A causal inference framework for cancer cluster investigations with Bayesian estimation of causal standardized incidence ratios using publicly available data

Rachel C. Nethery¹, Yue Yang¹, Anna J. Brown², Francesca Dominici¹
¹Department of Biostatistics, Harvard T.H. Chan School of Public Health
²Department of Statistics, University of Chicago

Abstract

Often, a community becomes alarmed when high rates of cancer are noticed, and residents suspect that the cancer cases are related to a known source of hazard. In response, the CDC recommends that departments of health perform a standardized incidence ratio (SIR) analysis to determine whether the observed cancer incidence is higher than expected, and, if significant results are found, to run follow-up studies to explore relationships with potential sources of hazard. Numerous statistical weaknesses, most notably the silent multiple comparisons problem, render results of this approach unreliable. We propose a novel causal inference approach to cancer cluster investigations, rooted in the potential outcomes framework, and we introduce a new estimand, called the causal SIR (cSIR), along with an estimation procedure, relying on causal matching methods and Bayesian hierarchical modeling, that can estimate the cSIR using publicly available data. In the proposed approach, a source of hazard representing a potential cause of increased cancer rates in the community is identified a priori, and the cSIR is defined as the ratio of the expected cancer incidence in the community of interest under the (factual) scenario of exposure to this hazard over the expected cancer incidence in the community under the (counterfactual) scenario of no such exposure. To estimate the cSIR, we first apply causal matching procedures to ensure that we are comparing the cancer incidence in the community of interest to the incidence that would have been expected for the community if it had not been exposed to the source of hazard but had remained otherwise the same. We then introduce a novel Bayesian model that jointly rectifies the spatial over-aggregation of publicly available cancer incidence data and utilizes data from the community of interest and its matched controls to estimate the cSIR. To illustrate the applicability of this approach, we will apply it to determine whether trichloroethylene vapor exposure has caused increased cancer incidence in Endicott, New York.

1 Introduction

Across the United States, citizens routinely recognize higher than expected rates of cancer in their community and request that the local health department conduct an investigation, with hopes of identifying a common cause. According to a review by Goodman et al. (2012), at least 2,876 cancer cluster investigations were conducted by health departments in the US between 1990 and 2011, most of which were initiated in response to community alarm at high numbers of cancer cases and fear of a connection between the cancer cases and a known environmental exposure. Shockingly few
of these investigations, however, have succeeded in identifying a link between the cancer cases and a common exposure, with the vast majority providing no clear answer to the concerned community.

1.1 Cancer cluster investigation protocol

The Centers for Disease Control and Prevention (CDC) has provided a protocol to guide health departments in responding to requests for cancer cluster investigations (Centers for Disease Control and Prevention, 2013). The first challenge in this process is that the definition of a cancer cluster is notoriously vague and contested. The CDC defines a cancer cluster as “a greater than expected number of cancer cases that occurs within a group of people in a geographic area over a defined period of time” (Centers for Disease Control and Prevention, 2013). They recommend that cancer cluster investigations proceed by first performing statistical analyses to determine whether the number of cancer cases experienced by the community represents a statistically significant elevation of cancer incidence, i.e., a number of cases significantly higher than what would be expected. If statistical significance is found, then the event constitutes a cancer cluster by their definition. Only if a cancer cluster is confirmed does the CDC suggest initiating investigations seeking to identify a common cause.

To formalize the CDC’s cancer cluster assessment statistical protocol, we first define \( Y \) as a random variable representing the observed number of cancer cases in the community of interest during a relevant time period. Cancer cluster investigations typically begin by computing an expected cancer number of cancer cases \( E = E(Y) \) for the community based on the incidence in some selected comparable or background population. Then, the observed number of cancer cases in the community is compared to the expected number by estimating the ratio \( S = \frac{Y}{E} \). \( S \) is called a standardized incidence ratio (SIR). An SIR estimate greater than 1 indicates an elevated cancer incidence in the community compared to what would be expected based on background incidence rates.

In order to test whether the estimated SIR is significantly greater than the null value of 1 (\( H_0 : S = 1 \)), statistical uncertainties must also be computed. Using the assumption that \( Y \sim \text{Poisson}(S \times E) \), confidence intervals and p-values can then be computed either exactly or approximately by invoking the relationship between the Poisson and Chi-Square distributions (Sahai and Khurshid, 1993). A p-value less than 0.05, or a lower confidence limit exceeding 1 generally results in the declaration of a cancer cluster. We note that, while not identical, this SIR estimation method has similarities to estimation using a simple Poisson regression model without confounding adjustment.

Only if this procedure reveals compelling evidence of a cancer cluster does the CDC recommend that health departments proceed to seek possible environmental causes. If the statistical evidence for a cancer cluster is strong and an epidemiological study to test for relationships between environmental factors and the cancer cases is deemed “warranted” and “feasible”, then the CDC suggests conducting such a study.

1.2 Critique of cancer cluster investigations

Although widely used, the protocol for cancer cluster analyses described above has recently been highly criticized on practical grounds, with objectors pointing to the fact that such analyses rarely lead to definitive identification of the cause(s) of the cluster. With no significant conclusion about the cause(s) of a potential cluster, simply the identification of one provides no guidance for the
concerned public in removing the cause or preventing future cases. In a review of cancer cluster investigations from 1990-2012 (Goodman et al., 2012), it is reported that out of the 428 considered, 72 clusters were confirmed, but only three were linked with hypothesized exposures and merely one revealed a clear cause. The many challenges in cancer cluster investigations have been summarized in recent years (Goodman et al., 2014). These include long and unclear latency period of cancer, missing rare disease incidence rate, difficulty in defining the relevant cluster area and population, population migration, and lack of confounder data in cancer registries, among others. While Goodman et al. (2014) propose several novel approaches to improve cancer cluster investigation, they do not address the most pressing statistical limitations that threaten the validity of cancer cluster analyses.

In addition to the questions surrounding the practicality of the current cancer cluster analysis protocol, the underlying statistical analysis has also come under attack. The most prominent statistical obstacle in cancer cluster analyses is the silent multiple comparison problem, also known as the Texas Sharpshooter problem (Bender et al., 1995). This issue arises due to the reliance on observed outcome data to inform the development of the statistical hypothesis, i.e., the hypothesis that a cancer cluster is present in the chosen community and time period is only formed in response to the observation that an unusually high number of cancer cases has occurred. Fundamental statistical principles dictate that, occasionally, the locations of cancer diagnoses will cluster together in space and time due to chance alone, i.e. not due to any common cause. Thus, we would expect, due only to chance, to occasionally encounter what appears to be an unusual excess of cancer cases within a small area. Thus, if we first evaluate the spatial distribution of the cancer cases and draw a boundary around a small area that appears to have a higher than expected cancer incidence, and then ask if that area is experiencing a higher cancer incidence than expected by chance, we inflate the probability of finding a false positive. Moreover, we are unable to effectively estimate how many multiple comparisons to adjust for, given the near infinite combination of different possible area boundaries, time periods and diseases that could be assessed (hence the term “silent” multiple comparisons). Therefore, even multiple comparisons adjusted p-values can range dramatically and thus are not reliable to determine the significance of a cluster.

