# research

June 24, 2022

## 1 Cancer incidence trends in young adults

### 1.1 Introduction

The idea of this project comes from a conversation I recently have with a friend of mine who worked in a cancer retreat center. The friend is a medical professional who worked at the center for many years. He noticed that over the past 10 years, there have been more relatively young patients in the center than it was before. The friend was inclined to attribute this observation to the deteriorating quality of air, food, and other environmental factors. I skeptically suggested that there are other possible explanations of this phenomenon. 10 years ago, the friend himself was younger and could perceive typical patients as older.

Cancer is a class of diseases in which some of the body's cells grow uncontrollably and spread to other organs and tissues. Most cancers form a tumor. According to modern concepts, cancer is caused by changes to genes that control the way our cells function, especially how they grow and divide. Despite intensive research and development of new treatments, cancer remains the leading causes of death worldwide. In 2018, there were 18.1 million new cases and 9.5 million cancer-related deaths worldwide. By 2040, the number of new cancer cases per year is expected to rise to 29.5 million and the number of cancer-related deaths to 16.4 million [1].

Cancer can be considered an age-related disease because the incidence of most cancers increases with age, rising more rapidly beginning in midlife. Despite the fact that the disease can occur at any age, more than half of all cancers occurred in adults aged  65 years [2]. Therefore, for a long time, cancer was considered a disease predominantly affecting the elderly. Given this age specificity, a possible increase in the number of diseases among younger age groups may be of great interest.

Figure 1. Invasive cancer incidence, by age, U.S., 2009 [2]

As a software developer I am interested in Digital health. The idea of using information technologies to enhance the efficiency of healthcare seems very promising. Collection and analysis of health data using the data-science approach could potentially improve our understanding of diseases such as cancer. Therefore, I decided to further investigate my friend's observation in this project. Using available data and data-science approach it seems possible to determine if young adults have become more often diagnosed with cancer in recent decades.

### 1.2 Literature Review

Before setting goals and objectives, I decided to study the available information on this topic. It is important to understand what has already been investigated and what has not yet been explored. I found several articles on the Internet in which the issue is studied. Below I provide a summary and key findings that are relevant in the context of this study.

**Incidence trends for twelve cancers in younger adults—a rapid review. Br J Cancer 126, 1374–1386 (2022). [5]**

This paper analyzed the epidemiological information of some types of cancer in young adult patients, and came to the following conclusion: "Overall, this review provides evidence that some cancers are increasingly being diagnosed in younger age groups, although the mechanisms remain unclear." [5] This is a meta-analysis of existing studies on different types of cancer, but my goal is to explore the big picture across all types of cancer. The sources used in the article were mainly originated from the United States.

**Trends in Cancer Incidence in US Adolescents and Young Adults, 1973-2015 [6].**

Some findings from this paper: "In this serial cross-sectional, US population-based study using cancer registry data from 497 452 AYAs, the rate of cancer increased by 29.6% from 1973 to 2015, with kidney carcinoma increasing at the greatest rate. Breast carcinoma and testicular cancer were the most common cancer diagnoses for female and male AYAs, respectively." [6]

The autors coclude: "In this cross-sectional, US population-based study, cancer in AYAs was shown to have a unique epidemiological pattern and is a growing health concern, with many cancer subtypes having increased in incidence from 1973 to 2015. Continued research on AYA cancers is important to understanding and addressing the distinct health concerns of this population." [6] AYA is stands for Adolescents and Young Adults.

The findings from this paper also support the idea that the increase in the number of diseases among young people in the United states has natural causes. However, the article is again focused on the United States.

**Cancer Stat Facts: Cancer Among Adolescents and Young Adults (AYAs) (Ages 15–39) [7].**

The results of this work once again confirm the increase in the number of diseases amoung young adults in the United Satates: "Using statistical models for analysis, rates of new cancer cases of any site among AYAs have been rising on average 0.3% each year over 2010–2019, the last 10 years of available data." [7]. This study is based on a database provided by National Cancer Institute of US. This is a high fidelity database, but it covers only ths U.S. This work uses great data visualization methods that I want to use for inspiration.

As a result of a review of the literature available on the Internet on this topic, I found that there is ample evidence of an increase in the number of diseases among young adults in the United States. However, there are no studies on other regions and on worldwide population.

## 1.3 Aims and Objectives

According to the available literature, there is a trend in the United States that young people are more likely to get cancer in recent decades. However, the question remained unexplored whether there is such a trend around the world. My goal within this project is to fill this gap. I want to know if this is a global trend or this is pecific for certain regions.

The literature cited above suggests that the causes for the trend observed in the US remain unclear. If a similar trend takes place in the rest of the world, this may help in finding its causes. Since in this case the reasons may be common to the entire planet, and not specific to a particular region. However, the search for the causes is beyond the scope of this study.

For the purpose of this project, a young adult is considered to be between the ages of 15 and 44 (inclusive). This range was chosen because people younger than 15 have other types of cancer with different epidemiological dynamics that are outside the scope of this study [3]. The age group over 45 also has its own epidemiological dynamics. It has long seen an increase in morbidity, but it is attributed mainly to an increase in overall life expectancy and a decrease in mortality from other causes [4]. This is also outside the scope of this study.

## 1.4 Dataset

### 1.4.1 Requirements

To meet the goals of the project, the dataset must include worldwide cancer incidence statistics. Data must have at least two dimensions: the numbers should be broken down by year of diagnosis and a patient's age group. To pinpoint the trend, we need data for at least two decades. The data should be in machine-readable format.

### 1.4.2 Datasets considered

There are several organizations that collect statistics on cancer. In particular, the previously mentioned National Cancer Institute (NCI) provides SEER database. SEER (Surveillance, Epidemiology, and End Results) is an authoritative source for cancer statistics in the United States. As previously metioned, this is a open and high fidelity database, but it covers only ths U.S. so it is not suitable for our research. Another problem is that this dataset is not generaly available in machine-readable format.

Another considered dataset is Cancer registration statistics for England provided by Office for National Statistics. The data includes cancer diagnoses and age-standardised incidence rates for all types of cancer by age and sex. It is available in a machine-readable format as Excel tables. Despite the fact that these are data for one geographic region, from these files I get an idea in what form such statistics can be provided at all.

### 1.4.3 Choosen dataset

After research of different sources concerning the subject I found that there is an organization International Agency for Research on Cancer (IARC) that is a part of the World Health Organization of the United Nations. Its role is to conduct and coordinate research into the causes of cancer. It also collects and publishes surveillance data regarding the occurrence of cancer worldwide.

