Exercise

Analysis of Global Burden of Disease (GBD) data for Saudi Arabia

Julian Flowers

2024-07-02

Contents

Introduction

In this exercise we will use data from the Global Burden of disease to explore trends in exposures and health outcomes in Saudi Arabia

In this exercise we will be using R and Tableau Public online to analyse and visualise the data. The source data is available here.

Get started

First we need to load the R packages for analysis.

```
needs(tidyverse)
```

Now we load the data.

Get the data

The code segment below shows how data is loaded.

```
gbd <- read_csv("https://github.com/julianflowers/spha/blob/main/IHME-GBD_2019_DATA-a5616352-1.csv?raw=
    filter(str_detect(rei_name, "Tobacco|Alcohol|Air|Env|Occupational|Dietary|Low"))
```

Exploring the data

There are numerous ways of exploring data in R. A first step is to review the variables in the dataset.

```
## Rows: 94,770
## Columns: 18
## $ measure_id    <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ measure_name  <chr> "Deaths", "Deaths", "Deaths", "Deaths", "Deaths", "Death~
```

```
## $ location_id   <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ location_name <chr> "Global", "Global", "Global", "Global", "Global", "Globa~
## $ sex_id        <dbl> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,~
## $ sex_name      <chr> "Both", "Both", "Both", "Both", "Both", "Both", "Both", ~
## $ age_id        <dbl> 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, ~
## $ age_name      <chr> "All ages", "All ages", "All ages", "All ages", "All age~
## $ cause_id      <dbl> 297, 297, 297, 297, 297, 297, 322, 322, 322, 322, 322, 3~
## $ cause_name    <chr> "Tuberculosis", "Tuberculosis", "Tuberculosis", "Tubercu~
## $ rei_id        <dbl> 98, 102, 98, 102, 98, 102, 85, 98, 102, 202, 85, 98, 102~
## $ rei_name      <chr> "Tobacco", "Alcohol use", "Tobacco", "Alcohol use", "Tob~
## $ metric_id     <dbl> 1, 1, 2, 2, 3, 3, 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 1,~
## $ metric_name   <chr> "Number", "Number", "Percent", "Percent", "Rate", "Rate"~
## $ year          <dbl> 1990, 1990, 1990, 1990, 1990, 1990, 1990, 1990, 1990, 19~
## $ val           <dbl> 3.094213e+05, 2.291413e+05, 1.741761e-01, 1.290618e-01, ~
## $ upper         <dbl> 3.622769e+05, 2.995508e+05, 2.039474e-01, 1.666499e-01, ~
## $ lower         <dbl> 2.565213e+05, 1.507260e+05, 1.448615e-01, 8.365103e-02, ~
```

The dataset consists a series of metrics for Saudi Arabia from the GBD Compare dataset for age ,sex, cause (level 3) and risk factor (level 3) by year. The dataset has 94770 records. The available metrics are Deaths, DALYs (Disability-Adjusted Life Years). For this analysis we will focus on Disability Adjusted Life Years (DALYs) rate, which is a summary measure of population health.

The causes included are

id	name
1	Tuberculosis
2	Lower respiratory infections
3	Otitis media
4	Upper respiratory infections
5	Meningitis
6	Cirrhosis and other chronic liver diseases
7	Parkinson's disease
8	Pancreatitis
9	Other pharynx cancer
10	Pancreatic cancer
11	Kidney cancer
12	Leukemia
13	Bladder cancer
14	Environmental heat and cold exposure
15	Ischemic heart disease
16	Alzheimer's disease and other dementias
17	Idiopathic epilepsy
18	Multiple sclerosis
19	Diarrheal diseases
20	Lip and oral cavity cancer
21	Nasopharynx cancer
22	Alcohol use disorders
23	Rheumatoid arthritis
24	Esophageal cancer
25	Mesothelioma
26	Neonatal disorders
27	Cervical cancer
28	Prostate cancer
29	Stomach cancer

