Exploratory Data Analysis

1 Personal Information

Name: Alexandra de Nooy
Student ID: 14581728
Email: alexandra.de.nooy@student.uva.nl or denooy.alex@gmail.com
Submitted on: 22/03/2023
Github link:

2 Data Context

This exploratory data analysis is conducted usinf R and RStudio. There are two main sets of data to be considered, linked to the two main sections of the project.

The first set of data represents the baseline or standard of care output runs produced by the existing patient pathway model. For the baseline there are 15 data output files (in csv format). Multiple files had been generated to account for model stochasticity, with the results of each simulation being based on a different set of random probabilities. The data files have a consistent structure and represent the population of indiviudals who move through the TB diagnostic patient pathway. In this, each column represents either a patient disease status or a point in the patient pathway that the indiviudal may or may not have reached.

The second data set consists of TB burden estimates for Kenya produced by the World Heath Organization as well as the accompanying data dictionary. The estimates cover a range of data variables and their estimated values between the years 2000 and 2022. Several key variables include estimates on TB incidence (new cases), notifications (diagnoses) and deaths. Estimates are also provided for different groups of individuals (for example HIV postive patients) and for different tpes of TB.

3 Data Description: Baseline TB model

4 Data Description: WHO Tuberculosis Data

4.1 Load and merge WHO TB burden data and data dictionary

```
# Path to directory
basePath = "/Users/adenooy/Library/CloudStorage/OneDrive-Personal/UVA/Thesis/MSc-Thesis/"

# Load data dictionary
datadict = read.csv(paste(basePath, "data/dynamic/TB_data_dictionary_2024-01-30.csv",
    sep = ""))
colnames(datadict)
```

```
## [1] "variable_name" "dataset"
## [3] "code_list"     "definition"
```

```r
print(datadict[1:3, ])
```

```
##     variable_name dataset code_list
## 1 budget_cpp_dstb  Budget
## 2  budget_cpp_mdr  Budget
## 3  budget_cpp_tpt  Budget
##
## 1 Average cost of drugs budgeted per patient for drug-susceptible TB treatment, excluding buffer sto
## 2                 Average cost of drugs budgeted per patient for MDR-TB treatment, excluding buffer sto
## 3        Average cost of drugs budgeted per patient for  TB preventive treatment, excluding buffer sto
```

```r
# Load TB data
tb_estimates = read_excel(paste(basePath, "data/dynamic/kenya_tb_burden.xlsx",
    sep = ""))
colnames(tb_estimates)
```

```
