A New Estimation Approach for Combining Epidemiological Data From Multiple Sources

Hui HUANG, Xiaomei MA, Rasmus WAAGEPETERSEN, Theodore R. HOLFORD, Rong WANG, Harvey RISCH, Lloyd MUELLER, and Yongtao GUAN

We propose a novel two-step procedure to combine epidemiological data obtained from diverse sources with the aim to quantify risk factors affecting the probability that an individual develops certain disease such as cancer. In the first step, we derive all possible unbiased estimating functions based on a group of cases and a group of controls each time. In the second step, we combine these estimating functions efficiently to make full use of the information contained in data. Our approach is computationally simple and flexible. We illustrate its efficacy through simulation and apply it to investigate pancreatic cancer risks based on data obtained from the Connecticut Tumor Registry, a population-based case–control study, and the Behavioral Risk Factor Surveillance System which is a state-based system of health surveys. Supplementary materials for this article are available online.

KEY WORDS: Estimating equation; Spatial epidemiology; Spatial point process.

1. INTRODUCTION

In the era of electronic records, to investigate risk factors for disease, it has become easier than ever for researchers to obtain epidemiological data from diverse sources. One widely used source of such data is the case–control study where the aim is to investigate risk factors affecting the probability that an individual develops a certain disease. Book-length treatment of this topic can be found in Schlesselman (1982). In its simplest form, a case–control study consists of the values of all risk factors under consideration for a representative sample of individuals in the study region who develop the health outcome (cases), as well as for a representative sample of individuals of the population giving rise to the cases (controls). Case–control studies are attractive because they provide information for an extensive list of risk factors, more than generally available from other data sources such as tumor registry data, which, other than patients' residential addresses, typically contain only basic demographic, diagnostic, and clinical variables. Case–control studies allow epidemiologists to perform a more comprehensive assessment of risk factors. Established statistical techniques are available for analyzing case–control data, in particular logistic regression analysis or conditional logistic regression analysis (for individually matched case–control studies).

Despite the strengths and popularity of case–control studies, their use alone can be inefficient when investigating spatial risk factors such as exposure to traffic. This is mainly because information obtained from other sources, such as tumor registry data or data from the Behavioral Risk Factor Surveillance System (BRFSS), which is a large state-based system of health surveys, may also contain important, and often richer or more precise, spatial or geographic information. Moreover, effects of spatial risk factors on cancer risk, if any, may be small. It is therefore desirable to be able to combine risk-factor data of diverse sources in a single analysis so as to increase the effective sample size.

The problem of combining epidemiological data obtained from multiple sources has been studied by several authors. For example, Prentice and Sheppard (1995) and Wakefield (2004) considered combining aggregated disease data with individual control or cohort data; Haneuse and Wakefield (2007, 2008a, 2008b) argued that the cohort-based approach of Wakefield (2004) was inefficient for investigation of rare disease outcomes such as cancer. They instead proposed a hybrid design to combine aggregated disease data with either case–control data or case data alone. Prentice and Sheppard's approach assumes significant variation in the risk factors across subpopulations over which the aggregated disease counts were made. The methods proposed in Wakefield (2004) and Haneuse and Wakefield (2007, 2008a, 2008b) concerned binary risk factors, but we may also be interested in continuous risk factors (e.g., exposure to traffic) in analyses. Diggle et al. (2010) developed procedures for combining individual-level disease data and spatially aggregated information on a population at risk. They required spatially aggregated information for all risk factors under consideration, which may not be generally possible. Best, Ickstadt, and Wolpert (2000) proposed a sophisticated parametric model to combine data at disparate spatial resolutions, where all data are related to a latent, spatially continuous stochastic process

Hui Huang is Postdoctoral Associate (E-mail: hhuanghui@gmail.com), and Yongtao Guan is Professor and corresponding author (Email: yguan@bus.miami.edu), Department of Management Science, University of Miami, Coral Gables, FL 33124. Xiaomei Ma is Associate Professor (E-mail: Xiaomei.Ma@yale.edu), Theodore Holford (E-mail: theodore.holford@yale.edu) and Harvey Risch (E-mail: harvey.risch@yale.edu) are Professor, and Rong Wang is Research Scientist (E-mail: r.wang@yale.edu), Yale School of Public Health, New Haven, CT 06520. Rasmus Waagepetersen is Professor, Department of Mathematical Sciences, Aalborg University, Fredrik Bajersvej 7G, DK-9220 Aalborg, Denmark (E-mail: rw@math.aau.dk). Lloyd Mueller is Senior Epidemiologist, Connecticut Department of Public Health, 410 Capitol Avenue, MS# 11HCQ, Hartford, CT 06134 (E-mail: Lloyd.Mueller@ct.gov). This research has been partially supported by NIH grants 1R01CA169043, 5R01ES017416, and 5R01CA098870, NSF grant DMS-0845368, the Danish Council for Independent Research—Natural Sciences grant 12-124675, "Mathematical and Statistical Analysis of Spatial Data," and the Centre for Stochastic Geometry and Advanced Bioimaging, funded by the Villum Foundation. The Connecticut Tumor Registry is supported by Contract No. HHSN261201300019I between the National Cancer Institute and State of Connecticut Department of Public Health. This study was approved by the Connecticut Department of Public Health (CDPH). Certain data used in this article were obtained from the CDPH. The authors assume full responsibility for analysis and interpretation of these data.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/jasa.

© 2014 American Statistical Association
Journal of the American Statistical Association
March 2014, Vol. 109, No. 505, Applications and Case Studies
DOI: 10.1080/01621459.2013.870904

representing unexplained spatial variation in risk. To fit the model, they used a computationally intensive Markov chain Monte Carlo (MCMC) implementation of Bayesian inference. Jackson, Best, and Richardson (2006) also used MCMC methods to fit their proposed hierarchical model for combining small area and individual-level data on exposures and health outcomes. In a social science case study, Angrist and Krueger (1992) considered instrumental variable inference for a linear model with moments for the instrumental variable method obtained from different datasets.

Multiple imputation (e.g., Rubin 2004) is a general strategy for handling inference in the presence of missing data, where missing data are imputed in complete-case estimating functions. In the context of combining different epidemiological datasets where variables from one dataset may be missing in another dataset, multiple imputation has been used, for example, by Gelman, King, and Liu (1998) and Schenker, Raghunathan, and Bondarenko (2010). One drawback of multiple imputation is the need of a joint model for all variables so that the missing quantities can be simulated given observed ones (e.g., Schafer 1999). Statistical (or file) matching (Little and Rubin 2002) is another possibility for creating complete datasets, where two incomplete datasets are merged based on a common variable appearing in both datasets. The validity of this procedure hinges on restrictive assumptions of conditional independence given the common variable used for matching. The approach in Robins, Rotnitzky, and Zhao (1994) is also based on complete-case estimating functions but removes estimating function components due to individuals with missing data. The remaining terms are then weighted with the inverse probability of missingness to retain unbiasedness.

We propose a novel two-step procedure to combine data obtained from diverse sources, with the aim to fit a parametric model to quantify risk factors affecting the probability that an individual develops certain disease such as cancer. In the first step, we derive all possible estimating functions based on data from one single source (e.g., a case–control study) or from two different sources. By considering only one or two data sources at a time, we can significantly reduce the difficulty in handling datasets with different patterns of missing covariates. In the second step, we combine all the available estimating functions efficiently to make full use of information contained in data. Our method of combining several estimating functions is related to the generalized method of moments used to combine various sources of information for microeconomic models and longitudinal surveys in respectively Imbens and Lancaster (1994) and Zhou and Kim (2012). Because our approach is constructed by forming unbiased estimating functions in terms of the risk factors, it avoids the use of the computationally intensive MCMC algorithms. In situations with missing data, we use single-imputation of the missing data in a set of modified complete-case estimating functions. Only crude estimators of the missing quantities are needed to maintain unbiasedness of the resulting estimating functions, as long as the variance of these estimators is asymptotically negligible. Our method does not assume conditional independence nor requires any precise knowledge on the correlation among the various covariates.

We organize the remainder of this article as follows. In Section 2, we describe the pancreatic cancer study that has motivated the present work. We provide necessary background in Section 3 and discuss our proposed methods in Section 4. We assess their numerical properties through simulation in Section 5. We apply the proposed methods in Section 6 to answer the substantive question, whether traffic-related exposures are related to risk of developing pancreatic cancer. We will use data obtained from a case–control study in Connecticut, the Connecticut Tumor Registry (CTR), and BRFSS. We conclude the article with a discussion in Section 7. Additional theoretical results are given in the Appendix.

2. DESCRIPTION OF THE DATA

Our core dataset was obtained from a classical case–control study for pancreatic cancer. We enrich this dataset with further cases from the CTR and further controls from BRFSS. Finally, we add traffic exposure data from the Connecticut Department of Transportation.

2.1 Case–Control Study of Pancreatic Cancer

The cases included in the case–control study were histologically and clinically validated, incident pancreatic cancer patients in Connecticut diagnosed between January 1, 2005, and August 31, 2009 (Risch et al. 2010). To identify case patients, study staff made frequent regular visits to each of the 30 general hospitals across the state. Consent to approaching patients was obtained from physicians or physician practices for 83% of 1092 potentially eligible newly diagnosed individuals aged 35–83 years at diagnosis of pancreatic cancer. Of 906 requested in-person interviews, 421 (46%) were successfully completed, after which 19 were found to be ineligible after further clinical review. The remaining patients were not able to be located or contacted ($n = 50$), were too ill or had died before study contact ($n = 333$), or refused study participation ($n = 121$). To identify potential control subjects, a pre-letter-assisted random-digit dialing (RDD) method was used over the same time frame. An address was sought for each randomly selected landline telephone number through reverse-directory lookup to mail a study letter before telephone contact for eligibility. Control subjects were frequency matched to case patients by gender and age group (35–51, 52–59, 60–64, 65–69, 70–74, 75–79, and 80–83 years). A total of 1137 potentially eligible control subjects was identified and 715 (63%) of them participated. Reasons for nonparticipation included inability to locate or contact ($n = 140$) and subject refusal ($n = 282$). All subjects were interviewed in person. At interview, participants provided signed informed consent, after which a structured questionnaire was used to collect information on a variety of potential risk factors. The study was approved by the Yale Human Investigation Committee.

2.2 The Connecticut Tumor Registry Data on Pancreatic Cancer

Connecticut is a small state geographically, yet has a dense population (about 3.5 million). The CTR is the oldest cancer registry in the United States and has been a Surveillance, Epidemiology, and End Results (SEER) program participating site since the SEER program commenced in 1973. The CTR has reciprocal reporting agreements with cancer registries in all adjacent states (and Florida, which is a popular vacation destination) to identify Connecticut residents with cancer diagnosed or treated in these states. CTR cases included in the present study fulfilled the following eligibility criteria: (1) incident cancer designated in the CTR as pancreatic, diagnosed between January 1, 2005, and August 31, 2009; (2) resident at diagnosis in the state of Connecticut; and (3) aged 35–83 years old. These criteria were set to correspond to those used in the case–control study. However, only a minority of pancreatic cancer cases in the CTR undergo rigorous research study-level validation of their primary site, thus blanket accession of CTR cases allows for some cases of cancer from other organs extending to the pancreas (e.g., Ampulla of Vater, common bile duct) or metastatic to it, to be included. The CTR subjects do include deceased cases and those not granted physician permission to be approached by the case–control study, thus their number is appreciably larger. For each CTR case, we have identified age, date of diagnosis, gender, race, Hispanic ethnic origin, and residential address at the time of diagnosis. A total of 2335 nominally pancreatic cancer patients was found (including the case–control study cases) and we have successfully geocoded the residential addresses of 2275 (97%) of them.