id	name
30	Road injuries
31	Lower extremity peripheral arterial disease
32	Falls
33	Diabetes mellitus
34	Larynx cancer
35	Colon and rectum cancer
36	Stroke
37	Rheumatic heart disease
38	Endocarditis
39	Chronic obstructive pulmonary disease
40	Encephalitis
41	Fire, heat, and hot substances
42	Poisonings
43	Animal contact
44	Upper digestive system diseases
45	Ovarian cancer
46	Aortic aneurysm
47	Exposure to forces of nature
48	Breast cancer
49	Self-harm
50	Atrial fibrillation and flutter
51	Other transport injuries
52	Other cardiovascular and circulatory diseases
53	Foreign body
54	Drowning
55	Chronic kidney disease
56	Exposure to mechanical forces
57	Asthma
58	Gallbladder and biliary diseases
59	Interpersonal violence
60	Other unintentional injuries (internal)
61	Hypertensive heart disease
62	Liver cancer
63	Cardiomyopathy and myocarditis
64	Sudden infant death syndrome
65	Non-rheumatic valvular heart disease
66	Pneumoconiosis
67	Tracheal, bronchus, and lung cancer
68	Blindness and vision loss
69	Idiopathic developmental intellectual disability
70	Low back pain
71	Age-related and other hearing loss

.

The risk factors (exposures) included are

id	name
1	Tobacco
2	Alcohol use
3	Air pollution

id	name
4	Environmental/occupational risks
5	Occupational risks
6	Dietary risks
7	Low physical activity

.

The data is all age, persons values.

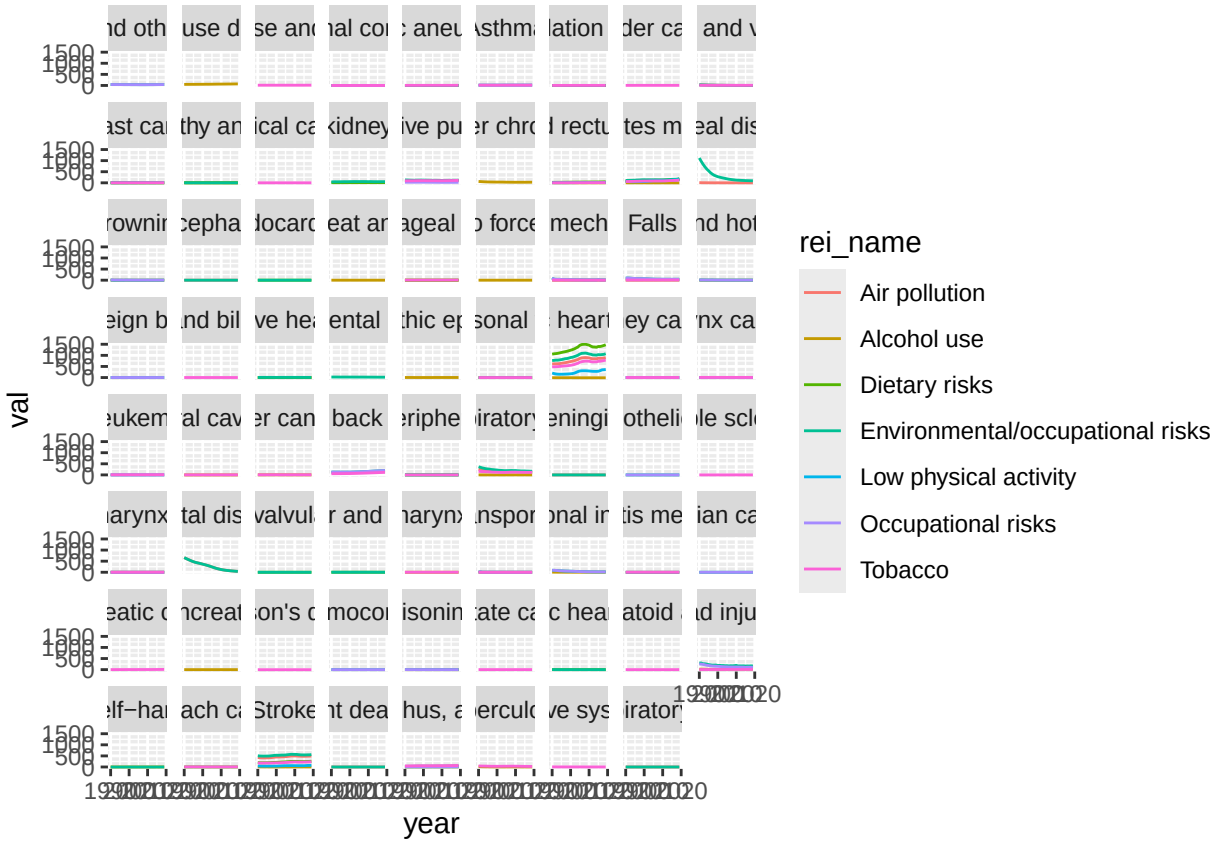
5 number summaries

We can create 5-number summaries of for each cause and exposure.

