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

John Maddox (University of Tennessee, Knoxville, TN 37659)
Ashley Rice (Oak Ridge National Laboratory, Oak Ridge, TN 37830)

An enduring public health goal is that of discovering environmental exposures that may explain cancer clusters. The Standardized Incidence Ratio (SIR), a ratio of observed and expected cancer incidence within a designated population, is commonly used to identify these localized regions of increased cancer incidence. However, the SIR is statistically flawed, because it assigns causality to suspected environmental factors within the area of interest independent of exposure. A similar but improved cluster-identifying tool, called the causal SIR (cSIR), has been proposed as a solution to the SIR-related problem. The cSIR accounts for exposure to chemicals of concern by establishing community exposure *before* identifying the cancer cluster. Thus, subsequent causal links between greater cancer incidence and exposures are statistically valid. Our research involved using public data to search for causal relationships between dioxin exposure and a lung- and brain-cancer cluster in Davis County, Utah. We implemented this search by calculating the cSIR metric for exposed census tracts. Before evaluating cancer incidence, we performed cosine-similarity matching on socioeconomic, demographic, and health-confounding variables to ensure similarity between populations in exposed and non-exposed census tracts. Then, we tested multiple methods to resolve granularity differences between a larger county-level cancer incidence and the smaller census tract-level covariate dataset including areal interpolation schemes, multiple-imputation schemes, and machine learning. Gradient boosting regressor machine learning models performed the best and were used to calculate the cSIR scores. Our analysis of cancer incidence between dioxin-exposed and the matched non-exposed census tracts using cSIR supported/did not support the hypothesis of a causal relationship between brain/lung/brain and lung cancer and dioxin exposure. However, the possibility of a causal link between dioxin exposure and increased cancer incidence needs to be confirmed. Overall, community-reported cancer clusters are relatively common, and if an elevated incidence is identified mathematically, future investigations should use cSIR to identify environmental causes.

Applications of causal inference concepts and machine learning methods to investigate cancer clusters

John Maddox[1] & Ashley E. Rice[2]
[1] *University of Tennessee, Knoxville, TN 37659*
[2] *Oak Ridge National Laboratory, Oak Ridge, TN 37830*

Abstract. An enduring public health goal is that of discovering environmental exposures that may explain cancer clusters. The Standardized Incidence Ratio (SIR), a ratio of observed and expected cancer incidence within a designated population, is commonly used to identify these localized regions of increased cancer incidence. However, the SIR is statistically flawed, because it assigns causality to suspected environmental factors within the area of interest independent of exposure. A similar but improved cluster-identifying tool, called the causal SIR (cSIR), has been proposed as a solution to the SIR-related problem. The cSIR accounts for exposure to chemicals of concern by establishing community exposure *before* identifying the cancer cluster. Thus, subsequent causal links between greater cancer incidence and exposures are statistically valid. Our research involved using public data to search for causal relationships between dioxin exposure and a lung- and brain-cancer cluster in Davis County, Utah. We implemented this search by calculating the cSIR metric for exposed census tracts. Before evaluating cancer incidence, we performed cosine-similarity matching on socioeconomic, demographic, and health-confounding variables to ensure similarity between populations in exposed and non-exposed census tracts. Then, we tested multiple methods to resolve granularity differences between a larger county-level cancer incidence and the smaller census tract-level covariate dataset including areal interpolation schemes, multiple-imputation schemes, and machine learning. Gradient boosting regressor machine learning models performed the best and were used to calculate the cSIR scores. Our analysis of cancer incidence between dioxin-exposed and the matched non-exposed census tracts using cSIR supported/did not support the hypothesis of a causal relationship between brain/lung/brain and lung cancer and dioxin exposure. However, the possibility of a causal link between dioxin exposure and increased cancer incidence needs to be confirmed. Overall, community-reported cancer clusters are relatively common, and if an elevated incidence is identified mathematically, future investigations should use cSIR to identify environmental causes.

