International Journal of Health Geographics

Review

Current practices in spatial analysis of cancer data: data characteristics and data sources for geographic studies of cancer

Francis P Boscoe[1], Mary H Ward*[2] and Peggy Reynolds[3]

Address: [1]New York State Cancer Registry, New York State Department of Health, Albany, NY, USA, [2]Occupational and Environmental Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS, Bethesda, MD, USA and [3]California Department of Health Services, Environmental Health Investigations Branch, Oakland, CA, USA

Email: Francis P Boscoe - fpb01@health.state.ny.us; Mary H Ward* - wardm@mail.nih.gov; Peggy Reynolds - PReynold@dhs.ca.gov

* Corresponding author

Published: 01 December 2004

International Journal of Health Geographics 2004, **3**:28 doi:10.1186/1476-072X-3-28

Received: 29 September 2004
Accepted: 01 December 2004

This article is available from: http://www.ij-healthgeographics.com/content/3/1/28

Abstract

The use of spatially referenced data in cancer studies is gaining in prominence, fueled by the development and availability of spatial analytic tools and the broadening recognition of the linkages between geography and health. We provide an overview of some of the unique characteristics of spatial data, followed by an account of the major types and sources of data used in the spatial analysis of cancer, including data from cancer registries, population data, health surveys, environmental data, and remote sensing data. We cite numerous examples of recent studies that have used these data, with a focus on etiological research.

Introduction

Understanding the spatial patterns of diseases in a population can provide insight as to their causes and controls. Indeed, this notion is at the very root of the field of epidemiology [1]. The recent explosion in data gathering, linkage and analysis capabilities fostered by computing technology, particularly geographic information systems (GIS), has greatly improved the ability to measure and assess these patterns. Large and complex georeferenced data sets are now readily available through Spatial Data Clearinghouses, facilitating analyses by researchers unaffiliated with the government agencies that have historically controlled data access. Meanwhile, increasingly sophisticated statistical tools have evolved to keep pace with the increased data availability and computing power.

The purpose of this article is to provide an overview of spatial data and its relevance to population-based cancer surveillance and research in the United States as of 2004. We begin by discussing a number of the distinctive characteristics of spatial data, which can sometimes hinder efforts to understand cancer etiology. We then proceed to describe the kinds of data sets that are available, accompanied by a survey of some applications using these data. Finally, we discuss several ongoing efforts to provide central repositories of geospatial data. Given the vast scope of cancer research taking place worldwide, our survey is necessarily partial, and we have chosen to emphasize etiology over other research themes with spatial dimensions, such as patterns of treatment or access to care [2].

Qualities of spatial data

Spatial data refer to data with locational attributes. Most commonly, locations are given in Cartesian coordinates referenced to the earth's surface. These coordinates may describe points, lines, areas or volumes. This need not be the only spatial framework; "relative spaces" may be defined in which distance is defined in terms of some other attribute, such as sociodemographic similarly or connectedness along transportation networks [3,4].

Spatial data have special qualities that require specialized statistical techniques and modeling approaches. A complete discussion of these special qualities is well beyond the scope of this article, but here we describe a number of the more compelling and recurring themes. For a focused discussion on the limitations on analysis that these data characteristics impose, see the companion piece to this article, "Current Practices in Spatial Analysis of Cancer Data: Flies in the Ointment, Or, The Limitations of Spatial Analysis" [5].

Individual humans represent the basic unit of spatial analysis in cancer research. Individuals are categorized as either having or not having a disease or attribute of a disease, and are assigned coordinates corresponding to the location of their place of residence, a technique known as geocoding. As with all measurements, geocoding involves some error. A growing body of literature is exploring the nature of this error and its potential to bias epidemiologic studies [6]. Among the topics that have been investigated are systematic problems with geographic reference files [7], the ramifications of different geocoding algorithms [8], positional accuracy [9] and how to handle non-residential addresses, such as rented post office boxes [10].

Assigning individuals to their place of residence also poses problems, although this is usually the only locational information that is available. Often the goal of a geographic analysis is to identify a common environmental exposure in a population, but exposures that are occupational or recreational may not necessarily reveal themselves in a residential analysis. Also, given the long latency period for many cancers and the mobility of the American population, the relevant exposure may be associated with a prior address. Difficult-to-measure behavioral risk factors such as smoking and diet often further confound attempts at geographic analysis.

Owing to confidentiality restrictions, researchers outside of central cancer registries typically do not have access to address-level data. In such instances, case data are aggregated by some functional or political unit such as census tract, county or ZIP code. Even when the case data are geocoded, population data must be aggregated, at least to the level of the census block, which is the smallest unit for which any population information is available. Knowing that there are four women with breast cancer living on the same street is not sufficient, by itself, to draw conclusions about whether the street displays an unusual incidence pattern; one must also know the number and ages of women without breast cancer on the same street. Short of conducting one's own thorough door-to-door census, this question cannot be answered, except by aggregating the street segments into blocks. When additional variables, such as measures of income or education, or also of interest, then still larger analytical units must be chosen.

The necessity for aggregating spatial data raises a whole set of analytic issues regarding the extent to which the act of aggregating introduces error and bias. It is theoretically possible to achieve dramatically different, even contradictory results, simply as a consequence of aggregating the data in a different fashion [11]. This is true not only for aggregations at different spatial scales, but also different aggregations at the same scale. Geographers have termed this the "modifiable areal unit problem" [12]. A special case of the modifiable areal unit problem is the ecological inference problem, which specifically refers to the lack of congruity between associations found in aggregated and individual-level data.

