Using R to analyse step count
Common pitfalls and how to get over them

Rladies Brisbane, 31st Mar 2021

Ms Huong Ly Tong PhD(c) MRes BHealth
PhD Researcher, Westmead Applied Research Centre
GitHub & Twitter: @lytong22

THE UNIVERSITY OF SYDNEY

Outline

- Background about the project and the step dataset
- Set up RStudio and GitHub
- Problem 1: Missing data
- Problem 2: Analyse longitudinal data
- Problem 3: Too few steps in a day?

Outline

Background about the project and the step dataset

Set up RStudio and GitHub

Problem 1: Missing data

Problem 2: Analyse longitudinal data

Problem 3: Too few steps in a day?

Background

Physical activity is important

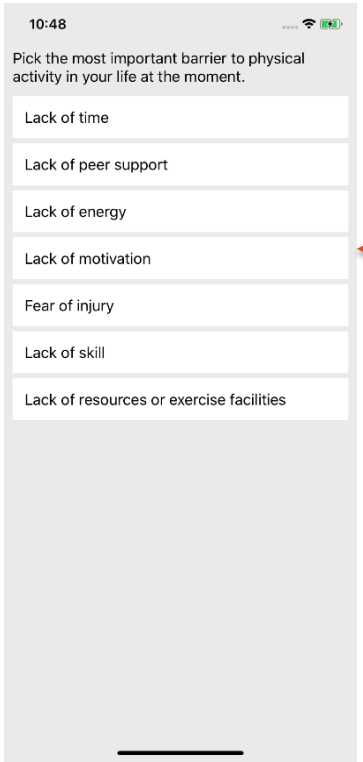
BUT… people **struggle** to stay active.

mobile applications (apps)

potential solution to facilitate physical activity

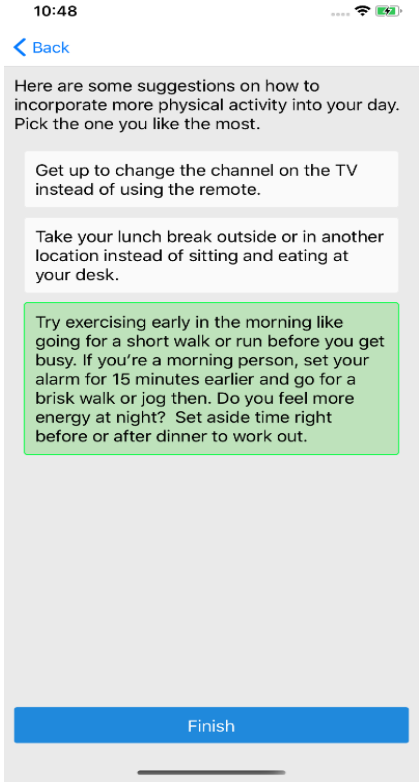
Background

HOW OUR BE.WELL APP ADDRESSES THESE GAPS



The personalised activity suggestion

Include users in the loop

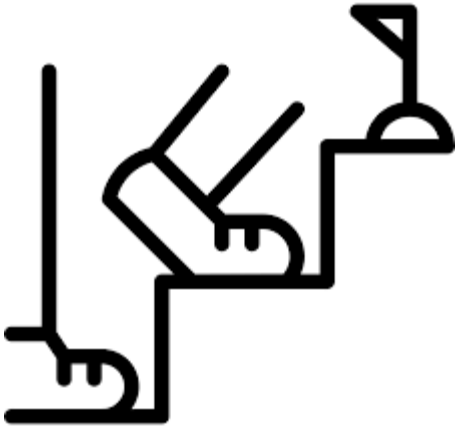


Background

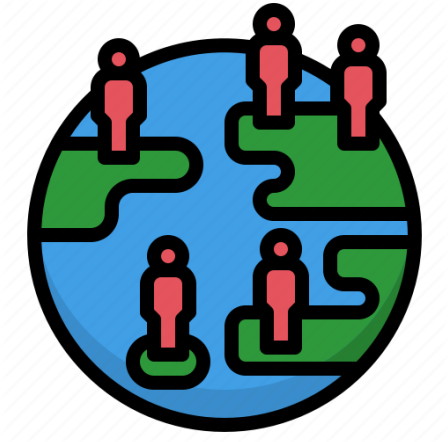
AIMS OF OUR STUDY

Test the impact of the be.well app on physical activity (i.e. daily step count)

