WEBINARIO

Introducción a R para ciencia de datos en salud

25 oct 2023, 8:00 (Brasília), 12:00 (Londres), 13:00 (Ciudad del Cabo), 17:00 (Dhaka)

MASHA

Ministério da Saúde
FIOCRUZ
Fundação Oswaldo Cruz

HDRGlobal
Health Data Research

icddr,b

THE GLOBAL HEALTH NETWORK
Enabling research by sharing knowledge

Housekeeping

- Firstly, this session is being recorded. The recording will be shared in the coming weeks on The Global Health Network platform.
- Due to the number of participants your microphones have been disabled.
- Please use the Chat function to introduce yourself or to report any technical issues that you may be experiencing.
- Please use the Q&A function (located in the toolbar at the bottom of the Zoom window) to post your comments or questions.
- Simultaneous translation will be provided into Spanish and Portuguese and English. Navigate to the toolbar, click on Language Interpretation and select your desired language input.

Agenda

12.00-12.10 - Welcome

12.10-12.30 - Overview of R programming language and its use in research, Miss Aashna Uppal

- Benefits and possibilities for using R for health research projects
- Live demonstration of R and RStudio

12.30-13.15 - Presentations from health data science project teams that are using R

- Analysis of stunting in Bangladesh, a case study of presenting findings in R. **Mr Md. Sojibul Islam**
- Using R to support data preparation and visualisation for a research study in Brazil. **Dr Soraida Aguilar**
- User-Centred Dashboards for COVID-19 Trends in Africa. **Dr Frank Kagoro**

13.15-13.30 - Question and Answers

Spotlight on R

The Global Health Data Science community hub has developed **Spotlight on: R for Health Data Research** which brings together freely available and helpful educational materials tailored to beginners in R for health data science.

This resource covers fundamental R concepts, data manipulation, analysis techniques and data visualisation, along with specialised packages and techniques employed in health research.

It is aimed at students, researchers, health care professionals or anyone who is interested in learning R programming.

Webinar: Getting Started with R for Health Data Science is a companion session to provide an opportunity for attendees to learn more about R through instructional presentations and case study examples

Miss Aashna Uppal

DPhil Student, The Global Health Network, Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford

Presentation

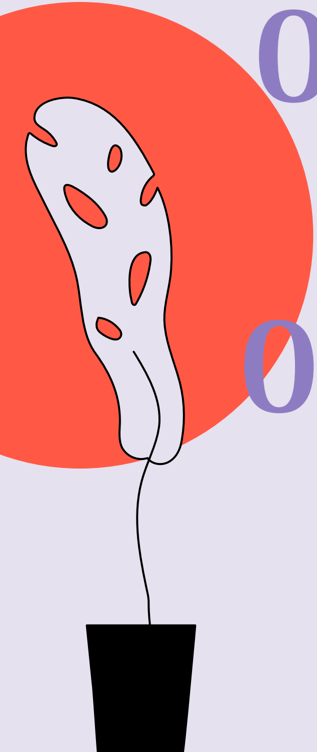
- Overview of the R programming language and its use in research
- Live demonstration of R and RStudio

Overview of R and its use in research

Aashna Uppal

THE GLOBAL HEALTH NETWORK
Enabling research by sharing knowledge

Table of contents



01

Introduction

What is R? What is RStudio?

R & RStudio

What is R & RStudio?

- R is a programming language
- RStudio is an Graphical User Interface (GUI), which is a fancy way of saying that you use RStudio to write code
- Think of it this way: **R is the writing, RStudio is the notebook**
- R is very powerful for statistical analysis and epidemiology, and it's free to use!

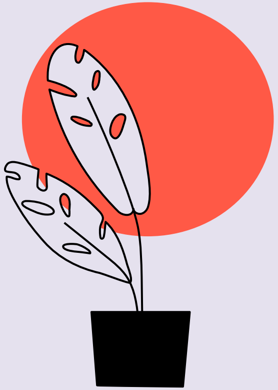
02

R versus

How does R compare to other languages/ software?

What are languages and software?

Languages are used for programming or creating software. **Software** are tools that help perform tasks or operations on a computer.



Common languages/software

	R	Python	SAS	SPSS
Free & Open Source	✓	✓	✗	✗
Point & Click	✗	✗	✓	✓
Customised Graphics	✓	✓	✗	✗
Packages	✓	✓	✗	✗

R versus Python

Which one to use?

- **R**
  - Statistics, epidemiology
  - Popular in research and academia
- **Python**
  - Machine learning, AI
  - Popular in web and software development

03

Possibilities

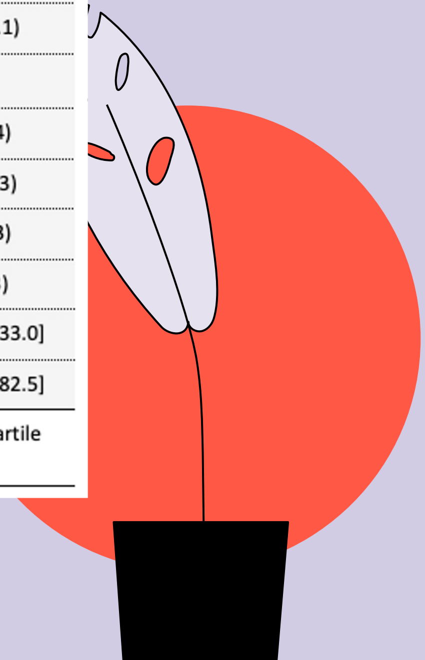
What kinds of outputs are possible with R?

https://www.r-bloggers.com/2021/11/publication-ready-tables-with-flextable-and-own-theme-in-r/

Table 1. Baseline characteristics of 686 patients enrolled in the German Breast Cancer Study Group between 1984 and 1989

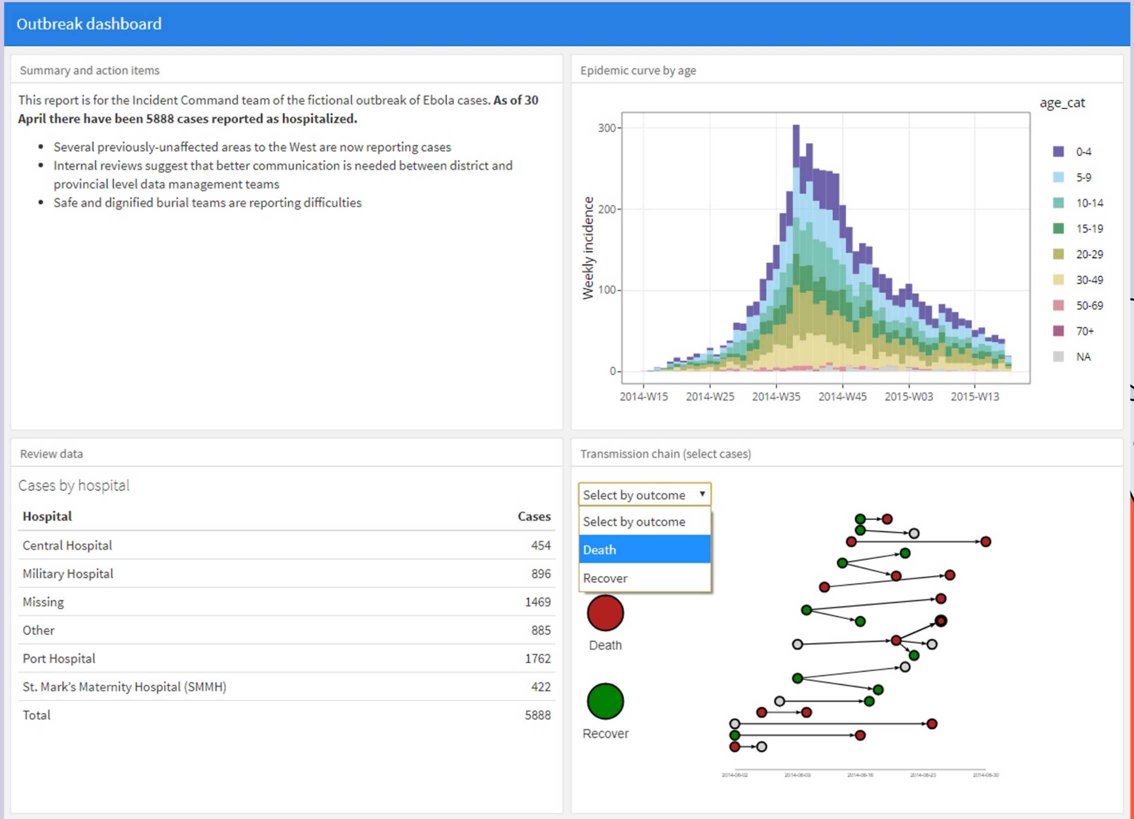
Variable	Overall	Placebo	Treated
No.	686	440	246
Age, years (mean (SD))	53.1 (10.1)	51.1 (10.0)	56.6 (9.4)
Postmenopausal	396 (57.7)	209 (47.5)	187 (76.0)
Tumor size, mm (mean (SD)	29.3 (14.3)	29.6 (14.4)	28.8 (14.1)
Tumor grade			
1	81 (11.8)	48 (10.9)	33 (13.4)
2	444 (64.7)	281 (63.9)	163 (66.3)
3	161 (23.5)	111 (25.2)	50 (20.3)
Positive lymph nodes, (n)	5.0 (5.5)	4.9 (5.6)	5.1 (5.3)
Progesterone receptors, fmol/L (median [IQR])	32.5 [7.0, 131.8]	32.0 [7.0, 130.0]	35.0 [7.2, 133.0]
Estrogen receptors, fmol/L (median [IQR])	36.0 [8.0, 114.0]	32.0 [8.0, 92.2]	46.0 [9.0, 182.5]

Numbers are No. (%) unless otherwise noted. SD = standard deviation, fmol/L = femtomole per liter, IQR = interquartile range



Publication Quality Tables

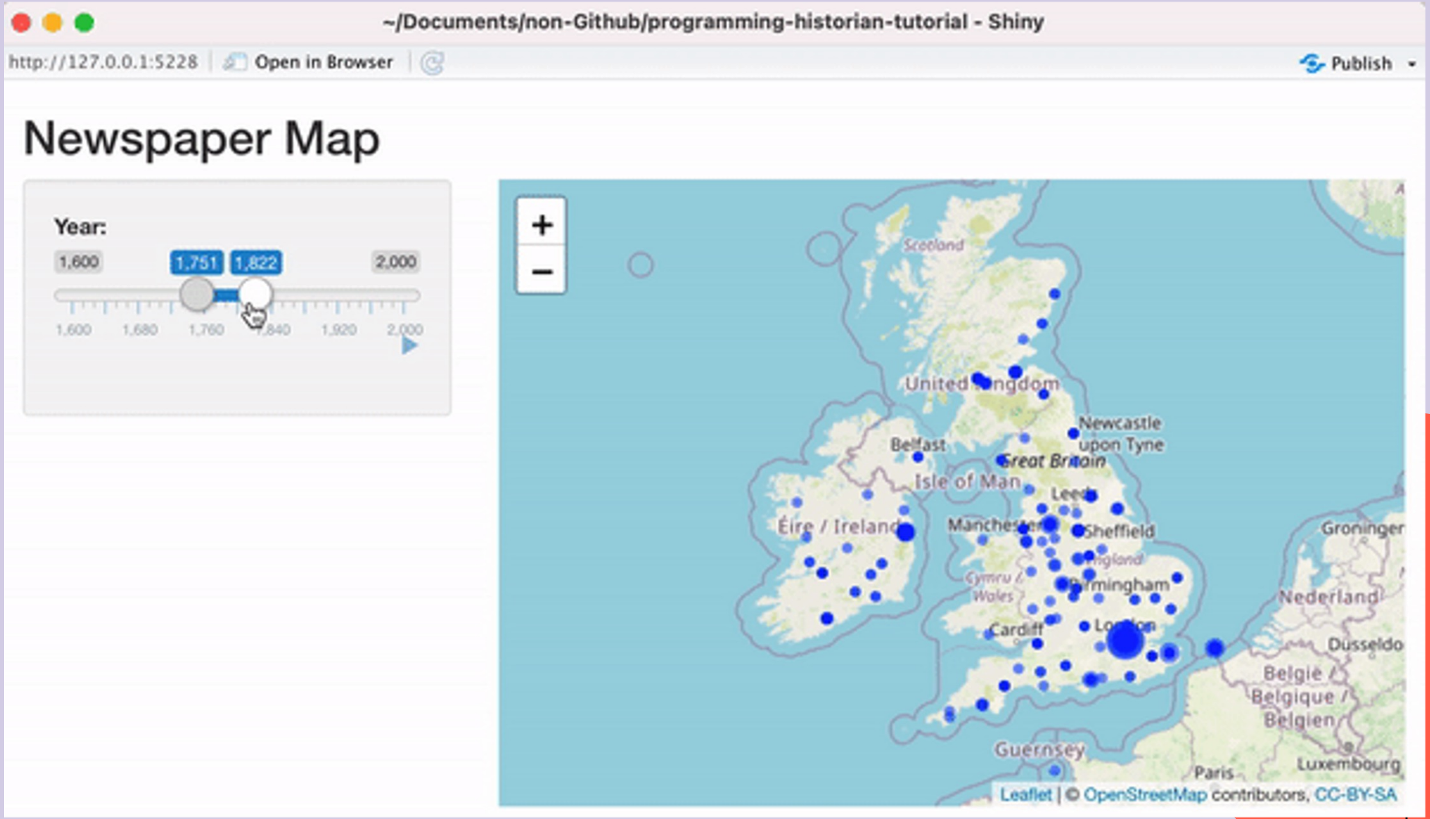
https://epirhandbook.com/en/dashboards-with-r-markdown.html