In practice, well-chosen scales and groupings can minimize the modifiable areal unit problem and allow reasonable consistency between aggregated and individual-level results [13]. There will always be exceptions to this, however, as evidenced by the many studies attempting to relate low-level indoor radon concentrations with lung cancer incidence. Individual-level studies have repeatedly found a positive correlation, while area-level studies have found a negative correlation at low radon levels [14-16]. Despite a general appreciation of how these discrepant results represent an example of aggregation bias, there is still active debate over what these results say about low-level radon risk [17,18].

Often analyses need to be performed on data that was collected at different spatial scales, such as a study using cancer cases aggregated by ZIP code and modeled air pollutant data at the census tract level. The resulting scale-translation problem is a recurring one that has inspired many independent solutions, and is known variously as areal interpolation, the polygon overlay problem, and the problem of inference with spatially misaligned data, among other terms [19]. The most naïve solution to this problem is to assume that each measured value is homogeneous within each spatial unit. Under this assumption, using our example, a ZIP code that is coincident with four census tracts would be broken into four polygons, each having the same cancer rate but different air pollutant values. More sophisticated cartographic overlay techniques have been developed that involve using covariate information to infer variation within spatial units. To date, these techniques have been primarily applied toward estimating population surfaces rather than cancer or other disease rate surfaces [20,21]. Hierarchical Bayesian and multi-level logit models have also shown promise [22-25].

Spatial autocorrelation is another distinctive quality of spatial data that requires the use of specialized analytic methods. Spatial autocorrelation is the tendency for nearby observations to have correlated attribute values. For most data sets involving the distribution of human populations and their characteristics, spatial autocorrelation is positive, meaning that neighboring individuals tend to have similar characteristics. Understanding the characteristics and qualities of spatial autocorrelation is essential to adequately model and interpret geographic patterns. For example, it is not appropriate to perform ordinary least squares regression on spatial data, because the presence of spatial autocorrelation means that the observations are not independent. Performing such a regression generally results in downwardly biased estimations of variance, which yields overstated levels of significance. In general, spatially autocorrelated data is less informative in a model than uncorrelated data. There is an ample literature on assessing and properly accounting for spatial autocorrelation in geographic analysis [26,27].

A final critically important characteristic of spatial data is spatial nonstationarity, or the tendency for relationships between and among variables to vary by geographic location [28]. First-order or strong stationarity refers to the degree to which measured values vary spatially, while second-order or weak stationarity refers to the degree to which the uncertainties in these measured values vary spatially. So-called global statistics ignore nonstationarity, suggesting that relationships across space are constant. The simple linear equation that has traditionally been used to express the relationship between rainfall and altitude is a well-known example. Local statistics, in contrast, take nonstationarity into account, at least first-order nonstationarity. Brunsdon et al. [29] used the technique of geographically weighted regression to demonstrate that both the slope and intercept of the rainfall-altitude equation vary considerably in space. The range and breadth of local statistics has seen rapid growth in recent years [27,30].

Local statistics are less adept at accounting for second-order nonstationarity. Indeed, many of these methods require the assumption of constant variance across space. Because of the uneven distribution of human populations, this assumption is seldom met for health data. Specifically, disease rates in areas with smaller numbers of cases are more variable than those in areas with larger numbers of cases, a property that has also been termed "variance instability" [31]. Variance instability is particularly pervasive on maps, since it is extremely difficult to design a map that is not visually biased toward either sparsely populated or densely populated areas [32,33]. A simple example is the tendency for rural counties to contain disproportionate numbers of unusually high or unusually low disease rates and thus visually dominate a choropleth map. The problem is compounded by the tendency of such counties to be large in size; for these reasons, maps of United States counties are often visually dominated by such states as Idaho, Nevada and Wyoming.

Efforts to include information about data uncertainty have shown promise, but have not seen widespread use [34]. One common way of addressing this problem is to produce smoothed maps, whereby the rate for a given area is influenced by the rates of neighboring areas. There are many algorithms available to accomplish this [35], ranging from conceptually straightforward spatial filters [36] to computationally-intensive Bayesian approaches [37,38]. Properly accounting for second-order spatial nonstationarity in maps and models remains an active research area.

Types and sources of data

In this section we described the primary types and sources of data most frequently used in the geographic analysis of cancer, along with examples of their application. These are summarized in Table 1.

Table 1: Sources of Cancer Registry Data

Dataset name	Source Agency	URL	Geographic Resolution
SEER*Stat, Cancer Mortality Maps and Graphs, State Cancer Profiles	National Cancer Institute	http://surveillance.cancer.gov/statistics	County
Florida Cancer Data System	University of Miami School of Medicine	http://fcds.med.miami.edu/inc/statistics.shtml	County
Cancer Incidence and Mortality Rates in Kentucky	Kentucky Cancer Registry	http://kcr.uky.edu	County
New York Cancer Incidence by ZIP code	NYS Department of Health	http://www.health.state.ny.us/nysdoh/cancer/csii/nyscsii.htm	ZIP code

1. Cancer registries

A cancer registry is a data collection system that tracks cancer cases that have been diagnosed or treated in a specific institution or geographic area. Cancer registries typically collect information from medical records provided by hospitals, doctors, other care facilities, medical laboratories, and/or insurers. Data collected by cancer registries is stored under secure conditions so as to protect confidentiality.

