Research Open

Preventive Medicine, Epidemiology and Public Health

Volume 2 Issue 3

Research Article

Identifying Possible Risk Factors of Poorly Understood Cancers – The Case for Using Health Information Technology

Julie Jaddoo MSHI[1]*, **Lincoln R. Sheets** MD, PhD, FAMIA[2], **Chester Lee Schmaltz** PhD[3], **Jeannette Jackson-Thompson** MSPH, PhD[4] **and Eduardo J. Simoes** MD, MSc, MPH, DLSHTM[5]

[1]PhD. Candidate, Informatics, MU Institute for Data Science and Informatics, University of Missouri, Columbia, MO, 65211.

[2]Assistant Professor of Health Informatics, School of Medicine, University of Missouri, Columbia, MO, 65211.

[3]Missouri Cancer Registry and Research Center, University of Missouri at Columbia, 1020 Hospital Dr, Columbia, MO 65211.

[4]Co-Director, Health and Behavioral Risk Research Center, Director Emeritus, Missouri Cancer Registry and Research Center, Research Associate Professor, Health Management and Informatics, University of Missouri (MU) School of Medicine, Core Faculty, MU Institute for Data Science & Informatics, 1095 Virginia Avenue, Suite 128, Columbia, MO 65211.

[5]Chair, Dr. Stuart Wesbury Distinguished Professor in Health Management and Informatics and HMI Alumni Distinguished Professor, Department of Health Management & Informatics, Interim Director of the Center for Biomedical Informatics, Director of the Health and Behavioral Risk Research Center, University of Missouri, School of Medicine, CE707 CS&E Bldg. One Hospital Dr. Columbia, MO 65212.

Corresponding author: Julie Jaddoo MSHI, Informatics, MU Institute for Data Science and Informatics, University of Missouri, Columbia, MO, 65211; **Email:** jpj8bd@mail.missouri.edu

Received: June 8, 2021; **Accepted:** June 10, 2021; **Published:** June 22, 2021

Introduction

Worldwide, cancer is the second leading cause of death, with one of every six deaths caused by cancer [1]. There were 17 million new cases and 9.6 million cancer deaths worldwide in 2018, including approximately 1.7 million new U.S. cases and 600,000 U.S. cancer deaths [3]. The total financial cost of cancer in 2010 was estimated at 1.16 trillion U.S. dollars [1].

There have been significant reductions in cancer mortality, thanks to improved screening, early detection, and better treatment [1]. However, the worldwide incidence of cancer is expected to increase to 27.5 million per year by 2040 [28], a 62% increase from 2018. The U.S. expects an increase to over 1.9 million new cases per year by 2020, largely due to an aging Caucasian population and a growing African American population [5].

The World Health Organization states that "30-50% of all cancer cases are preventable. Prevention offers the most cost-effective long-term strategy for the control of cancer" [26]. Cancer can be prevented by reducing exposure to environmental risk factors, modifying lifestyle factors that are linked to cancers, and protecting against the effects of risk-factor exposures [26].

Tobacco is one of the most widely known and most modifiable risk factors for cancer and the process in determining this illustrates the value of the systematic study of cancer causes [15]. Lung cancer is the most common cancer in the world after skin cancer and the most deadly [24]. Before smoking became widespread, lung cancer was rare; however, as cigarette production and smoking increased, lung cancer became much more common. Smoking tobacco was found to be associated with lung cancer around the mid-20th century when a study showed that smokers were more likely to have cancer than non-smokers [15]. This relationship was confirmed by epidemiological and prospective studies, experiments, pathological observations, and chemical analyses [15]. Smoking was also found to be a risk factor for many other types of cancers and diseases, and tobacco is now understood to be associated with 33% of cancers and 22% of cancer-related deaths worldwide [28]. Cigarette smoking is associated with 80%-90% of lung cancer deaths in the U.S. [25]. Deaths caused by smoking cigarettes have an average latency of about 25 years; lung cancer deaths are expected to reach about 2 million per year during the 2020s or the 2030s [15].

As a result of the overwhelming evidence that smoking is a causal factor for cancer, there have been many anti-smoking initiatives. These efforts include preventing smoking initiation, helping smokers quit the habit, and reducing exposure to second-hand smoke [11]. Smoking cessation reduces cancer risk and can improve outcomes for cancer patients. Smoking cessation can reduce lung cancer risk by as much as 85% after 16 years of cessation compared to non-cessation [13]. Due to tobacco control measures that were implemented in the U.S. in the mid-1950s, about 32% (795,851) of the lung cancer deaths that would have occurred during 1975-2000 were prevented; the benefits of these measures will continue [11]. These huge reductions in deaths, suffering, and costs were possible because good epidemiological evidence uncovered the link between smoking and cancer.