Keywords: Causal inference, cancer cluster, cSIR

I. INTRODUCTION

Determining causal relationships between a variable and outcome is a well-researched field, particularly in healthcare. From identifying thalidomide as the cause for birth defects to discovering new infectious diseases such as the prion Kuru, causal investigations continue to make advances in the fields of toxicology and epidemiology.[1] However, causal relationships are notoriously difficult to prove, so generally the results are evaluated considering epidemiologist Austin Bradford Hill's list of criteria.[2] Hill's criteria includes 1) strength of association between the exposure and outcome, 2) consistency of the outcome in varying populations and situations, 3) specific outcome from exposure, 4) temporal association between exposure and outcome, 5) dose-response relationship between an exposure and an outcome, 6) biological plausibility of the causal relationship, 7) coherence of current data with the supposed causal relationship, and 8) experimental support.[2] Lack of any specific criterion, aside from (4), does not necessarily negate causality as a preponderance of evidence is sufficient. However, with regards to (4), it is imperative that the exposure is present prior to the observed outcome. In general, satisfying more criteria strengthens the case for a causal relationship is present.

Given the nature of cancer as a disease, causal investigations into temporally and geographically localized increases in cancer incidence, or cancer clusters, are made even more complicated. The latency period between exposures to carcinogens and cancer development tends to span 10 to 30 years for most adult cancers, thus identifying a specific cause is an arduous task. [3] Shifting populational demographics, individual habits, and varying family health histories over this extended latency period also makes recording accurate data and accounting for all possible confounding influences much more difficult.[3] Unlike the identification of transmission routes, origins, and spread of communicable diseases, which exploits the relatively short latency between infection and presentation of disease, identification of causes and spread of cancer proves more challenging.

Currently, formal cancer cluster investigations are initiated following considerable community reports of a suspected cancer cluster.[4,5] A ratio between the observed and expected incidence of community cancer called the Standardized Incidence Ratio (SIR) is then calculated to confirm the community-reported high cancer incidence as a cancer cluster.[4,6] The SIR reports information about the observed incidence of cancer compared to the expected incidence of cancer within an area of a community-reported cancer cluster. [6] Statistically significant SIR values greater than 1.0 indicate that there is an increased observed cancer incidence, and values less than 1.0 indicate that there is a decreased observed cancer incidence. An SIR value of 1.0 means that the observed cancer incidence is equivalent to the expected cancer incidence. Once a statistically significant increase in cancer incidence is established, community concerns regarding possible local chemical, radiological, agricultural, or waste-disposal causes are reviewed under multivariate linear regression analyses to find which exposure, if any, likely led to

the increased cancer incidence.[7,8] Then, potential confounders such as socioeconomic and demographic features are accounted for before relative risks are estimated and associations between covariates and cancer incidence are tested. [6,7]

While this methodology is the current standard, the use of SIR and subsequent evaluation of environmental causes is methodologically flawed.[3,4] These investigations are likely to suffer from the "Texas Sharpshooter fallacy." [4] Much like a marksman defining his target around the largest cluster of hits after firing, the observed presence of high cancer incidence is used to define the potentially afflicted region that is then tested for high cancer incidence. This increases the likelihood of a false positive cancer cluster. Additionally, proposing no sources of exposure before a cancer cluster is confirmed restricts SIR to considering cancer cluster regions of interest independently of exposure.[3] Because of these issues, cancer cluster studies often fail to determine the causal role of environmental exposures with only three out of 428 total investigations in the past twenty years establishing some causal link.[4]

To resolve these issues prevalent in SIR, we can use an improved metric called the causal SIR (cSIR), which is the ratio between the observed cancer incidence in an exposed region and a covariate-matched unexposed region.[4] By searching for exposure rather than cancer incidence, cancer cluster investigations using the cSIR do not succumb to the Texas Sharpshooter Fallacy. Furthermore, integrating exposure into the diagnostic ratio makes causal links between exposure and increased cancer incidence statistically valid. An important note is that cSIR necessitates the use of the stable unit treatment value assumption (SUTVA) to produce significant results. SUTVA requires that there is only one form of exposure and that the exposure status of one region cannot influence the health outcomes of other regions.[4]