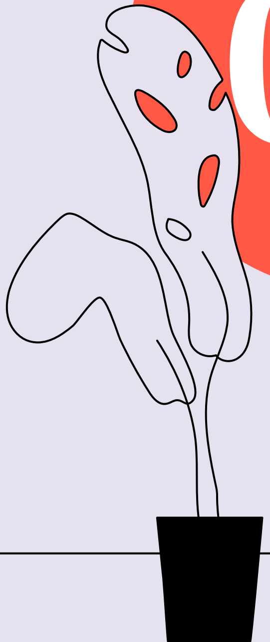
Hospital	Cases
Central Hospital	454
Military Hospital	896
Missing	1469
Other	885
Port Hospital	1762
St. Mark's Maternity Hospital (SMMH)	422
Total	5888

Dashboards

https://programminghistorian.org/en/lessons/shiny-leaflet-newspaper-map-tutorial#:~:text=Shiny%20is%20a%20library%20(a,down%20menus%2C%20and%20so%20forth.



Interactive Webpages



04 Demonstration

A simple data visualization in RStudio

THE GLOBAL HEALTH NETWORK
Enabling research by sharing knowledge

Thank you!

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, infographics & images by **Freepik**

Mr Md Sojibul Islam

Research Assistant, Non-Communicable Diseases and Nutrition Research Division, International Centre for Diarrhoeal Disease Research, Bangladesh (icddr,b)

Presentation

Analysis of stunting in Bangladesh, a case study of presenting findings in R

Reduction of the prevalence of stunting among children in Bangladesh and attribution of socio-demographic characteristics (2004-2017)

A case study of presenting findings in R programming language

Spotlight on R

Presented By
Md. Sojibul Islam
Research Assistant, Non-communicable Disease
Nutrition Research Division, icddr,b

6 September, 2023

icddr,b

Outline

- ❖ Background of the problems
- ❖ Why R programming language
- ❖ Objectives
- ❖ Methodology
- ❖ Presenting findings in R
  - ❖ Univariate analysis
  - ❖ Bivariate analysis
  - ❖ Modeling
- ❖ Conclusion

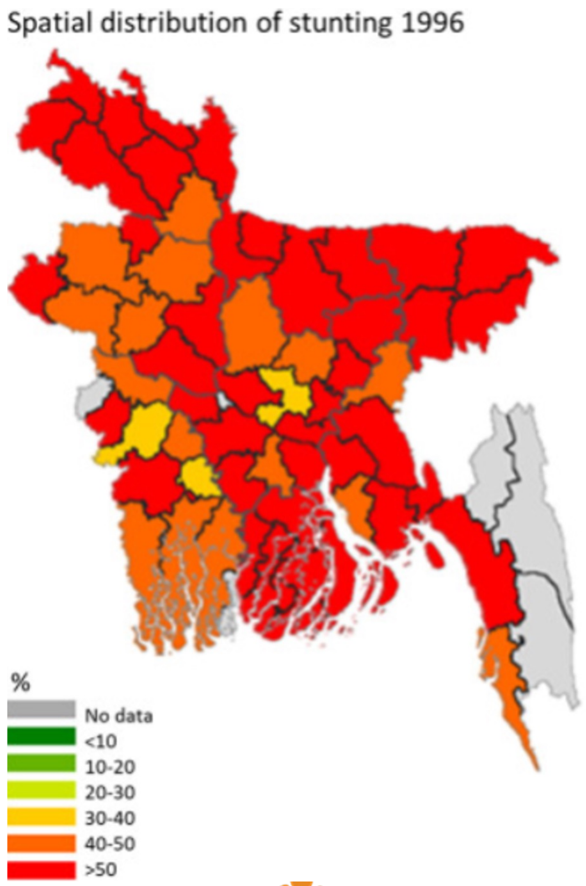
icddr,b

Background

Malnutrition is a major public health issue in developing countries.

According to UNICEF, 419 million children under 5 years old around the world are affected by stunting.

- Stunting is a condition that results from chronic malnutrition in early childhood, typically before the age of two. It is characterized by low height-for-age, reflecting a failure to reach one's full growth potential.
- This high prevalence of stunting in Asia is due to a variety of factors, including, poverty, food insecurity, and limited access to healthcare.



icddr,b

Why R programming

- R is a programming language
- R is very powerful for statistical analysis and epidemiology, and it’s free to use! You can use it to create:
  - ❑ Dashboards
  - ❑ Automated outbreak and situational analysis reports
  - ❑ Publication quality tables and figures

Objectives

Main Objective

The main objective is to present different statistical analysis about stunting status in Bangladesh in a smooth way by using application of modern packages in R Programming.

Specific Objectives

To find the reduction of childhood stunting prevalence in Bangladesh.

To identify the association between socio-demographic factor and stunting status in Bangladesh.

To determine potential socio-demographic factors affecting childhood stunting in Bangladesh.

icddr,b

Dataset for the case study: Bangladesh Demography and Health Survey

- The Bangladesh Demographic and Health Survey (BDHS) is a vital data collection effort that has been conducted periodically in Bangladesh over the past few decades. This comprehensive survey serves as a crucial resource for policymakers, researchers, and development organizations, providing valuable insights into various aspects of demographic and health-related information.
- This case study provides an overview of the combine demographic and health census data collected in Bangladesh over the past two decades, specifically in the years 2004, 2007, 2011, 2014, and 2017.
- Utilizing the large dataset, we will now proceed to conduct some statistical analysis by employing R to generate tables and insights.

icddr,b

Variables

Response Variable

Stunting Status variable have two categories:

- Stunted (HAZ < -2)
- Not Stunted (HAZ >= -2)

Socio-Demographic Factor

Age [Children Age in Month's]

Sex

- Male
- Female

Place of Residence

- Rural
- Urban

Mother's Education

- No Education
- Incomplete Primary
- Complete Primary
- Incomplete Secondary
- Complete Secondary
- Higher Education

Data analysis using R language

Analysis	Statistical technique	Use of R library
Univariate analysis	▪ Descriptive statistics for numerical variables ▪ Frequency Distribution table for categorical variables	library(tidyverse) library(gtsummary)
Bi-varaite analysis	• Cross tabulation • Chi square test • T test	library(tidyverse) library(gtsummary)
Multivariate Analysis	• Logistic regression • Forest plot	Library(arm) Library(forestmodels)

icddr,b

Tools and Techniques

R Programming Language (Version: 4.2.1)

- library(tidyverse) → **For Data Cleaning**
- library(gtsummary) → **Making all analysis table**
- library(arm) → **Building Logistic Regression models**
- library(forestmodel) → **Creating forest plot**
- library(DALEX) → **Creating forest plot**
- library(flextable) → **Export or Save analysis output into the Document**

icddr,b

Review Our Recently Published Dementia Paper

And intend to generate analysis tables and graphs as outlined below in our dementia research paper, aligning with our research objectives.

Prevalence of dementia among older age people and variation across different sociodemographic characteristics: a cross-sectional study in Bangladesh

Aliya Naheed,[a,*] Maliha Hakim,[b] Md Saimul Islam,[a] Md Badrul Islam,[c] Eugene Y. H. Tang,[d] Abdul Alim Prodhan,[e] Mohammad Robed Amin,[e,f] Blossom C. M. Stephan,[g,h,i] and Quazi Deen Mohammad[b,i]

[a]Initiative for Non Communicable Diseases, Health Systems and Population Studies Division, icddr,b, Mohakhali, Dhaka, 1000, Bangladesh
[b]National Institute of Neurosciences & Hospital, Dhaka, 1207, Bangladesh
[c]Laboratory Science and Services Division, icddr,b, Mohakhali, Dhaka, 1000, Bangladesh
[d]Population Health Sciences Institute, Newcastle University, UK
[e]Non Communicable Disease Control Program, Directorate General of Health Services, Dhaka, 1212, Bangladesh
[f]Department of Medicine, Dhaka Medical College and Hospital, Dhaka, 1000, Bangladesh
[g]Institute of Mental Health, Mental Health and Clinical Neurosciences, School of Medicine, University of Nottingham, Nottingham, UK
[h]Dementia Centre of Excellence, Curtin enAble Institute, Curtin University, Perth, Western Australia, Australia

Summary

Background Dementia is a significant global health issue, particularly for low-income and middle-income countries which majorly contribute to the dementia cases reported globally (67%). We estimated the prevalence of dementia among older people in Bangladesh and compared the estimate across different sociodemographic characteristics and divisions.

Methods A cross-sectional study was conducted in 2019 among individuals aged 60 years or older in seven administrative divisions in Bangladesh. Equal numbers of male and female participants were recruited from each division through a multi-stage random sampling technique. Recruitment was proportionally distributed in urban and rural areas in each division. Following consent, the Mini Mental State Examination (MMSE) was performed on all participants. Dementia was defined as an MMSE score of <24 out of 30. Data on age, sex, education, marital status, occupation, socioeconomic status, and type of community (urban or rural) were obtained using a structured questionnaire to compare the prevalence of dementia across different sociodemographic characteristics.

Findings Between January and December 2019, 2795 individuals were recruited including ~400 from each of the seven administrative divisions. The mean age was 67 years (SD: 7), 68% were from rural areas and 51% were female. The prevalence of dementia was 8.0% (95% CI: 7.0–8.9%) with variations across age, sex, education, marital status, occupation, and division. No variations in prevalence were observed across urban/rural locations or socioeconomic status. After adjusting for age, sex, education, occupation and marital status, the odds of dementia was two times higher in females than males (OR: 2.15, 95% CI: 1.43–3.28); nine times higher in people aged ≥90 years than people aged 60–69 years (OR: 9.62, 95% CI: 4.79–19.13), and three times higher in people with no education compared to those who had completed primary school (OR: 3.10, 95% CI: 1.95–5.17).

Interpretations The prevalence of dementia is high in Bangladesh and varies across sociodemographic characteristics with a higher prevalence among females, older people, and people with no education. There is an urgent need to identify the key risk factors for dementia in developing countries, such as Bangladesh, to inform the development of context-relevant risk reduction and prevention strategies.