Other cancer-prevention strategies that have grown out of accumulating epidemiological studies include reducing alcohol consumption [10], vaccinating against certain viruses [17], and

improving diet and exercise [4]. The International Agency for Research on Cancer (IARC) determined that alcohol was carcinogenic after reviewing studies that showed an association between alcohol consumption and certain cancers [8]. One study involving eight European countries estimated that for 2008, 3% of cases in women and 10% of cases in men were due to alcohol consumption [20]. A U.S. study determined that 3.2% - 3.7% (18,200 to 21,300) of all cancer deaths in 2009 were attributable to alcohol consumption [12].

Many viruses have been shown to cause or be associated with certain cancers [17]. Individuals and health care providers can take preventive steps such as vaccinations, follow-up treatment to minimize the risk of developing cancer, and screening to maximize chances of early detection of cancer [17]. And because obesity, diet, and sedentariness have proven to be risk factors that are related and modifiable, individuals can make lifestyle changes to reduce their cancer risk while gaining other health benefits [4].

There have been improvements in cancer survival rates due to improvements in cancer detection and treatment, but the progress made applies to relatively few cancers [16]. Also, this does not spare patients the ordeal, financial cost, and disability of cancer treatment. Screening guidelines are available for very few cancer types, so many cancers are detected at later stages and, therefore, have a lower survival rate [9]. In addition, incidence rates of some of these poorly understood cancer types are increasing. Cancer prevention is the least costly and most desirable approach to combat the expected increase in cancer incidence [26]. However, to achieve this we need epidemiological research that focuses on identifying risk factors for poorly understood cancer types.

A traditional epidemiological approach, such as the "Cancer Prevention Studies" (CPS), requires a large enough study group, long follow-up, and is costly [7]. Therefore, this approach is limiting, especially for poorly understood cancers, which tend to be rarer cancers. In addition, the research landscape has changed significantly. Information technology was one of the most significant technological developments of the twentieth century and has affected every aspect of our lives. It has made us very interconnected to people, activities, and the environment. Determining any effect of these connections is difficult due to the complexity and numerosity. Fortunately, these technological developments have also made significant advances that can be applied to health research. We have, are generating, and are capturing more data about many different aspects of our lives than ever before. We need to use current technology and data to overcome the limitations of the traditional epidemiological approach. We must develop reliable, efficient, and cost-effective research methods to identify possibly risk factors for poorly understood cancers.

Purpose

Our main objective in this study was to identify cancer types that represent a health burden, but for which environmental and lifestyle risk factors are poorly understood (i.e., without an established causal risk factor). We used a combination of inclusion and exclusion criteria to identify these cancers. The inclusion criteria were cancer types 1) without screening guidelines; 2) with low survival rates; and 3) with increasing incidence. Cancer screening aims to detect cancer before the individual becomes symptomatic, and early detection usually results in more successful treatment and greater survival [27]. Currently, only four types of cancer – breast, cervical, colorectal, and lung – have broadly accepted screening recommendations [16]. The cancers with low survival rates tend to be cancers that are more difficult to detect and to treat. An upward trending cancer indicates a growing concern that should be investigated to identify risk factors and reduce incidence. The primary exclusion criterion was cancer types with established causal risk factors. By default, cancer types with screening guidelines, without low survival rates, and/or without increasing incidence were excluded.

Secondarily, we propose a new methodological approach to study the etiology of these rare cancers that maximizes data utilization without the need for costly epidemiological studies, such as the "Cancer Prevention Studies". This design allows exploration of the relationships of selected cancer with various potential risk factors without the financial and feasibility barriers of traditional epidemiological designs.

Methods

We used data from the National Cancer Institute's (NCI's) Surveillance, Epidemiology, and End Results (SEER) Program. The SEER Program collects cancer incidence and survival data for every cancer case reported from population-based cancer registries covering approximately 34 percent of the U.S. population spanning 19 geographic areas. The program started in 1975 with nine registries and now collects data from twenty-one registries. Based on the broad coverage area, the data collected by the SEER Program is representative of the U.S. population [21].

SEER data are coded to ICD-O-3 and are grouped by major cancer site/histology [22]. The data have 102 groups in a hierarchical format, ranging from all sites to miscellaneous. SEER incidence data have both the rate (per 100,000 and age-adjusted to the 2000 U.S. population) and count. Survival data provide observed, expected, and relative rates. Incidence trend data show overall percent change, annual percent change (APC), and the rate for each year. For this study, we used incidence data for 2011 through 2016 and survival data for the five-year period 2011 to 2015.

Starting with all 102 groups of cancers from the SEER database, we compiled data for incidence, survival rate, and trend. We added data on whether the cancers had recommended screening guidelines. Our selection criteria for cancer groups representing a health burden were groupings that do not have recommended screening guidelines, had a positive APC over the 6-year period (2011 to 2016), and had a 5-year relative survival rate of less than 70%. From the cancers meeting all three criteria, we removed groupings that were poorly defined, groupings containing "Other" or "Not Otherwise Specified", and groupings that were too broadly defined (e.g.,"Female Genital System"). We removed "Liver and Intrahepatic Bile Duct" since it includes the sub-category "Liver" that did not meet the criteria, but the sub-category"Intrahepatic Bile Duct" is in the final list(Table 1).