There have been attempts to address the silent multiple comparison problem, including the introduction of Bayesian methodology (Coory et al., 2009) which suggests we include our uncertainty arising from the multiple comparison problem into our choice of prior distribution; however, the presence of these alternate approaches has failed to produce changes in the way cancer cluster analyses are carried out. In order to completely avoid the multiple comparison problem, it has also been suggested that the assessment of cancer clusters should abandon statistical analyses entirely and instead prioritize identification of an exposure that could represent a common cause of the cases (Coory and Jordan, 2013).

We take the position that statistical analyses can provide important insights to cancer cluster investigations; however, in order to produce useful and reliable inference, both the overall protocol and the statistical procedures must be overhauled. In this paper, we propose a number of changes to the cancer cluster investigation protocol in order to situate it within a causal inference framework, and we introduce a novel statistical modeling approach that allows for estimation of causal SIRs using publicly available cancer registry data. These methods provide a statistically rigorous approach to SIR analyses that can be adopted by the agencies that investigate cancer clusters. This approach will not only provide more reliable answers to concerned communities but will also directly answer the question of interest, i.e., whether a common exposure in the community caused
the increased cancer rates.

The most notable procedural change required for this causal inference framework is that potential or suspected causes of increased cancer incidence must be identified prior to any statistical analyses. In short, although an agency receives a report of a suspected cancer cluster within a certain community, it should move forward as if this were not known and search for possible common toxic exposures in the vicinity. Only after we have identified exposures in the community that represent potential causes should we construct statistical analyses, with the spatial and temporal boundaries of the area under study and the types of cancer investigated informed by realistic exposure hypotheses.

With a potential source of hazard in mind, causal questions can then be posed and investigated. We propose the use of a causal inference approach, rooted in the potential outcomes framework, to investigate whether a given exposure caused increased risk of certain cancers in the exposed population and time period. Using potential outcomes, we define a causal SIR (cSIR) and lay out the identifying assumptions, and we introduce a method for estimation that relies solely on publicly available data. Through the use of causal matching, this method ensures that the populations selected for comparison with the exposed population are (1) true “controls”, in the sense that they are not exposed to the same source of hazard, so that the expected incidence is not unduly inflated by inclusion of other exposed communities, and (2) highly similar to the exposed population in terms of all other relevant features, so that the threat of confounding is eliminated. We have also developed a novel Bayesian cSIR estimation model, to be applied to the matched data, that can accommodate spatial over-aggregation of the publicly available cancer incidence data for the matched controls.

In Section 2, we describe our causal inference framework for cancer cluster analyses and provide a method to estimate cSIRs when all data are appropriately spatially aggregated. In Section 3, we explain the challenges presented by publicly available cancer registry data and we extend our method to accommodate the spatial over-aggregation in these data. In Section 4, we use simulations to compare our methods to the existing SIR analysis methods used in cancer cluster investigations. In Section 5, we demonstrate our method with an application to investigate whether trichloroethylene vapor exposure in Endicott, New York has caused increases in kidney cancer incidence in the town. Finally, we discuss our findings and conclude in Section 6.

2 Methods

2.1 A priori identification of a possible cause

In this section, we will suggest revisions to the CDC’s cancer cluster assessment protocol that would remedy many of the statistical issues plaguing the current approach and, under certain assumptions, would provide estimates of the causal effect of an exposure on cancer incidence. We first remark that the CDC’s current approach to these analyses is statistically and epidemiologically backwards in that it tests for elevated cancer risk in a community prior to identifying putative sources of hazard that could be responsible for such elevation. Without a priori knowledge about the exposures that could have caused elevated cancer risk in a community, the hypotheses being tested in the statistical analysis are not well-defined. In particular, because no specific source(s) of exposure are postulated prior to analysis, the geographic region, time period, and disease types used to formulate the statistical hypothesis are at best defined arbitrarily, or at worst defined based on
observed distributions of cancer outcomes. If these features are arbitrarily specified, there is a high risk of failing to detect an elevated risk when one truly exists due to a focus on the wrong region, time period, or cancer type. Moreover, as discussed above, if the features are specified based on observed cancer outcomes, the statistical analysis will be subject to the problem of silent multiple comparisons, resulting in an inflated chance of detecting an elevated risk when one does not truly exist.

To avoid these pitfalls, we recommend that, in cancer cluster investigations, putative sources of hazard be identified prior to conducting any statistical analyses. With these potential sources identified a priori, statistical hypotheses can be shaped around appropriate geographic regions, time periods, and cancer types and will therefore provide more reliable and insightful results. Of course, even after potential sources of hazard are identified, figuring out the relevant cancer types and the at-risk population may not be trivial and may require input from medical and toxicological experts.

As discussed above, one complaint surrounding the current cancer cluster investigation protocol is that, even when statistical testing reveals evidence of a cluster, it rarely succeeds in identifying any common cause (Goodman et al., 2012). By testing hypotheses about specific putative sources of hazard, statistically significant results reveal a relationship between exposure to specific source(s) of hazard and cancer risk, thus providing greater insight into potential causes of any clusters. Below, we propose a novel approach to SIR analyses that, under certain assumptions and in conjunction with this a priori identification of putative source(s) of hazard, provides SIR estimates with a causal interpretation.

2.2 Causal inference: mimicking randomized trials with observational data

In scientific research, the gold standard for assessing links between a treatment or exposure (e.g., a new drug or environmental contaminants) and health outcomes is to conduct a randomized experiment. An example of a randomized experiment is controlled clinical trials, which is the gold standard in the context of drug approval. In a randomized clinical trial, scientists have control of the experiment and randomly choose which patients get the treatment and which patients get the placebo. Because of the randomization, the only difference between the treated and untreated groups should be the treatment assignment: on average the population of individuals that are treated and untreated will be similar in terms of other characteristics such as socio-economic status, age, gender, race etc. This means that a direct comparison of the average outcomes of the two groups should provide an unbiased estimate of the causal effect of treatment.

However, a randomized experiment to estimate the causal effects of exposure to contaminants on cancer generally cannot be conducted for ethical reasons. For example, if we wanted to estimate the causal effect of exposure to radiation on cancer risk, it would not be acceptable to randomly force some study participants to regularly be exposed to radiation. Instead, we must find a way to use natural and observed variation in exposure to radiation to learn causal effects.

