Complex Sampling Design in National Health and Morbidity Survey (NHMS)

Survey Package in R

Mohd Azmi Bin Suliman

azmi.suliman@moh.gov.my

Pusat Penyelidikan Penyakit Tak Berjangkit, Institut Kesihatan Umum

Sunday, 27 October 2024

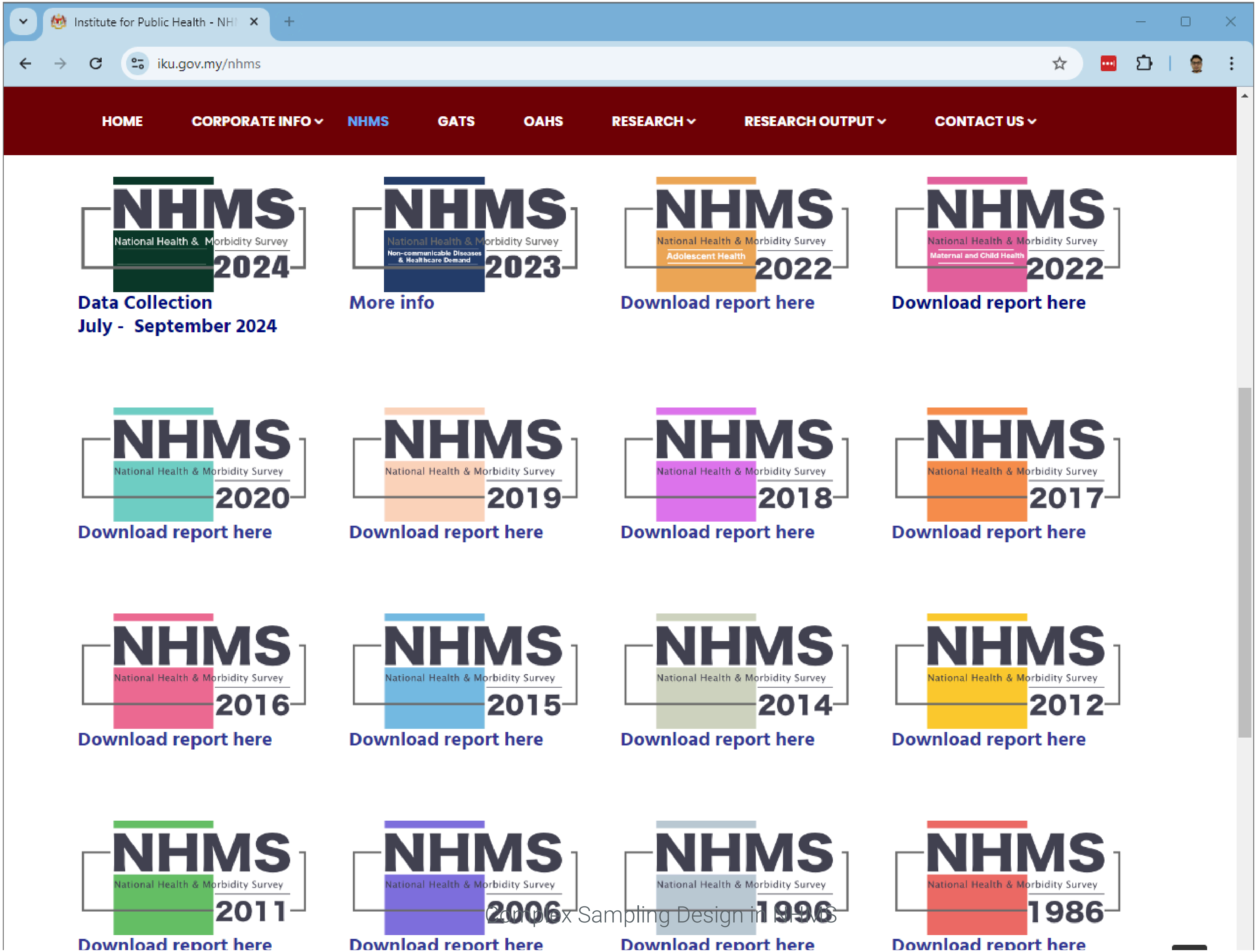
National Health and Morbidity Survey (NHMS)

Overview of NHMS

1. **Nationwide Health Survey**: Conducted by the Ministry of Health Malaysia to assess the health and healthcare needs of Malaysians.
2. **Regularly Conducted**: Since 1986, NHMS has been conducted with varying intervals, focusing on different health themes.
3. **Key Health Indicators**: Focuses on topics like Non-Communicable Diseases (NCDs), infectious diseases, and healthcare demand.
4. **Representative Sampling**: Nationally representative, covering different states, age groups, and ethnicities.
5. **Policy Impact**: NHMS findings guide national health policies and strategies.

NHMS Reports

- NHMS reports are available on the Institute for Public Health (IKU) website.