In this paper, we use the cSIR metric to reevaluate a 1993-2003 brain and lung cancer cluster in eight Utah census tracts previously identified with SIR.[8] However, the proposed chemical exposure to dioxin, a persistent organic pollutant released as a byproduct from the local Wasatch Energy Systems incineration plant, was not statistically validated as a cause.[8] No explicit methodological details were provided, but dioxin's causal role was left inconclusive.[8]

To calculate the cSIR, we compiled multiple datasets. To derive confounding variable distributions, we compiled 2000 census data and CDC Population-Level Analysis and Community Estimates (PLACES) model-based estimates of small area health measures which yielded a dataset with 58 socioeconomic, demographic and health measure features for each census tract. Toxic Release Inventory data and Superfund Site data from the Environmental Protection Agency were also compiled to quantify dioxin exposure at the census-tract level. Finally, the New York state cancer registry data for learning cancer incidence relationships at the census tract level, and Surveillance, Epidemiology, and End Results (SEER) for county-level cancer incidence data.

However, because the SEER data is reported with a county granularity, we were required to disaggregate this data and produce census-tract level estimates of cancer incidence. We tested multiple methods of disaggregation/imputation including area interpolation schemes, multiple imputation, and machine learning models to see which best estimated the census-tract level cancer. We found that a gradient-boosted regressor model trained on the New York census tract cancer incidence data performed the best and it was used to predict cancer incidence for census tracts beyond New York, under the constraint that census tracts within counties must sum to the total number of cases reported in SEER for each county. We then implemented a matching procedure to compare the cancer incidence rates of exposed and non-exposed census tracts that are most similar based on their confounder distribution, followed by calculation of cSIR for each census tract. Based on calculated cSIR values, there is an indication of causality between dioxin and this cancer cluster, but additional work is needed to rigorously evaluate these findings against Hill's criteria.

II. MATERIALS AND METHODS

A. Data collection

Calculating the cSIR for the eight dioxin-exposed Utah census tracts required the construction of a counterfactual group of non-exposed census tracts that had closely matched socioeconomic and demographic factors used to represent confounding variables. This counterfactual group acts as an estimated control group, intended to display the cancer incidence when no dioxin exposure was present. The purpose is to account for many differences between census tracts as possible such that the likelihood of an effect other than the exposure of dioxin on cancer incidence is minimized. To create the confounder distributions, we pulled data from the 2000 United States census data for 64,914 census tracts in the United States, since this is the census data most relevant to the Utah cancer cluster investigation from 1993-2003.[6] The socioeconomic and demographic factors collected from the 2000 census included 30 features consistent with those used in Nethery et al. [4] for a cancer cluster investigation in Endicott, NY and containing information such as race, age, income, education status, and others. Accounting for each factor at a granular level such as census tract reduces possible confounding influences and improves the quality of the subsequent matching.

We then appended the 2020 CDC PLACES of health measures to our dataset containing confounding factors for each United States census tract. However, census tract naming changes between the 2000 and 2020 census caused us to lose 13,374 census tracts. To mitigate this loss, we found census tract relationship files which relate census tract naming changes from 2020 to 2010 and 2010 to 2000. Using these relationship files, we related the PLACES health measures data from 2020 census tracts to 2000 census tracts. With this change, we preserved 58,293 census tracts with sociodemographic, economic, and health confounding factors on which to match.

Finally, we collected Toxic Release Inventory and Superfund Site data for dioxin in every census tract between 1982 and 2020. During matching, it is essential to ensure that the census tracts being matched to the eight census tracts of interest in Utah do not have dioxin exposure. Using R-package tigris, the geographic coordinates of the facilities were converted first to a census block, and then to a census tract, which were considered exposed if the census tract contained a census block with exposure. We found that 2,181 census tracts were exposed to dioxin and removed them from the possibility of being matched. Once the matches were made, we collected the SEER program lung and brain cancer incidence data from 2000 to 2005 in the matched census tracts. Unlike the census-tract level socioeconomic and dioxin exposure data, the SEER cancer incidence data are reported at a coarser granularity, the county level. The years 2000 to 2005 were chosen as this time frame minimized the time elapsed since the reported cancer cluster investigation period, while also maximizing the number states reporting county level cancer incidence.