Table 1: Cancers meeting inclusion criteria

		Age-adjusted Incidence Rate[1]	5-yr Relative Survival Rate	Rate/Trend (2011-2016 APC)
	Oral Cavity and Pharynx	11.05	63.13%	0.38
	Tongue	3.18	63.34%	1.77
	Gum and Other Mouth	1.55	59.00%	0.13
	Oropharynx	0.40	42.95%	2.62
Digestive System	Small Intestine	2.11	65.69%	0.99
	Appendix	0.85	69.13%	16.46
	Liver and Intrahepatic Bile Duct	7.82	17.12%	0.65
	Intrahepatic Bile Duct	0.84	7.83%	8.70
	Pancreas	12.23	7.67%	0.20
	Other Digestive Organs	0.60	9.37%	7.29
Respiratory System	Trachea, Mediastinum and Other Respiratory Organs	0.19	51.52%	0.74
	Soft Tissue including Heart	3.33	65.56%	0.62
	Female Genital System	26.41	68.98%	0.33
	Uterus, NOS	0.39	29.71%	1.50
	Other Female Genital Organs	0.55	56.75%	10.28
Male Genital System	Penis	0.38	67.52%	0.23
Urinary System	Other Urinary Organs	0.33	50.32%	5.98
Endocrine System	Other Endocrine including Thymus	0.75	64.33%	1.35
	Myeloma	6.35	46.73%	0.06
Leukemia	Acute Lymphocytic Leukemia	1.68	67.02%	0.57
	Chronic Myeloid Leukemia	1.75	63.73%	0.87
	Other oral cavity & pharynx - C148 (Overlapping lesion of lip, oral cavity)	0.08	50.77%	6.06
	Other digestive organs - C268 (Overlapping lesion of digestive system)	0.03	9.50%	9.67
	Other digestive organs-C269 (Malignant neoplasm of ill-defined sites)	0.56	8.72%	7.18
	Other digestive organs-C488 (Overlapping lesion of retroperitoneum and peritoneum)	0.01	51.17%	7.03

[1]Rate per 100,000 population, age-adjusted to the 2000 U.S. standard population

Shaded groupings were removed for reasons stated above

We implemented the primary exclusion criterion by reviewing the websites of the American Cancer Society (ACS) (Cancer.org) and the National Cancer Institute (NCI) (Cancer.gov) to identify risk factors, causal and non-causal, for each of the cancers initially selected based on epidemiological measures. None of the cancers that met the inclusion criteria had established causal factors, so none were excluded (Table 2).

Results

The main epidemiological measures for the 12 groupings of poorly understood cancers are presented in Table 1. The grouping of "Oral Cavity and Pharynx" is a major grouping with sub-groupings "Tongue" and "Oropharynx", and all three are in the final list. There are four groupings of the digestive system - "Small Intestine", "Appendix", "Intrahepatic Bile Duct", and "Pancreas". There are two leukemias – "Acute Lymphocytic Leukemia" and "Chronic Myeloid Leukemia". The other groupings are "Soft Tissue including Heart", "Penis", and "Myeloma".

The cancers included in the final list have varying statistics (Table 1). The age-adjusted incidence rates of poorly understood cancers in the final list range from 0.40 to 12.23 cases per year per 100,000 population. The incidence rates for all the 102 original groupings ranged from 0.00063 to 85.28 cases per year per 100,000 population. The final list of cancers has 5-year survival rates ranging from 7.67% for Pancreatic Cancer to 69.13% for Cancer of the Appendix. Cancer of the Appendix has the highest upward incidence trend (APC=16.46), followed by Intrahepatic Bile Duct (APC=8.70), Oropharynx (APC=2.62), and Tongue (APC=1.76). The other cancers have upward incidence trends of less than 1.00% APC. While there are some known risk factors for these cancers, there are no known causal risk factors.

Some of these cancers, such as soft tissue cancer including heart, have a more robust set of potential risk factors in the ACS classification of risk factors but a substantially larger number of potential risk factors classified by NCI. Others like "Acute Lymphocytic Leukemia" have a large number of potential risk factors in both ACS and NCI classifications. Tobacco, infections, radiation, and immunosuppressive medications were stated more as general risk factors for many of these poorly understood cancers. Some cancer-specific risk factors are viruses, diseases, syndromes, and poor nutrition, but the overall opinion is that very little is known about the causes of these cancers.