Variables	Total = 2795	With dementia (MMSE score <24)		Without dementia (MMSE score ≥24)		P-value[b]
		n = 223	%	n = 2573	%	
Gender						
Male	1369	57	4.2	1312	95.8	P < 0.001
Female	1426	166	11.6	1260	88.4	
Age group						
60–64 y	1183	59	5.0	1124	95.0	P < 0.001
65–69 y	699	42	6.0	657	94.0	
70–74 y	434	47	10.8	387	89.2	
75–79 y	261	31	11.9	230	88.1	
80–84 y	114	17	14.9	97	85.1	
85–89	60	9	15.0	51	85.0	
90–115	44	18	40.9	26	59.1	
Marital status						
Married	1642	80	4.9	1562	95.1	P < 0.001
Single[a]	1153	143	12.4	1010	87.6	
Education						
Completed priamry education	685	21	3.1	664	96.9	P < 0.001
Some education	939	56	6.0	883	94.0	
Never went to school	1171	146	12.5	1025	87.5	
Current engage in earning						
Yes	813	33	4.1	780	95.9	P < 0.001
No	1982	190	9.6	1792	90.4	
Place of residence						
Urban	896	70	7.8	826	92.2	0.824
Rural	1899	153	8.1	1746	91.9	
Socioeconomic status						
Lower	448	42	9.4	406	90.6	0.209
Lower middle	522	48	9.2	474	90.8	
Middle	646	41	6.3	605	93.7	
Upper middle	566	39	6.9	527	93.1	
Upper	613	53	8.6	560	91.4	
Division						
Rajshahi	409	59	14.4	350	85.6	P < 0.001
Rangpur	409	48	11.7	361	88.3	
Khulna	409	32	7.8	377	92.2	
Barisal	409	30	7.3	379	92.7	
Chattogram	341	23	6.7	318	93.3	
Sylhet	409	19	4.6	390	95.4	
Dhaka	409	12	2.9	397	97.1	

MMSE, Mini Mental State Examination. [a]Widowed, Separated, Unmarried, Divorced. [b]Applied chi-square test of independence.

Table 2: Prevalence of dementia across socio-demographic characteristics, type of community and divisions.

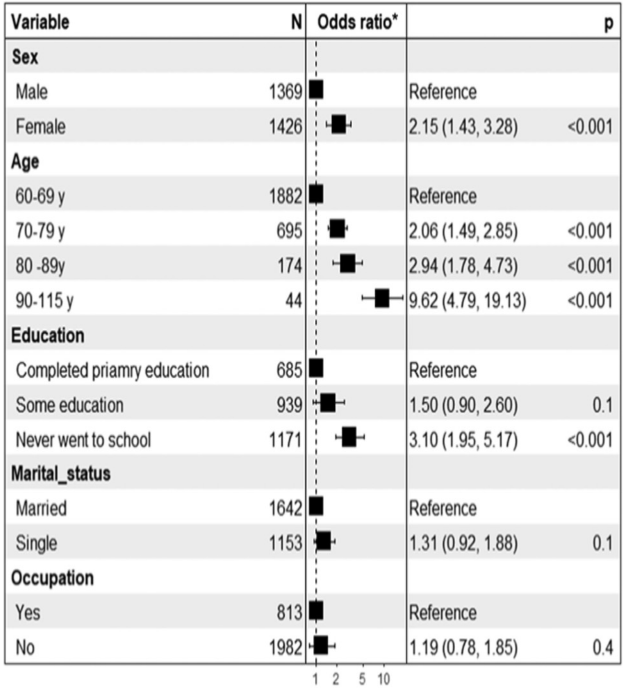


Fig. 3: Factors associated with and without dementia of older people (*Adjusted odds ratio with 95% CI.).

icddr,b

Preview Our Data Set

R Code

```
# View the Data
View(dt)
```

Output

Year	Age	Gender	Residence	Education	Stunting_Status
2017	6	Male	Rural	Incomplete Secondary	Not Stunted
2017	27	Male	Urban	Higher Education	Not Stunted
2017	51	Male	Rural	Incomplete Secondary	Not Stunted
2017	15	Male	Urban	Incomplete Secondary	Not Stunted
2017	12	Male	Urban	No Education	Not Stunted
2017	29	Male	Urban	Incomplete Secondary	Not Stunted
2017	13	Male	Urban	Complete Primary	Not Stunted
2017	34	Male	Urban	Incomplete Primary	Not Stunted
2017	59	Male	Rural	Incomplete Secondary	Not Stunted
2017	22	Female	Urban	Incomplete Primary	Not Stunted
2017	28	Male	Rural	Incomplete Secondary	Not Stunted
2017	21	Male	Rural	Higher Education	Not Stunted
2017	20	Male	Rural	Incomplete Secondary	Not Stunted
2017	0	Male	Urban	Incomplete Primary	Not Stunted
2017	11	Female	Urban	Incomplete Primary	Not Stunted
2017	30	Male	Urban	Higher Education	Not Stunted

Total Variables = 6

Total Observations = 33672

- BDHS 2004 Data = 5911
- BDHS 2007 Data = 5300
- BDHS 2011 Data = 7647
- BDHS 2014 Data = 6965
- BDHS 2018 Data = 7849

icddr,b

Univariate analysis in R

Univariate Analysis by Using R

Data

R Code

Output

Objectives 01 : To find the reduction of childhood stunting prevalence in Bangladesh.

Year	Age	Gender	Residence	Education	Stunting_Status
2017	6	Male	Rural	Incomplete Secondary	Not Stunted
2017	27	Male	Urban	Higher Education	Not Stunted
2017	51	Male	Rural	Incomplete Secondary	Not Stunted
2017	15	Male	Urban	Incomplete Secondary	Not Stunted
2017	12	Male	Urban	No Education	Not Stunted
2017	29	Male	Urban	Incomplete Secondary	Not Stunted
2017	13	Male	Urban	Complete Primary	Not Stunted
2017	34	Male	Urban	Incomplete Primary	Not Stunted
2017	59	Male	Rural	Incomplete Secondary	Not Stunted
2017	22	Female	Urban	Incomplete Primary	Not Stunted
2017	28	Male	Rural	Incomplete Secondary	Not Stunted
2017	21	Male	Rural	Higher Education	Not Stunted
2017	20	Male	Rural	Incomplete Secondary	Not Stunted
2017	0	Male	Urban	Incomplete Primary	Not Stunted
2017	11	Female	Urban	Incomplete Primary	Not Stunted
2017	30	Male	Urban	Higher Education	Not Stunted

File Edit Code View Plots Session Build Debug Profile Tools Help

```r
# Univariate analysis ####

dt %>%
  tbl_summary(by=Year)
```

Characteristic	**2004**, N = 5,911[1]	**2007**, N = 5,300[1]	**2011**, N = 7,647[1]	**2014**, N = 6,965[1]	**2017**, N = 7,849[1]
Age	29 (14, 44)	29 (15, 44)	31 (15, 46)	30 (14, 45)	28 (13, 44)
Gender					
Female	413 (7.0%)	482 (9.1%)	601 (7.9%)	649 (9.3%)	979 (12%)
Male	5,498 (93%)	4,818 (91%)	7,046 (92%)	6,316 (91%)	6,870 (88%)
Residence					
Rural	4,146 (70%)	3,450 (65%)	5,305 (69%)	4,777 (69%)	5,168 (66%)
Urban	1,765 (30%)	1,850 (35%)	2,342 (31%)	2,188 (31%)	2,681 (34%)
Education					
Complete Primary	585 (9.9%)	506 (9.6%)	974 (13%)	829 (12%)	865 (11%)
Complete Secondary	142 (2.4%)	359 (6.8%)	426 (5.6%)	457 (6.6%)	396 (5.0%)
Higher Education	354 (6.0%)	418 (7.9%)	608 (8.0%)	736 (11%)	1,303 (17%)
Incomplete Primary	1,275 (22%)	1,136 (22%)	1,356 (18%)	1,105 (16%)	1,411 (18%)
Incomplete Secondary	1,446 (24%)	1,441 (27%)	2,834 (37%)	2,762 (40%)	3,313 (42%)
No Education	2,109 (36%)	1,420 (27%)	1,449 (19%)	1,076 (15%)	561 (7.1%)
Unknown	0	20	0	0	0
Stunting_Status					
Not Stunted	2,896 (49%)	3,017 (57%)	4,523 (59%)	4,398 (63%)	5,365 (68%)
Stunted	3,015 (51%)	2,283 (43%)	3,124 (41%)	2,567 (37%)	2,484 (32%)

[1] Median (IQR); n (%)

Bivariate analysis in R

Bivariate Analysis by Using R

Data | **R Code** | **Output**

Objectives 02: To identify the association between socio-demographic factor and stunting status in Bangladesh.

Year	Age	Gender	Residence	Education	Stunting_Status
2017	6	Male	Rural	Incomplete Secondary	Not Stunted
2017	27	Male	Urban	Higher Education	Not Stunted
2017	51	Male	Rural	Incomplete Secondary	Not Stunted
2017	15	Male	Urban	Incomplete Secondary	Not Stunted
2017	12	Male	Urban	No Education	Not Stunted
2017	29	Male	Urban	Incomplete Secondary	Not Stunted
2017	13	Male	Urban	Complete Primary	Not Stunted
2017	34	Male	Urban	Incomplete Primary	Not Stunted
2017	59	Male	Rural	Incomplete Secondary	Not Stunted
2017	22	Female	Urban	Incomplete Primary	Not Stunted
2017	28	Male	Rural	Incomplete Secondary	Not Stunted
2017	21	Male	Rural	Higher Education	Not Stunted
2017	20	Male	Rural	Incomplete Secondary	Not Stunted
2017	0	Male	Urban	Incomplete Primary	Not Stunted
2017	11	Female	Urban	Incomplete Primary	Not Stunted
2017	30	Male	Urban	Higher Education	Not Stunted

```
# Add Overall Count

dt %>%
  tbl_summary(by=Stunting_Status) %>%
  add_overall()
```

Characteristic	Overall, N = 33,672[1]	Not Stunted, N = 20,199[1]	Stunted, N = 13,473[1]
Year			
2004	5,911 (18%)	2,896 (14%)	3,015 (22%)
2007	5,300 (16%)	3,017 (15%)	2,283 (17%)
2011	7,647 (23%)	4,523 (22%)	3,124 (23%)
2014	6,965 (21%)	4,398 (22%)	2,567 (19%)
2017	7,849 (23%)	5,365 (27%)	2,484 (18%)
Age	29 (14, 44)	26 (11, 44)	33 (20, 45)
Gender			
Female	3,124 (9.3%)	1,988 (9.8%)	1,136 (8.4%)
Male	30,548 (91%)	18,211 (90%)	12,337 (92%)
Residence			
Rural	22,846 (68%)	13,104 (65%)	9,742 (72%)
Urban	10,826 (32%)	7,095 (35%)	3,731 (28%)
Education			
Complete Primary	3,759 (11%)	2,062 (10%)	1,697 (13%)
Complete Secondary	1,780 (5.3%)	1,334 (6.6%)	446 (3.3%)
Higher Education	3,419 (10%)	2,809 (14%)	610 (4.5%)
Incomplete Primary	6,283 (19%)	3,295 (16%)	2,988 (22%)
Incomplete Secondary	11,796 (35%)	7,662 (38%)	4,134 (31%)
No Education	6,615 (20%)	3,027 (15%)	3,588 (27%)
Unknown	20	10	10

[1] n (%); Median (IQR)

Bivariate Analysis by Using R

R Code

```
# Add P Values

dt %>%
  tbl_summary(by=Stunting_Status) %>%
  add_overall() %>%
  add_p()
```