This data comes in datasets called Cancer Incidence in Five Continents (CI5). CI5 is the result of a long collaboration between the International Agency for Research on Cancer and the International Association of Cancer Registries. The series of monographs, published approximately every five years, has become the reference source of data on the international incidence of cancer. [8]

The whole dataset consists of separate publications (volumes) with data for different periods. These volumes are identified by Roman numerals (V, VI, VII, etc). The first 6 volumes (V, VI, VII, VIII, IX, and X) cover the period from 1973 till 2007. The last volume (XI) covers the period from 2008 till 2012. These publications are PDF files with very detailed reports on the incidence of cancer in different countries. Files with detailed source data are also provided along with reports. They are of primary interest to us.

The first 6 volumes are considered archived, they are downlable on this page. There are PDF reports and ZIP files with tabulated detailed (source) data. The latest volume is downlable on this page.

This dataset covers a large period of time from 1973 till 2012. The raw data is provided in machine-readable format (CSV, tabulated). There are detailed data on date of diagnosis, patient's age group, sex, geographic region, and cancer type. This is the most comprehensive source of information on cancer incidences in the world. Thus, this dataset is fully suitable for this study.

## 1.5 Method

After a more detailed examine of the dataset files, some problems were identified. The first problem is size of the dataset: more than 215 MB unziped. This is much more than the stated limit. The second problem is the extreme heterogeneity of the data. Each of these seven volumes has a different data format. Some files are in the form CSV, some are tab-separated. Different field names and other differences.

To overcome these difficulties, I decided to divide the data analysis process into several stages:

### 1.5.1 1. Preprocessing

Download raw files, cleansing, aggregate, and save only necessary information to an intermediate CSV file that serve as a source for the further stages. This file is relatively small and can be cached, so there is no need to run this stage more than once.

### 1.5.2 2. Processing

At this stage I no longer touch the raw files. The only source of the information is the intermediate CSV file I generate at the preprocessing stage. It is more precise cleansing and preparation for analysis.

### 1.5.3 3. Analysis

I want to explore the dataset through different lenses, in particular: * Dynamics of the total number of registered cancer cases in the world in the recent decades. * Dynamics of registered cancer cases in the world in younge adults in the recent decades. * Dynamics of the percentage of young people among all cases in the recent decades.

My assumption is that the evaluation of these metrics will be enough to answer this project's question: Is there a world trend that young people are more likely to be diagnosed with cancer in the recent decades.

## 1.6 Import necessary libraries

[1]:
```
!pip install pandas==1.4.2
!pip install matplotlib==3.5.2

import os
import requests
import zipfile
```

```
import re
import io
import codecs
from urllib.parse import urlparse
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
```

```
Requirement already satisfied: pandas==1.4.2 in
/Users/dima/opt/anaconda3/lib/python3.9/site-packages (1.4.2)
Requirement already satisfied: python-dateutil>=2.8.1 in
/Users/dima/opt/anaconda3/lib/python3.9/site-packages (from pandas==1.4.2)
(2.8.2)
Requirement already satisfied: numpy>=1.18.5 in
/Users/dima/opt/anaconda3/lib/python3.9/site-packages (from pandas==1.4.2)
(1.20.3)
Requirement already satisfied: pytz>=2020.1 in
/Users/dima/opt/anaconda3/lib/python3.9/site-packages (from pandas==1.4.2)
(2021.3)
Requirement already satisfied: six>=1.5 in
/Users/dima/opt/anaconda3/lib/python3.9/site-packages (from python-
dateutil>=2.8.1->pandas==1.4.2) (1.16.0)
Requirement already satisfied: matplotlib==3.5.2 in
/Users/dima/opt/anaconda3/lib/python3.9/site-packages (3.5.2)
Requirement already satisfied: pyparsing>=2.2.1 in
/Users/dima/opt/anaconda3/lib/python3.9/site-packages (from matplotlib==3.5.2)
(3.0.4)
Requirement already satisfied: cycler>=0.10 in
/Users/dima/opt/anaconda3/lib/python3.9/site-packages (from matplotlib==3.5.2)
(0.10.0)
Requirement already satisfied: fonttools>=4.22.0 in
/Users/dima/opt/anaconda3/lib/python3.9/site-packages (from matplotlib==3.5.2)
(4.25.0)
Requirement already satisfied: kiwisolver>=1.0.1 in
/Users/dima/opt/anaconda3/lib/python3.9/site-packages (from matplotlib==3.5.2)
(1.3.1)
Requirement already satisfied: pillow>=6.2.0 in
/Users/dima/opt/anaconda3/lib/python3.9/site-packages (from matplotlib==3.5.2)
(8.4.0)
Requirement already satisfied: numpy>=1.17 in
/Users/dima/opt/anaconda3/lib/python3.9/site-packages (from matplotlib==3.5.2)
(1.20.3)
Requirement already satisfied: packaging>=20.0 in
/Users/dima/opt/anaconda3/lib/python3.9/site-packages (from matplotlib==3.5.2)
(21.0)
Requirement already satisfied: python-dateutil>=2.7 in
/Users/dima/opt/anaconda3/lib/python3.9/site-packages (from matplotlib==3.5.2)
```

```
(2.8.2)
Requirement already satisfied: six in
/Users/dima/opt/anaconda3/lib/python3.9/site-packages (from
cycler>=0.10->matplotlib==3.5.2) (1.16.0)
```

### 1.7 Preprocessing

Download "detailed data" files from https://ci5.iarc.fr/ci5i-x/pages/download.aspx and https://ci5.iarc.fr/CI5-XI/Pages/download.aspx for each volumes.

