Proof of concept for health indicators

Julian Flowers

2024-07-12

Table of contents

# 1. Introduction

Outline an end-to-end process for creating public health indicators and generating public health profiles.

## 1.1 Global workflow

|  |
| --- |
| Figure 1.1: Workflow |

# 2. Pre-requisites

Using this resource requires:

* R
* RStudio
* quarto - download from <https://quarto.org/docs/download/>
* R packages
  + needs
  + tidyverse
  + quarto
  + sf
  + mapview
  + data.table
  + flextable
  + ggalluvial
  + GGally
  + myScrapers
  + epitools
  + PHEindicatormethods
* Install R packages - copy this code and run at the > prompt in the console
* ```
* install.packages(c(“needs”, “tidyverse”, “quarto”, “sf” , “mapview”, “ggalluvial” , “GGally”, “devtools”, “flextable”, “data.table”, “epitools”, “PHEindicatormethods”))
* library(needs, devtools)
* devtools::install\_github(“julianflowers/myScrapers”, dependencies = TRUE)
* library(myScrapers)
* ```

# 3. Indicator definitions

These are as extracted from the definitions as agreed for the proof of concept

## 3.1 Anti-microbial resistance

Percentage of bloodstream infections due to selected antimicrobial-resistant organisms (MRSA, Escherichia coli resistant to 3rd-generation cephalosporin (e.g., ESBL- E. coli))

### 3.1.1 Staphylococcus aureus

| title | denominator | numerator | metric | notes |
| --- | --- | --- | --- | --- |
| Percentage of bloodstream infections due to selected antimicrobial-resistant organisms: MRSA | Total No. of patients with growth of S. aureus in tested blood samples | No. of patients with growth of methicillin-resistant S. aureus in tested blood samples | Proportion |  |

### 3.1.2 Escherichia coli

| title | denominator | numerator | metric | notes |
| --- | --- | --- | --- | --- |
| Percentage of bloodstream infections due to selected antimicrobial-resistant organisms: E.coli | Total No. of patients with growth of E. coli in tested blood samples | No. of patients with growth of E.coli resistant to 3rd generation cephalosporins in tested blood samples | Proportion | Resistance to ceftazidine, cefotaxine |

## 3.2 Smoking

Current cigarette smoking among women aged 18-44 years

| title | denominator | numerator | metric | notes |
| --- | --- | --- | --- | --- |
| PCurrent cigarette smoking among women aged 18-44 years | Women aged 18-44 years who information about cigarette smoking (excluding unknowns and refusals). | Women aged 18-44 years who reported that they smoked ≥100 cigarettes in their lifetime and currently smoke every day or some days. | Rate with 95% confidence intervals | Data provided is attendances at smoking clinics rather than survey based smoking prevalence |

## 3.3 Injury

Nonfatal hospitalizations for all injuries

| title | denominator | numerator | metric | notes |
| --- | --- | --- | --- | --- |
| Nonfatal hospitalizations for all injuries | Midyear population for the calendar year under surveillance obtained from the KSA Census or suitable alternative. | Nonfatal hospitalizations with any of the following ICD-10-CM diagnostic codes in the principal diagnosis field | Rate (age standardised) with 95% confidence intervals | Data provided is attendances at smoking clinics rather than survey based smoking prevalence |

## 3.4 Flu vaccination coverage

### 3.4.1 Paediatric

| title | denominator | numerator | metric | notes |
| --- | --- | --- | --- | --- |
| Population vaccination coverage - Flu (pediatric group 6 months - 18 years old)) | The total number of children eligible for influenza vaccination | The total number of children in the respective eligible age cohort that have received at least one dose of influenza vaccine from 1 Oct in school, pharmacy, and general practice. | Proportion with 95% confidence intervals | Data provided is attendances at smoking clinics rather than survey based smoking prevalence |

### 3.4.2 At-risk groups

The dataset required for calculating coverage for at-risk groups is much more complex than for age-specific coverage rates because it needs details of underlying diseases e.g. asthma, COPD.

### 3.4.3 65+

| title | denominator | numerator | metric | notes |
| --- | --- | --- | --- | --- |
| Population vaccination coverage - Flu (65+)) | The total number of adults 65+ eligible for influenza vaccination | The total number of children in the respective eligible age cohort | Proportion with 95% confidence intervals | Data provided is attendances at smoking clinics rather than survey based smoking prevalence |

# 4. Data summaries

## 4.1 AMR

The dummy data is per patient sample based.

### 4.1.1 E. coli

|  |
| --- |
| Table 4.1: E. coli resistance rates to 3rd generation cephalosporins |

| age\_band | resistant | sensitive | total | sum\_res | sum\_tot | mean\_rate | value | lowercl | uppercl | confidence | method |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| [0,5) | 8 | 6 | 14 | 83 | 152 | 0.546 | 0.571 | 0.3259 | 0.786 | 95% | Wilson |
| [10,15) | 1 |  |  | 83 | 152 | 0.546 |  |  |  | 95% | Wilson |
| [15,20) | 1 | 2 | 3 | 83 | 152 | 0.546 | 0.333 | 0.0615 | 0.792 | 95% | Wilson |
| [20,25) | 2 | 1 | 3 | 83 | 152 | 0.546 | 0.667 | 0.2077 | 0.939 | 95% | Wilson |
| [25,30) | 1 |  |  | 83 | 152 | 0.546 |  |  |  | 95% | Wilson |
| [30,35) | 3 | 3 | 6 | 83 | 152 | 0.546 | 0.500 | 0.1876 | 0.812 | 95% | Wilson |
| [35,40) | 2 | 4 | 6 | 83 | 152 | 0.546 | 0.333 | 0.0968 | 0.700 | 95% | Wilson |
| [40,45) | 1 | 3 | 4 | 83 | 152 | 0.546 | 0.250 | 0.0456 | 0.699 | 95% | Wilson |
| [45,50) | 1 | 2 | 3 | 83 | 152 | 0.546 | 0.333 | 0.0615 | 0.792 | 95% | Wilson |
| [50,55) | 7 | 3 | 10 | 83 | 152 | 0.546 | 0.700 | 0.3968 | 0.892 | 95% | Wilson |
| [55,60) | 9 | 6 | 15 | 83 | 152 | 0.546 | 0.600 | 0.3575 | 0.802 | 95% | Wilson |
| [60,65) | 11 | 9 | 20 | 83 | 152 | 0.546 | 0.550 | 0.3421 | 0.742 | 95% | Wilson |
| [65,70) | 9 | 7 | 16 | 83 | 152 | 0.546 | 0.562 | 0.3318 | 0.769 | 95% | Wilson |
| [70,75) | 5 | 8 | 13 | 83 | 152 | 0.546 | 0.385 | 0.1771 | 0.645 | 95% | Wilson |
| [75,80) | 5 | 4 | 9 | 83 | 152 | 0.546 | 0.556 | 0.2667 | 0.811 | 95% | Wilson |
| [80,85) | 10 | 6 | 16 | 83 | 152 | 0.546 | 0.625 | 0.3864 | 0.815 | 95% | Wilson |
| [85,90) | 2 | 5 | 7 | 83 | 152 | 0.546 | 0.286 | 0.0822 | 0.641 | 95% | Wilson |
| [90,95) | 5 | 2 | 7 | 83 | 152 | 0.546 | 0.714 | 0.3589 | 0.918 | 95% | Wilson |
| [95,100) |  | 1 |  | 83 | 152 | 0.546 |  |  |  | 95% | Wilson |