Institute for Public Health - NHI
iku.gov.my/nhms
HOME
CORPORATE INFO
NHMS
GATS
OAHS
RESEARCH
RESEARCH OUTPUT
CONTACT US
NHMS National Health & Morbidity Survey 2024
Data Collection
July - September 2024
NHMS National Health & Morbidity Survey Non-communicable Diseases & Healthcare Demand 2023
More info
NHMS National Health & Morbidity Survey Adolescent Health 2022
Download report here
NHMS National Health & Morbidity Survey Maternal and Child Health 2022
Download report here
NHMS National Health & Morbidity Survey 2020
Download report here
NHMS National Health & Morbidity Survey 2019
Download report here
NHMS National Health & Morbidity Survey 2018
Download report here
NHMS National Health & Morbidity Survey 2017
Download report here
NHMS National Health & Morbidity Survey 2016
Download report here
NHMS National Health & Morbidity Survey 2015
Download report here
NHMS National Health & Morbidity Survey 2014
Download report here
NHMS National Health & Morbidity Survey 2012
Download report here
NHMS National Health & Morbidity Survey 2011
Download report here
NHMS National Health & Morbidity Survey 2006
Download report here
NHMS National Health & Morbidity Survey 1996
Download report here
NHMS National Health & Morbidity Survey 1986
Download report here

Census vs Survey

Census vs Survey

- **Census**: Collects data from every individual in a population. It's costly, time-consuming, and not feasible for large populations.
  - → e.g., DOSM conducts a Population and Housing Census every 10 years.
- **Survey**: Collects data from a sample of the population. More cost-effective and quicker but raises the question of representation.

Why Not Simple Random Sampling (SRS)?

- **Simple Random Sampling (SRS)**: Every individual theoretically has an equal chance of selection.
  - → Impractical for large, diverse populations.
  - → Assumes homogeneity, which leads to biases, especially with underrepresented groups.
- **Challenges of SRS**:
  - → Requires a complete list of the population for equal chance selection, which is often unavailable.

Simple Random Sampling (SRS) – Simulation

The Risk of Underrepresentation

- SRS may not represent minority groups adequately.
- Hypothetical Population, In a population of 1,000:
  - → 46% Malay, 33% Chinese, 25% Indian, 1% Borneo.

```r
sim_pop <- tibble(ethnicity = c(rep("Malay", 460),
                                rep("Chinese", 330),
                                rep("Indian", 250),
                                rep("Borneo", 10))) %>%
  mutate(ethnicity = fct_relevel(ethnicity, "Malay", "Chinese", "Indian"))

sim_pop %>%
  count(ethnicity) %>% mutate(pct = scales::label_percent()(n/1000))
```

```
# A tibble: 4 × 3
  ethnicity     n pct
  <fct>     <int> <chr>
1 Malay       460 46%
2 Chinese     330 33%
3 Indian      250 25%
4 Borneo       10 1%
```

The Risk of Underrepresentation

- Taking an SRS of 50 people, will the Borneo group (1%) be included?

```
sim_pop %>% sample_n(50) %>%
  count(ethnicity) %>% mutate(pct = scales::label_percent()(n/50))
```

```
# A tibble: 3 × 3
  ethnicity      n pct
  <fct>      <int> <chr>
1 Malay         21 42%
2 Chinese       13 26%
3 Indian        16 32%
```

- Let's try it again:

```
sim_pop %>% sample_n(50) %>%
  count(ethnicity) %>% mutate(pct = scales::label_percent()(n/50))
```

```
# A tibble: 3 × 3
  ethnicity      n pct
  <fct>      <int> <chr>
1 Malay         20 40%
2 Chinese       15 30%
3 Indian        15 30%
```

Key Takeaways from the Simulation

- And One More Time

```
sim_pop %>% sample_n(50) %>%
  count(ethnicity) %>% mutate(pct = scales::label_percent()(n/50))
```

```
# A tibble: 4 × 3
  ethnicity     n pct
  <fct>     <int> <chr>
1 Malay        16 32%
2 Chinese      18 36%
3 Indian       15 30%
4 Borneo        1 2%
```

- As shown in this short simulation, **Simple Random Sampling** may or may not select individuals from the **Borneo** group, which makes up only 1% of the population.
- To ensure that the **Borneo** group is properly represented in the sample, we may need to use **stratified sampling** to guarantee their inclusion.

