Data Extraction

In this notebook, we will extract abstract and summary parts from articles, then write to csv files

In [1]:
```
import pandas as pd
```

In [2]:
```
path = "./data/biolaysumm2024_data/"
filename = "eLife_train.jsonl"
df = pd.read_json(path + filename,
                  orient="records",
                  lines=True)
df.head()
```

Out[2]:

	lay_summary	article	headings	keywords	id
0	In the USA , more deaths happen in the winter ...	In temperate climates , winter deaths exceed s...	[Abstract, Introduction, Results, Discussion, ...	[epidemiology and global health]	elife-35500-v1
1	Most people have likely experienced the discom...	Whether complement dysregulation directly cont...	[Abstract, Introduction, Results, Discussion, ...	[microbiology and infectious disease, immunolo...	elife-48378-v2
2	The immune system protects an individual from ...	Variation in the presentation of hereditary im...	[Abstract, Introduction, Results, Discussion, ...	[microbiology and infectious disease, immunolo...	elife-04494-v1
3	The brain adapts to control our behavior in di...	Rapid and flexible interpretation of conflicti...	[Abstract, Introduction, Results, Discussion, ...	[neuroscience]	elife-12352-v2
4	Cells use motor proteins that to move organell...	Myosin 5a is a dual-headed molecular motor tha...	[Abstract, Introduction, Results, Discussion, ...	[structural biology and molecular biophysics]	elife-05413-v2

In [3]:
```
item = df.iloc[0]
item
```

Out[3]:
```
lay_summary    In the USA , more deaths happen in the winter ...
article        In temperate climates , winter deaths exceed s...
headings       [Abstract, Introduction, Results, Discussion, ...
keywords                        [epidemiology and global health]
id                                                elife-35500-v1
Name: 0, dtype: object
```

In [4]:
```
# count words
len(item["article"].split())
```

Out[4]: 3039

In [5]:
```
# split by paragraph
paras = item["article"].split("\n")
print(len(paras))
```

```
5
```

In [6]:
```
# the Abstract section is the first one
print(len(paras[0].split()))
paras[0]
```

```
171
```

Out[6]:
```
'In temperate climates , winter deaths exceed summer ones . However , there is lim
ited information on the timing and the relative magnitudes of maximum and minimum
mortality , by local climate , age group , sex and medical cause of death . We use
d geo-coded mortality data and wavelets to analyse the seasonality of mortality by
age group and sex from 1980 to 2016 in the USA and its subnational climatic region
s . Death rates in men and women ≥ 45 years peaked in December to February and wer
e lowest in June to August , driven by cardiorespiratory diseases and injuries . I
n these ages , percent difference in death rates between peak and minimum months d
id not vary across climate regions , nor changed from 1980 to 2016 . Under five ye
ars , seasonality of all-cause mortality largely disappeared after the 1990s . In
adolescents and young adults , especially in males , death rates peaked in June/Ju
ly and were lowest in December/January , driven by injury deaths . '
```

In [7]:
```
for pr in paras:
    print(pr)
    print("-------------")
```

In temperate climates , winter deaths exceed summer ones . However , there is limite
d information on the timing and the relative magnitudes of maximum and minimum morta
lity , by local climate , age group , sex and medical cause of death . We used geo-c
oded mortality data and wavelets to analyse the seasonality of mortality by age grou
p and sex from 1980 to 2016 in the USA and its subnational climatic regions . Death
rates in men and women ≥ 45 years peaked in December to February and were lowest in
June to August , driven by cardiorespiratory diseases and injuries . In these ages ,
percent difference in death rates between peak and minimum months did not vary acros
s climate regions , nor changed from 1980 to 2016 . Under five years , seasonality o
f all-cause mortality largely disappeared after the 1990s . In adolescents and young
adults , especially in males , death rates peaked in June/July and were lowest in De
cember/January , driven by injury deaths .