Estimating causal effects of an exposure from observational, non-randomized data is considerably more challenging, primarily due to the threat of confounding. A confounder is a factor that is associated with both the exposure and the outcome of interest. Because the confounder is associated with exposure, the exposed and unexposed groups will systematically differ on the confounder. Moreover, because the confounder is associated with the outcome, a direct comparison of the outcomes in the exposed and unexposed groups will not estimate the causal effect of exposure because some of the differences in the outcomes of the groups may be due to their differences in
confounder levels rather than their differences in exposure status. An abundance of literature in the field of causal inference now provides methods that, under some assumptions, can manipulate observational data to remove confounding, thereby mimicking a randomized experiment, so that the causal effect of exposure may be extracted.

Matching is one of the most well established causal inference approaches for eliminating confounding in observational data. Matching methods are entirely nonparametric and, therefore, require fewer assumptions than many causal inference methods. For each exposed unit in the data, matching methods typically aim to identify a pre-specified number of “matched control” units that are as similar as possible to the exposed unit in terms of observed confounders but are unexposed. Only matched units are retained for analysis, unmatched units are discarded. This procedure ensures that the distribution of all confounders are similar in the exposed and unexposed groups so that, as in a randomized experiment, the only systematic difference in the two groups is in their exposure status. Thus, by comparing the outcomes in the exposed and unexposed groups in this matched dataset, causal effects of exposure can be recovered. Methods like matching, which aim to remove confounding from the data without integrating information regarding outcomes, are known as “design phase methods”. Procedures applied subsequently using the outcomes to estimate causal effects are known as “analysis phase methods”.

2.3 The potential outcomes framework and the causal SIR

We now define notation that will be used to develop the causal SIR framework, which builds on the causal inference and matching ideas elaborated above. Causal inference methods are generally situated within the potential outcomes framework, as defined by Rubin (1974), which we will now extend to the cancer cluster analysis setting. Here we adopt notation similar to that used above, but we now (1) allow the exposed community of interest to be partitioned into numerous disjoint exposed regions and (2) consider data both from these exposed regions and their matched controls. We recommend partitioning the exposed area into as many sub-areas as the available data will permit, as this will allow better adjustment for confounding and provide greater power to detect effects of exposure. We let $Y_j$ be a random variable representing the observed number of cancer cases in region $j$ during the time period of interest, where $j = 1, \ldots, N$ index the complete matched dataset of the exposed regions and their controls. We let $T_j$ denote an indicator of exposure status, i.e., $T_j = 1$ if region $j$ was exposed to the hypothesized type of hazard and $T_j = 0$ otherwise. Let $X_j$ be a vector of observed confounder values for region $j$ (confounders would typically be socioeconomic, demographic, and behavioral variables, many of which are publicly available through the US census). Then the potential outcomes are $Y_j(T = 1)$, the number of cancer cases that would have been observed in region $j$ under exposure, and $Y_j(T = 0)$, the number of cancer cases that would have been observed in region $j$ under no exposure.

The fundamental problem of causal inference is that only one of the two potential outcomes can ever be observed for a given region, either the outcome under exposure or the outcome under no exposure. The unobserved potential outcome is called the counterfactual. As in nearly all causal inference analyses, we invoke the stable unit treatment value assumption (SUTVA) in order to ensure the existence of the potential outcomes (Rubin, 1980). SUTVA requires that the exposure be well-defined, i.e. that there is only a single “version” of exposure, and that the exposure status of a given region does not affect the outcome (cancer incidence) of other regions.
Using these potential outcomes, we now define the causal SIR (cSIR), our proposed estimand:

\[
cSIR = \frac{E[Y(T = 1)|T = 1]}{E[Y(T = 0)|T = 1]}
\]

We note that this estimand is similar to the classic SIR, but, because it adjusts for confounding and compares the incidence under exposure to the expected incidence under no exposure (instead of an expected incidence based on background rates), this estimand is endowed with a causal interpretation. As with the classic SIR analysis, we are interested in testing \(H_0 : cSIR = 1\), with \(cSIR = 1\) equivalent to \(E[Y(T = 1)|T = 1] = E[Y(T = 0)|T = 1]\), i.e., no causal effect of exposure in the exposed population. In order to estimate the cSIR from observed data, identification assumptions are required. These assumptions are nearly identical to those needed to estimate the average causal effect in classic causal inference settings– ignorability and causal consistency. First, cSIR identification relies on the classic assumption of no unobserved confounding, stated mathematically as \((Y(T = 1), Y(T = 0)) \perp \perp T | X\), i.e., conditional on observed confounders, \(X\), the assignment to exposure is independent of the potential outcomes. Moreover, we must assume that each unit of analysis (each region) was eligible to be exposed, i.e., exposure to this type of hazard could feasibly have occurred in any of the regions included in the dataset, and we must be able to identify suitable unexposed matches for each exposed region (this is analogous to the usual positivity assumption). Together, the assumptions of no unobserved confounding and positivity are known as ignorability. Finally, the causal consistency assumption states that \(Y_j = Y_j(T = 1) \times T_j - Y_j(T = 0) \times (1 - T_j)\), i.e. the observed outcome is equal to the potential outcome under the observed exposure level.

By applying these assumptions, we can see that the cSIR is identifiable from the observed data. Consider the numerator of the SIR, \(E[Y(T = 1)|T = 1]\). Note that

\[
E[Y(T = 1)|T = 1] = E_X [E[Y(T = 1)|T = 1, X]] = E_X [E[Y|T = 1, X]]
\]

where the last equality holds by causal consistency. Similarly, for the denominator, \(E[Y(T = 0)|T = 1] = E_X [E[Y(T = 0)|T = 1, X]]\). Now, we invoke the ignorability assumption, which states that treatment status is independent of the potential outcomes conditional on confounders, so that

\[
E[Y(T = 0)|T = 1, X] = E[Y(T = 0)|T = 0, X].
\]

Thus, we have

\[
E[Y(T = 0)|T = 1] = E_X [E[Y(T = 0)|T = 0, X]] = E_X [E[Y|T = 0, X]]
\]

by applying causal consistency as above. Thus, we see that both the numerator and denominator of the cSIR are identifiable and can be estimated with observed data.

### 2.4 Estimation of the cSIR

To estimate the cSIR, we recommend the use of matching in the design phase to remove confounding, followed by a Bayesian estimation procedure in the analysis stage to appropriately adjust for all sources of uncertainty. The ideas to apply matching and Bayesian estimation methods in SIR analyses are not entirely novel, although, to our knowledge, they have never been combined in the way proposed here to produce an SIR analysis with a causal interpretation. Dominici et al. (2007) described their use of matching to adjust for possible confounding in an SIR analysis performed for a legal case related to a brain cancer cluster. Both Wakefield and Morris (2001) and Coory et al. (2009) proposed a Bayesian approach to disease risk/SIR analyses. In Coory et al. (2009), the
Bayesian approach was motivated by the desire to resolve the silent multiple comparisons problem inherent in frequentist SIR analyses. We combine and expand on these ideas to estimate the cSIR. We begin by describing the design phase of our approach below.

