Spatial and Temporal Analysis of Cancer Prevalence and Mortality Rates in Canada: Aging and Cohort Effects.

Introduction
Cancer is the number one killer in Canada. According to Statistics Canada, an average of 21 people are diagnosed with some type of cancer every hour, and nine of these people eventually die from this cancer (Canadian Cancer Society, 2012). Although considerable research exists on the contextual determinants of cancer, little work has been done to investigate whether the importance of context varies with duration of exposure and age at exposure. In other words, no known studies identify whether the danger of living in a cancer ‘hotspot’ is a function of both how long and the age at which a person lived in that hotspot. Does living in a hotspot have a linear effect on the propensity to contract a certain form of cancer? Is it more or less dangerous to live in a hotspot as a younger person or as an older adult? Does this effect differ by socioeconomic characteristics? Given that roughly 1/3 of all people move in a five-year period (Haan, 2005), individuals encounter multiple residential environments throughout their lives, and the primary purpose of this project is to identify the effect that both past and present residences have on their risk of contracting prostate cancer, breast cancer, and colorectal cancer, three of the most prevalent forms of cancer in Canada.1

As shown by Krewski, Jerrett, et al. (2009), linking mortality information to a census database is powerful for studying various health and health-related issues. For example, it facilitates better understanding the impact of long-term exposure to air pollution on human health. Though only in pilot testing, the 1991 CCC has produced many important findings, see Wilkins, Tjepkema, Mustard and Choinière (2008), Peters and Tjepkema (2010), Ng (2011), and the references therein. The recently approved extended 1991-2011 linkage combines the 1991 census of population, Canadian Mortality Database (CMDB), Canadian Cancer Database (CCDB) and the Tax Summary Files (TSF). In addition to the extensive information from the 1991 census data, the tax summary files provide updated location information represented by postal code. The extended linkage contains more comprehensive information and thus will be a much more powerful tool for cancer and mortality-related research that focuses on the role that location has in influencing cancer outcomes.

The above studies are based on various pre-defined geographic regions, such as province, Census Metropolitan Area, or health region, without considering spatial correlation and spatial heterogeneity. In this paper we use the 1991 Canadian Census Cohort: mortality & cancer follow-up (1991 CCC) to: (i) identify incidence hotspots (clusters); (ii) identify critical age exposures to these hotspots; (iii) calculate cancer occurrence rates (the Cancer Registry only contains those that have been diagnosed, thus requiring age-sex specific population counts to compute incidence rates) according to both current and previous places of residence.

What Place of Residence Information Adds to Previous Research
Clearly, exposure to environmental risk factors are important for understanding the prevalence of prostate cancer, breast cancer, colorectal cancer, or nearly any other form of cancer. Several researchers show that environmental rather than genetic factors predominate in the causes of cancers; depending on the cancer type, only 2-10% of cancers are the result of a mutation in, or the operation of, a particular gene (see Davis, Donovan, and Herberman (2007)). This suggests

---

1 Source: http://www.cancer.ca/Canada-ide/About_cancer/Cancer_statistics/Stats_at_a_glance/General_cancer_stats.aspx. The fourth most prevalent form of cancer is lung cancer, but we are choosing not to model lung cancer since it is driven so heavily by smoking, acritical piece of information that is not on the file.
that the overwhelming majority of cancer-causing factors are either linked to individual behaviour or environmental context, or interactions between genes and these other contextual factors. As only one example of the importance of context, Milewski and Liu (2009) report that breast cancer rates in Fredericton, Saint John, and Moncton are significantly higher than the rates reported for their respective health regions. A weakness of their study – and of cancer research more generally – is that they cannot measure the amount of time that patients actually lived in these cities, making it difficult to truly evaluate environmental effects (it is theoretically possible that every cancer patient in the three cities moved from a hotspot in the previous year). As such, it is difficult to measure environmental effects without knowing the amount of time an individual actually lived in that environment.

This proposal is unique in that it will model individual characteristics (as contained in the 1991 census), as well as the effect of current and past environments. The 1991 CCC includes the tax file summary which provides accurate location of each individual and enables precise cluster detection in cancer incidence and mortality every year from 1991 to the year of diagnosis of cancer. Having information on place of residence from 1990-2007 enables us to identify whether age at exposure to a cancer hotspot subsequently increases the risk of contracting that cancer. It will be the first known study to identify whether not just exposure matters, but also age at exposure to a cancer hotspot. In the section below, we show how hotspots will be identified.

**Cancer Incidence Cluster Detection**

Spatial cluster detection is a typical spatial partition problem, and SatScan (Kulldorff, 2009) is a state-of-the-art tool for spatial data cluster detection. SatScan has been widely used in a variety of areas, such as health data mining, infectious disease, and crime surveillance. We identify departures from randomness with SatScan, designating areal units with a significantly higher prevalence rate a hotspot.

