

Final_Project_Code

Import Data and Libraries

```
library(readxl)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##   filter, lag
```

```
## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

```
data = read.csv("GlobalHealthStatistics.csv")
```

Exploratory Data Analysis

```
head(data)
```

##	Country	Year	Disease.Name	Disease.Category	Prevalence.Rate....
## 1	Italy	2013	Malaria	Respiratory	0.95
## 2	France	2002	Ebola	Parasitic	12.46
## 3	Turkey	2015	COVID-19	Genetic	0.91
## 4	Indonesia	2011	Parkinson's Disease	Autoimmune	4.68
## 5	Italy	2013	Tuberculosis	Genetic	0.83
## 6	Saudi Arabia	2011	Dengue	Bacterial	10.99
##	Incidence.Rate....	Mortality.Rate....	Age.Group	Gender	Population.Affected
## 1	1.55	8.42	0-18	Male	471007
## 2	8.63	8.75	61+	Male	634318
## 3	2.35	6.22	36-60	Male	154878
## 4	6.29	3.99	0-18	Other	446224
## 5	13.59	7.01	61+	Male	472908
## 6	6.49	4.64	61+	Female	479234
##	Healthcare.Access....	Doctors.per.1000	Hospital.Beds.per.1000	Treatment.Type	
## 1	57.74	3.34	7.58	Medication	
## 2	89.21	1.33	5.11	Surgery	
## 3	56.41	4.07	3.49	Vaccination	
## 4	85.20	3.18	8.44	Surgery	
## 5	67.00	4.61	5.90	Medication	
## 6	98.41	3.50	0.62	Therapy	
##	Average.Treatment.Cost..USD.	Availability.of.Vaccines.Treatment			
## 1	21064	No			
## 2	47851	Yes			
## 3	27834	Yes			
## 4	144	Yes			
## 5	8908	Yes			
## 6	42671	Yes			
##	Recovery.Rate....	DALYs	Improvement.in.5.Years....	Per.Capita.Income..USD.	
## 1	91.82	4493	2.16	16886	
## 2	76.65	2366	4.82	80639	
## 3	98.55	41	5.81	12245	
## 4	67.35	3201	2.22	49336	
## 5	50.06	2832	6.93	47701	
## 6	93.17	416	9.83	29597	
##	Education.Index	Urbanization.Rate....			
## 1	0.79	86.02			
## 2	0.74	45.52			
## 3	0.41	40.20			
## 4	0.49	58.47			
## 5	0.50	48.14			
## 6	0.46	56.50			

```
#convert column to factor
convert_factor <- function(data, colname){
  data[[colname]] = factor(data[[colname]])
}

convert_factor(data, 'Country')
convert_factor(data, 'Disease.Name')
convert_factor(data, 'Disease.Category')
convert_factor(data, 'Age.Group')
convert_factor(data, 'Gender')
convert_factor(data, 'Treatment.Type')
convert_factor(data, 'Availability.of.Vaccines.Treatment')
```

```
table(data$Country)
```

```
##
##      Argentina      Australia      Brazil      Canada      China      France
##      49798          49953          49687          50114          50066          49943
##      Germany        India      Indonesia      Italy        Japan        Mexico
##      50176          49760          49756          49839          49764          50080
##      Nigeria        Russia Saudi Arabia South Africa South Korea Turkey
##      50046          50532          49958          50408          50181          49901
##      UK            USA
##      50125          49913
```

```
table(data$Disease.Name)
```

```
##
## Alzheimer's Disease      Asthma      Cancer      Cholera
##      49823          50122          50285          50249
##      COVID-19      Dengue      Diabetes      Ebola
##      50404          50289          50020          49692
##      Hepatitis      HIV/AIDS      Hypertension      Influenza
##      49970          50274          49391          49919
##      Leprosy      Malaria      Measles Parkinson's Disease
##      50064          49948          49736          49708
##      Polio      Rabies      Tuberculosis      Zika
##      49956          49975          49862          50313
```

```
table(data$Disease.Category)
```

```
##
##      Autoimmune      Bacterial Cardiovascular      Chronic      Genetic
##      91153          90509          90968          90445          91153
##      Infectious      Metabolic      Neurological      Parasitic      Respiratory
##      90764          91332          91000          91178          90588
##      Viral
##      90910
```

```
summary(data)
```

```

##      Country                Year      Disease.Name      Disease.Category
## Length:1000000      Min.      :2000      Length:1000000      Length:1000000
## Class :character      1st Qu.:2006      Class :character      Class :character
## Mode  :character      Median :2012      Mode  :character      Mode  :character
##                               Mean  :2012
##                               3rd Qu.:2018
##                               Max.   :2024
## Prevalence.Rate.... Incidence.Rate.... Mortality.Rate.... Age.Group
## Min.      : 0.10      Min.      : 0.100      Min.      : 0.10      Length:1000000
## 1st Qu.: 5.09      1st Qu.: 3.840      1st Qu.: 2.58      Class :character
## Median :10.04      Median : 7.550      Median : 5.05      Mode  :character
## Mean      :10.05      Mean      : 7.555      Mean      : 5.05
## 3rd Qu.:15.01      3rd Qu.:11.280      3rd Qu.: 7.53
## Max.      :20.00      Max.      :15.000      Max.      :10.00
##      Gender      Population.Affected Healthcare.Access.... Doctors.per.1000
## Length:1000000      Min.      : 1000      Min.      : 50.00      Min.      :0.500
## Class :character      1st Qu.: 250491      1st Qu.: 62.47      1st Qu.:1.620
## Mode  :character      Median : 501041      Median : 75.00      Median :2.750
##                               Mean      : 500735      Mean      : 74.99      Mean      :2.748
##                               3rd Qu.: 750782      3rd Qu.: 87.49      3rd Qu.:3.870
##                               Max.      :1000000      Max.      :100.00      Max.      :5.000
## Hospital.Beds.per.1000 Treatment.Type      Average.Treatment.Cost..USD.
## Min.      : 0.500      Length:1000000      Min.      : 100
## 1st Qu.: 2.870      Class :character      1st Qu.:12538
## Median : 5.240      Mode  :character      Median :24980
## Mean      : 5.246      Mean      :25010
## 3rd Qu.: 7.620      3rd Qu.:37493
## Max.      :10.000      Max.      :50000
## Availability.of.Vaccines.Treatment Recovery.Rate.... DALYs
## Length:1000000      Min.      :50.00      Min.      : 1
## Class :character      1st Qu.:62.22      1st Qu.:1245
## Mode  :character      Median :74.47      Median :2499
##                               Mean      :74.50      Mean      :2499
##                               3rd Qu.:86.78      3rd Qu.:3750
##                               Max.      :99.00      Max.      :5000
## Improvement.in.5.Years.... Per.Capita.Income..USD. Education.Index
## Min.      : 0.000      Min.      : 500      Min.      :0.4000
## 1st Qu.: 2.500      1st Qu.: 25457      1st Qu.:0.5300
## Median : 5.000      Median : 50372      Median :0.6500
## Mean      : 5.003      Mean      : 50311      Mean      :0.6501
## 3rd Qu.: 7.510      3rd Qu.: 75195      3rd Qu.:0.7800
## Max.      :10.000      Max.      :100000      Max.      :0.9000
## Urbanization.Rate....
## Min.      :20.00
## 1st Qu.:37.47
## Median :54.98
## Mean      :54.99
## 3rd Qu.:72.51
## Max.      :90.00

```

```
data = as.data.frame(data)
colnames(data)
```

```
## [1] "Country"          "Year"
## [3] "Disease.Name"      "Disease.Category"
## [5] "Prevalence.Rate...." "Incidence.Rate...."
## [7] "Mortality.Rate...." "Age.Group"
## [9] "Gender"            "Population.Affected"
## [11] "Healthcare.Access...." "Doctors.per.1000"
## [13] "Hospital.Beds.per.1000" "Treatment.Type"
## [15] "Average.Treatment.Cost..USD." "Availability.of.Vaccines.Treatment"
## [17] "Recovery.Rate...." "DALYs"
## [19] "Improvement.in.5.Years...." "Per.Capita.Income..USD."
## [21] "Education.Index" "Urbanization.Rate...."
```

```
#change columns name
colnames(data) <- c('Country', 'Year', 'Disease_Name', 'Disease_Category', 'Prevalence_Rate', 'Incidence_Rate',
                    'Mortality_Rate', 'Age_Group', 'Gender', 'Population_Affected', 'Healthcare_Access',
                    'Doctors_per_1000', 'Hospital_Beds_per_1000', 'Treatment_Type', 'Average_Treatment_Cost',
                    'Availability_of_Vaccines_Treatment', 'Recovery_Rate', 'DALYs', 'Improvement_in_5_Years',
                    'Per_Capita_Income', 'Education_Index', 'Urbanization_Rate')
```

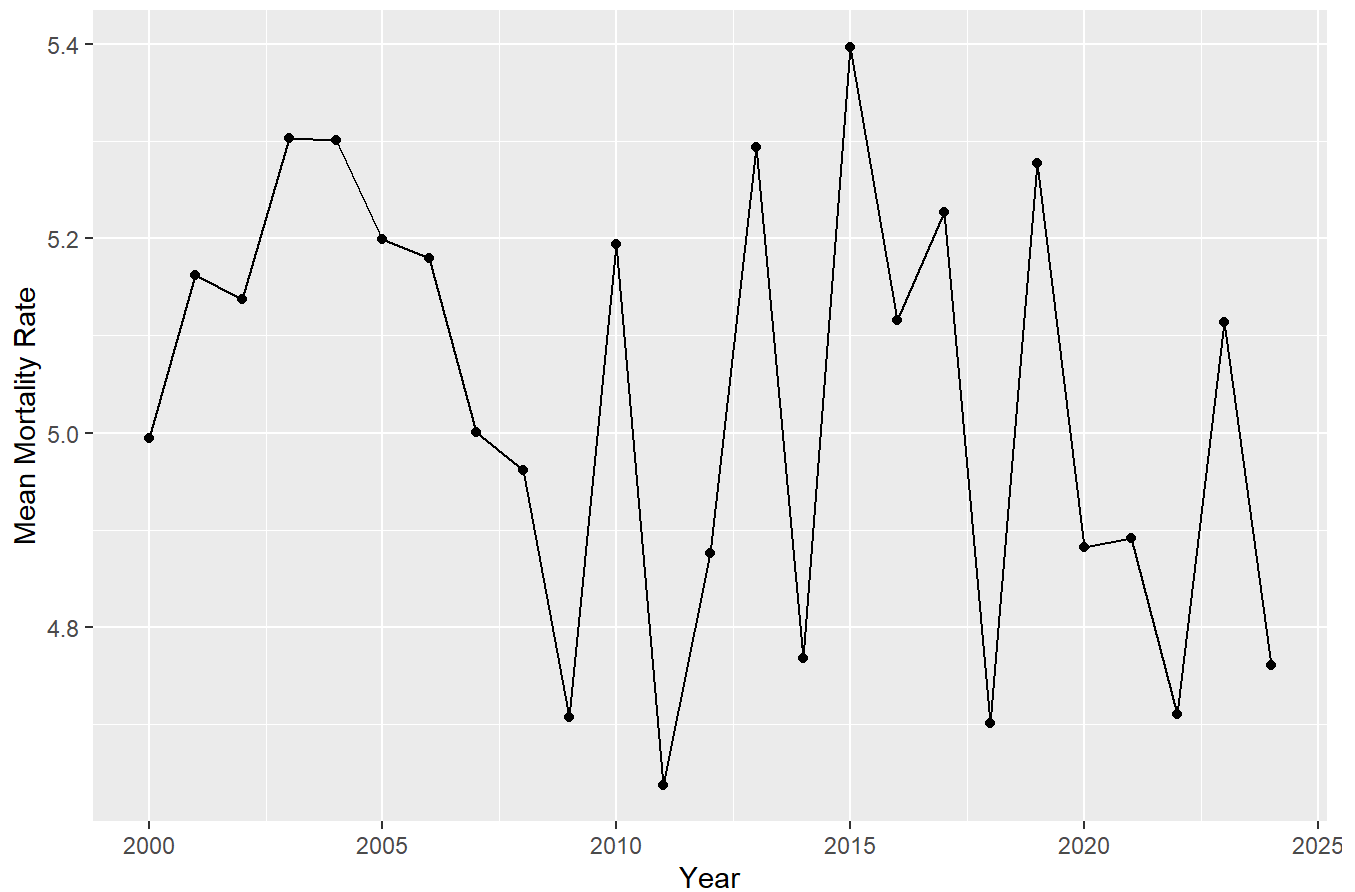
Americans' mortality rate