 It is well-established that death rates vary throughout the year , and in temperate
climates there tend to be more deaths in winter than in summer (Campbell , 2017; Fo
wler et al . , 2015; Healy , 2003; McKee , 1989) . It has therefore been hypothesiz
ed that a warmer world may lower winter mortality in temperate climates (Langford a
nd Bentham , 1995; Martens , 1998) . In a large country like the USA , which posses
ses distinct climate regions , the seasonality of mortality may vary geographically
, due to geographical variations in mortality , localized weather patterns , and reg
ional differences in adaptation measures such as heating , air conditioning and heal
thcare (Davis et al . , 2004; Braga et al . , 2001; Kalkstein , 2013; Medina-Ramón
and Schwartz , 2007) . The presence and extent of seasonal variation in mortality m
ay also itself change over time (Bobb et al . , 2014; Carson et al . , 2006; Sereta
kis et al . , 1997; Sheridan et al . , 2009) . A thorough understanding of the long
-term dynamics of seasonality of mortality , and its geographical and demographic pa
tterns , is needed to identify at-risk groups , plan responses at the present time a
s well as under changing climate conditions . Although mortality seasonality is well
-established , there is limited information on how seasonality , including the timin
g of minimum and maximum mortality , varies by local climate and how these features
have changed over time , especially in relation to age group , sex and medical cause
of death (Rau , 2004; Rau et al . , 2018) . In this paper , we comprehensively cha
racterize the spatial and temporal patterns of all-cause and cause-specific mortalit
y seasonality in the USA by sex and age group , through the application of wavelet a
nalytical techniques , to over three decades of national mortality data . Wavelets h
ave been used to study the dynamics of weather phenomena (Moy et al . , 2002) and
infectious diseases (Grenfell et al . , 2001) . We also used centre of gravity ana
lysis and circular statistics methods to understand the timing of maximum and minimu
m mortality . In addition , we identify how the percentage difference between death
rates in maximum and minimum mortality months has changed over time .

 The strengths of our study are its innovative methods of characterizing seasonality
of mortality dynamically over space and time , by age group and cause of death; usin
g wavelet and centre of gravity analyses; using ERA-Interim data output to compare t
he association between seasonality of death rates and regional temperature . A limit
ation of our study is that we did not investigate seasonality of mortality by socioe
conomic characteristics which may help with understanding its determinants and plann
ing responses .

 We used wavelet and centre of gravity analyses , which allowed systematically ident
ifying and characterizing seasonality of total and cause-specific mortality in the U
SA , and examining how seasonality has changed over time . We identified distinct se
asonal patterns in relation to age and sex , including higher all-cause summer morta
lity in young men (Feinstein , 2002; Rau et al . , 2018) . Importantly , we also s
howed that all-cause and cause-specific mortality seasonality is largely similar in
terms of both timing and magnitude across diverse climatic regions with substantiall