Output

Characteristic	**Overall**, N = 33,672[1]	**Not Stunted**, N = 20,199[1]	**Stunted**, N = 13,473[1]	**p-value**[2]
Year				<0.001
2004	5,911 (18%)	2,896 (14%)	3,015 (22%)	
2007	5,300 (16%)	3,017 (15%)	2,283 (17%)	
2011	7,647 (23%)	4,523 (22%)	3,124 (23%)	
2014	6,965 (21%)	4,398 (22%)	2,567 (19%)	
2017	7,849 (23%)	5,365 (27%)	2,484 (18%)	
Age	29 (14, 44)	26 (11, 44)	33 (20, 45)	<0.001
Gender				<0.001
Female	3,124 (9.3%)	1,988 (9.8%)	1,136 (8.4%)	
Male	30,548 (91%)	18,211 (90%)	12,337 (92%)	
Residence				<0.001
Rural	22,846 (68%)	13,104 (65%)	9,742 (72%)	
Urban	10,826 (32%)	7,095 (35%)	3,731 (28%)	
Education				<0.001
Complete Primary	3,759 (11%)	2,062 (10%)	1,697 (13%)	
Complete Secondary	1,780 (5.3%)	1,334 (6.6%)	446 (3.3%)	
Higher Education	3,419 (10%)	2,809 (14%)	610 (4.5%)	
Incomplete Primary	6,283 (19%)	3,295 (16%)	2,988 (22%)	
Incomplete Secondary	11,796 (35%)	7,662 (38%)	4,134 (31%)	
No Education	6,615 (20%)	3,027 (15%)	3,588 (27%)	
Unknown	20	10	10	

[1] n (%); Median (IQR)

[2] Pearson's Chi-squared test; Wilcoxon rank sum test

icddr,b

Bivariate Analysis by Using R

R Code

```r
# add event, C.I, and Labeling
dt %>%
  tbl_summary(by=Stunting_Status) %>%
  add_p() %>%
  add_overall() %>%
  add_n() %>%
  add_ci() %>%
  add_stat_label(
    label = all_continuous()~"Median (IQR)"
  )
```

Output

Characteristic	N	**Overall**, N = 33,672	**95% CI**[1]	**Not Stunted**, N = 20,199	**95% CI**[1]	**Stunted**, N = 13,473	**95% CI**[1]	**p-value**[2]
Year, n (%)	33,672							<0.001
2004		5,911 (18%)	17%, 18%	2,896 (14%)	14%, 15%	3,015 (22%)	22%, 23%	
2007		5,300 (16%)	15%, 16%	3,017 (15%)	14%, 15%	2,283 (17%)	16%, 18%	
2011		7,647 (23%)	22%, 23%	4,523 (22%)	22%, 23%	3,124 (23%)	22%, 24%	
2014		6,965 (21%)	20%, 21%	4,398 (22%)	21%, 22%	2,567 (19%)	18%, 20%	
2017		7,849 (23%)	23%, 24%	5,365 (27%)	26%, 27%	2,484 (18%)	18%, 19%	
Age, Median (IQR)	33,672	29 (14, 44)		26 (11, 44)	27, 28	33 (20, 45)	32, 33	<0.001
Gender, n (%)	33,672							<0.001
Female		3,124 (9.3%)	9.0%, 9.6%	1,988 (9.8%)	9.4%, 10%	1,136 (8.4%)	8.0%, 8.9%	
Male		30,548 (91%)	90%, 91%	18,211 (90%)	90%, 91%	12,337 (92%)	91%, 92%	
Residence, n (%)	33,672							<0.001
Rural		22,846 (68%)	67%, 68%	13,104 (65%)	64%, 66%	9,742 (72%)	72%, 73%	
Urban		10,826 (32%)	32%, 33%	7,095 (35%)	34%, 36%	3,731 (28%)	27%, 28%	
Education, n (%)	33,652							<0.001
Complete Primary		3,759 (11%)	11%, 12%	2,062 (10%)	9.8%, 11%	1,697 (13%)	12%, 13%	
Complete Secondary		1,780 (5.3%)	5.1%, 5.5%	1,334 (6.6%)	6.3%, 7.0%	446 (3.3%)	3.0%, 3.6%	
Higher Education		3,419 (10%)	9.8%, 10%	2,809 (14%)	13%, 14%	610 (4.5%)	4.2%, 4.9%	
Incomplete Primary		6,283 (19%)	18%, 19%	3,295 (16%)	16%, 17%	2,988 (22%)	21%, 23%	
Incomplete Secondary		11,796 (35%)	35%, 36%	7,662 (38%)	37%, 39%	4,134 (31%)	30%, 31%	
No Education		6,615 (20%)	19%, 20%	3,027 (15%)	15%, 15%	3,588 (27%)	26%, 27%	
Unknown		20		10		10		

[1] CI = Confidence Interval

[2] Pearson's Chi-squared test; Wilcoxon rank sum test

icddr,b

Bivariate Analysis by Using R

R Code

```
dt %>%
  tbl_summary(
    by = Stunting_Status,
    statistic = Age ~ "{mean} ({sd})",
    label = list(Age ~ "Age in Months",
                 Gender ~ "Sex of the Household Head",
                 Education ~"Mother's Education"
                 ),
    # missing = no,
    missing_text = "Missing Values",
    type =  list(Education="categorical",
                 Residence= "categorical"),

    sort = everything() ~ "frequency",
    percent = "col",
    digits = list (all_categorical() ~2,
                   all_continuous() ~1)) %>%
  add_p() %>%
  add_ci() %>%
  add_stat_label(
    label = all_continuous()~"Mean (SD)"
  ) %>%
  bold_p(t=0.05) %>%
  bold_labels()
```

Output

Characteristic	**Not Stunted**, N = 20,199	**95% CI**[1]	**Stunted**, N = 13,473	**95% CI**[1]	**p-value**[2]
Year, n (%)					**<0.001**
2017	5,365.00 (26.56%)	26%, 27%	2,484.00 (18.44%)	18%, 19%	
2011	4,523.00 (22.39%)	22%, 23%	3,124.00 (23.19%)	22%, 24%	
2014	4,398.00 (21.77%)	21%, 22%	2,567.00 (19.05%)	18%, 20%	
2004	2,896.00 (14.34%)	14%, 15%	3,015.00 (22.38%)	22%, 23%	
2007	3,017.00 (14.94%)	14%, 15%	2,283.00 (16.95%)	16%, 18%	
Age in Months, Mean (SD)	27.5 (18.0)	27, 28	32.4 (15.6)	32, 33	**<0.001**
Sex of the Household Head, n (%)					**<0.001**
Male	18,211.00 (90.16%)	90%, 91%	12,337.00 (91.57%)	91%, 92%	
Female	1,988.00 (9.84%)	9.4%, 10%	1,136.00 (8.43%)	8.0%, 8.9%	
Residence, n (%)					**<0.001**
Rural	13,104.00 (64.87%)	64%, 66%	9,742.00 (72.31%)	72%, 73%	
Urban	7,095.00 (35.13%)	34%, 36%	3,731.00 (27.69%)	27%, 28%	
Mother's Education, n (%)					**<0.001**
Incomplete Secondary	7,662.00 (37.95%)	37%, 39%	4,134.00 (30.71%)	30%, 31%	
No Education	3,027.00 (14.99%)	15%, 15%	3,588.00 (26.65%)	26%, 27%	
Incomplete Primary	3,295.00 (16.32%)	16%, 17%	2,988.00 (22.19%)	21%, 23%	
Complete Primary	2,062.00 (10.21%)	9.8%, 11%	1,697.00 (12.60%)	12%, 13%	
Higher Education	2,809.00 (13.91%)	13%, 14%	610.00 (4.53%)	4.2%, 4.9%	
Complete Secondary	1,334.00 (6.61%)	6.3%, 7.0%	446.00 (3.31%)	3.0%, 3.6%	
Missing Values	10		10		

[1] CI = Confidence Interval

[2] Pearson's Chi-squared test; Wilcoxon rank sum test

Multivariate Analysis by Using R

R Code

Output

Objectives 03: To determine potential socio-demographic factors affecting childhood stunting in Bangladesh.

Year	Age	Gender	Residence	Education	Stunting_Status
2017	6	Male	Rural	Incomplete Secondary	Not Stunted
2017	27	Male	Urban	Higher Education	Not Stunted
2017	51	Male	Rural	Incomplete Secondary	Not Stunted
2017	15	Male	Urban	Incomplete Secondary	Not Stunted
2017	12	Male	Urban	No Education	Not Stunted
2017	29	Male	Urban	Incomplete Secondary	Not Stunted
2017	13	Male	Urban	Complete Primary	Not Stunted
2017	34	Male	Urban	Incomplete Primary	Not Stunted
2017	59	Male	Rural	Incomplete Secondary	Not Stunted
2017	22	Female	Urban	Incomplete Primary	Not Stunted
2017	28	Male	Rural	Incomplete Secondary	Not Stunted
2017	21	Male	Rural	Higher Education	Not Stunted
2017	20	Male	Rural	Incomplete Secondary	Not Stunted
2017	0	Male	Urban	Incomplete Primary	Not Stunted
2017	11	Female	Urban	Incomplete Primary	Not Stunted
2017	30	Male	Urban	Higher Education	Not Stunted

```
# 3. Summarize the regression model

library(arm)

rm <- arm::bayesglm(
  Stunting_Status ~ Age+Year+Residence + Gender + Education,
              data = dt,
              family = binomial
              )

tbl_regression(rm, exponentiate = T)
```

Characteristic	OR[1]	95% CI[1]	p-value
Age	1.02	1.01, 1.02	<0.001
Year			
2004	—	—	
2007	0.79	0.73, 0.85	<0.001
2011	0.74	0.68, 0.79	<0.001
2014	0.66	0.62, 0.71	<0.001
2017	0.58	0.54, 0.62	<0.001
Residence			
Rural	—	—	
Urban	0.81	0.77, 0.85	<0.001
Gender			
Female	—	—	
Male	1.12	1.03, 1.21	0.005
Education			
Complete Primary	—	—	
Complete Secondary	0.43	0.38, 0.49	<0.001
Higher Education	0.30	0.27, 0.33	<0.001
Incomplete Primary	1.08	1.00, 1.18	0.055
Incomplete Secondary	0.69	0.64, 0.74	<0.001
No Education	1.27	1.17, 1.38	<0.001

[1] OR = Odds Ratio, CI = Confidence Interval

icddr,b

Multivariate Analysis by Using R

R Code

```
# Multivariate Modeling

glm(Stunting_Status ~ Age+ Residence + Year+ Gender + Education,
    data = dt, family = binomial) %>%
  tbl_regression(
    exponentiate=T
  ) %>%
  add_n()%>%
  add_significance_stars(
    hide_p = F, hide_se =F ,hide_ci = F) %>%
  # modify helpers
  modify_header(label="**Predictor**") %>%
  modify_caption("Table1. Cool Looking Table") %>%
  modify_footnote(
    ci= "CI= Credible Intervals are incredible ;",
    abbreviation = T) %>%
  bold_p(t=0.05) %>%
  bold_labels() %>%
  italicize_levels()
```

Output

Table1. Cool Looking Table

Predictor	**N**	**OR**[1,2]	**SE**[2]	**95% CI**[2]	**p-value**
Age	33,652	1.02***	0.001	1.01, 1.02	**<0.001**
Residence	33,652				
Rural		—	—	—	
Urban		0.81***	0.026	0.77, 0.85	**<0.001**
Year	33,652				
2004		—	—	—	
2007		0.79***	0.039	0.73, 0.85	**<0.001**
2011		0.74***	0.036	0.68, 0.79	**<0.001**
2014		0.66***	0.038	0.62, 0.71	**<0.001**
2017		0.58***	0.038	0.54, 0.62	**<0.001**
Gender	33,652				
Female		—	—	—	
Male		1.12**	0.041	1.03, 1.21	**0.005**
Education	33,652				
Complete Primary		—	—	—	
Complete Secondary		0.43***	0.065	0.38, 0.49	**<0.001**
Higher Education		0.30***	0.056	0.27, 0.33	**<0.001**
Incomplete Primary		1.08	0.042	1.00, 1.18	0.057
Incomplete Secondary		0.69***	0.039	0.64, 0.74	**<0.001**
No Education		1.27***	0.042	1.17, 1.38	**<0.001**

[1] *p<0.05; **p<0.01; ***p<0.001

[2] OR = Odds Ratio, SE = Standard Error, CI= Credible Intervals are incredible ;

icddr,b

Save Analysis Table by Using R

R Code

```r
# Save the Result in Word Document

library(flextable)
bi %>%
  as_flex_table() %>%
  save_as_docx(path = "Models.docx")
```