|  |
| --- |
| Figure 4.1: e. coli resistance |

### 4.1.2 S.aureus

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 4.2: MRSA   | age\_band | resistant | sensitive | total | sum\_res | sum\_tot | mean\_rate | value | lowercl | uppercl | confidence | method | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | [0,5) | 4 |  |  | 33 | 50 | 0.66 |  |  |  | 95% | Wilson | | [10,15) | 1 | 1 | 2 | 33 | 50 | 0.66 | 0.500 | 0.0945 | 0.905 | 95% | Wilson | | [15,20) |  | 2 |  | 33 | 50 | 0.66 |  |  |  | 95% | Wilson | | [20,25) |  | 1 |  | 33 | 50 | 0.66 |  |  |  | 95% | Wilson | | [25,30) |  | 2 |  | 33 | 50 | 0.66 |  |  |  | 95% | Wilson | | [30,35) | 3 | 3 | 6 | 33 | 50 | 0.66 | 0.500 | 0.1876 | 0.812 | 95% | Wilson | | [35,40) | 4 | 1 | 5 | 33 | 50 | 0.66 | 0.800 | 0.3755 | 0.964 | 95% | Wilson | | [40,45) |  | 1 |  | 33 | 50 | 0.66 |  |  |  | 95% | Wilson | | [45,50) | 1 | 2 | 3 | 33 | 50 | 0.66 | 0.333 | 0.0615 | 0.792 | 95% | Wilson | | [50,55) | 1 | 1 | 2 | 33 | 50 | 0.66 | 0.500 | 0.0945 | 0.905 | 95% | Wilson | | [55,60) | 2 | 6 | 8 | 33 | 50 | 0.66 | 0.250 | 0.0715 | 0.591 | 95% | Wilson | | [60,65) | 6 | 1 | 7 | 33 | 50 | 0.66 | 0.857 | 0.4869 | 0.974 | 95% | Wilson | | [65,70) | 1 | 1 | 2 | 33 | 50 | 0.66 | 0.500 | 0.0945 | 0.905 | 95% | Wilson | | [70,75) | 4 | 1 | 5 | 33 | 50 | 0.66 | 0.800 | 0.3755 | 0.964 | 95% | Wilson | | [75,80) | 1 | 2 | 3 | 33 | 50 | 0.66 | 0.333 | 0.0615 | 0.792 | 95% | Wilson | | [80,85) | 4 | 1 | 5 | 33 | 50 | 0.66 | 0.800 | 0.3755 | 0.964 | 95% | Wilson | | [85,90) | 1 | 1 | 2 | 33 | 50 | 0.66 | 0.500 | 0.0945 | 0.905 | 95% | Wilson | |

|  |
| --- |
| Figure 4.2: Staph aureus resistance rates |

## 4.2 Flu vaccination coverage

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 4.3: Counts of flu vaccination by age group, gender and region   | age\_groups | gender | Al Baha | Al Jawf | Al Qassim | Asir | Hail | Jazan | Makkah Al Mukarramah | Najran | Northern Frontier | Riyadh | Sharqiya | Tabuk | madina | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | 0-18 | female | 38 | 1 | 50 | 161 | 34 | 37 | 282 | 18 | 1 | 361 | 31 | 19 | 140 | | 0-18 | male | 2 |  | 18 | 33 | 6 | 20 | 188 | 2 |  | 122 | 23 | 45 | 91 | | 65+ | female | 1 |  |  | 49 | 3 | 2 | 64 | 3 |  | 19 | 1 | 6 | 24 | | 65+ | male |  |  |  | 10 |  | 4 | 29 | 5 |  | 13 |  | 2 | 14 | | other | female | 69 |  | 157 | 753 | 167 | 105 | 1,182 | 48 | 2 | 1,260 | 102 | 127 | 765 | | other | male | 9 | 1 | 74 | 295 | 37 | 120 | 938 | 50 | 1 | 536 | 199 | 229 | 443 | |

## 4.3 Smoking

Counts of female smoking clinic attendees by directorate for 15+ and 18-44 years

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 4.4: Counts of female attendees at smoking cessation clinics   | age\_1844 | directorate\_name | female | | --- | --- | --- | | 1 | AlAhsa | 73 | | 1 | Asir | 9 | | 1 | Baha | 28 | | 1 | Bisha | 6 | | 1 | Eastern | 551 | | 1 | Hafer AlBatin | 60 | | 1 | Hail | 22 | | 1 | Jazan | 145 | | 1 | Jeddah | 1,569 | | 1 | Jouf | 23 | | 1 | Madinah | 205 | | 1 | Makkah | 578 | | 1 | Najran | 82 | | 1 | Northern Borders | 4 | | 1 | Qassim | 22 | | 1 | Qunfotha | 13 | | 1 | Riyadh | 1,830 | | 1 | Tabuk | 291 | | 1 | Taif | 100 |  | age\_15 | directorate\_name | female | | --- | --- | --- | | 1 | AlAhsa | 93 | | 1 | Asir | 84 | | 1 | Baha | 28 | | 1 | Bisha | 6 | | 1 | Eastern | 579 | | 1 | Hafer AlBatin | 60 | | 1 | Hail | 22 | | 1 | Jazan | 158 | | 1 | Jeddah | 1,788 | | 1 | Jouf | 28 | | 1 | Madinah | 303 | | 1 | Makkah | 636 | | 1 | Najran | 85 | | 1 | Northern Borders | 7 | | 1 | Qassim | 48 | | 1 | Qunfotha | 13 | | 1 | Qurayyat | 4 | | 1 | Riyadh | 1,862 | | 1 | Tabuk | 436 | | 1 | Taif | 136 | |

## 4.4 Injury

|  |
| --- |
| Figure 4.3: Alluvial plot of counts of hospitalisation for injury be age, gender and ICD10 code |

# 5. Population demography

## 5.1 Population pyramids

|  |  |  |
| --- | --- | --- |
| |  | | --- | | (a) Absolute population |  |  | | --- | | (b) Relative population |   Figure 5.1: Population pyramids |

# 6. Data pre-processing

## 6.1 Getting started

These pages set out the process for loading, cleaning, reshaping and recoding provided datasets for calculating indicator values as a proof of concept, and creating a health profile from those indicators.

To facilitate reproducibility the analyses are written in R code in the form of a notebook and R scripts which do much of the pre-processing. The full code is made available as a Github repository at <https://github.com/julianflowers/poc>. The files can be cloned (downloaded)

* In RStudio, go to “File > New Project”
* Click on “Version Control:
* Checkout a project from a version control repository”
* Click on “Git:
* Clone a project from a repository”
* Fill in the info: URL: use HTTPS address
* Create as a subdirectory of: Browse to where you would like to create this folder

### 6.1.1 The indicators

4 sets of indicators are used for this PoC:

* Methicillin resistant Staph aureus and 3rd generation cephalosporin resistant E.coli - proportion of samples tested which are resistant
* Flu vaccination coverage rates
* Smoking rates in women
* Injury admission rates

Fully specifying the indicators

1. The proportion of blood samples which grow either Staph aureaus or E. coli, which are tested for antibiotic sensitivity and which are found to be resistant to oxacllin or 3rd generation cephalosporins respectively for time period X to Y, stratified by [area] / [time period] / [age group] / [gender]
2. The proportion of the population which has been vaccinated against flu for the time period X to Y stratified by [area] / [time period] / [age group] / [gender]
3. The rate of smoking in women aged 15+ / 18-44, per 100,000 female population for the time period X to Y stratified by [area] / [time period] / [age group] / [gender]
4. The rate of hospitalisation for injury for the time period X to Y stratified by [area] / [time period] / [age group] / [gender]

### 6.1.2 Pre-processing script

As a first step we will run a script which does a number of things.