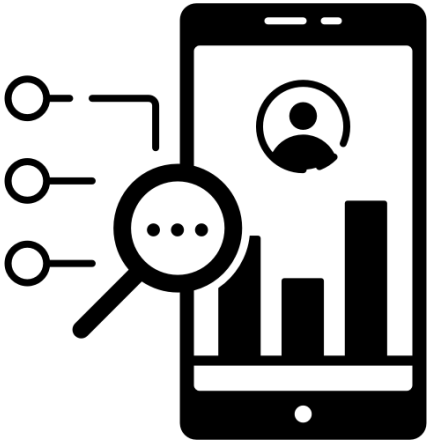
Dataset – 23 participants



- **Physical activity records**
  daily step counts



- **Demographic information**
  age, gender, weight, height

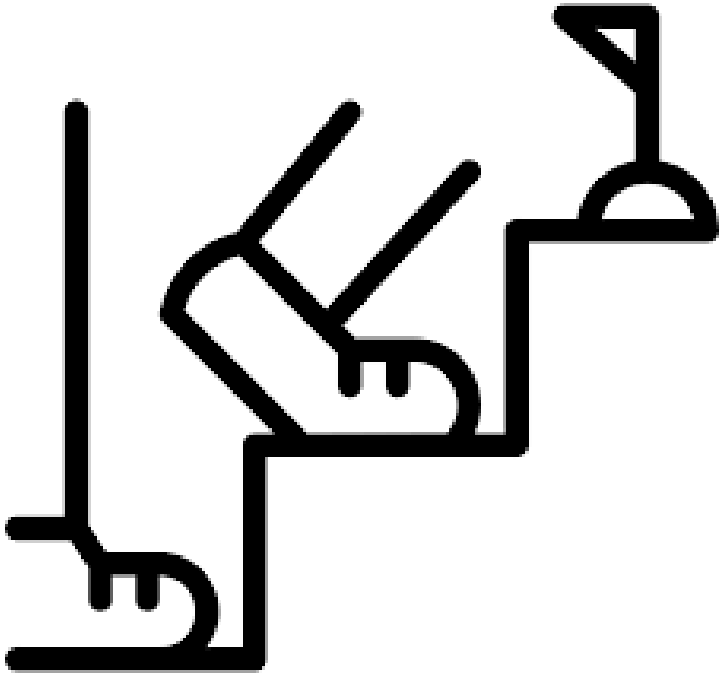


- **Additional information**
  phone model,
  amount of time participants spent on the phone in a day

Dataset – 23 participants

Daily step count

- Collected via HealthKit database on iPhone
- One-month pre-intervention
- Two-month during the intervention period



Outline

Background about the project and the step dataset

Set up RStudio and GitHub

Problem 1: Missing data

Problem 2: Analyse longitudinal data

Problem 3: Too few steps in a day?

Set up RStudio and GitHub

DEMO?

Chapter 12 Connect RStudio to Git and GitHub | Happy Git and GitHub for the useR (happygitwithr.com)

Outline

- Background about the project and the step dataset
- Set up RStudio and GitHub
- **Problem 1: Missing data**
- Problem 2: Analyse longitudinal data
- Problem 3: Too few steps in a day?

Problem: Missing data

Date	Steps
01/04/2020	529
02/04/2020	639
03/04/2020	756
04/04/2020	?

Example: steps data for a participant.

Question: how can we deal with missing data?