Historically, observed geographic differences in cancer incidence have been of great interest in trying to understand more about factors which may influence risk of these diseases. Such differences have served as the basis for studies of migrant populations and acculturation differences in migrant groups. They have been possible because cancer is one of the few chronic diseases for which high quality population-based disease surveillance systems have been in place for many years in many countries of the world.

Cancer registry data has been widely applied toward the production of cancer atlases [39], studies analyzing the spatial distribution of particular cancer sites [40], and studies assessing spatial clustering [41]. Most recently, cancer studies have been undertaken which build on the combined resources of cancer registry data and increasingly available GIS tools. Because address at diagnosis is available for most registry cases it can be geocoded and integrated in a GIS with social and environmental attribute information available at a variety of geographic scales. Examples of such approaches include studies of childhood cancer which examine rate differences in areas of low versus intense agricultural pesticide use [42], heavy traffic patterns [43], or high air pollution [44]. Alternatively, cancer registry data can serve to identify population-based cases for studies using case-control or cohort designs, which can in turn be integrated into a GIS for area attribute data. Examples of this approach include case-control studies of childhood leukemia and traffic patterns [45-48]. and a studies of breast cancer incidence associated with residence in high pesticide use areas in a large case-control study [49,50]. and in a large cohort study [51].

For these types of studies, cancer registry data offer both a number of strengths and limitations. Primary strengths include the comprehensiveness of geographic coverage, detailed information on disease subgroups, and rich covariable information on demographic characteristics for each newly diagnosed case of cancer. Because registry data are abstracted from medical records and reflect information for a snapshot in time, primary limitations include the lack of historical information on various factors of potential interest including residential mobility and relevant personal behaviors. Cancer registries typically collect information on the residential address for individuals newly diagnosed with cancer at the time of that diagnosis. Since this is the locational information which serves as the basis for national and international statistics on area cancer rates, it is also useful for looking at area characteristics associated with rate differences, although inferences about etiologic associations are limited for these long latency diseases, and even more so for residentially mobile populations.

The Surveillance, Epidemiology and End Results (SEER) program of the National Cancer Institute (NCI) offers county-level incidence data for its member registries, which cover part or all of eight states, through its SEER*Stat software. Because it provides direct access to individual cancer records, users must first sign a data access agreement. County-level mortality data for the entire United States, collected and maintained by the National Center for Health Statistics (NCHS), is also accessible through SEER*Stat. These data include all causes of death, not just cancer deaths. Selected county-level cancer data may also be accessed through the NCI's Cancer Mortality Maps and Graphs and State Cancer Profiles web sites. The latter was launched in 2003 and contains a host of innovative statistical graphics. Many individual state registries also offer additional geographically referenced data. For example, the Florida Cancer Data System web site allows users to generate a variety of county- and facility-level tables and county-level maps on demand. The Kentucky Cancer Registry also offers a county-level mapping application. New York State offers a limited set of ZIP code level data for the four most common cancer types in the mid-1990s. Currently, county-level cancer incidence data is not available nationally.

2. Population data

The United States Census Bureau is the principal source of data on the entire population; most countries have comparable agencies. Since cancer rates are calculated by dividing the number of cases by the number of people at risk, census data is frequently referred to as "denominator data". Census data are readily available in electronic format through the Census Bureau web site, http://www.census.gov. Data are available in three basic formats. American FactFinder is a web-based application that allows users to drill down through geographic levels to find data tables of interest. It is most useful for data queries that are well-focused. Data may also be downloaded through an ftp server. This method obtains raw text files that require computer code to be written before the data can be easily accessed or manipulated. This method is most useful for users with large data needs who are in possession of some database programming skills. The third approach is to purchase DVDs from the Census Bureau's Customer Serv-

ice center. The DVDs allow data output in many spreadsheet and database formats, facilitating the ability for users to process and analyze the data. There are also a large number of third-party vendors who offer similar services [52].

The four primary data files emanating from the 2000 census are named Summary File 1 through Summary File 4 (SF1–SF4). SF1 contains population counts by age, sex, race and ethnicity and basic housing characteristic information for the entire population, to the block level. SF2 contains similar information, detailed for ethnic subgroups, American Indian and Alaska Native tribes, and multiple-race individuals. These data are suppressed when the total number of individuals in a given geographic unit totals fewer than 100. SF3 contains detailed housing, demographic, and socioeconomic data to the census block group or census tract level, based on a long form that was sent to one in six households. Census block groups have an optimal population size of 1,500 and census tracts have an optimal population size of 4,000, though in practice populations vary widely. SF4 contains the same information as SF3 for detailed race and ethnic groups, with the same suppression rule as SF2. In addition to these four primary data files, the Census Bureau also provides digital cartographic boundary files for political entities in the country, as well as approximations of postal code boundaries known as ZIP code tabulation areas (ZCTAs).

The Census Bureau also conducts the American Community Survey (ACS), an ongoing survey designed to reach 3 million households each year nationwide. The goal of this survey is to allow the publication of detailed demographic and socioeconomic information more often than once a decade. Data for geographic units totaling more than 65,000 people will be released annually, while data for smaller geographic units will be based on either a three or five year moving average. It will replace the census long form, which will not be administered in 2010. There will undoubtedly be a challenging adjustment period as public health researchers begin to use ACS data.