B. Matching exposed and non-exposed census tracts

We matched our eight dioxin-exposed census tracts with groups of nonexposed census tracts that shared the most similar confounder distributions. To match census tracts, we used an adjusted cosine similarity metric. Other metrics, like the Euclidean metric and the Mahalanobis metric, that quantify similarity using distance can exacerbate negligible differences between vectors when working in spaces with many dimensions because distance tends to scale with dimensionality of the space. Cosine similarity negates this exacerbation by only looking at the angle between two vectors. Treating each census tract as an n-dimensional vector from the origin, traditional cosine similarity scores are simply the cosine of the angle between two vectors. Then, cosine similarity scores of 1, 0, and -1 implies the vectors are parallel, perpendicular, and facing opposite of each other, respectively.

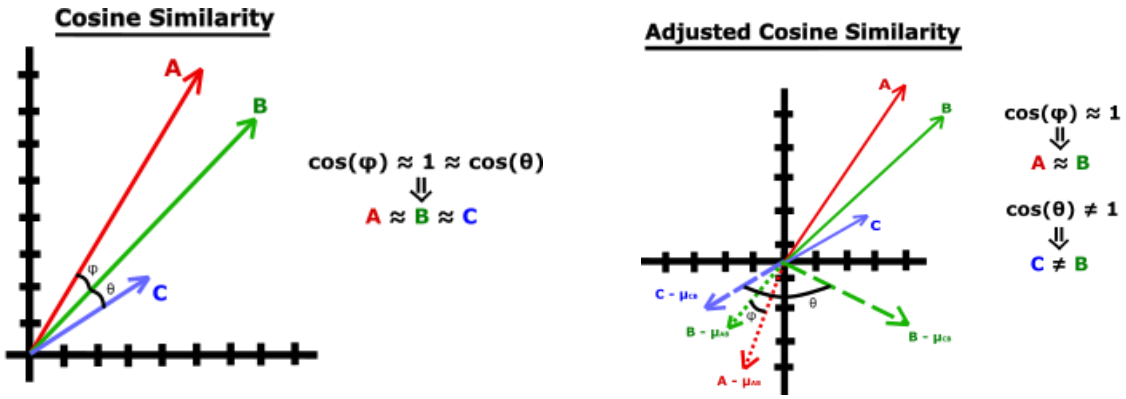

Figure 1: Using traditional cosine similarity, vectors A, B and C are all quantified as similar. With adjusted cosine similarity, vectors of similar angle *and* magnitude are quantified as similar.

However, using only angle to quantify similarity causes can cause census tracts massively different in a few dimensions to be classified as similar, as in Figure 1. To prevent this, we use the adjusted cosine similarity metric which subtracts the average of the two vectors from each of them before calculating the angle between them. Then, if vectors have similar angles but differ in magnitude in several dimensions, the adjusted cosine similarity will be closer to 0, as seen in Figure 1. This measure allowed us to rank the nonexposed census tracts by the adjusted similarity score between each of the eight exposed census tracts. From this ranking, we matched and selected census tracts with scores above ___ (a process known as ratio matching)[4] to each respective exposed census tract to offset the high variability in cancer incidence in the counterfactual scenario.

