**PROJECT REPORT**

**Topic- Assessing if government research funding meets the need in terms of fiscal burden of cancer**

**Business-related questions:**

* Question 1- What is the financial implication of cancer treatment across the (i) US states and (ii) for different cancer types/phases?
* Question 2- Which ethnic groups are most susceptible to cancer challenges?
* Question 3- What is the relationship between the government’s cancer research funding and patient’s out-of-pocket cost?

**Reason for picking this topic and question(s)**

Cancer is increasing worldwide. The formulation of plans related to cancer support and their success hinges upon understanding the interplay of different factors such as cost of treatment, number of people getting affected, ethnic susceptibility, and financial research support. It can potentially help to decrease cancer implication, including deaths due to cancer. A better understanding would lead to right distribution of funding and efforts, including cancer care and treatment for each cancer type.

Stakeholders interested in this question(s): Government, insurance Companies, hospitals, cancer support organization, cancer research institutes and public (for awareness)

Who are the results of your analysis ultimately for?

The results of the analysis are ultimately for government healthcare agencies that are planning to roll out the financial support program. The data can help the agency to understand the fiscal burden of cancer across different cancer types. Moreover, the agency can concentrate on specific cancer types based on the incidence/deaths or out-of-pocket cost per patient, and thus increase funding in such areas.

**2. Data:**

Data Source 1: https://gis.cdc.gov/Cancer/USCS/#/RaceEthnicity/

The dataset highlights Age-Adjusted Cancer Incidence Rates grouped by race and ethnicity. The dataset helps understand how each type of cancer impact different ethnicities or sex, and thus helps in better understanding Business Question 2.

The data in the dataset was data are compiled from selected cancer registries meeting U.S. Cancer Statistics data quality criteria. The variables in the data for Age-Adjusted Cancer Incidence Rates include All Races and Ethnicities, White, Black, American Indian/Alaska Native, Asian/Pacific Islander, and Hispanic.

Data Source 2: https://academic.oup.com/jnci/article/113/12/1670/6409890?login=false

The data source is Table 7 from the Annual Report to the “Nation on the Status of Cancer, Part 2: Patient Economic Burden Associated with Cancer Care” published in the Journal of the National Cancer Institute, Volume 113, Issue 12 (December 2021). The dataset highlights financial and time costs for patients of varying cancer types and helps in answering Business Question 1 and Question 3.

The variables in the datasets include Patient out-of-pocket costs by phase of care, patient time cost, and patient economic burden.

Dataset Source 3: https://www.cancer.gov/about-nci/budget/fact-book/data/research-funding

The dataset highlighting Funding for Research Areas is taken from the National Cancer Institute, a government source. It highlights that appropriated funds are spent based on different categories or classifications, including specific cancer sites and types. The dataset helps in the analysis of the business question 3. The data in the sources was collected from [NCI Funded Research Portfolio website](https://fundedresearch.cancer.gov/nciportfolio/).

The variables in the data sets include year-wise actual funding (2015-2018) and funding projections (2019-2020) for the cancer types.

Dataset Source 4: https://statecancerprofiles.cancer.gov/incidencerates/

Multiple data sets (21) were extracted from the data source with a focus on state-wise annual cancer incidence counts for different types of cancer. It is a government data source. Therefore, it would be used to support analysis for all the three business questions. As per the source, the data was collected from [State Cancer Registries](https://www.cdc.gov/cancer/npcr/index.htm) and [Surveillance, Epidemiology, and End Results](http://seer.cancer.gov/).

Variables in the data set include Age-adjusted incidence rate, CI Rank, Average Annual Count, and Recent Trend, Recent 5-year Trend in Incidence Rates. For the project, the focus would be on the Average Annual Count.

**3. Information quality:**

The data quality concerns include:

* Absence of reference or common column – The cancer types for each of the dataset were different. Thus, the records were restructured depending on the cancer sites.
* Missing Data: There are missing data in the dataset 1, which can impact the research quality. To fill such data, statistical measures (mean or median) will be used.
* Little added value content- The columns not adding any value to the business questions would be discarded accordingly.
* Time Difference- There is time variation for the datasets.
* Data Formatting- Combining and organizing the data by common reference points lead to the loss of certain values. The data mapping would be done to cover as much as values are clearly aligned.