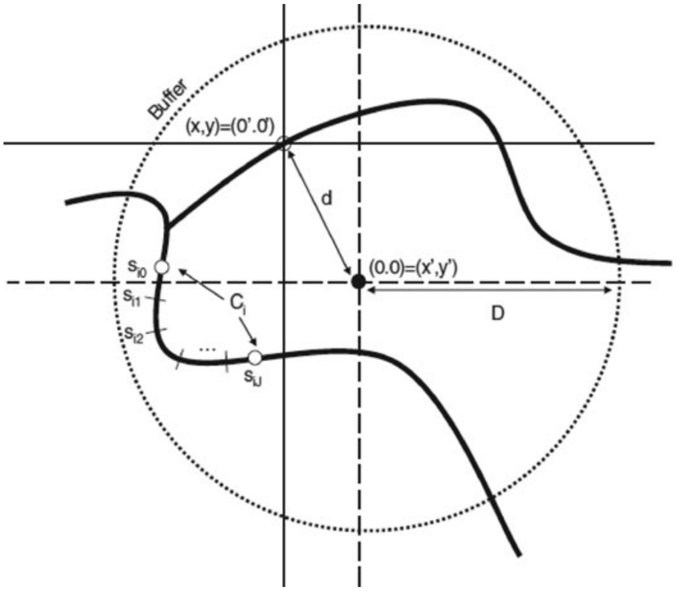

Figure 1. Calculation of the traffic exposure of a residence (bullet) along highways (bold curve) within a buffer circle with radius D.

2.3 The Behavioral Risk Factor Surveillance System Data

BRFSS is a state-based system of health surveys collecting information on health risk behaviors, preventive health practices, and health care access primarily related to chronic diseases and injury. BRFSS was established in 1984 by the Centers for Disease Control and Prevention; with more than 350,000 adults interviewed each year, it is the largest telephone health survey in the world. We have obtained the raw 2008 BRFSS survey data for Connecticut to gather information on lifestyle variables such as cigarette smoking. There were a total 6155 Connecticut residents 18 years or older who participated in the survey in 2008.

The 2008 BRFSS was conducted by using RDD to select study samples. The sampling frames between the BRFSS RDD and the case–control study RDD differed somewhat because the case–control study matched controls to the distribution of case sex and age. BRFSS also used post-survey weighting techniques to maximize the representativeness of the sampled data. The current BRFSS weighting formula, which can be found at *http://www.cdc.gov/brfss/technicalinfodata/weighting.htm*, accounts for differences in the basic probability of selecting among strata (i.e., subsets of area/prefix combinations), the number of residential telephone lines in the respondent's household, the number of adults in the household, and the age-by-sex or age-by-race-by-sex distribution in the population in general (not in the cancer cases) so as to adjust for overcoverage and nonresponse.

The BRFSS data provide extremely rich information on lifestyle variables such as cigarette smoking. Because the survey targets the general population, BRFSS subjects should also be treated as controls. However, unlike the CTR data and the case–control data, residential locations of the study participants, and consequently traffic-related exposures that are derived based on residential locations, are only available at the zip code level. In the present study, the 4,459 (72%) BRFSS participants aged 35–83 years old with available zip codes were included in the analysis.

2.4 Traffic Data

Traffic data for Connecticut are available from the state's department of transportation. We have obtained data on annualized average daily traffic (ADT) on all numbered state highways (including interstates) for the year 2007; the ADT was measured on segments of roadways where major changes in volume occur.

For a subject at residence $\mathbf{u}$, we may define exposure to traffic as an integral of the traffic volume over all points of highways within a buffer surrounding $\mathbf{u}$. In practice, however, this integral has to be estimated. The ADT estimates are given along highway segments, as shown by the highways displayed in Figure 1. The open circles therein represent nodes dividing lines into segments with a common ADT. The ith segment, $C_i, i = 1, \ldots, n$, would be specified by the starting and ending nodes. We divide C_i into short subsegments, $\mathbf{s}_{i0}, \mathbf{s}_{i1}, \mathbf{s}_{i2}, \ldots, \mathbf{s}_{iJ_i}$. In our analysis, we use

$$\sum_{i=1}^{n} \tau_i \sum_{j=0}^{J_i} I(||\mathbf{s}_{ij} - \mathbf{u}|| \le D)\Delta \mathbf{s}_{ij}, \qquad (1)$$

as the traffic exposure variable, where τ_i is the common ADT value on C_i, $I(\cdot)$ is an indicator function with $|| \cdot ||$ denoting the Euclidean distance, D is a prespecified constant used to define the buffer extent, and $\Delta \mathbf{s}_{ij}$ is the length of the jth subsegment of C_i.

2.5 Rationale for Combining Data

Motor vehicle emissions are a major source of air pollution. To date, numerous studies have been conducted to investigate whether exposure to traffic is related to risk of developing cancer (e.g., Pearson, Wachtel, and Ebi 2000; Raaschou-Nielsen et al. 2001, 2011; Reynolds et al. 2002; Visser, van Wijnen, and van Leeuwen 2004; Beelen et al. 2008). While some of these studies concluded a significant association of exposure to traffic with the risk for certain types of cancer, others were not able to detect such an association. For example, Pearson, Wachtel, and Ebi (2000) concluded significant associations between their distance-weighted traffic density metrics and all childhood cancers (including childhood leukemia), but Raaschou-Nielsen et al. (2001) found that traffic-related air pollution at the residence did not appear to cause leukemias, central nervous system tumors, or non-Hodgkins lymphomas in children. Beelen et al. (2008) found an association of exposure to traffic with lung cancer for nonsmokers but no association for ex- or current smokers in their study. They argued that the failure to detect a significant association for the latter might be because the effect was too small to measure when compared against the much stronger association between cigarette smoking and lung cancer risk.

The majority of the existing studies was based on data collected from either case–control or cohort/case cohort studies. As a result, the total numbers of incident cancers included were typically much smaller than those recorded in tumor registries. Given that the effect of traffic may be small, as argued by Beelen et al. (2008) in the case of lung cancer for ex- and current smokers, it is desirable to increase the sample size by including all available cases. However, as we pointed out earlier, an analysis based on tumor registry data alone is not useful due to the very few risk factors available. For example, Reynolds et al. (2002) estimated rate ratios for childhood cancer incidence using tumor registry data but were able to adjust only for age, sex, and race; Visser, van Wijnen, and van Leeuwen (2004) conducted a separate smoking survey to show that smoking was not confounded with traffic effect, but still they did not include any other risk factors.

Our primary interest is to study whether exposure to traffic is related to risk of developing pancreatic cancer. Raaschou-Nielsen et al. (2011) found no evidence that traffic-related air pollution increased the risk for pancreatic cancer, based on a large Danish cohort study. We have conducted a preliminary analysis based on our case–control data but also failed to detect any significant association. However, given the additional CTR data, we are interested in supplementing them to the case–control data so as to increase the sample size but without suffering the limitations with using tumor registry data alone. We wish to further include the BRFSS data and investigate whether a significant association can now be detected with these new additional data.

2.6 Definition of Risk Factors

Let $\mathbf{Z}(\mathbf{s})$ be a $p \times 1$ vector of risk factors for an individual at location $\mathbf{s}$. For ease of presentation we suppose that $\mathbf{Z}(\mathbf{s}) = [\mathbf{Z}_d(\mathbf{s})', \mathbf{Z}_l(\mathbf{s})', \mathbf{Z}_t(\mathbf{s})']'$, where $\mathbf{Z}_d(\cdot)$, $\mathbf{Z}_l(\cdot)$, and $\mathbf{Z}_t(\cdot)$ denote demographic, lifestyle, and traffic-related exposure variables, respectively. Moreover, we set the first element of $\mathbf{Z}_d(\cdot)$ to be unity. In our pancreatic cancer data application (Section 6), in addition to the value one, $\mathbf{Z}_d$ will contain age, age squared, and indicator variables for sex and race, $\mathbf{Z}_l$ will consist of indicator variables for smoking and education status, and $\mathbf{Z}_t$ will be a one-dimensional measure of traffic exposure. The availability and accuracy of the different risk factors can vary with the source of data. A typical scenario is summarized in Table 1 and will be assumed throughout the article.

Table 1. Information on risk factors in the pancreatic cancer data.

Source	$\mathbf{Z}_d$	$\mathbf{Z}_l$	$\mathbf{Z}_t$
CTR	Available	Missing	Available
BRFSS	Available	Available	Available at the zip code level
Case–control data	Available	Available	Available

In what follows, we use $\mathbf{Z}_{-d}$, $\mathbf{Z}_{-l}$, and $\mathbf{Z}_{-t}$ to denote the risk factors that are in $\mathbf{Z}$ but not in $\mathbf{Z}_d$, $\mathbf{Z}_l$, and $\mathbf{Z}_t$, respectively. Similarly, we use $\boldsymbol{\beta}_{\text{ind}}$ to denote the regression coefficients associated with $\mathbf{Z}_{\text{ind}}$, where ind $= d, l, t, -d, -l, -t$.

3. CASE–CONTROL STUDY WITH COMPLETE RISK FACTORS

In this section, we consider an estimating function for a simple yet generic setting with only one case group and one control group. We further assume that $\mathbf{Z}(\cdot)$ is entirely observed for all available cases and controls. The estimating functions to be derived in this setting will serve as the starting point for the construction of estimating functions in the more complex situations with varying patterns of incomplete covariate data, see Section 4.

Let N and M be two spatial point processes that have generated the random spatial locations of cases and controls over a geographic region, W. We assume that the control process M is an inhomogeneous spatial Poisson process with intensity $\alpha(\mathbf{s})\lambda_0(\mathbf{s})$, where $\alpha(\mathbf{s})$ is the probability for an individual at $\mathbf{s}$ to be included in the controls and $\lambda_0(\mathbf{s})$ represents a spatially varying population density. In general, there is some prior knowledge about the sampling design used to select the controls so we assume that $\alpha(\cdot)$ is known. However, we do not need any specific knowledge of $\lambda_0(\cdot)$.

We assume that conditional on a nonnegative (random) intensity measure $\Lambda(\mathbf{s})$, the case process N is also an inhomogeneous spatial Poisson process. We further assume that

$$\lambda(\mathbf{s}; \boldsymbol{\beta}) \equiv \mathrm{E}[\Lambda(\mathbf{s})] = \lambda_0(\mathbf{s}) \exp[\mathbf{Z}(\mathbf{s})'\boldsymbol{\beta}] \qquad (2)$$

for some unknown regression coefficients $\boldsymbol{\beta}$. Intuitively speaking, $\exp[\mathbf{Z}(\mathbf{s})'\boldsymbol{\beta}]$ is the probability that an individual residing at $\mathbf{s}$ develops cancer. Hence, the parameter vector $\boldsymbol{\beta}$ provides a direct interpretation on cancer risk.