##  [1] "country"
##  [2] "iso2"
##  [3] "iso3"
##  [4] "iso_numeric"
##  [5] "g_whoregion"
##  [6] "year"
##  [7] "e_pop_num"
##  [8] "e_inc_100k"
##  [9] "e_inc_100k_lo"
## [10] "e_inc_100k_hi"
## [11] "e_inc_num"
## [12] "e_inc_num_lo"
## [13] "e_inc_num_hi"
## [14] "e_tbhiv_prct"
## [15] "e_tbhiv_prct_lo"
## [16] "e_tbhiv_prct_hi"
## [17] "e_inc_tbhiv_100k"
## [18] "e_inc_tbhiv_100k_lo"
## [19] "e_inc_tbhiv_100k_hi"
## [20] "e_inc_tbhiv_num"
## [21] "e_inc_tbhiv_num_lo"
## [22] "e_inc_tbhiv_num_hi"
## [23] "e_mort_exc_tbhiv_100k"
## [24] "e_mort_exc_tbhiv_100k_lo"
## [25] "e_mort_exc_tbhiv_100k_hi"
## [26] "e_mort_exc_tbhiv_num"
## [27] "e_mort_exc_tbhiv_num_lo"
## [28] "e_mort_exc_tbhiv_num_hi"
## [29] "e_mort_tbhiv_100k"
## [30] "e_mort_tbhiv_100k_lo"
## [31] "e_mort_tbhiv_100k_hi"
## [32] "e_mort_tbhiv_num"
## [33] "e_mort_tbhiv_num_lo"
## [34] "e_mort_tbhiv_num_hi"
```

```
## [35] "e_mort_100k"
## [36] "e_mort_100k_lo"
## [37] "e_mort_100k_hi"
## [38] "e_mort_num"
## [39] "e_mort_num_lo"
## [40] "e_mort_num_hi"
## [41] "cfr"
## [42] "cfr_lo"
## [43] "cfr_hi"
## [44] "cfr_pct"
## [45] "cfr_pct_lo"
## [46] "cfr_pct_hi"
## [47] "c_newinc_100k"
## [48] "c_cdr"
## [49] "c_cdr_lo"
## [50] "c_cdr_hi"
```

```r
print(tb_estimates[1:3, ])
```

```
## # A tibble: 3 x 50
##   country iso2  iso3  iso_numeric g_whoreg~1  year
##   <chr>   <chr> <chr>       <dbl> <chr>      <dbl>
## 1 Kenya   KE    KEN           404 AFR         2000
## 2 Kenya   KE    KEN           404 AFR         2001
## 3 Kenya   KE    KEN           404 AFR         2002
## # ... with 44 more variables: e_pop_num <dbl>,
## #   e_inc_100k <dbl>, e_inc_100k_lo <dbl>,
## #   e_inc_100k_hi <dbl>, e_inc_num <dbl>,
## #   e_inc_num_lo <dbl>, e_inc_num_hi <dbl>,
## #   e_tbhiv_prct <dbl>, e_tbhiv_prct_lo <dbl>,
## #   e_tbhiv_prct_hi <dbl>,
## #   e_inc_tbhiv_100k <dbl>, ...
```

```r
# Merge tb data with data dictionary
tbData = tb_estimates %>%
    gather("variable_name", "value", 7:50) %>%
    left_join(datadict)
```

```
## Joining, by = "variable_name"
```

```r
# remove unnecessary regional columns, blank code_list
# column
tbData = subset(tbData, select = -c(iso2, iso3, iso_numeric,
    g_whoregion, code_list))
print(tbData[1:5, ])
```

```
## # A tibble: 5 x 6
##   country  year variable_~1  value dataset defin~2
##   <chr>   <dbl> <chr>        <dbl> <chr>   <chr>
## 1 Kenya    2000 e_pop_num   3.09e7 Estima~ Estima~
## 2 Kenya    2001 e_pop_num   3.18e7 Estima~ Estima~
## 3 Kenya    2002 e_pop_num   3.28e7 Estima~ Estima~
```

```
## 4 Kenya      2003 e_pop_num     3.38e7 Estima~ Estima~
## 5 Kenya      2004 e_pop_num     3.48e7 Estima~ Estima~
## # ... with abbreviated variable names
## #   1: variable_name, 2: definition
```