Table 2: Application of exclusion criterion using information from the ACS website

	What causes the cancer?	**Risk Factors**
Oral Cavity and Pharynx	*"Oral cavity and oropharyngeal cancers start in the mouth (including the tongue) or throat."* [2] *"Doctors and scientists cannot say for sure what causes each case of oral cavity or oropharyngeal cancer."* [3]	• Tobacco • Alcohol • Betel quid • Gutka • HPV • Ultraviolet (UV) light • Poor nutrition • Weakened immune system • Graft-versus-host disease (GVHD) • Genetic syndromes - (Fanconi anemia, Dyskeratosis congenita), • Lichen planus [4]
Tongue		
Oropharynx		
Small Intestine	*"The 4 major types of small intestine cancers are:* *Adenocarcinomas: …* *Carcinoid tumors: …see Gastrointestinal Carcinoid Tumors.* *Lymphomas: …see Non-Hodgkin Lymphoma.* *Sarcomas: …The most common sarcomas in the intestine are known as gastrointestinal stromal tumors (GISTs)."* [5] *"What Causes Small Intestine Cancer (Adenocarcinoma)?* *...not much is known about exactly what causes these cancers."* [6] *"What Causes Gastrointestinal Carcinoid Tumors?* *… Changes in many different genes are usually needed to cause carcinoid tumors."* [7] *"What Causes Non-Hodgkin Lymphoma?* *… but the cause of most lymphomas is not known."* [8] *"What Causes Soft Tissue Sarcomas?* *Scientists don't know exactly what causes most soft tissue sarcomas, but they have found some risk factors that can make a person more likely to develop these cancers…"* [9]	• African American • Smoking and alcohol use • Celiac disease • Colon cancer • Crohn's disease • Inherited syndromes: – Familial adenomatous polyposis (FAP) – Lynch syndrome (hereditary nonpolyposis colorectal cancer, or HNPCC) – Peutz-Jeghers syndrome (PJS) – MUTYH-associated polyposis – Cystic fibrosis (CF) [10]
Appendix	*"What Is a Gastrointestinal Carcinoid Tumor?* *Gastrointestinal carcinoid tumors are a type of cancer that forms in the lining of the gastrointestinal (GI) tract."* [11]	• Genetic syndromes: – Multiple endocrine neoplasia, type I – Neurofibromatosis type 1

[2] https://www.cancer.org/cancer/oral-cavity-and-oropharyngeal-cancer/about/what-is-oral-cavity-cancer.html

[3] https://www.cancer.org/cancer/oral-cavity-and-oropharyngeal-cancer/causes-risks-prevention/what-causes.html

[4] https://www.cancer.org/cancer/oral-cavity-and-oropharyngeal-cancer/causes-risks-prevention/risk-factors.html

[5] https://www.cancer.org/cancer/small-intestine-cancer/about/what-is-small-intestine-cancer.html

[6] https://www.cancer.org/cancer/small-intestine-cancer/causes-risks-prevention/what-causes.html

[7] https://www.cancer.org/cancer/gastrointestinal-carcinoid-tumor/causes-risks-prevention/what-causes.html

[8] https://www.cancer.org/cancer/non-hodgkin-lymphoma/causes-risks-prevention/what-causes.html

[9] https://www.cancer.org/cancer/soft-tissue-sarcoma/causes-risks-prevention/what-causes.html

[10] https://www.cancer.org/cancer/small-intestine-cancer/causes-risks-prevention/risk-factors.html

	What causes the cancer?	**Risk Factors**
	"*What Causes Gastrointestinal Carcinoid Tumors?* *Changes in 4 tumor suppressor genes are responsible for many inherited cases of carcinoid tumors…Most carcinoid tumors are caused by sporadic changes (mutations) in oncogenes or tumor suppressor genes..."* 7	• Other genetic syndromes (tuberous sclerosis complex, von Hippel Lindau disease and familial small intestinal neuroendocrine tumor) • African American • Family history of any type of cancer [12]
Intrahepatic Bile Duct	*"…All of these ducts within the liver are called intrahepatic bile ducts. "* [13] *"We don't know the exact cause of most bile duct cancers…."* [14]	• Certain diseases of the liver or bile ducts • Primary sclerosing cholangitis (PSC), • Bile duct stones • Choledochal cyst disease • Liver fluke infections • Cirrhosis • Hepatitis B virus or hepatitis C virus • Inflammatory bowel disease • Hispanic Americans • Obesity • Being overweight • Non-alcoholic fatty liver disease • Exposure to Thorotrast • Family history of bile duct cancer • Diabetes • Alcohol [15]
Pancreas	*"We don't know what causes pancreatic cancer. But we do know many of the risk factors for this cancer…"* [16]	• Tobacco use • Being overweight • Diabetes • Chronic pancreatitis • Workplace exposure to certain chemicals • African Americans • Family history • Inherited genetic syndromes – Hereditary breast and ovarian cancer syndrome – Hereditary breast cancer – Familial atypical multiple mole melanoma (FAMMM) syndrome – Familial pancreatitis – Lynch syndrome – Peutz-Jeghers syndrome [17]
Soft Tissue including Heart	*"What Is a Soft Tissue Sarcoma?* *… Soft tissue sarcomas can develop in soft tissues like fat, muscle, nerves, fibrous tissues, blood vessels, or deep skin tissues…*[18]	• Radiation given to treat other cancers • Family cancer syndromes • Neurofibromatosis