To estimate $\boldsymbol{\beta}$, let $N(d\mathbf{s})$ and $M(d\mathbf{s})$ denote the number of cases and controls in an infinitesimal region $d\mathbf{s}$ containing $\mathbf{s}$. It is easy to see that

$$\mathrm{E}[N(d\mathbf{s})] = \lambda_0(\mathbf{s}) \exp[\mathbf{Z}(\mathbf{s})'\boldsymbol{\beta}] d\mathbf{s} \quad \text{and} \quad \mathrm{E}[M(d\mathbf{s})] = \alpha(\mathbf{s})\lambda_0(\mathbf{s}) d\mathbf{s}.$$

Hence, $\mathrm{E}[\Delta(d\mathbf{s}; \boldsymbol{\beta})] = 0$, where

$$\Delta(d\mathbf{s}; \boldsymbol{\beta}) = N(d\mathbf{s}) - \frac{\exp[\mathbf{Z}(\mathbf{s})'\boldsymbol{\beta}]}{\alpha(\mathbf{s})} M(d\mathbf{s}). \qquad (3)$$

Following well-established theories on estimating equations (Crowder 1986), a consistent estimator for $\boldsymbol{\beta}$ can be obtained by solving the following estimating equations:

$$\int_W \mathbf{h}(\mathbf{s}; \boldsymbol{\beta}) \Delta(d\mathbf{s}; \boldsymbol{\beta}) = \mathbf{0}_p, \qquad (4)$$

where $\mathbf{h}(\mathbf{s};\boldsymbol{\beta})$ is a $p \times 1$ real vector-valued function of $\boldsymbol{\beta}$ and $\mathbf{0}_p$ is a $p \times 1$ vector of zeros. If we set

$$\mathbf{h}(\mathbf{s};\boldsymbol{\beta}) = \mathbf{Z}(\mathbf{s})\frac{\alpha(\mathbf{s})}{\alpha(\mathbf{s}) + \exp[\mathbf{Z}(\mathbf{s})'\boldsymbol{\beta}]}, \tag{5}$$

then Equation (4), using sum notation, can be written as

$$\mathbf{U}(\boldsymbol{\beta}) \equiv \sum_{\mathbf{s}\in(N\cap W)} \mathbf{Z}(\mathbf{s})\frac{\alpha(\mathbf{s})}{\alpha(\mathbf{s}) + \exp[\mathbf{Z}(\mathbf{s})'\boldsymbol{\beta}]} - \sum_{\mathbf{s}\in(M\cap W)} \mathbf{Z}(\mathbf{s})\frac{\exp[\mathbf{Z}(\mathbf{s})'\boldsymbol{\beta}]}{\alpha(\mathbf{s}) + \exp[\mathbf{Z}(\mathbf{s})'\boldsymbol{\beta}]} = \mathbf{0}_p. \tag{6}$$

Note that $\mathbf{U}(\boldsymbol{\beta})$ in (6) is essentially the score function of the conditional likelihood proposed by Diggle and Rowlingson (1994). In the case where N and M are both Poisson processes, (5) is optimal in the sense of minimizing variance of parameter estimates (Rathbun 2012). Other forms of $\mathbf{h}(\cdot)$ can be used but we will consider only (5) here, as our proposed methods can be easily generalized to the new settings.

4. DATA FROM MULTIPLE SOURCES WITH POTENTIAL SELECTION BIAS AND INCOMPLETE RISK FACTORS

Let N_1 and M_1 be two spatial point processes that generated the locations of cases and controls in our case–control study. Let N_2 and M_2 denote two other spatial point processes for the additional cases and controls in the CTR and BRFSS data. We assume that M_1 and M_2 are independent of each other and also of N_1 and N_2. We use $\alpha_1(\cdot)$ and $\alpha_2(\cdot)$ to denote the counterpart sampling probabilities to $\alpha(\cdot)$ defined in Section 3 for M_1 and M_2.

The first problem that we need to handle is selection bias when N_1 is not a simple random sample from $N_c = N_1 \cup N_2$. In a typical case–control study, not all contacted cases have equal likelihood to participate. If the participation rates are correlated with any of the risk factors, estimates of effects of those risk factors may be biased. The second problem is how to adapt the methodology in Section 3 to obtain unbiased estimating functions based on pairs of case and control datasets with potentially incomplete covariate data. Finally, we need to derive a method to combine efficiently the estimating functions obtained for the different pairs.

4.1 Handling Selection Bias

Suppose that N_1 is a sample from N_c where the probability for a case to be included in N_1 varies with the case's own characteristics. Specifically, we assume that the probability $\pi(\mathbf{s})$ of including a case $\mathbf{s}$ in N_1 is of the form

$$\pi(\mathbf{s};\boldsymbol{\eta}) = \frac{\exp[\mathbf{Y}(\mathbf{s})'\boldsymbol{\eta}]}{1 + \exp\,[\mathbf{Y}(\mathbf{s})'\boldsymbol{\eta}]}, \tag{7}$$

where $\boldsymbol{\eta}$ is an unknown vector of parameters and $\mathbf{Y}(\mathbf{s})$ is a vector of covariates assumed to be known for all diseased subjects. For example, we may set $\mathbf{Y}(\cdot)$ as a combination of the demographic variables and the traffic-related exposure variables. We perform a standard logistic regression analysis to estimate $\boldsymbol{\eta}$. In what follows, let $\hat{\boldsymbol{\eta}}$ and $\boldsymbol{\eta}_0$ denote the resulting estimator and the true value of $\boldsymbol{\eta}$, respectively.

4.2 Deriving Estimating Functions for the Pancreatic Cancer Data

We derive the estimating functions in our setting based on all possible pairs of case and control processes, that is, (N_1, M_1), (N_1, M_2), (N_2, M_1), and (N_2, M_2). For the pair (N_1, M_1), $\mathbf{Z}(\cdot)$ is fully observed. We may therefore modify $\mathbf{U}(\boldsymbol{\beta})$ defined in (6) as

$$\mathbf{U}_{11}(\boldsymbol{\beta},\hat{\boldsymbol{\eta}}) \equiv \sum_{\mathbf{s}\in(N_1\cap W)} \mathbf{Z}(\mathbf{s})\frac{\alpha_1(\mathbf{s})}{\alpha_1(\mathbf{s}) + \pi(\mathbf{s};\hat{\boldsymbol{\eta}})\exp[\mathbf{Z}(\mathbf{s})'\boldsymbol{\beta}]} - \sum_{\mathbf{s}\in(M_1\cap W)} \mathbf{Z}(\mathbf{s})\frac{\pi(\mathbf{s};\hat{\boldsymbol{\eta}})\exp[\mathbf{Z}(\mathbf{s})'\boldsymbol{\beta}]}{\alpha_1(\mathbf{s}) + \pi(\mathbf{s};\hat{\boldsymbol{\eta}})\exp[\mathbf{Z}(\mathbf{s})'\boldsymbol{\beta}]}. \tag{8}$$

Here $\alpha_1(\cdot)$ is used because $\mathrm{E}[M_1(d\mathbf{s})] = \alpha_1(\mathbf{s})\lambda_0(\mathbf{s})d\mathbf{s}$ and the additional term $\pi(\cdot)$ is used because $\mathrm{E}[N_1(d\mathbf{s})] = \pi(\mathbf{s};\boldsymbol{\eta}_0)\mathrm{E}[N_c(d\mathbf{s})]$.

To derive the estimating functions for the remaining pairs, let $\hat{\mathbf{Z}}_t(\cdot)$ and $\hat{\mathbf{Z}}_l(\cdot)$ be some estimators for the traffic-related exposure variables in $\mathbf{Z}_t(\cdot)$ and the lifestyle variables in $\mathbf{Z}_l(\cdot)$, respectively. It is not necessary for either $\hat{\mathbf{Z}}_t(\cdot)$ or $\hat{\mathbf{Z}}_l(\cdot)$ to be consistent estimators, but we require their variances to be asymptotically negligible. For example, we may set $\hat{\mathbf{Z}}_t(\mathbf{s}) = \mathbf{Z}_t[\mathbf{c}(\mathbf{s})]$, where $\mathbf{c}(\mathbf{s})$ is the centroid of the zip code containing $\mathbf{s}$, hence $\hat{\mathbf{Z}}_t(\cdot)$ is nonrandom. However, $\hat{\mathbf{Z}}_t(\cdot)$ still varies spatially and contains information about one's exposure level. We will describe strategies to estimate $\mathbf{Z}_l(\cdot)$ in the simulation and real data analysis.

For the pair (N_1, M_2), note that the traffic-related exposure variables in $\mathbf{Z}_t(\cdot)$ are missing for the BRFSS data (i.e., M_2). Let $\hat{\mathbf{Z}}_{12}(\cdot) = [\mathbf{Z}_d(\cdot)', \mathbf{Z}_l(\cdot)', \hat{\mathbf{Z}}_t(\cdot)]'$ be an estimate for the complete set of risk factors $\mathbf{Z}(\cdot)$. We consider

$$\begin{aligned}\mathrm{U}_{12}(\boldsymbol{\beta},\hat{\boldsymbol{\eta}}) &= \sum_{\mathbf{s}\in(N_1\cap W)} \hat{\mathbf{Z}}_{12}(\mathbf{s})\frac{\alpha_2(\mathbf{s})}{\alpha_2(\mathbf{s}) + \pi(\mathbf{s};\hat{\boldsymbol{\eta}})\exp[\hat{\mathbf{Z}}_{12}(\mathbf{s})'\boldsymbol{\beta}]}\frac{\exp[\hat{\mathbf{Z}}_t(\mathbf{s})'\boldsymbol{\beta}_t]}{\exp[\mathbf{Z}_t(\mathbf{s})'\boldsymbol{\beta}_t]} \\ &\quad - \sum_{\mathbf{s}\in(M_2\cap W)} \hat{\mathbf{Z}}_{12}(\mathbf{s})\frac{\pi(\mathbf{s};\hat{\boldsymbol{\eta}})\exp[\mathbf{Z}(\mathbf{s})'\boldsymbol{\beta}]}{\alpha_2(\mathbf{s}) + \pi(\mathbf{s};\hat{\boldsymbol{\eta}})\exp[\hat{\mathbf{Z}}_{12}(\mathbf{s})'\boldsymbol{\beta}]}\frac{\exp[\hat{\mathbf{Z}}_t(\mathbf{s})'\boldsymbol{\beta}_t]}{\exp[\mathbf{Z}_t(\mathbf{s})'\boldsymbol{\beta}_t]}.\end{aligned} \tag{9}$$

Compared to $\mathbf{U}(\boldsymbol{\beta})$ defined in (6), $\hat{\mathbf{Z}}_{12}(\cdot)$ is used instead of $\mathbf{Z}(\cdot)$ due to the missing data on $\mathbf{Z}_t(\cdot)$ in M_2. Similar to the derivation of $\mathbf{U}_{11}$, $\alpha_2(\cdot)$ is used because $\mathrm{E}[M_2(d\mathbf{s})] = \alpha_2(\mathbf{s})\lambda_0(\mathbf{s})d\mathbf{s}$ and $\pi(\cdot)$ is used because $\mathrm{E}[N_1(d\mathbf{s})] = \pi(\mathbf{s};\boldsymbol{\eta}_0)\mathrm{E}[N_c(d\mathbf{s})]$. By multiplying the additional term $\exp[\hat{\mathbf{Z}}_t(\mathbf{s})'\boldsymbol{\beta}_t]/\exp[\mathbf{Z}_t(\mathbf{s})'\boldsymbol{\beta}_t]$, we see that

$$\exp[\mathbf{Z}(\mathbf{s})'\boldsymbol{\beta}]\frac{\exp[\hat{\mathbf{Z}}_t(\mathbf{s})'\boldsymbol{\beta}_t]}{\exp[\mathbf{Z}_t(\mathbf{s})'\boldsymbol{\beta}_t]} = \exp[\hat{\mathbf{Z}}_{12}(\mathbf{s})'\boldsymbol{\beta}],$$

which can be calculated for the control subjects in M_2 even though the traffic-related exposure variables in $\mathbf{Z}_t$ are missing for these subjects.