Complex Sampling Design in NHMS

NHMS Complex Sampling

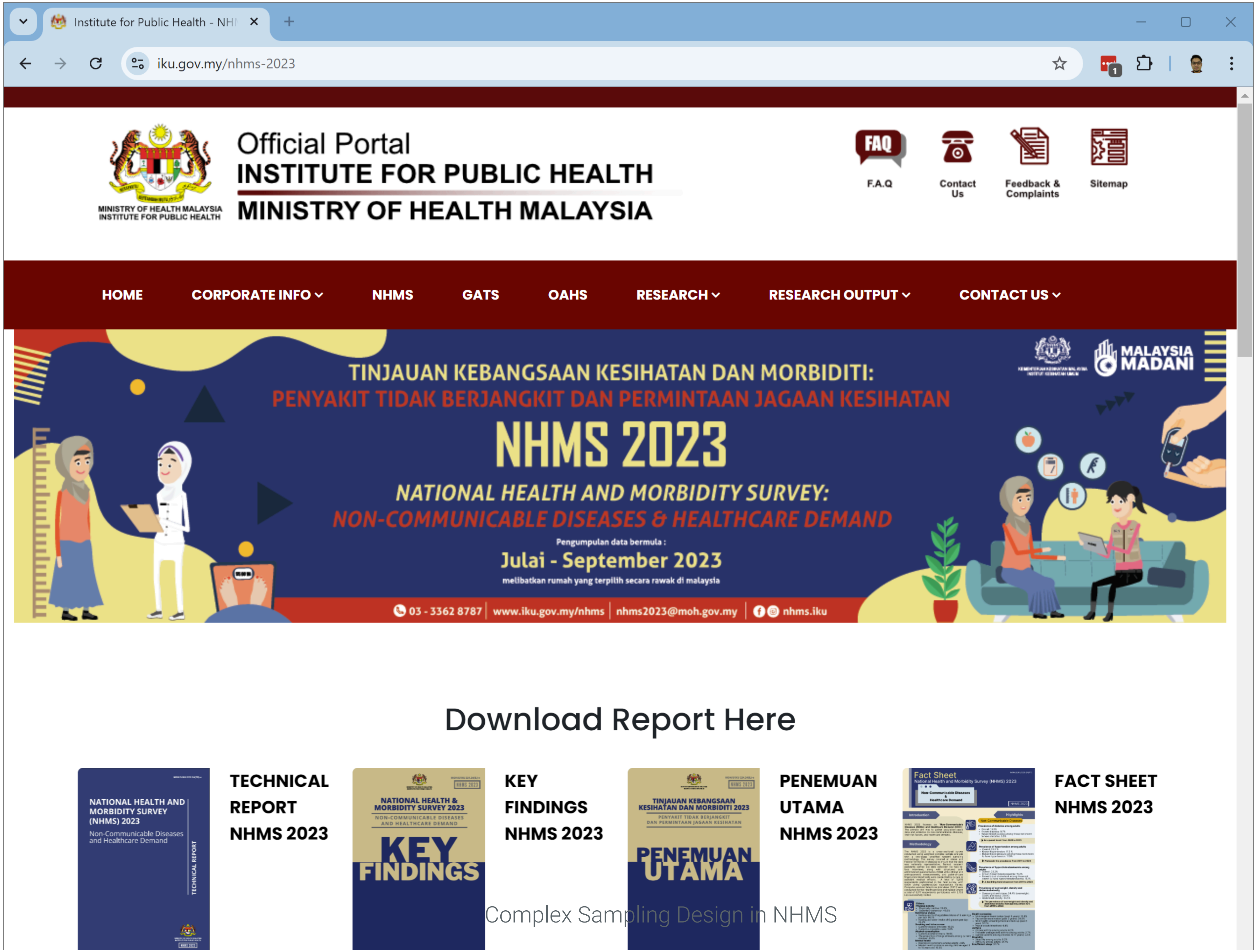
- NHMS applies stratification (State and Urban/Rural) and clustering (DOSM's enumeration blocks) to ensure representation.
- Two-stage Sampling:
  - → **Primary Sampling Unit (PSU)**: Enumeration Blocks (EBs).
  - → **Secondary Sampling Units (SSU)**: Living Quarters (LQs) within EBs.
- Impact on Sampling: Stratification and clustering affect sampling probabilities, requiring the use of sampling weights.

NHMS 2023

NHMS 2023 Overview

- Theme: **Non-communicable Diseases (NCDs) & Healthcare Demand**.
- Data collected from 11 July to 29 September 2023.
    - → 5,006 households visited.
    - → 13,616 respondents, representing the Malaysian adult population (~22 million).

NHMS 2023 Overview


Institute for Public Health - NH
iku.gov.my/nhms-2023
MINISTRY OF HEALTH MALAYSIA
INSTITUTE FOR PUBLIC HEALTH
Official Portal
INSTITUTE FOR PUBLIC HEALTH
MINISTRY OF HEALTH MALAYSIA
FAQ
F.A.Q
Contact Us
Feedback & Complaints
Sitemap
HOME
CORPORATE INFO
NHMS
GATS
OAHS
RESEARCH
RESEARCH OUTPUT
CONTACT US
MALAYSIA MADANI
TINJAUAN KEBANGSAAN KESIHATAN DAN MORBIDITI:
PENYAKIT TIDAK BERJANGKIT DAN PERMINTAAN JAGAAN KESIHATAN
NHMS 2023
NATIONAL HEALTH AND MORBIDITY SURVEY:
NON-COMMUNICABLE DISEASES & HEALTHCARE DEMAND
Pengumpulan data bermula :
Julai - September 2023
melibatkan rumah yang terpilih secara rawak di malaysia
03 - 3362 8787 | www.iku.gov.my/nhms | nhms2023@moh.gov.my | nhms.iku
Download Report Here
NATIONAL HEALTH AND MORBIDITY SURVEY (NHMS) 2023
Non-Communicable Diseases and Healthcare Demand
TECHNICAL REPORT
TECHNICAL REPORT NHMS 2023
NATIONAL HEALTH & MORBIDITY SURVEY 2023
NON-COMMUNICABLE DISEASES AND HEALTHCARE DEMAND
KEY FINDINGS
KEY FINDINGS NHMS 2023
TINJAUAN KEBANGSAAN KESIHATAN DAN MORBIDITI 2023
PENYAKIT TIDAK BERJANGKIT DAN PERMINTAAN JAGAAN KESIHATAN
PENEMUAN UTAMA
PENEMUAN UTAMA NHMS 2023
Fact Sheet
FACT SHEET NHMS 2023

NHMS 2023: Findings

- NHMS 2023 included various modules focusing on **Non-Communicable Diseases (NCDs)** and **healthcare demand**.
- The **cholesterol module** was conducted using WHO's **STEPwise approach**, which is a standardized method for collecting and analysing health data.
- Among the respondents, **4,353 individuals** were identified as having **raised total cholesterol** levels.
- This represents a **33.3% prevalence** of raised cholesterol, translating to an estimated **7.6 million Malaysian adults** with high cholesterol.