[11] https://www.cancer.org/cancer/gastrointestinal-carcinoid-tumor/about/what-is-gastrointestinal-carcinoid.html

[12] https://www.cancer.org/cancer/gastrointestinal-carcinoid-tumor/causes-risks-prevention/risk-factors.html

[13] https://www.cancer.org/cancer/bile-duct-cancer/about/what-is-bile-duct-cancer.html

[14] https://www.cancer.org/cancer/bile-duct-cancer/causes-risks-prevention/what-causes.html

[15] https://www.cancer.org/cancer/bile-duct-cancer/causes-risks-prevention/risk-factors.html

[16] https://www.cancer.org/cancer/pancreatic-cancer/causes-risks-prevention/what-causes.html

[17] https://www.cancer.org/cancer/pancreatic-cancer/causes-risks-prevention/risk-factors.html

	What causes the cancer?	Risk Factors
	"What Causes Soft Tissue Sarcomas? *Scientists don't know exactly what causes most soft tissue sarcomas… Researchers still don't know why most soft tissue sarcomas develop in people who have no apparent risk factors."* [19]	• Gardner syndrome • Li-Fraumeni syndrome • Retinoblastoma • Werner syndrome • Gorlin syndrome • Tuberous sclerosis • Damaged lymph system [29]
Penis	*"The exact cause of most penile cancers is not known. But scientists have found that it's linked with a number of other conditions."* [21]	• Human papillomavirus (HPV) infection • Not being circumcised • Phimosis • Smegma • Smoking and other tobacco use • UV light treatment of psoriasis • AIDS [22]
Myeloma	*"Scientists still do not know exactly what causes most cases of multiple myeloma..."* [23]	• African Americans • Family history of myeloma • Obesity • monoclonal gammopathy of undetermined significance (MGUS) • solitary plasmacytoma [24]
Acute Lymphocytic Leukemia	*"Some people with acute lymphocytic leukemia (ALL) have one or more of the known risk factors, but many do not. Even when a person has one or more risk factors, it can be very hard to know if it actually caused the leukemia."* [25]	• Radiation exposure • Certain chemical exposures • Certain viral infections – human T-cell lymphoma/leukemia virus-1 (HTLV-1) – Epstein-Barr virus (EBV) • Certain genetic syndromes – Down syndrome – Klinefelter syndrome – Fanconi anemia – Bloom syndrome – Ataxia-telangiectasia – Neurofibromatosis – Li-Fraumeni syndrome • Race/ethnicity - more common in whites than African Americans • Having an identical twin with ALL[26]
Chronic Myeloid Leukemia	*"There are no proven risk factors for CML."* [27]	• Radiation exposure [28]

[18] https://www.cancer.org/cancer/soft-tissue-sarcoma/about/soft-tissue-sarcoma.html

[19] https://www.cancer.org/cancer/soft-tissue-sarcoma/causes-risks-prevention/what-causes.html

[20] https://www.cancer.org/cancer/soft-tissue-sarcoma/causes-risks-prevention/risk-factors.html

[21] https://www.cancer.org/cancer/penile-cancer/causes-risks-prevention/what-causes.html

[22] https://www.cancer.org/cancer/penile-cancer/causes-risks-prevention/risk-factors.html

[23] https://www.cancer.org/cancer/multiple-myeloma/causes-risks-prevention/what-causes.html

[24] https://www.cancer.org/cancer/multiple-myeloma/causes-risks-prevention/risk-factors.html

[25] https://www.cancer.org/cancer/acute-lymphocytic-leukemia/causes-risks-prevention/what-causes.html

[26] https://www.cancer.org/cancer/acute-lymphocytic-leukemia/causes-risks-prevention/risk-factors.html

[27] https://www.cancer.org/cancer/chronic-myeloid-leukemia/causes-risks-prevention/what-causes.html

[28] https://www.cancer.org/cancer/chronic-myeloid-leukemia/causes-risks-prevention/risk-factors.html

Discussion

Our study identified 12 cancers as poorly understood because a causal risk factor has not yet been identified. These 12 cancers, though not among the cancers with highest health burden in the United States and worldwide — a ranking mostly reserved for lung, colorectal, prostate, breast and cervical cancer — represent a moderate health burden if their count is taken in aggregate. Therefore, preventing these cancers and improving population health will be possible if we identify their causal risk factors (exposures).