For the pair (N_2, M_1), note that the lifestyle variables in $\mathbf{Z}_l(\cdot)$ are missing for cases in the CTR but not in the case–control study (i.e., N_2). Let $\hat{\mathbf{Z}}_{21}(\mathbf{s}) = [\mathbf{Z}_d(\cdot)', \hat{\mathbf{Z}}_l(\cdot)', \mathbf{Z}_t(\cdot)]'$ be an estimate for the complete set of risk factors in $\mathbf{Z}(\cdot)$. We consider

$$\begin{aligned}\mathbf{U}_{21}(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}}) &= \sum_{\mathbf{s}\in(N_2\cap W)} \hat{\mathbf{Z}}^*_{21}(\mathbf{s}) \frac{\alpha_1(\mathbf{s})}{\alpha_1(\mathbf{s}) + [1-\pi(\mathbf{s};\hat{\boldsymbol{\eta}})]\exp[\hat{\mathbf{Z}}_{21}(\mathbf{s})'\boldsymbol{\beta}]} \\ &\quad - \sum_{\mathbf{s}\in(M_1\cap W)} \hat{\mathbf{Z}}^*_{21}(\mathbf{s}) \frac{[1-\pi(\mathbf{s};\hat{\boldsymbol{\eta}})]\exp[\mathbf{Z}(\mathbf{s})'\boldsymbol{\beta}]}{\alpha_1(\mathbf{s}) + [1-\pi(\mathbf{s};\hat{\boldsymbol{\eta}})]\exp[\hat{\mathbf{Z}}_{21}(\mathbf{s})'\boldsymbol{\beta}]},\end{aligned} \tag{10}$$

where $\hat{\mathbf{Z}}^*_{21}(\cdot)$ is either $\hat{\mathbf{Z}}_{21}(\cdot)$ or a subset of it obtained by removing component(s) of $\hat{\mathbf{Z}}_l(\cdot)$ that are highly correlated with other components of $\mathbf{Z}(\cdot)$. Such a scenario may occur if any component of $\hat{\mathbf{Z}}_l(\cdot)$ is a linear combination of components in $\mathbf{Z}_d(\cdot)$. In particular, if $\hat{\mathbf{Z}}_l(\cdot)$ is constant, then it is linear in $\mathbf{Z}_d(\cdot)$ since the first element of $\mathbf{Z}_d(\cdot)$ is always one. Nevertheless, new information on $\mathbf{Z}_d(\cdot)$ and $\mathbf{Z}_t(\cdot)$ is still obtained by including the new data source N_2. Compared to $\mathbf{U}(\boldsymbol{\beta})$ defined in (6), $\hat{\mathbf{Z}}_{21}(\cdot)$ and $\hat{\mathbf{Z}}^*_{21}(\cdot)$ are used instead of $\mathbf{Z}(\cdot)$ due to the missing data on $\mathbf{Z}_l(\cdot)$ in N_2. Similar to the derivation of $\mathbf{U}_{11}$, $\alpha_1(\cdot)$ is used because $\mathrm{E}[M_1(d\mathbf{s})] = \alpha_1(\mathbf{s})\lambda_0(\mathbf{s})d\mathbf{s}$ and $[1-\pi(\cdot)]$ is used because $\mathrm{E}[N_2(d\mathbf{s})] = [1-\pi(\mathbf{s};\boldsymbol{\eta}_0)]\mathrm{E}[N_c(d\mathbf{s})]$.

For the last pair (N_2, M_2), note that the traffic-related exposure variables in $\mathbf{Z}_t(\cdot)$ are missing for the BRFSS data (i.e., M_2) and the lifestyle variables in $\mathbf{Z}_l(\cdot)$ are missing cases in the CTR but not in the case–control study (i.e., N_2), respectively. Define $\hat{\mathbf{Z}}_{22}(\mathbf{s}) = [\mathbf{Z}_d(\cdot)', \hat{\mathbf{Z}}_l(\cdot)', \hat{\mathbf{Z}}_t(\cdot)]'$. In light of the previous derivations of $\mathbf{U}_{12}(\cdot)$ and $\mathbf{U}_{21}(\cdot)$, we consider

$$\begin{aligned}\mathbf{U}_{22}(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}}) &= \sum_{\mathbf{s}\in(N_2\cap W)} \hat{\mathbf{Z}}^*_{22}(\mathbf{s}) \frac{\alpha_2(\mathbf{s})}{\alpha_2(\mathbf{s}) + [1-\pi(\mathbf{s};\hat{\boldsymbol{\eta}})]\exp[\hat{\mathbf{Z}}_{22}(\mathbf{s})'\boldsymbol{\beta}]} \\ &\quad \times \frac{\exp[\hat{\mathbf{Z}}_t(\mathbf{s})'\boldsymbol{\beta}_t]}{\exp[\mathbf{Z}_t(\mathbf{s})'\boldsymbol{\beta}_t]} \\ &\quad - \sum_{\mathbf{s}\in(M_2\cap W)} \hat{\mathbf{Z}}^*_{22}(\mathbf{s}) \frac{[1-\pi(\mathbf{s};\hat{\boldsymbol{\eta}})]\exp[\hat{\mathbf{Z}}_{12}(\mathbf{s})'\boldsymbol{\beta}]}{\alpha_2(\mathbf{s}) + [1-\pi(\mathbf{s};\hat{\boldsymbol{\eta}})]\exp[\hat{\mathbf{Z}}_{22}(\mathbf{s})'\boldsymbol{\beta}]},\end{aligned} \tag{11}$$

where similar to $\hat{\mathbf{Z}}^*_{21}(\cdot)$ used in (10), $\hat{\mathbf{Z}}^*_{22}(\cdot)$ is either $\hat{\mathbf{Z}}_{22}(\cdot)$ or a subset of it.

4.3 Combining Estimating Functions for the Pancreatic Cancer Data

Define $\mathbf{U}_c(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}}) = [\mathbf{U}_{11}(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}})', \mathbf{U}_{12}(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}})', \mathbf{U}_{21}(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}})', \mathbf{U}_{22}(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}})']'$. Note that $\mathbf{U}_c(\cdot)$ is asymptotically unbiased for zero if the variances of $\hat{\mathbf{Z}}_t(\cdot)$, $\hat{\mathbf{Z}}_l(\cdot)$ and $\hat{\boldsymbol{\eta}}$ are all asymptotically negligible. Define

$$\mathbf{D}(\boldsymbol{\beta}) = \mathrm{E}\left[\frac{\partial \mathbf{U}_c(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}})}{\partial \boldsymbol{\beta}}\right] \quad \text{and} \quad \mathbf{V}(\boldsymbol{\beta}) = \mathrm{var}\left[\mathbf{U}_c(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}})\right].$$

In Web Appendix A, we derive consistent estimators for all components of $\mathbf{D}(\boldsymbol{\beta})$ under $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, where $\boldsymbol{\beta}_0$ denotes the true value of $\boldsymbol{\beta}$. In Web Appendix B, we derive theoretical expressions for $\mathbf{V}(\boldsymbol{\beta})$, which involve the extra variability caused by $\hat{\boldsymbol{\eta}}$, as well as consistent estimators for all its components.

To combine the derived estimating functions, one strategy is to follow Heyde (1997) to consider

$$\tilde{\mathbf{U}}(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}}) = \mathbf{D}(\boldsymbol{\beta})'\mathbf{V}(\boldsymbol{\beta})^{-1}\mathbf{U}_c(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}}). \tag{12}$$

Solving $\tilde{\mathbf{U}}(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}}) = 0$ with respect to $\boldsymbol{\beta}$ corresponds to minimizing the generalized least-squares criterion $\mathbf{U}_c(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}})^{-1}\mathbf{V}(\boldsymbol{\beta})^{-1}$ $\mathbf{U}_c(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}})$ except that we replace the derivative $\frac{\partial \mathbf{U}_c(\boldsymbol{\beta},\hat{\boldsymbol{\eta}})}{\partial \boldsymbol{\beta}}$ with its expectation. Hence (12) is closely related to the generalized method of moments (Hansen 1982). However, for this approach, $\mathbf{V}(\cdot)^{-1}$ may be difficult to evaluate because (1) the estimating functions in $\mathbf{U}_c(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}})$ tend to be highly correlated with each other and as a result $\mathbf{V}(\cdot)^{-1}$ may be unstable and (2) the dimension of $\mathbf{V}(\cdot)$ can be high depending on the numbers of risk factors and pairs of processes to be considered. To mitigate this problem, we instead combine these estimating functions sequentially in three steps.

Step 1. Define $\mathbf{U}_0(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}}) = [\mathbf{U}_{11}(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}})', \mathbf{U}_{12}(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}})']'$. We follow (12) to consider

$$\tilde{\mathbf{U}}_1(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}}) = \mathbf{D}_0(\boldsymbol{\beta})'\mathbf{V}_0(\boldsymbol{\beta})^{-1}\mathbf{U}_0(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}}),$$

where $\mathbf{D}_0(\cdot)$ and $\mathbf{V}_0(\cdot)$ are submatrices of $\mathbf{D}(\cdot)$ and $\mathbf{V}(\cdot)$ that are associated with $\mathbf{U}_{11}(\cdot)$ and $\mathbf{U}_{12}(\cdot)$.

Step 2. Define $\mathbf{U}_1(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}}) = [\tilde{\mathbf{U}}_1(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}})', \mathbf{U}_{21}(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}})']'$. We follow (12) again to consider

$$\tilde{\mathbf{U}}_2(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}}) = \mathbf{D}_1(\boldsymbol{\beta})'\mathbf{V}_1(\boldsymbol{\beta})^{-1}\mathbf{U}_1(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}}),$$

where

$$\mathbf{D}_1(\boldsymbol{\beta}) = \mathrm{E}\left[\frac{\partial \mathbf{U}_1(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}})}{\partial \boldsymbol{\beta}}\right] \text{ and } \mathbf{V}_1(\boldsymbol{\beta}) = \mathrm{Var}\left[\mathbf{U}_1(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}})\right].$$

In Web Appendix C, we show that $\mathbf{D}_1(\cdot)$ and $\mathbf{V}_1(\cdot)$ can be written in terms of the submatrices of $\mathbf{D}(\cdot)$ and $\mathbf{V}(\cdot)$ that are associated with $\mathbf{U}_{11}(\cdot)$, $\mathbf{U}_{12}(\cdot)$, and $\mathbf{U}_{21}(\cdot)$.

Step 3. Define $\mathbf{U}_2(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}}) = [\tilde{\mathbf{U}}_2(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}})', \mathbf{U}_{22}(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}})']'$. Similar to Step 2, we now follow (12) to derive $\tilde{\mathbf{U}}_3(\cdot)$; see Web Appendix C for details. Solve $\tilde{\mathbf{U}}_3(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}}) = \mathbf{0}_p$ to estimate $\boldsymbol{\beta}$.

The covariance matrices involved in all the above steps are at most $2p \times 2p$ in dimension so it is not difficult to calculate their inverses. Moreover, the proposed procedure can be easily modified to include any additional estimating functions in the analysis. Nevertheless, to apply the method we would need to decide the order in which the estimating functions are sequentially combined. The order that we considered above is natural because the level of completeness in terms of the covariates drops in the order of $\mathbf{U}_{11}(\cdot)$, $\mathbf{U}_{12}(\cdot)$, $\mathbf{U}_{21}(\cdot)$, and $\mathbf{U}_{22}(\cdot)$. Our own simulation experience also suggests that the actual order has only limited impact on the performance of the resulting estimators.

We finally note that instead of plugging in an estimate for $\boldsymbol{\eta}$, one could alternatively follow Zhou and Kim (2012) and include the logistic regression score for estimating $\boldsymbol{\eta}$ in $\mathbf{U}_c$ as yet another estimating function so that $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ can be estimated simultaneously. However, this would add to the already considerable computational complexity and we prefer to stick with the simpler plug-in approach.

4.4 Model Diagnostics

Our model diagnostics consist of two tasks. The first is to evaluate the validity of the assumed intensity model (2), and the second is to assess whether N_c is Poisson.