1. Imports the datasets for each indicator
2. Loads KSA 2022 census data by age, gender and region[[1]](#footnote-74)
3. Makes variable names consistent
4. Recodes region names so that they match the names used in the Census 2022 data
5. Maps directorate names (used in smoking data) to region[[2]](#footnote-76)
6. Ensures numeric variables (e.g. age) are converted to numbers
7. Creates a set of intermediate data tables
8. Saves file of reshaped population data,

The intermediate tables can be reused for further analysis and generating profiles.

The script can be run by typing

`source("~/proof-of-concept/scripts/pre-process.R")`

at the prompt in the console

This generates a set of objects in the R environment

objects()

[1] "RETICULATE\_PYTHON"

Objects called dfs reflect pre-processing of the original datasets. Objects containing names, contain the different variables used for age, gender and area variables across datasets. Those called sc are used to map locations of smoking clinics as part of recoding directorates to regions for the smoking data (see below).

The object regional\_counts\_complete is a data frame of regional age\_band, gender specific counts for each indicator and forms the basis of indicator generation. Note, this includes region-age-band-gender combinations for which there is no data (because these combinations are not present in the original data - although they maybe in the full datasets).

regional\_counts\_complete <- read\_csv("data/regional\_counts.csv")  
  
regional\_counts\_complete |>  
 head() |>  
 flextable::flextable()

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | Region | age\_band | sum\_f\_Smoking | sum\_f\_flu | sum\_f\_injury | sum\_m\_Smoking | sum\_m\_flu | sum\_m\_injury | | --- | --- | --- | --- | --- | --- | --- | --- | | `Asir | [0,5) |  | 140 | 160 |  | 26 | 0 | | `Asir | [5,10) | 0 | 7 |  | 1 | 1 |  | | `Asir | [10,15) | 5 | 4 |  | 355 | 2 |  | | `Asir | [15,20) | 39 | 13 |  | 2,744 | 4 |  | | `Asir | [20,25) |  | 79 |  |  | 25 |  | | `Asir | [25,30) | 1 | 76 |  | 21 | 22 |  |   Table 6.1: Region age-band specific counts by gender and indicator |

### 6.1.3 Mapping health directorates to regions for smoking data

The smoking data provided is clinic based data disaggregated at the level of health directorate. There are 20 directorates in KSA, and 13 regions.

Available population data is at regional level, so to generate population denominators for the smoking data I have mapped directorates to regions as follows:

* I used a dataset which contained spatial locations of healthcare facilities (smoking cessation clinics) at health directorate level[[3]](#footnote-79)
* For each location I extracted spatial coordinates (longitude and latitude)
* I obtained a regional boundary file (shape file) from <https://data.humdata.org/dataset/41ce9023-1d21-4549-a485-94316200aba0/resource/99834c81-ad34-415e-91c5-af053d8e55b4/download/sau_capp_adm1_1m_ocha.zip>
* I spatially joined the clinic location and boundary files (see
* This created a lookup table for directorates and regions and enabled the smoking data to be recoded to regions and calculation of rates using the census regional population estimates

sa\_bound <- read\_sf("data/ksa\_bound.gml")  
sc\_ll\_sf <- read\_sf("data/smok\_loc.gml")  
  
sc\_ll\_sf |>  
 ggplot() +  
 geom\_sf(data = sa\_bound) +  
 geom\_sf(aes(colour = name))

|  |
| --- |
| Figure 6.1: Location map of smoking clinics mapped to KSA regional boundaries |

# 7. AMR walkthrough

## 7.1 Introduction

This document outlines a stepwise approach to calculating AMR indicators from dummy data kindly supplied by PHA.

There are x steps

1. EDA (exploratory data analysis of raw data) - this involves cleaning, visualisation and creation of relevant variables.
2. Review of indicator definitions
   * Numerator
   * Denominator
3. Method for calculating numerator and denominator values from dataset. The outline uses R code for reproducibility and flexibility.
4. Calculating indicator values and uncertainty intervals
5. Suggested indicator visualisations (if appropriate).

## 7.2 AMR indicators

### 7.2.1 MRSA

Percentage of bloodstream infection due to methicillin-resistant Staphylococcus aureus (MRSA)

Numerator: No. of patients with growth of methicillin-resistant S. aureus in tested blood samples

Denominator: Total No. of patients with growth of S. aureus in tested blood samples

### 7.2.2 E. coli

Percentage of bloodstream infection due to 3rd-generation cephalosporin resistant E. coli

Numerator: No. of patients with growth of 3rd-generation cephalosporin resistant E. coli in tested blood samples

Denominator: Total No. of patients with growth of E. coli in tested blood samples

### 7.2.3 Import data

df <- amr

334 observations

## 7.3 Data preparation

### 7.3.1 calculate 5-year age bands

amr <- amr[, `:=` (age\_band = cut(age\_year, breaks = seq(0, 100, 5), right = FALSE))][]  
  
head(amr)

record\_number sample\_no patient\_mrn location  
 <num> <char> <char> <char>  
1: 1 ###### ##### Outpatient  
2: 17 ###### ##### Inpatient  
3: 20 ###### ##### Inpatient  
4: 25 ###### ##### Inpatient  
5: 43 ###### ##### Outpatient  
6: 63 ###### ##### Outpatient  
 patient\_hospitalized  
 <char>  
1: Patient had NOT been admitted for more than 2 days in the past 30 days  
2: Patient has been hospitalized for 2 days or less  
3: Patient has been hospitalized for more than 2 days  
4: Patient has been hospitalized for 2 days or less  
5: Patient had NOT been admitted for more than 2 days in the past 30 days  
6: Patient had NOT been admitted for more than 2 days in the past 30 days  
 specific\_location age\_year community\_origin site first\_name second\_name  
 <char> <num> <char> <char> <char> <char>  
1: Emergency Room 0 Community Origin Blood #### #####  
2: Intensive Care Unit 71 Community Origin Blood #### #####  
3: Intensive Care Unit 44 Hospital Origin Blood #### #####  
4: Intensive Care Unit 67 Community Origin Blood #### #####  
5: Emergency Room 67 Community Origin Blood #### #####  
6: Emergency Room 92 Community Origin Blood #### #####  
 family\_name national\_iqama\_id nationality pathogen\_name minocycline  
 <char> <char> <char> <char> <lgcl>  
1: #### ########## ##### Escherichia coli NA  
2: #### ########## ##### Escherichia coli NA  
3: #### ########## ##### Escherichia coli NA  
4: #### ########## ##### Escherichia coli NA  
5: #### ########## ##### Escherichia coli NA  
6: #### ########## ##### Escherichia coli NA  
 tigecycline ampicillin penicillin\_g oxacillin cefoxitin cefotaxime  
 <lgcl> <char> <lgcl> <char> <char> <char>  
1: NA R NA <NA> <NA> R  
2: NA R NA <NA> <NA> NA  
3: NA S NA <NA> <NA> S  
4: NA R NA <NA> <NA> R  
5: NA R NA <NA> <NA> NA  
6: NA R NA <NA> <NA> NA  
 ceftazidime ceftriaxone cefixime cefepime doripenem ertapenem imipenem  
 <char> <char> <lgcl> <char> <char> <char> <char>  
1: R R NA R NA S S  
2: S S NA S NA S S  
3: S S NA S NA S S  
4: R R NA R NA S S  
5: I S NA S NA S S  
6: R R NA R R S S  
 meropenem co\_trimoxazole azithromycin amikacin gentamicin ciprofloxacin  
 <char> <char> <lgcl> <lgcl> <lgcl> <char>  
1: S S NA NA NA S  
2: S S NA NA NA S  
3: S S NA NA NA S  
4: S R NA NA NA S  
5: S S NA NA NA S  
6: S R NA NA NA R  
 levofloxacin colistin spectinomycin age\_band  
 <char> <char> <lgcl> <fctr>  
1: S NA NA [0,5)  
2: S S NA [70,75)  
3: S NA NA [40,45)  
4: S NA NA [65,70)  
5: S S NA [65,70)  
6: R S NA [90,95)