y different summer and winter temperatures . Insights of this kind would not have been possible analysing data averaged over time or nationally , or fixed to pre-specified frequencies . Prior studies have noted seasonality of mortality for all-cause mortality and for specific causes of death in the USA (Feinstein , 2002; Kalkstein , 2013; Rau , 2004; Rau et al . , 2018; Rosenwaike , 1966; Seretakis et al . , 1997) . Few of these studies have done consistent national and subnational analyses , and none has done so over time , for a comprehensive set of age groups and causes of death , and in relation to regional temperature differences . Our results on strong seasonality of cardiorespiratory diseases deaths and weak seasonality of cancer deaths , restricted to older ages , are broadly consistent with these studies (Feinstein , 2002; Rau et al . , 2018; Rosenwaike , 1966; Seretakis et al . , 1997) , which had limited analysis on how seasonality changes over time and geography (Feinstein , 2002; Rau et al . , 2018; Rosenwaike , 1966) . Similarly , our results on seasonality of injury deaths are supported by a few prior studies (Feinstein , 2002; Rau et al . , 2018; Rosenwaike , 1966) , but our subnational analysis over three decades revealed variations in when injury deaths peaked and in how seasonal differences in these deaths have changed over time in relation to age group which had not been reported before . A study of 36 cities in the USA , aggregated across age groups and over time , also found that excess mortality was not associated with seasonal temperature range (Kinney et al . , 2015) . In contrast , a European study found that the difference between winter and summer mortality was lower in colder Nordic countries than in warmer southern European nations (Healy , 2003; McKee , 1989) (the study’s measure of temperature was mean annual temperature which differed from the temperature difference between maximum and minimum mortality used in our analysis although the two measures are correlated) . The absence of variation in the magnitude of mortality seasonality indicates that different regions in the USA are similarly adapted to temperature seasonality , whereas Nordic countries may have better environmental (e . g . housing insulation and heating) and health system measures to counter the effects of cold winters than those in southern Europe . If the observed absence of association between the magnitude of mortality seasonality and seasonal temperature difference across the climate regions also persists over time , the changes in temperature as a result of global climate change are unlikely to affect the winter-summer mortality difference . The cause-specific analysis showed that the substantial decline in seasonal mortality differences in adolescents and young adults was related to the diminishing seasonality of (unintentional) injuries , especially from road traffic crashes , which are more likely to occur in the summer months (Liu et al . , 2005) and are more common in men . The weakening of seasonality in boys under five years of age was related to two phenomena: first , the seasonality of death from cardiorespiratory diseases declined , and second , the proportion of deaths from perinatal conditions , which exhibit limited seasonality (Figure 9—figure supplement 2 and Figure 10—figure supplement 3) , increased (MacDorman and Gregory , 2015) . In contrast to young and middle ages , mortality in older ages , where death rates are highest , maintained persistent seasonality over a period of three decades (we note that although the percent seasonal difference in mortality has remained largely unchanged in these ages , the absolute difference in death rates between the peak and minimum months has declined because total mortality has a declining long-term trend) . This finding demonstrates the need for environmental and health service interventions targeted towards this group irrespective of geography and local climate . Examples of such interventions include enhancing the availability of both environmental and medical protective factors , such as better insulation of homes , winter heating provision and flu vaccinations , for the vulnerable older population (Katiyo et al . , 2017) . Social interventions , including regular visits to the isolated elderly during peak mortality periods to ensure that they are optimally prepared for adverse conditions , and responsive and high-quality emergency care , are also important to protect this vulnerable group (Healy , 2003; Lerchl , 1998; Katiyo et al . , 2017) . Emergent new technologies , such as always-connected hands-free communications devices wi th the outside world , in-house cameras , and personal sensors also provide an opportunity to enhance care for the older , more vulnerable groups in the population , especially in winter when the elderly have fewer social interactions (Morris , 2013) . Such interventions are important today , and will remain so as the population ages and climate change increases the within- and between-season weather variability .