4.4.1 Evaluation of the Intensity Model. We develop residual diagnostic tools based on the complete CTR data (i.e., N_c) and the BRFSS data (i.e., M_2) to check the validity of the assumed intensity model (2). Let $X(\cdot)$ be a statistic derived from of $\mathbf{Z}_d(\cdot)$. For example, $X(\cdot)$ can be age, which is included in $\mathbf{Z}_d(\cdot)$, or is age in a given sex-by-race category. Define the cumulative residuals

$$Q(x) = \sum_{\mathbf{s}\in(N_c\cap W)} \frac{\exp[\hat{\mathbf{Z}}_t(\mathbf{s})'\hat{\boldsymbol{\beta}}_t]}{\exp[\mathbf{Z}_t(\mathbf{s})'\hat{\boldsymbol{\beta}}_t]} I[X(\mathbf{s})\le x] - \sum_{\mathbf{s}\in(M_2\cap W)} \frac{\exp[\hat{\mathbf{Z}}_t(\mathbf{s})'\hat{\boldsymbol{\beta}}_t + \mathbf{Z}_{-t}(\mathbf{s})'\hat{\boldsymbol{\beta}}_{-t}]}{\alpha_2(\mathbf{s})} I[X(\mathbf{s})\le x],$$

where $\hat{\boldsymbol{\beta}}_t$ and $\hat{\boldsymbol{\beta}}_{-t}$ are the estimates of $\boldsymbol{\beta}_t$ and $\boldsymbol{\beta}_{-t}$ from our proposed sequential approach, respectively. We plot $Q(x)$ against x to assess the overall fit. If our proposed model fits the data well, then $Q(x)$ should be close to zero for all x. For inference, we use bootstrap to construct confidence bands for $Q(x)$. To do so, we keep the memberships of the case subjects (i.e., whether they belong to N_1 or N_2) unchanged from the original data.

Let $W_l : l = 1, \ldots, L$ be a partition of W. In our analysis, W_l's are the complete set of zip code regions in Connecticut. We also consider the residuals

$$R_l = \sum_{\mathbf{s}\in(N_c\cap W_l)} \frac{\exp[\hat{\mathbf{Z}}_t(\mathbf{s})'\hat{\boldsymbol{\beta}}_t]}{\exp[\mathbf{Z}_t(\mathbf{s})'\hat{\boldsymbol{\beta}}_t]} - \sum_{\mathbf{s}\in(M_2\cap W_l)} \frac{\exp[\hat{\mathbf{Z}}_t(\mathbf{s})'\hat{\boldsymbol{\beta}}_t + \mathbf{Z}_{-t}(\mathbf{s})'\hat{\boldsymbol{\beta}}_{-t}]}{\alpha_2(\mathbf{s})}.$$

We plot R_l against components of $\hat{\mathbf{Z}}_{t,l}$ to assess whether the assumed functional forms of the traffic-related exposure variables $\mathbf{Z}_t(\cdot)$ are appropriate, where $\hat{\mathbf{Z}}_{t,l}$ are the traffic-related exposure variables derived at the centroid of the lth zip code. If the proposed model fits the data well, then R_l should vary randomly across zero.

4.4.2 Detecting Non-Poisson Behavior. To detect non-Poisson behavior, we follow Diggle et al. (2007) to consider second-order statistics related to the K-function. Specifically, for any positive number r, define

$$K_1(r) = \sum_{\mathbf{s},\mathbf{u}\in(N_c\cap W)}^{\neq}\!\!\sum \frac{I(||\mathbf{s}-\mathbf{u}|| \le r)}{\rho(\mathbf{s};\hat{\boldsymbol{\beta}})\rho(\mathbf{u};\hat{\boldsymbol{\beta}})},$$

where $\neq$ signifies summation over distinct events and

$$\rho(\mathbf{s};\hat{\boldsymbol{\beta}}) = \begin{cases} \exp[\mathbf{Z}(\mathbf{s})'\hat{\boldsymbol{\beta}}] & \text{if } \mathbf{s}\in N_1; \\ \exp[\mathbf{Z}_{-l}(\mathbf{s})'\hat{\boldsymbol{\beta}}_{-l} + \hat{\mathbf{Z}}_l(\mathbf{s})'\hat{\boldsymbol{\beta}}_l] & \text{if } \mathbf{s}\in N_2. \end{cases}$$

Under the Poisson assumption, the second-order intensity function $\lambda_2(\mathbf{s},\mathbf{u}) = \lambda(\mathbf{s};\boldsymbol{\beta})\lambda(\mathbf{u};\boldsymbol{\beta})$. Then by a direct application of Campbell's Theorem (Møller and Waagepetersen 2004) and ignoring some higher-order terms, we can show that for $\mathbf{s}\neq\mathbf{u}$,

$$\mathrm{E}\{N_c(d\mathbf{s})N_c(d\mathbf{u})/[\rho(\mathbf{s};\hat{\boldsymbol{\beta}})\rho(\mathbf{u};\hat{\boldsymbol{\beta}})]\} \approx \lambda_0(\mathbf{s})\lambda_0(\mathbf{u})\psi(\mathbf{s},\mathbf{u};\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\eta}})d\mathbf{s}d\mathbf{u}, \qquad (13)$$

where

$$\begin{aligned}\psi(\mathbf{s},\mathbf{u};\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\eta}}) &= \pi(\mathbf{s};\hat{\boldsymbol{\eta}})\pi(\mathbf{u};\hat{\boldsymbol{\eta}}) + 2\pi(\mathbf{s};\hat{\boldsymbol{\eta}})[1-\pi(\mathbf{u};\hat{\boldsymbol{\eta}})] \\ &\quad \times \exp\{[\mathbf{Z}_l(\mathbf{s}) - \hat{\mathbf{Z}}_l(\mathbf{s})]'\hat{\boldsymbol{\beta}}_l\} + [1-\pi(\mathbf{s};\hat{\boldsymbol{\eta}})][1-\pi(\mathbf{u};\hat{\boldsymbol{\eta}})] \\ &\quad \times \exp\{[\mathbf{Z}_l(\mathbf{s})-\hat{\mathbf{Z}}_l(\mathbf{s})]'\hat{\boldsymbol{\beta}}_l + [\mathbf{Z}_l(\mathbf{u})-\hat{\mathbf{Z}}_l(\mathbf{u})]'\hat{\boldsymbol{\beta}}_l\}.\end{aligned}$$

In terms of the control process M_1, we may then define

$$K_2(r) = \sum_{\mathbf{s},\mathbf{u}\in(M_1\cap W)}^{\neq}\!\!\sum \frac{I(||\mathbf{s}-\mathbf{u}|| \le r)\psi(\mathbf{s},\mathbf{u};\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\eta}})}{\alpha_1(\mathbf{s})\alpha_1(\mathbf{u})}.$$

Define $T(r) = K_1(r)/K_2(r)$. By (13), $T(r)$ should be close to one for a given $r > 0$ under the Poisson assumption. We again use bootstrap to construct confidence bands for $T(r)$. As before, we keep the memberships of the case subjects unchanged, but we update the sampling probabilities $\alpha_1(\cdot)$ based on the resampled control data.

5. SIMULATION STUDY

5.1 Simulation Design

We conduct a simulation study to assess the performance of our proposed estimation method. Specifically, we generate realizations of cases and controls from inhomogeneous spatial Poisson processes over a 1 × 1 square region W. The intensity functions for the control processes are set to be $\alpha_i(\mathbf{s})\lambda_0(\mathbf{s})$ for $i = 1, 2$, where $\alpha_1(\mathbf{s}) = 0.0007$, $\alpha_2(\mathbf{s}) = 0.0045$, and $\lambda_0(\mathbf{s}) = 1{,}000{,}000$ for all $\mathbf{s}\in W$. The expected numbers of control events are therefore equal to 700 for M_1 and 4500 for M_2. The constant probabilities $\alpha_1(\cdot)$ and $\alpha_2(\cdot)$ imply that all subjects in the population have the same probability to be selected in a control process.

The intensity function for the overall case process is $\lambda(\mathbf{s};\boldsymbol{\beta}) = \lambda_0(\mathbf{s})\exp[\mathbf{Z}(\mathbf{s})'\boldsymbol{\beta}]$, where $\boldsymbol{\beta} = (\beta_0, \beta_d, \beta_l, \beta_t)' = (-6.5655, 1, 1, 1)'$, $\mathbf{Z}(\mathbf{s}) = [1, Z_d(\mathbf{s}), Z_l(\mathbf{s}), Z_t(\mathbf{s})]$, and $Z_d(\cdot)$, $Z_l(\cdot)$ and $Z_t(\cdot)$ denote demographic, lifestyle, and traffic exposure variables, respectively. We generate $\mathbf{Z}(\cdot)$ over a 100 × 100 pixel grid and assume that its value is constant over each pixel. Specifically,

- $Z_d(\mathbf{s})$'s are independent and identically distributed (iid) normal random variables with mean 0 and standard deviation 0.5;
- $Z_l(\mathbf{s}) = 0.2 * Z_d(\mathbf{s}) + 0.2 * \epsilon(\mathbf{s}) + e(\mathbf{s})$, where $\epsilon(\mathbf{s})$'s are a realization of a zero-mean and unit-variance Gaussian random field with an exponential covariance function $\exp(-20r)$ with r being the lag distance, and $e(\mathbf{s})$'s are iid normal random variables with mean 0 and standard deviation 0.4.
- $Z_t(\mathbf{s})$'s are a realization of a zero-mean and unit-variance Gaussian random field with an exponential covariance function $\exp(-10r)$.

The above construction yields a correlation of 0.2 between the lifestyle variable $Z_l(\mathbf{s})$ and the demographic variable $Z_d(\mathbf{s})$. We assign a case to be in N_1 randomly with the probability given by (7), with $\mathbf{Y}(\mathbf{s}) = [1, Z_d(\mathbf{s})]'$ and $\boldsymbol{\eta} = [-1.75, 0.5]'$. The remaining cases are included in N_2. Based on the simulation setup, the expected number of cases is 400 in N_1 and 1900 in N_2. Note that the expected numbers of cases and controls are similar to their counterparts in the real data, see Section 2. We assume that $Z_d(\cdot)$ is observed for all cases and controls, but $Z_l(\cdot)$ and $Z_t(\cdot)$ are missing for N_2 and M_2, respectively. This setup mimics that being described in Section 2.6.

To implement our proposed method, it is necessary to obtain $\hat{Z}_l(\cdot)$ and $\hat{Z}_t(\cdot)$. For the former, we first fit the regression model $E[Z_l(\mathbf{s})] = \theta_0 + \theta_1 Z_d(\mathbf{s})$ based on the data given in M_1, and then set the predicted value given $Z_d(\cdot)$ as $\hat{Z}_l(\cdot)$. For the latter, we divide W into 10×10 nonoverlapping equal square subblocks and define $\hat{Z}_t(\mathbf{s})$ as the average of $Z_t(\cdot)$ for the subblock that contains $\mathbf{s}$. Since $\hat{Z}_l(\cdot)$ is linear in $Z_d(\cdot)$, it will not be included in $\hat{\mathbf{Z}}^*_{21}(\cdot)$ and $\hat{\mathbf{Z}}^*_{22}(\cdot)$ that are used to form the estimating functions defined in (10) and (11).

5.2 Simulation Results

We generate 1000 realizations of $\{N_1, N_2, M_1, M_2\}$ and apply our proposed method to estimate $\boldsymbol{\beta}$ as described in Section 4.3. For comparison, we also conduct the estimation by combining different subsets of these pairs (see Table 2 and the discussion below for details). For the analysis based on (N_1, M_1) alone, we also run a logistic regression analysis without adjusting for $\pi(\cdot)$, since this is the commonly adopted approach in practice. For each estimator, we calculate its empirical mean and standard error (values in round brackets), and estimate the standard error (values in square brackets) using a standard sandwich estimator.