```
dis_cat_USA = data %>% filter(Country == 'USA') %>% group_by(Disease_Category, Year) %>% summarise(mean_mortality = mean(Mortality_Rate, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'Disease_Category'. You can override using
## the `.groups` argument.
```

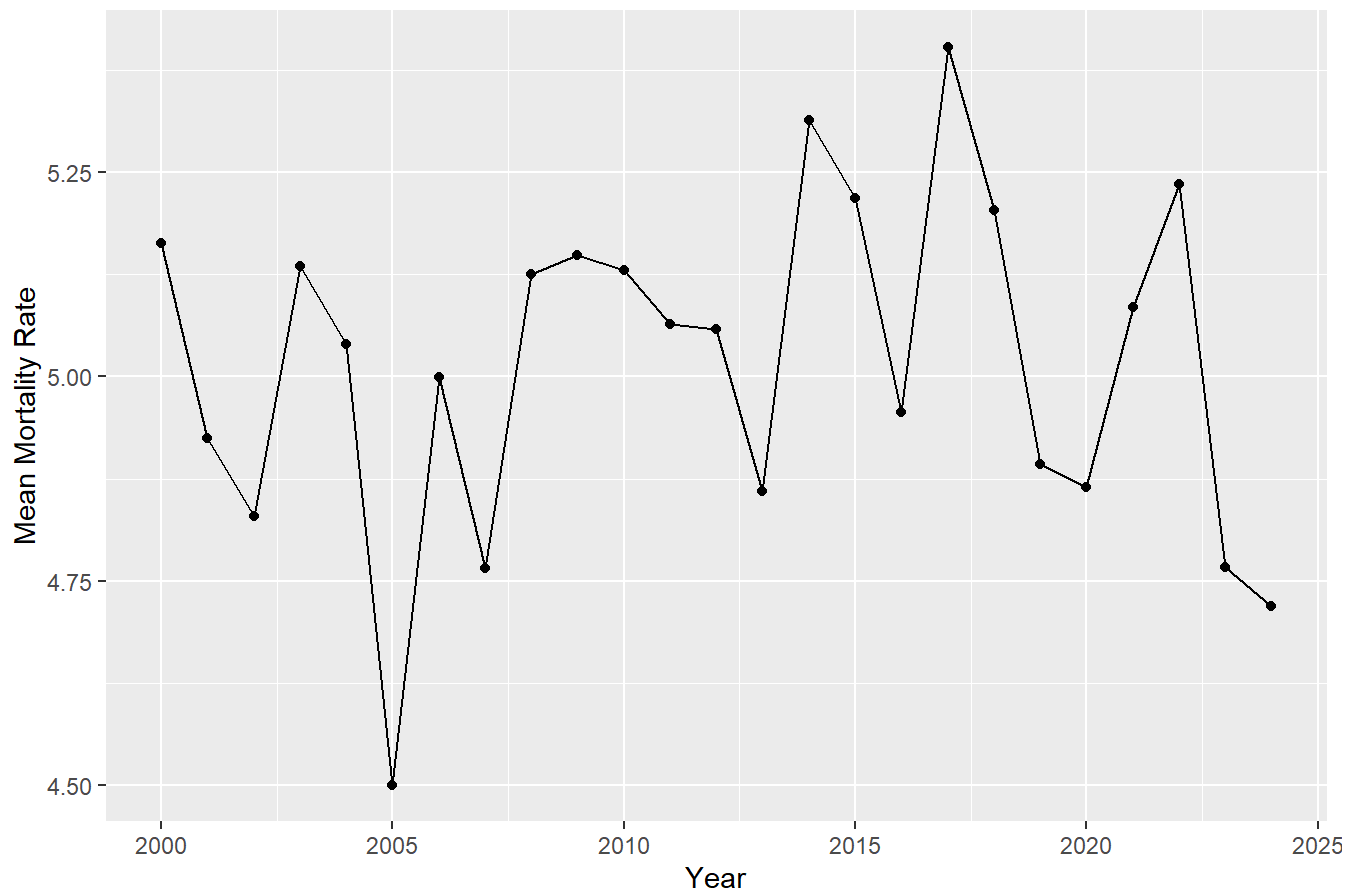
```
# Cardiovascular
dis_cat_USA %>%
  filter(Disease_Category == 'Cardiovascular') %>%
  ggplot(aes(x = Year, y = mean_mortality)) +
  geom_line()+
  geom_point()+
  labs(title = "Mean Mortality Rate by Cardiovascular Category in USA",
       y = "Mean Mortality Rate")
```

Mean Mortality Rate by Cardiovascular Category in USA



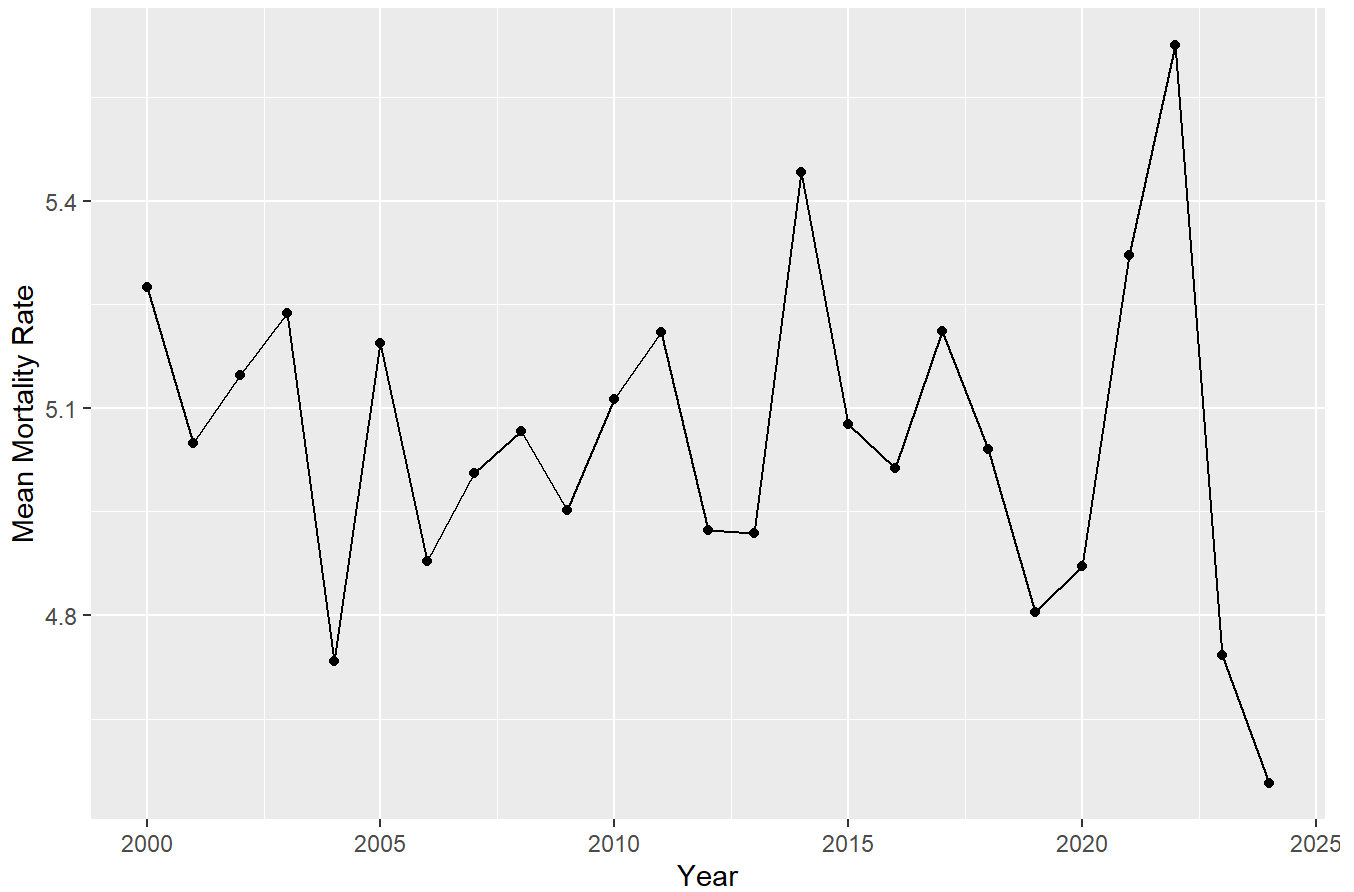
```
# Autoimmune
dis_cat_USA %>%
  filter(Disease_Category == 'Autoimmune') %>%
  ggplot(aes(x = Year, y = mean_mortality)) +
  geom_line()+
  geom_point()+
  labs(title = "Mean Mortality Rate by Autoimmune Category in USA",
        y = "Mean Mortality Rate")
```

Mean Mortality Rate by Autoimmune Category in USA



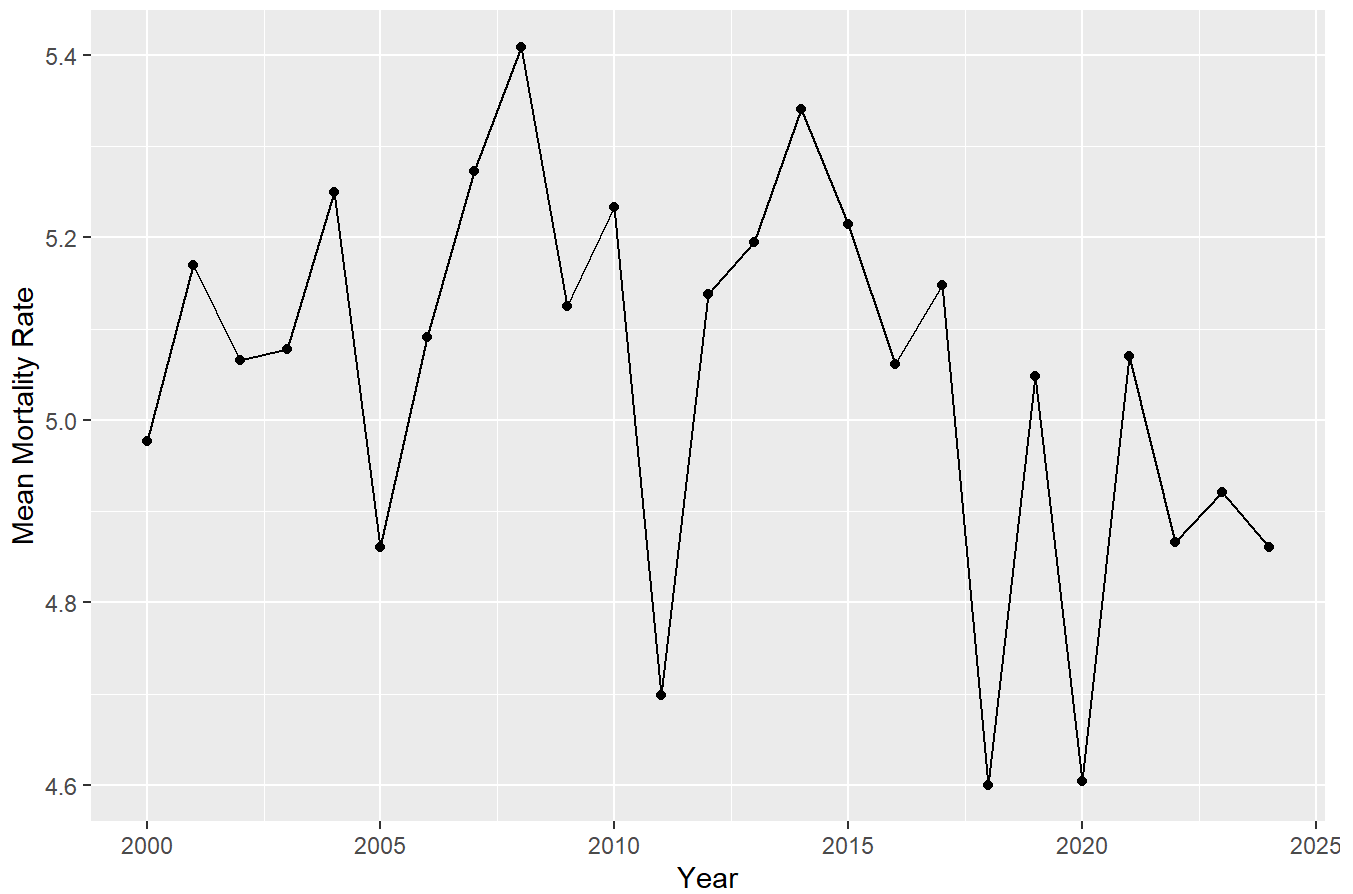
```
# Bacterial
dis_cat_USA %>%
  filter(Disease_Category == 'Bacterial') %>%
  ggplot(aes(x = Year, y = mean_mortality)) +
  geom_line()+
  geom_point()+
  labs(title = "Mean Mortality Rate by Bacterial Category in USA",
        y = "Mean Mortality Rate")