Simulation and Analysis of Complex Sampling Design

Purpose of Simulation

- Objective: Demonstrate how complex sampling design is applied in practice, mimicking the National Health and Morbidity Survey (NHMS) setup.
- Population Data: We simulate data using population estimates from OpenDOSM to replicate the adult population (ages 20-59) for Malaysia.
- Disease Data: Simulated based on characteristics such as BMI, age, gender, and ethnicity to study cholesterol prevalence.

Simulating Population

- Target Population:
  - → We focus on three main ethnicities (Malay, Chinese, Indian), and simulate both male and female participants between 20-59 years of age.

Simulating Population

- OpenDOSM Population Data: Used for population distribution across states and districts, forming the basis for the simulated population.

```
library(arrow)

pop_mydist <- read_parquet("https://storage.dosm.gov.my/population/population_district.parquet'
  filter(date == dmy("01/01/23"),
         sex != "overall",
         age %in% c("20-24", "24-29", "30-34", "35-39", "40-44",
                    "45-49", "50-54", "55-59"),
         ethnicity %in% c("bumi_malay", "chinese", "indian")) %>%
  rename(gender = sex) %>%
  mutate(gender = fct_recode(gender,
                             "Male" = "male",
                             "Female" = "female"),
         ethnicity = fct_recode(ethnicity,
                                "Malay" = "bumi_malay",
                                "Chinese" = "chinese",
                                "Indian" = "indian"),
         population = population * 1000)
```

```r
pop_mydist
```

```
# A tibble: 6,720 × 7
   state district   date       gender age   ethnicity population
   <chr> <chr>      <date>     <fct>  <chr> <fct>          <dbl>
 1 Johor Batu Pahat 2023-01-01 Male   20-24 Malay          15600
 2 Johor Batu Pahat 2023-01-01 Male   30-34 Malay          13400
 3 Johor Batu Pahat 2023-01-01 Male   35-39 Malay          13300
 4 Johor Batu Pahat 2023-01-01 Male   40-44 Malay          11700
 5 Johor Batu Pahat 2023-01-01 Male   45-49 Malay           9100
 6 Johor Batu Pahat 2023-01-01 Male   50-54 Malay           8200
 7 Johor Batu Pahat 2023-01-01 Male   55-59 Malay           7500
 8 Johor Batu Pahat 2023-01-01 Female 20-24 Malay          15000
 9 Johor Batu Pahat 2023-01-01 Female 30-34 Malay          13000
10 Johor Batu Pahat 2023-01-01 Female 35-39 Malay          13000
# ℹ 6,710 more rows
```

```r
pop_mydist %>%
  summarise(population = sum(population)) %>%
  mutate(population = scales::label_comma()(population))
```

```
# A tibble: 1 × 1
  population
  <chr>
1 12,816,400
```

Simulating Population

- Stratification by Zone:
  - → Malaysia is divided into five zones (Utara, Selatan, Timur, Tengah, Borneo).
  - → For each zone, two districts are randomly selected.

```
set.seed(121)
selected_district <- pop_mydist %>%
  distinct(state, district) %>%
  mutate(zone = case_when(state %in% c("Johor", "Melaka", "Negeri Sembilan") ~ "Selatan",
                          state %in% c("Kedah", "Perak", "Perlis", "Pulau Pinang") ~ "Utara",
                          state %in% c("Kelantan", "Pahang", "Terengganu") ~ "Timur",
                          state %in% c("Selangor", "W.P. Kuala Lumpur", "W.P. Putrajaya") ~ "Ter
                          state %in% c("Sabah", "Sarawak", "W.P. Labuan") ~ "Borneo")) %>%
  group_by(zone) %>%
  slice_sample(n = 2) %>%
  ungroup() %>%
  relocate(zone, .before = 1)

selected_district
```

```
# A tibble: 10 × 3
   zone    state        district
   <chr>   <chr>        <chr>
 1 Borneo  Sarawak      Pusa
 2 Borneo  Sarawak      Asajaya
 3 Selatan Melaka       Jasin
 4 Selatan Johor        Muar
 5 Tengah  Selangor     Kuala Selangor
 6 Tengah  Selangor     Ulu Selangor
 7 Timur   Terengganu   Kuala Nerus
 8 Timur   Kelantan     Bachok
 9 Utara   Pulau Pinang Barat Daya
10 Utara   Perak        Bagan Datuk
```

Simulating Participants

- Sample Size: 40 participants are selected from each of the 10 districts.
- Variables Simulated: Variables such as gender, age, ethnicity, BMI, and hba1c are simulated to reflect realistic population characteristics.
- Disease Data: The hba1c variable is used to categorize participants as diabetic or non-diabetic.

Define Simulation (simstudy package)

```
#| eval: false

library(simstudy)

def <- defData(varname = "gender", dist = "binary", formula = 0.5,
               link = "identity") %>%
  defData(varname = "age", dist = "uniform", formula = "20;59") %>%
  defData(varname = "ethnicity", dist = "categorical",
          formula = "0.57;0.29;0.14") %>%
  defData(varname = "BMI", dist = "normal", formula = 26, variance = 2.6^2) %>%
  defData(varname = "height", dist = "normal", formula = 165, variance = 5) %>%
  defData(varname = "PAhour", dist = "uniform", formula = "2;6") %>%
  defData(varname = "hba1c", dist = "normal", variance = 1.4^2,
          formula = "2.4 + 0.05 * age + 0.1 * BMI - 0.15 * PAhour")
```