### 7.3.2 remove non-relevant data

This step removes identifiers (names, record IDs)

amr <- amr |> select(-c(family\_name, first\_name, sample\_no, patient\_mrn, second\_name, national\_iqama\_id, nationality))

### 7.3.3 create per test file (long data)

* this create a *per test* dataset rather than a per patient sample dataset

amr\_long <- amr |>  
 pivot\_longer(names\_to = "antibiotic", values\_to = "resistance", cols = minocycline:spectinomycin) |> setDT()

### 7.3.4 Recode 3rd generation cephalosporins

* this step adds a new variable which labels 3rd generation cephalosporins

amr\_long <- amr\_long[, gen\_3 := case\_when(str\_detect(antibiotic, "cef") ~ "3rd-gen", TRUE ~ "other")][]

## 7.4 Data summarisation and description (EDA)

* first generate a high level tabular summary

gtsummary::tbl\_summary(amr)

* represent this visually - we’ll use decomposition trees

amr\_freq <- amr\_long[pathogen\_name == "Escherichia coli", .N, by = .(age\_band, gen\_3, resistance, pathogen\_name, community\_origin)]  
  
collapsibleTreeSummary(amr\_freq,   
 c( "community\_origin", "gen\_3","resistance"),   
 root = "E. coli",   
 nodeSize = "N",   
 attribute = "N",   
 fontSize = 16,   
 collapsed = FALSE)

|  |
| --- |
| Figure 7.1: Decomposition tree for E. coli |

## 7.5 Numerators and denominators

To calculate indicators we need to calculate

* patients with blood stream infection
* samples with antibiotic resistance

amr\_long

Index: <pathogen\_name>  
 record\_number location  
 <num> <char>  
 1: 1 Outpatient  
 2: 1 Outpatient  
 3: 1 Outpatient  
 4: 1 Outpatient  
 5: 1 Outpatient  
 ---   
7678: 1210 Inpatient  
7679: 1210 Inpatient  
7680: 1210 Inpatient  
7681: 1210 Inpatient  
7682: 1210 Inpatient  
 patient\_hospitalized  
 <char>  
 1: Patient had NOT been admitted for more than 2 days in the past 30 days  
 2: Patient had NOT been admitted for more than 2 days in the past 30 days  
 3: Patient had NOT been admitted for more than 2 days in the past 30 days  
 4: Patient had NOT been admitted for more than 2 days in the past 30 days  
 5: Patient had NOT been admitted for more than 2 days in the past 30 days  
 ---   
7678: Patient has been hospitalized for more than 2 days  
7679: Patient has been hospitalized for more than 2 days  
7680: Patient has been hospitalized for more than 2 days  
7681: Patient has been hospitalized for more than 2 days  
7682: Patient has been hospitalized for more than 2 days  
 specific\_location age\_year community\_origin site pathogen\_name  
 <char> <num> <char> <char> <char>  
 1: Emergency Room 0 Community Origin Blood Escherichia coli  
 2: Emergency Room 0 Community Origin Blood Escherichia coli  
 3: Emergency Room 0 Community Origin Blood Escherichia coli  
 4: Emergency Room 0 Community Origin Blood Escherichia coli  
 5: Emergency Room 0 Community Origin Blood Escherichia coli  
 ---   
7678: Non Intensive Unit 96 Hospital Origin Blood Staphylococcus aureus  
7679: Non Intensive Unit 96 Hospital Origin Blood Staphylococcus aureus  
7680: Non Intensive Unit 96 Hospital Origin Blood Staphylococcus aureus  
7681: Non Intensive Unit 96 Hospital Origin Blood Staphylococcus aureus  
7682: Non Intensive Unit 96 Hospital Origin Blood Staphylococcus aureus  
 age\_band antibiotic resistance gen\_3  
 <fctr> <char> <char> <char>  
 1: [0,5) minocycline <NA> other  
 2: [0,5) tigecycline <NA> other  
 3: [0,5) ampicillin R other  
 4: [0,5) penicillin\_g <NA> other  
 5: [0,5) oxacillin <NA> other  
 ---   
7678: [95,100) gentamicin <NA> other  
7679: [95,100) ciprofloxacin R other  
7680: [95,100) levofloxacin R other  
7681: [95,100) colistin NA other  
7682: [95,100) spectinomycin <NA> other

## 7.6 Calculate resistance rates

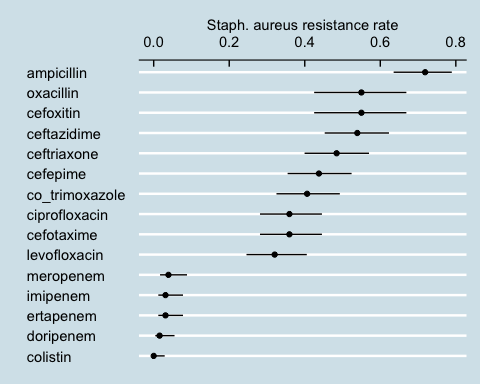
* calculate proportion of tests resistant
* calculate confidence interval (using Wilson’s score method for proportions via the PHEindicatormethods R package)

amr\_long[pathogen\_name == "Escherichia coli" & !is.na(resistance), .N, by = .(resistance, gen\_3)] |>  
 pivot\_wider(names\_from = resistance, values\_from = N) |>  
 rowwise() |>  
 mutate(total\_tests = sum(c\_across(R:I), na.rm = TRUE),   
 resistance\_rate = R / total\_tests) |>  
 flextable::flextable()

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 7.1: E. coli resistance rates to 3rd generation cephalosporins   | gen\_3 | R | NA | S | I | total\_tests | resistance\_rate | | --- | --- | --- | --- | --- | --- | --- | | other | 300 | 193 | 646 | 7 | 1,146 | 0.2617801 | | 3rd-gen | 269 | 56 | 192 | 4 | 521 | 0.5163148 | |

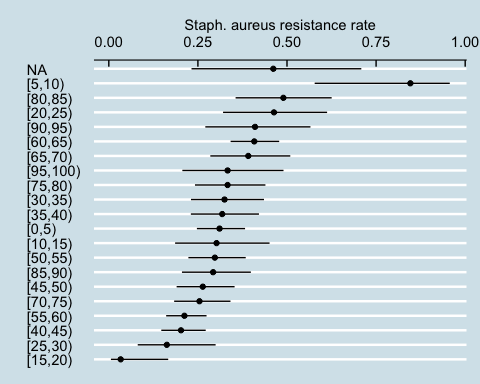
### 7.6.1 by antibiotic

options(digits = 2)  
  
amr\_res\_ci\_sa <- amr\_long[pathogen\_name == "Staphylococcus aureus" & !is.na(resistance), .N, by = .(antibiotic, resistance)] |>  
 pivot\_wider(names\_from = resistance, values\_from = N, values\_fill = 0) |>  
 rowwise() |>  
 mutate(total\_tests = sum(c\_across(S:I), na.rm = TRUE),   
 resistance\_rate = R / total\_tests)  
  
phe\_proportion(amr\_res\_ci\_sa, R, total\_tests) |>  
 bind\_cols(amr\_res\_ci\_sa) |>  
 ggplot() +  
 geom\_point(aes(reorder(antibiotic, value), value)) +  
 geom\_linerange(aes(antibiotic, ymin = lowercl, ymax = uppercl)) +  
 coord\_flip() +  
 labs(y = "Staph. aureus resistance rate", x = "") +   
 scale\_y\_continuous(position = "right")