cause_name	rei_name	age_name	sex_name	mean	0%	25%	50%	75%	100%
Age-related and other hearing loss	Environmental/occupational risks	All ages	Both	38.1	35.8	36.4	38.0	39.5	41.6
Age-related and other hearing loss	Occupational risks	All ages	Both	38.1	35.8	36.4	38.0	39.5	41.6
Alcohol use disorders	Alcohol use	All ages	Both	57.0	48.0	51.2	55.2	62.5	69.5
Alzheimer's disease and other dementias	Tobacco	All ages	Both	10.9	10.6	10.6	10.7	10.8	12.3
Animal contact	Alcohol use	All ages	Both	0.0	0.0	0.0	0.0	0.0	0.0
Animal contact	Environmental/occupational risks	All ages	Both	0.2	0.1	0.2	0.2	0.3	0.4

Visualisation

One way of presenting mutlidimensional data is to present each dimension separatly. In visualisation, this is known as **faceting**. This is shown in the code below.

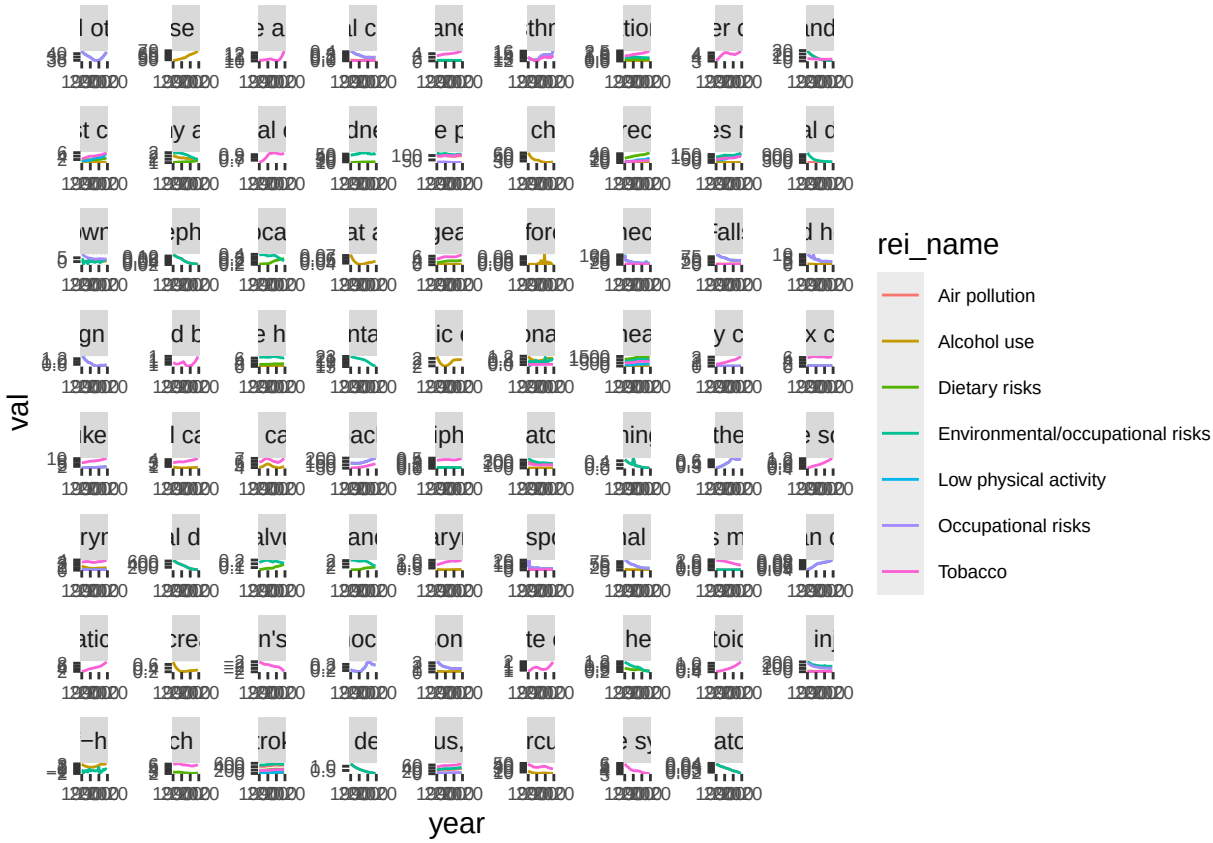


Qn: How can this chart be improved?

