Fatality and Disease Analyzation and Statistics

Kamran Qureshi
Naveen Ithikkat
Qinya Wang
Rithvik Mundra
Tianweibao Zheng

About the clients

CDC

- The Centers for Disease Control and Prevention is the nation’s health protection agency
- They conduct science and provides health information that protects our nation against health threats
- **The CDC saves lives**

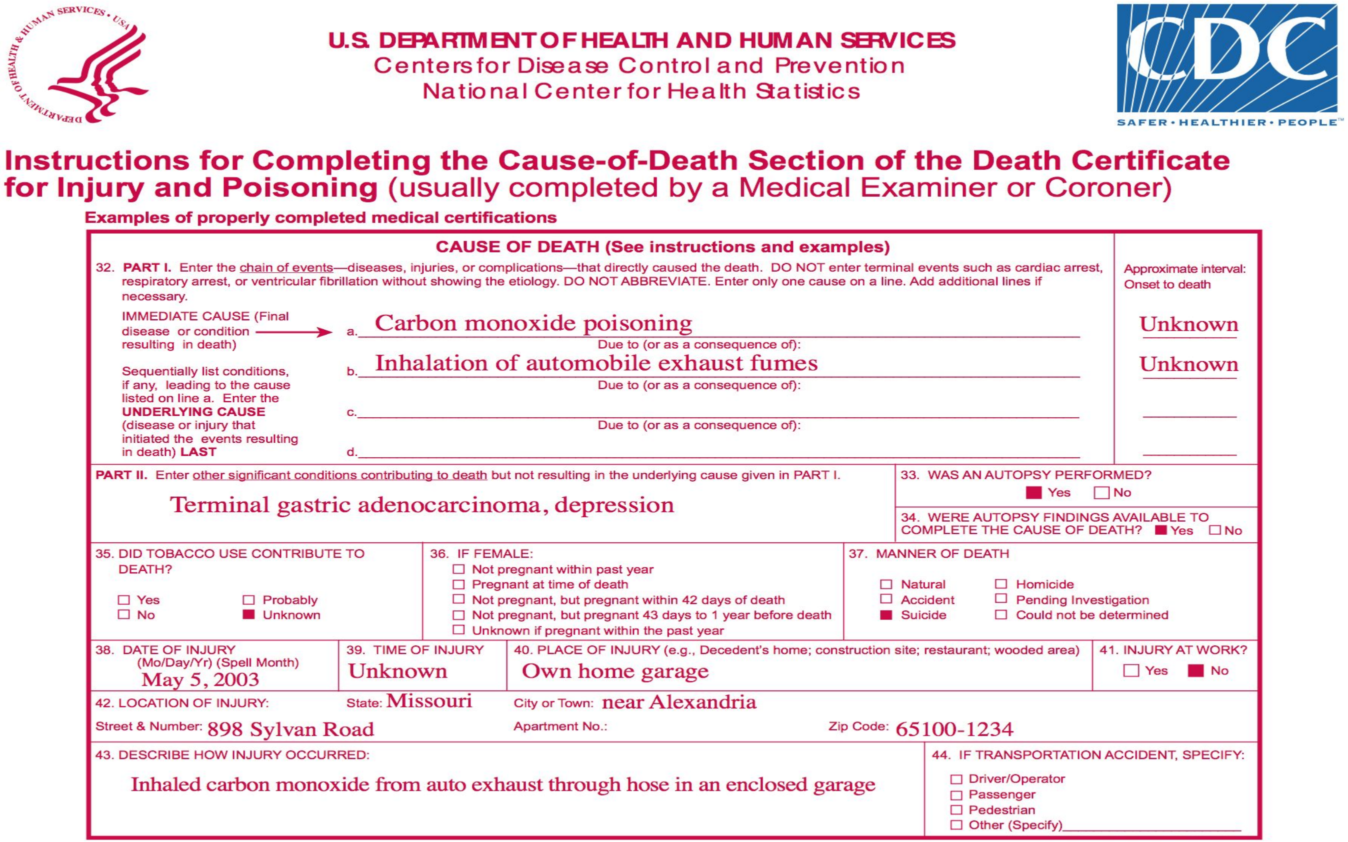
Elder Research

- A data science and predictive analytics company
- Develops analytic solutions and provides analytics consulting

Client objective

- Use US mortality data to find trends in fatal diseases and injuries that co-occur (meaning two or more conditions that are commonly listed together on death certificates as factors that contributed to or caused the death).