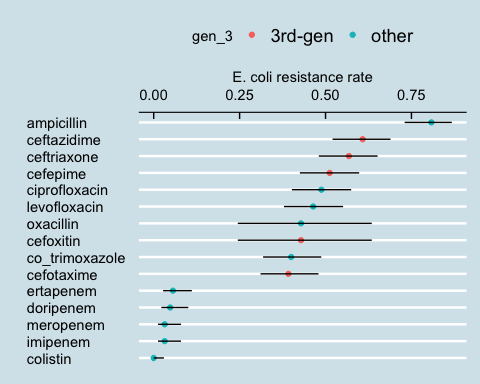
### 7.6.2 by age

amr\_res\_ci\_age <- amr\_long[pathogen\_name == "Staphylococcus aureus" & !is.na(resistance), .N, by = .(age\_band, resistance)] |>  
 pivot\_wider(names\_from = resistance, values\_from = N, values\_fill = 0) |>  
 rowwise() |>  
 mutate(total\_tests = sum(c\_across(S:I), na.rm = TRUE),   
 resistance\_rate = R / total\_tests)  
  
phe\_proportion(amr\_res\_ci\_age, R, total\_tests) |>  
 bind\_cols(amr\_res\_ci\_age) |>  
 ggplot() +  
 geom\_point(aes(reorder(age\_band, value), value)) +  
 geom\_linerange(aes(age\_band, ymin = lowercl, ymax = uppercl)) +  
 coord\_flip() +  
 labs(y = "Staph. aureus resistance rate", x = "") +   
 scale\_y\_continuous(position = "right")



## 7.7 E. coli

amr\_res\_ci\_ec <- amr\_long[str\_detect(pathogen\_name, "coli") & !is.na(resistance), .N, by = .(antibiotic, resistance, gen\_3)] |>  
 pivot\_wider(names\_from = resistance, values\_from = N, values\_fill = 0) |>  
 rowwise() |>  
 mutate(total\_tests = sum(c\_across(R:I), na.rm = TRUE),   
 resistance\_rate = R / total\_tests)  
  
phe\_proportion(amr\_res\_ci\_ec, R, total\_tests) |>  
 bind\_cols(amr\_res\_ci\_ec) |>  
 ggplot() +  
 geom\_point(aes(reorder(antibiotic, value), value, colour = gen\_3)) +  
 geom\_linerange(aes(antibiotic, ymin = lowercl, ymax = uppercl)) +  
 coord\_flip() +  
 labs(y = "E. coli resistance rate", x = "") + scale\_y\_continuous(position = "right")



# 8. Smoking

## 8.1 Workflow

|  |
| --- |
| Figure 8.1: Proposed workflow for calculating smoking rates |

### 8.1.1 Load census population and create age bands

## create age band  
  
pops <- setDT(pops)[, `:=` (`18-44` = dplyr::between(age, 18, 44), `15+` = age >= 15, `80+` = age >= 80, age\_band = cut(age, seq(0, 110, 5), right = FALSE))][]  
  
pop\_age <- pops[, .(n = .N, sumpops = sum(Population, na.rm = TRUE)), by = .(Region, age\_band, Gender)][order(age\_band)]  
  
pop1844F <- pops[, .(n = .N, sumpops = sum(Population, na.rm = TRUE)), by = .(Region, `18-44`, Gender)][`18-44` == "TRUE" & Gender == "Female",]  
  
pop15F <- pops[, .(n = .N, sumpops = sum(Population, na.rm = TRUE)), by = .(Region, `15+`, Gender)][`15+` == "TRUE" & Gender == "Female",]

### 8.1.2 Load smoking data and calculate age bands

Note: it is not clear if the dummy data is a random sample of clinic attendance data.

smok\_age <- smoking[, `:=` (`18-44` = dplyr::between(age, 18, 44), `15+` = age >= 15, `80+` = age >= 80, age\_band = cut(age, seq(0, 110, 5), right = FALSE))][]  
  
smok\_age\_bands <- smok\_age[, .(n = .N, smokers = sum(n, na.rm = TRUE)), by = .(Region, age\_band, Gender)][order(age\_band)]  
  
smok1844F <- smok\_age[, .(n = .N, smokers = sum(n, na.rm = TRUE)), by = .(Region, `18-44`, Gender)][`18-44` == "TRUE" & Gender == "female",]  
  
smok15F <- smok\_age[, .(n = .N, smokers = sum(n, na.rm = TRUE)), by = .(Region, `15+`, Gender)][`15+` == "TRUE" & Gender == "female",]

### 8.1.3 Join datasets and calculate rates

This step uses Byar’s method for confidence interval for rates

smok\_rates <- pop1844F |>  
 left\_join(smok1844F, by = "Region") |>   
 phe\_rate(x = smokers, n = sumpops)   
  
smok\_rates |>  
 dplyr::select(Region, smokers, population = sumpops, value, lowercl, uppercl) |>  
 flextable::flextable()

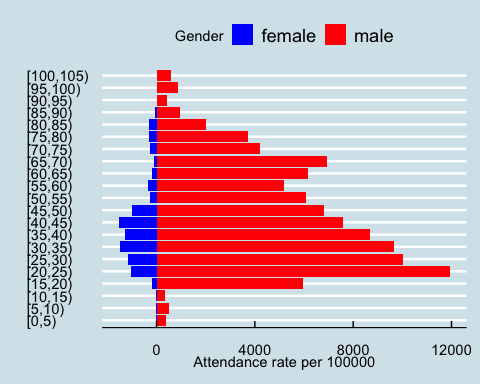
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | Region | smokers | population | value | lowercl | uppercl | | --- | --- | --- | --- | --- | --- | | 'Asir |  | 366,324 |  |  |  | | Al Bahah | 28 | 61,969 | 45.183882 | 30.017729 | 65.30573 | | Al Hudud ash Shamaliyah | 4 | 70,149 | 5.702148 | 1.553643 | 14.59976 | | Al Jawf | 23 | 103,536 | 22.214495 | 14.077587 | 33.33412 | | Al Madinah al Munawwarah |  | 387,944 |  |  |  | | Al Mintaqah ash Sharqiyah |  | 877,403 |  |  |  | | Al Qasim |  | 249,081 |  |  |  | | Ar Riyadh |  | 1,609,493 |  |  |  | | Ha'il | 22 | 136,642 | 16.100467 | 10.086571 | 24.37741 | | Jazan |  | 267,342 |  |  |  | | Makkah al Mukarramah |  | 1,507,301 |  |  |  | | Najran | 82 | 105,915 | 77.420573 | 61.572876 | 96.10050 | | Tabuk | 291 | 163,453 | 178.032829 | 158.162585 | 199.70848 |   Figure 8.2: Smoking clinic attendance rates (18-44) |

smok\_rates |>  
 write\_csv("data/smok\_rates.csv")  
   
smok\_rates |>  
 ggplot() +  
 geom\_col(aes(reorder(Region, value), value)) +  
 geom\_linerange(aes(x = Region, ymin = lowercl, ymax = uppercl)) +  
 coord\_flip() +  
 labs(x = " ",  
 y = "Rate per 100.000")

|  |
| --- |
| Figure 8.3: Smoking clinic attendance rates (18-44) |

### 8.1.4 Age-specific attendance rates

smok\_age\_agg <- smok\_age\_bands[, .(tot\_smokers = sum(smokers)), by = .(age\_band, Gender)][]  
  
options(scipen = 999, digitds = 2)  
  