Output

Characteristic	0, N = 20,199	95% CI[1]	1, N = 13,473	95% CI[1]	p-value[2]
Year, n (%)					**<0.001**
2017	5,365.00 (26.56%)	26%, 27%	2,484.00 (18.44%)	18%, 19%	
2011	4,523.00 (22.39%)	22%, 23%	3,124.00 (23.19%)	22%, 24%	
2014	4,398.00 (21.77%)	21%, 22%	2,567.00 (19.05%)	18%, 20%	
2004	2,896.00 (14.34%)	14%, 15%	3,015.00 (22.38%)	22%, 23%	
2007	3,017.00 (14.94%)	14%, 15%	2,283.00 (16.95%)	16%, 18%	
Age in Months, Mean (SD)	27.5 (18.0)	27, 28	32.4 (15.6)	32, 33	**<0.001**
Sex of the Household Head, n (%)					**<0.001**
Male	18,211.00 (90.16%)	90%, 91%	12,337.00 (91.57%)	91%, 92%	
Female	1,988.00 (9.84%)	9.4%, 10%	1,136.00 (8.43%)	8.0%, 8.9%	
Residence, n (%)					**<0.001**

Multivariate Analysis by Using R

R Code

```
library(forestmodel)S
library(DALEX)


models <- glm(Stunting_Status ~ Age+ Residence + Year+ Gender + Education,
              data = dt, family=binomial(link="logit"))

forest_model(models,factor_separate_line=TRUE)
```

Forest Plot Output

Variable	N	Odds ratio	p
Age	33652	1.02 (1.01, 1.02)	<0.001
Residence			
Rural	22831	Reference	
Urban	10821	0.81 (0.77, 0.85)	<0.001
Year			
2004	5911	Reference	
2007	5280	0.79 (0.73, 0.85)	<0.001
2011	7647	0.74 (0.68, 0.79)	<0.001
2014	6965	0.66 (0.62, 0.71)	<0.001
2017	7849	0.58 (0.54, 0.62)	<0.001
Gender			
Female	3124	Reference	
Male	30528	1.12 (1.03, 1.21)	0.005
Education			
Complete Primary	3759	Reference	
Complete Secondary	1780	0.43 (0.38, 0.49)	<0.001
Higher Education	3419	0.30 (0.27, 0.33)	<0.001
Incomplete Primary	6283	1.08 (1.00, 1.18)	0.057
Incomplete Secondary	11796	0.69 (0.64, 0.74)	<0.001
No Education	6615	1.27 (1.17, 1.38)	<0.001

icddr,b

Conclusion

- R programming language provides effective packages for epidemiological data analysis
- The univariate, bivariate and multivariate model can be performed in a tabular format with few effort of r pakages.

Dr Soraida Aguilar

Assistente de Pesquisa, Divisão de Pesquisa em Doenças Não Transmissíveis e Nutrição, Centro Internacional de Pesquisa em Doenças Diarreicas, Bangladesh (icddr,b)

Presentation

Using R to support data preparation and visualization for a research study in Brazil

Using R to support Data Analytics

Usando R para respaldar el Análisis de Datos

Soraida Aguilar

Agenda

01 Estudio de investigación

De que estamos hablando

02 Flujo de trabajo para preparación de los datos

Pasos para preparar nuestros datos

03 Preparación de los datos

Uniendo todos los conjuntos de datos

04 Visualización de los datos

Dataframes y figuras

Estudio de investigación

The impact of the first year of COVID-19 vaccination strategy in Brazil

Bajo revisión

Flujo de trabajo de preparación de datos

01 Colecta de datos

02 Limpieza de datos

03 Transformación enriquecimiento de los datos

04 Validación de los datos

Preparación de los datos | Data

```r
library(dplyr)
library(tidyverse)
library(ggplot2)
library(scales)
library(lubridate)
library(purrr)
library(pracma)
library(patchwork)
library(RColorBrewer)
library(broom)
library(tidymodels)
library(rms)
library(ggsci)
Sys.setlocale("LC_TIME", "English")


##########################################################################################
#------------------------------- D A T A S E T   P R E P A R A T I O N -------------------------
##########################################################################################

#---------------- Data reading

# Deaths - SIVEP
obitos <- read.csv("srag_adults_covid_hosp_2022-07-11.csv", header = TRUE, sep = ",", dec = ",")

# Vaccination by state - SI-PNI
vacinacao <- read.csv("vw_vacc_date_state_age_dose_2022-02-02.csv", header = TRUE, sep = ",")

# Population by age group and state
piramide_etaria <- read.csv2("df_population_city_sex.csv", sep = ",", header = TRUE)

# Age-adjusted
df_ageSex_adjusted <- read.csv2("who_pop_std_rates.csv", sep = ",", header = TRUE)
```

Pasos 1-3 | Steps 1-3

- Hospitalizaciones y muertes hospitalarias por COVID-19 COVID-19 hospitalizations and in-hospital deaths
- Datos de vacunación Vaccine data
- Datos de Población por estado - por edad/sexo Population per state - per Age/Sex data
- Datos de población ajustados

Preparación de los datos | Data

```r
# Filtreing deaths and creating age group
data_obitos <- obitos %>%
  mutate(state = SG_UF, data = date_desf, deaths = EVOLUCAO) %>%
  mutate(faixa_etaria = case_when(NU_IDADE_N < 10  ~ "0-9",
                                  NU_IDADE_N >= 10 & NU_IDADE_N < 20 ~ "10-19",
                                  NU_IDADE_N >= 20 & NU_IDADE_N < 50 ~ "20-49",
                                  NU_IDADE_N >= 50 & NU_IDADE_N < 60 ~ "50-59",
                                  NU_IDADE_N >= 60 & NU_IDADE_N < 70 ~ "60-69",
                                  TRUE ~ "70+")) %>%
  select(state, data, faixa_etaria, deaths) %>%
  group_by(data, faixa_etaria) %>%
  filter(deaths == "Death") %>%
  mutate(deaths = 1) %>%
  summarise(newDeaths = sum(deaths))


aux_1 <- data_obitos %>%
  spread(faixa_etaria, newDeaths) %>%
  replace(is.na(.),0) %>%
  mutate(`<60` = `20-49` + `50-59`,
         `>=60` = `60-69` + `70+`)
aux_1 <- aux_1[-c(1),]
aux_1$data = lubridate::ymd(aux_1$data)


aux_2 <- data.frame(data = seq(from = lubridate::ymd(as.Date(aux_1$data[1])),
                               to = lubridate::ymd(as.Date("2021/12/31")),
                               by = "day"), valor = 0)

aux_3 <- left_join(aux_2, aux_1, by = c("data") )%>%
  replace(is.na(.),0)
aux_3 <- aux_3[,-c(2)]
```

Pasos 1-3 | Steps 1-3

- Hospitalizaciones y muertes hospitalarias por COVID-19
  COVID-19 hospitalizations and in-hospital deaths
- Selección de Variables
  Variables selection
- Transformar y enriquecer los datos
  Transforming and enrich the data

Preparación de los datos | Data preparation

```r
# Filtering single dose
data_vacinacao <- vacinacao %>%
  filter(uf != "" & uf != "XX") %>%
  mutate(state = uf, data = data_aplicacao, faixa_etaria = idade_grupo) %>%
  select(state, data, faixa_etaria, total, vacina_dose) %>%
  group_by(data, faixa_etaria) %>%
  filter(vacina_dose == "D1", faixa_etaria != "0-4" &  faixa_etaria != "5-9" &
           faixa_etaria != "10-14" &  faixa_etaria != "15-19" ) %>%
  summarise(vaccinated = sum(total))


aux_4 <- data_vacinacao %>%
  spread(faixa_etaria, vaccinated) %>%
  replace(is.na(.),0)
aux_4$data = lubridate::ymd(aux_4$data)

aux_4$`20-49` <- aux_4$`20-29` + aux_4$`30-39` + aux_4$`40-49`
aux_4$`70+` <- aux_4$`70-79` + aux_4$`80+`
aux_4$`<60` <- aux_4$`20-29` + aux_4$`30-39`  + aux_4$`40-49` + aux_4$`50-59`
aux_4$`>=60` <- aux_4$`60-69` + aux_4$`70-79`  + aux_4$`80+`
aux_4 <- aux_4 %>% select(data, "20-49", "50-59", "60-69", "70+", "<60", ">=60")

aux_4 <- as.data.frame(aux_4)
names(aux_4) <- c("data","a1","a2","a3","a4","a5","a6")
```

Pasos 1-3 | Steps 1-3

- Datos de vacunación
  Vaccine data
- Limpieza de los datos
  Clean data
- Transformar y enriquecer los datos
  Transforming and enrich the data

Preparación de los datos | Data preparation

```r
# Vaccine coverage, mortality and  deaths-ageadjusted
dataset_completo <- dataset_joined %>% mutate(
  `cobertura_20-49` = `20-49_vac`/faixa$`20-49`,
  `cobertura_50-59` = `50-59_vac`/faixa$`50-59`,
  `cobertura_60-69` = `60-69_vac`/faixa$`60-69`,
  `cobertura_70+` = `70+_vac`/faixa$`70+`,
  `cobertura_<60` = `<60_vac`/faixa$`<60`,
  `cobertura_>=60` = `>=60_vac`/faixa$`>=60`,

  `rate_20-49` = 100000*`20-49`/faixa$`20-49`,
  `rate_50-59` = 100000*`50-59`/faixa$`50-59`,
  `rate_60-69` = 100000*`60-69`/faixa$`60-69`,
  `rate_70+` = 100000*`70+`/faixa$`70+`,
  `rate_<60` = 100000*`<60`/faixa$`<60`,
  `rate_>=60` = 100000*`>=60`/faixa$`>=60`,

  `ageAdj_20-49` = 100000*(`20-49`/faixa$`20-49`)*
    (pop_ageAdjusted$`20-49`/total_ageAdjusted[[1]]),
  `ageAdj_50-59` = 100000*(`50-59`/faixa$`50-59`)*
    (pop_ageAdjusted$`50-59`/total_ageAdjusted[[1]]),
  `ageAdj_60-69` = 100000*`60-69`/faixa$`60-69`*
    (pop_ageAdjusted$`60-69`/total_ageAdjusted[[1]]),
  `ageAdj_70+` = 100000*(`70+`/faixa$`70+`)*
    (pop_ageAdjusted$`70+`/total_ageAdjusted[[1]]),
  `ageAdj_<60` = 100000*(`<60`/faixa$`<60`)*
    (pop_ageAdjusted$`<60`/total_ageAdjusted[[1]]),
  `ageAdj_>=60` = 100000*(`>=60`/faixa$`>=60`)*
    (pop_ageAdjusted$`>=60`/total_ageAdjusted[[1]])
```

Pasos 1-3 | Steps 1-3

- Enriqueciendo los datos: coberturas de vacunación, tasas de mortalidad, muertes ajustadas por edad
  Enriching the data: vaccination coverage, mortality rates, age-adjusted deaths

Preparación de los datos | Data preparation

Obteniendo un conjunto de datos unificado | Getting a unified data set

```r
# Joint all the data sets
conjunto_obitos_rate$faixa_etaria = conjunto_obitos$faixa_etaria
conjunto_vacina_doses$faixa_etaria = conjunto_obitos$faixa_etaria
conjunto_vacina_coverage$faixa_etaria = conjunto_obitos$faixa_etaria
conjunto_obitos_ageAdjusted$faixa_etaria = conjunto_obitos$faixa_etaria

join_1 <- left_join(conjunto_obitos, conjunto_obitos_ageAdjusted, by = c("data","faixa_etaria"))
join_2 <- left_join(conjunto_vacina_doses, conjunto_vacina_coverage, by = c("data","faixa_etaria"))
conjunto_dados <- left_join(join_1, join_2, by = c("data","faixa_etaria"))

df_obitos <- left_join(conjunto_dados, conjunto_obitos_rate, by = c("data","faixa_etaria"))
```