Understanding the data



U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Centers for Disease Control and Prevention
National Center for Health Statistics

CDC SAFER • HEALTHIER • PEOPLE™

Instructions for Completing the Cause-of-Death Section of the Death Certificate for Injury and Poisoning (usually completed by a Medical Examiner or Coroner)

Examples of properly completed medical certifications

CAUSE OF DEATH (See instructions and examples)

32. **PART I.** Enter the chain of events—diseases, injuries, or complications—that directly caused the death. DO NOT enter terminal events such as cardiac arrest, respiratory arrest, or ventricular fibrillation without showing the etiology. DO NOT ABBREVIATE. Enter only one cause on a line. Add additional lines if necessary.

Approximate interval: Onset to death

IMMEDIATE CAUSE (Final disease or condition resulting in death) → a. Carbon monoxide poisoning — Unknown

Due to (or as a consequence of):

Sequentially list conditions, if any, leading to the cause listed on line a. Enter the **UNDERLYING CAUSE** (disease or injury that initiated the events resulting in death) **LAST**

b. Inhalation of automobile exhaust fumes — Unknown

Due to (or as a consequence of):

c. ______________________

Due to (or as a consequence of):

d. ______________________

PART II. Enter other significant conditions contributing to death but not resulting in the underlying cause given in PART I.

Terminal gastric adenocarcinoma, depression

33. WAS AN AUTOPSY PERFORMED?
■ Yes □ No

34. WERE AUTOPSY FINDINGS AVAILABLE TO COMPLETE THE CAUSE OF DEATH? ■ Yes □ No

35. DID TOBACCO USE CONTRIBUTE TO DEATH?
□ Yes □ Probably
□ No ■ Unknown

36. IF FEMALE:
□ Not pregnant within past year
□ Pregnant at time of death
□ Not pregnant, but pregnant within 42 days of death
□ Not pregnant, but pregnant 43 days to 1 year before death
□ Unknown if pregnant within the past year

37. MANNER OF DEATH
□ Natural □ Homicide
□ Accident □ Pending Investigation
■ Suicide □ Could not be determined

38. DATE OF INJURY (Mo/Day/Yr) (Spell Month) May 5, 2003

39. TIME OF INJURY Unknown

40. PLACE OF INJURY (e.g., Decedent's home; construction site; restaurant; wooded area) Own home garage

41. INJURY AT WORK? □ Yes ■ No

42. LOCATION OF INJURY: State: Missouri City or Town: near Alexandria

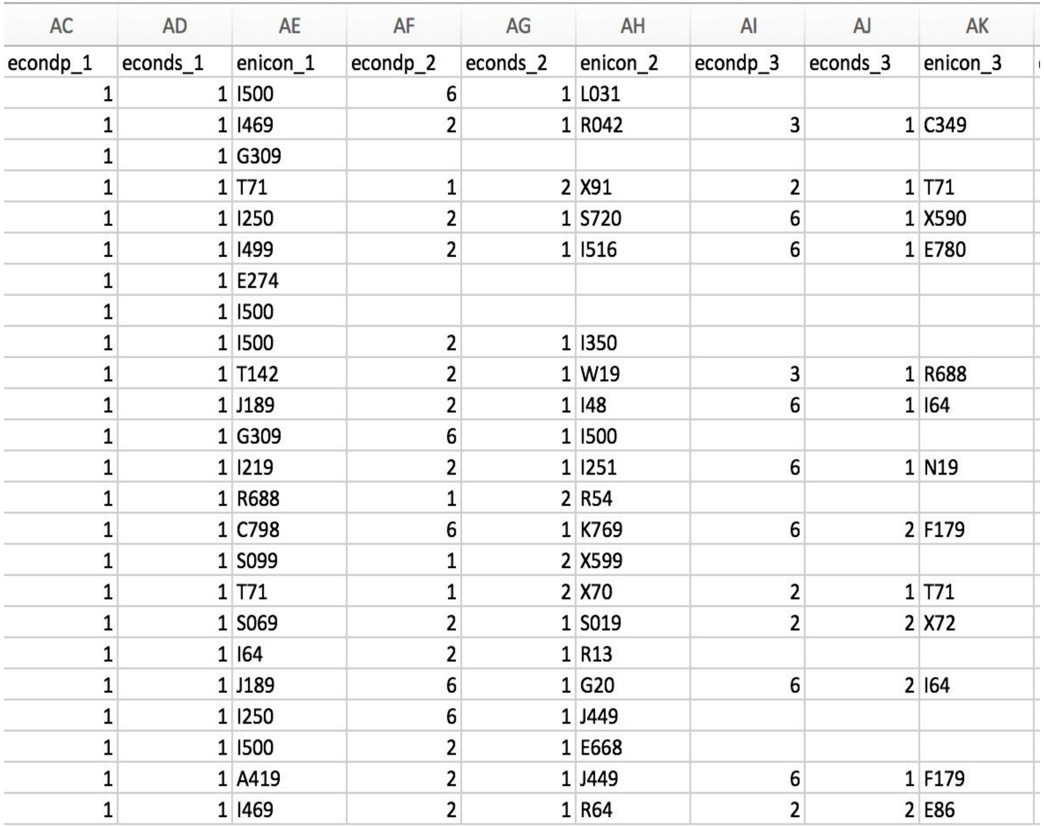
Street & Number: 898 Sylvan Road Apartment No.: Zip Code: 65100-1234