4.2   Exploring new incident infections (all infections and HIV)

Incident infections are the number of estimated people being infected with TB each year. The number of new infections is an estimate and is different from the number of reported cases or diagnoses - which is reliant on the identification, testing and treating of people with TB. This data represent a key element in the transmision model and it is important in understanding the past dynamics of TB in kenya and provides an idea on the current trend.

HIV is an important factor to consider, given that Kenya has relatively high HIV/TB coinfection and because HIV impacts the likelihood of contracting TB, becoming infectious or of becoming severely ill.

```r
# select relevant variables related to incidence
inc_data = tbData %>%
    filter(variable_name %in% c("e_inc_num", "e_inc_num_lo",
        "e_inc_num_hi")) %>%
    mutate(var = "e_inc_num")
hiv_inc = tbData %>%
    filter(variable_name %in% c("e_inc_tbhiv_num", "e_inc_tbhiv_num_lo",
        "e_inc_tbhiv_num_hi")) %>%
    mutate(var = "e_inc_tbhiv_num")
hiv_perc_inc = tbData %>%
    filter(variable_name %in% c("e_tbhiv_prct", "e_tbhiv_prct_lo",
        "e_tbhiv_prct_hi")) %>%
    mutate(var = "e_tbhiv_prct")

# label upper, lower and mean estimates
all_inc = rbind(inc_data, hiv_inc)
all_inc$est_type = "Mean"
all_inc$est_type[grepl("_lo", all_inc$variable_name, fixed = TRUE) ==
    TRUE] = "Lower"
all_inc$est_type[grepl("_hi", all_inc$variable_name, fixed = TRUE) ==
    TRUE] = "Upper"

hiv_perc_inc$est_type = "Mean"
hiv_perc_inc$est_type[grepl("_lo", hiv_perc_inc$variable_name,
    fixed = TRUE) == TRUE] = "Lower"
hiv_perc_inc$est_type[grepl("_hi", hiv_perc_inc$variable_name,
    fixed = TRUE) == TRUE] = "Upper"

# Incident cases (all and HIV)
ggplot(all_inc, aes(x = year, y = value, group = variable_name,
    color = est_type)) + geom_point() + geom_line() + theme_bw() +
    xlab("Year") + ylab("Number of cases") + labs(title = "Estimated number of new cases per year") +
    theme(legend.position = "bottom") + facet_wrap(. ~ var)
```
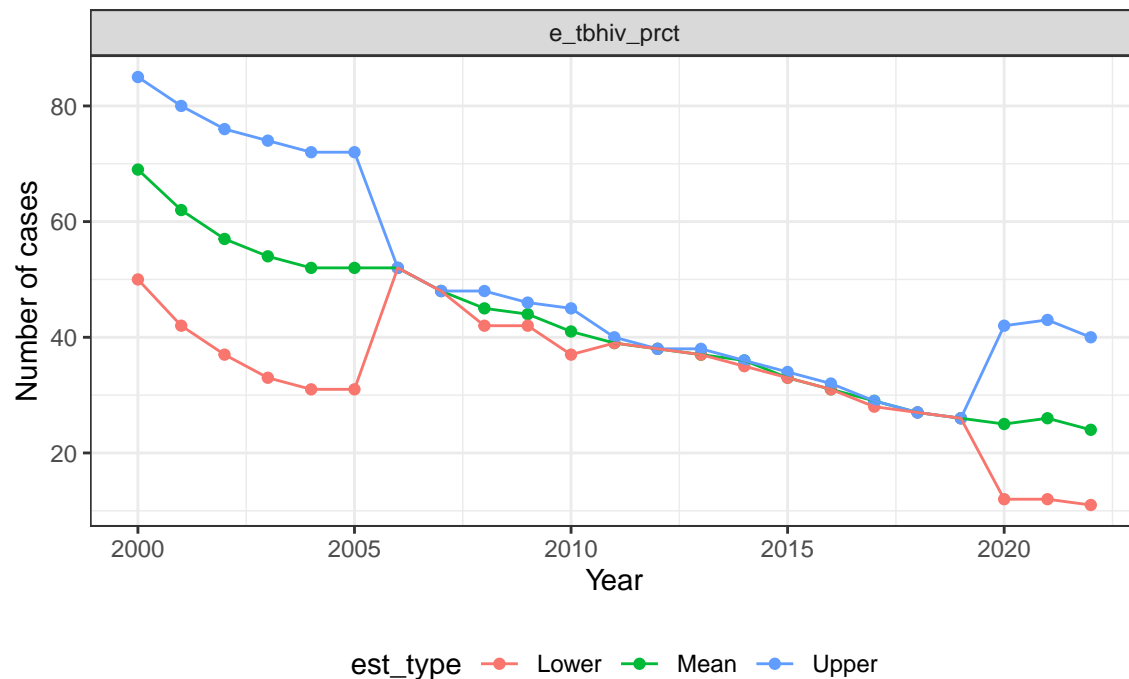
Estimated number of new cases per year
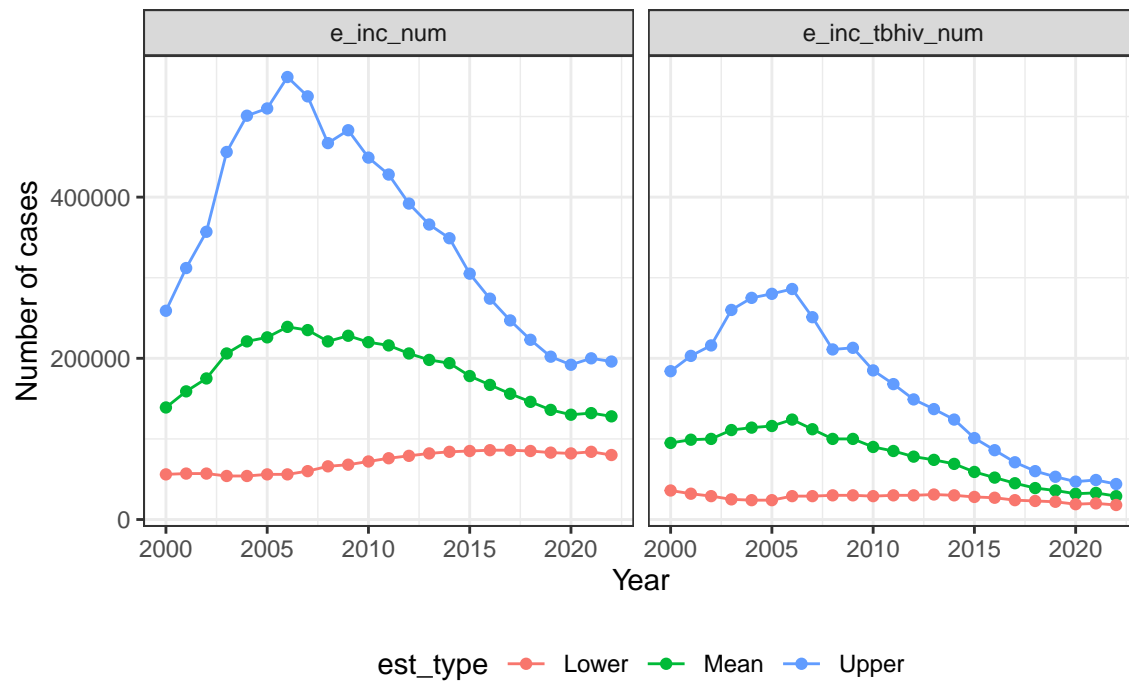

```r
# Percentage of new cases HIV positive
ggplot(hiv_perc_inc, aes(x = year, y = value, group = variable_name,
    color = est_type)) + geom_point() + geom_line() + theme_bw() +
    xlab("Year") + ylab("Number of cases") + labs(title = "Estimated number of new cases per year") +
    theme(legend.position = "bottom") + facet_wrap(. ~ var)
```