Identifying associations between exposures and cancer can be done through cohort or case-control studies [18, 19]. A cohort study can provide strong evidence of causal associations. ACS' Cancer Prevention Studies (CPS-I, CPS-II, CPS-3) are large scale, prospective, cohort studies [7]. These studies required large scale recruitment of subjects, survey completion by the subjects, and long follow-up periods. Initially, the aim was to determine the relationship between smoking and mortality from diseases. CPS-3 aims to determine the causes and protectants of cancer by looking at lifestyle, exposure, biology, and environment [7]. The results of CPS-I and II have identified significant factors that affect health and diseases progression. This information identified the areas to focus resources in order to combat diseases. These studies have provided valuable information that has helped to improve health; however, they require significant manpower and follow-up time. CPS-I had 68,000 volunteers in 25 states, a cohort of almost one million participants (men and women) and ran from 1959 through 1972. The CPS-II cohort was established in 1982 through recruitment of 1.2 million men and women in 50 states by 77,000 volunteers. CPS-3, with over 30,000 volunteers, enrolled over 304,000 participants from across the United States and Puerto Rico from 2006 through 2013. CPS-II and CPS-3 are still ongoing [7, 14].

Two common features of the poorly understood cancers identified in this study are they are rare and have low survival rates (< 70%). These features limit the choices of study designs; however, the low survival rates indicate their severity and justify the need for studying these cancers. A cohort study of these cancers would be challenging. The first challenge for studying these cancers is finding a large enough number of eligible, willing subjects to form a reliable study group. Second, based on the long follow-up required, the expected loss of study subjects might make any results obtained unreliable. Third, these challenges would increase the costs of studying these cancers. In addition, any study results obtained might not be useful due to low power. This justifies a case-control approach to investigating these cancers. A case-control approach selects subjects based on the outcome (e.g., presence/absence of one of these rare cancers) and measures the prior exposure event retrospectively. Compared to a cohort study, this approach would require fewer subjects, less time, and less funding.

We intend to use a case-control design and informatic-derived analytical techniques to identify potential risk factors for these poorly understood and somewhat rare cancers. Our aim is to combine various secondary datasets that traditionally have not been analyzed together for the purpose of performing exploratory data analyses and subsequent generation of hypotheses about unknown risk factors for these cancers. We believe this approach is novel due to the use of only secondary data and informatic-derived imputation methods and analyses. Using logistic regression to impute missing attributes in the dataset will produce a more complete dataset with sociodemographic, behavioral, and environmental attributes. The application of geographic information system (GIS) analyses, association mining, cluster analyses, and contrast mining to this dataset could reveal valid relationships.

The term "poorly understood" is often used to describe many different aspects of diseases, ranging from etiology to outcomes. However, criteria for assigning the term to any aspect of disease have not been established. There is research on individual cancer types and sub-types that are termed "poorly understood", but the publications do not provide objective justification for assigning the term. We believe this approach is also relatively novel due to the use of set measures for selecting poorly understood cancers.

We aim to include in the study multiple types of factors (environmental, behavioral, sociodemographic, clinical) against multiple types of poorly understood cancers as in the Environmental Public Health Tracking Network (EPHTN) of Wisconsin conducted by Hanrahan et al, 2004. The Wisconsin EPHTN was established to generate and test hypotheses for environmental causes of childhood cancers [6]. However, by using more types of factors, this proposed study can also examine interactions of the factors against cancer type(s) in all age groups.

Using data from the Missouri Cancer Registry, University of Missouri (MU) Healthcare, U.S. Census, Behavioral Risk Factor Surveillance System, and the Environmental Protection Agency, we will create datasets that have data on cancer incidence, health care records, demographics, behavioral risk factors, and environmental factors.

We will start by identifying the records of new cases of the cancers of interest in the Missouri Cancer Registry (i.e., incidence cancer cases). We will then identify if these patients also exist in the MU Healthcare electronic health records (EHR). For these matches, we will link and merge the records for the individuals, including cancer diagnosis and all available sociodemographic attributes, in both datasets as well as medications and procedure information from the MU healthcare system dataset. From the EHR dataset, we will also select un-matched patients (non-cancer patients or patients without a cancer of interest) that have similar demographic characteristics to the matched subjects. The selected patients will form control pools from which we will select our controls.

The data for the individual subjects and controls will lack values for many sociodemographic, behavioral, and environmental attributes of interest in this research but will have geographic identifiers that will be used to impute such values. Using demographic, behavioral, and environmental attributes from individuals in similar sociodemographic categories as the study cancer cases but from other datasets and the geographical identifiers available in both datasets, an

imputation process will be used to ascribe values of these attributes to the study cancer cases. These geographic identifiers will be used to ascribe extrapolated and imputed values of demographic and environmental attributes to the cancer cases. We will use logistic regression for this imputation analysis. The resulting datasets will be significantly enriched for hypothesis generation analyses of the associations between cancer and potential risk factors that otherwise could not have been studied. This type of analytical approach is only hypothesis generating because of the many possible biases originating from the extrapolation and imputation processes.