43. DESCRIBE HOW INJURY OCCURRED:

Inhaled carbon monoxide from auto exhaust through hose in an enclosed garage

44. IF TRANSPORTATION ACCIDENT, SPECIFY:
□ Driver/Operator
□ Passenger
□ Pedestrian
□ Other (Specify)______________

Understanding the data (cont.)



AC	AD	AE	AF	AG	AH	AI	AJ	AK
econdp_1	econds_1	enicon_1	econdp_2	econds_2	enicon_2	econdp_3	econds_3	enicon_3
1	1	I500	6	1	L031			
1	1	I469	2	1	R042	3	1	C349
1	1	G309						
1	1	T71	1	2	X91	2	1	T71
1	1	I250	2	1	S720	6	1	X590
1	1	I499	2	1	I516	6	1	E780
1	1	E274						
1	1	I500						
1	1	I500	2	1	I350			
1	1	T142	2	1	W19	3	1	R688
1	1	J189	2	1	I48	6	1	I64
1	1	G309	6	1	I500			
1	1	I219	2	1	I251	6	1	N19
1	1	R688	1	2	R54			
1	1	C798	6	1	K769	6	2	F179
1	1	S099	1	2	X599			
1	1	T71	1	2	X70	2	1	T71
1	1	S069	2	1	S019	2	2	X72
1	1	I64	2	1	R13			
1	1	J189	6	1	G20	6	2	I64
1	1	I250	6	1	J449			
1	1	I500	2	1	E668			
1	1	A419	2	1	J449	6	1	F179
1	1	I469	2	1	R64	2	2	E86

ICD 10

10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD), a medical classification list by the World Health Organization(WHO).

It contains codes for diseases, signs and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or diseases.

The code set allows more than **14,400 different codes** and permits the tracking of many new diagnoses

Understanding the data (cont.)

- Used mortality data of 10 years (2006-2015)
- Approx 25 million records
- Approx 10 GB data
- A record can contain up to 20 causes of death (specified as 4-char ‘ICD’ codes)
- To find patterns of co-occurring codes, ‘association rules’ to be used