Three of the four datasets used are from government sources, considered reputable and trustworthy due to the strategic process from data collection to analysis.

Given that several data quality issues would be addressed using different techniques, it is suitable for addressing questions using the methods and tools planned to be used during the project.

**4. Methods and tools:**

**Methods:** Data Discovery, Data Transformation, Data Enrichment, Data Formatting, Data Profiling, and Data Pre-processing

**Tools:** Jupyter Notebook, Excel (for formatting and dashboard), and Lucid chart

**Theories:** The Iterative process of wrangling and analysis is used as a theory for the project. (Kandel, 2011)

**Language:** Python

**Challenges:**

* Researching to identify the key cancer sites and grouped the similar cancer types to a single cancer site from each dataset accordingly.
* Extracting heatmap using Python case in Jupyter Pandas
* The missing data, columns with little value and combining different datasets presented challenges for the project.

**5. Data Wrangling Process:**

Business Question 1:

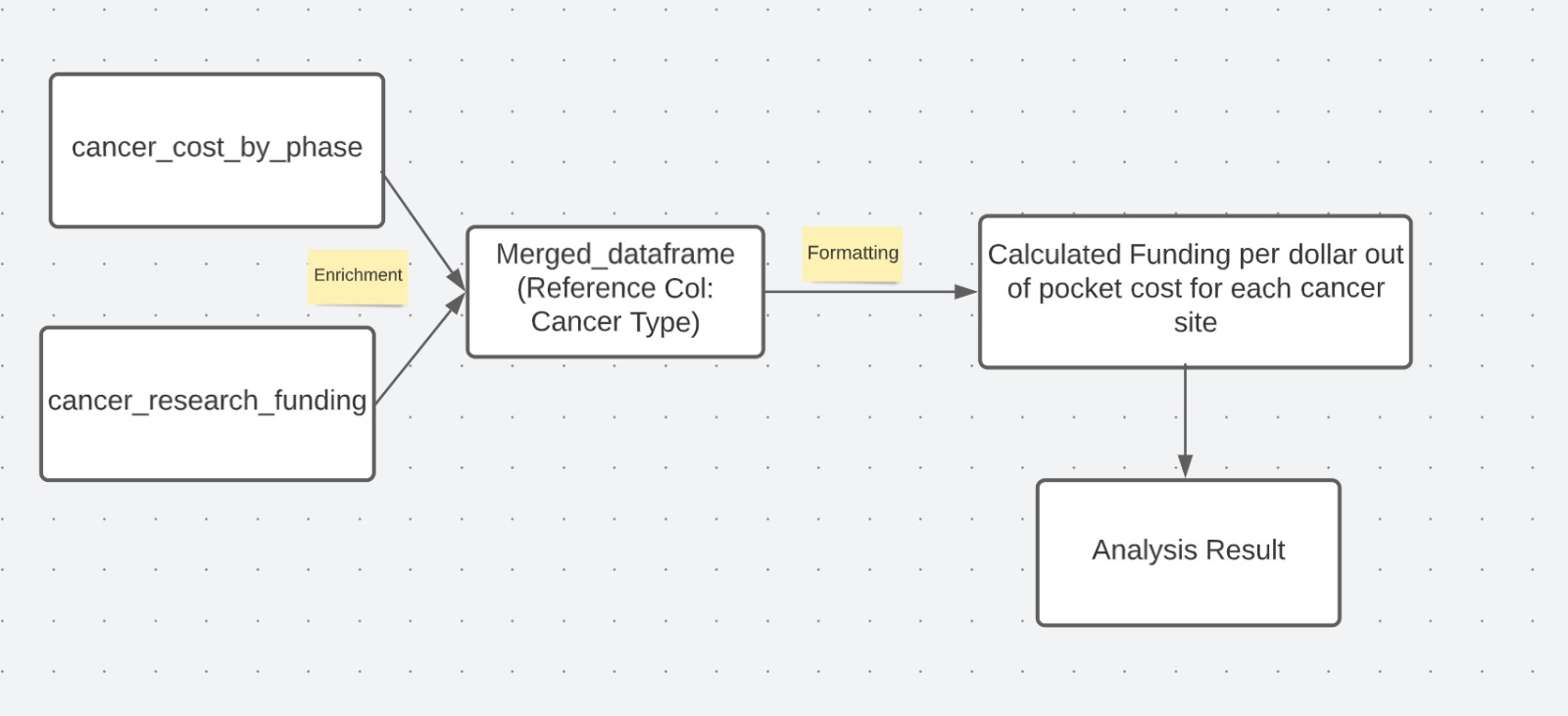
Diagram

Description automatically generated

Business Question 2:

We considered Demographics of different races and ethnicities dataset and performed several data wrangling steps such as grouping the data, and identifying minimum, maximum and mean values for each race.

Business Question 3:



**Data Wrangling Steps**

**Data Discovery**

Identified trends or patterns in each of the four datasets., along with several issues such as missing values, absence of common column for the dataset.

Trends and patterns identified from the datasets includes:

* Dataset 1: (Demographics data)

The columns Year, Sex, AgeGroup, RateType and DataType had single value for all records.

The data is for 2019, highlighting cancer count, cancer rate (age-adjusted) and population for each of the races and ethnicities for all ages and both male and females. Few of the data values in the Rate, Count and Population columns were suppressed.

* Dataset 2: (Cost\_Phase data)

The sum of cost from all the phases including initial, continuing, End of life cancer death and End-of-life other cause of cause of death almost totalled Total-out-of pocket costs for each of the cancer. The cost was in million dollars and there was no missing value for any record.

* Dataset 3: (Funding data)

The variables in the datasets include year\_wise expenditure of appropriated research funds for different types of cancer as well as AIDS and Total NCI Budget. There were no missing values in the dataset.