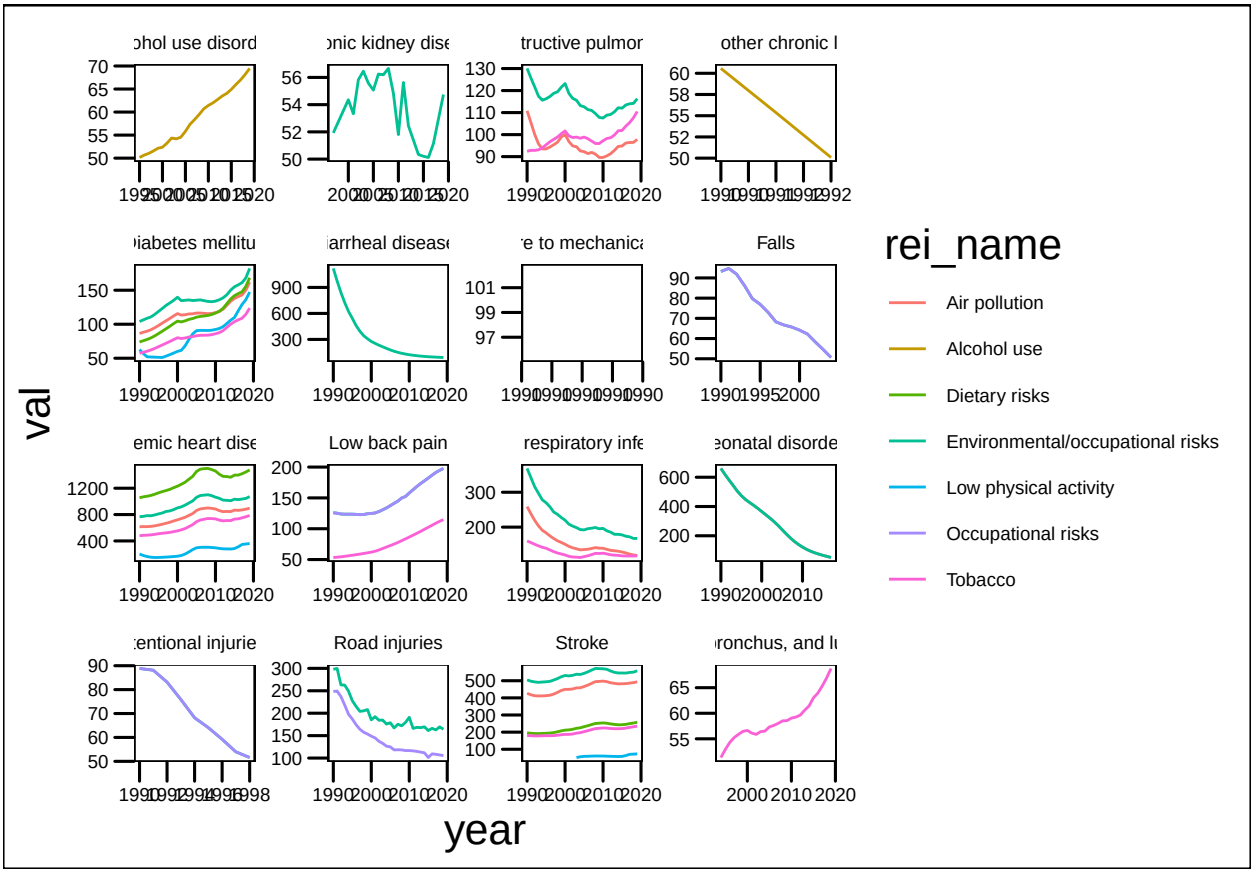
There are a number of problems with this initial visualsation.

- The different scales of the data masks trends in mamy of the cause.
- The cause names are too long

We can modify the scaling so each chart is scaled separately as below.