Technical setup

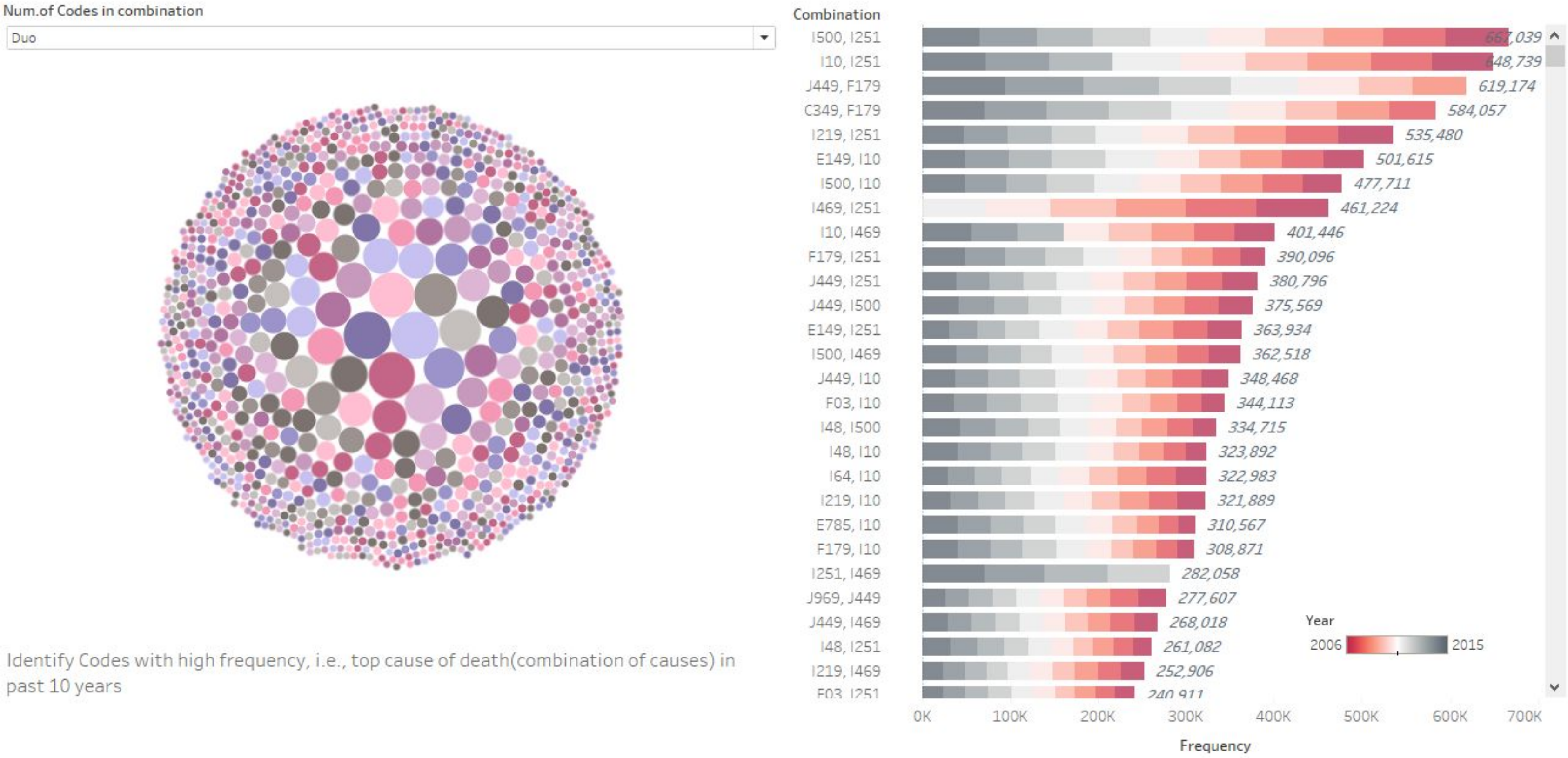
- R for data cleaning
- Python with Spark on AWS - for scalable analysis (as data increases YoY)
- 1 master, 10 worker nodes setup on AWS (each 16CPU, 30GB memory)
- FPGrowth (Frequent Pattern Growth) algorithm
  - ‘Support’ parameter (used 0.0001)
  - ‘Partitions’ parameter (used 480)
- Tableau for visualizations

Key Findings

Top co-occurring code pairs

Rank	Combination		Code1 Desc	Code2 Desc	Frequency
1	I500	I251	Congestive heart disease	Atherosclerotic heart disease, Coronary	667,039
2	I10	I251	Essential (primary) hypertension	Atherosclerotic heart disease, Coronary	648,739
3	J449	F179	Chronic obstructive pulmonary disease	Mental/behavioural disorders due to use of tobacco	619,174
4	C349	F179	Malignant neoplasm: Bronchus or lung	Mental/behavioural disorders due to use of tobacco	584,057
5	I219	I251	Acute myocardial infarction	Atherosclerotic heart disease, Coronary	535,480
6	E149	I10	Diabetes mellitus without complications	Essential (primary) hypertension	501,615
7	I500	I10	Congestive heart disease	Essential (primary) hypertension	477,711
8	I469	I251	Cardiac arrest	Atherosclerotic heart disease, Coronary	461,224
9	I10	I469	Essential (primary) hypertension	Cardiac arrest	401,446
10	F179	I251	Mental/behavioural disorders due to use of tobacco	Atherosclerotic heart disease, Coronary	390,096