The goal of our matching procedure is to obtain a set of unexposed regions that have as similar as possible to the confounder distribution in the exposed regions. Thus, the first step in the design phase is to identify a (hopefully large) set of unexposed communities in which to search for matches. This step may prove challenging, because, for some types of hazard, reliable data about which geographic regions are exposed may be difficult or impossible to obtain. Moreover, certain contaminants (such as air pollutants) may be assumed to be universally present at some low level, making it impossible to find truly unexposed communities. In such cases, the “controls” selected for matching could instead be communities with exposure levels below a certain threshold, where the threshold should be selected so that exposures below that level are believed to have no effect on health.

Because we allow the exposed community to be partitioned into numerous regions, we must expand on the matching strategy used by Dominici et al. (2007), who sought matched controls for a single exposed region. Examples of matching procedures that identify a control(s) for each of many exposed/treated units are abundant in the causal inference literature (Rubin, 1973; Rubin and Thomas, 2000; Abadie and Imbens, 2006; Ho et al., 2007; Abadie and Imbens, 2011; Iacus et al., 2011) and are reviewed by Stuart (2010). In practice, we recommend utilizing the matching procedure that provides the best covariate balance between the exposed and unexposed regions, i.e. the smallest standardized differences in means. Covariate balance is an indication that, as in a randomized trial, the distributions of observed confounders are similar in the exposed and unexposed groups. Most matching procedures allow for ratio matching– finding multiple matches for each exposed region, and, because small area cancer incidence rates are often unstable, we suggest applying ratio matching in this context in order to obtain as much information as possible about the expected cancer incidence under no exposure.

After matching, if cancer incidence data for both case and matched control regions are available and aggregated at the desired level for analysis, we may simply use a loglinear modeling approach to estimate the cSIR. In the next section, we propose an extension to the loglinear model approach that can be used when the cancer incidence data for the potential matched controls are spatially over-aggregated, as would usually be the case when relying on publicly available cancer incidence data. The loglinear model approach for estimating disease risk relative to a point source with aggregated data was introduced by Diggle et al. (1997) from a frequentist perspective and by Wakefield and Morris (2001) from a Bayesian perspective. This approach assumes that the locations of cases in the region of study (for us the US) follow a point process model– specifically an inhomogeneous Poisson process with intensity at location $l$ a function of the intensity of the non-case generating Poisson process at $l$, as well as other covariates. Under the null hypothesis, the intensity of the case generating process at $l$ is not related to the exposure of $l$ to the purported source of hazard. For greater detail on these assumptions, see Diggle et al. (1997). Our approach will rely on this same set of background assumptions.

The loglinear model for estimation of the cSIR should include both exposure status and confounder variables as predictors. If matching procedures are entirely successful at removing all differences in confounder distributions between the exposed and unexposed, then adjustment for the confounders in the analysis phase is not needed. However, in practice, matching typically does not remove these differences entirely, thus adjustment for the confounders in analysis phase modeling is recommended.
The model, then, has the form

$$\log(E[Y_j]) = \alpha_0 + \alpha_1 T_j + \alpha_2 X_j + \log(N_j)$$

where $N_j$ is the population size in region $j$, and $\log(N_j)$ is an offset term used to account for potential differences in population size across the regions. If the cancer incidence data are collected over different time periods for different regions, the offset could represent person-time rather than population size. Because the sample size for this model (the number of regions in the matched dataset) will generally be small, we recommend a Bayesian approach to model fitting, which will generally provide more stable estimates than frequentist models. The cSIR estimate is $\exp(\hat{\alpha}_1)$, and uncertainties and confidence regions follow accordingly. If cancer incidence data exhibit excess zeros or extra-Poisson variation, existing extensions to the Bayesian Poisson regression model to accommodate these deviations can be used.

3 A cSIR estimation procedure to accommodate spatial over-aggregation in publicly available cancer incidence data

While improvements are being made in the accessibility of cancer incidence data, data privacy and patient identifiability concerns mean that publicly available data must be aggregated to a degree that renders it too coarse to perform the desired analyses in many cases. In this section, we introduce approaches to estimating the cSIR when the available cancer incidence data for some or all of the potential matched controls regions are unsuitably spatially aggregated. This means that the geographic units for which cancer incidences are available are incompatible with the desired units for matching and analysis.

The most commonly used publicly available cancer incidence data are collected and published by the Surveillance, Epidemiology, and End Results (SEER) Program of the US National Cancer Institute (National Cancer Institute Surveillance, Epidemiology, and End Results Program, 2018). Today, SEER compiles cancer diagnosis records from a total of 18 state and city cancer registries. The SEER data contain records of all diagnosed cancer cases in nine states/cities dating back to 1975 or prior. Four additional state/city registries have contributed data dating back to 1992, and five additional city/state registries have contributed data dating back to 2000. For each record in the SEER data, demographic data and diagnosis information are provided, as well as the year of diagnosis and the person’s county of residence at the time of diagnosis. Thus, the lowest available level of spatial aggregation for these data is the county level. However, sources of hazard that are suspected to contribute to higher cancer incidence in a community generally have effects that are much more spatially concentrated than the county level; thus, in most cases, the exposed community and/or the regions formed by its partitioning will be much smaller (and therefore more homogeneous) than a county and would likely be better compared to zip codes, census tracts, or census block groups (CBGs). Yet, while obtaining confounder data (i.e. socioeconomic and demographic characteristics) at the zip code, census tract, or CBG level is typically easy and therefore matching is possible at these levels, cancer incidences from SEER are only available at the county level.

Recently, some states have begun to make small area cancer data publicly available. Illinois provides state-wide zip code level cancer incidence data for 11 different anatomic site groupings (Illinois State Cancer Registry, 2017). These incidences can be stratified by sex, diagnosis year
(1986-2015, aggregated over 5 year periods), disease stage, and age at diagnosis (0-14, 15-44, 45-64, 65+). New York state provides state-wide CBG level cancer incidences for 23 different anatomic sites over the period 2005-2009 (Boscoe et al., 2016). These CBG incidences are not stratified on any demographic features. We note that, to ensure the protection of privacy, publicly available small area cancer incidence data are aggregated over fairly large time periods, multiple disease classifications, and demographic strata, which may render the incidence values not directly comparable to those of interest in the exposed community.

Inspired by the publicly available SEER, Illinois, and New York data, our approach to cSIR estimation begins with the assumption that, in many areas where matched control regions are sought, cancer incidence data with appropriate temporal and demographic aggregations is over-aggregated spatially—available only at the county level. Meanwhile, in fewer areas, cancer incidence data with possibly inappropriate temporal and demographic aggregation is available at the desired level of spatial aggregation (or a smaller level, since aggregation up to larger spatial regions is straightforward). We note that this is a generalization of the case when cancer incidence data for all matched controls are available at the desired spatial aggregation, and a simplified version of the proposed methodology, which we described in the previous section, can be used to estimate the cSIR in this more straightforward scenario. Throughout this section, we assume that the cancer incidences in each region of the exposed community of interest are known exactly and that all confounder data (for matching) are available at the desired level of spatial aggregation. Without loss of generality, we assume that the desired spatial level of the analysis is the CBG and that sources of cancer incidence data outside the community of interest include only the publicly available ones described above.