[2]:
```
# list of URL of volume files from  https://ci5.iarc.fr/ci5i-x/pages/download.
 ↪aspx and
# https://ci5.iarc.fr/CI5-XI/Pages/download.aspx
urls = (
    "https://ci5.iarc.fr/CI5-X/CI5-Xd.zip",
    "https://ci5.iarc.fr/ci5i-x/old/vol9/CI5-IXd.zip",
    "https://ci5.iarc.fr/ci5i-x/old/vol8/CI5-VIIId.zip",
    "https://ci5.iarc.fr/ci5i-x/old/vol7/CI5-VIId.zip",
    "https://ci5.iarc.fr/ci5i-x/old/vol6/CI5-VI.zip",
    "https://ci5.iarc.fr/ci5i-x/old/vol5/CI5-V.zip",
    "https://ci5.iarc.fr/CI5-XI/CI5-XI.zip"
)

# create a new dir CI5 to save volume files
os.makedirs("CI5", exist_ok=True)

def download_if_not_exists(url, path):
    """Download a file from the url and save to the path only if file does not␣
 ↪already exist"""
    if os.path.exists(path):
        print(f"File {path} exists in cache")
    else:
        print(f"Downloading {url}...")
        response = requests.get(url)
        print(f"Save to {path}")
        open(path, "wb").write(response.content)

def unzip(file_path, target_dir):
    """Extract ZIP file file_path to target_dir"""
    print(f"Extract {file_path} to {target_dir}")
    with zipfile.ZipFile(file_path, "r") as zip_ref:
        zip_ref.extractall(target_dir)

for url in urls:
    url_path = urlparse(url).path
    file_name = os.path.basename(url_path)
    file_path = os.path.join("CI5", file_name)
    volume_name = os.path.splitext(file_name)[0]
```

```
    volume_path = os.path.join("CI5", volume_name)

    if os.path.exists(volume_path):
        print(f"Volume {volume_name} exists in cache")
    else:
        download_if_not_exists(url, file_path)
        unzip(file_path, os.path.join("CI5", volume_name))
```

```
Volume CI5-Xd exists in cache
Volume CI5-IXd exists in cache
Volume CI5-VIIId exists in cache
Volume CI5-VIId exists in cache
Volume CI5-VI exists in cache
Volume CI5-V exists in cache
Volume CI5-XI exists in cache
```

### 1.7.1 Volume V

[3]:
```
# registry.txt is a file with a list
v_registry_df = pd.read_csv("CI5/CI5-V/registry.txt", sep="\t", index_col=0)
v_cases_df = pd.read_csv("CI5/CI5-V/cases.csv", index_col=0)

v_df = v_registry_df.join(v_cases_df, how="inner", lsuffix="_registry")
v_df["PERIOD"] = v_df["PERIOD_1"].astype(str) + '-' + v_df["PERIOD_2"].
 ↪astype(str)
v_df = v_df[["PERIOD",␣
 ↪"N0_4","N5_9","N10_14","N15_19","N20_24","N25_29","N30_34","N35_39","N40_44","N45_49","N50_
v_df = v_df.groupby(["PERIOD"]).sum()

# v_df # uncomment to display data for this volume
```

### 1.7.2 Volume VI

[4]:
```
vi_registry_df = pd.read_csv("CI5/CI5-VI/registry.txt", sep="\t", index_col=0,␣
 ↪names=["REGISTRY", "PERIOD_1", "PERIOD_2", "NAME"])
vi_cases_df = pd.read_csv("CI5/CI5-VI/cases.csv", index_col=0)
vi_df = vi_registry_df.join(vi_cases_df, how="inner", lsuffix="_registry")
vi_df["PERIOD"] = vi_df["PERIOD_1"].astype(str) + '-' + vi_df["PERIOD_2"].
 ↪astype(str)
vi_df = vi_df[["PERIOD",␣
 ↪"N0_4","N5_9","N10_14","N15_19","N20_24","N25_29","N30_34","N35_39","N40_44","N45_49","N50_
vi_df = vi_df.groupby(["PERIOD"]).sum()
vi_df.rename(columns={"N85": "N85+", "N_unk": "N_UNK"}, inplace=True)

# vi_df # uncomment to display data for this volume
```

### 1.7.3 Volume VIId

[5]:
```
viid_registry_df = pd.read_csv("CI5/CI5-VIId/registry.txt", index_col=0,␣
↪names=["REGISTRY", "NAME"])
# remove the first row as it is a broken header
viid_registry_df = viid_registry_df[1:]

viid_name_split_df = viid_registry_df["NAME"].str.extract(r"(.
↪+)\s+\((\d+)-(\d+)\)", expand=True)
viid_name_split_df
viid_registry_df[["PERIOD_1", "PERIOD_2"]] = viid_name_split_df[[1,2]].
↪rename(columns= {1: "PERIOD_1", 2: "PERIOD_2"})
viid_registry_df.index =  viid_registry_df.index.astype(int)

viid_cases_df = pd.read_csv("CI5/CI5-VIId/CI5VII.csv", names=["REGISTRY",␣
↪"SEX", "CANCER_NUMBER", "AGE", "CASES_COUNT", "PERSON_YEARS"])
viid_cases_df["AGE"].replace({1: "N0_4", 2: "N5_9", 3: "N10_14", 4: "N15_19", 5:
↪ "N20_24", 6: "N25_29", 7: "N30_34", 8: "N35_39", 9: "N40_44", 10: "N45_49",␣
↪11: "N50_54", 12: "N55_59", 13: "N60_64", 14: "N65_69", 15: "N70_74", 16:␣
↪"N75_79", 17: "N80_84", 18: "N85+", 19: "N_UNK"}, inplace=True)
viid_cases_df = viid_cases_df.groupby(["REGISTRY", "AGE"])["CASES_COUNT"].sum().
↪to_frame().reset_index().pivot(index="REGISTRY", columns="AGE",␣
↪values="CASES_COUNT")

viid_df = viid_registry_df.join(viid_cases_df, how="inner", lsuffix="_registry")
viid_df["PERIOD"] = viid_df["PERIOD_1"].astype(str) + '-' + viid_df["PERIOD_2"].
↪astype(str)
viid_df = viid_df[["PERIOD",␣
↪"N0_4","N5_9","N10_14","N15_19","N20_24","N25_29","N30_34","N35_39","N40_44","N45_49","N50_
↪groupby(["PERIOD"]).sum()

# viid_df # uncomment to display data for this volume
```