C. Disaggregation, multiple imputation, and machine learning model testing

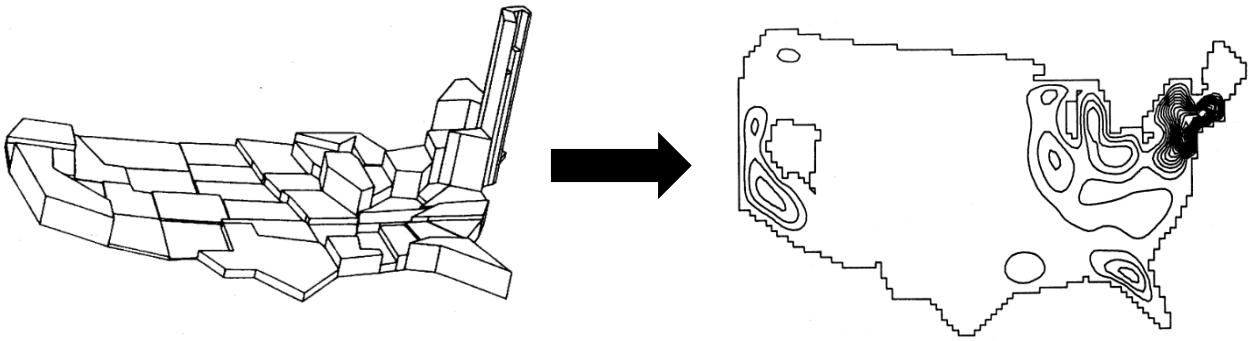

Figure 2: Pycnophylactic interpolation smooths large granularity data into a smooth density map which allows for smaller granularity estimation. Credit: (Tobler, 1979)

Because SEER data is reported at the county-level, we needed a methodology to turn county-level cancer incidence into census-tract cancer incidence predictions. We began by looking into common Census Bureau practices that seek to disaggregate high granularity data, which included areal interpolation. Two areal interpolation schemes that we decided to test included weighted area interpolation and pycnophylactic interpolation. Weighted area interpolation takes a source and target dataset with differing spatial granularities, overlays them, and attributes featues from the source to regions in the target dataset proportionally according to their share of the total area. Pycnophylactic interpolation (Figure 2) takes a spatial dataset with some feature and converts the discrete divisions into a continuous smooth spatial distribution of that feature.[12] Besides this, we also considered multiple imputation schemes, which average a series of machine learning predictions to estimate missing data. We investigated Multiple Imputation by Chained Equations and Multiple Imputation using xgboost which are implemented in the Python packages sklearn and miceforest, respectively.

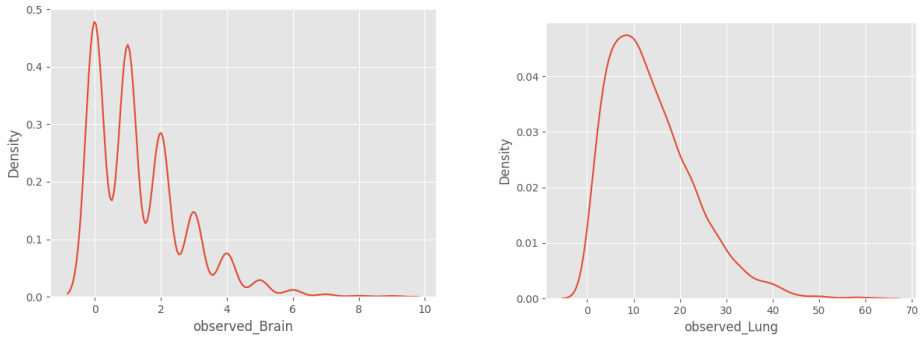

Figure 3: Base rate for the observed brain and observed lung cancer incidence from the New York census tracts that were used to train the machine learning models.

Finally, we also tested machine learning models in imputing cancer incidence. To do so, we trained a series of supervised learning models on the New York small area cancer data to predict census-tract-level cancer incidence in matched census tracts outside New York (Figure 3). The models we examined included the Gradient Boosting Regressor, Ada Boost Regressor, Bagging Regressor, Extra Trees Regressor, Histogram-based Gradient Boosting Regressor, and Random Forest Regressor.[9] We performed a five-fold cross validation on each method with a grid search to find the most optimal hyperparameters and selected the model with the best performance based on mean squared error (*MSE, smaller is better*) and the squared correlation coefficient (R^2, *larger is better).*

III. RESULTS

A. Disaggregation/Imputation method results and prediction performance

Disaggregation/Imputation Scheme	Mean Squared Error (MSE)
Weighted Area Interpolation	
Pycnophylactic Interpolation	
Multiple Imputation by Chained Equations	
Multiple Imputation using xgboost	
Machine Learning Model	

Table 1: Disaggregation/Imputation methods were compared using the mean squared error between county sums of census tract estimated cancer incidence and actual SEER county-level cancer incidence. Machine learning models were found to perform the best with a mean squared error of ____.