We used data on all 85 , 854 , 176 deaths in the USA from 1980 to 2016 from the National Center for Health Statistics (NCHS) . Age , sex , state of residence , month of death , and underlying cause of death were available for each record . The underlying cause of death was coded according to the international classification of diseases (ICD) system (9th revision of ICD from 1980 to 1998 and 10th revision of ICD thereafter) . Yearly population counts were available from NCHS for 1990 to 2016 and from the US Census Bureau prior to 1990 (Ingram et al . , 2003) . We calculated monthly population counts through linear interpolation , assigning each yearly count to July . We also subdivided the national data geographically into nine climate regions used by the National Oceanic and Atmospheric Administration (Figure 18 and Table 2) (Karl and Koss , 1984) . On average , the Southeast and South are the hottest climate regions with average annual temperatures of 18 . 4°C and 18°C respectively; the South also possesses the highest average maximum monthly temperature (27 . 9°C in July) . The lowest variation in temperature throughout the year is that of the Southeast (an average range of 17 . 5°C) . The three coldest climate regions are West North Central , East North Central and the Northwest (7 . 6°C , 8 . 0°C , 8 . 2°C respectively) . Mirroring the characteristics of the hottest climate regions , the largest variation in temperature throughout the year is that of the coldest region , West North Central (an average range of 30 . 5°C) , which also has the lowest average minimum monthly temperature (−6 . 5°C in January) . The other climate regions , Northeast , Southwest , and Central , possess similar average temperatures (10°C to 14°C) and variation within the year of (23°C to 26°C) , with the Northeast being the most populous region in the United States (with 19 . 8% total population in 2016) . Data were divided by sex and age in the following 10 age groups: 0–4 , 5–14 , 15-24 , 25–34 , 35–44 , 45–54 , 55–64 , 65–74 , 75–84 , 85+ years . We calculated monthly death rates for each age and sex group , both nationally and for sub-national climate regions . Death rate calculations accounted for varying length of months , by multiplying each month’s death count by a factor that would make it equivalent to a 31 day month . For analysis of seasonality by cause of death , we mapped each ICD-9 and ICD-10 codes to four main disease categories (Table 1) and to a number of subcategories which are presented in the Supplementary Note . Cardiorespiratory diseases and cancers accounted for 56 . 4% and 21 . 2% of all deaths in the USA , respectively , in 1980 , and 40 . 3% and 22 . 4% , respectively , in 2016 . Deaths from cardiorespiratory diseases have been associated with cold and warm temperatures (Basu , 2009; Basu and Samet , 2002; Bennett et al . , 2014; Braga et al . , 2002; Gasparrini et al . , 2015) . Injuries , which accounted for 8% of all deaths in the USA in 1980 and 7 . 3% in 2016 , may have seasonality that is distinct from so-called natural causes . We did not further divide other causes because the number of deaths could become too small to allow stable estimates when divided by age group , sex and climate region . We obtained data on temperature from ERA-Interim , which combines predictions from a physical model with ground-based and satellite measurements (Dee et al . , 2011) . We used gridded four-times-daily estimates at a resolution of 80 km to generate monthly population-weighted temperature by climate region throughout the analysis period . We used wavelet analysis to investigate seasonality for each age-sex group . Wavelet analysis uncovers the presence , and frequency , of repeated maxima and minima in each age-sex-specific death rate time series (Hubbard , 1998; Torrence and Compo , 1998) . In brief , a Morlet wavelet , described in detail elsewhere (Cazelles et al . , 2008) , is equivalent to using a moving window on the death rate time series and analysing periodicity in each window using a short-form Fourier transform , hence generating a dynamic spectral analysis , which allows measu

```
ring dynamic seasonal patterns , in which the periodicity of death rates may disappe
ar , emerge , or change over time . In addition to coefficients that measure the fre
quency of periodicity , wavelet analysis estimates the probability of whether the da
ta are different from the null situation of random fluctuations that can be represen
ted with white ( an independent random process ) or red ( autoregressive of order on
e process ) noise . For each age-sex group , we calculated the p-values of the prese
nce of 12 month seasonality for the comparison of wavelet power spectra of the entir
e study period ( 1980–2016 ) with 100 simulations against a white noise spectrum , w
hich represents random fluctuations . We used the R package WaveletComp ( version 1
. 0 ) for the wavelet analysis . Before analysis , we de-trended death rates using a
polynomial regression , and rescaled each death rate time series so as to range betw
een 1 and −1 . To identify the months of maximum and minimum death rates , we calcul
ated the centre of gravity and the negative centre of gravity of monthly death rates
. Centre of gravity was calculated as a weighted average of months of deaths , with
each month weighted by its death rate; negative centre of gravity was also calculate
d as a weighted average of months of deaths , but with each month was weighted by th
e difference between its death rate and the year’s maximum death rate . In taking th
e weighted average , we allowed December ( month 12 ) to neighbour January ( month 1
) , representing each month by an angle subtended from 12 equally-spaced points arou
nd a unit circle . Using a technique called circular statistics , a mean ( θ- ) of t
he angles ( θ1 , θ2 , θ3… , θn , ) representing the deaths ( with n the total number
of deaths in an age-sex group for a particular cause of death ) is found using the r
elation below:θ-=arg∑j=1nexp ( iθj ) , where arg denotes the complex number argument
and θj denotes the month of death in angular form for a particular death j . The out
come of this calculation is then converted back into a month value ( Fisher , 1995 )
. Along with each circular mean , a 95% confidence interval ( CI ) was calculated by
using 1000 bootstrap samples . The R package CircStats ( version 0 . 2 . 4 ) was use
d for this analysis . For each age-sex group and cause of death , and for each year
, we calculated the percent difference in death rates between the maximum and minimu
m mortality months . We fitted a linear regression to the time series of seasonal di
fferences from 1980 to 2016 , and used the fitted trend line to estimate how much th
e percentage difference in death rates between the maximum and minimum mortality mon
ths had changed from 1980 to 2016 . We weighted seasonal difference by the inverse o
f the square of its standard error , which was calculated using a Poisson model to t
ake population size of each age-sex group through time into account . This method gi
ves us a p-value for the change in seasonal difference per year , which we used to c
alculate the seasonal difference at the start ( 1980 ) and end ( 2016 ) of the perio
d of study . Our method of analysing seasonal differences avoids assuming that any s
pecific month or group of months represent highest and lowest number of deaths for a
particular cause of death , which is the approach taken by the traditional measure o
f Excess Winter Deaths . It also allows the maximum and minimum mortality months to
vary by age group , sex and cause of death .
-------------
```