## m:f as smoking ratios  
  
smok\_rates\_gender <- pop\_age |>  
 dplyr::select(-Region) |>  
 mutate(Gender = recode(Gender, "Female" = "female", "Male" = "male")) |>  
 left\_join(smok\_age\_agg, by = c("age\_band", "Gender")) |>  
 group\_by(age\_band, Gender) |>  
 reframe(nat\_smok = sum(tot\_smokers, na.rm = TRUE),   
 nat\_pop = sum(sumpops)) |>  
 phe\_rate(x = nat\_smok, n = nat\_pop) |>   
 mutate(value = ifelse(Gender == "female", -value, value)) |>  
 ggplot() +  
 geom\_col(aes(age\_band, value, fill = Gender)) +  
 coord\_flip() +  
 scale\_fill\_manual(values = c("blue", "red")) +  
 labs(x = "", y = "Attendance rate per 100000")  
  
smok\_rates\_gender



smok\_gender\_ratio <- pop\_age |>  
 dplyr::select(-Region) |>  
 mutate(Gender = recode(Gender, "Female" = "female", "Male" = "male")) |>  
 left\_join(smok\_age\_agg, by = c("age\_band", "Gender")) |>  
 group\_by(age\_band, Gender) |>  
 reframe(nat\_smok = sum(tot\_smokers, na.rm = TRUE),   
 nat\_pop = sum(sumpops)) |>  
 phe\_rate(x = nat\_smok, n = nat\_pop) |>  
 slice(7:36) |>  
 group\_by(age\_band) |>  
 reframe(ratio = value[2] / value[1]) |>  
 mutate(mean\_ratio = mean(ratio))

The relative attendance rate for males is 17.32 times that of females. Smoking clinic attendance rates could be used as a proxy for population smoking rates if the relationship between attendance and prevalence is known . For example, survey data suggests that KSA male smoking rates are around 20% which implies a female smoking rate based on 1.15%. In fact, available survey data (Algabbani et al. (2018)), and estimates from the Global Burden of Disease ((“Prevalence of Smoking,” n.d.; Xiang et al. 2023)), suggest a smoking prevalence in females over 15 of 2%,

# 9. Flu vaccination coverage

## 9.1 Workflow

|  |
| --- |
| Figure 9.1: Proposed workflow for calculating flu vaccination coverage |

## 9.2 Clean data

Recode region names Add new age bands

flu\_data <- flu\_data |>  
 mutate(paed = dplyr::between(age, 0, 18),   
 older = age >= 65) |>  
 mutate(Region = case\_when(str\_detect(Region, "Baha") ~ "Al Bahah",   
 str\_detect(Region, "Qu?assim") ~ "Al Qasim",  
 str\_detect(Region, "Asir") ~ "`Asir",  
 str\_detect(Region, "Hail") ~ "Ha'il",  
 str\_detect(Region, "[Mm]adin") ~ "Al Madinah al Munawwarah",  
 str\_detect(Region, "Jizan") ~ "Jazan",  
 str\_detect(Region, "Makka") ~ "Makkah al Mukarramah",  
 str\_detect(Region, "Sharqiya" ) ~ "Al Mintaqah ash Sharqiyah",  
 str\_detect(Region, "Northern Frontier") ~ "Al Hudud ash Shamaliyah",  
 str\_detect(Region, "Riyad") ~ "Ar Riyadh",  
 TRUE ~ Region ))

## 9.3 Calculate regional numerators and denominators

* Calculate region-by-gender populations for 0-18 and 65+
* Calculate region-by-gender vaccinations for 0-18 and 65+

pop\_paed <- popns |>  
 group\_by(Region, Gender, paed) |>  
 reframe(pop = sum(Population)) |>  
 filter(paed == "TRUE") |>  
 mutate(Gender = recode(Gender, "Female" = "female", "Male" = "male"),   
 Region = recode(Region, "`Asir" = "'Asir"))  
  
pop\_older <-popns |>  
 group\_by(Region, Gender, older) |>  
 reframe(pop = sum(Population)) |>  
 filter(older == "TRUE") |>  
 mutate(Gender = recode(Gender, "Female" = "female", "Male" = "male"),   
 Region = recode(Region, "`Asir" = "'Asir"))  
  
## fills all region-gender categories and remove unknown  
paed\_flu <- flu\_data |>  
 filter(paed == "TRUE") |>  
 count(Region, Gender) |>  
 complete(Region, Gender) |>  
 filter(Gender != "unknown") |>  
 mutate( Region = recode(Region, "`Asir" = "'Asir"))  
  
older\_flu <- flu\_data |>  
 filter(older == "TRUE") |>  
 count(Region, Gender) |>  
 complete(Region, Gender) |>  
 filter(Gender != "unknown") |>  
 mutate( Region = recode(Region, "`Asir" = "'Asir"))

## 9.4 Join populations and vaccination data

paed\_nd <- paed\_flu |>  
 left\_join(pop\_paed) |>  
 select(-paed)  
  
older\_nd <- older\_flu |>  
 left\_join(pop\_older) |>  
 select(-older)

## 9.5 Calculate coverage by age group

paed\_coverage <- paed\_nd |>  
 phe\_proportion(x = n, n = pop) |>  
 mutate(age = "0-18")   
  
older\_coverage <- older\_nd |>  
 phe\_proportion(x = n, n = pop) |>  
 mutate(age = "65+")  
  
bind\_rows(paed\_coverage, older\_coverage) |>  
 write\_csv("data/flu\_coverage.csv")

## 9.6 Tabulate

options(digits = 3)  
  
paed\_coverage |>  
 select(-c(statistic, method)) |>  
 head() |>  
 flextable()  
older\_coverage |>  
 select(-c(statistic, method)) |>  
 head() |>  
 flextable()

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 9.1: Flu vaccination coverage   | Region | Gender | n | pop | value | lowercl | uppercl | confidence | age | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | 'Asir | female | 161 | 335,011 | 0.0004806 | 0.00041187 | 0.0005607 | 95% | 0-18 | | 'Asir | male | 33 | 347,751 | 0.0000949 | 0.00006757 | 0.0001333 | 95% | 0-18 | | Al Bahah | female | 38 | 57,230 | 0.0006640 | 0.00048382 | 0.0009112 | 95% | 0-18 | | Al Bahah | male | 2 | 59,077 | 0.0000339 | 0.00000928 | 0.0001234 | 95% | 0-18 | | Al Hudud ash Shamaliyah | female | 1 | 70,350 | 0.0000142 | 0.00000251 | 0.0000805 | 95% | 0-18 | | Al Hudud ash Shamaliyah | male |  | 73,068 |  |  |  | 95% | 0-18 |  | Region | Gender | n | pop | value | lowercl | uppercl | confidence | age | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | 'Asir | female | 49 | 32,224 | 0.001521 | 0.0011505 | 0.002010 | 95% | 65+ | | 'Asir | male | 10 | 36,288 | 0.000276 | 0.0001497 | 0.000507 | 95% | 65+ | | Al Bahah | female | 1 | 7,791 | 0.000128 | 0.0000227 | 0.000727 | 95% | 65+ | | Al Bahah | male |  | 6,752 |  |  |  | 95% | 65+ | | Al Madinah al Munawwarah | female | 24 | 31,099 | 0.000772 | 0.0005187 | 0.001148 | 95% | 65+ | | Al Madinah al Munawwarah | male | 14 | 31,231 | 0.000448 | 0.0002671 | 0.000752 | 95% | 65+ | |