First, we sought to identify the best disaggregation/imputation scheme by comparing mean squared error of county sums to the SEER data following census tract cancer incidence estimation. While interpolation schemes maintained the county cancer

incidence as an upper bound for the sum of census tract cancer incidence estimates, the mean square error was simply too large to consider areal interpolation as a viable method to disaggregating our county-level data. We believe this is because these methods have no information on population distribution. Rather, they simply seek to divide up a feature proportionally by area or smooth out discrete boundaries to create a smooth distribution of that feature. Additionally, our map was not completely contiguous due to some states not having county-level cancer incidence data which may have interfered with these methods' interpolation procedures. Multiple imputation methods performed better, but still could not match the performance of the best machine learning model trained for imputation. With this in mind, we tested a multitude of models and recorded both mean square error and R^2 as performance measurements in TABLE 2.

Based on the performance in TABLE 2, we opted for the Gradient-Boosting Regressor for both brain and lung cancer incidence prediction. The lung cancer incidence model had an R^2 value of ____ and a mean squared error of ____. On the other hand, the brain cancer incidence model had a worse R^2 value of ____ and a mean squared error of _____. This large discrepancy between models is likely due to the differences in the number of cases, with lung cancer being ten times more common. Using Shapley values for each covariate, we found that the most important features were total population, proportion of the population that is white, and proportion of the population that was over the age of 25.

Model	**MSE (Lung)**	**R^2 (Lung)**	**MSE (Brain)**	**R^2 (Brain)**
Gradient Boosting Regressor				
Ada Boost Regressor				
Bagging Regressor				
Extra Trees Regressor				
Histogram-based Gradient Boosting Regressor				
Random Forest Regressor				

Table 2: The best machine learning models were the Gradient Boosting Regressor for both brain and lung cancer. The machine learning model performance for every model trained to predict brain and lung cancer incidence at the census tract level was evaluated by mean-squared error (MSE) and a squared correlation coefficient (R^2) value.

B. cSIR scores and Interpretation

In our cSIR evaluation, we found that ___ out of eight census tracts had a cSIR value greater than one for brain cancer incidence and ___ out of eight had a cSIR value greater than one for lung cancer incidence (Table 2). While the causal link must be confirmed by future research, the cSIR scores support dioxin's potential causal role for the brain/lung/brain and lung cancer cluster.

Census Tract	cSIR (Lung Cancer, Brain Cancer)
1251.03	(#, #)
1251.04	(#,#)
1258.04	(#, #)
1258.05	(#, #)
1258.06	(#, #)
1259.04	(#, #)
1259.05	(#, #)
1259.06	(#, #)

Table 2: The cSIR scores for both lung and brain cancer in each exposed Utah census tract.

To appropriately assess the presence of a causal relationship, we must also evaluate the cSIR results against Hill's criteria of causality.[2] Strength of association was determined by analysis with the cSIR. Both temporality and biological plausibility were confirmed as the dioxin exposure happened before the increased cancer incidence and dioxin has been confirmed as a carcinogen.[11] Consistency, coherence, and experimental support all remain undetermined as more work is necessary to establish experimental evidence and consensus regarding dioxin exposure's link to increased cancer incidence. Specificity is also not present because the exposure was hypothesized to increase the likelihood of two separate health outcomes: brain and lung cancer.

Criteria	Satisfied?
Strength of Association	Undetermined
Consistency	Undetermined
Specificity	No
Temporality	Yes

Dose-Response	No
Biological Plausibility	Yes
Coherence	Undetermined
Experimental Support	Undetermined

Table 3: Evaluation of the causal criteria to determine whether a preponderance of evidence exists supporting a causal relationship between dioxin exposure and lung/brain cancer.

IV. DISCUSSION

Here we utilized the causal estimand, cSIR, to reevaluate an identified lung and brain cancer cluster in eight Utah census tracts where the causal role of dioxin exposure was left undetermined. After assembling census data on census tract socioeconomic, health and demographic features, we matched the exposed census tracts adjusted cosine similarity score above ___. Then, using the SEER county level cancer incidence for matched census tracts as an upper bound, we used the gradient boosting regressor model trained on New York census tract-level cancer incidence data to predict cancer incidence in each matched census tract. This allowed us to compute the cSIR scores which indicated that dioxin exposure from the Wasatch Energy Systems incineration plant was a component in the increased brain/lung/brain and lung cancer incidence observed in ___ of the eight Utah census tracts. However, considering the lack of a preponderance of Hill's causal criteria which are satisfied, future work is necessary to confirm whether this support for a causal link between dioxin and brain and lung cancer clustering is valid.