In [8]:

```python
# function to extract first paragraph of the text
def get_abstract(text):
    """
        return abstract (first paragraph) of the text
    """
    result = ""
    result = text.split("\n")[0]
    return result

get_abstract(item["article"])
```

Out[8]:

```
'In temperate climates , winter deaths exceed summer ones . However , there is lim
ited information on the timing and the relative magnitudes of maximum and minimum
mortality , by local climate , age group , sex and medical cause of death . We use
d geo-coded mortality data and wavelets to analyse the seasonality of mortality by
age group and sex from 1980 to 2016 in the USA and its subnational climatic region
s . Death rates in men and women ≥ 45 years peaked in December to February and wer
e lowest in June to August , driven by cardiorespiratory diseases and injuries . I
n these ages , percent difference in death rates between peak and minimum months d
id not vary across climate regions , nor changed from 1980 to 2016 . Under five ye
ars , seasonality of all-cause mortality largely disappeared after the 1990s . In
adolescents and young adults , especially in males , death rates peaked in June/Ju
ly and were lowest in December/January , driven by injury deaths . '
```

In [9]:

```python
# function to extract first paragraph of the text
def get_conclusion(text):
    """
        return conclusion (second last paragraph) of the text
    """
    result = ""
    result = text.split("\n")[-2][:1000]  # limit to 1,000 characters
    return result

get_conclusion(item["article"])
```

Out[9]:

```
' We used wavelet and centre of gravity analyses , which allowed systematically id
entifying and characterizing seasonality of total and cause-specific mortality in
the USA , and examining how seasonality has changed over time . We identified dist
inct seasonal patterns in relation to age and sex , including higher all-cause sum
mer mortality in young men ( Feinstein , 2002; Rau et al . , 2018 ) . Importantly
, we also showed that all-cause and cause-specific mortality seasonality is largel
y similar in terms of both timing and magnitude across diverse climatic regions wi
th substantially different summer and winter temperatures . Insights of this kind
would not have been possible analysing data averaged over time or nationally , or
fixed to pre-specified frequencies . Prior studies have noted seasonality of morta
lity for all-cause mortality and for specific causes of death in the USA ( Feinste
in , 2002; Kalkstein , 2013; Rau , 2004; Rau et al . , 2018; Rosenwaike , 1966; Se
retakis et al . ,'
```

In [10]:

```python
# apply to dataset
df["abstract"] = df["article"].apply(get_abstract)
df["conclusion"] = df["article"].apply(get_conclusion)
df.head()
```

Out[10]:

	lay_summary	article	headings	keywords	id	abstract	conclu
0	In the USA , more deaths happen in the winter ...	In temperate climates , winter deaths exceed s...	[Abstract, Introduction, Results, Discussion, ...	[epidemiology and global health]	elife-35500-v1	In temperate climates , winter deaths exceed s...	We u wavelet centr gra analy
1	Most people have likely experienced the discom...	Whether complement dysregulation directly cont...	[Abstract, Introduction, Results, Discussion, ...	[microbiology and infectious disease, immunolo...	elife-48378-v2	Whether complement dysregulation directly cont...	Mechan advanc understan
2	The immune system protects an individual from ...	Variation in the presentation of hereditary im...	[Abstract, Introduction, Results, Discussion, ...	[microbiology and infectious disease, immunolo...	elife-04494-v1	Variation in the presentation of hereditary im...	We re that HOIL esse during
3	The brain adapts to control our behavior in di...	Rapid and flexible interpretation of conflicti...	[Abstract, Introduction, Results, Discussion, ...	[neuroscience]	elife-12352-v2	Rapid and flexible interpretation of conflicti...	We u intracra potentia me
4	Cells use motor proteins that to move organell...	Myosin 5a is a dual-headed molecular motor tha...	[Abstract, Introduction, Results, Discussion, ...	[structural biology and molecular biophysics]	elife-05413-v2	Myosin 5a is a dual-headed molecular motor tha...	Label-size a few ter nm traditi