To estimate the cSIR in this setting, we propose a two-stage Bayesian model to be applied to the matched dataset. This model (1) combines publicly available small area cancer data, SEER data, and socio-economic and demographic data to predict CBG cancer incidences in the matched control regions and (2) using these predictions and the observed incidence data from the exposed regions, fits the loglinear model described in the previous section to estimate the cSIR.

3.1 Stage 1: Prediction model

The goal of the prediction stage is to use the limited publicly available information about CBG cancer incidence to learn about relationships between CBG incidence and community characteristics and then to use those relationships, along with more plentiful SEER county cancer incidence data, to predict cancer incidences for all CBGs in each SEER-participating state. This entails building a prediction model with the New York CBG cancer data as the outcome and many CBG level community characteristics as predictors. In order to be used to predict at the CBG level in SEER states and account for the observed county level cancer incidences, our model must also incorporate the constraint that the CBG predicted incidences within a given county should sum to the observed county incidence. Finally, because this model is fit only to New York data but is employed for prediction across all SEER states, we must make an additional assumption that the results of this model are transportable, or equivalently, that the model has external validity (Singleton et al., 2014).

Amending notation a bit, we let $Y_{ij}$ and $Z_{ij}$ denote, respectively, the cancer incidence and the vector of predictors for CBG $j$ within county $i$. The variables in $Z$ are likely to largely overlap with those in $X$, but the sets may not be identical. Let $N_{cty}$ be the number of counties in the sample and $n_i$ be the number of CBGs in county $i$. We assume $Y_{ij} \sim \text{Poisson}(\lambda_{ij})$ and
\[ \log(\lambda_{ij}) = Z_{ij}\beta + \log(N_{ij}), \text{ where } N_{ij} \text{ is the population size (or person-time, if needed).} \]

Then, it is a well-known result that \((Y_{i1}, \ldots, Y_{in_i})|\sum_{h=1}^{n_i} Y_{ih} = K_i \sim \text{Multinomial}(K_i, \pi_i), \pi_i = (\pi_{i1}, \ldots, \pi_{in_i}), \]

with \(\pi_{ij} = \frac{\lambda_{ij}}{\lambda_{i1} + \cdots + \lambda_{in_i}}\). \(K_i\) corresponds to the cancer incidence in county \(i\) and \(\pi_{ij}\) the proportion of county \(i\)'s cancer incidence that falls into CBG \(j\). Thus, in order to account for the constraint that CBG incidence predictions should sum to their county’s observed incidence, we will develop the model around a multinomial likelihood, letting the \(\pi_{ij}\) be a function of the predictors, \(Z_{ij}\).

The form of the prediction model is similar to a classic loglinear model, but includes a non-traditional offset that imposes the constraint that the estimated multinomial probabilities must sum to one. It follows from the properties laid out above that \(\pi_{ij}\) should have the following relationship to the predictors:

\[ \log(\pi_{ij}) = Z_{ij}\beta + \log(N_{ij}) - \log(\sum_{h=1}^{n_i} e^{Z_{ih}\beta} N_{ih}) \]

where the final term is an offset which imposes the constraint. Note that this implies that

\[ \pi_{ij} = \frac{e^{Z_{ij}\beta} N_{ij}}{\sum_{h=1}^{n_i} e^{Z_{ih}\beta} N_{ih}} \]

so that the CBG probabilities within a county sum to one, as desired. Combining this with the multinomial assumption results in the following data likelihood:

\[ L(\beta|Y, Z) = \prod_{i=1}^{N_{cty}} \prod_{j=1}^{n_i} K_i! \frac{(e^{Z_{ij}\beta} N_{ij})^{Y_{ij}}}{Y_{ij}!(\sum_{h=1}^{n_i} e^{Z_{ih}\beta} N_{ih})^{Y_{ij}}} \]

Now, from a Bayesian perspective, we can fit this model to the New York CBG cancer incidence data through the use of a simple Metropolis sampler. The resulting posterior summaries of \(\beta\) speak to the associations between a CBG’s features and the proportion of the cancer incidence of its larger county that it accounts for.

Moreover, for any CBG in the SEER states, where \(Z_{ij}\) and \(K_i\) are known, we can obtain posterior predictive samples of its cancer incidence from the multinomial distribution with parameters \(K_i\) and the modeled \(\pi_i\). We note that we only need posterior predictive samples from the matched control CBGs in order to do cSIR estimation. However, because the model relies on normalization of the CBG proportions within counties, in order to obtain the multinomial posterior predictive samples for any CBG we must utilize the predictor data from all the other CBGs in its encompassing county, as well as the observed SEER incidence for the county. Assuming matched control CBGs have been selected prior to the model fitting, we denote the posterior predictive samples for the matched control CBGs as \(\{Y_{ctl}^{(1)}, \ldots, Y_{ctl}^{(M)}\}\), and these get passed into the estimation stage of the model.

### 3.2 Stage 2: Estimation model

Using the observed incidences in the exposed CBGs and the posterior predictive samples of the incidence in the matched control CBGs, we estimate the cSIR in the second stage of the model. As above, we employ a Bayesian loglinear model, now integrating in the sampled outcomes for the controls at each iteration of the sampler. By including the full distribution of predicted cancer
incidences in the estimation stage, rather than a single summarized predicted value, our cSIR estimate will correctly capture the additional variability generated by the use of predicted cancer incidences for the matched controls.

Let \( Y^{(m)} \) be a vector containing the observed cancer incidence for the exposed regions and the \( m^{th} \) posterior predictive sample of the cancer incidence for control regions, \( Y^{(m)}_{ctl} \). Then in each iteration of the Metropolis sampler for the estimation model, we plug in a different \( Y^{(m)} \) sample, i.e. for \( m = 1, \ldots, M \) we collect a posterior sample of \( \{\alpha_0, \alpha_1, \alpha_2\} \) from

\[
\log(E[Y^{(m)}]) = \alpha_0 + \alpha_1 T + \alpha_2 X + \log(N).
\]

The cSIR and its uncertainties are estimated from this model as described in Section 2.4.

We note that, in its current form, stage 1 of this method relies on the Poisson distribution and is not equipped to handle zero-inflated cancer data. Thus, the reasonableness of a Poisson data likelihood should be assessed before applying this method to data. Moreover, prediction models based on publicly available data have not yet been validated for any cancer type and the results and model fit should be evaluated on a case-by-case basis. Validation of these prediction models is an important topic for future work.