* Dataset 4: (States data)

There were several columns with the Average Annual Count column being the focus of the project. There were few missing values for Recent Trend, Recent 5-Year Trend, Lower 95% Confidence Interval and Upper 95% Confidence Interval.

**Data Formatting**

Filtering required columns

* Selected Average annual count column to create a dataframe cancer\_count\_states. It was done using excel and upon several datasets.
* Identified 13 cancer sites that covers records from all datasets comprehensively. Grouped the data from each dataset to match these cancer sites (Question 1, Question 2 and Question 3).
* Dropped the last row (including due to system error) from Dataset 1 (Question 1).
* Identified the required columns and dropped other columns from the above combined dataframe. (Question 1)
* Created the subset by taking only CancerType','Race','Rate','Count','Population’ and leaving out 'Year', 'Sex', 'AgeGroup', 'RateType' and 'DataType'. (Question 2)
* Grouped data for each demographics using the get\_group() function. (Question 2)

Structuring

* Removing dollar sign and comma from Dataset 2 (Question 1 and Question 3)
* Rounded the values in the ‘Merged\_dataframe’ to two decimal places. (Question 3)
* Set race as the index using the set\_index() function. (Question 2)
* Changed rate data type to float using astype() function. (Question 2)
* Plotted the bar graphic for each cancer type and associated rate using white.plot.bar() function. (Question 2)
* Replaced ‘Data Suppressed’ with 0 for various records using replace() function. (Question 2)

Renaming the columns

* Renamed the columns in the ‘Merged\_dataframe’. (Question 3)

**Data Profiling**

* Submitted data profiling reports of updated datasets after pre formatting in excel manually.
* Data source- Data was sourced from government sources and peer-reviewed journal article, both of which are reputable sources.
* Citations and Descriptions- The sources have citations and descriptions on how the data was obtained.
* Process (Ongoing/Past)- Data collection is not ongoing.
* Description- Description of the columns was provided in the data sources.
* Found number of rows, number of columns, missing values from the data profiling report.
* Found mean rate for each race using the mean () function. (Question 2)
* Found largest and minimum value and associated cancer types using nlargest() and nsmallest() function, respectively. (Question 2)