At present, the level of information available for intercensal time points is quite limited, and derives from Census Bureau estimates at the state or county level. These estimates are used in the calculation of cancer rates by federal and state agencies, although some research has shown that they are not especially reliable, particularly county-level estimates for specific race groups [53]. Various private vendors publish intercensal estimates for areas smaller than counties, though it is impossible to verify their accuracy. Since many vendors use the Census Bureau estimates as controls (for example, vendor estimates of ZIP code populations in a county must add to the Census Bureau estimate for that county), vendor estimates necessarily suffer from the same limitations as the Census Bureau estimates. Finally, some state governments publish their own population estimates. Generally, these estimates are thought to represent improvements over the Census Bureau estimates because of higher levels of local knowledge and a broader use of data sources. We are unaware of any independent efforts to evaluate these claims, however. Examples include the population estimates and projections published by the California Department of Finance, and those by the Epidemiology Program of the Cancer Research Center of Hawaii. The latter population estimates were developed in response to a concern that the Native Hawaiian population was substantially undercounted in previous censuses, and are used by the NCI in calculating national cancer rates.

The 2000 census allowed respondents to select more than one race, although cancer data are only beginning to be collected in this manner. As a result, population data from 2000 must be "bridged" back to the earlier single-race categories to allow comparisons with earlier data. NCHS developed a sophisticated bridging algorithm taking into account age, sex, distribution of single-race groups within counties, and other covariates [54]. This algorithm is reflected in the 1991–2003 population projections and estimates that are published on the NCI web site and included in their statistical software. The Census Bureau itself uses a simpler algorithm in its estimates, allocating equal proportions of each multiple-race combination to the constituent single races [55]. Given the multiplicity of population estimates and methods for calculating them that are available, it is important to be aware of the sources of these data, and how they may influence the confidence associated with a particular research result. This is especially true for small-area analyses, where uncertainties are highest.

In addition to the issues noted above, it is important to realize that even the decennial census counts are not as accurate as popularly believed. The census represents an attempt to enumerate the population as of a single date, but invariably some people are missed or double-counted. These undercounts and overcounts are differential by race, socioeconomic status, and geographic area, potentially biasing cancer rates [56,57].

Countless epidemiologic and geographic studies make use of census data in some capacity, including most studies that report cancer rates for geographic areas. It is also quite common to use census data where individual-level data are not available, particularly for indicators of socioeconomic status [58-60], educational attainment [61] and housing characteristics [7]. Table 2 summarizes the population data sources described in this section.

Table 2: Sources of Population Data

Dataset name	Source Agency	URL	Geographic Resolution
2000 Census Summary Files 1–4	US Census Bureau	http://www.census.gov/main/www/cen2000.html	Census Tract, Block Group or Block (varies by data element)
American Community Survey	US Census Bureau	http://www.census.gov/acs/www/	Areas with populations >65,000
E-1 City/County Population Estimates, with Annual Percent Change	California Department of Finance	http://www.dof.ca.gov/html/Demograp/E-1text.htm	City/County
US Population Data, 1969–2001	National Cancer Institute	http://seer.cancer.gov/popdata/download.html	County

Table 3: Sources of survey data. Survey data recorded at the ZIP code level are designed to give valid estimates of risk factor distributions at the State level.

Dataset name	Source Agency	URL	Geographic Resolution
Behavioral Risk Factors Surveillance Survey (BRFSS)	Centers for Disease Control	http://www.cdc.gov/brfss	ZIP code
National Health and Nutrition Examination Survey (NHANES), National Health Care Survey (NHCS), National Health Interview Survey (NHIS), National Immunization Survey (NIS), National Survey of Family Growth (NSFG).	National Center for Health Statistics	http://www.cdc.gov/nchs/datawh/ftpserv/ftpdata/ftpdata.htm	Metropolitan Statistical Area, National Region
California Tobacco Survey	California Department of Health Services	http://www.surveyresearchgroup.com/clients.asp?ID=10	ZIP code
California Women's Health Survey	California Department of Health Services	http://www.dhs.ca.gov/director/owh/survey.htm http://www.surveyresearchgroup.com/clients.asp?ID=11	ZIP code
California Health Information Survey	UCLA Center for Health Policy Research	http://www.chis.ucla.edu/	ZIP code

3. Surveys

In addition to the Census Bureau as a primary source of sociodemographic attribute data, special survey data can provide valuable information on these characteristics for population groups in some areas. Perhaps one of the best known such surveys is the CDC-sponsored Behavioral Risk Factor Surveillance System (BRFSS), which is touted as the "world's largest telephone survey". Designed in the 1980s to track trends in behavioral risk factors at the state level, this ongoing system of national surveys also provides subarea and subgroup information within some of the larger states. Some researchers have estimated county-level behavioral risk factor prevalence by combining the statewide BRFSS data with county-level demographic data [62,63]. A mapping application to view BRFSS response data at the state and metropolitan level is also available http://apps.nccd.cdc.gov/gisbrfss/.

Another well-known national survey is the NCHS's National Health and Nutrition Examination Survey (NHANES), which has been in place since 1960 and combines questionnaire information with a national physical examination and biomonitoring program. NCHS also sponsors a National Health Care Survey (NHCS), a National Health Interview Survey (NHIS), a National Immunization Survey (NIS), and a National Survey of Family Growth (NSFG). Similarly designed large-scale efforts to track temporal and area differences for targeted health behaviors within a state include California's Tobacco Survey, Women's Health Survey, and Health Information Survey (Table 3).

Although population survey data has not been extensively incorporated into GIS studies to date, these resources may in the future provide some opportunity to characterize

regional differences in behavioral risk profiles targeted for specific health outcomes.