```


Mean Mortality Rate by Bacterial Category in USA



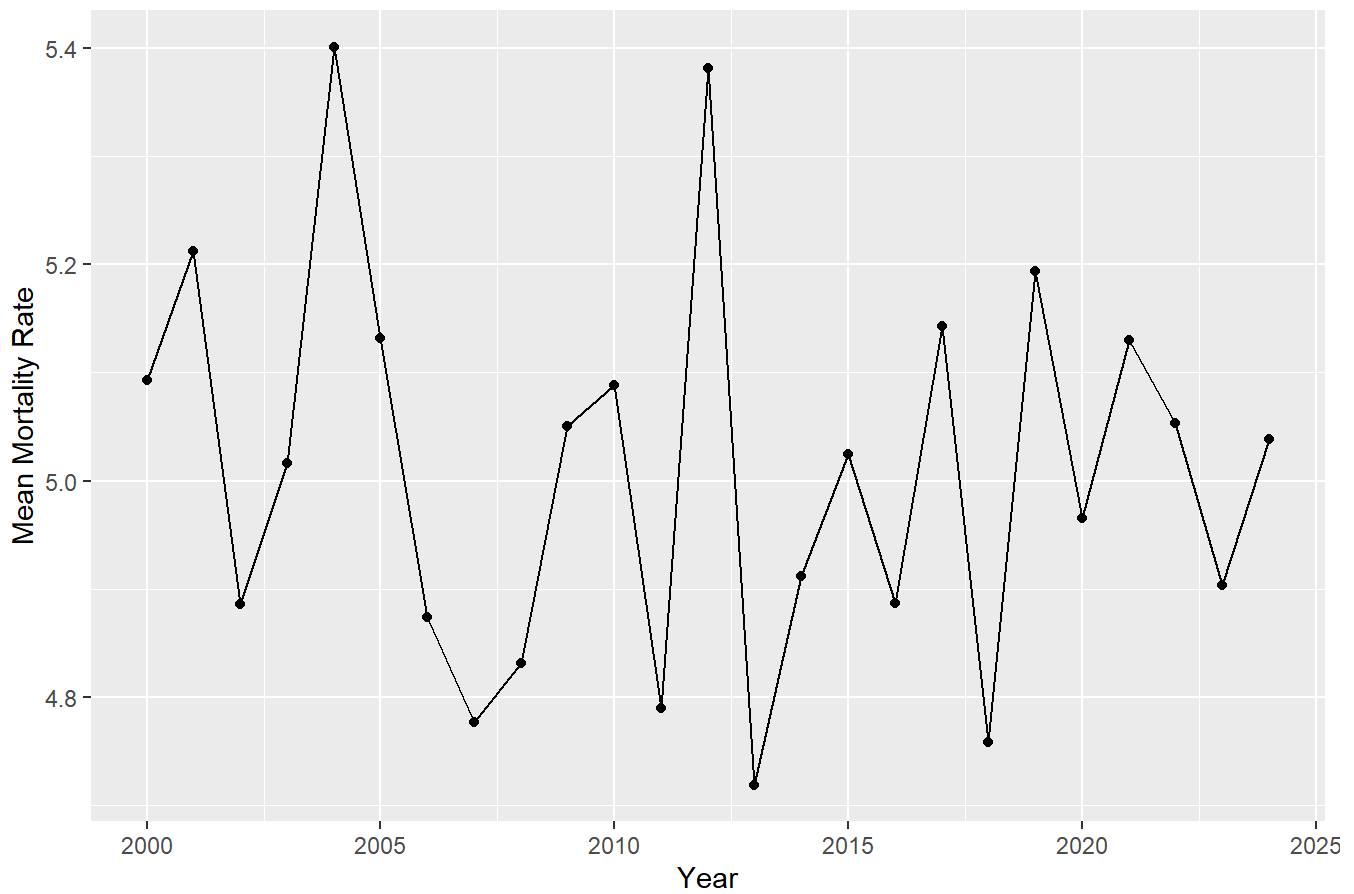
```
# Chronic
dis_cat_USA %>%
  filter(Disease_Category == 'Chronic') %>%
  ggplot(aes(x = Year, y = mean_mortality)) +
  geom_line()+
  geom_point()+
  labs(title = "Mean Mortality Rate by Chronic Category in USA",
        y = "Mean Mortality Rate")
```

Mean Mortality Rate by Chronic Category in USA



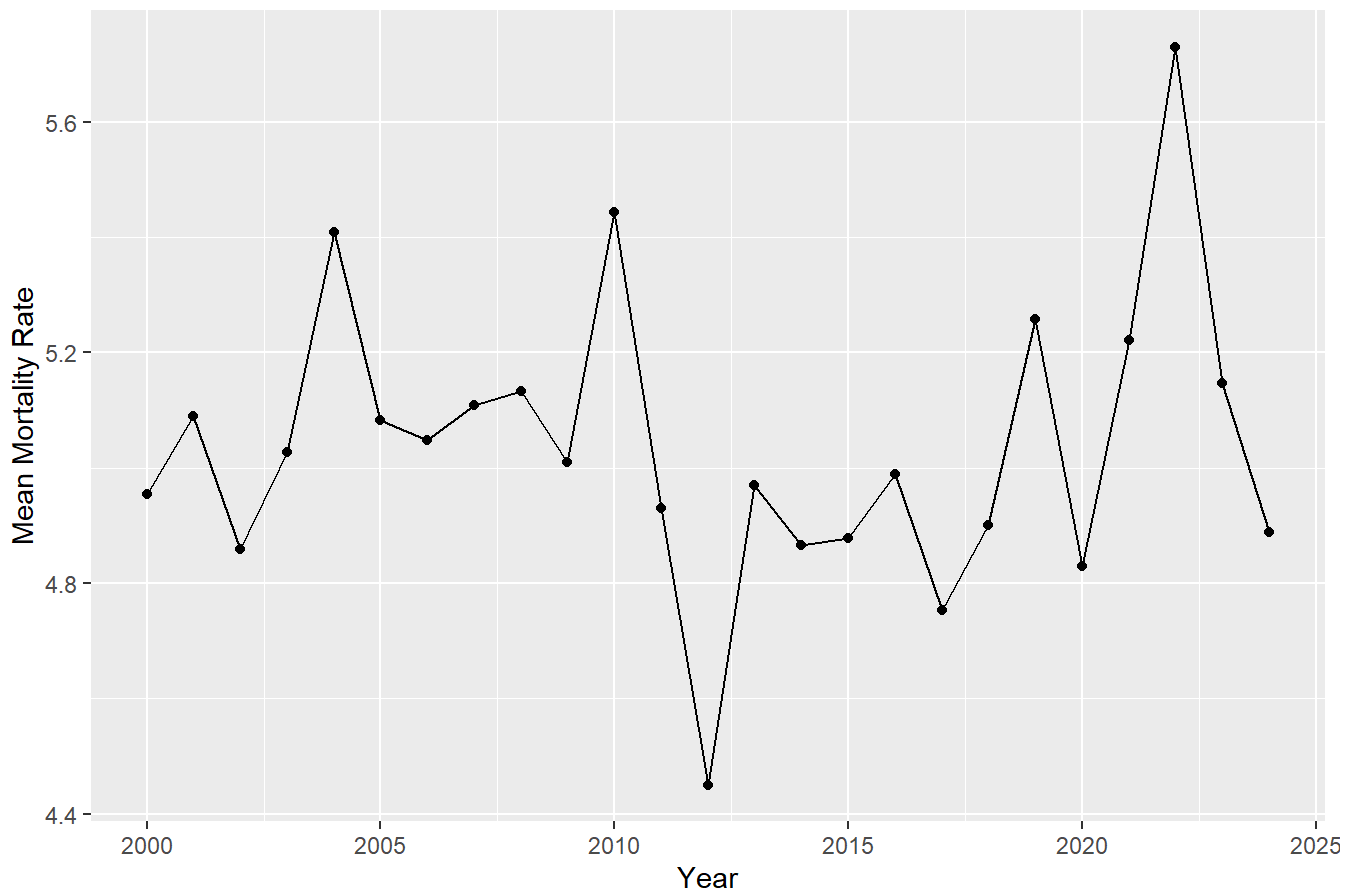
```
# Genetic
dis_cat_USA %>%
  filter(Disease_Category == 'Genetic') %>%
  ggplot(aes(x = Year, y = mean_mortality)) +
  geom_line()+
  geom_point()+
  labs(title = "Mean Mortality Rate by Genetic Category in USA",
        y = "Mean Mortality Rate")
```

Mean Mortality Rate by Genetic Category in USA



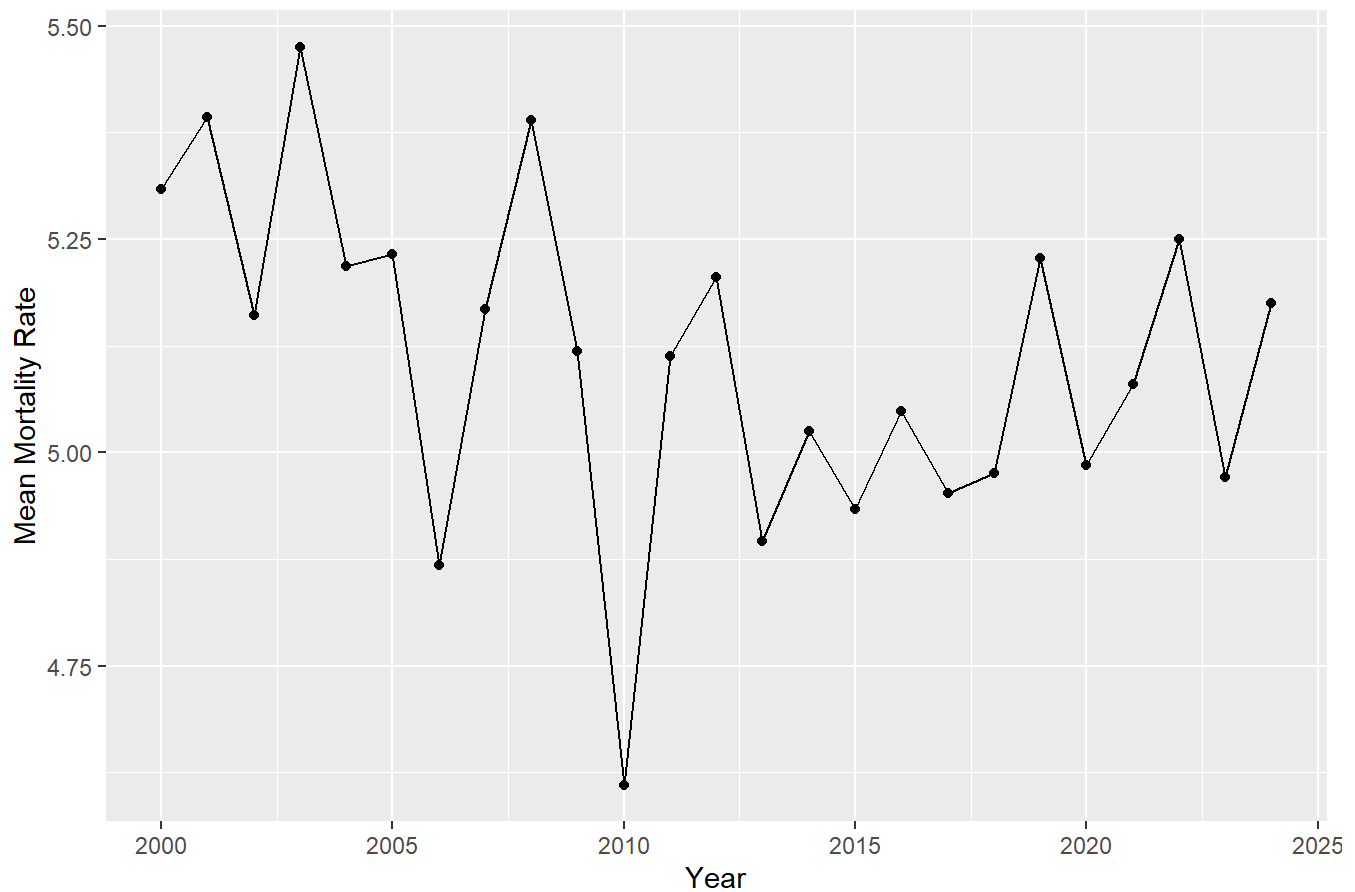
```
# Infectious
dis_cat_USA %>%
  filter(Disease_Category == 'Infectious') %>%
  ggplot(aes(x = Year, y = mean_mortality)) +
  geom_line()+
  geom_point()+
  labs(title = "Mean Mortality Rate by Infectious Category in USA",
        y = "Mean Mortality Rate")
```

Mean Mortality Rate by Infectious Category in USA



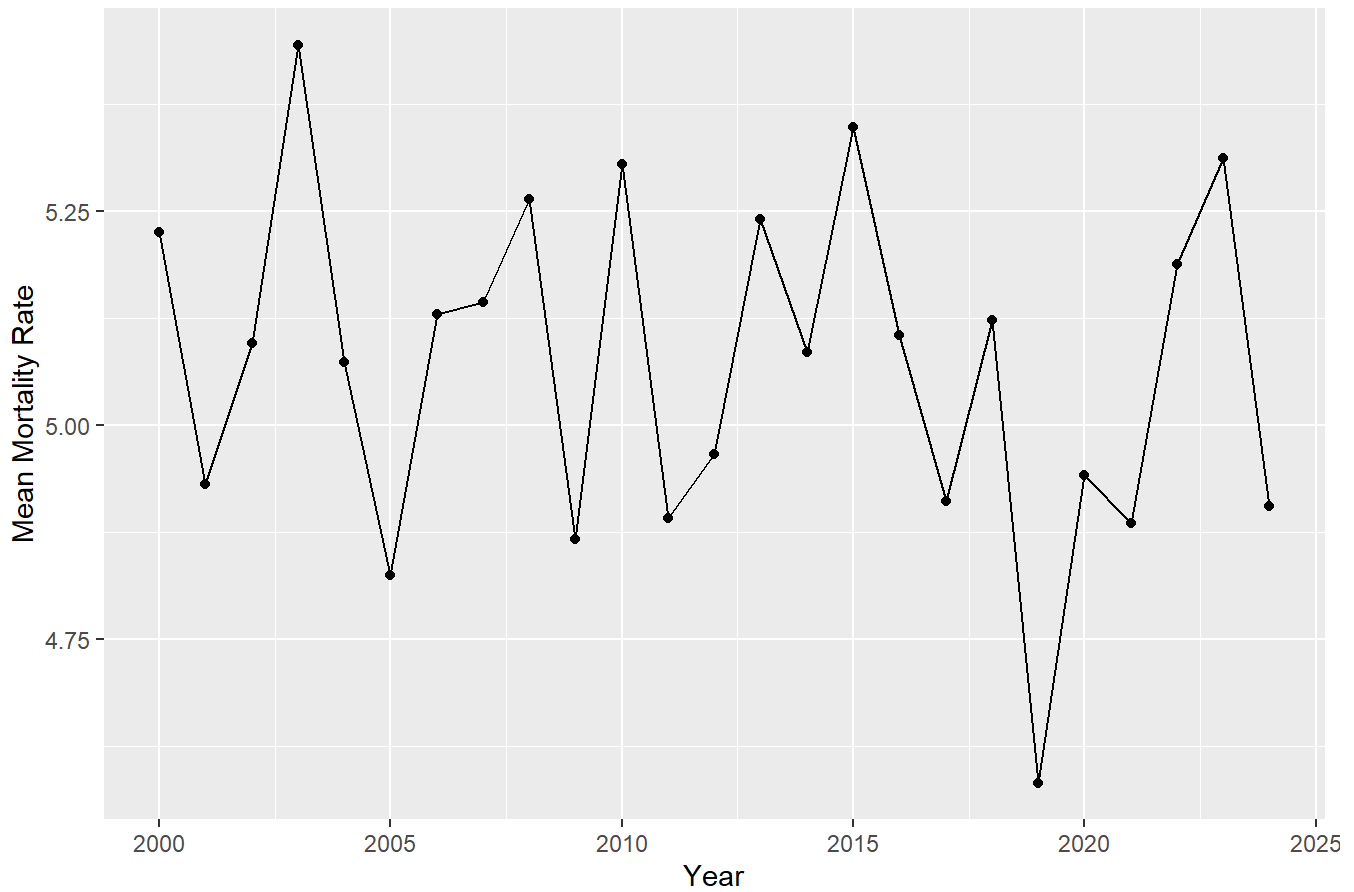
```
# Metabolic
dis_cat_USA %>%
  filter(Disease_Category == 'Metabolic') %>%
  ggplot(aes(x = Year, y = mean_mortality)) +