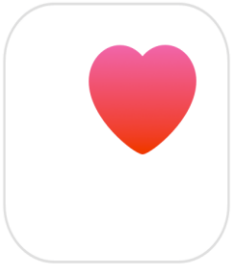
Solution

1. Participants: Extract and send XML data file

- Tap “Export All Health Data”

2. RStudio

- Libraries: xml2, tidyverse, lubridate, ggplot2



Solution

To read xml file

library(xml2)

xml_in <- read_xml(file.path("…"))

```
## {xml_document}
## <HealthData locale="en_GB">
##  [1] <ExportDate value="2021-03-22 10:31:51 +0000"/>
##  [2] <Me HKCharacteristicTypeIdentifierDateOfBirth="" HKCharacteristic
##  [3] <Record type="HKQuantityTypeIdentifierStepCount" sourceName="MD's
##  [4] <Record type="HKQuantityTypeIdentifierStepCount" sourceName="MD's
##  [5] <Record type="HKQuantityTypeIdentifierStepCount" sourceName="MD's
```

Solution

Let's look at one record

```
record <- xml_find_all(xml_in, "//Record")
record[[1]]
```

<Record type="HKQuantityTypeIdentifierStepCount" sourceName="Alex's phone" unit="count" creationDate="2020-06-21 12:57:31 +0000" startDate="2020-06-21 12:31:17 +0000" endDate="2020-06-21 12:33:00 +0000" value="30">

Solution

Let's pull out the data I need: record type, date, source and value

```
record_df <- map_dfr(  # rowbind to dataframe
  c(date = "creationDate", source = “sourceName”, type = "type",
    steps = "value"),
  ~xml_attr(records, .x)
)
```

Solution

```
glimpse(record_df)  # preview
```

```
## Rows: 560,001
## Columns: 4
## $ date <chr> “2020-06-15 12:57:31 +0000”…
## $ source <chr> “Alex’s iPhone”
## $ type <chr> “HKQuantityTypeIdentifierStepCount”…
## $ steps <chr> “30”
```

Solution

NEW PROBLEM

Which record type?

pull(distinct(record_df, type))

```
##  [1] "HKQuantityTypeIdentifierStepCount"
##  [2] "HKQuantityTypeIdentifierDistanceWalkingRunning"
##  [3] "HKQuantityTypeIdentifierActiveEnergyBurned"
##  [4] "HKQuantityTypeIdentifierFlightsClimbed"
##  [5] "HKQuantityTypeIdentifierHeadphoneAudioExposure"
##  [6] "HKQuantityTypeIdentifierWalkingDoubleSupportPercentage"
##  [7] "HKQuantityTypeIdentifierWalkingSpeed"
##  [8] "HKQuantityTypeIdentifierWalkingStepLength"
##  [9] "HKQuantityTypeIdentifierWalkingAsymmetryPercentage"
## [10] "HKCategoryTypeIdentifierSleepAnalysis"
## [11] "HKCategoryTypeIdentifierMindfulSession"
```

Solution

NEW PROBLEM

Which source?

```
pull(distinct(record_df, source))
## [1] "Alex's iPhone"
## [2] "Alex's Apple Watch"
```

Use filter function:

```
data <- record_df %>%
    filter (type =="HKQuantityTypeIdentifierStepCount") %>%
    filter (sourceName=="Alex's iPhone")
```

Solution

GET DAILY STEP COUNT

```
data<- mutate(date = as.Date(date), steps = as.integer(steps)) %>%
    group_by(date) %>%
    summarise(steps=sum(steps))
```

