Applications of Causal Inference Concepts and Machine Learning Methods to Investigate Cancer Clusters

John L. Maddox[1] & Ashley E. Rice[2]

[1]University of Tennessee, Knoxville, TN 37996, [2]Oak Ridge National Laboratory, Oak Ridge, TN 37830

Contact: lmaddox5@utk.edu, riceae@ornl.gov

Introduction

Identifying the cause of significant localized increases in populational cancer incidence, or cancer clusters, remains a vital public health issue. A ratio between observed and expected incidence of community cancer called the Standardized Incidence Ratio (SIR) is used to confirm the community-reported high cancer incidence as a cancer cluster. Only then, are possible environmental causes analyzed, which restricts SIR to considering regions of interest independently of exposure and fails to determine the causal role of most environmental factors [1]. Using the improved causal SIR (cSIR), we calculate the ratio between the observed cancer incidence in an exposed region and a covariate-matched unexposed region [2]. By integrating exposure into the diagnostic ratio, causal links between exposure and increased cancer incidence are made statistically sound. Improving the results of cancer cluster studies necessitates the implementation of cSIR.

Problem Statement

We used the cSIR to reevaluate the conclusions of an identified brain and lung cancer cluster (Fig.1) from 1997-2003 in which investigation found no causal link between the increased cancer incidence and local dioxin exposure.

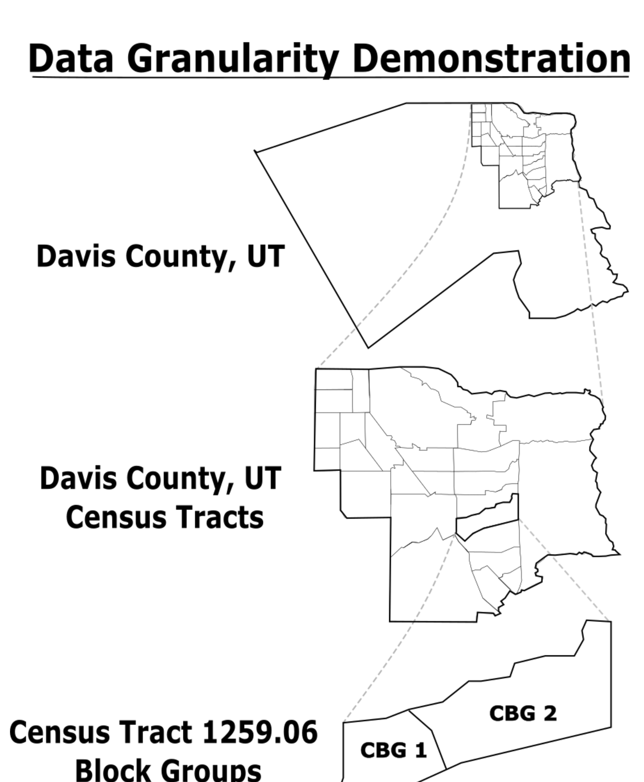

Figure 1: Block group-level and county-level data were transformed into census tract-level granularity.

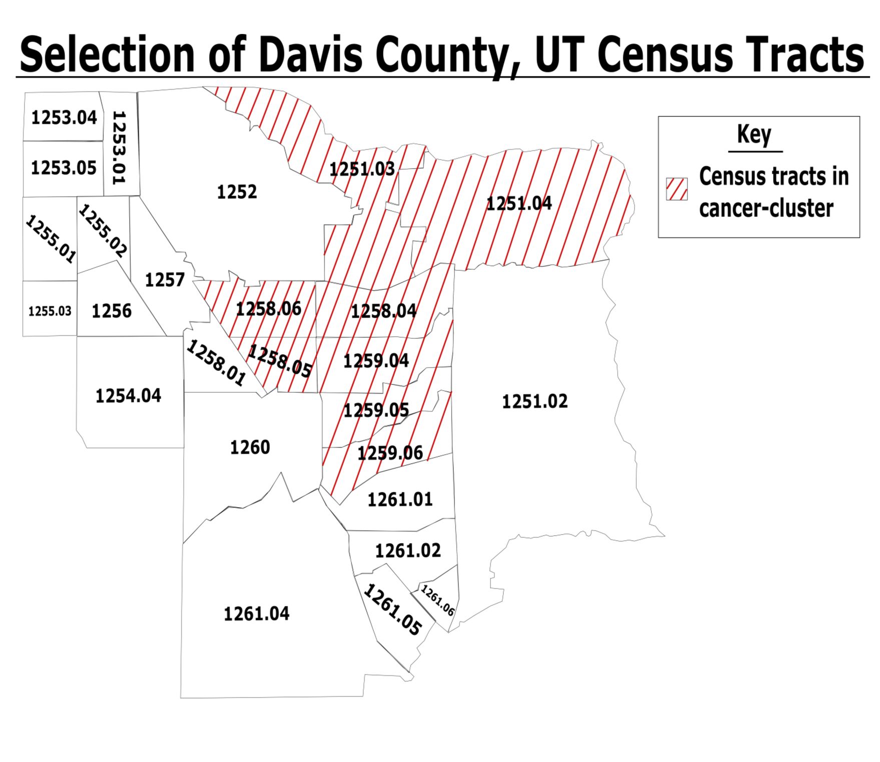

Figure 2: Selection of the eight Davis County, Utah census tracts containing both the confirmed brain and lung cancer cluster we are reinvestigating.