  geom_line()+
  geom_point()+
  labs(title = "Mean Mortality Rate by Metabolic Category in USA",
        y = "Mean Mortality Rate")
```

Mean Mortality Rate by Metabolic Category in USA



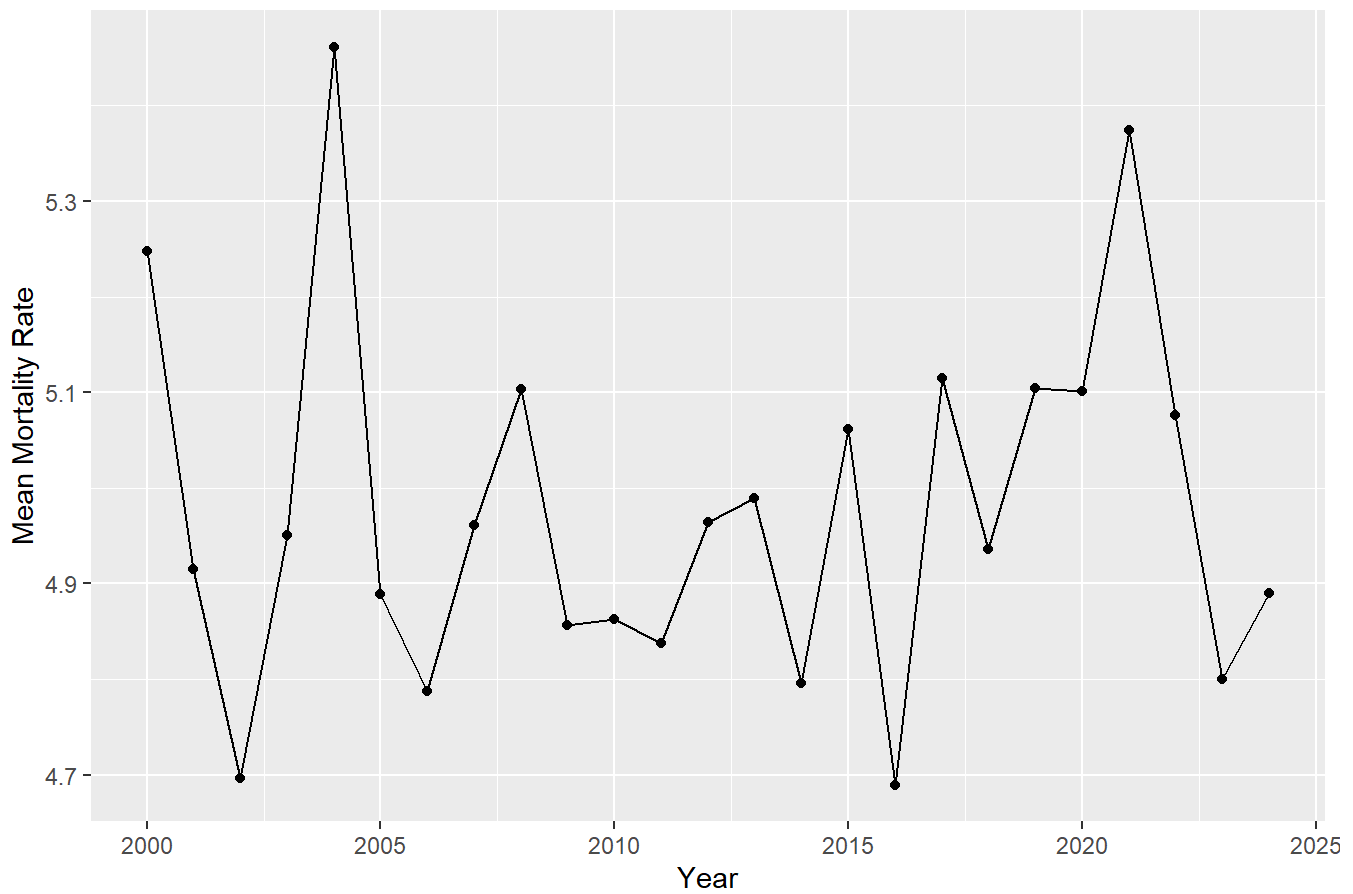
```
# Neurological
dis_cat_USA %>%
  filter(Disease_Category == 'Neurological') %>%
  ggplot(aes(x = Year, y = mean_mortality)) +
  geom_line()+
  geom_point()+
  labs(title = "Mean Mortality Rate by Neurological Category in USA",
        y = "Mean Mortality Rate")
```

Mean Mortality Rate by Neurological Category in USA



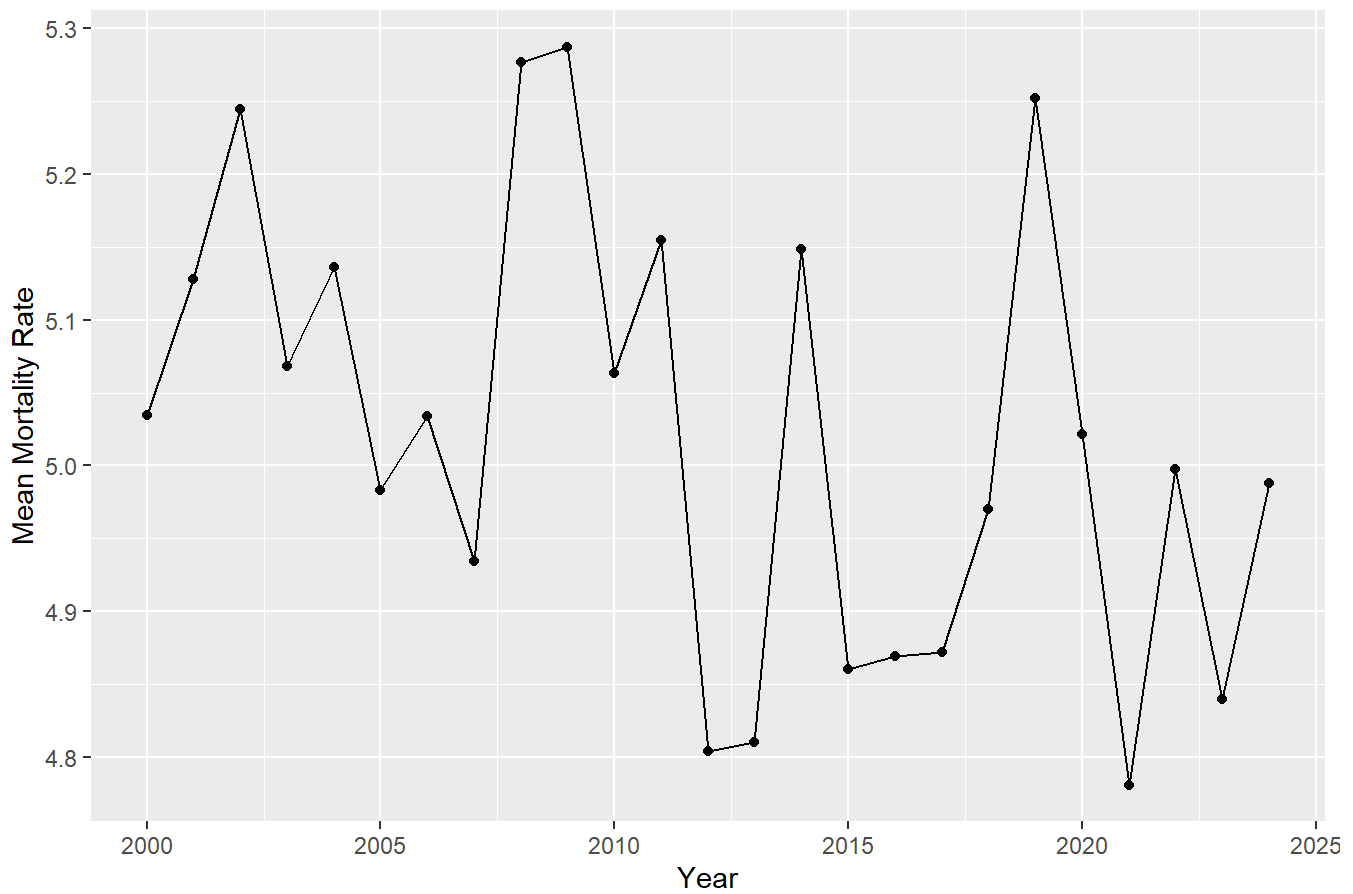
```
# Parasitic
dis_cat_USA %>%
  filter(Disease_Category == 'Parasitic') %>%
  ggplot(aes(x = Year, y = mean_mortality)) +
  geom_line()+
  geom_point()+
  labs(title = "Mean Mortality Rate by Parasitic Category in USA",
        y = "Mean Mortality Rate")
```

Mean Mortality Rate by Parasitic Category in USA



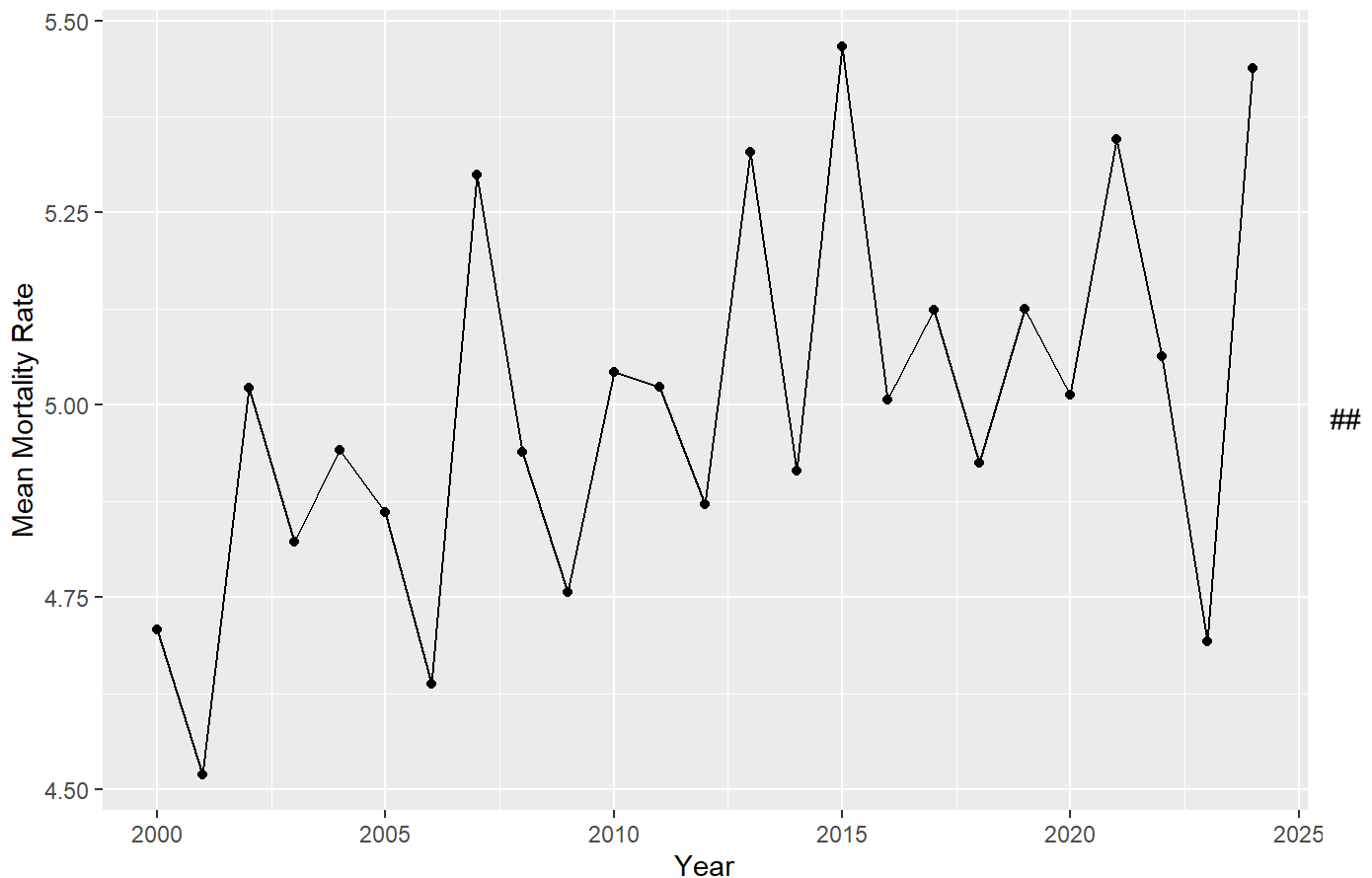
```
# Respiratory
dis_cat_USA %>%
  filter(Disease_Category == 'Respiratory') %>%
  ggplot(aes(x = Year, y = mean_mortality)) +
  geom_line()+
  geom_point()+
  labs(title = "Mean Mortality Rate by Respiratory Category in USA",
        y = "Mean Mortality Rate")
```

Mean Mortality Rate by Respiratory Category in USA