4. Environmental data

Over the past several decades there has been a large increase in the availability of spatially registered environmental data in the United States and other countries. Much of these data have been collected as a result of environmental regulations or government-funded research efforts. Examples of US programs to collect spatial data on concentrations or releases of pollutants in the environment include the United States Geological Survey (USGS) National Assessment of Water Quality program (NAWQA) http://water.usgs.gov/nawqa, the Environmental Protection Agency (EPA) National Air Toxics Assessment database http://www.epa.gov/ttn/atw, and EPA's Toxic Release Inventory program http://www.epa.gov/tri. EPA has organized environmental data in an umbrella database called Envirofacts Data Warehouse http://www.epa.gov/enviro/. Some states have extensive efforts to collect additional environmental data. An example is California's Pesticide Use Reporting program http://www.cdpr.ca.gov/docs/pur/purmain.htm) that requires reporting of all agricultural pesticide use at the level of Public Land Survey System sections (a unit approximately one square mile in area).

There are several issues to consider in using these data for assigning "exposure" in epidemiologic studies. Monitoring data collected for regulatory purposes should be carefully evaluated for its usefulness for estimating individual exposures. The fate and transport of the chemicals in the environment should also be considered. Simple proximity measures to sites of chemical releases may not adequately describe the transport of the chemical in the environment. The likely route of exposure should be considered along with the biological plausibility for an association between the exposure and disease under study. Finally, much of the environmental monitoring data was collected within the past decade and reconstructing exposure over longer periods more relevant to cancer incidence will be challenging.

Environmental databases have begun to be used in epidemiology studies of cancer to determine if disease mortality or incidence rates are higher in areas with specific environmental exposures (i.e., ecologic study designs) or as a means of classifying individuals with respect to their potential exposure in an analytic epidemiologic study design (i.e., case-control, cohort studies). With few exceptions, the residence location is used as the geographic location for assigning exposure. Below we provide an overview of the various types of spatially registered exposure data and include examples of their use in epidemiologic studies of cancer.

a. Water quality data

The US EPA is responsible for regulating public drinking water supplies. A water supply is regulated if it has 5 or more connections or serves at least 25 people. Routine monitoring is required for a variety of contaminants and naturally occurring elements including disinfection by-products, arsenic, nitrate, certain pesticides and volatile organic chemicals. States are required to report violations of the Maximum Contaminant Levels (MCL) to EPA. Since 1996, EPA has been required to maintain a National Contaminant Occurrence Database (NCOD) using occurrence data for both regulated and unregulated contaminants in public water systems. The majority of historical public water supply measurement data, however, reside with the states. Some states record the latitude and longitude of the locations where the water samples were taken (location in the distribution system, point of entry to the distribution system, or water source location). The location information is typically not publicly available but may be available to researchers with appropriate approvals.

The water quality data are reported by utility and to be useful for epidemiologic studies a linkage to the towns served must be established. In larger metropolitan areas multiple utilities may serve a city or, conversely, one utility may serve multiple towns and subdivisions. Therefore, establishing an accurate linkage between the study participant's addresses and water utilities is essential to avoid misclassification of exposure. Long-term exposure metrics can be calculated when a lifetime water source history is collected. Examples of studies using public supply water quality monitoring data include studies of disinfection by-products [64-66]., nitrate [67,68]., radionuclides [69,70]., and arsenic [71,72]. Contaminants such as disinfection by-products and volatile organic compounds vary in concentration across a public supply distribution system. GIS-based modeling efforts have been used to improve estimates of exposure at individual residences [73,74].

In contrast to public water supplies, private domestic wells are not regulated and there are no monitoring requirements, although well owners may be required to provide some water quality information upon the sale of a property in some states. Some states have conducted representative surveys of private well water quality [75]. A nationwide survey was conducted by EPA in 1988–1990 [76,77]. The US Centers for Disease Control (CDC) conducted a survey of coliform bacteria, nitrate, and atrazine in private wells in nine Midwestern States http://www.cdc.gov/nceh/emergency/WellWater/default.htm. The paucity of historical water quality data for private wells limits the exposure assessment for epidemiologic studies of cancer in this population.

Table 4: Sources of Water Quality data

Database name	Source Agency	URL	Geographic Resolution
National Contaminant Occurrence Database	EPA	http://www.epa.gov/safewater/data/ncod.html	Public water utility
National Water Quality Assessment (NAWQA) Data Warehouse	USGS	http://water.usgs.gov/nawqa/data http://waterdata.usgs.gov/nwis/qw	Latitude and longitude
Legacy Data Center/STORET	EPA	http://www.epa.gov/STORET/dbtop.html	Latitude and longitude

Table 5: Sources of Air Quality Data

Dataset name	Source Agency	URL	Geographic Resolution
Air Quality System database	EPA	http://www.epa.gov/air/data/aqsdb.html	Monitoring stations (latitude, longitude)
National Emissions Inventory	EPA	http://www.epa.gov/air/data/neidb.html	Varies (point locations, county level)

The USGS NAWQA program has been collecting information on nutrients, pesticides, volatile organic compounds, radionuclides, and major ions in more than 50 river basins and aquifers since 1991. All of the measurement data include spatial attributes. Because the goal of this research effort is to understand ambient water quality (not necessarily the same as drinking water quality) these data may not be of direct use in epidemiologic studies. However, the NAWQA data may be useful in modeling efforts to estimate contaminant levels in private wells. EPA also maintains two data management systems containing water quality information collected by federal, state, and private groups for surface and ground waters in all 50 states. The Legacy Data Center (LDC) is an archived database with data dating from the early 20th century up to the end of 1998. STORET contains data collected beginning in 1999, along with older data documented data from the LDC. Table 4 summarizes the sources of water quality data.

b. Air pollutants

The EPA collects and processes monitoring data from states on six criteria air pollutants (carbon monoxide, nitrogen dioxide, ozone, sulfur dioxide, particulate matter [PM10 and PM2.5], lead) and hazardous air pollutants, of which 188 have been identified. The hazardous air pollutants (HAP), also known as air toxics, are those for which there is some evidence of an increased risk for cancer or adverse reproductive outcomes. Routine monitoring of HAPs is not required and the monitoring data that exists is sparsely distributed compared with the criteria air pollutants. The data are maintained in the Air Quality Systems database.