4 Simulations

In this section, we conduct simulations to compare the current protocol for SIR estimation to our proposed method. Primarily, we intend to demonstrate how the use of matched controls and Bayesian estimation methods, under the assumptions laid out above, leads to stable and unbiased estimation of the effect of an exposure on cancer incidence. Recall that our proposed approach calls for a priori specification of a potential cause of increased cancer incidence, and determination of the area and time period under study should be based on exposure hypotheses. In practice, these procedures would likely lead to a different area/time period under study than the CDC’s protocol, which bases these considerations only on where/when the high cancer incidence is reported. However, in these simulations, for comparability, we consider the same area/time period under study for all methods compared.

We simulate data under 4 conditions: (1) no exposure effect and no confounding, (2) no exposure effect and confounding, (3) exposure effect and no confounding, and (4) exposure effect and confounding. Moreover, within simulations 3 and 4, we simulate data exhibiting varying strengths of exposure effects (simulations 3(a)-3(d) and 4(a)-4(d)) to test the power of our method to detect effects, an important consideration given that the number of regions composing the exposed community is generally small. We also structure the simulations using real confounder data from the SEER states so that we are ensured that the simulated data reflect the complexity of real data.

For each SEER state CBG, we collect the following variables: percent of the population age 65+, percent of the population male, percent of the population white, percent of the adult population unemployed, average commute time, median household income, dollars spent on smoking products as a portion of per capita income, and percent of total dollars spent on food that was spent on food outside the home. We then utilize these variables in different ways to construct the exposure and outcome variables for simulations (1)-(4). Let \( X \) denote a matrix containing this set of confounders for each CBG in the SEER states, \( T \) denote the vector of exposure indicators for each CBG, and \( Y \) denote the vector of cancer incidences for each CBG. In each simulation, we generate \( T \) and \( Y \)
using the models $\log(P(T = 1)) = \gamma_0 + \gamma_1 X$ and $\log(E[Y]) = \alpha_0 + \alpha_1 T + \alpha_2 X + \log(N)$, where $N$ is a vector containing the population in each CBG. Below, we specify the parameter values utilized in each simulation setting:

Simulation 1:
$\gamma_0 = -1.15$, $\gamma_1 = 0$, $\alpha_0 = -5.99$, $\alpha_1 = 0$, $\alpha_2 = 0$

Simulation 2:
$\gamma_0 = 0$, $\gamma_1 = \{0.0009, 0.015, 0.003, -0.001, -0.01, 0.004, 0.002, -0.01\}$, $\alpha_0 = -5$, $\alpha_1 = 0$, $\alpha_2 = \{0.007, 0.015, 0.03, -0.001, -0.02, 0.004, 0.002, -0.005\}$

Simulation 3:
$\gamma_0 = -1.15$, $\gamma_1 = 0$, $\alpha_0 = -5$, $\alpha_2 = \{0.007, 0.015, 0.03, -0.001, -0.02, 0.004, 0.002, -0.005\}$
(a) $\alpha_1 = 0.1$, True SIR=1.1
(b) $\alpha_1 = 0.26$, True SIR=1.3
(c) $\alpha_1 = 0.41$, True SIR=1.5
(d) $\alpha_1 = 0.69$, True SIR=2

Simulation 4:
$\gamma_0 = 0$, $\gamma_1 = \{0.0009, 0.015, 0.003, -0.001, -0.01, 0.004, 0.002, -0.01\}$, $\alpha_0 = -5$, $\alpha_2 = \{0.007, 0.015, 0.03, -0.001, -0.02, 0.004, 0.002, -0.005\}$
(a) $\alpha_1 = 0.1$, True SIR=1.1
(b) $\alpha_1 = 0.26$, True SIR=1.3
(c) $\alpha_1 = 0.41$, True SIR=1.5
(d) $\alpha_1 = 0.69$, True SIR=2

$T \sim Bernoulli(P(T = 1))$ is held fixed across the 5,000 simulations while a different $Y \sim Poisson(E[Y])$ is simulated in each. In each scenario, we select 10 exposed CBGs to represent the regions within the community under study, which are also held fixed across simulations.

The CDC’s recommended SIR analysis is implemented by using the cancer incidence in all the SEER state CBGs outside the community of interest to compute the expected incidence (i.e., background rate). We also test a variant of the CDC’s method that does make some effort at adjustment for confounding– we fit a frequentist Poisson regression model to the data from all the CBGs, using the confounding variables as covariates, but estimating the SIR as the exponentiated parameter estimate corresponding to an indicator of inclusion in the community of interest, rather than the exposure indicator. We abbreviate this approach as PR. Finally, we implement our cSIR estimation method as described in Section 2.4 (note that here we are assuming appropriate spatial aggregation of the cancer incidence data), by first applying 20:1 mahalanobis distance nearest neighbor ratio matching and then fitting the Bayesian loglinear model. The results are given in Table 1, which provides for each method the bias in the point estimate, the coverage rate of the true SIR for 95% confidence/credible intervals, and the coverage rate of the null value of 1 for 95% confidence/credible intervals.

In all the simulations, the cSIR demonstrates small bias and coverage of the true SIR value near 95%. Appropriately, coverage of the null value (related to power) decreases as the true SIR increases. We begin to see reasonable power to detect elevated SIRs when the true SIR is between 1.5 and 2. For the two competing methods, absolute bias values frequently exceed 0.5. Coverage trends differ for CDC and PR. Due to high bias, the standard CDC approach often gives low coverage of the true SIR. Its coverage of the null value is erratic and does not reflect trends in the true SIR, i.e., coverage of the null does not consistently decrease as the true SIR value increases.
Table 1: Simulation results comparing the proposed cSIR method with the standard cancer cluster SIR estimation method (CDC) and a similar Poisson regression approach (PR). Shown are the bias in the point estimate, the coverage rate of the true SIR for 95% confidence/credible intervals, and the coverage rate of the null value of 1 for 95% confidence/credible intervals.