```
# Viral
dis_cat_USA %>%
  filter(Disease_Category == 'Viral') %>%
  ggplot(aes(x = Year, y = mean_mortality)) +
  geom_line()+
  geom_point()+
  labs(title = "Mean Mortality Rate by Viral Category in USA",
        y = "Mean Mortality Rate")
```


Mean Mortality Rate by Viral Category in USA



International Mortality Rate

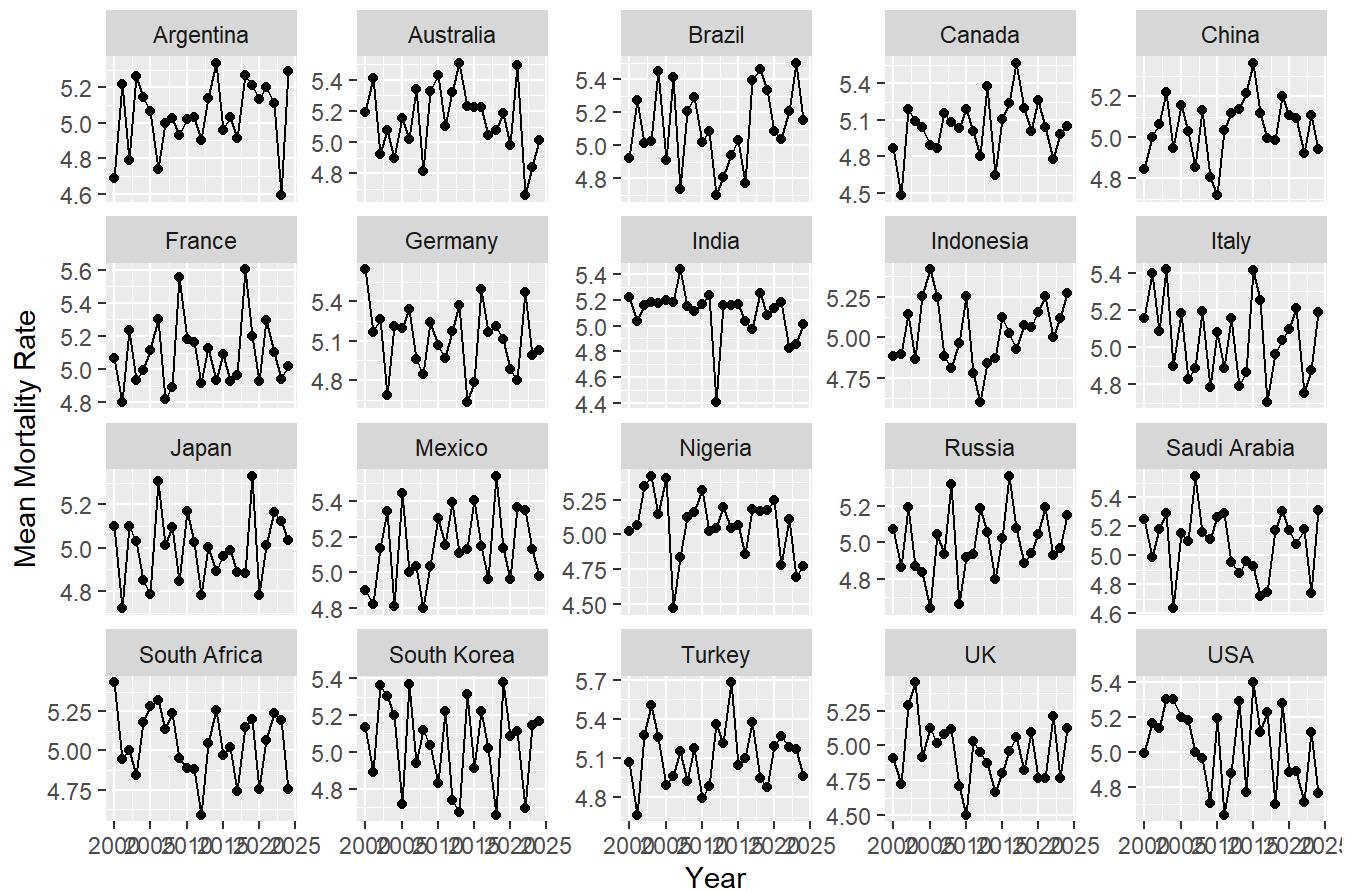
```
dis_cat_inter = data %>% group_by(Disease_Category, Year, Country) %>% summarise(mean_mortality = mean(Mortality_Rate, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'Disease_Category', 'Year'. You can
## override using the `.groups` argument.
```