EPA compiles HAP emissions from stationary sources (points and areas) and mobile sources in a National Toxics Inventory (NTI) database (now combined with the National Emissions Trends data in the National Emissions Inventory database), which is updated at three-year intervals. To do the updates, EPA obtains emissions inventories from state environmental agencies and supplemental data from other sources, including the Toxic Release Inventory. The first nationwide inventory was in 1996. The spatial scale of the emissions data varies by type of source. Location information for point sources emissions is available, whereas area-source emissions are estimated at the county level. Using a dispersion model EPA has estimated the annual average HAP concentrations for each census tract in the contiguous US [78]. These datasets are summarized in Table 5.

Air pollutant monitoring data has been used in studies of lung cancer, which have generally employed some type of dispersion model to estimate exposure for metropolitan areas or census tracts [79-81]. Recently the modeled concentrations of HAP have been used to evaluate childhood cancer incidence [44]. Other studies have also evaluated traffic density and childhood cancer incidence [43].

c. Agricultural Pesticides

In the United States the U.S. Department of Agriculture (USDA) is the main federal agency responsible for collect-

Table 6: Sources of Pesticide Data

Dataset name	Source Agency	URL	Geographic Resolution
Agricultural Chemical Use	USDA	http://usda.mannlib.cornell.edu/usda/usda.html	State
California Pesticide Use Reporting database	California Department of Pesticide Regulation	http://www.cdpr.ca.gov/docs/pur/purmain.htm	Public Land Survey Section (approximately one square mile)

ing information on pesticide use on crops and livestock. The availability of historical agricultural pesticide use data in the US has been reviewed [82]. The first comprehensive survey of pesticide use on crops occurred in 1964 [83] and periodic surveys were conducted thereafter through the 1970s. These early surveys only provided national or regional estimates of crop-specific use for individual pesticides. From 1986 onwards, the USDA surveys produced state-specific estimates of pesticide use on field crops in the major producing states and from 1990 onwards, biannual state-specific estimates of pesticide use on fruits and vegetables were also available.

Several states have collected their own pesticide use information but most data collection efforts have been recent. Oregon enacted legislation requiring reporting of agricultural pesticide use beginning in 2002; however, insufficient funding was provided for additional years. State pesticide use data are most comprehensive for California, which has had some type of mandatory reporting for agricultural pesticides since the 1950s, currently overseen by the California Department of Pesticide Regulation. Beginning in 1969, information about restricted-use pesticides was made public. In 1990, a new law required growers to report all pesticide use on crops on a monthly basis, including the pesticide name and manufacturer, crop treated, the public land survey section where the pesticide was applied, the date and time of application, number of acres treated, method of application, and application rates. The availability of this detailed pesticide use data at the spatial scale of a section led to the development of methods to link the use data to cancer incidence data [84] for use in an ecologic study of childhood cancer at the census tract level [42]. The California data have also been used in a case-control study of pancreas cancer [85], cohort study of breast cancer [51], and an as-yet unpublished case-control study of childhood cancer. Methods have also been developed to estimate potential pesticide exposure at residences by linking pesticide use data to crop maps [86,87]. Pesticide "exposure" is assigned to homes that have crop fields within distances that reflect likely pesticide drift. Table 6 summarizes the sources of pesticide data.

d. Industrial releases and hazardous waste

The Emergency Planning and Community Right to Know Act of 1986 in the United States requires certain industries to report to EPA annually their releases and waste management activities involving specific toxic chemicals. The data are available to the public in a database called the Toxics Release Inventory (TRI). Manufacturing, metal mining, coal mining, and electric generating facilities must report the estimated mass of toxic chemicals released into the environment (air, water, land, or underground injection), treated on-site, or shipped off-site for further waste treatment. Reporting is required only for facilities that meet certain minimum criteria in terms of the pounds of toxic chemical produced or processed; persistent chemicals that bioaccumulate are subject to lower minimum reporting requirements. The regulations do not require environmental monitoring, so much of the data are estimates of releases. Location information is reported by the business and is not verified by EPA. Some of the strengths and limitations of these data for environmental health studies has been described [88,89].

Canada also requires reporting of emissions of chemicals rated by the International Agency for Research on Cancer as likely, probable, and possible human carcinogens for 64 industrial sectors [90]. These data form part of the Canadian Environmental Quality Database, which also contains a national inventory of municipal waste disposal sites, municipal drinking water data, air quality data, and historical industrial location and productivity data [91]. A large multi-province case-control study of 18 cancer sites was conducted with the aim of linking residential histories by postal code to the environmental database for cancer surveillance. To date, one analysis of residential proximity to 7 types of heavy industries and risk of non-Hodgkin lymphoma (NHL) has been published. Residential proximity within 3.2 km of copper smelters and <0.8 km of sulfite pulp mills was associated with an increased risk of NHL [92] after adjusting for employment in the industries evaluated. Earlier case-control studies of NHL [93] and leukemia [94] found elevated risks for residing close to industrial sites but these studies relied on a self-reported assessment of the distance of the residence from industrial facilities which may be subject to recall bias.