<table>
<thead>
<tr>
<th>Simulation</th>
<th>True SIR</th>
<th>Bias</th>
<th>Coverage</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>True SIR</td>
<td>Null</td>
<td></td>
</tr>
<tr>
<td>Simulation 1</td>
<td>True SIR=1</td>
<td>CDC -0.00</td>
<td>0.96</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PR 0.20</td>
<td>0.99</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cSIR -0.02</td>
<td>0.94</td>
<td>-</td>
</tr>
<tr>
<td>Simulation 2</td>
<td>True SIR=1</td>
<td>CDC 0.55</td>
<td>0.13</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PR -0.27</td>
<td>0.99</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cSIR -0.00</td>
<td>0.95</td>
<td>-</td>
</tr>
<tr>
<td>Simulation 3a</td>
<td>True SIR=1.1</td>
<td>CDC -0.33</td>
<td>0.43</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PR -0.06</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cSIR -0.03</td>
<td>0.95</td>
<td>0.94</td>
</tr>
<tr>
<td>Simulation 3b</td>
<td>True SIR=1.3</td>
<td>CDC -0.43</td>
<td>0.25</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PR 0.39</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cSIR -0.04</td>
<td>0.94</td>
<td>0.74</td>
</tr>
<tr>
<td>Simulation 3c</td>
<td>True SIR=1.5</td>
<td>CDC -0.54</td>
<td>0.09</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PR 1.23</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cSIR -0.05</td>
<td>0.94</td>
<td>0.39</td>
</tr>
<tr>
<td>Simulation 3d</td>
<td>True SIR=2</td>
<td>CDC -0.84</td>
<td>0.00</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PR 0.73</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cSIR -0.07</td>
<td>0.93</td>
<td>0.01</td>
</tr>
<tr>
<td>Simulation 4a</td>
<td>True SIR=1.1</td>
<td>CDC 0.56</td>
<td>0.13</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PR 0.02</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cSIR -0.01</td>
<td>0.95</td>
<td>0.88</td>
</tr>
<tr>
<td>Simulation 4b</td>
<td>True SIR=1.3</td>
<td>CDC 0.54</td>
<td>0.17</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PR 1.07</td>
<td>1.00</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cSIR -0.01</td>
<td>0.95</td>
<td>0.45</td>
</tr>
<tr>
<td>Simulation 4c</td>
<td>True SIR=1.5</td>
<td>CDC 0.51</td>
<td>0.26</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PR 3.76</td>
<td>0.99</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cSIR -0.01</td>
<td>0.94</td>
<td>0.08</td>
</tr>
<tr>
<td>Simulation 4d</td>
<td>True SIR=2</td>
<td>CDC 0.35</td>
<td>0.58</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PR 3.26</td>
<td>1.00</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cSIR -0.02</td>
<td>0.94</td>
<td>0.00</td>
</tr>
</tbody>
</table>

The frequentist PR model, on the other hand, is unstable due to the small number of samples in the community under study and this instability leads to extraordinarily wide confidence intervals and, therefore, highly conservative coverage, both of the true SIR and of the null value. These results demonstrate that our approach, based in causal inference procedures, provides increased reliability and stability in comparison to existing alternative cancer cluster investigation procedures.
5 An investigation of kidney cancer incidence in Endicott, New York

Endicott, New York was the home of the first IBM manufacturing complex. A spill of thousands of gallons of a mixture of chemicals by IBM in 1979 has plagued the town for decades. According to the NY State Department of Environmental Conservation New York State Department of Environmental Conservation (2018) (DEC), trichloroethylene (TCE), a metal degreaser and a known carcinogen, was the spilled contaminant that migrated the furthest outside the IBM plant and into the surrounding community, carried via groundwater. In 2002, an investigation mandated by the DEC discovered that the TCE that had migrated into the soil in residential areas was evaporating and the resultant vapors entering indoor air in homes at dangerous levels, a process known as vapor intrusion. How long prior to 2002 the community had been exposed to TCE vapor intrusion remains unknown. TCE exposure is known to cause kidney cancer, but evidence in human studies has also indicated that it is associated with lymphomas and childhood leukemia and liver, biliary tract, bladder, esophageal, prostate, cervical, and breast cancers (Environmental Protection Agency, 2011).

In 2006, the NY State Department of Health (DOH) conducted an investigation of cancer rates in Endicott between 1980-2000 using a SIR analysis. They found rates of kidney and testicular cancer were significantly higher than background rates (New York State Department of Health, 2006). To our knowledge, no follow up investigation has been conducted to determine whether residential exposure to the TCE vapors, detected in 2002, after the end date of the DOH study, has led to increased cancer rates in the community. Using our proposed cSIR method, and relying exclusively on publicly available data, we investigate whether TCE exposure caused increased incidence of kidney cancer in Endicott during 2005-2009.

We define the appropriate study area boundaries based on the work of these previous investigations, which determined the boundaries of the area affected by TCE vapor intrusion (New York State Department of Health, 2006). Because we wish to define the area as a set of CBGs, we choose to use in our analysis all CBGs fully or partially overlapping the exposed area. This leads to eight exposed CBGs.

TCE vapor exposure is only known to have been present in Endicott in 2002 and after, it is unclear when it began affecting the community; therefore, the appropriate time period to study is not obvious. Such uncertainties are likely to plague any cancer cluster investigation. The time period under study here, 2005-2009, was chosen primarily based on cancer data availability, and, because kidney cancer latency periods are relatively short compared to other cancers (Yuan et al., 2010), any effects of TCE vapor exposure from from the early 2000s or prior may be already be detectable during this time period.

5.1 Data

We have collected publicly available 2005-2009 county level kidney cancer incidence data from SEER and publicly available 2005-2009 CBG level kidney cancer incidence data from NY. These data are described in detail in Section 3. We obtained potential TCE exposure information for the entire US from the EPA’s publicly available Toxics Release Inventory (TRI) data (Environmental Protection Agency, 2018b) and Superfund site data (Environmental Protection Agency, 2018a). The use of over 650 toxic chemicals, including TCE, is tracked by the EPA, and business manufacturing or using more than a specified threshold amount of any one of these chemicals (and meeting certain other
criteria) is required to submit yearly release reports to the EPA. Current and historic information about the location of these businesses, as well as the chemical types and amounts used by each, is provided to the public via the TRI data. The geocoded locations of all the Superfund hazardous waste sites, many of which have been contaminated by TCE, are also available through the EPA.

We employed the TRI and Superfund site location data to create a binary indicator of potential TCE exposure, around or before the time of Endicott’s TCE vapor exposure, for each CBG in the US. We classify a CBG as potentially exposed to TCE if (a) a facility using TCE in or before 2000 or a Superfund site is/was located within its boundaries or (b) a facility using TCE in or before 2000 or a Superfund site is/was located within 2 miles of its centroid. We allow a CBG to serve as a potential matched control for the Endicott CBGs if it is classified as having no potential for exposure to TCE.

Finally, for each CBG in each state for which we have cancer incidence data, we have collected data from the US Census on the following potential confounders of the association between TCE exposure and cancer incidence (in parentheses, the names used for the remainder of the paper): percent of the population age 65+ (P65+), percent of the population male (PMale), percent of the population white (PWhite), rural indicator (Rural), percent of the adult population unemployed (Unemploy), average commute time (Commute), median household income (Income), total dollars spent on smoking products as a portion of per capita income (MoneySmoke), percent of total dollars spent on food that was spent on food outside the home (MoneyFood), percent of the population that reports exercising at least 2 times per week (Exercise), and percent of the population working in the agriculture, mining, construction, or manufacturing industries (Industry). Because confounders should precede exposure, these confounder data come from the year 2000, just prior to the time that TCE vapor exposure was detected in Endicott.