In [11]:

```
# drop the article column to reduce file size
output_df = df.drop(columns=["article"])
output_df.head()
```

Out[11]:

	lay_summary	headings	keywords	id	abstract	conclusion
0	In the USA , more deaths happen in the winter ...	[Abstract, Introduction, Results, Discussion, ...	[epidemiology and global health]	elife-35500-v1	In temperate climates , winter deaths exceed s...	We used wavelet and centre of gravity analyse...
1	Most people have likely experienced the discom...	[Abstract, Introduction, Results, Discussion, ...	[microbiology and infectious disease, immunolo...	elife-48378-v2	Whether complement dysregulation directly cont...	Mechanistic advances in our understanding of ...
2	The immune system protects an individual from ...	[Abstract, Introduction, Results, Discussion, ...	[microbiology and infectious disease, immunolo...	elife-04494-v1	Variation in the presentation of hereditary im...	We report that HOIL-1 is essential during inf...
3	The brain adapts to control our behavior in di...	[Abstract, Introduction, Results, Discussion, ...	[neuroscience]	elife-12352-v2	Rapid and flexible interpretation of conflicti...	We used intracranial field potentials to meas...
4	Cells use motor proteins that to move organell...	[Abstract, Introduction, Results, Discussion, ...	[structural biology and molecular biophysics]	elife-05413-v2	Myosin 5a is a dual-headed molecular motor tha...	Label-sizes of a few tens of nm are tradition...

In [12]:
```
filename
```

Out[12]: 'eLife_train.jsonl'

In [13]:
```
path = "./data/extracted/"
filename = "eLife_train.csv"
print("Writing output to", filename)
output_df.to_csv(path + filename)
print("Completed")
```

```
Writing output to eLife_train.csv
Completed
```

Apply data extraction for all datasets

In [14]:
```
# testing filename conversion jsonl -> csv
filename = "eLife_train.jsonl"
print(filename)
print(filename[:filename.rfind(".")] + "_extracted.csv") # should be eLife_train.cs
```

```
eLife_train.jsonl
eLife_train_extracted.csv
```

In [15]:
```
file_path = "./data/biolaysumm2024_data/"
file_names = ["eLife_train.jsonl", "eLife_val.jsonl", "eLife_test.jsonl",
              "PLOS_train.jsonl", "PLOS_val.jsonl", "PLOS_test.jsonl"
```

```
            ]

output_path = "./data/extracted/"

print("Abstract text extraction:")
print("=============================")
for filename in file_names:
    print("Processing file =", filename)
    df = pd.read_json(file_path+filename,
                       orient="records",
                       lines=True)
    print("Number of records =", len(df))

    # apply get_abstract function
    print("Getting abstracts")
    df["abstract"] = df["article"].apply(get_abstract)
    print("Getting conclusions")
    df["conclusion"] = df["article"].apply(get_conclusion)
    print("Completed")
    output_df = df.drop(columns=["article"])
    output_filename = filename[:filename.rfind(".")] + "_extracted.csv"
    print("Writing output to", output_filename)
    output_df.to_csv(output_path + output_filename)
    print("Completed")
    print("--------------------")

print("======= completed ========")
```

```
Abstract text extraction:
=============================
Processing file = eLife_train.jsonl
Number of records = 4346
Getting abstracts
Getting conclusions
Completed
Writing output to eLife_train_extracted.csv
Completed
--------------------
Processing file = eLife_val.jsonl
Number of records = 241
Getting abstracts
Getting conclusions
Completed
Writing output to eLife_val_extracted.csv
Completed
--------------------
Processing file = eLife_test.jsonl
Number of records = 142
Getting abstracts
Getting conclusions
Completed
Writing output to eLife_test_extracted.csv
Completed
--------------------
Processing file = PLOS_train.jsonl
Number of records = 24773
Getting abstracts
Getting conclusions
Completed
Writing output to PLOS_train_extracted.csv
Completed
--------------------
Processing file = PLOS_val.jsonl
Number of records = 1376
Getting abstracts
Getting conclusions
Completed
Writing output to PLOS_val_extracted.csv
Completed
--------------------
Processing file = PLOS_test.jsonl
Number of records = 142
Getting abstracts
Getting conclusions
Completed
Writing output to PLOS_test_extracted.csv
Completed
--------------------
======= completed ========
```