Table 7: Sources of Hazardous Waste Data

Dataset name	Source Agency	URL	Geographic Resolution
Toxics Release Inventory	EPA	http://www.epa.gov/tri/	Latitude, longitude
HazDat	ATSDR	http://www.atsdr.cdc.gov/hazdat.html	Latitude, longitude
RCRAInfo	EPA	http://www.epa.gov/enviro/html/rcris/	Latitude, longitude
Permit Compliance System (PCS)		http://www.epa.gov/enviro/html/pcs/	

Table 8: Sources of Remote Sensing Data

Dataset name	Source Agency	URL	Geographic Resolution
Digital orthophoto quadrangles	USGS	http://edc.usgs.gov/products/aerial.html	1:12,000
Satellite imagery	USGS	http://edc.usgs.gov/products/satellite.html	1 meter to 1 km
National Landcover Dataset (NLCD) 1992 Multi-resolution Land Characteristics (MRLC) 2000	USGS	http://edc.usgs.gov/products/landcover.html	30 meters

The EPA maintains information on the location of waste handlers, waste treatment facilities and waste sites that are regulated under the Resource and Conservation Recovery Act (RCRA) and the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA), also known as the Superfund law in the RCRAInfo database available through the Envirofacts Data Warehouse. Information on the location of companies issued permits to discharge waste into rivers is maintained in the Permit Compliance System database (also available through Envirofacts). These data sources are summarized in Table 7.

The U.S. Agency for Toxic Substances and Disease Registry (ATSDR) was established by Congress in 1980 under CERCLA. Since 1986, ATSDR has been required to conduct a public health assessment at each of the sites on the EPA National Priorities List, waste sites deemed to be the most hazardous. The aim of these evaluations is evaluate exposure to hazardous substances and health effects among the population living in vicinity of the site [95]. The location of the sites and information on specific contaminants by the type of media (soil, air, water) in which they were measured are available from the ATSDR HazDat database web site. Limitations of these monitoring data for cancer studies include the limited historical measurement data. A few studies have evaluated cancer incidence among those potentially exposed to hazardous waste sites [96] or municipal waste sites and incinerators [97,98].

The reconstruction of historical exposure to releases from industries and waste sites is difficult for studies of cancers of long latency. A few studies have evaluated proximity and residence duration near sites. Long duration of residence within one-half mile of a chemical plant manufacturing PCBs was positively correlated with blood serum PCB concentrations [99]. However, none of the epidemiologic studies to date determined whether proximity resulted in meaningful exposure to chemicals from the sites. Confounding by socioeconomic status should also be evaluated because manufacturing and waste facilities are more likely to be located in neighborhoods of lower socioeconomic status [100] and socioeconomic status is associated with the incidence of some cancers.

5. Remote sensing/aerial imaging

Remotely sensed data include images of the earth and our atmosphere obtained by satellites or aircraft. The usefulness of the information depends largely on the technology used to obtain the imagery and the additional processing that has been done to georeference the data. The USGS Earth Resources Observation Systems Data Center (EDC) is the major U.S. storehouse of these data. Aerial photography has been available since the early part of the twentieth century. Digital Orthophoto Quadrangles (DOQs)

which are digital images of aerial photos which combine the image characteristics of a photo with the georeferenced qualities of a map are available through EDC from 1987 through the present. DOQs are available in black and white, natural color, or color-infrared images and have 1-meter ground resolution. Satellite imagery useful for land cover characterization includes the multispectral Landsat imagery available as early as 1972. USGS has created historical land use and land cover data derived from 1970s and 1980s aerial photography (the Land Use and Land Cover Data). A national land cover datasets (NLCD) derived from Landsat multispectral imagery for 1992 is available. The Multi-resolution Land Characteristics (MRLC) national dataset which represents land cover in 2000 is currently being developed. Table 8 summarizes these data sources. Applications of these data to studies of cancer have included mapping residences on crop maps to estimate their probable exposure to agricultural pesticides [49,87,101].

Centralized geospatial data availability

The data sources we have described are available from a multitude of federal and state agencies. The National Cancer Institute's Geographic Information Systems web site http://gis.cancer.gov offers links to many of these sources, as well as links to freely available geographical tools and resources. There have also been several initiatives to try and compile spatial data into a shared, centralized information system [102]. Such centralized systems offer the promise of standardized data coding systems, file formats and geographic boundary definitions. They also facilitate the sharing of metadata, or descriptive information about the data. The leader in this endeavor has been the Federal Geographic Data Committee http://www.fgdc.gov. The FGDC is a consortium of federal agencies with the charge of developing the National Spatial Data Infrastructure (NSDI), a set of technologies, policies, standards and procedures that facilitate the creation and sharing of geospatial data. Among the achievements of the FGDC is the establishment of the National Spatial Data Clearinghouse, a central catalog of links to geospatial data and metadata. In 2003, an enhanced web portal http://www.geodata.gov was launched to further facilitate access to this data. Many states have echoed the national clearinghouse with clearinghouses of their own. The New York GIS Clearinghouse http://www.nysgis.state.ny.us, for example, boasts over 400 member institutions providing links to thousands of datasets.