Age is one of the most obvious factors affecting risk for almost all diseases, especially cancers. It is well-known that two populations may show differences in incidence or mortality for a given disease as a consequence of differences in age distributions alone. This is commonly addressed by working with age-standardized rates, and in this research we detect spatial clusters in age standardized incidence rates and mortality rates by using postal code, subsequently grouping postal codes to identify a hotspot at a higher level of geography. For this purpose, we first extract the average latitude and longitude of each 6-digital postal code and extract the physical distances between postal codes, and then use the software SatScan to detect the potential clusters for age standardized new cancer incidence rates and age standardized mortality rate within a predefined geographic area across a five-year period. Then, we identify an individual’s exposure level to that region, using both age of exposure and duration as indicators of exposure.

More practically, the procedure runs as follows: A scanning window is placed at different coordinates with radii that vary from zero to some set upper limit, and for each location and each size of window, if it’s inside statistics is significantly higher than outside, a cluster is reported. The detected clusters are represented by sets of postal codes. Similarly, we will detect the potential spatial clusters that have lower age standardized incidence rate and mortality rate. Once the high rate and lower rate clusters are detected, we then investigate the neighborhood and environmental characteristics to identify the most significant risk factors.

**Cancer Type Specific Spatial-temporal Clustering Analysis**

In practice, each cancer type may have its own environmental causes. Since the 1991 CCC contains specific cancer type, we can identify hotspots particular to prostate cancer, breast
cancer, and colorectal cancer using the spatial-temporal cluster detection function of Satscan (Vieira, et al. 2008). For each of the three types of cancer (prostate cancer, breast cancer, and colorectal cancer), we use Satscan to detect its spatial temporal clusters. The idea is to explore the time-space domain of the data using cylindrical windows with a circular or ellipse geographic base and with the height of the cylinder corresponding to some interval in time. For each location, each size of space-time cylinder if it’s inside statistic of interest is significantly higher/lower than outside, a spatial-temporal cluster is reported. This scanning can detect spots that have higher/lower incidence rate during certain period time or abrupt change in time. Once clusters are identified, the next step is to investigate their neighborhood and environmental characteristics to identify its risk factors.

Because of mobility, a new cancer incidence reported in one place may have been contracted elsewhere years ago. By containing postal code information on every individual age 25 and older from 1990-2006, the 1991 CCC is a powerful tool to track the mobility of each individual. From the linkage, we can extract residential history for each individual.

**Methodology**

**Sample**
We transform the 1991 CCC into a person-period dataset, where each record denotes a year of observation (there is a maximum of 18 observations per individual). Individuals contribute person period records as long they do not have the cancer of interest in the particular model. Upon receiving a diagnosis for a particular form of cancer, an individual will contribute no additional records to the dataset, for they have experienced the outcome of interest. We treat 1991 data as tombstone data, which we attach to each person period record. We treat place of residence (hotspot/non hotspot), age, and the age*hotspot interaction as time-varying, as described in greater detail below.

**Analytical Technique**

This paper identifies the importance of exposure to cancer hotspots as a risk for being diagnosed with prostate cancer, breast cancer, or colorectal cancer, where exposure will be defined as both duration of exposure and age at exposure. This makes it a prime candidate for discrete-time event history analysis techniques. Possible to estimate with STATA (by using the XT suite of commands), the method focuses on estimating the probability of occurrence of an event at a given point in time (which in this case will be one of the three forms of cancer). Event history analysis models are preferable to traditional regression approaches because they allow for the occurrence of time-varying characteristics (such as age, residence in a hotspot, and duration of residence in that hotspot) while accounting for within-individual correlations in error terms.

Change in individual propensities for one of the three cancer diagnoses will be measured and analyzed for individual between 1990 and 2007, according to the following equation:

\[
C_{it}^* = \alpha + \beta_1 X_{it} + \beta_2 Age_{it} + \beta_3 Year_{it} + \beta_4 Hotspot_{it} + \beta_5 HotspotYear_{it} \\
+ \beta_6 Age \times Hotspot + \beta_7 Age \times HotspotYear + \epsilon_{it}
\]
Where $i$ denotes the individual and $t$ indicates the year of observation. The dependent variable is a dichotomous indicator, set to 1 if an individual has one of the three cancers, 0 otherwise (in each of the three separate models, the probability of a different cancer will be estimated). $X_i$ is a vector of socio-demographic characteristics taken from the 1991 census (which are all be treated as fixed), $\text{Age}_i$ denotes individual $i$’s age in continuous years at time $t$, $\text{Year}_i$ reflects the year of observation, $\text{Hotspot}_{it}$ is a dichotomous variable set to 1 if a person is currently living in a hotspot, 0 otherwise. $\text{HotspotYears}$ is a count variable which measures the number of years that an individual has lived in a cancer hotspot up to time $t$, and $\text{age}^*\text{hotspot}$ is an interaction between age and current residence in a hotspot, and measures the age-specific effect of living in a cancer hotspot. $\text{Age}^*\text{HotspotYears}$ measures the differential effect of duration in a hotspot across age groups.

References


Kulldorff, M. (2009), SatScan user manual, (http://www.satscan.org/)