We will also analyze the enriched dataset using geographic information system (GIS) analyses, association mining, cluster analyses, contrast mining, and statistical analyses. GIS analyses can determine the proximity to regulated environmental activities. Association mining will identify associations between cancer(s) and the attributes within the dataset. Cluster analysis will be used to group cancers based on similarities and might identify different cancers that have one or more common factors. Contrast mining will be used to identify differences among different cancers and cancer groups by comparing the factors associated with each. Statistical analyses will be used to model relationships within the dataset and determine the odds ratios and 95% confidence intervals for associations within the dataset.

The results of this design and analyses approach are expected to benefit prevention and control strategies for these rare cancers. Currently, the rarity of these cancers and the prohibitive costs of established epidemiological studies of cancer etiology make it infeasible for research-funding institutions to support studies of these cancers. The findings of this study and the accompanying big-data driven case-control study should help guide research agencies' decisions to fund further investigation into specific cancers and risk factors relationships they postulate.

If progress is not made regarding cancer prevention and control, the medical cost of cancer in the U.S. could rise to $207 billion by 2020 [2]. The increasing burden of cancer will have an even greater impact on low- and middle-income countries. These countries already bear the burden of 70% of cancer deaths, are at a financial disadvantage due to the significant financial cost of cancer, and lack the resources to detect and adequately treat cancer [1, 23].

Conclusion

This study is a first step toward our overall research goal to identify possible causal risk factors for poorly understood cancers. This first step systematically identified the cancer types that are severe and trending up but for which the risk factors are poorly understood. A major limitation is the low incidence for these cancers. This low power makes it highly infeasible to study these specific cancers using cohort studies. We propose a novel approach to generate hypotheses for the associations of these poorly understood cancers with multiple risk factors. This approach circumvents historical limitations of cost and feasibility of implementation that culminated with these cancers being currently classified as poorly understood.

We propose to use current health information technology and data to develop methods to overcome the limitations of the traditional epidemiological approach and identify possible risk factors for these poorly understood cancers. A big-data driven approach to identifying risk factors maximizes the size of the study group, is more cost effective than the traditional approach, eliminates the problem of lost study subjects, and reduces the time to obtain results. We believe it is the least costly and most feasible approach to identify risk factors for poorly understood cancers.

References

1. Cancer (2012, September 8) World Health Organization. https://www.who.int/news-room/fact-sheets/detail/cancer.

2. Cancer costs projected to reach at least $158 billion in 2020 (2011, January 11) National Institutes of Health (NIH). https://www.nih.gov/news-events/news-releases/cancer-costs-projected-reach-least-158-billion-2020

3. Cancer Facts & Figures 2018 | American Cancer Society. (2018) Cancer Facts & Figures 2018. https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2018.html

4. Cancer Prevention Overview (PDQ®)–Patient Version—National Cancer Institute (nciglobal,ncienterprise) (2009, June 22) [PdqCancerInfoSummary]. https://www.cancer.gov/about-cancer/causes-prevention/patient-prevention-overview-pdq

5. CDC - Expected New Cancer Cases and Deaths in 2020 (2019, January 31). https://www.cdc.gov/cancer/dcpc/research/articles/cancer_2020.htm

6. Hanrahan, L. P., Anderson, H. A., Busby, B., Bekkedal, M., Sieger, T., Stephenson, L., Knobeloch, L., Werner, M., Imm, P., & Olson, J (2004) Wisconsin's Environmental Public Health Tracking Network: Information Systems Design for Childhood Cancer Surveillance. Environmental Health Perspectives 112(14), 1434–1439. https://doi.org/10.1289/ehp.7150 [crossref]

7. History of the Cancer Prevention Studies. (n.d.). American Cancer Society. Retrieved October 7, 2019, from https://www.cancer.org/research/we-conduct-cancer-research/behavioral-and-epidemiology-research-group/history-cancer-prevention-study.html

8. Humans, I. W. G. on the E. of C. R. to. (2010) Summary of Data Reported. In Alcohol Consumption and Ethyl Carbamate. International Agency for Research on Cancer. https://www.ncbi.nlm.nih.gov/books/NBK326555/ [crossref]

9. Late-stage cancer detection in the USA is costing lives—The Lancet (2010, December 4). The Lancet. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(10)62195-2/fulltext

10. Limit Alcohol Consumption. (n.d.). American Institute for Cancer Research. Retrieved October 7, 2019, from https://www.aicr.org/cancer-prevention/recommendations/limit-alcohol-consumption/

11. Moolgavkar, S. H., Holford, T. R., Levy, D. T., Kong, C. Y., Foy, M., Clarke, L., Jeon, J., Hazelton, W. D., Meza, R., Schultz, F., McCarthy, W., Boer, R., Gorlova, O., Gazelle, G. S., Kimmel, M., McMahon, P. M., de Koning, H. J., & Feuer, E. J (2012) Impact of Reduced Tobacco Smoking on Lung Cancer Mortality in the United States During 1975–2000. JNCI Journal of the National Cancer Institute 104(7), 541–548. https://doi.org/10.1093/jnci/djs136