Top co-occurring code pairs

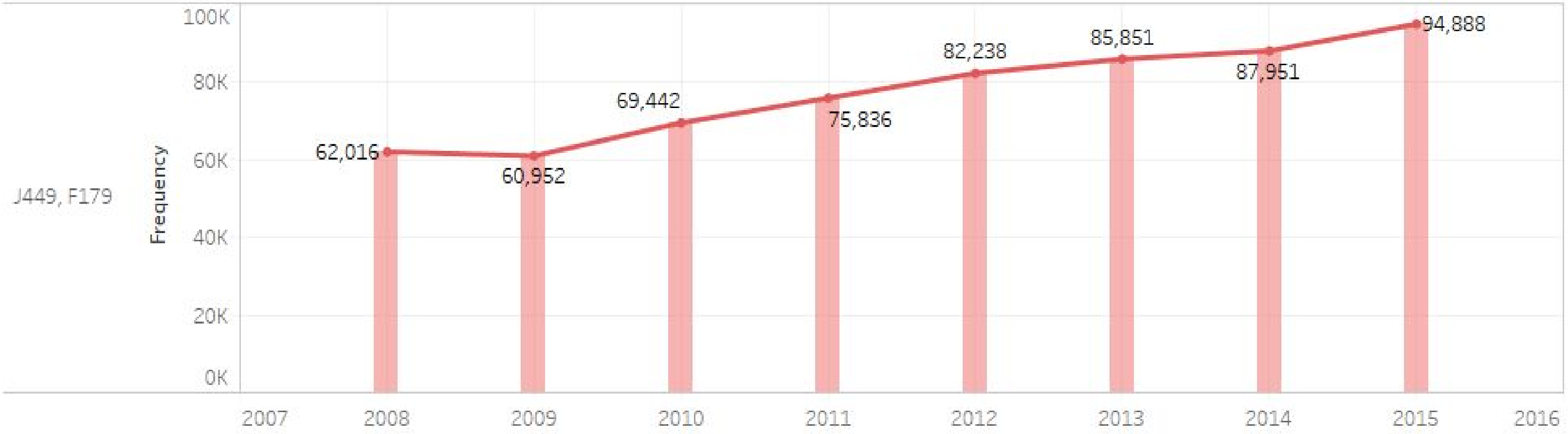


Num.of Codes in combination

Duo ▾

Identify Codes with high frequency, i.e., top cause of death(combination of causes) in past 10 years

Trends in (selected) co-occurring code pairs

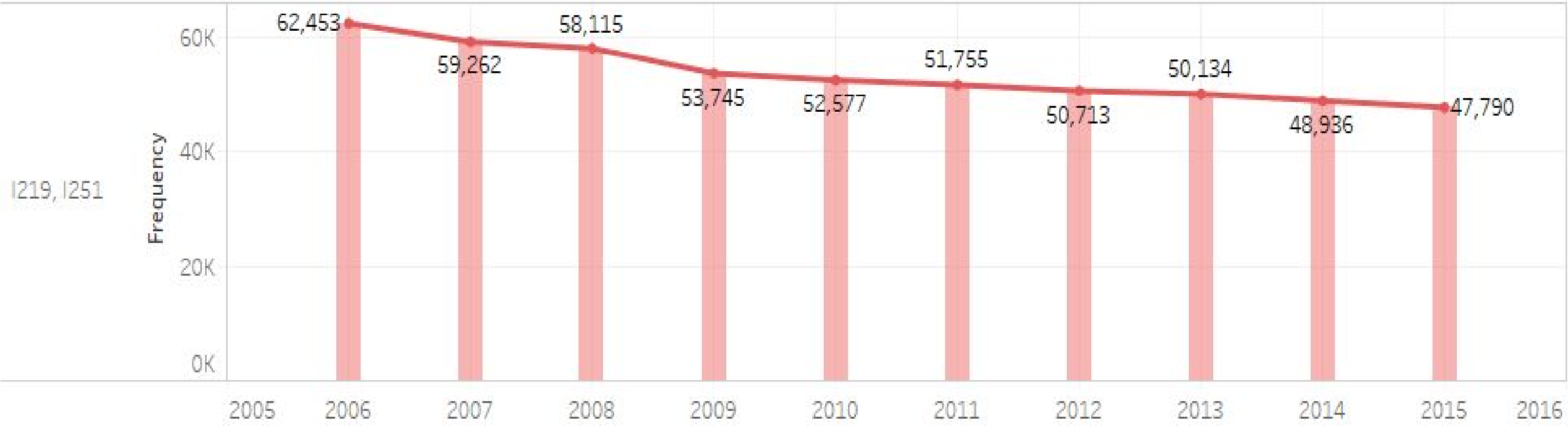


Highest increase (~62k to ~95k):

J449 - Chronic obstructive pulmonary disease, unspecified

F179 - Mental and behavioural disorders due to use of tobacco: Unspecified mental and behavioural disorder