Generate Dataset (simstudy package)

```
#| eval: false

set.seed(121)
simnhmsds0 <- genData(400, def) %>%
  mutate(gender = case_when(gender == 0 ~ "Male",
                            gender == 1 ~ "Female"),
         agegp = cut(age,
                     breaks = c(19, 29, 39, 49, 59),
                     labels = c("20-29", "30-39", "40-49", "50-59")),
         ethnicity = case_when(ethnicity == 1 ~ "Malay",
                               ethnicity == 2 ~ "Chinese",
                               ethnicity == 3 ~ "Indian"),
         weight = BMI * (height/100)^2,
         across(.cols = c(hba1c, weight),
                .fns = ~ round(., 1)),
         across(.cols = c(height),
                .fns = ~ round(., 2)),
         district = rep(1:10, each = 40),
         dm_dx = cut(hba1c,
```

```r
head(simnhmsds0, 15)
```

```
# A tibble: 15 × 13
      id zone   state district gender   age agegp ethnicity height weight PAhour
   <int> <chr>  <chr> <chr>    <chr>  <int> <fct> <chr>      <int>  <dbl>  <int>
 1     1 Borneo Sara… Pusa     Male      28 20-29 Chinese      163   56.7      2
 2     2 Borneo Sara… Pusa     Female    53 50-59 Indian       164   63.2      5
 3     3 Borneo Sara… Pusa     Female    53 50-59 Chinese      164   71.1      3
 4     4 Borneo Sara… Pusa     Female    39 40-49 Malay        164   76.7      2
 5     5 Borneo Sara… Pusa     Female    32 30-39 Chinese      161   69.8      3
 6     6 Borneo Sara… Pusa     Male      28 20-29 Malay        168   71.5      5
 7     7 Borneo Sara… Pusa     Male      28 20-29 Chinese      164   63.9      3
 8     8 Borneo Sara… Pusa     Female    51 50-59 Indian       167   66.9      4
 9     9 Borneo Sara… Pusa     Male      23 20-29 Chinese      164   64.4      2
10    10 Borneo Sara… Pusa     Male      23 20-29 Malay        166   79.3      3
11    11 Borneo Sara… Pusa     Female    47 40-49 Malay        164   73.3      5
12    12 Borneo Sara… Pusa     Female    37 30-39 Malay        170   74.8      3
13    13 Borneo Sara… Pusa     Female    30 30-39 Malay        163   66.9      5
14    14 Borneo Sara… Pusa     Male      37 30-39 Chinese      164   69.5      4
15    15 Borneo Sara… Pusa     Male      30 30-39 Indian       163   73.9      5
# ℹ 2 more variables: hba1c <dbl>, dm_dx <int>
```

Simulating Non-response

- Non-response Adjustment: Different response rates are applied by zone (e.g., 36/40 for Utara), and the sampling weights are adjusted accordingly.
- Simulation of Non-response: The dataset is adjusted to reflect these response rates, ensuring the final dataset accounts for real-world data collection challenges.

Simulating Non-response

```r
sample_sizes <- list("Utara" = 36, "Selatan" = 34,
                     "Timur" = 38, "Tengah" = 32, "Borneo" = 38)

simnhmsds_split <- simnhmsds0 %>%
  group_split(zone, .keep = TRUE)

set.seed(121)
simnhmsds_final <- map(simnhmsds_split, function(data) {
  zone <- unique(data$zone)
  data %>%
    group_by(district) %>%
    slice_sample(n = sample_sizes[[zone]]) %>%
    ungroup()
}) %>%
  bind_rows() %>%
  arrange(id) %>%
  mutate(success = 1,
         .before = 1)
```

Sampling Weights Calculation

- Design Weights (W1):
  - → Calculated as the inverse probability of selecting a district within its respective zone.
  - → Ensures that smaller districts are adequately represented in the final analysis.
- Non-response Adjustment Factor (F):
  - → The inverse of the response rate for each district.
  - → Adjusts the design weight to account for missing data due to non-response.

Sampling Weights Calculation

- Post-stratification Adjustment (PS)
  - → Ensures that the sample reflects the actual population distribution by gender, age group, and ethnicity.
  - → Uses the population data to adjust for any over- or under-representation in the sample.

Design Weight (W1)

- Design Weights (W1):
  - → Calculated as the inverse probability of selecting a district within its respective zone.
  - → Ensures that smaller districts are adequately represented in the final analysis.

Design Weight (W1)

```
design_weight <- pop_mydist %>%
  distinct(state, district) %>%
  mutate(zone = case_when(state %in% c("Johor", "Melaka", "Negeri Sembilan") ~ "Selatan",
                          state %in% c("Kedah", "Perak", "Perlis", "Pulau Pinang") ~ "Utara",
                          state %in% c("Kelantan", "Pahang", "Terengganu") ~ "Timur",
                          state %in% c("Selangor", "W.P. Kuala Lumpur", "W.P. Putrajaya") ~ "Ter
                          state %in% c("Sabah", "Sarawak", "W.P. Labuan") ~ "Borneo")) %>%
  count(zone) %>%
  rename(total_district = n) %>%
  mutate(selected_district = 2,
         f1 = selected_district / total_district,
         W1 = 1/f1)

design_weight
```

```
# A tibble: 5 × 5
  zone     total_district selected_district     f1    W1
  <chr>             <int>             <dbl>  <dbl> <dbl>
1 Borneo               68                 2 0.0294  34  
2 Selatan              20                 2 0.1     10  
3 Tengah               11                 2 0.182    5.5
4 Timur                30                 2 0.0667  15  
5 Utara                31                 2 0.0645  15.5
```

Design Weight (W1)