Visualización de los datos | Data Visualization

- Dataframes

	data	faixa_etaria	deaths	deaths_ageAdj	vacina	cobertura	deaths_pop
438	2021-05-23	20-49	343	0.2336583125	9598103	0.09823391	0.351050953
439	2021-05-24	20-49	343	0.2336583125	9887947	0.10120039	0.351050953
440	2021-05-25	20-49	347	0.2363831908	10253243	0.10493909	0.355144842
441	2021-05-26	20-49	350	0.2384268495	10610579	0.10859632	0.358215259
442	2021-05-27	20-49	356	0.2425141669	11031985	0.11290930	0.364356092
443	2021-05-28	20-49	361	0.2459202648	11567970	0.11839495	0.369473452
444	2021-05-29	20-49	370	0.2520512409	11832747	0.12110487	0.378684702
445	2021-05-30	20-49	377	0.2568197779	11870780	0.12149413	0.385849007
446	2021-05-31	20-49	383	0.2609070953	12268110	0.12556069	0.391989840
447	2021-06-01	20-49	389	0.2649944127	12800006	0.13100450	0.398130673
448	2021-06-02	20-49	396	0.2697629497	13487786	0.13804374	0.405294978
449	2021-06-03	20-49	399	0.2718066084	13759615	0.14082583	0.408365395
450	2021-06-04	20-49	403	0.2745314867	14173804	0.14506494	0.41245928

Visualización de los datos | Data Visualization

- Generación de figuras
  Figures generation

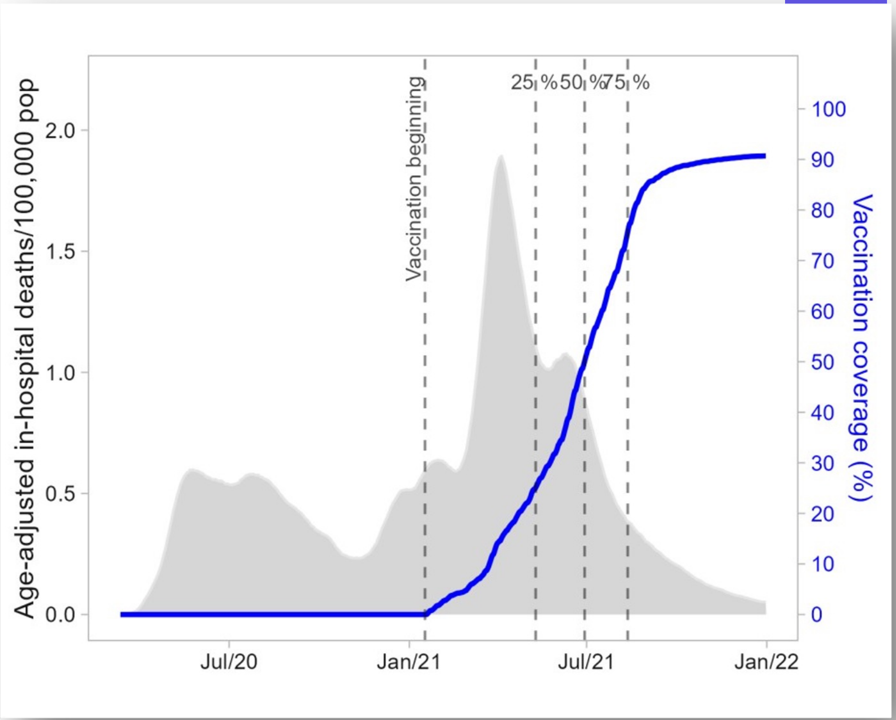
```r
# Extraindo as datas em que foram atingidos os
c0 <- lubridate::ymd("2021-01-17") # Inicio da
c10 <- df_filtro_brasil$data[which(df_filtro_b
c20 <- df_filtro_brasil$data[which(df_filtro_b
c25 <- df_filtro_brasil$data[which(df_filtro_b
c30 <- df_filtro_brasil$data[which(df_filtro_b
c40 <- df_filtro_brasil$data[which(df_filtro_b
c50 <- df_filtro_brasil$data[which(df_filtro_b
c60 <- df_filtro_brasil$data[which(df_filtro_b
c70 <- df_filtro_brasil$data[which(df_filtro_b
c75 <- df_filtro_brasil$data[which(df_filtro_b
c80 <- df_filtro_brasil$data[which(df_filtro_b
c90 <- df_filtro_brasil$data[which(df_filtro_b

corte_cobertura_1 <- data.frame(corte = c(c0,
scaleFactor <- max(df_filtro_brasil$deaths_age
```

```r
p1 <- df_filtro_brasil %>%
  ggplot() +
  geom_area(aes(x = data, y = deaths_ageAdj), color = "gray90", alpha = 0.2 ) +
  geom_line(aes(x = data, y = scaleFactor*cobertura),
            size = 1, color = "blue") +
  geom_vline(data = corte_cobertura_1, aes(xintercept = corte),
             linetype = "dashed" , color = "grey5", alpha = 0.5) +
  geom_text(data = corte_cobertura_1,
            aes(x = lubridate::as_date(corte)-12, y = 1.8,
                label = c("Vaccination beginning", "", "", "")),
            color = "grey5", size = 3, alpha = 0.8, angle = 90) |
  geom_text(data = corte_cobertura_1,
            aes(x = lubridate::as_date(corte)-1, y = 2.2,
                label = c("", "25 %", "50 %", "75 %")),
            color = "grey5", size = 3, alpha = 0.8) +
  scale_y_continuous(sec.axis = sec_axis(~.*(100/scaleFactor),
                                         breaks = seq(0,100, by = 10),
                                         name = "Vaccination coverage (%)"),
                     breaks = seq(0,2.5, by = 0.5)) +
  scale_x_date(labels = date_format("%b/%y"), breaks = waiver()) +
  ylab("Age-adjusted in-hospital deaths/100,000 pop") + xlab("") +
  labs(subtitle = "") +
  theme_light() +
  theme(axis.title.y.left = element_text(color = "grey8"),
        axis.text.y.left = element_text(color = "grey8"),
        axis.title.y.right = element_text(color = "blue"),
        axis.text.y.right = element_text(color = "blue"),
        axis.text.x = element_text(color = "grey8"),
        plot.background = element_rect(color = "white"),
        panel.grid.major = element_line(color = "white"),
        panel.grid.minor = element_line(color = "white"))
```

Visualización de los datos | Data Visualization

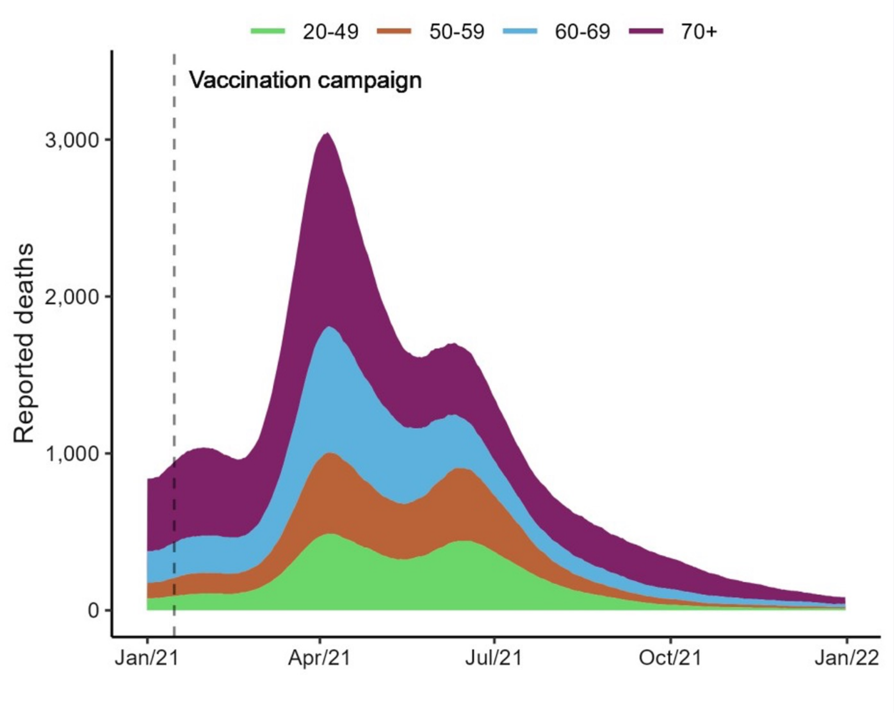
- Generación de figuras
  Figures generation

```
p1 <- df_filtro_brasil %>%
  ggplot() +
  geom_area(aes(x = data, y = deaths_ageAdj), color = "gray90", alpha = 0.2 ) +
  geom_line(aes(x = data, y = scaleFactor*cobertura),
            size = 1, color = "blue") +
  geom_vline(data = corte_cobertura_1, aes(xintercept = corte),
             linetype = "dashed" , color = "grey5", alpha = 0.5) +
  geom_text(data = corte_cobertura_1,
            aes(x = lubridate::as_date(corte)-12, y = 1.8,
                label = c("Vaccination beginning", "", "", "")),
            color = "grey5", size = 3, alpha = 0.8, angle = 90) |
  geom_text(data = corte_cobertura_1,
            aes(x = lubridate::as_date(corte)-1, y = 2.2,
                label = c("", "25 %", "50 %", "75 %")),
            color = "grey5", size = 3, alpha = 0.8) +
  scale_y_continuous(sec.axis = sec_axis(~.*(100/scaleFactor),
                                         breaks = seq(0,100, by = 10),
                                         name = "Vaccination coverage (%)"),
                     breaks = seq(0,2.5, by = 0.5)) +
  scale_x_date(labels = date_format("%b/%y"), breaks = waiver()) +
  ylab("Age-adjusted in-hospital deaths/100,000 pop") + xlab("") +
  labs(subtitle = "") +
  theme_light() +
  theme(axis.title.y.left = element_text(color = "grey8"),
        axis.text.y.left = element_text(color = "grey8"),
        axis.title.y.right = element_text(color = "blue"),
        axis.text.y.right = element_text(color = "blue"),
        axis.text.x = element_text(color = "grey8"),
        plot.background = element_rect(color = "white"),
        panel.grid.major = element_line(color = "white"),
        panel.grid.minor = element_line(color = "white"))
```



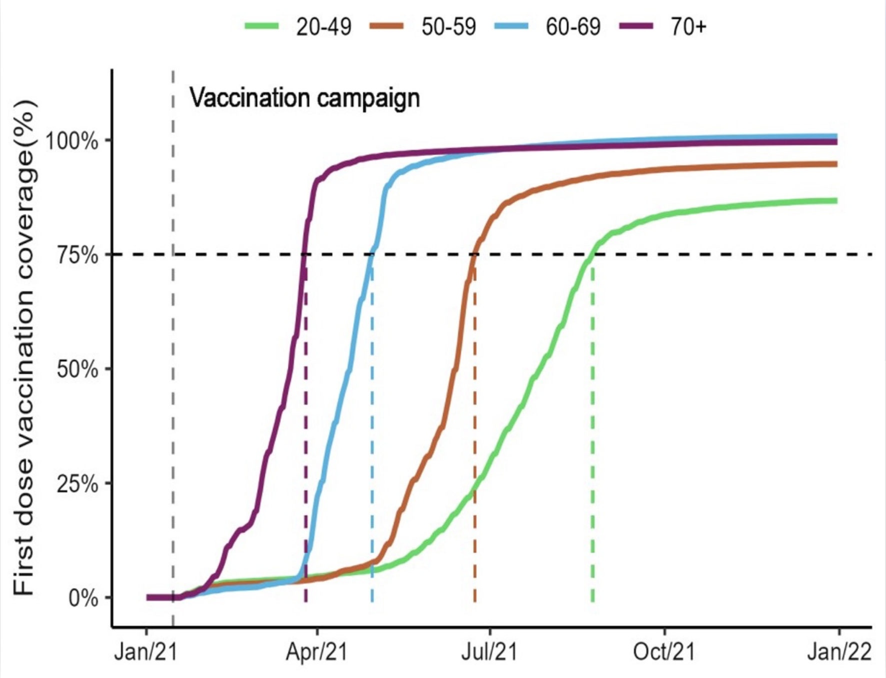
Visualización de los datos | Data Visualization

```r
colores <- c("#6BD76BFF", "#BA6338FF", "#5DB1DDFF", "#802268FF")
names(colores) <- c("20-49", "50-59", "60-69", "70+")
c0 <- lubridate::ymd("2021-01-17")

p3 <- df_obitos %>% filter(faixa_etaria %in% c("20-49", "50-59", "60-69", "70+"),
                           data > "2020-12-31") %>%
  mutate(faixa_etaria = factor(faixa_etaria,
                               levels = c("70+", "60-69", "50-59", "20-49"))) %>%
  ggplot() +
  geom_area(aes(x = data, y = deaths, fill = faixa_etaria)) +
  geom_vline(aes(xintercept = c0), linetype = "dashed" ,
             color = "black", alpha = 0.5) +
  geom_text(aes(x = lubridate::as_date(c0)+108, y = 3400,
                label = c("Vaccination campaign")), color = "grey5",
            size = 3.4, alpha = 0.8) +
  scale_fill_manual(guide = "none", values = colores) +
  scale_x_date(labels = date_format("%b/%y"), breaks = waiver()) +
  scale_y_continuous(labels = comma) +
  ylab("Reported deaths") + xlab("") + labs(subtitle = "") +
  labs(subtitle = "", color = "\n", fill = "\n") +
  theme_classic() +
  theme(axis.title.y.left = element_text(color = "grey8"),
        axis.text.y.left = element_text(color = "grey8"),
        axis.title.y.right = element_text(color = "blue"),
        axis.text.y.right = element_text(color = "blue"),
        axis.text.x = element_text(color = "grey8"),
        plot.background = element_rect(color = "white"),
        panel.grid.major = element_line(color = "white"),
        panel.grid.minor = element_line(color = "white"),
        legend.position = "top")
```