```
plot_international <- function(disease){
  dis_cat_inter %>%
    filter(Disease_Category == disease) %>%
    ggplot(aes(x = Year, y = mean_mortality)) +
    geom_line()+
    geom_point()+
    facet_wrap(~ Country, scales = "free_y")+
    labs(title = paste("Mean Mortality Rate by", disease , "Category in 20 countries"),
         y = "Mean Mortality Rate")
}
```

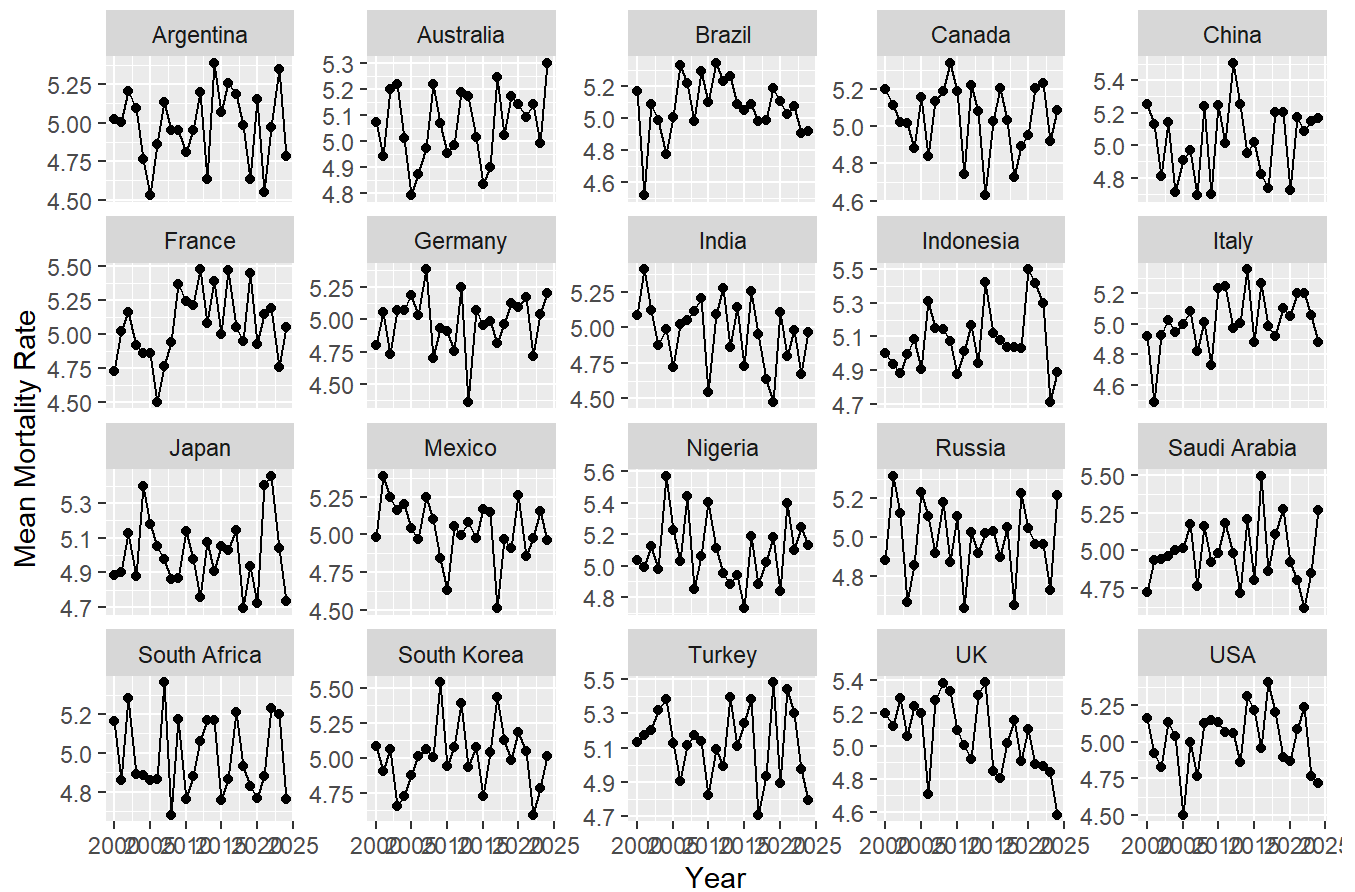
```
plot_international('Cardiovascular')
```

Mean Mortality Rate by Cardiovascular Category in 20 countries



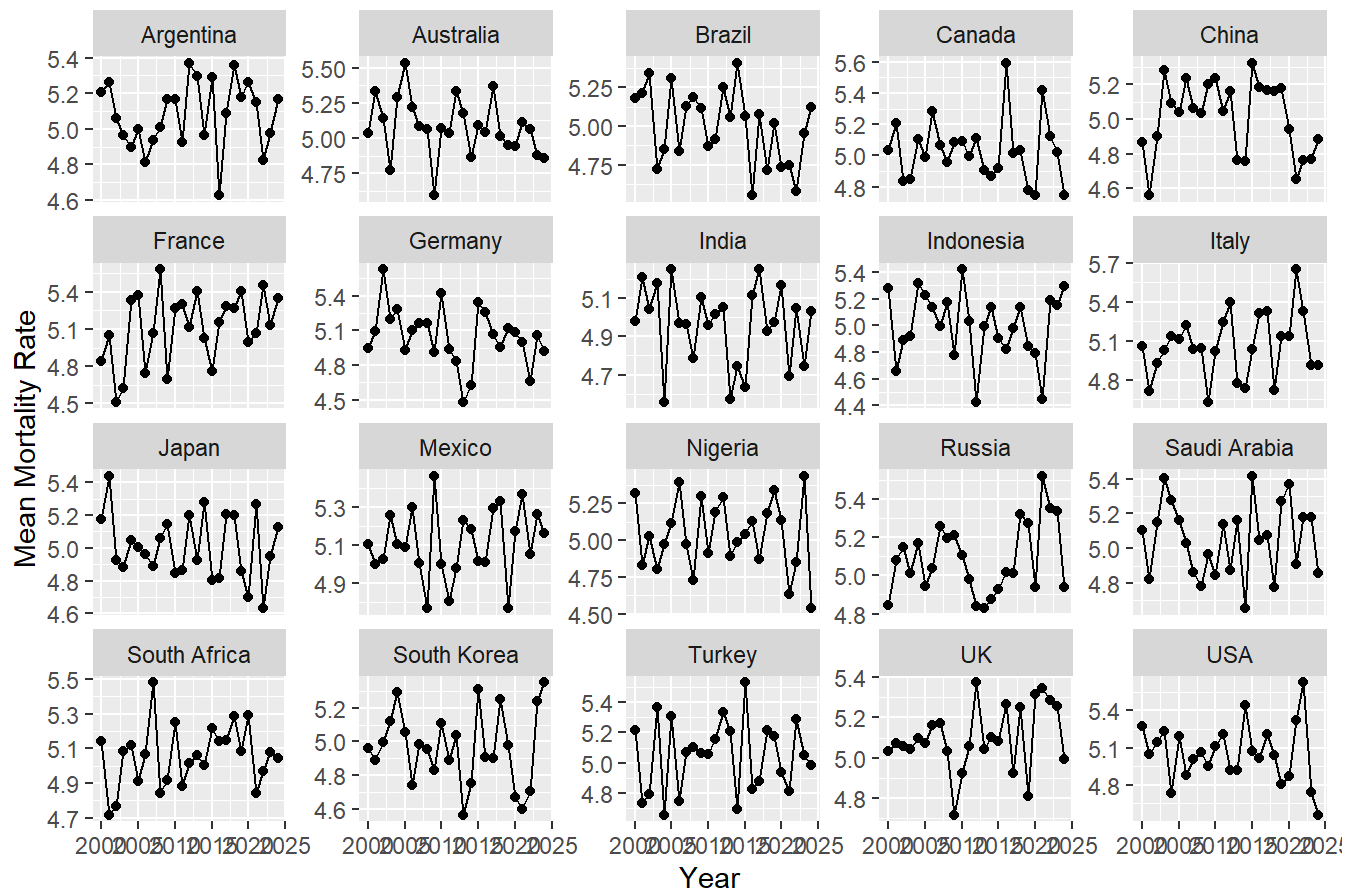
```
plot_international('Autoimmune')
```