### 1.7.4 Volume VIIId

[6]:
```
viiid_registry_df = pd.read_table("CI5/CI5-VIIId/registry.txt", index_col=0)
viiid_registry_df = viiid_registry_df.index.str.extract(r"\s*(\d+)\s+(.
↪*)\((\d+)(-(\d+))?\)", expand=True).drop(columns=3)
viiid_registry_df = viiid_registry_df.rename(columns= {0: "REGISTRY", 1:␣
↪"NAME", 2: "PERIOD_1", 4: "PERIOD_2"}).set_index("REGISTRY")
viiid_registry_df.index = viiid_registry_df.index.astype(int)

viiid_cases_df = pd.read_csv("CI5/CI5-VIIId/CI5-VIII.csv", names=["REGISTRY",␣
↪"SEX", "CANCER_NUMBER", "AGE", "CASES_COUNT", "PERSON_YEARS"])
```

```
viiid_cases_df["AGE"].replace({1: "N0_4", 2: "N5_9", 3: "N10_14", 4: "N15_19",␣
↪5: "N20_24", 6: "N25_29", 7: "N30_34", 8: "N35_39", 9: "N40_44", 10:␣
↪"N45_49", 11: "N50_54", 12: "N55_59", 13: "N60_64", 14: "N65_69", 15:␣
↪"N70_74", 16: "N75_79", 17: "N80_84", 18: "N85+", 19: "N_UNK"}, inplace=True)
viiid_cases_df = viiid_cases_df.groupby(["REGISTRY", "AGE"])["CASES_COUNT"].
↪sum().to_frame().reset_index().pivot(index="REGISTRY", columns="AGE",␣
↪values="CASES_COUNT")


viiid_df = viiid_registry_df.join(viid_cases_df, how="inner",␣
↪lsuffix="_registry")
viiid_df = viiid_df[["PERIOD_1", "PERIOD_2",␣
↪"N0_4","N5_9","N10_14","N15_19","N20_24","N25_29","N30_34","N35_39","N40_44","N45_49","N50_

# Fix a broken record for Taiwan
viiid_df.loc[81, 'PERIOD_1'] = 1993
viiid_df.loc[81, 'PERIOD_2'] = 1997

viiid_df["PERIOD"] = viiid_df["PERIOD_1"].astype(str) + '-' +␣
↪viiid_df["PERIOD_2"].astype(str)
viiid_df = viiid_df.groupby(["PERIOD"]).sum()

# viiid_df # uncomment to display data for this volume
```

### 1.7.5 Volume IXd

[7]:
```
ixd_registry_df = pd.read_table("CI5/CI5-IXd/registry.txt", names=["REGISTRY",␣
↪"NAME"], index_col=0)
ixd_registry_df = ixd_registry_df["NAME"].str.extract(r"\s*(.
↪*)\s*\((\d+)-(\d+)\)", expand=True)
ixd_registry_df = ixd_registry_df.rename(columns= {0: "NAME", 1: "PERIOD_1", 2:␣
↪"PERIOD_2"})

registry_dfs = []
for registry in ixd_registry_df.index:
    df = pd.read_csv(f"CI5/CI5-IXd/{registry}.csv", names=["SEX",␣
↪"CANCER_NUMBER", "AGE", "CASES_COUNT", "PERSON_YEARS"])
    df['REGISTRY'] = registry
    registry_dfs.append(df)

ixd_cases_df = pd.concat(registry_dfs)
ixd_cases_df["AGE"].replace({1: "N0_4", 2: "N5_9", 3: "N10_14", 4: "N15_19", 5:␣
↪"N20_24", 6: "N25_29", 7: "N30_34", 8: "N35_39", 9: "N40_44", 10: "N45_49",␣
↪11: "N50_54", 12: "N55_59", 13: "N60_64", 14: "N65_69", 15: "N70_74", 16:␣
↪"N75_79", 17: "N80_84", 18: "N85+", 19: "N_UNK"}, inplace=True)
```

```
ixd_cases_df = ixd_cases_df.groupby(["REGISTRY", "AGE"])["CASES_COUNT"].sum().
 ↪to_frame().reset_index().pivot(index="REGISTRY", columns="AGE",␣
 ↪values="CASES_COUNT")


ixd_df = ixd_registry_df.join(ixd_cases_df, how="inner", lsuffix="_registry")
ixd_df["PERIOD"] = ixd_df["PERIOD_1"].astype(str) + '-' + ixd_df["PERIOD_2"].
 ↪astype(str)
ixd_df = ixd_df[["PERIOD",␣
 ↪"N0_4","N5_9","N10_14","N15_19","N20_24","N25_29","N30_34","N35_39","N40_44","N45_49","N50_
ixd_df = ixd_df.groupby(["PERIOD"]).sum()

# ixd_df # uncomment to display data for this volume
```

### 1.7.6 Volume Xd

[8]:
```
# A file CI5/CI5-Xd/registry.txt contains unicode errors which cause error when␣
 ↪read it directly with pd.read_table()
with codecs.open("CI5/CI5-Xd/registry.txt", 'r', 'utf8', errors="ignore") as ff:
    content = ff.read()

xd_registry_df = pd.read_table(io.StringIO(content), names=["REGISTRY",␣
 ↪"NAME"], index_col=0)
xd_registry_df
xd_registry_df = xd_registry_df["NAME"].str.extract(r"\s*(.*)\s*\((\d+)-(?:
 ↪\d+,\d+-)?(\d+)\)", expand=True)
xd_registry_df
xd_registry_df = xd_registry_df.rename(columns= {0: "NAME", 1: "PERIOD_1", 2:␣
 ↪"PERIOD_2"})

registry_dfs = []
for registry in xd_registry_df.index:
    df = pd.read_csv(f"CI5/CI5-Xd/{registry}.csv", names=["SEX",␣
 ↪"CANCER_NUMBER", "AGE", "CASES_COUNT", "PERSON_YEARS"])
    df['REGISTRY'] = registry
    registry_dfs.append(df)

xd_cases_df = pd.concat(registry_dfs)

xd_cases_df["AGE"].replace({1: "N0_4", 2: "N5_9", 3: "N10_14", 4: "N15_19", 5:␣
 ↪"N20_24", 6: "N25_29", 7: "N30_34", 8: "N35_39", 9: "N40_44", 10: "N45_49",␣
 ↪11: "N50_54", 12: "N55_59", 13: "N60_64", 14: "N65_69", 15: "N70_74", 16:␣
 ↪"N75_79", 17: "N80_84", 18: "N85+", 19: "N_UNK"}, inplace=True)
xd_cases_df = xd_cases_df.groupby(["REGISTRY", "AGE"])["CASES_COUNT"].sum().
 ↪to_frame().reset_index().pivot(index="REGISTRY", columns="AGE",␣
 ↪values="CASES_COUNT")
```

```
xd_df = xd_registry_df.join(xd_cases_df, how="inner", lsuffix="_registry")
xd_df["PERIOD"] = xd_df["PERIOD_1"].astype(str) + '-' + xd_df["PERIOD_2"].
 ↪astype(str)
xd_df = xd_df[["PERIOD",␣
 ↪"N0_4","N5_9","N10_14","N15_19","N20_24","N25_29","N30_34","N35_39","N40_44","N45_49","N50_
xd_df = xd_df.groupby(["PERIOD"]).sum()

# xd_df # uncomment to display data for this volume
```