## 9.7 Visualise

paed\_coverage |>  
 ggplot() +  
 geom\_col(aes(Region, value, fill = Gender)) +  
 geom\_linerange(aes(Region, ymin = lowercl, ymax = uppercl)) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1) ) +  
 facet\_wrap(~ Gender) +  
 scale\_fill\_manual(values = c("blue", "red"))  
  
  
older\_coverage |>  
 ggplot() +  
 geom\_col(aes(Region, value, fill = Gender)) +  
 geom\_linerange(aes(Region, ymin = lowercl, ymax = uppercl)) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1) ) +  
 facet\_wrap(~ Gender) +  
 scale\_fill\_manual(values = c("blue", "red"))

|  |  |  |
| --- | --- | --- |
| |  | | --- | | (a) Paediatric coverage |  |  | | --- | | (b) 65+ coverage |   Figure 9.2: Flu vaccination coverage |

# 10. Injury

## 10.1 Injury

First, load injury and population data

age code diagnosis Gender   
 Min. : 1 Length:10380 Length:10380 Length:10380   
 1st Qu.:27 Class :character Class :character Class :character   
 Median :38 Mode :character Mode :character Mode :character   
 Mean :39   
 3rd Qu.:53   
 Max. :84   
 NA's :773   
 Region   
 Length:10380   
 Class :character   
 Mode :character

|  |
| --- |
| Figure 10.1: Age distribution of injury admissions |

Overall there are 10380 records, of which 773 (7.45% ) have missing dates of of birth. These are excluded from analyis

The data contains values for 12 of the 13 regions.

## 10.2 Aggregate injury data

| Region | diagnosis | code | age\_band | Female | Male |
| --- | --- | --- | --- | --- | --- |
| Makkah | Open wound of head | S01 | [25,30) |  | 20 |
| Northern Frontier | Open wound of head | S01 | [20,25) | 15 |  |
| Northern Frontier | Open wound of head | S01 | [25,30) | 10 |  |
| Northern Frontier | Open wound of head | S01 | [30,35) | 16 | 6 |
| Northern Frontier | Open wound of head | S01 | [35,40) |  | 2 |
| Northern Frontier | Open wound of head | S01 | [40,45) | 8 | 2 |

## 10.3 Recode region names to match census names

[1] "Al Jawf" "Jazan" "Najran"

[1] "0-4" "5-9" "10-14" "15-19" "20-24" "25-29" "30-34" "35-39" "40-44"  
[10] "45-49" "50-54" "55-59" "60-64" "65-69" "70-74" "75-79" "80+"

[1] [40,45) [70,75) [35,40) [65,70) [15,20) [80,85) [55,60) [25,30) [30,35)  
[10] [60,65) [50,55) [0,5) [45,50) [10,15) [20,25) [75,80)  
24 Levels: [0,5) [5,10) [10,15) [15,20) [20,25) [25,30) [30,35) ... [115,120)

[1] 0-4 5-9 10-14 15-19 20-24 25-29 30-34 35-39 40-44 45-49 50-54 55-59  
[13] 60-64 65-69 70-74 75-79 80+   
17 Levels: 0-4 5-9 10-14 15-19 20-24 25-29 30-34 35-39 40-44 45-49 ... 80+

## 10.4 Calculate directly age standardised rates for injury admission by region and gender

Use epitools::ageadjust.direct.

To iterate over regions and gender we’ll split by gender and use the nest\_by function to group data by region and apply standardiastion function.

|  |
| --- |
| Figure 10.2: Age standardised injury admission rate per 100,000 by region and gender |

# 11. Health profile (1)

## 11.1 Putting it all together

Health profiles combine multiple indicators to allow comparison, usually between population or administrative units e.g. regions, hospitals, ethnic groups, age groups and provide tools to facilitate a range of comparison.

### 11.1.1 Workflow

#### 11.1.1.1 Preparing data

To generate profile-style presentations per-indicator datasets need to be configured in a standard way and pooled. The minimum meta-data (variable names) for each indicator would be:

Meta-data for profile data

| Indicator name |
| --- |
| Area name |
| Age group |
| Gender |
| Value |
| Count (numerator) |
| Denominator |
| Metric type e.g. proportion, rate |
| Unit e.g. per 100,000, % |
| Lower uncertainty interval |
| Upper uncertainty interval |
| Uncertainty e.g. 95%, 99.8% |

|  |
| --- |
| Figure 11.1: Proposed workflow for creating health profiles |

UK health profiles (for example, Public Health Outcome Framework) offer a wide range of comparison:

* Between and within areas
* Between age groups and genders

## 11.2 Visualisation

| Comparison | Visualisation | Comment |
| --- | --- | --- |
| Single area, multiple indicators | Spine or bullet chart | Indicator values are scaled so that they can be directly compared. Single area (unit) values are usually encoded as a dot or line, plotted on a bar representing the range of values for all areas / units. |
| Multiple areas, single indicator | Column / bar chart | Plots can be ordered and plotted horizontally to enable rapid appraisal of relative performance. Confidence intervals can be added |
| Indicator pairs / multiple indicators | Scatter plot or scatter plot matrices | May include smooths (lines representing fit to the data - common options are linear, loess and generalised additive models) |
| Spatial variation | Choropleth maps | Requires boundary files and GIS tools |
| Overview | Heatmap / tile chart | Colour cells with statistical significance categories (e.g. high / low) |
|  |  |  |

We will take the toolkit for plotting provided by the UKHSA which generates the charts used in the online version if the PHOF as a basis for plotting profiles. This is available at https://github.com/ukhsa-collaboration/fingertipscharts.

Note: in the PHOF statistical significance is defined by whether the global average value (e.g. KSA) is contained with the confidence limits of the unit values (e.g. ’Asir region). It does not calculate limits for the global value. The rationale for this is explained in Public Health Data Science (2018)

## install fingertipscharts  
##   
  
if(!require("fingertipscharts")) devtools::install\_github("https://github.com/ukhsa-collaboration/fingertipscharts", force = TRUE)  
  
library(fingertipscharts)

To use fingertipscharts data has to be in the right format.

Bar charts can be created using the compare\_areas function

smok\_rates |>  
 compare\_areas(area = region, value = value, lowerci = lowercl, upperci = uppercl, title = "Rate of attendance at smoking clinics 18-44: F")  
  
flu\_coverage |>  
 filter(gender == "female" & age == "0-18") |>  
 compare\_areas(area = region, value = value, lowerci = lowercl, upperci = uppercl, fill = gender, title = "Flu vaccination coverage 0-18; F")

|  |  |  |
| --- | --- | --- |
| |  | | --- | | (a) Smoking rates |  |  | | --- | | (b) Paediatric flu coverage |   Figure 11.2: Using `compare-areas |

## 11.3 Data overview

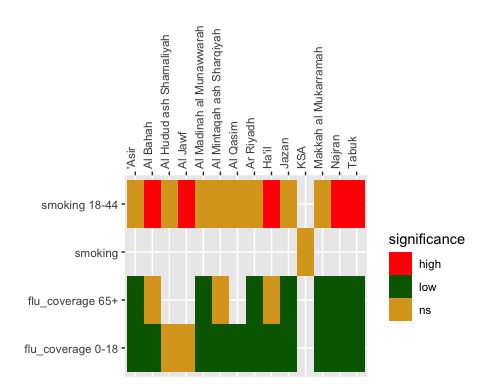
PHOF uses a heatmap (known as a “tartan rug” to display area by indicator summaries. The term tartan rug is used to refer to the colour scheme (red-amber-green) to denote statistical significance. The

prof\_data <- bind\_rows(smok\_rates, flu\_coverage) |>  
 select(-c(11:13))   
  
## calculate ksa values  
ksa\_values <- prof\_data |>  
 group\_by(indicator) |>  
 reframe(ksa\_values = 100000 \* sum(n, na.rm = TRUE) / sum(pop))  
  