We can simplify further by filtering out causes where DALYs exceed a threshold - lets use DALYs > 50.

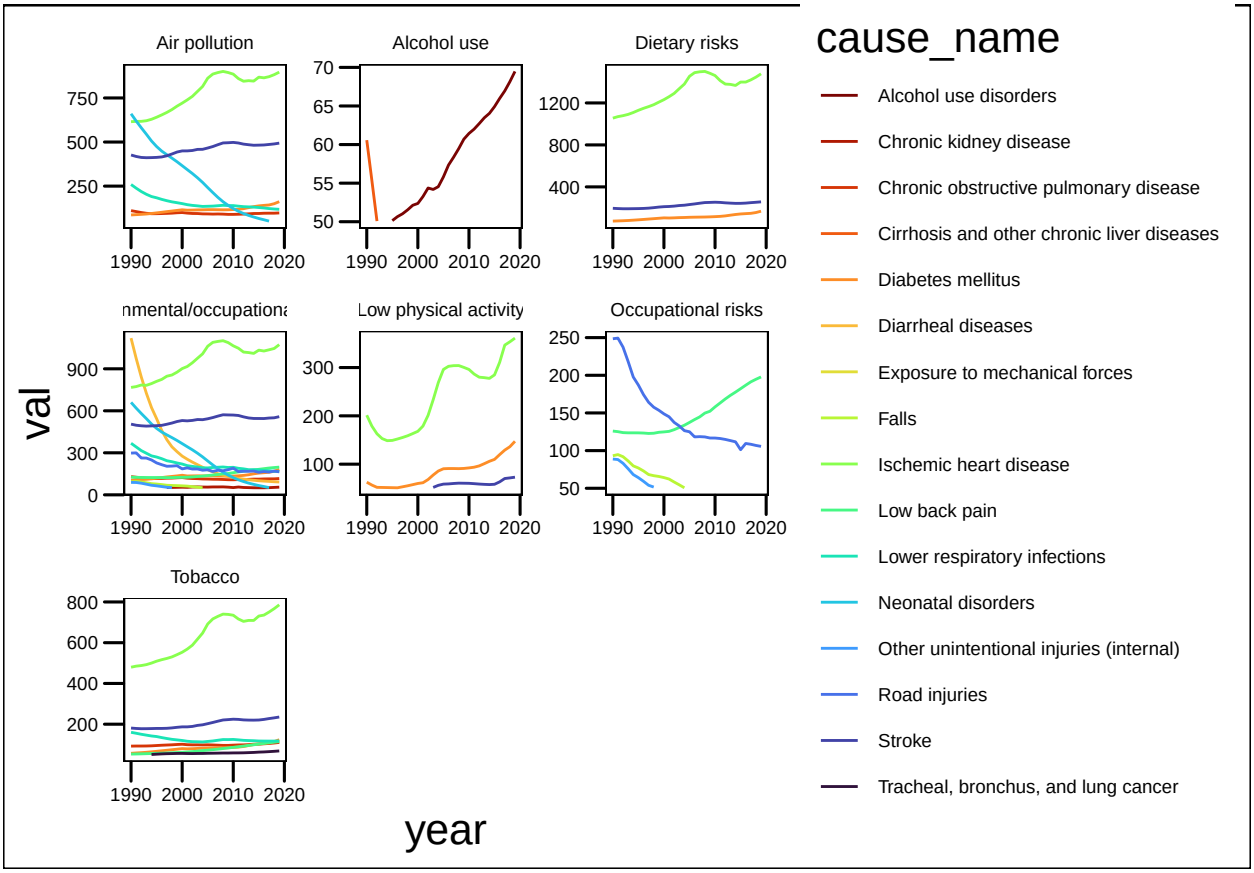


This provides greater clarity of trends in burden of disease by cause, stratified by underlying risk factors.

Qn:What trends do you see?

Risk factors

Rather than stratifying by risk, we can plot risk stratified by disease. This only requires a minor change in the code.



This is a more complex picture but highlights the pattern of attributable risk of exposures. For example, a growing contribution of air pollution to ischaemic heart disease.