5.2 CBG kidney cancer incidence prediction and cSIR estimation

Our first step in estimating the cSIR is to identify matched control CBGs for the Endicott CBGs. Mahalanobis distance nearest neighbor matching is applied to select the five best matches for each Endicott CBG among the SEER state CBGs. 5:1 ratio matching was chosen because it provided a reasonable compromise between our desire for (1) good confounder balance and (2) a substantial number of controls to stabilize the estimation. In Table 2, we show the standardized difference in means between the exposed and controls for each confounder before and after matching (Rural, a binary variable, is not shown in the table but is matched on exactly). Although the standardized differences in means after matching are not all less than the generally recommended (but arbitrary) threshold of 0.2, the matched data are much more balanced on all variables except one than the unmatched data. Moreover, because we are also adjusting for the confounders in the model applied after matching, we are not concerned about minor deviations from perfect balance.

Due to the spatial over-aggregation of the SEER data, the next step in the analysis is to apply the model introduced in Section 3 to jointly predict the CBG kidney cancer incidence for the matched controls and fit the loglinear model for cSIR estimation. We fit the prediction model using all the CBG kidney cancer incidence data from New York and the confounder variables described above as predictors, collecting 200,000 posterior samples. The resulting parameter estimates reveal significant positive associations between each of the following variables and the proportion of the county’s cancer incidence that falls into a given CBG: MoneySmoke, P65+, PMale, PWhite, Income, and Industry. A significant negative association was found for MoneyFood. We use this model to collect posterior predictive samples of the cancer incidence for the matched control CBGs for use
Table 2: Standardized difference in means for each confounder before and after matching

<table>
<thead>
<tr>
<th>Before Matching</th>
<th>After Matching</th>
</tr>
</thead>
<tbody>
<tr>
<td>MoneyFood</td>
<td>-0.51</td>
</tr>
<tr>
<td>MoneySmoke</td>
<td>0.93</td>
</tr>
<tr>
<td>P65+</td>
<td>0.67</td>
</tr>
<tr>
<td>PMale</td>
<td>-0.20</td>
</tr>
<tr>
<td>PWhite</td>
<td>7.30</td>
</tr>
<tr>
<td>Unemploy</td>
<td>-0.22</td>
</tr>
<tr>
<td>Commute</td>
<td>-3.44</td>
</tr>
<tr>
<td>Income</td>
<td>-6.16</td>
</tr>
<tr>
<td>Industry</td>
<td>1.73</td>
</tr>
<tr>
<td>Exercise</td>
<td>-2.10</td>
</tr>
</tbody>
</table>

in the estimation stage of the model.

In the cSIR estimation model, all confounders are included besides Rural, because all CBGs in the matched dataset are non-rural. The resulting cSIR and 95% credible interval are 4.05 (2.42, 6.53), which provides evidence that TCE vapor exposure was strongly related to kidney cancer rates in Endicott between 2005 and 2009. If we are willing to make the assumptions of SUTVA, ignorability, and causal consistency described in Section 2.3, then these results are endowed with a causal interpretation. However, the assumption of no unobserved confounding, which is untestable, could be violated in this analysis. Because the Endicott, New York area was historically very industrial, the area under study (or portions of it) may be contaminated by chemicals other than TCE that are associated with increased kidney cancer risk. Our analysis does not directly adjust for exposure to other chemicals, thus, if the TCE-exposed community is also exposed to other chemicals and the matched control regions in the analysis are not, the increased kidney cancer risk could be due to these other exposures rather than TCE vapor exposure. However, the area surrounding the IBM spill site has been subjected to a great deal of environmental testing and no other chemicals have been found to be widely present in the area under study here. Thus, it is unlikely that any chemical exposure confounding would have an effect large enough to attenuate the strong TCE vapor effect observed here.

6 Discussion

In this paper, we have introduced a causal inference framework for cancer cluster analyses, which relies on a priori identification of sources of hazard that could cause increased cancer incidence. By constructing statistical analyses around exposure hypotheses rather than observed cancer outcomes, the silent multiple comparisons problem associated with the traditional approach to cancer cluster investigations is resolved so that statistically valid results are possible. Moreover, this approach allows us to directly ask and answer the question of interest—whether exposure to a specific hazard caused increased cancer incidence in a community.

Using the potential outcomes framework, we develop a causal analog of the standardized incidence ratio typically used to evaluate cancer clusters, the causal SIR, and provide identifying assumptions. We propose a two-stage Bayesian model that resolves the problem of spatial over-
aggregation in cancer incidence data. This model, applied to a matched dataset, allows the cSIR
can be estimated from publicly available data. In simulations, our statistical approach was shown
to provide dramatically improved results, i.e., less bias and better coverage, than the current
approach to SIR analyses. Finally, we demonstrated the use of our method by applying it to
investigate whether TCE vapor exposure, resulting from a chemical spill dating back the the 1970s,
statistically significant cSIR estimate suggests that TCE vapor exposure indeed caused increased
incidence, which is consistent with the known causal relationship between TCE exposure and kidney
cancer.

While these methods provide an improvement over existing methods for cancer cluster investigation,
they have only just begun to be developed and still have numerous limitations. First, these methods
need more testing and development in the setting of very rare cancer with many zeros counts in
small areas. As we discuss in Section 3, our method for addressing spatial over-aggregation relies
on a Poisson likelihood and must be extended to handle zero-inflation and extra-Poisson variation.
Additionally, more work is needed to adapt this framework to the setting in which multiple
exposures affecting a community may have synergistic affect on cancer incidence. Finally, in some
contexts, small sample size of the matched data and a potentially large number of confounders may
mean that causal methods for \( p > N \) need to be integrated into this approach.

Likely the most challenging aspects of applying these methods in real cancer cluster investigations
will be (1) determining the area and time period exposed to a given source and (2) collecting
reliable data. With regards to the former, we remark that the exposure hypotheses on which
analyses are based do not have to be perfect nor unanimously agreed upon. First, multiple
different potential exposures can be considered and analyzed (separately), i.e., a single exposure for
investigation need not be settled on from the beginning. Moreover, while some research should be
done regarding the area, time period, and cancer types reasonably associated with a given exposure,
these determinations need not be set in stone in order to proceed with statistical analyses. Different
reasonable specifications of area, time period, and cancer types could be tested and the results
multiple comparisons adjusted accordingly, using standard multiple comparisons corrections like
Bonferroni.

Obtaining reliable data to carry out these analyses is a less forgiving endeavor. While confounder
data is readily available from the census, cancer incidence and exposure data are scarce. As
described here, a few states are beginning to take the lead in public release of small area cancer
incidence data. If this movement spreads, it stands to deliver huge improvements to the efficiency
and reliability of cancer cluster investigations. Exposure data is similarly sparse, and its reliability
is often dubious. For instance, the TRI data only represents businesses using large amounts of
certain chemicals, and businesses self-report usage to the TRI database. Moreover, the TRI data
do not capture events like spills of chemicals that may put communities at highest risk. In order
to carry out cancer cluster investigations with maximal rigor, more work is needed both to collect
better data and to make the data more easily accessible.

References

treatment effects. *econometrica* 74(1), 235–267.