We also apply the methods proposed by Prentice and Sheppard (1995) and Diggle et al. (2010). For both methods, we first divide W into 100 equal subsquare regions, $W_k : k = 1, \ldots, 100$. For Prentice and Sheppard (1995), the total count of events $N_c = N_1 \cup N_2$ in each subsquare is known in addition to the control data M_1, but no information on $\mathbf{Z}(\cdot)$ is available for any case event in N_c. The unweighted version of the estimator is used due to its ease of implementation as well as its good performance (Prentice and Sheppard 1995). For Diggle et al. (2010), we assume that the summary measures $\int_{W_k} \lambda_0(\mathbf{s})\mathbf{Z}(\mathbf{s})d\mathbf{s}$ are also available, for $k = 1, \ldots, 100$. These summary measures are then combined with the case events in N_1 to estimate $\boldsymbol{\beta}$. Diggle et al.'s (2010) method requires that the complete set of $\mathbf{Z}(\cdot)$ must be known for the cases. As a result, N_2 cannot be used.

Table 2. Simulation results. The two rows for each of the first three methods show the empirical means and standard errors from 1,000 simulations; the three rows for each of the remaining methods show the empirical means and standard errors from 1,000 simulations, and the estimated sandwich standard errors

Method	β_0	β_d	β_t	β_l
Prentice & Sheppard	−6.5792	1.0377	1.0073	1.0296
	(0.0776)	(0.1840)	(0.1683)	(0.1996)
Diggle et al.	−8.4938	1.4567	0.9996	1.0296
	(0.1009)	(0.1996)	(0.1724)	(0.2013)
(N_1, M_1) Logistic	−0.6656	1.4282	1.0107	1.0040
	(0.1021)	(0.1743)	(0.1781)	(0.1844)
(N_1, M_1) Adjusted for $\boldsymbol{\eta}$	−6.5729	1.0099	1.0110	1.0041
	(0.0773)	(0.1450)	(0.1781)	(0.1844)
	[0.0774]	[0.1476]	[0.1787]	[0.1885]
$(N_1, M_1) + (N_2, M_1)$	−6.5723	1.0062	1.0092	1.0086
	(0.0679)	(0.1257)	(0.1310)	(0.1854)
	[0.0679]	[0.1270]	[0.1330]	[0.1876]
$(N_1, M_1) + (N_1, M_2)$	−6.5763	1.0027	1.0061	1.0051
	(0.0548)	(0.0752)	(0.1381)	(0.1309)
	[0.0536]	[0.0755]	[0.1341]	[0.1285]
$(N_1, M_1) + (N_2, M_2)$	−6.5731	1.0014	1.0003	0.9978
	(0.0520)	(0.0698)	(0.0719)	(0.1771)
	[0.0521]	[0.0700]	[0.0713]	[0.1787]
$(N1, M1) + (N1, M2)$	−6.5721	1.0028	0.9964	0.9988
$+(N2, M1) + (N2, M2)$	(0.0440)	(0.0640)	(0.0711)	(0.1308)
	[0.0428]	[0.0633]	[0.0695]	[0.1261]

The simulation results are shown in Table 2. For both the conventional logistic regression analysis approach and Diggle et al.'s approach, we can see that $\hat{\beta}_d$ is severely biased. This is because the probability of selecting N_1 from N_c, which depends on $Z_d(\cdot)$, is not adjusted by these approaches. All the remaining estimators are approximately unbiased, and our proposed estimator has the smallest standard error for all regression coefficients. Moreover, the standard error estimates obtained from the sandwich method are reasonably close to their empirical counterparts.

To understand how the inclusion of additional data may affect accuracy of the resulting estimators, we supplement (N_1, M_1) each time in the estimation with only one of the remaining three pairs, that is, (N_1, M_2), (N_2, M_1), and (N_2, M_2). In all these situations, the standard error of $\hat{\beta}_d$ is significantly reduced and the largest reduction occurs when (N_2, M_2) is included. This is expected because all these new pairs contain additional information on $Z_d(\cdot)$ and the most information is provided by (N_2, M_2). Some further findings and possible explanations are: when (N_1, M_2) is included, the standard errors of $\hat{\beta}_l$ and $\hat{\beta}_t$ are both reduced due to the new individual-level and coarsened information in M_2 on $Z_l(\cdot)$ and $\hat{Z}_t(\cdot)$, respectively; when (N_2, M_1) is included, the additional information on $Z_t(\cdot)$ from N_2 leads to a better estimator for β_t; when (N_2, M_2) is included, information on $\hat{Z}_t(\cdot)$ that is available in both N_2 and M_2 helps improve the estimation of β_t. For the latter two scenarios, $\hat{Z}_l(\cdot)$ is needed to set up the necessary estimating functions. However, because $\hat{Z}_l(\cdot)$ is linear in $Z_d(\cdot)$, it does not provide any additional information on $Z_l(\cdot)$. As a result, the standard errors of $\hat{\beta}_l$ are similar to that based on (N_1, M_1) alone.

We have conducted additional simulations with different sample sizes for both cases and controls. The primary findings described above persisted. Furthermore, when including all four pairs, we have assessed the effects of doing so in different sequential orders. The results are nearly identical so we omit a detailed presentation. We have also considered solving $\tilde{\mathbf{U}}(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}}) = \mathbf{0}_p$ directly, where $\tilde{\mathbf{U}}(\cdot)$ is given in (12). The resulting estimator can be much more variable than that obtained from our proposed sequential approach. Moreover, the estimated standard errors based on the sandwich method tend to underestimate the true standard errors.

6. DATA ANALYSIS

6.1 Definition of Risk Factors

We define the demographic risk factors $\mathbf{Z}_d(\cdot) = [1, \text{age}, \text{age}^2, \text{sex}, \text{race}]'$, where sex = 1 for male and 0 for female, and race = 1 for white and 0 for others. For the lifestyle variables, we define $\mathbf{Z}_l(\cdot) = [\text{smoking}, \text{education}]'$, where smoking = 1 if one ever smoked and 0 otherwise, and education = 1 if the subject received some college or above education and 0 otherwise.

To estimate $\mathbf{Z}_l(\cdot)$, we employ two different strategies. For smoking, we run a logistic regression analysis based on $M_c = M_1 \bigcup M_2$, where M_1 and M_2 are controls in the case–control study and the BRFSS, respectively. The response variable is a subject's smoking status and the predictors are the demographic variables $\mathbf{Z}_d(\cdot)$. We then define an estimate as the fitted value given $\mathbf{Z}_d(\cdot)$. For education, we still run a logistic regression analysis based on M_c. The response variable is the education status, but both $\mathbf{Z}_d(\cdot)$ and $\tilde{Z}_{\text{edu}}(\cdot)$ are used as predictors, where $\tilde{Z}_{\text{edu}}(\cdot)$ is the percentage of population aged 25 and up that had received some college or above education at the zip code level from the 2000 U.S. Census.

The traffic-related exposure variable is derived as described in Section 2.4, with the subsegment length $\Delta \mathbf{s}_{ij} = 50$ m and the radius of the circular buffer zone $D = 2000$ m. A power transformation of 0.25 is used to reduce the skewness. We have also considered $D = 1000, 3000$ m but obtained similar results. We rewrite $\mathbf{Z}_t$ as Z_t in what follows since only one traffic-related exposure variable is considered. As described in Section 4, $\hat{Z}_t(\cdot)$ is defined as the exposure derived at the centroid of the zip code region that the subject resided in. Both age and traffic exposure have been standardized in the subsequent analysis. We define $\mathbf{Y}(\cdot) = [\mathbf{Z}_d(\cdot), Z_t(\cdot)]'$ when estimating the selection probability $\pi(\cdot)$.

6.2 Derivation of Sampling Probabilities

Since the controls in the case–control study (i.e., M_1) were selected to frequency match the age and sex distributions of the case process N_1, we derive the sampling probability $\alpha_1(\mathbf{s})$ given the subject's age and sex information. To do so, we first obtain the age-by-sex distribution based on the Census for the following ten age groups: 35–40, 41–45, 46–50, 51–55, 56–60, 61–65, 66–70, 71–75, 76–80, and 81–83. Let $S(\mathbf{s})$ denote the age-by-sex stratum that the subject at $\mathbf{s}$ is in. Then, we define $\alpha_1(\mathbf{s})$ as

$$\alpha_1(\mathbf{s}) = \frac{\text{number of subjects from } M_1 \text{ in } S(\mathbf{s})}{\text{total number of subjects in the population in } S(\mathbf{s})}.$$

Similarly, we define

$$\alpha_2(\mathbf{s}) = \frac{\text{number of subjects from } M_2 \text{ in } S(\mathbf{s})}{\text{total number of subjects in the population in } S(\mathbf{s})}.$$

It may be overly simplistic to derive the sampling probabilities based on the age-by-sex distribution alone, since other factors such as the number of residential telephone lines in the respondent's household and whether the telephone number(s) are in directory listings will also affect one's probability to be selected in a study. Further, the likelihood of participation of selected potential controls also generally depends upon socioeconomic and other factors that are typically difficult to account for in case–control analyses since they are not measured well (or may be unknown). The BRFSS data do assign differential weights to the sampled subjects by taking some of the various factors into consideration. Let $w(\mathbf{s})$ denote the weight assigned to an individual at $\mathbf{s}$. The weights can be viewed as the inverse of the sampling probabilities. We therefore define

$$\alpha_2^*(\mathbf{s}) = \frac{1}{w(\mathbf{s})}.$$

Although $\alpha_2^*(\cdot)$ can better describe the sampling probabilities than $\alpha_2(\cdot)$, it is not available for the cases since information such as the number of residential telephone lines is typically not recorded for cases. Moreover, the selection probability $\pi(\cdot)$ depends on the traffic variable $Z_t(\cdot)$ and therefore cannot be calculated for the controls in the BRFSS data. As a result, the estimating functions $\mathbf{U}_{12}(\cdot)$ and $\mathbf{U}_{22}(\cdot)$ given in (9) and (11) cannot be calculated. For $\mathbf{U}_{12}(\cdot)$, we modify it as

$$\begin{aligned}
&\mathbf{U}_{12}^*(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}}) \\
&= \sum_{\mathbf{s} \in (N_1 \cap W)} \hat{\mathbf{Z}}(\mathbf{s}) \frac{\alpha_2(\mathbf{s})}{\alpha_2(\mathbf{s}) + \tilde{\pi}(\mathbf{s}) \exp[\hat{\mathbf{Z}}(\mathbf{s})'\boldsymbol{\beta}]} \\
&\quad \cdot \frac{\exp[\hat{Z}_t(\mathbf{s})'\boldsymbol{\beta}_t]\tilde{\pi}(\mathbf{s})}{\exp[Z_t(\mathbf{s})'\boldsymbol{\beta}_t]\pi(\mathbf{s}; \hat{\boldsymbol{\eta}})} \\
&\quad - \sum_{\mathbf{s} \in (M_2 \cap W)} \hat{\mathbf{Z}}(\mathbf{s}) \frac{\tilde{\pi}(\mathbf{s}) \exp[\hat{\mathbf{Z}}(\mathbf{s})'\boldsymbol{\beta}]}{\alpha_2(\mathbf{s}) + \tilde{\pi}(\mathbf{s}) \exp[\hat{\mathbf{Z}}(\mathbf{s})'\boldsymbol{\beta}]} \cdot \frac{\alpha_2(\mathbf{s})}{\alpha_2^*(\mathbf{s})}, \qquad (14)
\end{aligned}$$

where $\tilde{\pi}(\cdot)$ is an alternative estimate of $\pi(\cdot)$ obtained by defining $\mathbf{Y}(\cdot)$ as the demographic variables $\mathbf{Z}_d(\cdot)$ alone. A similar modification can be made straightforwardly for $\mathbf{U}_{22}(\cdot)$ and we denote the resulting new estimating function by $\mathbf{U}_{22}^*(\cdot)$.