Trends in (selected) co-occurring code pairs

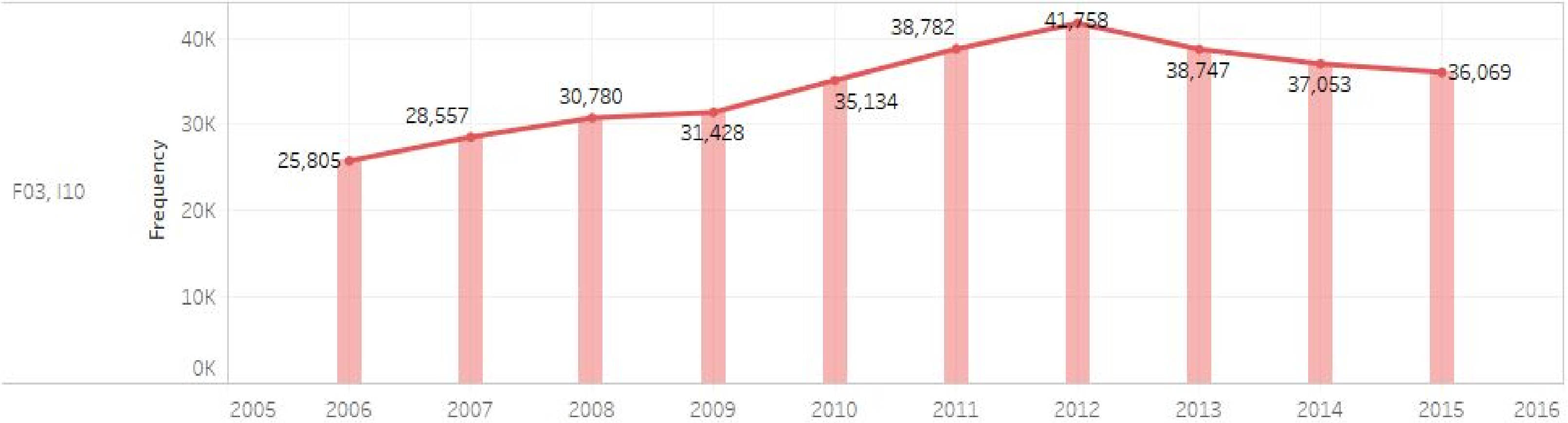


Highest decrease (~62k to ~48k):

I219 - Acute Myocardial Infarction, Unspecified

I251 - Atherosclerotic Heart Disease

Trends in (selected) co-occurring code pairs



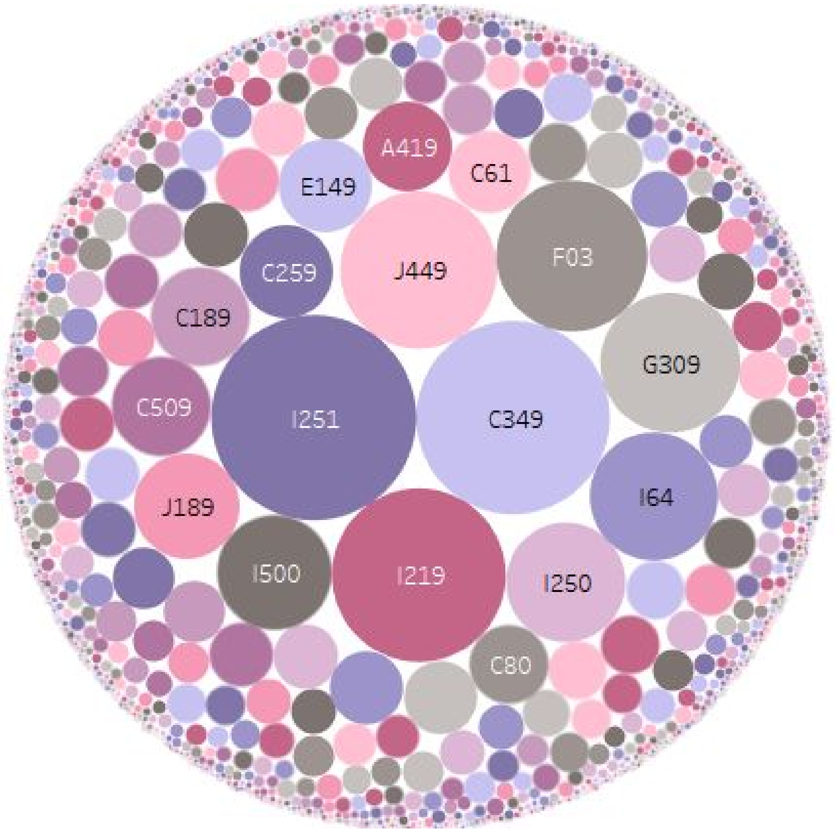
Most interesting (increasing till 2012 and then decreasing):
F03 - Unspecified dementia
I10 - Essential (primary) hypertension