### 1.7.7 Volume XI

[9]:
```
# A file CI5/CI5-XI/registry.txt contains unicode errors which cause error when␣
 ↪read it directly with pd.read_table()
with codecs.open("CI5/CI5-XI/registry.txt", 'r', 'utf8', errors="ignore") as ff:
    content = ff.read()

xi_registry_df = pd.read_table(io.StringIO(content), names=["REGISTRY",␣
 ↪"NAME"], index_col=0)
xi_registry_df = xi_registry_df["NAME"].str.extract(r"\s*(.*)\s*\((\d+)-(?:
 ↪\d+,\d+-)?(\d+)\)", expand=True)
xi_registry_df = xi_registry_df.rename(columns= {0: "NAME", 1: "PERIOD_1", 2:␣
 ↪"PERIOD_2"})

xi_cases_df = pd.read_csv("CI5/CI5-XI/cases.csv", index_col=0)
xi_df = xi_registry_df.join(xi_cases_df, how="inner")
xi_df["PERIOD"] = xi_df["PERIOD_1"].astype(str) + '-' + xi_df["PERIOD_2"].
 ↪astype(str)
xi_df = xi_df[["PERIOD",␣
 ↪"N0_4","N5_9","N10_14","N15_19","N20_24","N25_29","N30_34","N35_39","N40_44","N45_49","N50_
xi_df = xi_df.groupby(["PERIOD"]).sum()
xi_df.rename(columns={"N85": "N85+", "N_unk": "N_UNK"}, inplace=True)

# xi_df # uncomment to display data for this volume
```

### 1.7.8 Save results

Merge data from all volumes and save to onle intermediate CSV file that serve as a source for the further stages. This file is relatively small and can be cached, so there is no need to run Preprocessing stage more than once.

[10]:
```
df = pd.concat([v_df, vi_df, viid_df, viiid_df, ixd_df, xd_df, xi_df])
df.to_csv("preprocessed.csv")
```

### 1.8 Processing

[11]:
```
df = pd.read_csv("preprocessed.csv", index_col=0)
years = df.index.str.extract("(\d+)-(\d+)", expand=True)
years_count = (years[1].astype(int) - years[0].astype(int) + 1)
years_count.set_axis(df.index, inplace=True)
df = df.floordiv(years_count, axis=0)
df.reset_index(inplace=True)
df.insert(1, "YEAR", df["PERIOD"].str.extract("(\d+)-(\d+)", expand=True).
 ↪apply(lambda x : list(range(int(x[0]), int(x[1])+1)), axis=1))
df = df.explode("YEAR")
df.reset_index(inplace=True)
df.drop(columns=["index", "PERIOD"], inplace=True)
df = df.groupby("YEAR").sum()
df
```

[11]:

| YEAR | N0_4 | N5_9 | N10_14 | N15_19 | N20_24 | N25_29 | N30_34 | N35_39 | N40_44 | \ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1973 | 10 | 6 | 8 | 10 | 19 | 22 | 31 | 45 | 61 | |
| 1974 | 10 | 6 | 8 | 10 | 19 | 22 | 31 | 45 | 61 | |
| 1975 | 10 | 6 | 8 | 10 | 19 | 22 | 31 | 45 | 61 | |
| 1976 | 10 | 6 | 8 | 10 | 19 | 22 | 31 | 45 | 61 | |
| 1977 | 5122 | 3576 | 2861 | 4333 | 5335 | 9055 | 14381 | 14798 | 20635 | |
| 1978 | 23687 | 14430 | 14395 | 23668 | 35280 | 57622 | 86279 | 109677 | 160021 | |
| 1979 | 33195 | 21014 | 21826 | 34713 | 49778 | 81630 | 131024 | 171260 | 251940 | |
| 1980 | 33537 | 21246 | 22105 | 35069 | 50232 | 82236 | 131970 | 172317 | 253521 | |
| 1981 | 33596 | 21279 | 22110 | 35107 | 50363 | 82472 | 132099 | 172609 | 254019 | |
| 1982 | 28646 | 18058 | 19466 | 31104 | 45516 | 74113 | 118786 | 159537 | 233640 | |
| 1983 | 14323 | 8580 | 8825 | 14912 | 24124 | 39342 | 60955 | 92053 | 126046 | |
| 1984 | 14644 | 8766 | 9021 | 15167 | 24470 | 39879 | 61825 | 93342 | 127566 | |
| 1985 | 14776 | 8880 | 9148 | 15334 | 24710 | 40222 | 62338 | 94217 | 128912 | |
| 1986 | 15401 | 9353 | 9700 | 16198 | 25592 | 41644 | 64130 | 97421 | 133783 | |
| 1987 | 14790 | 8936 | 9183 | 15399 | 24653 | 40377 | 61859 | 93530 | 129387 | |
| 1988 | 33066 | 20534 | 19500 | 31279 | 49357 | 84211 | 128168 | 183983 | 264724 | |
| 1989 | 34131 | 21324 | 20207 | 32487 | 51303 | 87184 | 132708 | 190791 | 275653 | |
| 1990 | 33833 | 21241 | 20287 | 32274 | 50587 | 85294 | 130296 | 188269 | 273080 | |
| 1991 | 31271 | 19965 | 19095 | 29995 | 46570 | 78847 | 121335 | 174505 | 249060 | |
| 1992 | 30227 | 19019 | 18192 | 28889 | 45077 | 76794 | 118195 | 169690 | 242220 | |
| 1993 | 30285 | 18654 | 17880 | 28478 | 45057 | 76890 | 117284 | 168367 | 241516 | |
| 1994 | 31740 | 19834 | 18963 | 30225 | 47668 | 81217 | 124178 | 178806 | 257340 | |
| 1995 | 33383 | 20932 | 19911 | 31621 | 49871 | 84740 | 129553 | 186416 | 269326 | |
| 1996 | 32890 | 20575 | 19579 | 31159 | 49067 | 83412 | 127574 | 183774 | 265586 | |
| 1997 | 30968 | 19186 | 18534 | 29371 | 46081 | 78984 | 121401 | 176401 | 256343 | |
| 1998 | 109869 | 67159 | 71856 | 109507 | 159256 | 256801 | 410626 | 652089 | 1034790 | |
| 1999 | 113248 | 69384 | 74033 | 112888 | 164045 | 265312 | 425714 | 677323 | 1075382 | |
| 2000 | 113583 | 69647 | 74311 | 113248 | 164411 | 266074 | 426959 | 679528 | 1078524 | |
| 2001 | 113338 | 69500 | 74158 | 112984 | 163948 | 265333 | 425738 | 677500 | 1075331 | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 2002 | 111920 | 68492 | 73042 | 111355 | 161320 | 261234 | 419482 | 667729 | 1060644 |
| 2003 | 147703 | 83873 | 95843 | 152205 | 233787 | 354176 | 543922 | 840667 | 1414114 |
| 2004 | 149042 | 84738 | 96867 | 153800 | 236081 | 357640 | 549438 | 849831 | 1428941 |
| 2005 | 149862 | 85460 | 97611 | 154913 | 237795 | 360237 | 553254 | 855754 | 1437484 |
| 2006 | 149206 | 84958 | 97018 | 154091 | 236665 | 358422 | 550570 | 851708 | 1431312 |
| 2007 | 148366 | 84396 | 96382 | 153074 | 235059 | 355661 | 545967 | 844602 | 1421165 |
| 2008 | 117091 | 64448 | 73361 | 125555 | 205446 | 339283 | 510044 | 784525 | 1308915 |
| 2009 | 117455 | 64678 | 73558 | 125827 | 205954 | 340136 | 511389 | 786952 | 1312801 |
| 2010 | 118613 | 65366 | 74268 | 127048 | 207955 | 343313 | 516568 | 795756 | 1326236 |
| 2011 | 117624 | 64754 | 73540 | 125905 | 206119 | 340252 | 511136 | 787020 | 1312199 |
| 2012 | 116821 | 64222 | 72926 | 124858 | 204470 | 337685 | 506939 | 780578 | 1302184 |