## join to data  
prof\_data <- prof\_data |>  
 left\_join(ksa\_values)  
  
## code for statistical significance  
  
prof\_data <- prof\_data |>  
 mutate(significance = case\_when(lowercl > ksa\_values ~"high" ,  
 uppercl < ksa\_values~ "low",  
 TRUE ~ "ns"))  
  
head(prof\_data) |>  
 gt::gt()

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 11.1: Required data format example   | indicator | region | age | gender | n | pop | value | lowercl | uppercl | timeperiod | ksa\_values | significance | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | smoking 18-44 | 'Asir | 18-44 | female | NA | 366324 | NA | NA | NA | 2023 | 7.618658 | ns | | smoking 18-44 | Al Bahah | 18-44 | female | 28 | 61969 | 45.183882 | 30.017729 | 65.30573 | 2023 | 7.618658 | high | | smoking 18-44 | Al Hudud ash Shamaliyah | 18-44 | female | 4 | 70149 | 5.702148 | 1.553643 | 14.59976 | 2023 | 7.618658 | ns | | smoking 18-44 | Al Jawf | 18-44 | female | 23 | 103536 | 22.214495 | 14.077587 | 33.33412 | 2023 | 7.618658 | high | | smoking 18-44 | Al Madinah al Munawwarah | 18-44 | female | NA | 387944 | NA | NA | NA | 2023 | 7.618658 | ns | | smoking 18-44 | Al Mintaqah ash Sharqiyah | 18-44 | female | NA | 877403 | NA | NA | NA | 2023 | 7.618658 | ns | |

### 11.3.1 Heatmap

prof\_data |>  
 ggplot() +  
 geom\_tile(aes(region, indicator, fill = significance)) +  
 scale\_fill\_manual(values = c("red", "darkgreen", "goldenrod")) +  
 scale\_x\_discrete(position = "top") +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 0)) +  
 labs(y = "", x = "")



### 11.3.2

# 12. Health profile (2)

Maps

### 12.0.1 Boundary file

A regional boundary file is downloaded as part of the data pre-processing step and is contained in the object sa\_bound.

This can be plotted as a static or interactive map:

ksa\_bound |>  
 ggplot() +  
 geom\_sf()  
  
mapview(ksa\_bound)

|  |  |  |
| --- | --- | --- |
| |  | | --- | | (a) static |  |  | | --- | | (b) interactive |   Figure 12.1: KSA boundary map |

### 12.0.2 Plot data

Attribute data (e..g area rates) needs to be combined with the boundary data in order to map them. Taking the example of injury….

inj <- read\_csv("data/injury\_dsr.csv")

### 12.0.3 Rename regions to link to spatial file

setdiff(inj$region, ksa\_bound$ADM1\_EN)

[1] "'Asir" "Al Madinah al Munawwarah"   
[3] "Al Mintaqah ash Sharqiyah" "Al Qasim"   
[5] "Ar Riyadh" "Jazan"   
[7] "Makkah al Mukarramah"

ksa\_bound$ADM1\_EN

[1] "`Asir" "Al Bahah"   
 [3] "Al Hudud ash Shamaliyah" "Al Jawf"   
 [5] "Al Madinah" "Al Quassim"   
 [7] "Ar Riyad" "Ash Sharqiyah"   
 [9] "Ha'il" "Jizan"   
[11] "Makkah" "Najran"   
[13] "Tabuk"

## recode area names and join to boundary data  
inj\_sf <- inj |>  
 mutate(region1 = recode(region,   
 "'Asir" = "`Asir",   
 "Al Madinah al Munawwarah" = "Al Madinah",  
 "Al Qasim" = "Al Quassim",   
 "Makkah al Mukarramah" = "Makkah",   
 "Al Mintaqah ash Sharqiyah" = "Ash Sharqiyah",   
 "Ar Riyadh" = "Ar Riyad",   
 "Jazan" = "Jizan"),   
 region1 = ifelse(region1 != region, region1, region)  
 )  
  
inj\_sf <- left\_join(ksa\_bound, inj\_sf, by = c("ADM1\_EN" = "region1"))   
  
inj\_sf |>  
 filter(is.na(gender))

Simple feature collection with 1 feature and 19 fields  
Geometry type: MULTIPOLYGON  
Dimension: XY  
Bounding box: xmin: 34.4943 ymin: 24.53653 xmax: 40.17387 ymax: 28.98889  
Geodetic CRS: WGS 84  
# A tibble: 1 × 20  
 gml\_id Shape\_Leng Shape\_Area ADM1\_EN ADM1\_PCODE ADM1\_REF ADM1ALT1EN ADM1ALT2EN  
\* <chr> <dbl> <dbl> <chr> <chr> <chr> <chr> <chr>   
1 ksa\_b… 34.6 9.05 Tabuk SA07 <NA> <NA> <NA>   
# ℹ 12 more variables: ADM0\_EN <chr>, ADM0\_PCODE <chr>, date <date>,  
# validOn <date>, validTo <date>, geometryProperty <MULTIPOLYGON [°]>,  
# region <chr>, crude.rate <dbl>, adj.rate <dbl>, lci <dbl>, uci <dbl>,  
# gender <chr>

### 12.0.4 Plot smoking rates as choropleth map

## convert back to spatial data and plot  
inj\_sf |> st\_as\_sf() |>  
 filter(!is.na(gender)) |>  
 ggplot() +  
 geom\_sf(data = ksa\_bound) +  
 geom\_sf(aes(fill = 100000 \* adj.rate)) +  
 ggplot2::geom\_sf\_label(aes(label = ADM1\_EN)) +  
 facet\_wrap(~ gender) +  
 theme\_void() +  
 scale\_fill\_viridis\_c(name = "DSR Injury\nper 100,000")  
  
mapview(inj\_sf |> st\_as\_sf() |>  
 filter(gender == "female") |> mutate(adj.rate = 100000 \* adj.rate), zcol = "adj.rate")

|  |  |  |
| --- | --- | --- |
| |  | | --- | | (a) Static |  |  | | --- | | (b) Interactive |   Figure 12.2: Smoking rate map |

# References

Algabbani, Aljoharah Mohammed, Rasha Almubark, Nora Althumiri, Amani Alqahtani, and Nasser BinDhim. 2018. “The Prevalence of Cigarette Smoking in Saudi Arabia in 2018.” *Food and Drug Regulatory Science Journal* 1 (1): 1. <https://doi.org/10.32868/rsj.v1i1.22>.

“Prevalence of Smoking.” n.d. <https://doi.org/10.1787/301887638251>.

Public Health Data Science. 2018. “PHDS Guidance - Confidence Intervals - Public Health Profiles.” Public Health England.

Xiang, Hong, Deshi Dong, Linlin Lv, and Xufeng Tao. 2023. “Global Burden of Disease Study 2019 Indicates That Smoking Gradually Becomes a Key Driver of the Burden of Pancreatic Cancer in Developing Regions.” In. IntechOpen. <https://doi.org/10.5772/intechopen.1003616>.

1. Note. The census data was downloaded from <https://tableau.saudicensus.sa/#/views/TA3-PopulationbydetailedAgebyRegionGovernorateNationalityandGenderAR_16850208449070/PopulationbydetailedAgebyRegionGovernorateNationalityandGenderARCSV.csv> and variable and region names translated to English using ChatGPT4o [↑](#footnote-ref-74)
2. This uses the spatial locations of smoking clinics which include directorate names to map to KSA regional boundaries [↑](#footnote-ref-76)
3. [https://www.moh.gov.sa/en/Ministry/Projects/TCP/Pages/default.aspx](https://www.moh.gov.sa/en/Ministry/Projects/TCP/Pages/default.aspx%22) [↑](#footnote-ref-79)