Extra resources: https://www.rostrum.blog/2021/03/23/xml-health/

Outline

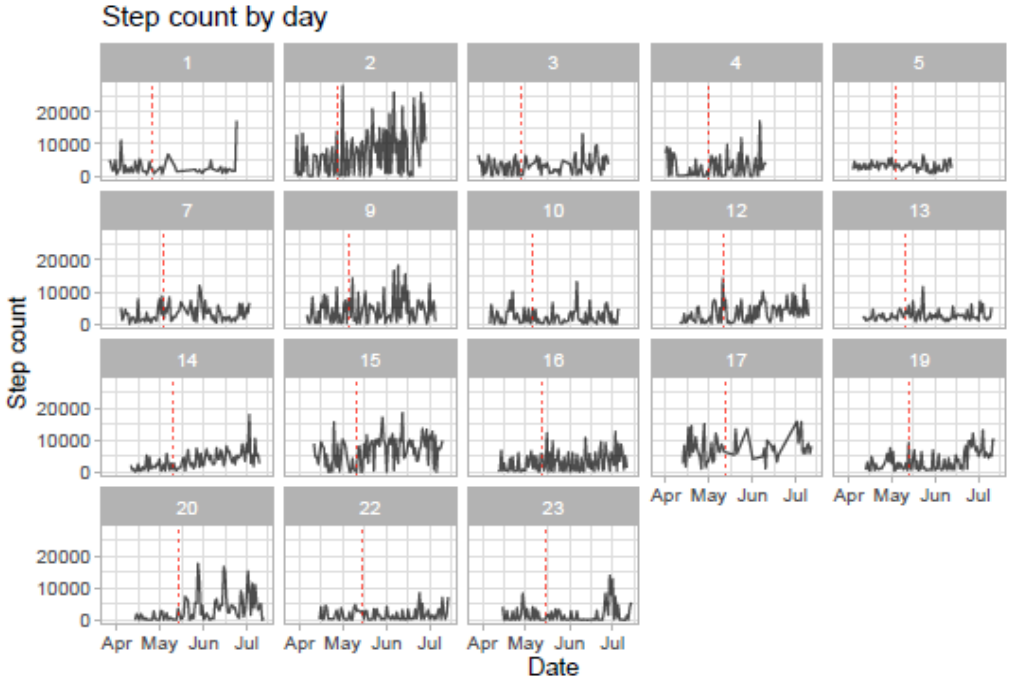
- Background about the project and the step dataset
- Set up RStudio and GitHub
- Problem 1: Missing data
- **Problem 2: Analyse longitudinal data**
- Problem 3: Too few steps in a day?

Problem: Longitudinal data

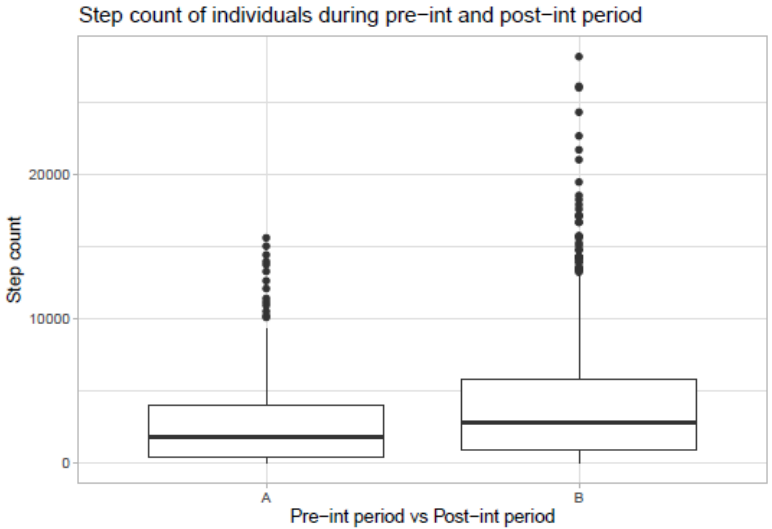
study_id	user_id	date	steps
10	uqwPBcx7	6/04/2020	119
10	uqwPBcx7	7/04/2020	6015
10	uqwPBcx7	8/04/2020	3181
10	uqwPBcx7	9/04/2020	24
10	uqwPBcx7	10/04/2020	1110
10	uqwPBcx7	11/04/2020	4074
10	uqwPBcx7	12/04/2020	2654
10	uqwPBcx7	13/04/2020	109
10	uqwPBcx7	14/04/2020	1906
10	uqwPBcx7	15/04/2020	188
10	uqwPBcx7	16/04/2020	106
10	uqwPBcx7	17/04/2020	155
10	uqwPBcx7	18/04/2020	4025

Visualisation first!

```
ggplot(data, aes(x=date, y=steps)) +
  geom_line(alpha=0.7)  +
  geom_vline(aes(xintercept =
start_date +     30), alpha=0.7,
linetype="dotted",  color="red",
show_guide = TRUE) +
  ggtitle('Step count by day') +
  xlab('Date') +
  ylab('Step count') +
  theme_light() +
  theme(legend.position = "none") +
  facet_wrap( ~ study_id)
```



Visualisation first!