Visualización de los datos | Data Visualization

```r
p5 <- df_obitos %>%
  filter(faixa_etaria %in% c("20-49", "50-59", "60-69", "70+"), data > "2020-12-31") %>%
  mutate(cobertura = if_else(faixa_etaria == "20-49" & cobertura > 1, 1, cobertura)) %>%
  ggplot() +
  geom_line(aes(x = data, y = cobertura, color = faixa_etaria), size = 1) +
  geom_hline(aes(yintercept = 0.75), linetype = "dashed") +
  geom_vline(aes(xintercept = c0), linetype = "dashed" ,
             color = "black", alpha = 0.5) +
  geom_segment(aes(x = c75.1, y = -0.01, xend =  c75.1, yend = 0.75),
               linetype = "dashed" , color = "#6BD76BFF", alpha = 0.5) +
  geom_segment(aes(x = c75.2, y = -0.01, xend =  c75.2, yend = 0.75),
               linetype = "dashed" , color = "#BA6338FF", alpha = 0.5) +
  geom_segment(aes(x = c75.3, y = -0.01, xend =  c75.3, yend = 0.75),
               linetype = "dashed" , color = "#5DB1DDFF", alpha = 0.5) +
  geom_segment(aes(x = c75.4, y = -0.01, xend =  c75.4, yend = 0.75),
               linetype = "dashed" , color = "#802268FF", alpha = 0.5) +
  geom_text(aes(x = lubridate::as_date(c0)+108, y = 1.1,
                label = c("Vaccination campaign")), color = "grey5", size = 3.4, alpha = 0.8) +
  scale_x_date(labels = date_format("%b/%y"), breaks = waiver()) + #breaks = "2 month"
  scale_y_continuous(labels = scales::percent_format(), breaks = seq(0,100, by = 0.25),) +
  scale_color_manual(values = colores) +
  ylab("First dose vaccination coverage(%)") + xlab("") + labs(subtitle = "") +
  labs(subtitle = "", color = "\n") +
  theme_classic() +
  theme(axis.title.y.left = element_text(color = "grey8"),
        axis.text.y.left = element_text(color = "grey8"),
        axis.title.y.right = element_text(color = "blue"),
        axis.text.y.right = element_text(color = "blue"),
        axis.text.x = element_text(color = "grey8"),
        plot.background = element_rect(color = "white"),
        panel.grid.major = element_line(color = "white"),
        panel.grid.minor = element_line(color = "white"),
        legend.position = "top")
```



Visualización de los datos | Data Visualization

- Dataframes

```
# General descriptive table
library(gtsummary)
library(gt)

descritive <-
  tbl_summary(data = df_CQ019_EQ5D_DM_final_filter,
              missing = "ifany",
              missing_text = "NA",
              percent = "row",
              include = c(LONGCOVID_CQ019, LONGCOVID_EQ5, SEX, AGE,
                          COUNTRY, CONTINENT),
              label = list(LONGCOVID_CQ019 = "LONGCOVID_CQ019", SEX = "Sex",
                           AGE = "Age",
                           COUNTRY = "Country", CONTINENT = "Continent"
              )
  )%>%
  modify_header(label = "**Feature**") %>%
  as_gt() %>%
  tab_header(title = "Descriptive Analysis",
             subtitle = "Long COVID19 Worldwide Dataset"
  )
descritive

descritive%>%
  gt::gtsave(filename = "Descriptive_Analysis_simples_1.rtf")
```

Descriptive Analysis

Long COVID19 Worldwide Dataset

		Outcome – CQ019	
Feature	**Overall**, N = 10,997[1]	**Recovered**, N = 7,880[1]	**Unrecovered**, N = 3,117[1]
Sex			
F	5,915 (100%)	4,161 (70%)	1,754 (30%)
M	5,001 (100%)	3,658 (73%)	1,343 (27%)
U	81 (100%)	61 (75%)	20 (25%)
Age			
[18 - 30)	946 (100%)	743 (79%)	203 (21%)
[30 - 40)	1,649 (100%)	1,266 (77%)	383 (23%)
[40 - 50)	2,344 (100%)	1,699 (72%)	645 (28%)
[50 - 60)	2,740 (100%)	1,854 (68%)	886 (32%)
[60 - 70)	2,128 (100%)	1,461 (69%)	667 (31%)
[70 - 80)	1,004 (100%)	734 (73%)	270 (27%)
>= 80	186 (100%)	123 (66%)	63 (34%)
Country			
BRA	468 (100%)	357 (76%)	111 (24%)
COL	119 (100%)	82 (69%)	37 (31%)
FRA	54 (100%)	52 (96%)	2 (3.7%)
GBR	2,203 (100%)	938 (43%)	1,265 (57%)
GIB	314 (100%)	227 (72%)	87 (28%)
GMB	8 (100%)	5 (62%)	3 (38%)
IND	1,364 (100%)	1,263 (93%)	101 (7.4%)
ISR	600 (100%)	430 (72%)	170 (28%)
ITA	387 (100%)	291 (75%)	96 (25%)
NOR	5,459 (100%)	4,220 (77%)	1,239 (23%)
PRT	3 (100%)	2 (67%)	1 (33%)
SDN	2 (100%)	0 (0%)	2 (100%)
ZAF	16 (100%)	13 (81%)	3 (19%)
Continent			
Africa	26 (100%)	18 (69%)	8 (31%)
Asia	1,964 (100%)	1,693 (86%)	271 (14%)
Europe	8,420 (100%)	5,730 (68%)	2,690 (32%)
South America	587 (100%)	439 (75%)	148 (25%)

¡Muchas gracias! Thank you!

¿Preguntas?

Any Question?

Dr Frank Kagoro

Dr Frank Kagoro - Research Fellow, University of Cape Town & Research Scientist, Ifakara Health Institute, Tanzania

Presentation

User-Centred Dashboards for COVID-19 Trends in Africa

User-Centred Visualisations – Application in Malaria and COVID-19

Dr Frank M. Kagoro MSc
Research Fellow, University of Cape Town
Research Scientist, Ifakara Health Institute

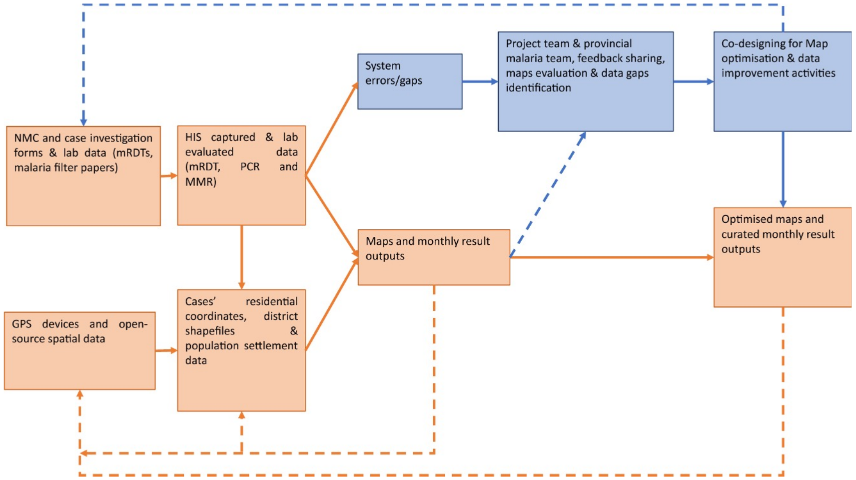
Adapted from DS-I Africa Virtual State of Data Science Series October 2020

Why visualisations?

- Understanding data, analytics and findings
- Simplifying BIG data workflows and outputs
- Simplifying/enabling communication of analytical findings
- Increasing usability of data and analytical products

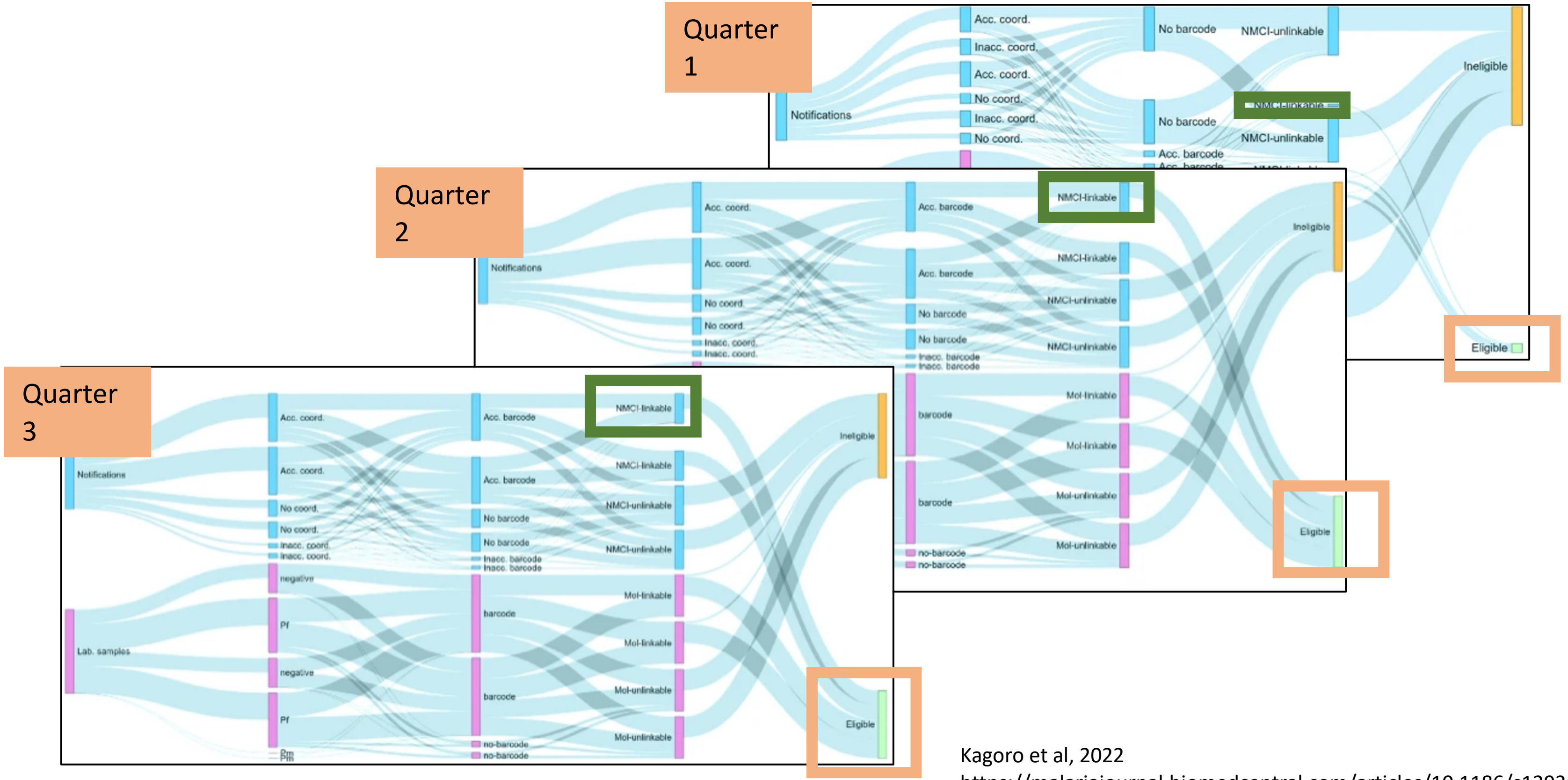
Problem 1: Malaria surveillance

- How best can we assess linkage and communicate antimalarial drug resistance (different sources, timelines)?