**Data Pre-processing**

* Calculated the number of missing values for each column in the Dataset 3. (Question 3)
* Filled the missing values with the mean value of each column in the Dataset 3. (Question 3)
* Calculated the number of missing values for each column in the Dataset 2. (Question 1 and 3)
* Filled the missing values with the mean value of each column in the Dataset 2. (Question 1 and 3)
* Combined the Dataset 2 and Dataset 3 and stored the value in the dataset named ‘Merged\_dataframe’. (Question 3)
* Checked missing values in the columns using the for-loop and is.null() function. (Question 2)

**Data Enrichment**

* Merged cancer\_count\_states dataframe and dataset 2 and stored the data in dataframe named merged\_states\_cost. (Question 1)
* Added the new column ‘Per Person out of pocket cost (in millions)’ by dividing ‘Per Person out of pocket cost (in millions)’ by ‘Total-out-of-pocket costs (in millions). (Question 1)
* Multiplied states column (containing each cancer type count) with Per Person out of pocket cost to get financial impact on each state due to different cancer types. (Question 1)

**Validation Rules and checks**

For All Business Questions

Rule 1: Check if any Missing or Null values present in the dataset.

**Purpose:** We need to find the missing and Null values so that we can replace the missing values with average using mean imputation method to get accurate results.

**Validation Rule Type:** This is a system rule as we have included code in the python code to check if there are any missing values.

**Specific Process and Tasks:** This validation rule takes place before we retrieve the data for each cancer type.

**Example:** Find if any missing values present.

**Code:**

for c in Dataset1\_index.columns:

miss = Dataset1\_index[c].isnull().sum()

if miss>0:

print("{} has {} missing value(s)".format(c,miss))

else:

print("{} has No missing value(s)".format(c))

For Question 2

**Rule 2:** Check and drop cancer type as All Cancer Sites Combined and All Sites (comparable to ICD-O-2)

**Purpose:** The focus is on different cancer types and not on All sites Combined data.

**Validation Rule Type:** This is a system rule as we have python code to check and drop the unrequited records (All Sites Combined and All Sites (comparable to ICD-O-2) ).

**Specific Process and Tasks:** This validation takes place at the start of the python code before we find any missing values present.

**Example:** Drops the unrequited records of All sites combined.

**Code:**

Dataset1.drop(Dataset1[Dataset1["CancerType"] == 'All Cancer Sites Combined'].index, inplace = True)

Dataset1.drop(Dataset1[Dataset1["CancerType"] == "All Sites (comparable to ICD-O-2)"].index, inplace = True)

**6. Analysis and results:**

**Business Question 1:**

What is the financial implication of cancer treatment across the various cancer states and for different cancer types/phases?

**Part 1:** Financial implication of cancer treatment across various cancer states.

**Steps:**

* To analyse this question, we considered datasets cancer count by states and cancer cost by phase and merged them.
* Dropped the unrequired columns and calculated per person out of pocket cost by dividing total out of pocket cost by total affected cancer people of USA.
* Then calculated the total cost of each individual state by multiplying per person out of pocket cost with total count of people affected in each state.

**Result:**

California is the most spending country on different cancer types as more no of people getting affected in the state and Nevada the least spending state on cancer.

**Graph:**

Chart, waterfall chart

Description automatically generated

We have also created heat map for the whole data for each individual state with respect to each cancer type and attached along with the documents submitted.

**Part 2:** Financial implications of cancer treatment for different cancer types/phases.

* We have considered dataset cancer cost phase wise and plotted bar graph of required columns like Initial, Continuing, End of life cancer death, End of life other cause of death.

**Result:**

|  |  |  |
| --- | --- | --- |
| Phases | Highest amount Spent on Cancer Type | Least amount spent on Cancer Type |
| Initial | Breast Cancer | Skin |
| Continuing | Breast Cancer | Mouth |
| End of life cancer death | Lung Cancer | Throat |
| End of life other cause of death | Male Reproductive Cancer | Throat |

**Graph**

Chart, bar chart

Description automatically generated

**Business Question 2:**

Which ethnic groups are most susceptible to cancer challenges?