- **Do not follow a normal distribution!**

Solution 1: Compare monthly median

DEFINE STUDY PERIOD: PRE/POST INTERVENTION

```
# 1: calculate date difference from current date - start date
data <- data %>% mutate(date_count = date - start_date + 1)

# 2: define period type: if date_count =<30 => 0 (baseline)
# date_count >30 & <61 => 1 (first month in the intervention); date_count >= 61 => 2 (2nd month)
data$period <- ifelse(data$date_count <31,0,
                      ifelse(data$date_count >=31 & data$date_count <61,1,
                             ifelse(data$date_count>=61,2,
                                    NA)))
```

Solution 1: Compare monthly median

CALCULATE MEDIAN FOR EACH USER IN EACH MONTH

```
# 3: calculate median for each user in each month
# 3.1: calculate baseline period
baseline <- data %>% filter(period == 0)
baseline <- baseline %>%
  group_by(study_id) %>%
  summarize(median_baseline = median(steps, na.rm = TRUE))

# 3.2: calculate 1st month
month1 <- data %>% filter(period == 1)
month1 <- month1 %>%
  group_by(study_id) %>%
  summarize(median_month1 = median(steps, na.rm = TRUE))

# 3.3: calculate 2nd month
month2 <- data %>% filter(period == 2)
month2 <- month2%>%
  group_by(study_id) %>%
  summarize(median_month2 = median(steps, na.rm = TRUE))
```

Solution 1: Compare monthly median

PERFORM WILCOXON SIGNED-RANK TEST

```
# 4: perform wilcox test on median
# 4.1: compare baseline and 1st month
wilcox.test(step_data$median_month1,step_data$median_baseline,
            paired = TRUE, alternative = "two.sided", conf.int = TRUE)
```

```
        Wilcoxon signed rank exact test

data:  step_data$median_month1 and step_data$median_baseline
V = 128, p-value = 0.06654
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
  -58.0 1844.5
sample estimates:
(pseudo)median
         696.5
```

Solution 1: Compare monthly median

PERFORM WILCOXON SIGNED-RANK TEST

```
# 4.2: compare baseline and 2nd month
wilcox.test(step_data$median_month2,step_data$median_baseline,
            paired = TRUE, alternative = "two.sided", conf.int = TRUE)
```

```
        Wilcoxon signed rank exact test

data:  step_data$median_month2 and step_data$median_baseline
V = 158, p-value = 0.0006714
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
  782.243 3112.000
sample estimates:
(pseudo)median
          1953
```

Conclusion: Step count increased significantly towards the end of the study.

Outline


Background about the project and the step dataset


Set up RStudio and GitHub


Problem 1: Missing data


Problem 2: Analyse longitudinal data


Problem 3: Too few steps in a day?

Problem: Too few steps in a day

study_id	user_id	date	steps
23	iQupzXynI	10/06/2020	1
2	IedFqakIT	8/04/2020	3
11	abk3RkrHI	6/05/2020	4
20	AGuyeKpz	12/07/2020	4
2	IedFqakIT	10/04/2020	4
4	OL1F2UT6	31/05/2020	8
4	OL1F2UT6	1/05/2020	10
15	pafyuWn6	1/06/2020	11
2	IedFqakIT	16/05/2020	11
4	OL1F2UT6	29/04/2020	11
2	IedFqakIT	21/04/2020	13
15	pafyuWn6	7/05/2020	14
23	iQupzXynI	13/06/2020	16
2	IedFqakIT	14/04/2020	16
4	OL1F2UT6	27/04/2020	16
2	IedFqakIT	11/06/2020	18

Possible explanations:

- Injury-cannot move
- Forget to carry phones

Solution: Sensitivity analysis

Data <- filter(steps >=100)

Then carry out the same analysis

Acknowledgement

Dr Liliana Laranjo

Dr Juan Quiroz

Mr Jason Dalmazzo

Mr Alexander Southern

Mr Joshua Irawan

Ms Yoonah Kim

Dr Baki Kocaballi

Dr Dana Rezazadegan

Mr Vitaliy Kim

Dr Kiran Ijaz

Ms Agustina Briatore

Ms Kim Phuong Dao

@lytong22

hton5658@uni.sydney.edu.au