| YEAR | N45_49 | N50_54 | N55_59 | N60_64 | N65_69 | N70_74 | N75_79 | N80_84 | \ |
|---|---|---|---|---|---|---|---|---|---|
| 1973 | 96 | 133 | 164 | 202 | 246 | 272 | 236 | 222 | |
| 1974 | 96 | 133 | 164 | 202 | 246 | 272 | 236 | 222 | |
| 1975 | 96 | 133 | 164 | 202 | 246 | 272 | 236 | 222 | |
| 1976 | 96 | 133 | 164 | 202 | 246 | 272 | 236 | 222 | |
| 1977 | 31154 | 44179 | 62395 | 71547 | 106349 | 107060 | 80649 | 47391 | |
| 1978 | 243303 | 368573 | 518873 | 592673 | 729451 | 729843 | 591088 | 377385 | |
| 1979 | 385579 | 600621 | 897084 | 1049378 | 1336703 | 1391126 | 1151620 | 729088 | |
| 1980 | 388155 | 604319 | 900688 | 1053558 | 1340169 | 1394676 | 1154306 | 730718 | |
| 1981 | 389112 | 606023 | 903490 | 1056167 | 1343126 | 1398013 | 1156649 | 731929 | |
| 1982 | 356797 | 560806 | 841288 | 989353 | 1237456 | 1295790 | 1082688 | 689155 | |
| 1983 | 180290 | 262396 | 387423 | 515274 | 549490 | 613962 | 536408 | 351052 | |
| 1984 | 182556 | 265930 | 392364 | 521369 | 554753 | 619864 | 541675 | 354462 | |
| 1985 | 184870 | 269149 | 397119 | 527010 | 559931 | 625819 | 550246 | 356122 | |
| 1986 | 191861 | 278445 | 410231 | 542900 | 576868 | 642835 | 566034 | 367329 | |
| 1987 | 184866 | 268094 | 394795 | 522709 | 555406 | 614493 | 542987 | 353668 | |
| 1988 | 345487 | 470771 | 669744 | 944003 | 1171203 | 1105849 | 1049082 | 726927 | |
| 1989 | 359742 | 489689 | 696670 | 980815 | 1218062 | 1150158 | 1091607 | 757127 | |
| 1990 | 349858 | 474651 | 678879 | 961551 | 1193814 | 1134710 | 1062635 | 739380 | |
| 1991 | 318957 | 428341 | 608318 | 854093 | 1043063 | 988568 | 910571 | 626738 | |
| 1992 | 309791 | 414893 | 588845 | 831950 | 1020496 | 972564 | 891551 | 614270 | |
| 1993 | 312343 | 418150 | 595535 | 842444 | 1049968 | 1003198 | 937555 | 649123 | |
| 1994 | 333252 | 454216 | 647198 | 912436 | 1124048 | 1062185 | 995695 | 687910 | |
| 1995 | 349026 | 473601 | 674399 | 950008 | 1170498 | 1107042 | 1037481 | 718334 | |
| 1996 | 344457 | 466203 | 663047 | 934466 | 1154424 | 1094495 | 1024348 | 708829 | |
| 1997 | 330487 | 448391 | 639885 | 904350 | 1116913 | 1061567 | 985473 | 682078 | |
| 1998 | 1510481 | 2172050 | 2692099 | 3235231 | 3937117 | 4418724 | 4078922 | 2764978 | |
| 1999 | 1561816 | 2233873 | 2767725 | 3330664 | 4036844 | 4509502 | 4146379 | 2797931 | |
| 2000 | 1566000 | 2238210 | 2772049 | 3336099 | 4043485 | 4515550 | 4150632 | 2801272 | |
| 2001 | 1560711 | 2230368 | 2761893 | 3322234 | 4024375 | 4494458 | 4131426 | 2789238 | |
| 2002 | 1539036 | 2199217 | 2722677 | 3269048 | 3956009 | 4420729 | 4065541 | 2751098 | |
| 2003 | 2227405 | 3183196 | 4220899 | 4810798 | 5384873 | 5460588 | 5277478 | 4020967 | |
| 2004 | 2250259 | 3215619 | 4263991 | 4859524 | 5443156 | 5524762 | 5337290 | 4063161 | |
| 2005 | 2262419 | 3232463 | 4286835 | 4887540 | 5474882 | 5559138 | 5370255 | 4088752 | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 2006 | 2253789 | 3221046 | 4272041 | 4869425 | 5452180 | 5534281 | 5347174 | 4072169 |
| 2007 | 2240043 | 3201627 | 4244280 | 4833894 | 5407847 | 5486033 | 5301596 | 4038558 |
| 2008 | 2205168 | 3379021 | 4398763 | 5424524 | 5775077 | 5359155 | 4738410 | 3791839 |
| 2009 | 2210790 | 3385754 | 4406648 | 5431726 | 5781314 | 5365506 | 4743722 | 3795163 |
| 2010 | 2229392 | 3407371 | 4435555 | 5462061 | 5810498 | 5397250 | 4772031 | 3814841 |
| 2011 | 2207831 | 3376812 | 4394625 | 5411259 | 5753653 | 5336475 | 4713624 | 3769572 |
| 2012 | 2193493 | 3358216 | 4371705 | 5383837 | 5723832 | 5305315 | 4683903 | 3747025 |