Mean Mortality Rate by Autoimmune Category in 20 countries



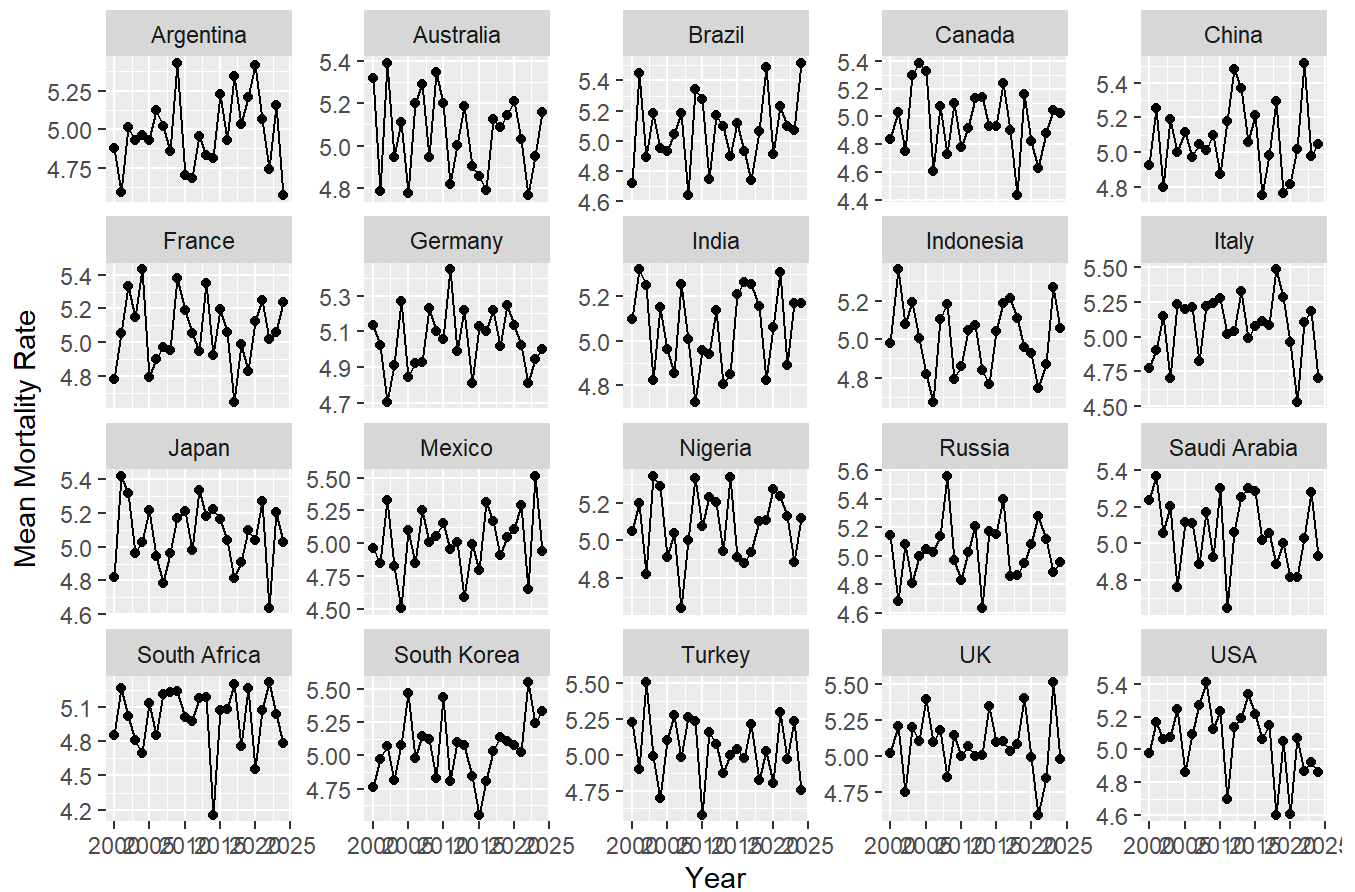
```
plot_international('Bacterial')
```

Mean Mortality Rate by Bacterial Category in 20 countries



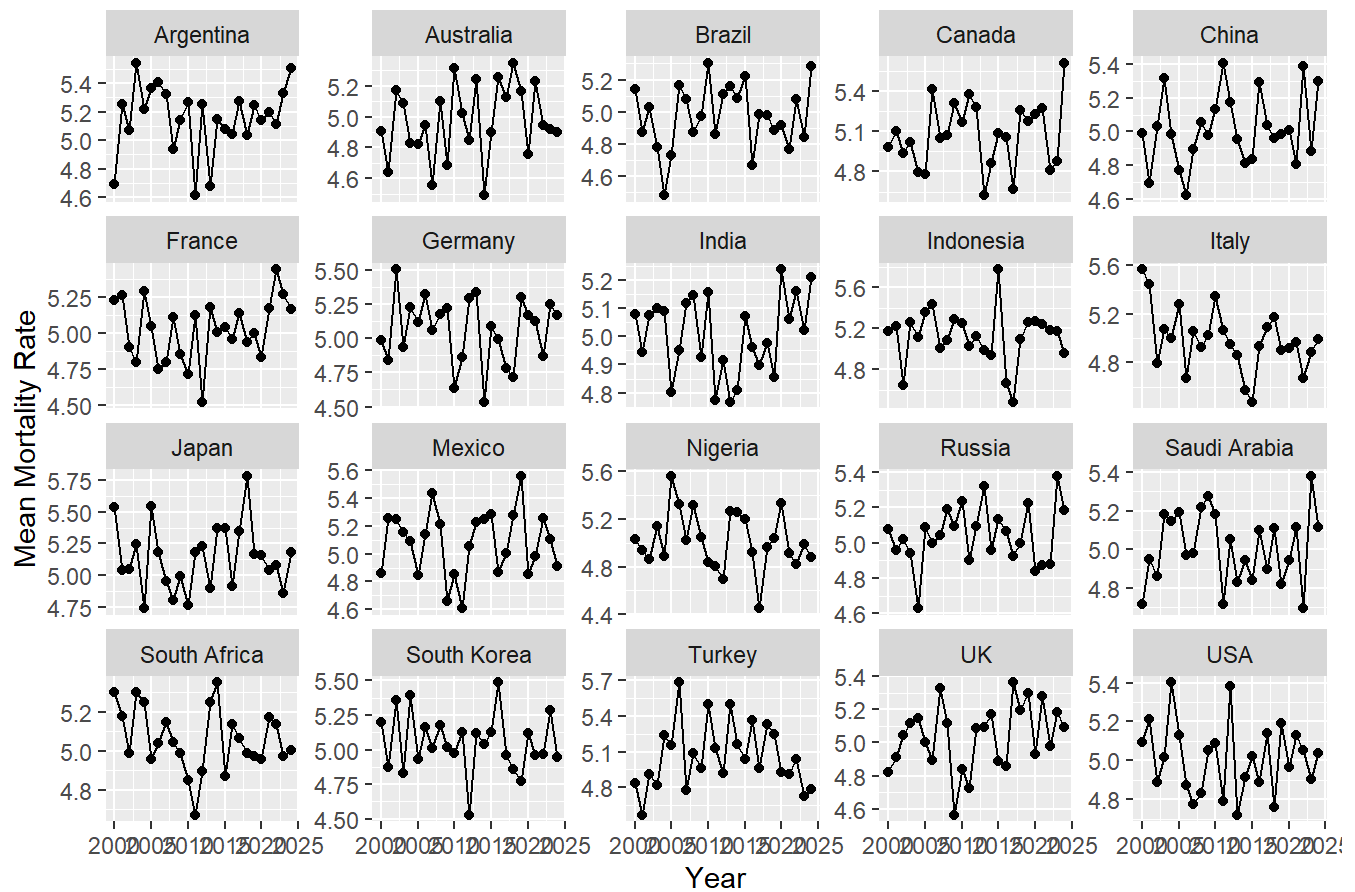
```
plot_international('Chronic')
```

Mean Mortality Rate by Chronic Category in 20 countries



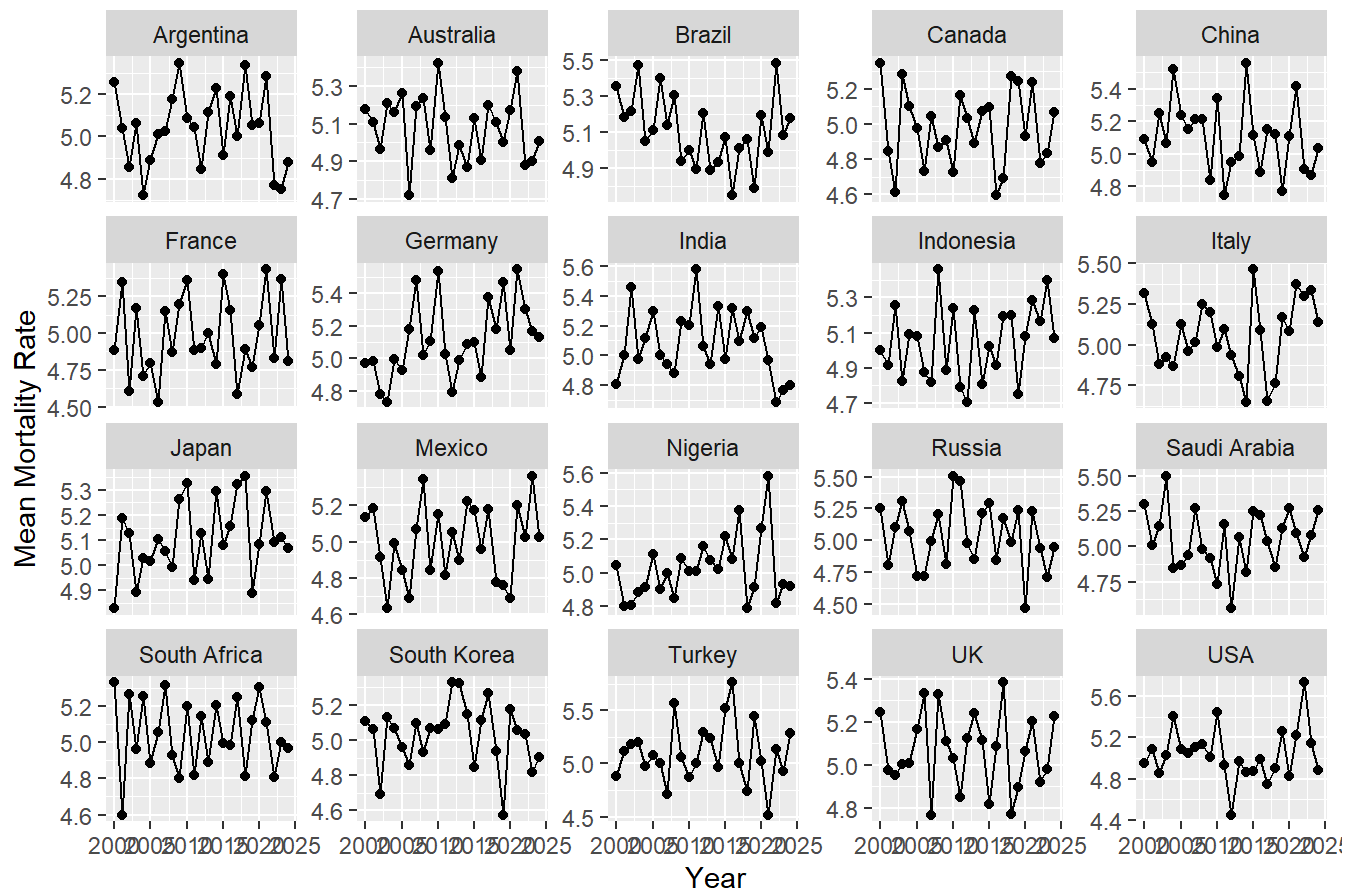
```
plot_international('Genetic')
```

Mean Mortality Rate by Genetic Category in 20 countries



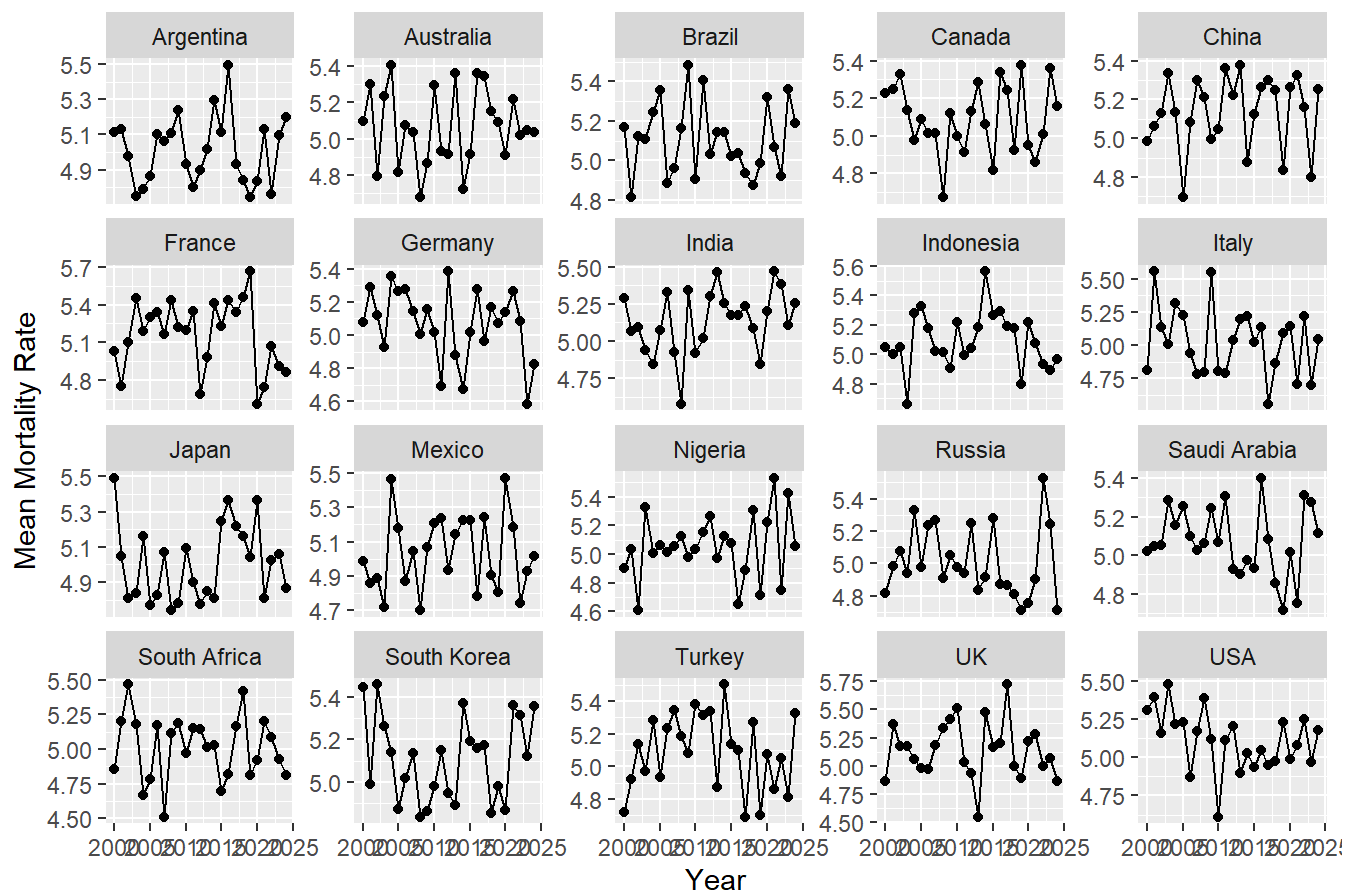
```
plot_international('Infectious')
```

Mean Mortality Rate by Infectious Category in 20 countries