- We use survey and srvyr package to recalculate the design weight by district

```
library(survey)
library(srvyr)

district_ws <- simnhmsds_final %>%
  left_join(.,
            design_weight %>%
              select(zone, W1),
            by = "zone") %>%
  as_survey_design(id = 1,
                   weight = W1) %>%
  group_by(district) %>%
  summarise(district_w1 = survey_total(success),
            .groups = "drop")
```

Design Weight (W1)

```
district_ws
```

```
# A tibble: 10 × 3
   district        district_w1 district_w1_se
   <chr>                 <dbl>          <dbl>
 1 Asajaya                1292          198. 
 2 Bachok                  570           87.5
 3 Bagan Datuk             558           88.3
 4 Barat Daya              558           88.3
 5 Jasin                   340           55.5
 6 Kuala Nerus             570           87.5
 7 Kuala Selangor          176           29.7
 8 Muar                    340           55.5
 9 Pusa                   1292          198. 
10 Ulu Selangor            176           29.7
```

Non-response Adjustment Factor (F)

- Non-response Adjustment Factor (F):
  - → The inverse of the response rate for each district.
  - → Adjusts the design weight to account for missing data due to non-response.

```r
nonresponse_weight <- selected_district %>%
  mutate(fnr = case_when(zone == "Utara" ~ 36/40,
                         zone == "Selatan" ~ 34/40,
                         zone == "Timur" ~ 38/40,
                         zone == "Tengah" ~ 32/40,
                         zone == "Borneo" ~ 38/40),
         Fw = 1/fnr)

nonresponse_weight
```

Non-response Adjustment Factor (F)

```
# A tibble: 10 × 5
   zone    state         district           fnr    Fw
   <chr>   <chr>         <chr>            <dbl> <dbl>
 1 Borneo  Sarawak       Pusa              0.95  1.05
 2 Borneo  Sarawak       Asajaya           0.95  1.05
 3 Selatan Melaka        Jasin             0.85  1.18
 4 Selatan Johor         Muar              0.85  1.18
 5 Tengah  Selangor      Kuala Selangor    0.8   1.25
 6 Tengah  Selangor      Ulu Selangor      0.8   1.25
 7 Timur   Terengganu    Kuala Nerus       0.95  1.05
 8 Timur   Kelantan      Bachok            0.95  1.05
 9 Utara   Pulau Pinang  Barat Daya        0.9   1.11
10 Utara   Perak         Bagan Datuk       0.9   1.11
```

Non-response Adjustment Factor (F)

- The non-response adjustment factor (F) is calculated for each district based on the response rate.

```r
district_adw <- nonresponse_weight %>%
  left_join(.,
            district_ws %>%
              select(district, district_w1),
            by = "district") %>%
  mutate(district_adw = district_w1 * Fw)

district_adw
```

```
# A tibble: 10 × 7
   zone    state         district         fnr    Fw district_w1 district_adw
   <chr>   <chr>         <chr>          <dbl> <dbl>       <dbl>        <dbl>
 1 Borneo  Sarawak       Pusa            0.95  1.05        1292         1360
 2 Borneo  Sarawak       Asajaya         0.95  1.05        1292         1360
 3 Selatan Melaka        Jasin           0.85  1.18         340          400
 4 Selatan Johor         Muar            0.85  1.18         340          400
 5 Tengah  Selangor      Kuala Selangor  0.8   1.25         176          220
 6 Tengah  Selangor      Ulu Selangor    0.8   1.25         176          220
 7 Timur   Terengganu    Kuala Nerus     0.95  1.05         570          600
 8 Timur   Kelantan      Bachok          0.95  1.05         570          600
 9 Utara   Pulau Pinang  Barat Daya      0.9   1.11         558          620
10 Utara   Perak         Bagan Datuk     0.9   1.11         558          620
```

Post-stratification Adjustment (PS)

- Post-stratification Adjustment (PS)
  - → Ensures that the sample reflects the actual population distribution by gender, age group, and ethnicity.
  - → Uses the population data to adjust for any over- or under-representation in the sample.

Total Population by Post-strat Group

```r
ps_pop <- pop_mydist %>%
  mutate(agegp = case_when(age %in% c("20-24", "24-29") ~ "20-29",
                           age %in% c("30-34", "35-39") ~ "30-39",
                           age %in% c("40-44", "45-49") ~ "40-49",
                           age %in% c("50-54", "55-59") ~ "50-59")) %>%
  group_by(gender, agegp, ethnicity) %>%
  summarise(population = sum(population),
            .groups = "drop")

ps_pop %>%
  mutate(popcoma = scales::label_comma()(population))
```

```
# A tibble: 24 × 5
   gender agegp ethnicity population popcoma
   <fct>  <chr> <fct>          <dbl> <chr>
 1 Female 20-29 Malay         777700 777,700
 2 Female 20-29 Chinese       241200 241,200
 3 Female 20-29 Indian         76800 76,800
 4 Female 30-39 Malay        1445300 1,445,300
 5 Female 30-39 Chinese       529300 529,300
 6 Female 30-39 Indian        166100 166,100
 7 Female 40-49 Malay        1101900 1,101,900
 8 Female 40-49 Chinese       515800 515,800
 9 Female 40-49 Indian        146700 146,700
10 Female 50-59 Malay         799700 799,700
# ℹ 14 more rows
```

Post-stratification Adjustment (PS)

