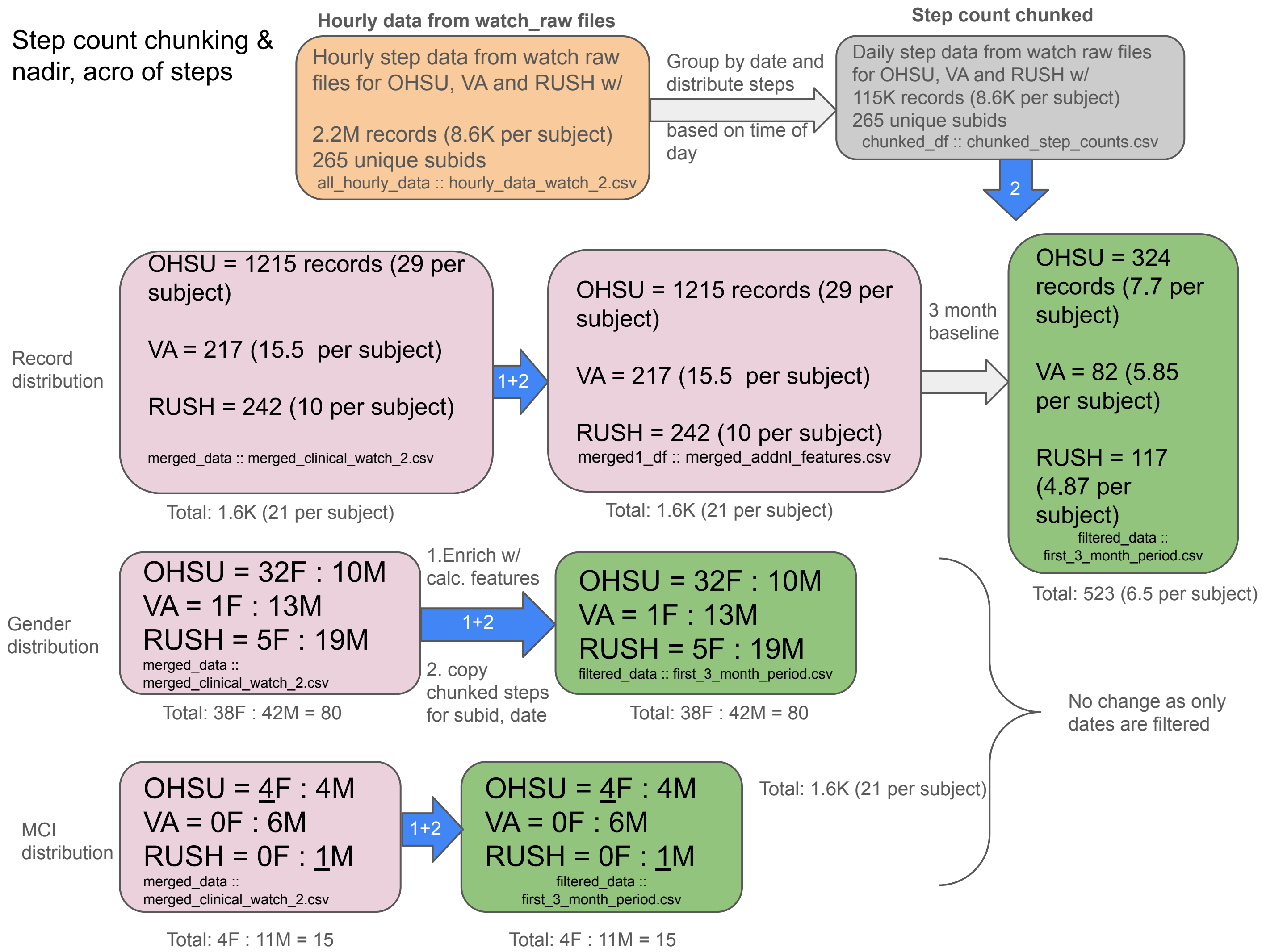
Total: 5F : 12M = 17

OHSU = 4F : 4M
VA = 0F : 6M
RUSH = 0F : 1M
merged_data :: merged_clinical_watch_2.csv