Finally, this analysis is not without limitations. The most significant limitation is our lack of the observed census tract cancer incidence data for the eight exposed census tracts, which forced us to use our machine learning model to predict cancer incidence here. Future work will request the actual cancer frequencies from the Utah Cancer Registry. Besides this, variability in the cancer incidence rates was also reduced when averages were taken in the cSIR analysis, potentially impacting the cSIR scoring metric.

In addition, TRI data and Superfund sites might not represent all dioxin exposures across the United States. Because of this, it is possible that SUTVA may be violated by an unaccounted influencing variable defining a different exposure.[2] Additionally, identified census tract trends might not accurately represent true cancer incidence trends between individuals. One can try to remedy this by using the lowest granularity of cancer incidence data possible. Census block groups are often ideal, as they are the closest census group to the individual.[8] Standardization of reporting cancer incidence by census block groups like it is done in Illinois and New York would also improve the reliability of cancer cluster investigation results. It is also possible that we have overestimated the total

number exposed census tracts. We have TRI data beginning in 1998, which is later than the time of exposure suspected to be the factor in the UT cancer cluster, but still prior to the observed cancer cluster. We also have Superfund site data beginning in 1982 but it is important to note that the Superfund data doesn't include "non-disasters." So, even the Wasatch Energy Systems complex, a facility known for releasing dioxin into the surrounding area is not included in the Superfund site data. We made the most inclusive estimate possible by largely overestimating exposures from all the data that we had.

Regarding future work, improvements can likely be made by modifying the imputation or disaggregation method to get census-tract level cancer data. Use of multiple disaggregation methods simultaneously with prepared ancillary data about population distribution would likely allow these methods to perform better compared to our machine learning model. Additionally, the field of knowledge guided machine learning can produce constrained predictions during the modeling procedure, removing the need for a transformation after machine learning predictions are made.[13]

V. ACKNOWLEDGEMENTS

This manuscript has been co-authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

Works Cited:

[1] P.P. Liberski, B. Sikorska, S. Lindenbaum, L.G. Goldfarb, C. McLean, J.A. Hainfellner, P. Brown, Journal of Neuropathology & Experimental Neurology **71**, 2 (2012).

[2] A.B. Hill, Proc R Soc Med **58**, 295 (1965).

[3] An Investigation of Cancer Incidence In Census Tracts – 1251.03, 1251.04, 1258.04, 1258.05, 1258.06, 1259.04, 1259.05, And 1259.06, 1978-2001 (Centers for Disease Control and Prevention, 2007), pp. 1-22.

[4] R.C. Nethery, Y. Yang, A.J. Brown, and F. Dominici, J R Stat Soc Ser A Stat Soc **183**, 1253 (2020).

[5] M. Goodman, J.S. Naiman, D. Goodman, and J.S. LaKind, Crit Rev Toxicol **42**, 474 (2012).

[6] N.E. Breslow and N.E. Day, IARC Sci Publ 1 (1987).

[7] J. Luo, M. Hendryx, and A. Ducatman, J Environ Public Health **2011**, 463701 (2011).

[8] S. Alanee, J. Clemons, W. Zahnd, D. Sadowski, and D. Dynda, Anticancer Res **35**, 4009 (2015).

[9] F. Pedregosa, G., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplass, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Journal of Machine Learning Research, **12**, (2011).

[10] L.S. Shapley, Contributions to the Theory of Games, **2** (1953).

[11] N. Marinković, D. Pašalić, G. Ferenčak, B. Gršković, and A.S. Rukavina, Hig Rada Toksikol, **61**, 4 (2010).

[12] W. Tobler, Journal of the American Statistical Association, **74** (1979).