- The adjusted weight is attached back to our dataset, and post-stratification weight is calculated using survey and srvyr package.

```
ps_weight <- simnhmsds_final %>%
  left_join(.,
            district_adw %>%
              select(district, district_adw),
            by = join_by(district)) %>%
  as_survey_design(id = 1,
                   weights = district_adw) %>%
  group_by(gender, agegp, ethnicity) %>%
  summarise(ps_adw = survey_total(success),
            .groups = "drop") %>%
  select(-ps_adw_se) %>%
  left_join(simnhmsds_final %>%
              count(gender, agegp, ethnicity),
            .,
            by = join_by(gender, agegp, ethnicity)) %>%
  left_join(.,
            ps_pop,
            by = join_by(gender, agegp, ethnicity)) %>%
  mutate(fps = population / ps_adw,
```

Post-stratification Adjustment (PS)

- The adjusted weight is attached back to our dataset, and post-stratification weight is calculated using survey and srvyr package.

```
ps_weight
```

```
# A tibble: 24 × 8
   gender agegp ethnicity     n ps_adw population   fps final_weight
   <chr>  <chr> <chr>     <int>  <dbl>      <dbl> <dbl>        <dbl>
 1 Female 20-29 Chinese      15   8220     241200  29.3      0.0341
 2 Female 20-29 Indian        6   3380      76800  22.7      0.0440
 3 Female 20-29 Malay        21  14380     777700  54.1      0.0185
 4 Female 30-39 Chinese      16  10420     529300  50.8      0.0197
 5 Female 30-39 Indian        6   4560     166100  36.4      0.0275
 6 Female 30-39 Malay        19  15420    1445300  93.7      0.0107
 7 Female 40-49 Chinese       4   1840     515800 280.       0.00357
 8 Female 40-49 Indian        8   3820     146700  38.4      0.0260
 9 Female 40-49 Malay        33  21300    1101900  51.7      0.0193
10 Female 50-59 Chinese      14   8020     428200  53.4      0.0187
# ℹ 14 more rows
```

Attaching Final Weight to Dataset

- The final weight is attached to the dataset for further analysis.

```
simnhmsds_weight <-  simnhmsds_final %>%
  left_join(.,
            ps_weight %>%
              select(gender, agegp, ethnicity, final_weight),
            by = join_by(gender, agegp, ethnicity))

simnhmsds_weight
```

```
# A tibble: 356 × 15
   success    id zone  state district gender   age agegp ethnicity height weight
     <dbl> <int> <chr> <chr> <chr>    <chr>  <int> <chr> <chr>      <int>  <dbl>
 1       1     1 Born… Sara… Pusa     Male      28 20-29 Chinese      163   56.7
 2       1     2 Born… Sara… Pusa     Female    53 50-59 Indian       164   63.2
 3       1     3 Born… Sara… Pusa     Female    53 50-59 Chinese      164   71.1
 4       1     4 Born… Sara… Pusa     Female    39 40-49 Malay        164   76.7
 5       1     6 Born… Sara… Pusa     Male      28 20-29 Malay        168   71.5
 6       1     7 Born… Sara… Pusa     Male      28 20-29 Chinese      164   63.9
 7       1     8 Born… Sara… Pusa     Female    51 50-59 Indian       167   66.9
 8       1     9 Born… Sara… Pusa     Male      23 20-29 Chinese      164   64.4
 9       1    10 Born… Sara… Pusa     Male      23 20-29 Malay        166   79.3
10       1    11 Born… Sara… Pusa     Female    47 40-49 Malay        164   73.3
# ℹ 346 more rows
# ℹ 4 more variables: PAhour <int>, hba1c <dbl>, dm_dx <int>,
#   final_weight <dbl>
```

Analyzing the Data

Survey Design Object:

- The `svydesign()` function from the survey package is used to define the complex survey design.
- We account for stratification, clustering, and weighting to accurately estimate population parameters.
- Parameters:
  - → `ids`: cluster id. ~1 if no cluster
  - → `probs` or `weights`: sampling probability or weight, use only one
  - → `strata`: strata id. NULL (or leave unspecified) if no strata
  - → `data`: dataset

Unweighted Design

```r
unwt_dsg <- svydesign(ids = ~1,
                      weights = 1,
                      data = simnhmsds_weight)

summary(unwt_dsg)
```

```
Independent Sampling design (with replacement)
svydesign(ids = ~1, weights = 1, data = simnhmsds_weight)
Probabilities:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      1       1       1       1       1       1 
Data variables:
 [1] "success"      "id"           "zone"         "state"        "district"    
 [6] "gender"       "age"          "agegp"        "ethnicity"    "height"      
[11] "weight"       "PAhour"       "hba1c"        "dm_dx"        "final_weight"
```

Weighted Design

```r
wtds_dsg <- svydesign(ids = ~district,
                      weights = ~final_weight,
                      strata = ~zone,
                      data = simnhmsds_weight)

summary(wtds_dsg)
```

```
Stratified 1 - level Cluster Sampling design (with replacement)
With (10) clusters.
svydesign(ids = ~district, weights = ~final_weight, strata = ~zone,
    data = simnhmsds_weight)
Probabilities:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  13.48   35.83   52.25   54.56   65.99  280.33
Stratum Sizes:
           Borneo Selatan Tengah Timur Utara
obs            76      68     64    76    72
design.PSU      2       2      2     2     2
actual.PSU      2       2      2     2     2
Data variables:
 [1] "success"      "id"           "zone"         "state"        "district"
 [6] "gender"       "age"          "agegp"        "ethnicity"    "height"
[11] "weight"       "PAhour"       "hba1c"        "dm_dx"        "final_weight"
```