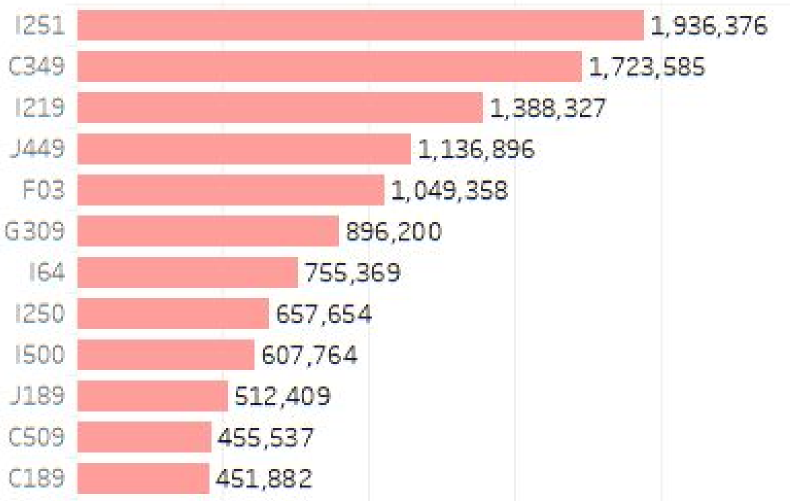
Other Findings

Top ‘underlying causes of death’

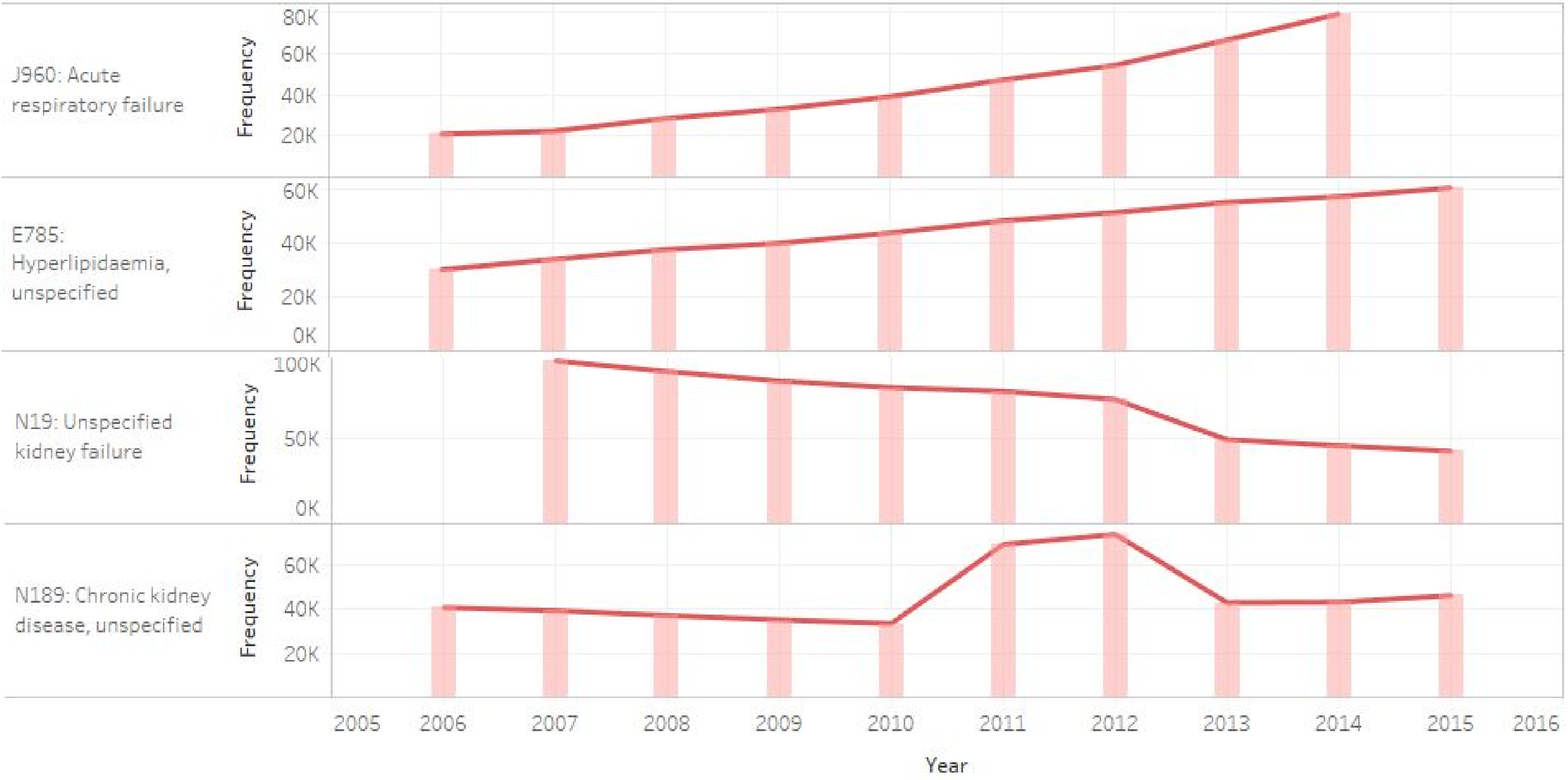
Rank	UCOD	Description	Count
1	I251	Atherosclerotic heart disease of native coronary artery	1,936,376
2	C349	Malignant neoplasm: Bronchus or lung, unspecified	1,723,585
3	I219	Acute myocardial infarction, unspecified	1,388,327
4	J449	Chronic obstructive pulmonary disease, unspecified	1,136,896
5	F03	Unspecified dementia	1,049,358
6	G309	Alzheimer's disease, unspecified	896,200
7	I64	Stroke, not specified as haemorrhage or infarction	755,369
8	I250	Atherosclerotic cardiovascular disease	657,654
9	I500	Congestive heart failure	607,764
10	J189	Pneumonia	512,409