Total: 4F : 11M = 15

Used as source step count
chunking and computing nadir,
acro of steps

Step count chunking & nadir, acro of steps



Hourly data from watch_raw files

Hourly step data from watch raw files for OHSU, VA and RUSH w/

2.2M records (8.6K per subject)
265 unique subids
all_hourly_data :: hourly_data_watch_2.csv

Group by date and distribute steps based on time of day

Step count chunked

Daily step data from watch raw files for OHSU, VA and RUSH w/
115K records (8.6K per subject)
265 unique subids
chunked_df :: chunked_step_counts.csv

Record distribution

OHSU = 1215 records (29 per subject)

VA = 217 (15.5 per subject)

RUSH = 242 (10 per subject)

merged_data :: merged_clinical_watch_2.csv

Total: 1.6K (21 per subject)

OHSU = 1215 records (29 per subject)

VA = 217 (15.5 per subject)

RUSH = 242 (10 per subject)
merged1_df :: merged_addnl_features.csv

Total: 1.6K (21 per subject)

3 month baseline

OHSU = 324 records (7.7 per subject)

VA = 82 (5.85 per subject)

RUSH = 117 (4.87 per subject)
filtered_data :: first_3_month_period.csv

Total: 523 (6.5 per subject)

Gender distribution

OHSU = 32F : 10M
VA = 1F : 13M
RUSH = 5F : 19M
merged_data :: merged_clinical_watch_2.csv

Total: 38F : 42M = 80

1.Enrich w/ calc. features

2. copy chunked steps for subid, date

OHSU = 32F : 10M
VA = 1F : 13M
RUSH = 5F : 19M
filtered_data :: first_3_month_period.csv

Total: 38F : 42M = 80

MCI distribution

OHSU = 4F : 4M
VA = 0F : 6M
RUSH = 0F : 1M
merged_data :: merged_clinical_watch_2.csv

Total: 4F : 11M = 15

OHSU = 4F : 4M
VA = 0F : 6M
RUSH = 0F : 1M
filtered_data :: first_3_month_period.csv

Total: 4F : 11M = 15

Total: 1.6K (21 per subject)

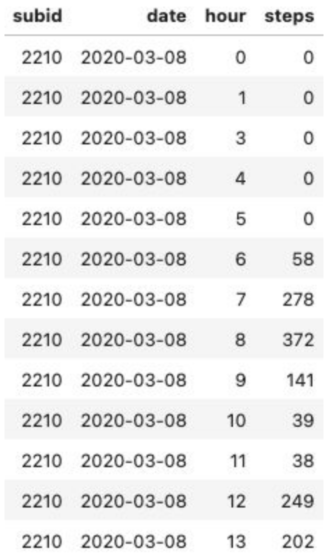
No change as only dates are filtered

Calculated features

For each subid and report_date in step+sleep+demo df,

Read hourly steps from watch_raw files for OHSU, VA and RUSH

Compute acro, nadir, amp, IV, IS and total steps on weekly basis.
calculated_features_df :: calculated_features.csv

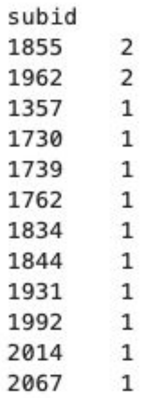


subid	date	hour	steps
2210	2020-03-08	0	0
2210	2020-03-08	1	0
2210	2020-03-08	3	0
2210	2020-03-08	4	0
2210	2020-03-08	5	0
2210	2020-03-08	6	58
2210	2020-03-08	7	278
2210	2020-03-08	8	372
2210	2020-03-08	9	141
2210	2020-03-08	10	39
2210	2020-03-08	11	38
2210	2020-03-08	12	249
2210	2020-03-08	13	202