Estimating Population Prevalence

- We calculate estimates for key outcomes (e.g., prevalence of diabetes) using weighted data to ensure valid, representative conclusions.

```r
svymean(x = ~dm_dx,
          design = wtds_dsg,
          na.rm = T)
```

```
         mean     SE
dm_dx 0.50992 0.0373
```

Variance Estimation:

- Variance is estimated using complex sampling design techniques to ensure accurate confidence intervals for population estimates.

```r
svyciprop(formula = ~dm_dx,
          design = wtds_dsg) %>%
  attr(., "ci")
```

```
     2.5%     97.5%
0.4147657 0.6043703
```

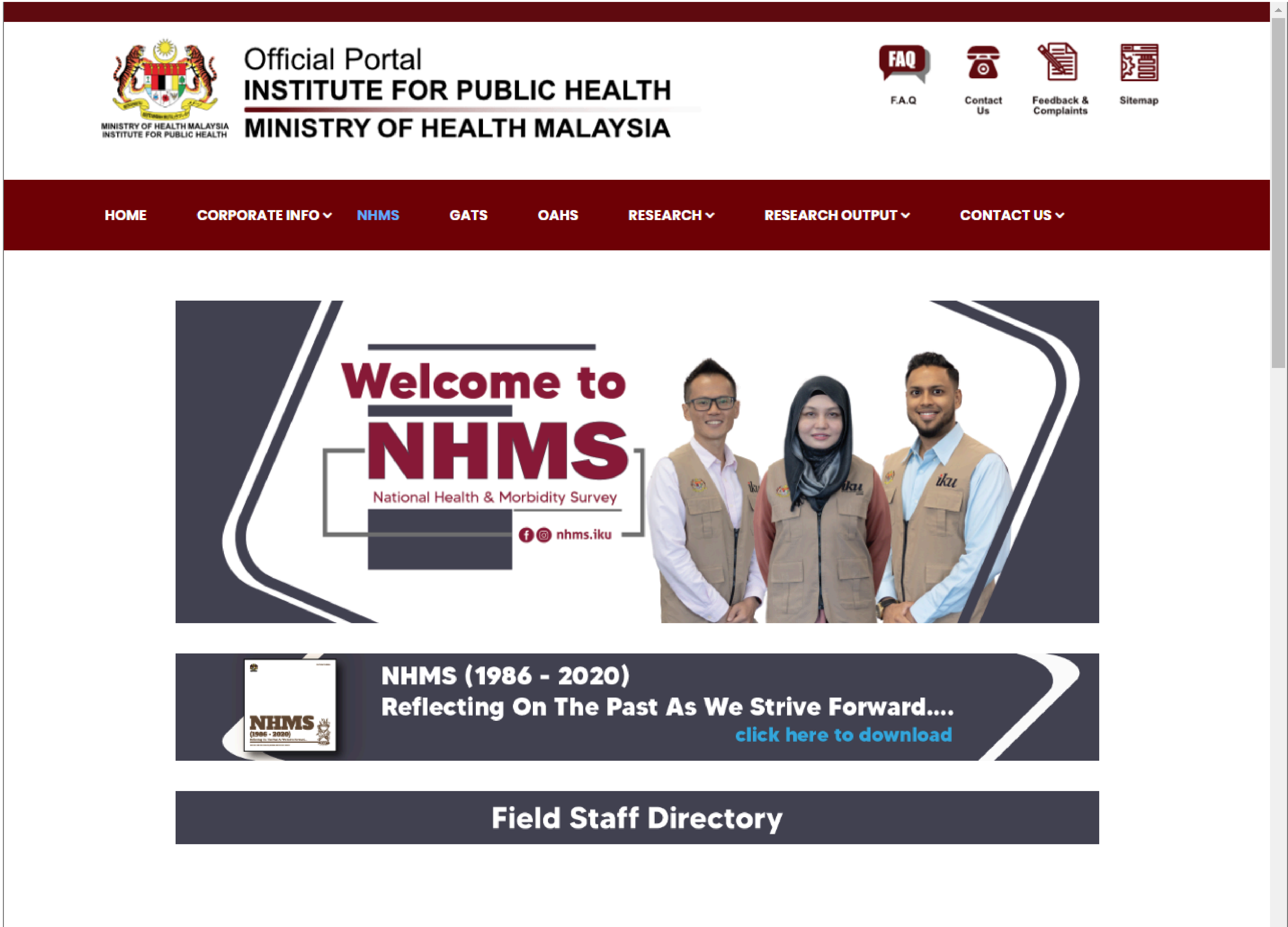
Subgroup Analysis

- For subpopulation analysis, we can use `svyby( )` function

```
svyby(formula = ~dm_dx,
      by = ~gender,
      design = wtds_dsg,
      FUN = svymean,
      na.rm = T)
```

```
       gender     dm_dx         se
Female Female 0.5350983 0.03554017
Male     Male 0.4739481 0.05523103
```

Thank you



Official Portal
INSTITUTE FOR PUBLIC HEALTH
MINISTRY OF HEALTH MALAYSIA
MINISTRY OF HEALTH MALAYSIA INSTITUTE FOR PUBLIC HEALTH
F.A.Q
Contact Us
Feedback & Complaints
Sitemap
HOME
CORPORATE INFO
NHMS
GATS
OAHS
RESEARCH
RESEARCH OUTPUT
CONTACT US
Welcome to
NHMS
National Health & Morbidity Survey
nhms.iku
NHMS (1986 - 2020)
Reflecting On The Past As We Strive Forward....
click here to download
Field Staff Directory