**Steps:**

1. To answer this question, we considered demographic dataset. In the dataset we have five different race types. These include: 'All Races and Ethnicities' 'American Indian/Alaska Native’, ‘Asian/Pacific Islander' 'Black' 'Hispanic' 'White'.
2. Filtered the columns required for the analysis on how different cancer types affecting various ethnic groups and dropped unrequired records.
3. Later, subsetting the dataset according to race types and analyzed the type of cancer affecting more and less no of people in each individual race type.

**Results:**

Most and Least affected Races with all cancers combined in USA:

Most Affected Race– white (75%)

Least Affected Race– American Indian (1%)

Chart, sunburst chart

Description automatically generated

**Table:**

Table

Description automatically generated

**Business Question 3:**

What is the relationship between the government’s cancer research funding and patient’s out-of-pocket cost?

**Steps:**

1. To answer this business question, we have considered two datasets cancer cost by phase and government cancer research funding and merged them.
2. We found funding per dollar out of pocket cost for the year 2019 by taking the ratio of 2019 research funding and total out of pocket cost of individual cancer type.
3. From the result obtained we can conclude that government’s funding with respect to the total amount of patient’s money spent is highest for the mouth cancer and lowest for the blood cancer.
4. If government allocates research funding on basis of the financial impact, the focus on blood cancer should increase.

**Graph:**

Chart

Description automatically generated

**7. External material:**

1. The shift in risk factor profiles and disease classifications, advancements in cancer detection and treatment, and demographic shifts is leading to continuous evolution of cancer settings in the USA (Islami, et al., 2020).

2. While high cost of cancer treatment is critical challenge for the patients, the cost of hardship experienced by the patients may vary depending upon the age, ethnicity, education, income. There is no “one size fits all” cancer treatment cost, and many private insurance plans are required by the current law to limit annual patient spending (American Cancer Society, 2020).

3. The exorbitant cost of medicine which known to cause physical toxicity such as pain, nausea, fatigue, etc. have potential impact of mental health. Cost burden, often referred to as financial toxicity can be related to fiscal or economic ruins, with younger adults more prone to the catastrophe (Gubar, 2019).

4. Cancer and chronic obstructive pulmonary disease are considered more identifiable diseases in terms of biology, diagnosis, and pathogenesis relative to infectious diseases. It has driven significant investments from government, industries, and philanthropists into the cancer research (Trasta, 2018).

5. Governments have been focussed on supporting cancer programs. Europe’s Beating cancer plan aims to focus on prevention, early detection, diagnosis, and treatment; and improving quality of life for cancer patients (European Commission, 2021). Meanwhile, the US president Biden relaunched ‘cancer moon-shot’ in a bid to reduce cancer death rate by 50% in 25 years (McGinley, 2022).

6. The mortality rate for patients with hematologic cancers and solid tumours admitted to the intensive care unit is quite high. However, this limited survival is achieved at a very high cost. Thus, given the focus on financial implication, potential outcomes and the possibility of withdrawing life-supportive therapy could be discussed with the patient and family (Schapira, et al., 1993).

7. The rising burden of healthcare-related costs for the patients impacts the treatment. To save on the out-of-pocket-costs, a significant number of patients either took less the prescribed amount of medication or avoided taking the medication altogether (Zafar, et al., 2013).

8. The cancer cost can have outside the treatment cost such as that spent on problem-solving counselling. In such scenario, telephone counselling could be used as a cost-effective strategy to prevent the onset of depression and improved the use of more favourable coping behaviours (Downe-Wamboldt, et al., 2007).

9. The COVID-19 pandemic put the longer-term effects on grants and funding for cancer research on the radar, with investigators becoming increasingly concerned. The impact of the economic crisis has domino effect on the cancer research funding as the institutions and philanthropic organisations are affected by the economic crisis (Caruso, 2020).