Top ‘underlying causes of death’

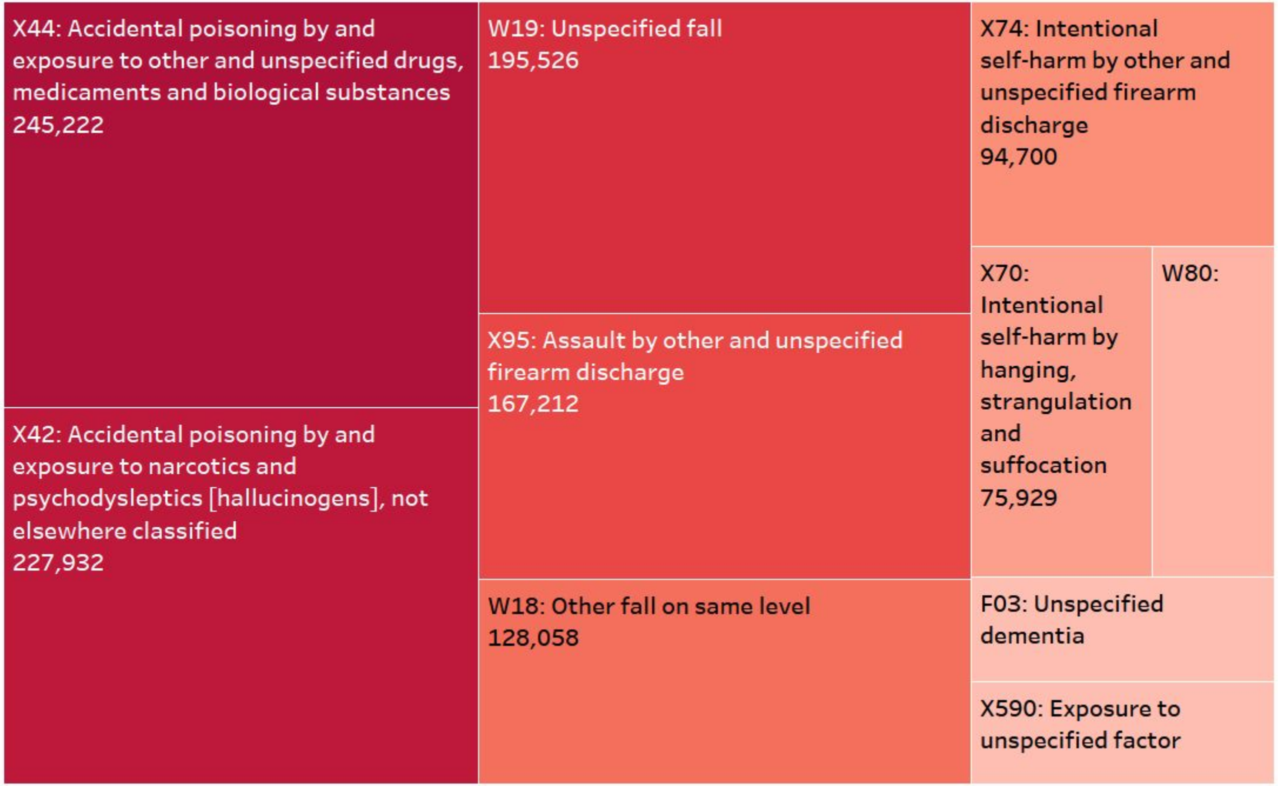




Trends in (selected) codes



Top occurring codes where injury was involved



Thank You