| YEAR | N85+ | N_UNK |
|---|---|---|
| 1973 | 151 | 0 |
| 1974 | 151 | 0 |
| 1975 | 151 | 0 |
| 1976 | 151 | 0 |
| 1977 | 26982 | 2501 |
| 1978 | 255397 | 9787 |
| 1979 | 490048 | 5988 |
| 1980 | 490600 | 7248 |
| 1981 | 491552 | 6278 |
| 1982 | 468374 | 4575 |
| 1983 | 235475 | 2441 |
| 1984 | 237503 | 2555 |
| 1985 | 238419 | 2623 |
| 1986 | 245540 | 3942 |
| 1987 | 237392 | 2999 |
| 1988 | 503246 | 5257 |
| 1989 | 522876 | 5659 |
| 1990 | 516884 | 6833 |
| 1991 | 432399 | 6800 |
| 1992 | 425106 | 4785 |
| 1993 | 451701 | 4457 |
| 1994 | 475222 | 4996 |
| 1995 | 496236 | 6556 |
| 1996 | 490322 | 6812 |
| 1997 | 474303 | 4984 |
| 1998 | 2229700 | 21824 |
| 1999 | 2250814 | 21829 |
| 2000 | 2252117 | 21887 |
| 2001 | 2240884 | 21729 |
| 2002 | 2205894 | 20301 |
| 2003 | 3110408 | 23134 |
| 2004 | 3135004 | 23149 |
| 2005 | 3155764 | 23178 |
| 2006 | 3144711 | 22851 |
| 2007 | 3121475 | 22492 |
| 2008 | 3353773 | 1109 |
| 2009 | 3355792 | 1109 |

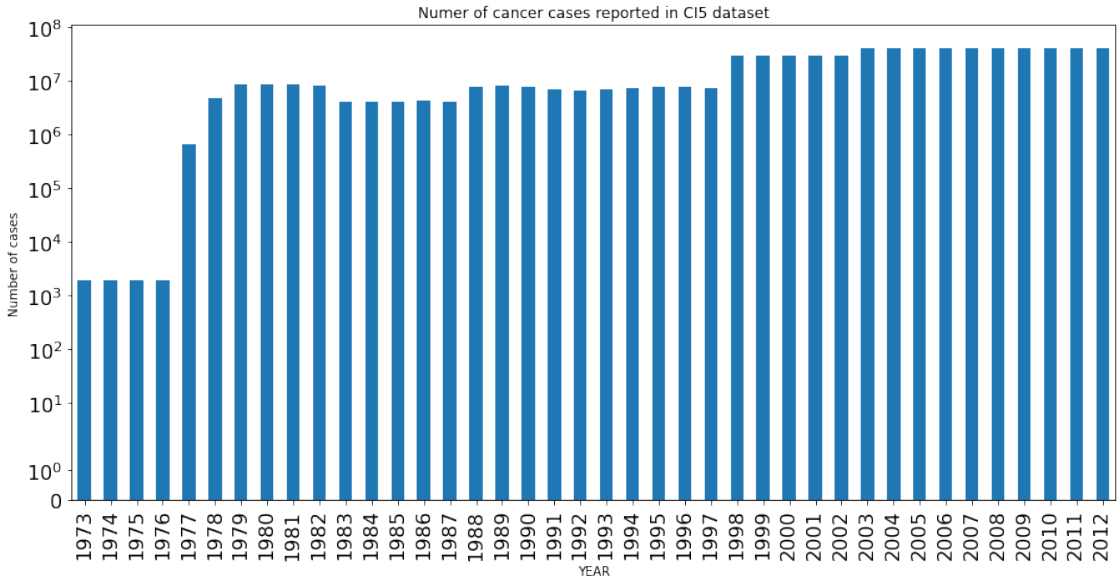
```
2010   3367913    1459
2011   3333067    1446
2012   3315440    1375
```

### 1.9 Analysis

Let's visualize dynamics of the total number of registered cancer cases in the world:

[12]:
```
total_cases = df.sum(axis=1)
total_cases.plot(kind="bar", logy="sym", figsize=(15,7), ylabel="Number of␣
↪cases",
                 title="Numer of cancer cases reported in CI5 dataset",␣
↪fontsize=16)
```

[12]:
```
<AxesSubplot:title={'center':'Numer of cancer cases reported in CI5 dataset'},
xlabel='YEAR', ylabel='Number of cases'>
```



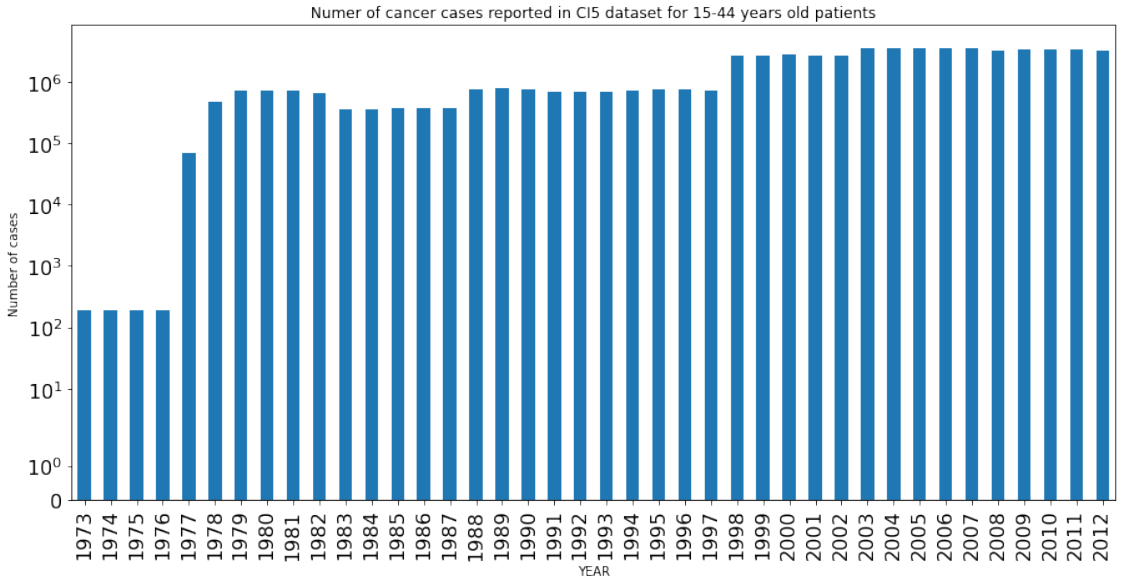

As we can see from this graph, the number of reported cancer cases increases in steps over time. Remember, our dataset consists of individual reports (volumes) that were published at different times. Different reports had different coverage of regions. These steps show the boundaries of different volumes. As we can see, over time the coverage has improved and this was reflected in more reported cases on the right side of the graph. Unfortunately, due to this feature, we cannot rely on absolute values in our dataset.

