A Statistical Approach to Identify Environmental and Demographic Causes of Cancer Incidences Across US Counties

Approximately 80% of all cancer incidences are classified as sporadic and are primarily caused by external factors. Preliminary findings illustrate wide geographical variances in lung cancer occurrence rates, strongly suggesting that sporadic cancer incidence is correlated with factors related to the local demography and environment. This study focuses on identifying such factors by examining their correlations with cancer incidences (lung, colorectal, pancreatic, and all cancers).

Two types of data were analyzed: 69 demographic factors and 274 chemical pollutants. For demographic data, only features with |Pearson's linear correlation coefficient| > 0.3 were used, and they were normalized to enable comparison. A linear regressor was built to predict cancer incidences, and features with $p < 0.05$ were chosen as statistically significant (10.1 features on average across cancers). The coefficients were compared to rank factors based on their significance. The approach was validated qualitatively with several key features that are known causes (eg: age) and quantitatively confirmed through the high predictive power of the linear regression model (10.74% error) compared to a model that simply predicted the average rate of cancer incidences (23.74% error). While some of the features identified may not be direct causes of cancer, they are equally important (ie: poverty does not cause cancer, but resulting unsanitary conditions may).

The chemical emissions data is extremely sparse since each county may only have a few of many possible industries. Chemical levels were batched into 0's, signifying no contamination, and 1's, signifying a level higher than 3 times the average. For each chemical, if the difference between the average cancer incidences with and without contamination was greater than two standard deviations of the incidences with no contamination, the chemical was counted as having statistical significance — in practice, the majority had a difference greater than three standard deviations. On average, 10.1 chemicals were marked as statistically significant. Some of these are known carcinogens, while others are not and should potentially be reinvestigated. Even if certain chemicals marked as statistically significant do not directly cause cancer, they may be produced along with truly harmful chemicals not in the dataset. The chemical data analysis was qualitatively validated for lung cancer through remarkable visual similarity between a map of counties with high lung cancer rates and a map of counties producing chemicals marked as statistically significant for lung cancer.

Analyzing data by county is a novel and powerful tool in disease prevention as it offers a means of identifying sources for increased incidence rates. This type of analysis grants each county the ability to pinpoint and address its specific problems (eg: creating stricter sanitation legislation or restricting emissions of benzidine). Hence, analysis of big data by locality can lead to a major step forward in cancer prevention.