6.3 Results

We first estimate $\boldsymbol{\eta}$ based on Equation (7), which yields

$$\hat{\boldsymbol{\eta}} = [-1.8484, -0.3380, -0.2729, 0.2003, 0.3182, -0.1316].$$

The result suggests that younger, male, white patients were more likely to participate in the case–control study and that cases included in the case–control study were less exposed to traffic than those in the CTR but not in the case–control study. The demographic variables in $\mathbf{Z}_d(\cdot)$ are in fact also correlated with the lifestyle variables in $\mathbf{Z}_l(\cdot)$. Given these observations, biased estimates for β_t and $\boldsymbol{\beta}_l$ could be obtained if we do not adjust for $\boldsymbol{\eta}$.

We assume the intensity model (2) given in Section 3, with the covariate vector $\mathbf{Z}(\cdot)$ therein as being defined in Section 6.1. Our main purpose is to estimate the regression parameters $\boldsymbol{\beta}$ by combining the data from the following different sources: the case–control data (i.e., N_1 and M_1), the CTR data excluding the cases included in the case–control study (i.e., N_2), and the BRFSS data (i.e., M_2). To apply our proposed method, we combine the available pairs sequentially in the order (N_1, M_1), (N_1, M_2), (N_2, M_1), and (N_2, M_2). For comparison, we also conduct the estimation based on different subsets of these pairs (see Table 3 for details). For the analysis based on (N_1, M_1) alone, we also run a logistic regression analysis to estimate $\boldsymbol{\beta}$ without adjusting for $\boldsymbol{\eta}$.

Figure 2 shows plots of cumulative residuals $Q(x)$ versus age in each sex-by-race group, and Figure 3 plots zip-code residuals R_l versus zip-code level traffic, where $Q(x)$ and R_l are as defined in Section 4.4. The confidence bands of $Q(x)$ include the zero line inside for all x, suggesting a satisfactory overall fit. The plot of the zip-code residuals is centered around zero and also shows

Table 3. Data analysis result. The two rows for each method show the parameter estimates and the estimated bootstrap standard errors

	Intercept	Age	Age^2	Sex	Race	Traffic	Smoke	Education
(N_1, M_1) Logistic	−0.249	0.175	0.047	0.017	−0.287	0.064	0.456	−0.641
	(0.332)	(0.136)	(0.080)	(0.144)	(0.268)	(0.074)	(0.150)	(0.150)
(N_1, M_1) Adjusted for η	−5.557	0.911	−0.057	0.319	−0.446	0.081	0.431	−0.665
	(0.220)	(0.057)	(0.036)	(0.062)	(0.205)	(0.054)	(0.158)	(0.153)
$(N_1, M_1) + (N_2, M_1)$	−5.416	0.779	−0.129	0.277	−0.551	0.157	0.435	−0.689
	(0.214)	(0.046)	(0.033)	(0.061)	(0.194)	(0.050)	(0.162)	(0.164)
$(N_1, M_1) + (N_1, M_2)$	−6.079	0.799	−0.136	0.288	−0.120	0.127	0.495	−0.354
	(0.215)	(0.054)	(0.059)	(0.097)	(0.191)	(0.065)	(0.154)	(0.136)
$(N_1, M_1) + (N_2, M_2)$	−6.013	0.747	−0.181	0.285	−0.009	0.233	0.424	−0.518
	(0.199)	(0.038)	(0.025)	(0.054)	(0.114)	(0.051)	(0.159)	(0.124)
$(N_1, M_1) + (N_1, M_2)$	−6.166	0.778	−0.164	0.251	−0.089	0.217	0.508	−0.235
$+(N_2, M_1) + (N_2, M_2)$	(0.134)	(0.034)	(0.021)	(0.048)	(0.099)	(0.042)	(0.144)	(0.132)

no systematic pattern, and hence the functional form of $Z_t(\cdot)$ appears to be appropriate.

Figure 4 shows the plot of $T(r)$. The values of $T(r)$ are approximately constant and are reasonably close to one for all r. Moreover, the confidence band contains one throughout the plotted range, which indicates that the assumption of Poisson process is acceptable.

For variance estimation, we need to incorporate the complex sampling design used to produce the BRFSS data. For our data, a disproportionate stratified sampling design was used involving four geographic region-by-household density strata. Each BRFSS weight is calculated as the product of three components: the stratum weight, which accounts for differences in the basic probability of selection among strata, a raw weighting factor, which adjusts for variations in the numbers of residential telephone numbers and adults in the respondent's household, and a post-stratification weight, which adjusts for noncoverage and nonresponse based on the age-by-sex distribution in the population at the county level. We follow Lahiri (2003) to use a bootstrap procedure to estimate the variance. Specifically, we first sample with replacement the same number of subjects within each stratum. By doing so, both the stratum weight and the raw weighting factor associated with each newly selected sample are unchanged. We then adjust the post-stratification weights by comparing the age-by-sex distributions in the population and in the resampled data. We resample the remaining data, that is, N_1, N_2, and M_1, as described in Section 4.4. For each resampled dataset, we apply our proposed model estimation procedures to estimate $\boldsymbol{\beta}$.

The estimation results are shown in Table 3. Values in parentheses are the estimated standard errors calculated by the bootstrap method. The results from our proposed sequential

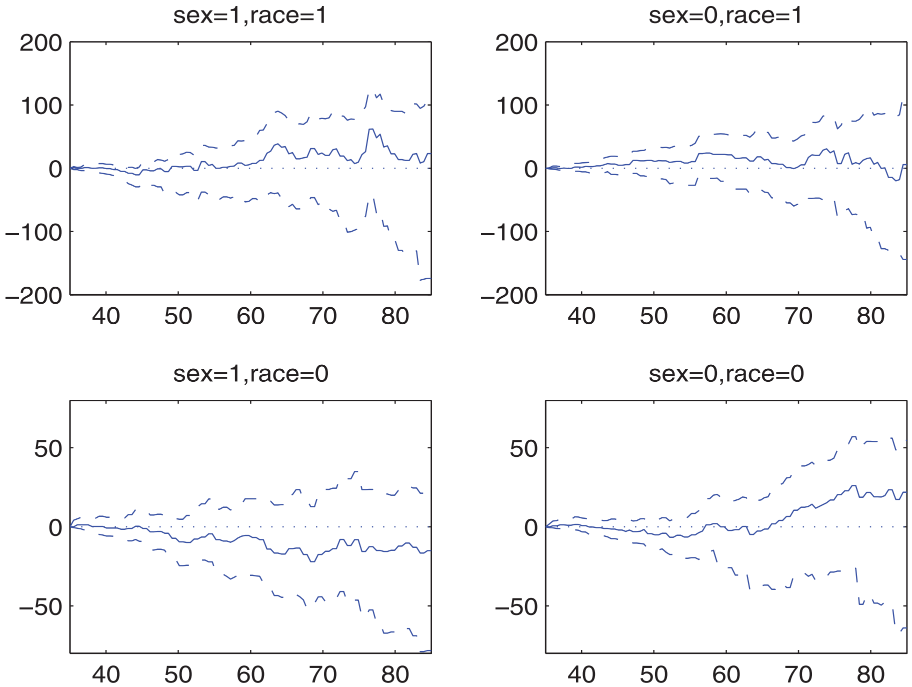

Figure 2. Plots of cumulative residuals versus age in each sex-by-race group. In each plot, the solid line is $Q(x)$ (y-axis) versus age (x-axis), the dashed lines are the confidence band, the dotted line is the constant line of zero.

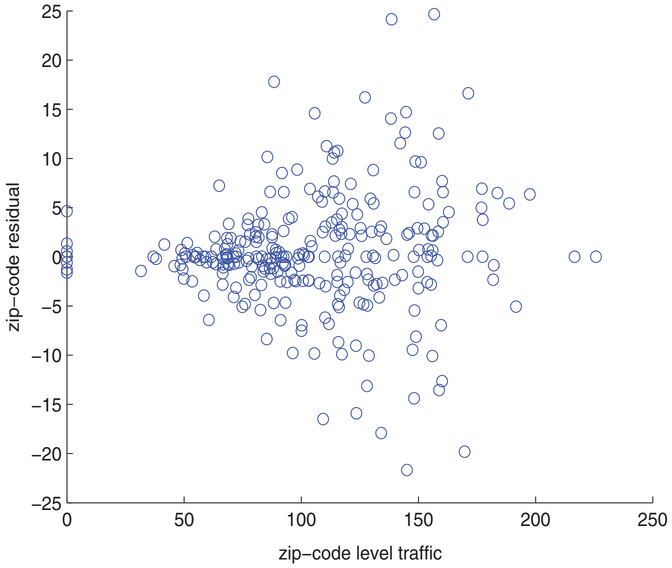

Figure 3. Plot of zip-code residuals versus zip-code level traffic.

approach suggest that exposure to traffic is a significant risk factor for pancreatic cancer. After adjusting for the effect of age, sex, race, smoking, and education, a unit change in the standardized traffic exposure variable increases the risk by a factor of 1.2427 (90% confidence interval 1.1594 to 1.3312). A significant relationship is also detected in all other analyses except the two based on (N_1, M_1) alone. These observations demonstrate the benefits of including the additional CTR and BRFSS data.

Our analysis also revealed a significant effect of smoking and education. Specifically, after adjusting all the other factors, the pancreatic cancer risk increases by a factor of 1.6622 (90% confidence interval 1.3114–2.1062) if one ever smoked but decreases by a factor of 0.7905 (90% confidence interval 0.6363–0.9823) if one has received college or above education. The finding on smoking is consistent with other findings in literature (Risch et al. 2010). The decrease of pancreatic cancer risk with education may be due to unobserved socioeconomic factors that confound with education.

The results also suggest that both age and sex are significantly related to the risk of developing pancreatic cancer. After controlling for other risk factors, males have an increased risk versus females by a factor of 1.2850 (90% confidence interval 1.1877–1.3909). The risk increases with age, but the interpretation is less straightforward due to the term for age squared. However, neither age nor sex is significant from the conventional logistic regression analysis based on (N_1, M_1) alone. This is because the controls were frequency matched to the cases in the case–control study by age and sex.

Our proposed sequential approach suggests that race is not related to risk for pancreatic cancer. However, when the BRFSS data (i.e., M_2) are not included, the results suggest that white race has a significantly lower risk for pancreatic cancer than other races. This observation is likely a result of the large number of white subjects included in M_1.