Materials and Methods

Data Name	Number of Records
SEER Cancer Incidence	1,085 counties
Census Data	65,444 Census Tracts
TRI Dioxin Exposure	1,572 Census Tracts
NY Cancer Incidence	4,603 Census Tracts
PLACES Health Data	58,668 Census Tracts

Table 1: Public data gathered to calculate the cSIR with their respective sizes and granularity.

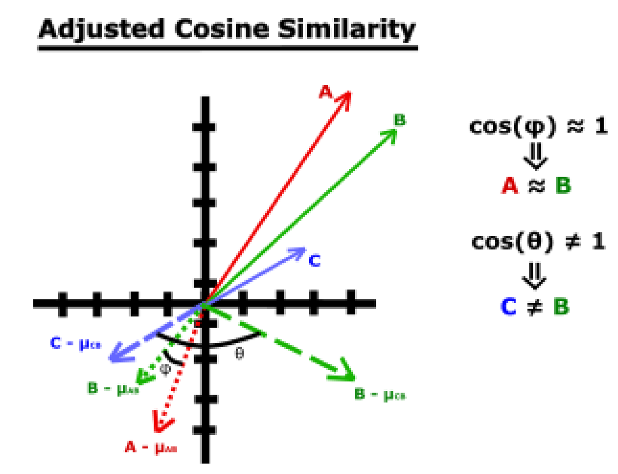

Figure 3: Adjusted cosine similarity quantifies how alike two vectors are in angle and magnitude.

We compiled census, health, superfund site, and Toxic Release Inventory (TRI) dioxin exposure data for all US census tracts (Table 1). After removing dioxin-exposed census tracts, we used adjusted cosine similarity matching (Fig. 3) with a threshold of 0.85 to find the counterfactual group, or unexposed census tracts that are similar in socioeconomic, health, and demographic covariates. These covariates included percentage impoverished, education level, race, smoking status, career industry and more. The cSIR for brain and lung cancer incidence was then calculated in each exposed census-tract (Fig. 4).

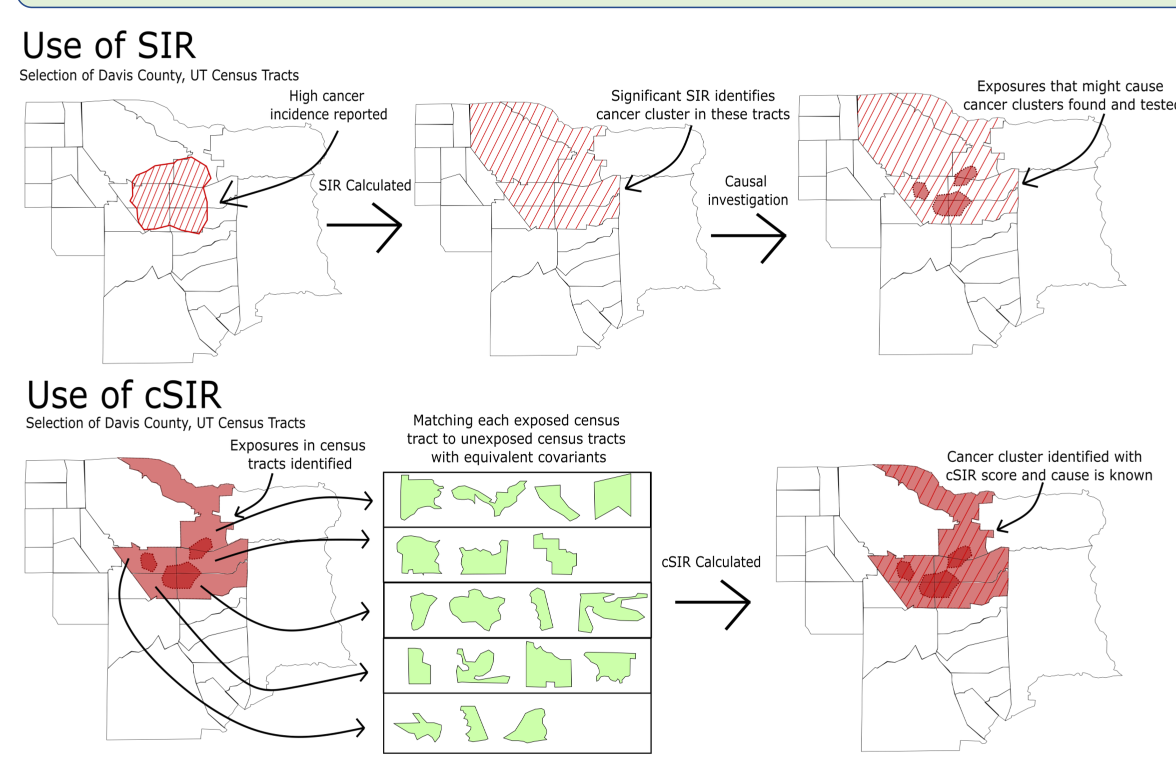

Figure 4: Differences in SIR and cSIR implementation. SIR finds cancer clusters before exposure, while cSIR finds exposure before the cancer cluster.