Kagoro et al, 2022
https://malariajournal.biomedcentral.com/articles/10.1186/s12936-022-04224-4

Are we improving over time? Where?



Kagoro et al, 2022
https://malariajournal.biomedcentral.com/articles/10.1186/s12936-022-04224-4

Monthly PDF reports

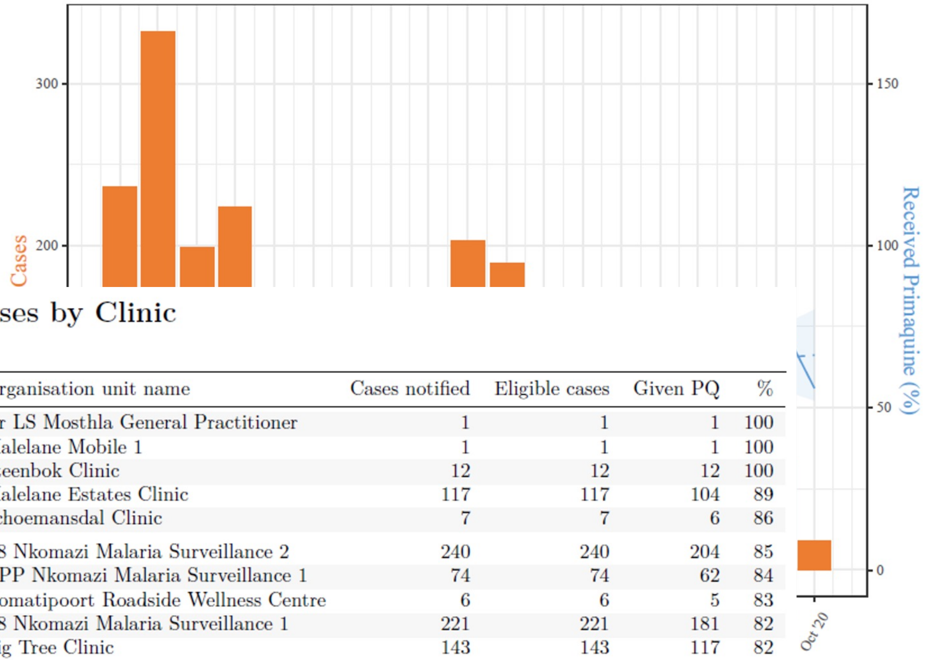
MPM Primaquine (Section 21) Summary - (23 February, 2021)

Formatted by - CCOAT, UCT

- Summary
- Cases by month
- Cases by Clinic
- Side effects/progress

1.1 Reporting Time

Start date	End date	Total Cases	Infants	Pregnant W.	Eligible
2019-04-01	2020-11-18	2665	41	0	2624

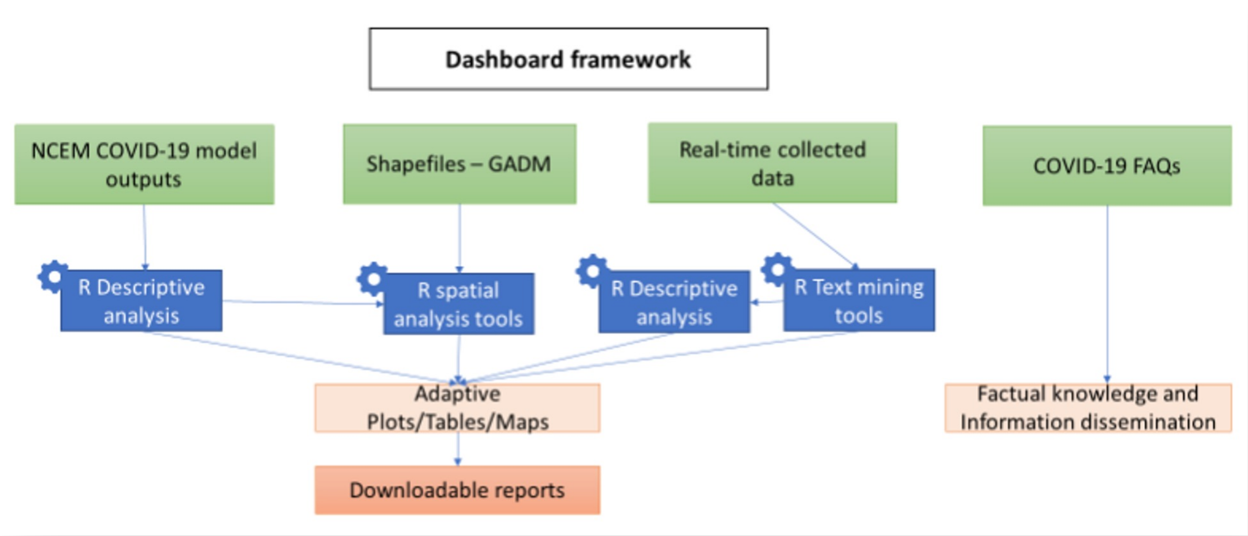


3 Cases by Clinic

Organisation unit name	Cases notified	Eligible cases	Given PQ	%
Dr LS Mosthla General Practitioner	1	1	1	100
Malelane Mobile 1	1	1	1	100
Steenbok Clinic	12	12	12	100
Malelane Estates Clinic	117	117	104	89
Schoemansdal Clinic	7	7	6	86
E8 Nkomazi Malaria Surveillance 2	240	240	204	85
HPP Nkomazi Malaria Surveillance 1	74	74	62	84
Komatipoort Roadside Wellness Centre	6	6	5	83
E8 Nkomazi Malaria Surveillance 1	221	221	181	82
Fig Tree Clinic	143	143	117	82
Komatipoort Clinic	231	231	184	80
E8 Nkomazi Malaria Mobile 1 (Basic)	307	307	242	79
HPP Nkomazi Malaria 3 (Basic)	128	128	101	79
Mgobodi CHC	9	9	7	78
Richtershoek Clinic	23	23	18	78

Problem 2:

- The South African COVID-19 Modelling Consortium was formed and provided projections of estimated COVID-19 cases, hospitalisations and deaths to support national and provincial response.
- How do we better visualise and use the estimates for training the rapid responders and other decision-makers and actors?



https://pubmed.ncbi.nlm.nih.gov/37093784

NCEM Dashboard

[redacted]

NCEM Dashboard WELCOME PROJECTIONS RESOURCES LOGIN

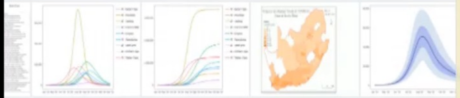
NCEM DASHBOARD

South Africa's first case of COVID-19 was recorded on 5 March 2020, with confirmed cases increasing to 1,353 by 31 March. The simulation period for the NCEM Dashboard begins on 1 April.

The National COVID-19 Epi Model (NCEM) Dashboard has been developed by the South African COVID-19 Modelling Consortium to provide interactive projections of estimated COVID-19 cases, hospitalisations and deaths to support policy and planning in South Africa over the coming months.

The projections are generated using the NCEM mathematical modelling sim... African data and using parameter estimates jointly agreed upon by the SA COVID-19 Modelling Consortium.

Due to the rapidly changing nature of the outbreak globally and in South Africa, the projections are updated regularly as new data become available.



Disclaimer

Due to the rapidly changing nature of the outbreak globally and in South Africa, the projections are updated regularly and should be interpreted with caution. The models have been developed using data that is subject to a high degree of uncertainty. All models are simplifications of reality that are designed to describe and predict system behaviour and are justified by the assumptions and data with which they are developed.

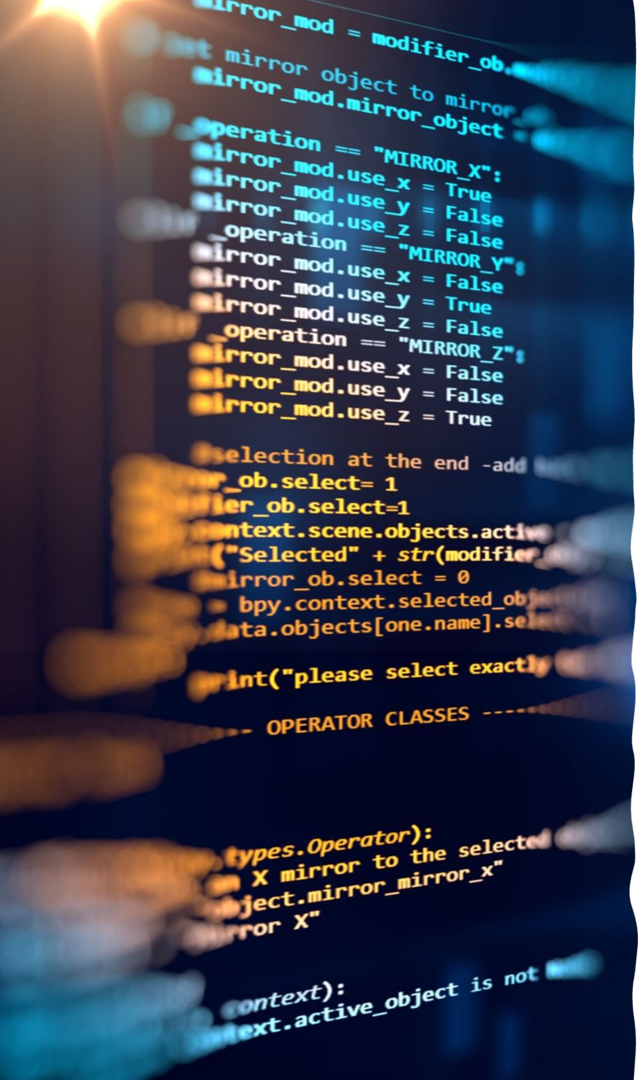
For official statistics and information on COVID-19, please visit http://www.sacoronavirus.co.za

PLEASE NOTE THAT THIS IS A PRELIMINARY RELEASE AND SHOULD BE TREATED AS CONFIDENTIAL

This app may time-out if left idle too long, which will cause the screen to grey-out. To use the app again, refresh the page. This will reset all previously-selected input options.

This app is best supported by Chrome and Edge browsers

I understand and wish to continue.



Opportunities of using R for visualisations

- Versatile language e.g., analysis, reproducible reports, powerpoints, webs, blogs, developing simple web apps
- Free open-source tools e.g. R programming, provide room for different professionals, organisations or groups of individuals without computer sciences background (who are interested in coding) to package their innovative solutions for the public good
- Foster collaborations (Mathematicians, Epidemiologists, policymakers etc). working together to address public health challenges

Questions & Answers

MASHA

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

HDRGlobal
Health Data Research

icddr,b

THE GLOBAL HEALTH NETWORK
Enabling research by sharing knowledge

Thank you

MASHA

Ministério da Saúde
FIOCRUZ
Fundação Oswaldo Cruz

HDRGlobal
Health Data Research

icddr,b

THE GLOBAL HEALTH NETWORK
Enabling research by sharing knowledge