7. DISCUSSION

We have proposed a new method for combining epidemiologic data that are obtained from diverse sources. The proposed approach allows us to make full use of all available information, regardless of source. It is computationally simple and can also be easily generalized to more complex settings. Our simulation shows that our method can yield estimators with smaller variances than those based on only a subset of the available data

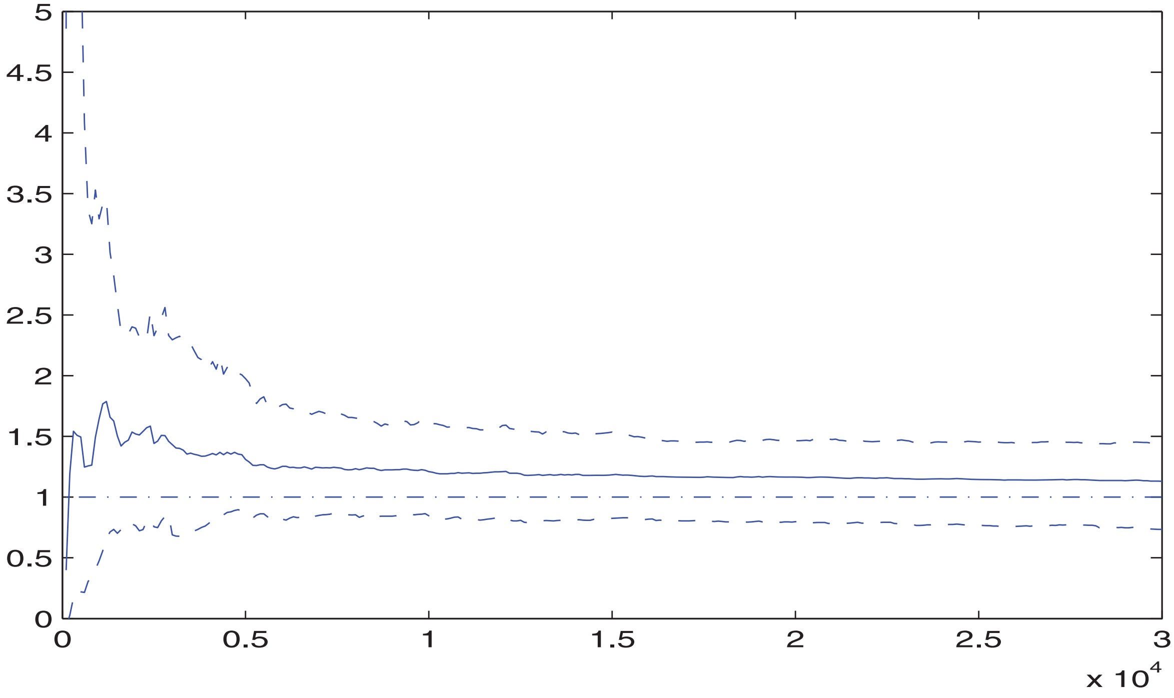

Figure 4. Test for Poisson point process. The solid line is $T(r)$ of data, the dashed lines are the confidence band, and the dot-dashed line is the constant line of one.

sources. In the substantive application, we have supplemented a standard population-based case–control study with Tumor Registry data and BRFSS data. The inclusion of these additional data can significantly enhance study power to detect effects of risk factors over the conventional case–control analysis approach. In particular, our analysis of the Connecticut pancreatic cancer data provides evidence for a positive association between traffic-related exposure and disease risk, while such a conclusion cannot be made with the case–control data alone.

For many population-based case–control studies, individual residential histories are available and can be used to construct trajectories of past traffic-related exposures. Since we do not have such data, we derive our traffic-related exposure based on residential locations at the time of diagnosis (for cases) or interview (for controls). A small number (3%) of patients also did not have geocodable addresses on file, due to the use of post-office boxes and possibly because of transcription errors when the addresses were recorded by tumor registrars. We do not include these data in our analysis. Given the successfully geocoded addresses, we define traffic-related exposure broadly based on geographic proximity to highways. In reality, vehicle emissions include both particulate and gaseous pollutants, and the impact of different pollutants on health outcomes could be influenced by the local environment (e.g., wind, rain) and mode of transmission (e.g., inhalation vs. ingestion).

When forming our proposed estimating functions (8)–(11), we require that all necessary risk factors must be available in either the case or the control data being considered or in both. The unbiasedness of these estimating functions is maintained even if there are unmeasured confounders in some of the data sources. For example, the tumor registry data do not provide any information on smoking which can be a potential confounder. However, we form the estimating functions (10) and (11) by combining the tumor registry data with the control data in a case–control study and the BRFSS data, both of which contain information on smoking. By doing so, we can still obtain unbiased estimating functions. Nevertheless, biased estimates can be resulted in if there are omitted confounders in both the case and control data sources. In our application, residential proximity to highways may also be associated with other factors such as socioeconomic status. Although we have controlled for education, a commonly used proxy measure for socioeconomic status, residual confounding remains a possibility.

As pointed out by one referee, our real data analysis results reveal that the parameter estimates can vary with the data sources being included, even though the effect of the risk factors is common as specified by (2). In some situations, conflicting results may even be obtained; see the discussion in the last paragraph of Section 6 regarding the effect of race as an example. Such variations are due to the different forms of potential selection bias in the sampling designs used to collect the data from different sources. For the case data, we are able to account for the selection bias through the use of (7), because we have information for all available cases given the tumor registry data. Should data from an unbiased sampling design be available for the controls, then a similar mechanism can be potentially developed and incorporated in the estimation process to correct for the bias associated with a given control data source. More research is needed on this topic, since that would require a more careful design of future epidemiological studies and also a nontrivial extension of our proposed approach. Nevertheless, it is reasonable to believe that the BRFSS data can produce less biased results than the controls collected by individual investigators in a case–control study, because more sophisticated sampling designs and statistical tools are often used to mitigate the bias for the BRFSS data; hence, we believe that our estimates are more objective than those obtained from the commonly used approach based on case–control data alone. In terms of the effect of traffic, our main conclusion on its significance persisted across the different data sources, despite the variations in the actual estimates.

SUPPLEMENTARY MATERIALS

WebAppendix

[Received February 2012. Revised September 2013.]

REFERENCES

Angrist, J. D., and Krueger, A. B. (1992), "The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables With Moments From Two Samples," *Journal of the American Statistical Association*, 87, 328–336. [12]

Beelen, R., Hoek, G., van den Brandt, P. A., Goldbohm, R. A., Fischer, P., Schouten, L. J., Armstrong, B., and Brunekreef, B. (2008), "Long-Term Exposure to Traffic-Related Air Pollution and Lung Cancer Risk," *Epidemiology*, 19, 702–710. [13,14]

Best, N. G., Ickstadt, K., and Wolpert, R. L. (2000), "Spatial Poisson Regression for Health and Exposure Data Measured at Disparate Resolutions," *Journal of the American Statistical Association*, 95, 1076–1088. [11]

Crowder, M. (1986), "On Consistency and Inconsistency of Estimating Equations," *Econometric Theory*, 2, 305–330. [14]

Diggle, P. J., Gómez-Rubio, V., Brown, P. E., Chetwynd, A. G., and Gooding, S. (2007), "Second-Order Analysis of Inhomogeneous Spatial Point Processes Using Case–Control Data," *Biometrics*, 63, 550–557. [17]

Diggle, P. J., Guan, Y., Hart, C., Paize, F., and Stanton, M. (2010), "Estimating Individual-Level Risk in Spatial Epidemiology Using Spatially Aggregated Information on the Population at Risk," *Journal of the American Statistical Association*, 105, 1394–1402. [11,18]

Diggle, P. J., and Rowlingson, B. S. (1994), "Conditional Approach to Point Process Modelling of Elevated Risk," *Journal of the Royal Statistical Society*, Series A, 157, 433–440. [15]

Gelman, A., King, G., and Liu, C. (1998), "Not Asked and Not Answered: Multiple Imputation for Multiple Surveys," *Journal of the American Statistical Association*, 93, 846–857. [12]

Haneuse, S., and Wakefield, J. (2007), "Hierarchical Models for Combining Ecological and Case–Control Data," *Biometrics*, 63, 128–136. [11]

——— (2008a), "Geographic-Based Ecological Correlation Studies Using Supplemental Case–Control Data," *Statistics in Medicine*, 27, 864–887. [11]

——— (2008b), "The Combination of Ecological and Case–Control Data," *Journal of the Royal Statistical Society,* Series B, 70, 73–93. [11]

Hansen, L. P. (1982), "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054. [16]

Heyde, C. C. (1997), *Quasi Likelihood and Its Application a General Approach to Optimal Parameter Estimation*, New York: Springer-Verlag, Inc. [16]

Imbens, G. W., and Lancaster, T. (1994), "Combining Micro and Macro Data in Microeconometric Models," *The Review of Economic Studies*, 61, 655–680. [12]

Jackson, C., Best, N., and Richardson, S. (2006), "Improving Ecological Inference Using Individual-Level Data," *Statistics in Medicine*, 25, 2136–2159. [12]

Little, R., and Rubin, D. (2002), *Statistical Analysis With Missing Data*, New York: Wiley. [12]

Møller, J., and Waagepetersen, R. (2004), *Statistical Inference and Simulation for Spatial Point Process*, London: Chapman and Hall. [17]

Pearson, R. L., Wachtel, H., and Ebi, K. L. (2000), "Distance-Weighted Traffic Density in Proximity to a Home is a Risk Factor for Leukemia and Other Childhood Cancers," *Journal of the Air & Waste Management Association*, 50, 175–180. [13]

Prentice, R. L., and Sheppard, L. (1995), "Aggregate Data Studies of Disease Risk Factors," *Biometrika*, 82, 113–125. [11,18]

Raaschou-Nielsen, O., Andersen, Z. J., Hvidberg, M., Jensen, S. S., Ketzel, M., Sørensen, M., Hansen, J., Loft, S., Overvad, K., and Tjønneland, A. (2011), "Air Pollution From Traffic and Cancer Incidence: A Danish Cohort Study," *Environmental Health*, 10, 67–77. [13,14]

Raaschou-Nielsen, O., Hertel, O., Thomsen, B. L., and Olsen, J. H. (2001), "Air Pollution From Traffic at the Residence of Children With Cancer," *American Journal of Epidemiology*, 153, 433–443. [13]

Rathbun, S. L. (2012), "Optimal Estimation of Poisson Intensity With Partially Observed Covariates," *Biometrika*, 100, 277–281. [15]

Reynolds, P., Behren, J. V., Gunier, R. B., Goldberg, D. E., Hertz, A., and Smith, D. (2002), "Traffic Patterns and Childhood Cancer Incidence Rates in California, United States," *Cancer Causes and Control*, 13, 665–673. [13,14]

Risch, H. A., Yu, H., Lu, L., and Kidd, M. S. (2010), "ABO Blood Group, Helicobacter Pylori Seropositivity, and Risk of Pancreatic Cancer: A Case–Control Study," *Journal of the National Cancer Institute*, 102, 502–505. [12,21]

Robins, J. M., Rotnitzky, A., and Zhao, L. (1994), "Estimation of Regression Coefficients When Some Regressors are Not Always Observed," *Journal of the American Statistical Association*, 89, 846–866. [12]

Rubin, D. B. (2004), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley. [12]

Schafer, J. L. (1999), "Multiple Imputation: A Primer," *Statistical Methods in Medical Research*, 8, 3–15. [12]

Schenker, N., Raghunathan, T., and Bondarenko, I. (2010), "Improving on Analyses of Self-Reported Data in a Large-Scale Health Survey by Using Information From an Examination-Based Survey," *Statistics in Medicine*, 29, 533–545. [12]

Schlesselman, J. J. (1982), *Case–Control Studies: Design, Conduct, Analysis*, New York: Oxford University Press. [11]

Visser, O., van Wijnen, J. H., and van Leeuwen, F. E. (2004), "Residential Traffic Density and Cancer Incidence in Amsterdam, 1989–1997," *Cancer Causes and Control*, 15, 331–339. [13,14]

Wakefield, J. (2004), "Ecological Inference for 2 × 2 Tables" (with discussion), *Journal of the Royal Statistical Society,* Series A, 167, 385–445. [11]

Zhou, M., and Kim, J. K. (2012), "An Efficient Method of Estimation for Longitudinal Surveys With Monotone Missing Data," *Biometrika*, 99, 631–648. [12,16]