Let's visualize dynamics of registered cancer cases in the world in younge adults in the recent decades:

[13]:
```
young_adults_cases = df["N15_19"] + df["N20_24"] + df["N25_29"] + df["N30_34"]␣
↪+ df["N35_39"] + df["N40_44"]
```

```
young_adults_cases.plot(kind="bar", logy="sym", figsize=(15,7), ylabel="Number␣
↪of cases", fontsize=16,
                    title="Numer of cancer cases reported in CI5 dataset for␣
↪15-44 years old patients")
```

[13]:
```
<AxesSubplot:title={'center':'Numer of cancer cases reported in CI5 dataset for
15-44 years old patients'}, xlabel='YEAR', ylabel='Number of cases'>
```

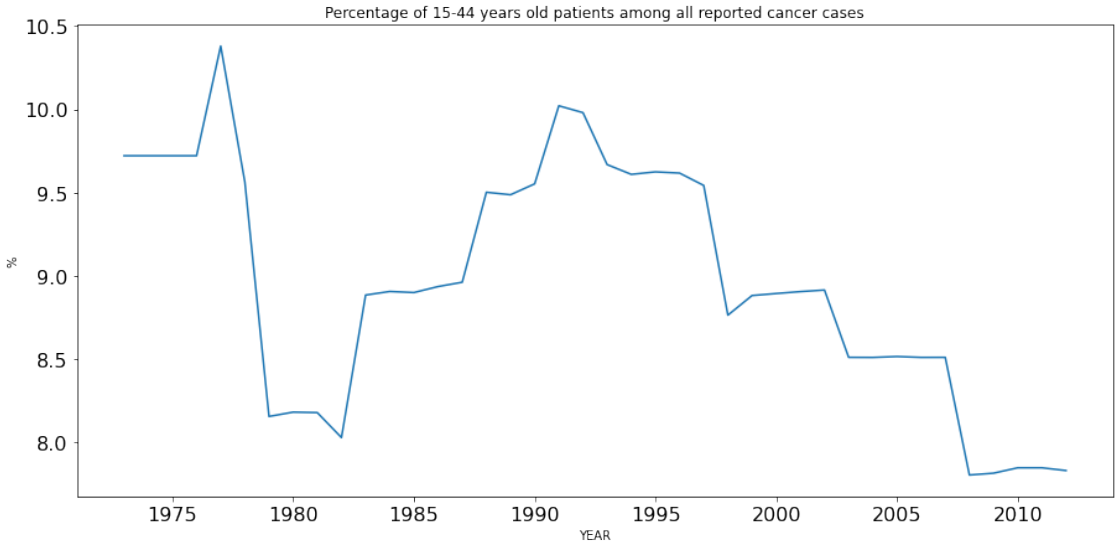



Here we see almost the same figure as previous one. From this we can conclude that more cases of cancer have been reported among younger patients in recent decades. However, this feature of the dataset may not reflect the actual dynamics. To better understand this data, you need to look at relative numbers.

Let's visualize dynamics of the percentage of young people among all cases in the recent decades.

[14]:
```
(young_adults_cases / total_cases * 100).plot(figsize=(15,7), ylabel="%",␣
↪fontsize=16,
                                        title="Percentage of 15-44 years old␣
↪patients among all reported cancer cases")
```

[14]:
```
<AxesSubplot:title={'center':'Percentage of 15-44 years old patients among all
reported cancer cases'}, xlabel='YEAR', ylabel='%'>
```




The graph shows that the percentage of young patients among all recorded cancer cases is around 9% and not very volatile over time. There is no trend for a significant increase in the proportion of young patients in recent years.

## 1.10 Conclusion

In the reviewed literature, there is evidence of an increase in the incidence of cancer among young patients in some regions, in particular in the USA. The aim of this project was to analyze the data for the entire world to understand whether such a trend is taking place worldwide, or if it is a local phenomenon. After analyzing the data, I came to the conclusion that on a global scale there is no trend towards an increase in the proportion of young people among all cases of cancer. This result may mean that the causes for the increase in cancer cases in the considered regions are of a local nature and are not the result of any global changes, such as climate change.

## 1.11 References

[1] "What Is Cancer?" by National Cancer Institute (2021, May 5) [Online]. Available: https://www.cancer.gov/about-cancer/understanding/what-is-cancer

[2] "Age and Cancer Risk" Am J Prev Med. 2014 Mar; 46(3 0 1): S7–15. [Online]. Available: https://doi.org/10.1016/j.amepre.2013.10.029

[3] "Childhood Cancers" by National Cancer Institute (2021, April 12) [Online]. Available: https://www.cancer.gov/types/childhood-cancers

[4] "The Challenging Landscape of Cancer and Aging: Charting a Way Forward" by Norman E. Sharpless, M.D. (2018, January 24) [Online]. Available: https://www.cancer.gov/news-events/cancer-currents-blog/2018/sharpless-aging-cancer-research

[5] di Martino, E., Smith, L., Bradley, S.H. et al. Incidence trends for twelve cancers in younger adults—a rapid review. Br J Cancer 126, 1374–1386 (2022). [Online]. Available: https://doi.org/10.1038/s41416-022-01704-x

[6] Scott AR, Stoltzfus KC, Tchelebi LT, et al. Trends in Cancer Incidence in US Adolescents and Young Adults, 1973-2015. JAMA Netw Open. 2020;3(12):e2027738. [Online]. Available: https://doi.org/10.1001/jamanetworkopen.2020.27738

[7] Cancer Stat Facts: Cancer Among Adolescents and Young Adults (AYAs) (Ages 15–39) by National Cancer Institute (2022) [Online]. Available: https://seer.cancer.gov/statfacts/html/aya.html

[8] CI5: CANCER INCIDENCE IN FIVE CONTINENTS by International Agency for Research on Cancer (IARC) [Online]. Available: https://ci5.iarc.fr/Default.aspx