NAME: Katherine Freeland
THESIS ADVISOR: Michael Sonksen
DEPARTMENT: Mathematics and Statistics

PROPOSAL

An Exploration of Cancer Reoccurrence and Mortality Trends Using the SEER Data Set

INTRODUCTION

Using the SEER data set, a massive publicly available cancer database, we wish to explore several questions concerning cancer reoccurrence and mortality trends. The data set captures a wide scope of demographic information, as well as treatment and diagnostic information that we believe will enhance our understanding of treatment effects and mortality prediction.

DATA DESCRIPTION

Beginning in 1973, the Surveillance Epidemiology and End Results (SEER) Program at the National Cancer Institute began to collect data on cancer cases throughout the country. As collection continued, the data set expanded to include many more geographic areas and demographic information. The current data set includes information collected between 1973 and 2009. Presently, the data set has registries from 18 geographic areas with records from over 2 million patients. Registries associated with the SEER program collect data on primary tumor site, tumor morphology and stage at diagnosis, first course of treatment, and follow-up for vital status for patients with many types of cancer. The SEER data is unique in its comprehensive collection of information about the stage of cancer at diagnosis as well as patient survival. SEER registries currently cover approximately 28% of the US population.

RESEARCH PROBLEM

Using this data set, we hope to explore possible links between radiation and cancer reoccurrence. Scientists at Lawrence Berkeley National Laboratory recently completed a study showing exposure to radiation prematurely aged cells, causing them to stop dividing, and thus creating a more hospitable environment to pre-cancerous cells [3]. In recent months, new research has also been done to investigate chemotherapy resistance. A 2012 article, from *Nature Medicine,* shows connections between DNA-damaging cancer treatments (i.e. chemotherapy) and the production of a specific protein that enables cancerous cell growth [2]. We want to see if the SEER data, through

examination of patient records, gives epidemiological evidence to support these claims.

We also will examine cancer specific mortality trends in the SEER data. We suspect more impoverished regions, with less access to quality healthcare, have greater cancer mortality. The 2011 America's Health Rankings, sponsored by the United Health Foundation, ranked states using indicators such as lack of health insurance, childhood poverty, premature death, etc. to determine a general 'healthfulness' score [1]. We plan to compare these rankings to cancer mortality trends observed within the SEER data set. We are also interested in developing a model to predict cancer mortality. A goal of physicians is to predict how long a patient will survive after initial diagnosis. We will develop a model, using variables captured by the SEER data set, to better predict survival time.

METHODS

To examine the relationship between radiation and cancer reoccurrence, as well as the relationship between chemotherapy and drug resistance, we will use propensity scores. In a perfect world, we would test these claims by conducting a randomized experiment. Unfortunately, we do not have the ability to conduct such an experiment. We can only conduct an observational study, using the SEER data set, which lacks the requisite randomization to establish causation in the classic Rubin Causal Model sense. Propensity scores are a methodology which allow for the estimation of treatment effects when treatment assignment is not random.

To assess cancer mortality trends across states we will use logistic regression with random effects. Logistic regression is commonly used to predict mortality in the public health literature and random effects will allow us to model the correlation between different states (which we believe exists). Lastly, the Cox proportional hazards model will be used to predict years until mortality in cancer patients. Because this data includes current patients, we will not always observe when (and if) the patient has died of cancer. In other words, we have censored data. Cox proportional hazards models are especially useful when the data is censored. The model relates the time of an event, in this case death, to a number of explanatory variables (covariates). The model assesses the impact of these covariates in predicting survival time. After fitting this model, we will be able to predict, for a given patient's information, an estimated time until death from cancer.

REFERENCES

1.) "America's Health Rankings." United Health Foundation, Dec. 2011. Web. 18 Aug. 2012.<http://www.americashealthrankings.org/SiteFiles/Reports/AHR%202011edition.pdf>.

2.) Sun, Yu, Judith Campisi, Celestia Higano, Tomasz M. Beer, Peggy Porter, Ilsa Coleman, Lawrence True, and Peter S. Nelson. "Treatment-induced Damage to the Tumor Microenvironment Promotes Prostate Cancer Therapy Resistance through WNT16B." *Nature Medicine* (2012).

3.) Yarris, Lynn. "Study Raises New Concerns About Radiation and Breast Cancer." *Berkeley Lab News Center*. <http://newscenter.lbl.gov/news-releases/2010/05/13/new-concerns-about-radiation-and-breast-cancer/>.

DATA CITATION

Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Incidence - SEER 9 Regs Research Data, Nov 2011 Sub (1973-2009) <Katrina/Rita Population Adjustment> - Linked To County Attributes - Total U.S., 1969-2010 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released April 2012, based on the November 2011 submission.