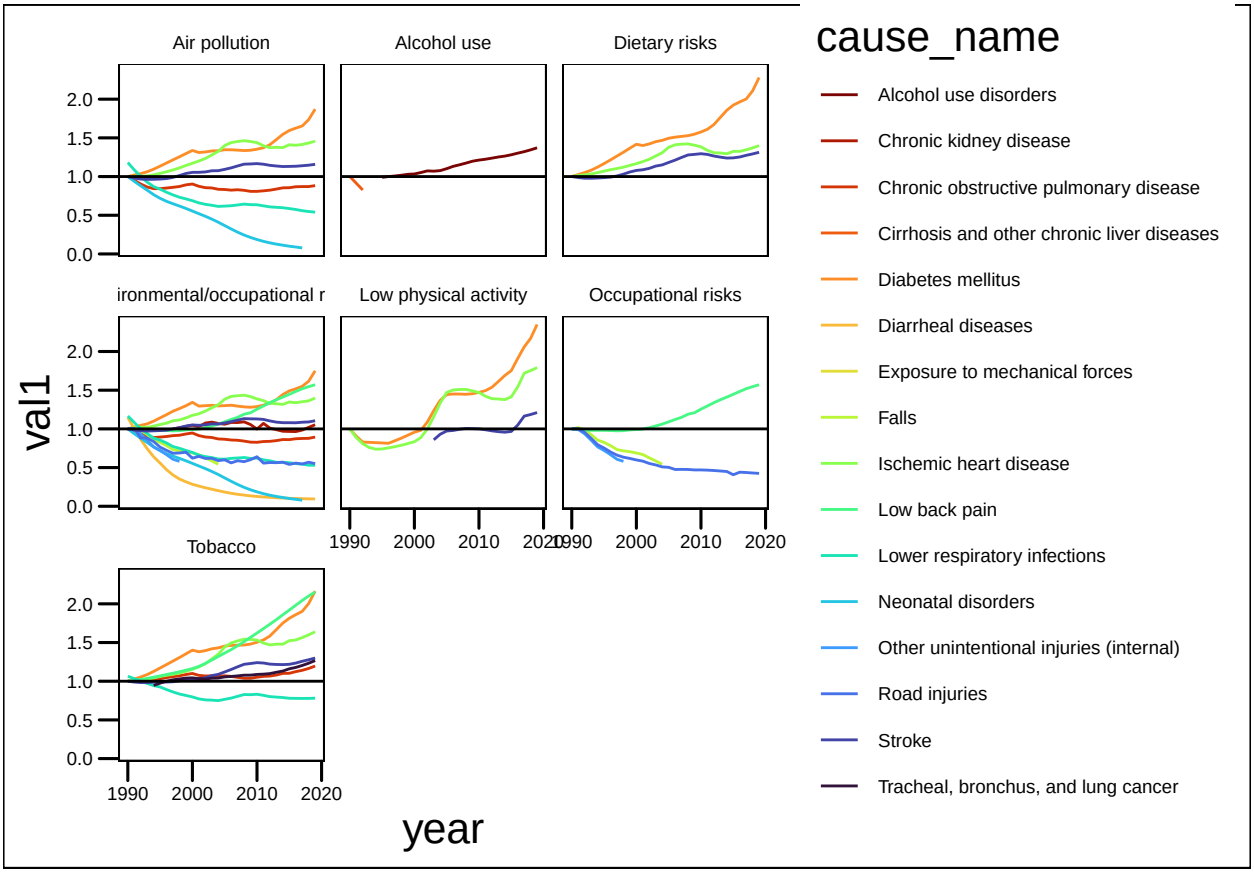
Qn: What other trends emerge?

Further improvements

One disadvantage of scaling each cause separately is that we cannot generate inter-cause comparison.

Qn: Why might this be useful? How might it be achieved?

Advanced


Air pollution
Alcohol use
Dietary risks
ironmental/occupational r
Low physical activity
Occupational risks
Tobacco
val1
year
0.0
0.5
1.0
1.5
2.0
1990
2000
2010
2020
cause_name
Alcohol use disorders
Chronic kidney disease
Chronic obstructive pulmonary disease
Cirrhosis and other chronic liver diseases
Diabetes mellitus
Diarrheal diseases
Exposure to mechanical forces
Falls
Ischemic heart disease
Low back pain
Lower respiratory infections
Neonatal disorders
Other unintentional injuries (internal)
Road injuries
Stroke
Tracheal, bronchus, and lung cancer