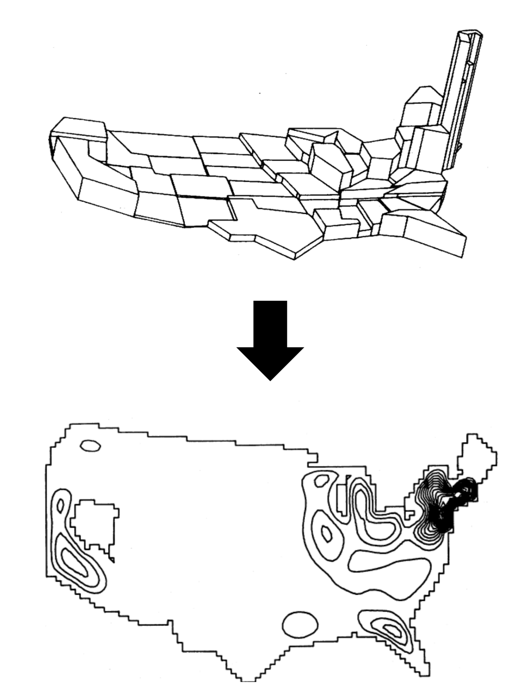

Figure 5: Pycnophylactic areal interpolation smooths out large granularity data into a density map [3]. Credit: (Tobler, 1979)

We extracted cancer incidence in the unexposed census tracts from the Surveillance, Epidemiology, and End Results (SEER) program, but SEER cancer incidence data is reported at the county-level. We addressed this data granularity issue (Fig. 1) with multiple methods: machine learning, areal interpolation (Fig. 5), and multiple imputation schemes. Gradient boosting regressor models outperformed all other methods (Table 2), so they were used to predict brain and lung cancer incidence. These models were trained on the compiled data and observed New York census tract cancer incidence to predict lung and brain cancer incidence in all US census tracts. Predictions were adjusted by a factor of $\frac{\textit{SEER county incidence}}{\textit{predicted county incidence}}$ to reduce model error.

Results

Imputation Method	MSE (Brain, Lung)
Weighted Area Interpolation	(30484.73, 2729274.31)
Pycnophylactic Interpolation	(15466.53, 1418138.34)
Multiple Imputation by Chained Equations (MICE)	(387.55, 23405.58)
Multiple Imputation using LightGBM	(360.58, 28102.18)
Gradient Boosting Regressor Model	(**272.92**, **19430.55**)

Table 2: Mean square error for each imputation schema.

On average, the gradient boosting regressor models used to impute cancer incidence had an R^2 value of 0.25 for brain cancer and 0.65 for lung cancer incidence. Only our lung cancer model performed adequately in predicting cancer incidence according to the SEER data. In our cSIR evaluation, we found that seven out of eight census tracts had a cSIR value above 1.0 for brain cancer incidence and zero out of eight had a cSIR value above 1.0 for lung cancer incidence. **While the causal link must be confirmed by future research, the cSIR scores did support dioxin's potential causal role for the brain cancer cluster (Table 3).**

Census-Tract	cSIR (Brain, Lung)	Census-Tract	cSIR (Brain, Lung)
1251.03	(**1.64**, 0.42)	1258.06	(**1.65**, 0.54)
1251.04	(0.64, 0.20)	1259.04	(**1.10**, 0.43)
1258.04	(**1.20**, 0.39)	1259.05	(**1.73**, 0.50)
1258.05	(**1.29**, 0.27)	1259.06	(**1.13**, 0.37)

Table 3: cSIR scores for each Davis County, UT census-tracts.

Conclusion

It is imperative that possible environmental causes behind cancer clusters are thoroughly investigated. Future work can improve the imputation scheme through combinations of interpolation methods or use of Knowledge Guided Machine Learning. Cancer cluster investigations should produce valuable information that determines the safety of extended living in exposed communities. To ensure cancer cluster investigations produce statistically valid results, widespread application of the cSIR is vital.

References:
1. Goodman, M., Naiman, J. S., Goodman, D., & LaKind, J. S. (2012). Cancer clusters in the USA: what do the last twenty years of state and federal investigations tell us?. *Critical reviews in toxicology*, *42*(6), 474–490. https://doi.org/10.3109/10408444.2012.675315
2. Nethery, R. C., Yang, Y., Brown, A. J., & Dominici, F. (2020). A causal inference framework for cancer cluster investigations using publicly available data. *Journal of the Royal Statistical Society. Series A, (Statistics in Society)*, *183*(3), 1253–1272. https://doi.org/10.1111/rssa.12567
3. Tobler, Waldo. (1979). Smooth Pycnophylactic Interpolation for Geographic Regions. *Journal of the American Statistical Association*. 74. 519-30. 10.1080/01621459.1979.10481647.

This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internships program.