10. President Biden's budget proposal for Fiscal Year 2023 includes unchanged spending for the NIH and a $200 million reduction in funding for the NCI. The decline in financing might jeopardize the existing infrastructure for biomedical research in the United States, as well as weaken present attempts to advance scientific understanding for cancer management and other key (basic and translational) investigations. The American Society of Clinical Oncology (ASCO) is concerned about President Joe Biden's proposed financing for the Advanced Research Projects Agency for Health (ARPA-H), which would come at the expense of funds for the US National Institutes of Health (NIH) and National Cancer Institute (NCI) (Das, 2022).

*Many researchers have focussed upon the evolution of cancer landscape in the USA and highlighted financial burden that different ethnic groups face. The mental pressure yet remains another critical factor, focussed by another research. The high mortality rate from cancer is regarded by several researchers as a driving force to high funding for the cancer research.*

**8. Additional data and analysis:**

The lack of time and resources, the project scope was limited to identifying the relationship between the government’s cancer research funding and patient’s out-of-pocket cost. Further work could be undertaken which would enhance project scope and its applicability

* Using the health insurance data to enhance the value and identify how well people from each ethnicities are financially prepared to bear the cancer cost
* Using data collected first-hand from several hospitals to identify the year-wise pattern in the cancer cases.
* Using the beds and equipment data from hospitals to identify how well the US health system is prepared to provide cancer treatment.

**9. Contribution of individual Member**

**Anam Siddiqui** – Finding Datasets for the project, worked on Business Question 1 including data wrangling and Analysis of graphs, PPT and Project Report.

**Venkata Asha Deepika Penagganti** -Finding Datasets for the project, worked on business Question 2 including data wrangling and Analysis of graphs, PPT and Project Report.

**Deepak Sri Ram Murthy** - Finding Datasets for the project, worked on business Question 3 including data wrangling and Analysis of graphs, PPT and Dashboard.

**References:**

* American Cancer Society, 2020. *The Costs of Cancer,* s.l.: s.n.
* Gubar, S., 2019. *The Financial Toxicity of Illness.* Available at: <https://www.nytimes.com/2019/02/21/well/live/the-financial-toxicity-of-illness.html>
* Caruso, C., 2020. COVID-19 Hits Cancer Research Funding. *Cancer Discovery,* 10(6), p. 756.
* Das, M., 2022. Biden's proposed investment in cancer research sparks concerns. *The LANCER,* 23(5), p. 576.
* Downe-Wamboldt, B. L. et al., 2007. The Effects and Expense of Augmenting Usual Cancer Clinic Care With Telephone Problem-Solving Counseling. *Cancer Nursing,* 30(6), pp. 441-453.
* Gubar, S., 2019. *The Financial Toxity of Illness.* Available at: https://www.nytimes.com/2019/02/21/well/live/the-financial-toxicity-of-illness.html
* Islami, F., Siegel, R. L. & Jemal, A., 2020. The changing landscape of cancer in the USA — opportunities for advancing prevention and treatment. *Nature Reviews clinical oncology,* Volume 17, pp. 631-649.
* Kandel, S., Heer, J., Plaisant, C., Kennedy, J., Van Ham, F., Riche, N. H., ... & Buono, P. (2011). Research directions in data wrangling: Visualizations and transformations for usable and credible data. Information Visualization, 10(4), 271-288.
* McGinley, L., 2022. *Biden relaunches ‘cancer moonshot,’ aiming to reduce death rate by 50 percent in 25 years.* Available at: https://www.washingtonpost.com/health/2022/02/02/biden-cancer-moonshot-renewal/
* Schapira, D. V., FRCPC, M., Studnicki, J. & Bradham, D. D., 1993. Intensive Care, Survival, and Expense of Treating Critically III Cancer Patients. Jama, 269(6), pp. 783-786.
* Trasta, A., 2018. Where does public funding for cancer research go. *EMBO Reports,* 19(19), p. 3.