Fig: Hourly data from watch_raw files

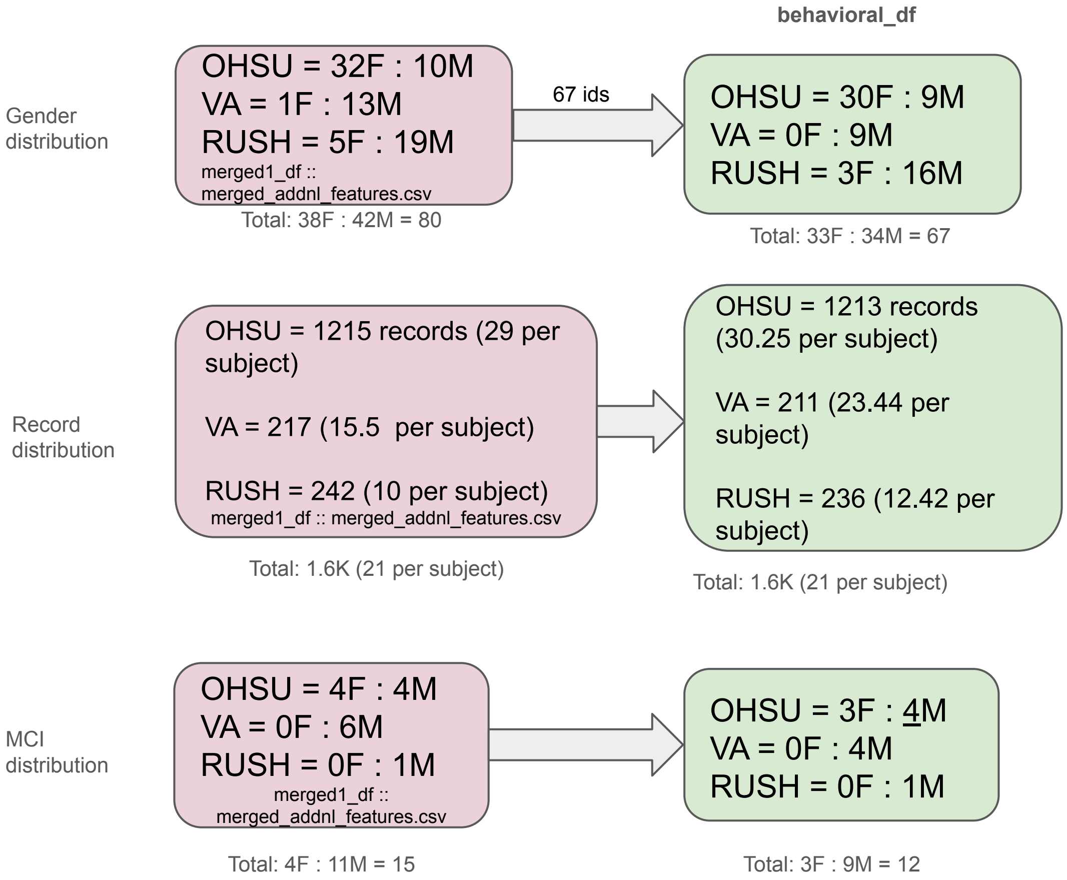
Exclusion criteria for trend analysis

- Data is not restricted to 3 months
- Subjects with less than 97 steps (in place when processing step data)
- Subjects with less than 3 records (12)



```
subid
1855    2
1962    2
1357    1
1730    1
1739    1
1762    1
1834    1
1844    1
1931    1
1992    1
2014    1
2067    1
```

Fig: (12) Subjects with less than 3 records



behavioral_df

Gender distribution

OHSU = 32F : 10M
VA = 1F : 13M
RUSH = 5F : 19M
merged1_df :: merged_addnl_features.csv

Total: 38F : 42M = 80

67 ids

OHSU = 30F : 9M
VA = 0F : 9M
RUSH = 3F : 16M

Total: 33F : 34M = 67

Record distribution

OHSU = 1215 records (29 per subject)

VA = 217 (15.5 per subject)

RUSH = 242 (10 per subject)
merged1_df :: merged_addnl_features.csv

Total: 1.6K (21 per subject)

OHSU = 1213 records (30.25 per subject)

VA = 211 (23.44 per subject)

RUSH = 236 (12.42 per subject)

Total: 1.6K (21 per subject)

MCI distribution

OHSU = 4F : 4M
VA = 0F : 6M
RUSH = 0F : 1M
merged1_df :: merged_addnl_features.csv

Total: 4F : 11M = 15

OHSU = 3F : 4M
VA = 0F : 4M
RUSH = 0F : 1M

Total: 3F : 9M = 12