12. Nelson, D. E., Jarman, D. W., Rehm, J., Greenfield, T. K., Rey, G., Kerr, W. C., Miller, P., Shield, K. D., Ye, Y., & Naimi, T. S (2013) Alcohol-Attributable Cancer Deaths and Years of Potential Life Lost in the United States. American Journal of Public Health 103(4), 641–648. https://doi.org/10.2105/AJPH.2012.301199 [crossref]

13. Novello, A. C. (1990) Surgeon General's report on the health benefits of smoking cessation. Public Health Reports 105(6), 545–548. [crossref]

14. Patel, A. V., Jacobs, E. J., Dudas, D. M., Briggs, P. J., Lichtman, C. J., Bain, E. B., Stevens, V. L., McCullough, M. L., Teras, L. R., Campbell, P. T., Gaudet, M. M., Kirkland, E. G., Rittase, M. H., Joiner, N., Diver, W. R., Hildebrand, J. S., Yaw, N. C., & Gapstur, S. M (2017) The American Cancer Society's Cancer Prevention Study 3 (CPS-3): Recruitment, study design, and baseline characteristics. Cancer 123(11), 2014–2024. https://doi.org/10.1002/cncr.30561. [crossref]

15. Proctor, R. N (2012) The history of the discovery of the cigarette–lung cancer link: Evidentiary traditions, corporate denial, global toll. Tobacco Control 21(2), 87–91. https://doi.org/10.1136/tobaccocontrol-2011-050338.

16. Recommendations (2017, September) U. S. Preventive Services Taskforce. https://www.uspreventiveservicestaskforce.org/uspstf/topic_search_results?topic_status=P&category%5B%5D=15&type%5B%5D=5&searchterm=cancer+screening

17. Risk Factors: Infectious Agents - National Cancer Institute (nciglobal,ncienterprise) (2015, April 29) [CgvArticle]. https://www.cancer.gov/about-cancer/causes-prevention/risk/infectious-agents.

18. Rothman, Kenneth J (1998a) Case-Control Studies / Kenneth J. Rothman, Sander Greenland. In Modern epidemiology / Kenneth J. Rothman, Sander Greenland ; with 15 contributors (2nd ed., pp. 93–114). Lippincott-Raven.

19. Rothman, Kenneth J (1998b) Cohort Studies / Kenneth J. Rothman, Sander Greenland. In Modern epidemiology / Kenneth J. Rothman, Sander Greenland ; with 15 contributors (2nd ed., pp. 79–92). Lippincott-Raven.

20. Schütze, M., Boeing, H., Pischon, T., Rehm, J., Kehoe, T., Gmel, G., Olsen, A., Tjønneland, A. M., Dahm, C. C., Overvad, K., Clavel-Chapelon, F., Boutron-Ruault, M.-C., Trichopoulou, A., Benetou, V., Zylis, D., Kaaks, R., Rohrmann, S., Palli, D., Berrino, F., … Bergmann, M. M (2011). Alcohol attributable burden of incidence of cancer in eight European countries based on results from prospective cohort study. BMJ 342. https://doi.org/10.1136/bmj.d1584.

21. SEER_Overview.pdf. (n.d.). Retrieved June 7, 2020, from https://seer.cancer.gov/about/factsheets/SEER_Overview.pdf.

22. Site Recode—SEER Recodes. (n.d.). Surveillance, Epidemiology, and End Results Program. Retrieved October 7, 2019, from https://seer.cancer.gov/siterecode/icdo3_dwhoheme/index.html.

23. Torre, L. A., Siegel, R. L., Ward, E. M., & Jemal, A. (2016) Global Cancer Incidence and Mortality Rates and Trends—An Update. Cancer Epidemiology and Prevention Biomarkers 25(1), 16–27. https://doi.org/10.1158/1055-9965.EPI-15-0578 [crossref]

24. Ultraviolet (UV) radiation and skin cancer (2017, October 16). https://www.who.int/news-room/q-a-detail/ultraviolet-(uv)-radiation-and-skin-cancer. [crossref]

25. What Are the Risk Factors for Lung Cancer? | CDC (2019, November 18). https://www.cdc.gov/cancer/lung/basic_info/risk_factors.htm

26. WHO | Cancer prevention (2019) WHO; World Health Organization. http://www.who.int/cancer/prevention/en/

27. Why is early diagnosis important? (2015, April 2) Cancer Research UK. https://www.cancerresearchuk.org/about-cancer/cancer-symptoms/why-is-early-diagnosis-important

28. Worldwide cancer statistics (2019). Cancer Research UK. https://www.cancerresearchuk.org/health-professional/cancer-statistics/worldwide-cancer

Citation:

Jaddoo J, Sheets LR, Schmaltz CL, et. al (2021) Identifying Possible Risk Factors of Poorly Understood Cancers – The Case for Using Health Information Technology. *Prev Med Epid Public Heal* Volume 2(3): 1-9.