**Appendix**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Cancer Types** | **Dataset 1** | **Dataset 2** | **Dataset 3** | **Dataset 4** |
| Brain/Head/ Neck | Brain and Other Nervous System |  | Brain & CNS | Brain and ONS |
| Cranial Nerves and Other Nervous System |  | Head and Neck Cancers |  |
| Brain |  |  |  |
| Breast Cancer | Female Breast | Breast | Breast Cancer | Breast\_Female\_In\_situ |
| Male and Female Breast in situ |  |  | Breast\_Femaie |
| Male and Female Breast |  |  |  |
| blood cancer | Myeloma | lymphoma | Leukemia | Leukemia |
| Lymphomas | Leukemia | Hodgkin Disease | NonHodgkin\_Lymphoma |
| Other Leukemias | Hodgkin | Multiple Myeloma |  |
| Acute Lymphocytic | Hodgkin | Non-Hodgkin Lymphoma |  |
| Chronic Myeloid | lymphoma |  |  |
| Chronic Lymphocytic |  |  |  |
| Hodgkin Lymphoma |  |  |  |
| Acute Myeloid |  |  |  |
| Non-Hodgkin Lymphoma |  |  |  |
| Leukemias |  |  |  |
| lung cancer | Pleura | Lung | Lung Cancer | Lung\_Bronchus |
| Respiratory System |  |  |  |
| Lung and Bronchus |  |  |  |
| Mesothelioma |  |  |  |
| Trachea Mediastinum and Other Respiratory Organs |  |  |  |
| throat | Hypopharynx | Thyroid |  | Thyroid |
| tonsil |  |  |  |
| Thyroid |  |  |  |
| Larynx |  |  |  |
| stomach\_D | Colon and Rectum |  | Liver Cancer | Pancreas |
| Colon excluding Rectum | Colorectal | Colorectal Cancer | Stomach |
| Digestive System |  | Stomach Cancer | Colon\_Rectum |
| Esophagus |  | Pancreatic Cancer | Esophagus |
| Gallbladder |  |  | Liver Bile Duct |
| Liver and Intrahepatic Bile Duct |  |  |  |
| Other Biliary |  |  |  |
| Other Endocrine including Thymus |  |  |  |
| Other Digestive Organs |  |  |  |
| Pancreas |  |  |  |
| Peritoneum Omentum and Mesentery |  |  |  |
| Endocrine system |  |  |  |
| Rectum and Rectosigmoid Junction |  |  |  |
| Retroperitoneum |  |  |  |
| Small Intestine |  |  |  |
| Stomach |  |  |  |
| mouth | Floor of Mouth | Oral cavity |  | Oral Cavity\_Pharynx |
| Gum and Other Mouth |  |  |  |
| Lip |  |  |  |
| Oral Cavity and Pharynx |  |  |  |
| Oropharynx |  |  |  |
| Other Oral Cavity and Pharynx |  |  |  |
| Salivary Gland |  |  |  |
| Tongue |  |  |  |
| eye | Eye and Orbit |  |  |  |
| Female Genitals | Other Female Genital Organs | Cervix | Ovarian Cancer | Cervix |
| Cervix | Uterine | Uterine Cancer | Ovary |
| Corpus | Ovary | Cervical Cancer | Uterus |
| Corpus and Uterus NOS | Corpus |  |  |
| Female Genital System |  |  |  |
| Ovary |  |  |  |
| Uterus NOS |  |  |  |
| Vagina |  |  |  |
| Vulva |  |  |  |
| Male Genitals and reproductive system | Male Genital System |  |  |  |
| Other Male Genital Organs |  |  |  |
| Penis |  |  |  |
| Prostate | Prostate | Prostate Cancer | Prostate |
| Testis |  |  |  |
| Urinary Bladder and all | Urinary Bladder | Bladder |  | Bladder |
| Other Urinary Organs |  |  |  |
| Urinary System |  |  |  |
| Ureter |  |  |  |
| skin | Kaposi Sarcoma | Melanoma | Melanoma | Skin |
| Melanomas of the Skin |  |  |  |
| Skin excluding Basal and Squamous |  |  |  |
| Other Non-Epithelial Skin |  |  |  |
| kidney | Kidney and Renal Pelvis | Kidney | NA | Kidney\_Renal\_Pelvic |
| nose | Nose Nasal Cavity and Middle Ear |  |  |  |
| Nasopharynx |  |  |  |