[13] A. Daw, R. Thomas, C. Carey, J. Read, A. Appling, and A. Karpatne, Society for Industrial and Applied Mathematics, **74** (2020).

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.

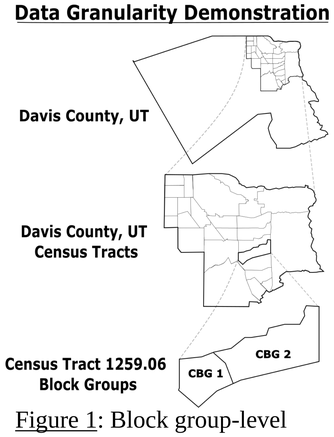

Figure 1: Block group-level and county-level data were transformed into census tract-level granularity.

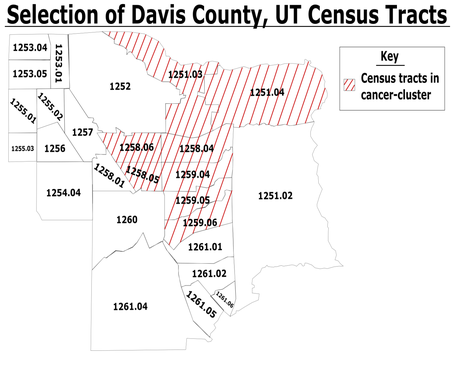

Figure 2: Selection of the eight Davis County, Utah census tracts containing both the confirmed brain and lung cancer cluster we are reinvestigating.

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.

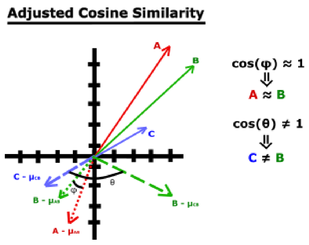

Figure 3: Adjusted cosine similarity quantifies how alike two vectors are in angle and magnitude.

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 an angle threshold of ____ 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).

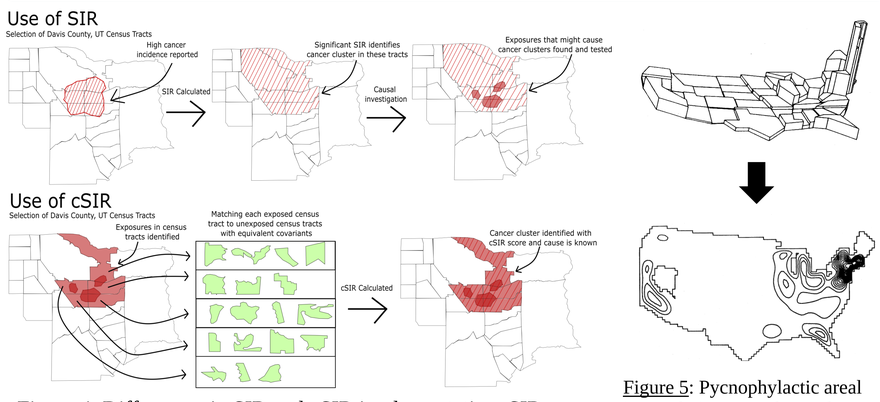

Figure 4: Differences in SIR and cSIR implementation. SIR finds cancer clusters before exposure, while cSIR finds exposure before the cancer cluster.

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 $\frac{SEER\ county\ incidence}{predicted\ county\ incidence}$ to reduce model error.

Results

Imputation Method	Mean Square Error
Weighted Area Interpolation	#
Pycnophylactic Interpolation	#
Multiple Imputation by Chained Equations (MICE)	#
Multiple Imputation using LightGBM	#
Gradient Boosting Regressor Model	#

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 ___ out of eight census tracts had a cSIR value above 1.0 for brain cancer incidence and ____ 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/didn't support dioxin's potential causal role for the brain/lung/brain and lung cancer cluster (Table 3).**

Census-Tract	cSIR (Brain, Lung)
1251.03	(#, #)
1251.04	(#, #)
1258.04	(#, #)
1258.05	(#, #)

Census-Tract	cSIR (Brain, Lung)
1258.06	(#, #)
1259.04	(#, #)
1259.05	(#, #)
1259.06	(#, #)

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.