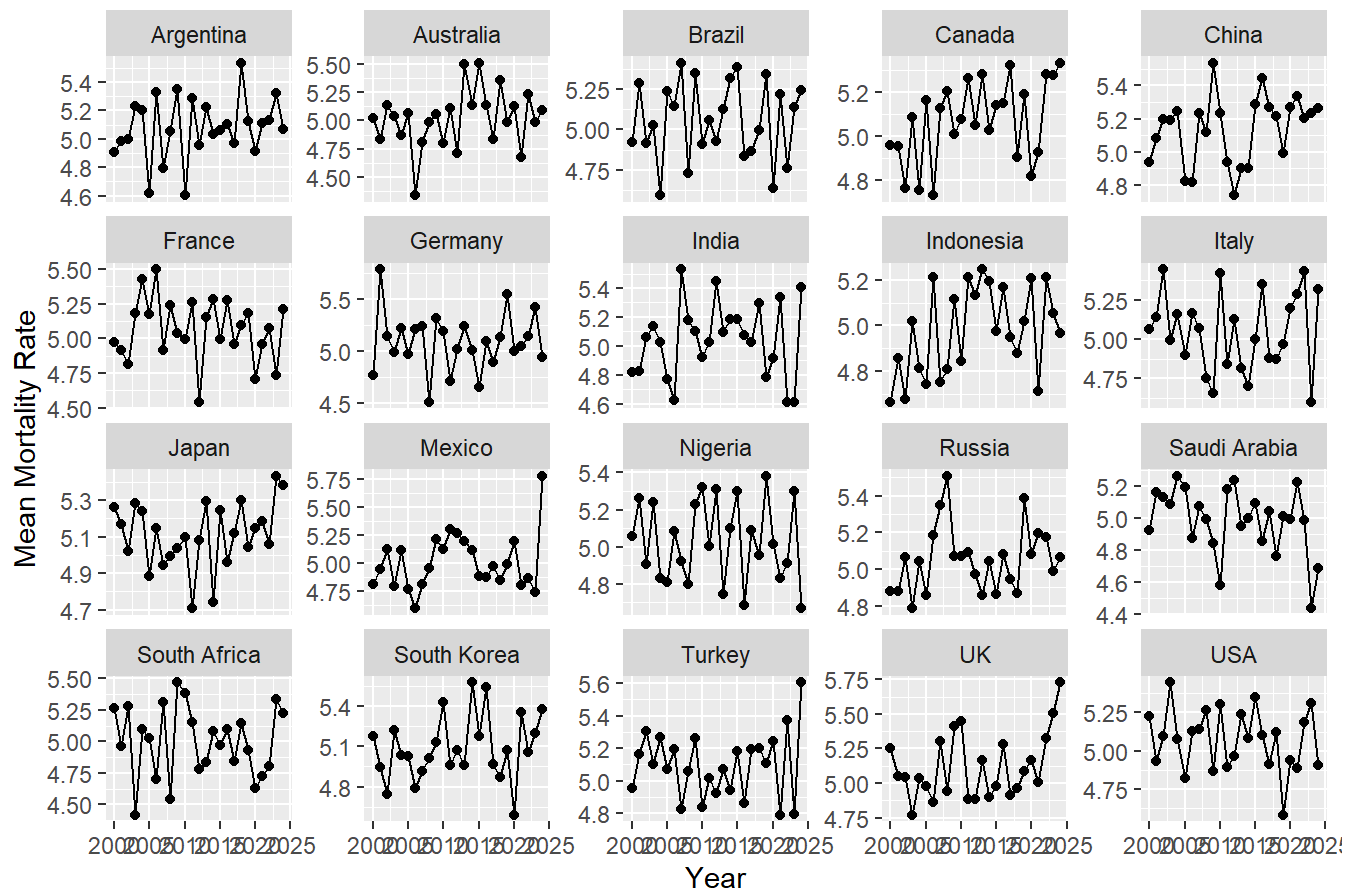
```
plot_international('Metabolic')
```

Mean Mortality Rate by Metabolic Category in 20 countries



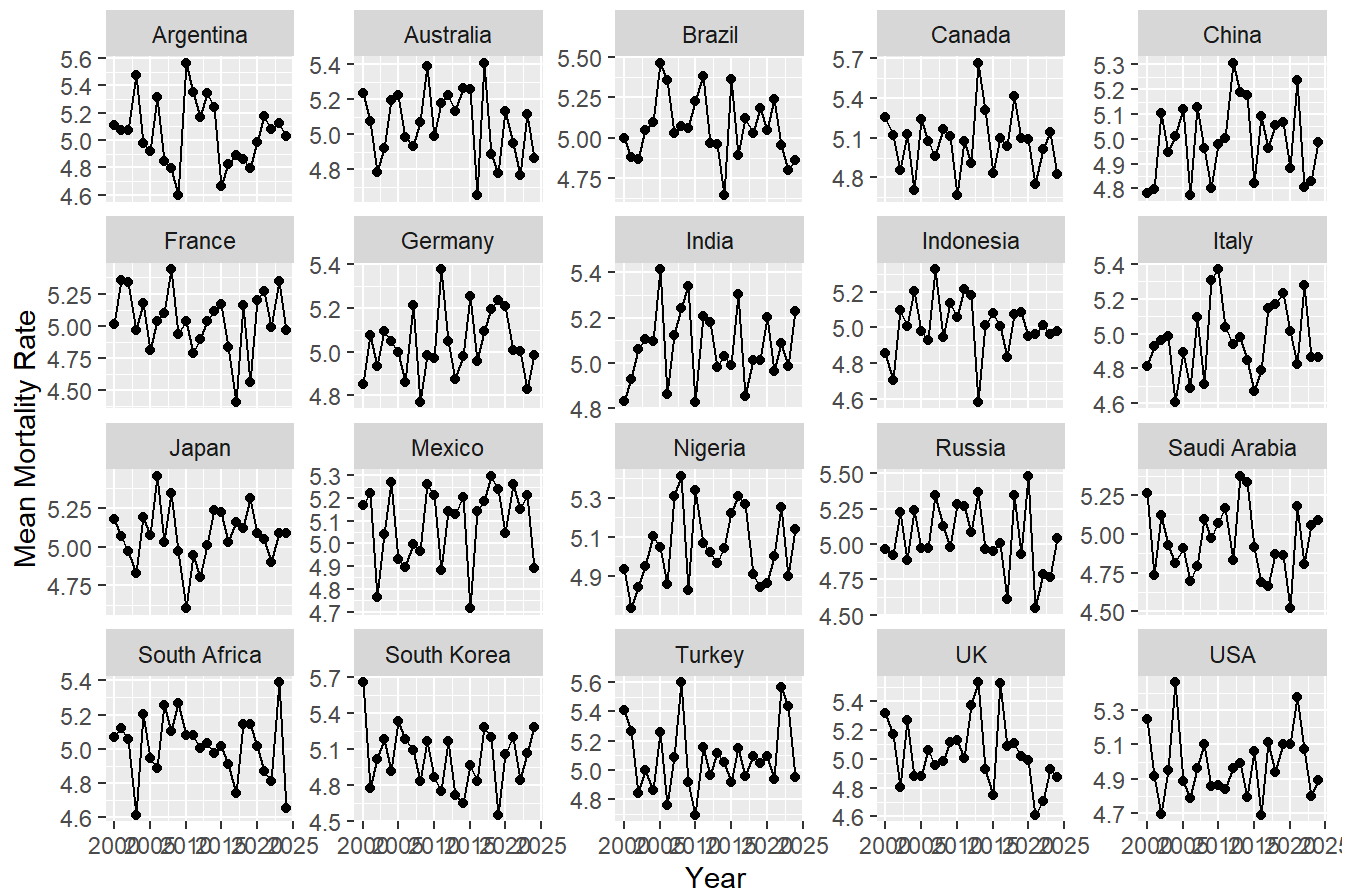
```
plot_international('Neurological')
```


Mean Mortality Rate by Neurological Category in 20 countries



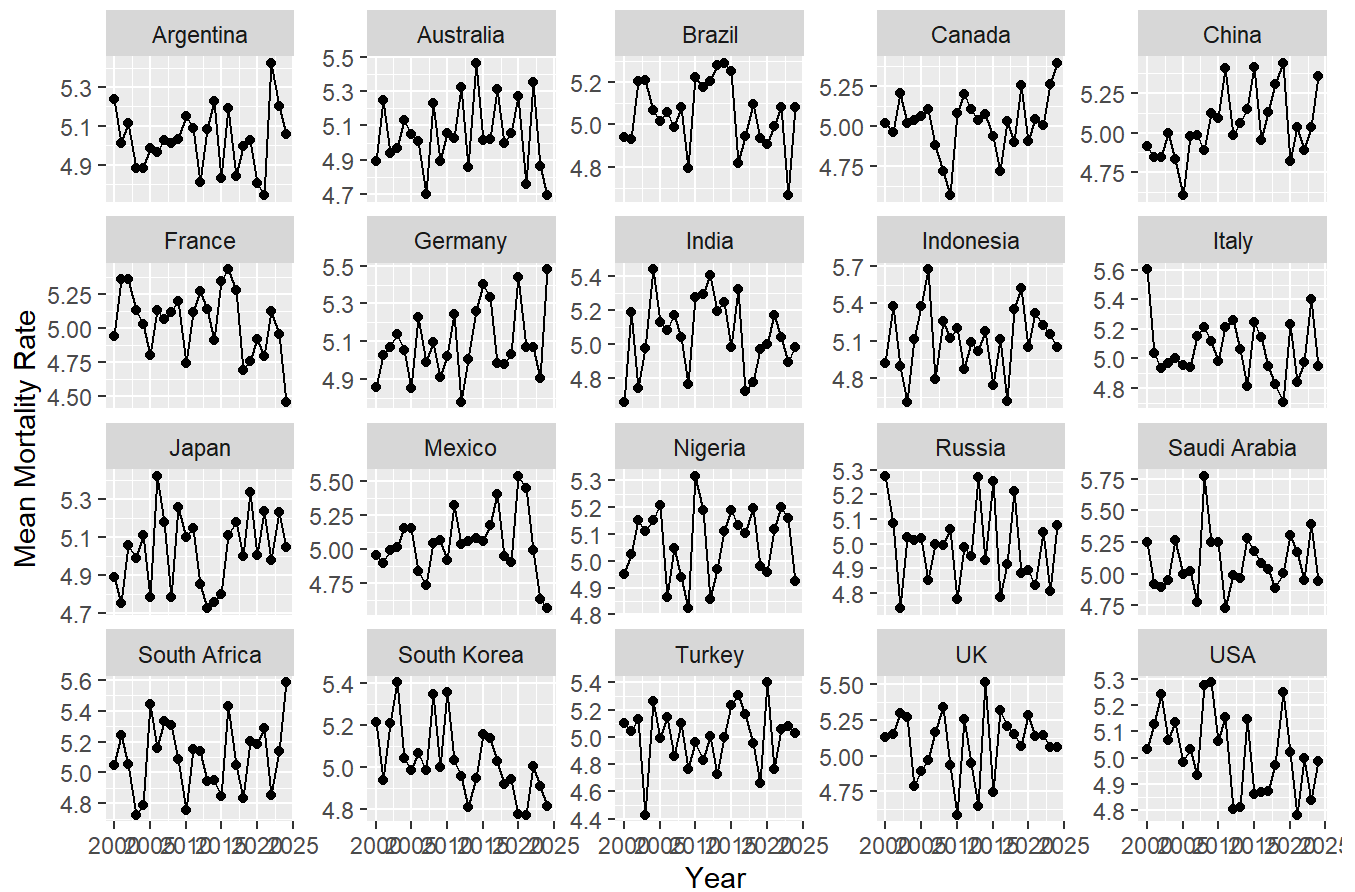
```
plot_international('Parasitic')
```

Mean Mortality Rate by Parasitic Category in 20 countries



```
plot_international('Respiratory')
```

Mean Mortality Rate by Respiratory Category in 20 countries



```
plot_international('Viral')
```

Mean Mortality Rate by Viral Category in 20 countries