Estimated number of new cases per year

4.3 Exploring mean estimates of key factors - population, incidence, case detection, mortality (for whole population)

Key definitions from WHO indicator metadata registry and estimate methedology appendix
- Case Detection rate (%) : Proportion of estimated new and relapse TB (incident) cases diagnosed in a year
-Number of deaths: Product of incidence and case fatality rate
-Case fatality rate: risk of death among people with active (incident) TB, adapte dto account fro low covergae/reporting

```
# Collect key factors relevant to understanding the
# dynamics of TB transmission
key_fact = tbData %>%
    filter(variable_name %in% c("e_pop_num", "e_inc_num",
        "e_inc_rr_num", "e_mort_num", "c_cdr", "cfr", "cfr_pct")) %>%
    select(variable_name, year, value) %>%
    spread(variable_name, value)


# Conduct additional caluclations fo rates or use
# rates to estimate numbers
key_fact$calc_cfr_inc = 100 * (key_fact$e_mort_num/key_fact$e_inc_num)
key_fact$calc_case_detect = key_fact$c_cdr/100 * key_fact$e_inc_num
key_fact$calc_cfr_case = 100 * (key_fact$e_mort_num/key_fact$calc_case_detect)
key_fact$deaths_cal = key_fact$cfr * key_fact$e_inc_num

# Add user friendly labels
```

```
variable = c("e_pop_num", "e_inc_num", "e_inc_rr_num", "e_mort_num",
    "c_cdr", "cfr", "cfr_pct", "calc_case_detect")
labels = c("Population", "Incident Cases", "Incident RR Cases",
    "Deaths", "Case Detection Rate", "Case fatality Rate",
    "Case Fatality Rate (%)", "Cases Detected")
labs = data.frame(cbind(variable, labels))

# Plot key factors with absolute numbers (other than
# population)
key_fact_num = key_fact %>%
    select(year, e_inc_num, calc_case_detect, e_mort_num) %>%
    gather("variable", "value", 2:4) %>%
    left_join(labs)
```

```
## Joining, by = "variable"
```

```
ggplot(key_fact_num, aes(x = year, y = value, group = variable,
    colour = labels)) + geom_point() + geom_line() + theme_bw() +
    labs(title = "Number of incident cases, cases detected and deaths") +
    ylab("Number of individuals") + theme(legend.position = "bottom")
```
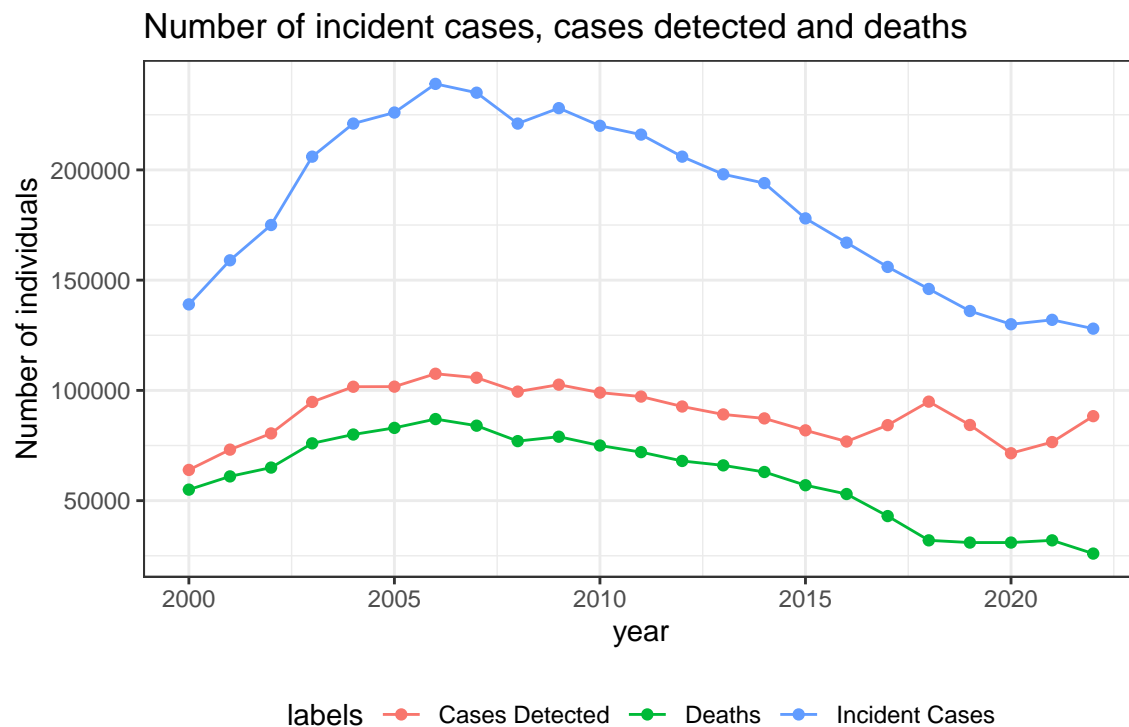


```
# Plot key factors which are rates
key_fact_rate = key_fact %>%
    select(year, c_cdr, cfr_pct) %>%
    gather("variable", "value", 2:3) %>%
    left_join(labs)
```

```
## Joining, by = "variable"
```

```
ggplot(key_fact_rate, aes(x = year, y = value, group = variable,
    colour = labels)) + geom_point() + geom_line() + theme_bw() +
    labs(title = "Case detection and case fatality rates over time") +
    ylab("Percent(%)") + theme(legend.position = "bottom")
```

Case detection and case fatality rates over time