The cancer data collection community has yet to fully engage this resource. As of January 2004, no cancer incidence or mortality data was available through the national clearinghouse. The keyword "cancer" provided only a link to the Environmental Defense Scorecard, a web site from which various environmental data sets can be accessed, particularly those published by the EPA http://www.scorecard.org. Most of the very limited data in the "human health and disease" category accessible through the web portal consisted of hospital and other health facility locations for a smattering of states. In some cases, the steps required to make cancer data available through the national clearinghouse would be modest. For example, the NCI's mortality data, geographic boundary files, and associated metadata used in its Cancer Mortality Maps and Graphs web site are easily accessed and downloaded, and only minor modifications would be required to make them compliant with FGDC standards.

The DataWeb http://www.TheDataWeb.org is another centralized online data resource, consisting of a network of online data libraries created in a collaboration between the CDC and the US Census Bureau. The libraries consist of both microdata and aggregate data in numerous categories. Available health data includes NHANES and NHIS survey data and county-level mortality. Information in DataWeb is accessed through DataFerret, an application that prepares data sets for the user to download. It allows users to select a "databasket" of variables and then recode those variables as needed. Users develop and customize data tables and may download them to their desktop in a variety of common formats.

Conclusion

In this article we have surveyed the distinctive characteristics of spatial data, along with commonly available sources of data relevant to etiologic cancer research. Spatial analysis is invaluable for data exploration, identification of geographic patterns, generation of new hypotheses, and providing supporting evidence about existing hypotheses. A geographic perspective will be increasingly relevant as GIS software, spatial analytic methods, and the availability and quality of geographically referenced data continues to improve.

References

1. Lee CV, Irving JL: **Sources of spatial data for community health planning.** *Journal of Public Health Management and Practice* 1999, **5:**7-22.
2. Burkitt DP: **Geography of a disease: purpose and possibilities from geographical medicine.** In *Biocultural aspects of disease* Edited by: Rothschild HR. New York, Academic Press; 1981.
3. Gould P, Wallace R: **Spatial structures and scientific paradoxes in the AIDS pandemic.** *Geografiska Annaler B* 1994, **76:**105-116.
4. Miller HJ, Wentz EA: **Representation and spatial analysis in geographic information systems.** *Annals of the Association of American Geographers* 2003, **93:**574-594.
5. Jacquez GM: **Current practices in spatial analysis of cancer data: flies in the ointment, or, the limitations of spatial analysis.** *International Journal of Health Geographics* 2004, **3:**22.
6. Krieger N: **Place, space, and health: GIS and epidemiology.** *Epidemiology* 2003, **14:**384-385.
7. Boscoe FP, McLaughlin CC: **The effect of seasonal residence on cancer incidence rates.** *Journal of Registry Management* 2002, **29:**3-7.

8. McElroy JA, Remington PL, Trentham-Dietz A, Robert SA, Newcomb PA: **Geocoding addresses from a large population-based study: lessons learned.** *Epidemiology* 2003, **14:**399-407.
9. Bonner MR, Han D, Nie J, Rogerson P, Vena JE, Freudenheim JL: **Positional accuracy of geocoded addresses in epidemiologic research.** *Epidemiology* 2003, **14:**408-412.
10. Hurley SE, Saunders TM, Nivas R, Hertz A, Reynolds P: **Post office box addresses: a challenge for geographic information system-based studies.** *Epidemiology* 2003, **14:**386-391.
11. Fotheringham AS, Wong DWS: **The modifiable areal unit problem in multivariate statistical analysis.** *Environment and Planning A* 1991, **23:**1025-1044.
12. Openshaw S: ***The modifiable areal unit problem.*** In *Concepts and techniques in modern geography 38* Norwich, Geobooks; 1984.
13. Holt D, Steel DG, Tranmer M, Wrigley N: **Aggregation and ecological effects in geographically based data.** *Geographical Analysis* 1996, **28:**244-261.
14. Cohen BL: **Lung cancer rate versus mean radon level in US counties of various characteristics.** *Health Physics* 1997, **72:**114-119.
15. Lagarde F, Pershagen G: **Parallel analyses of individual and ecologic data on residential radon, cofactors, and lung cancer in Sweden.** *American Journal of Epidemiology* 1999, **149:**268-274.
16. Darby S, Deo H, Doll R, Whitley E: **A parallel analysis of individual and ecological data on residential radon and lung cancer in south-west England.** *Journal of the Royal Statistical Society A* 2001, **164:**193-203.
17. Lubin JH: **On the discrepancy between epidemiologic studies in individuals of lung cancer and residential radon and Cohen's ecologic regression.** *Health Physics* 1998, **75:**4-10.
18. Darby S, Hill D, Doll R: **Radon: a likely carcinogen at all exposures.** *Annals of Oncoogy* 2001, **12:**1341-1351.
19. Gotway CA, Young LJ: **Combining incompatible spatial data.** *Journal of the American Statistical Association* 2002, **97:**632-648.
20. Dobson JE, Bright EA, Coleman PR, Durfee RC, Worley BA: **LandScan: a global population database for estimating populations at risk.** *Photogrammetric Engineering and Remote Sensing* 2000, **66:**849-857.
21. Mennis J: **Generating surface models of population using dasymetric mapping.** *Professional Geographer* 2003, **55:**31-42.
22. Mugglin AS, Carlin BP, Zhu L, Conlon E: **Bayesian areal interpolation, estimation, and smoothing: an inferential approach for geographic information systems.** *Environment and Planning A* 1999, **31:**1337-1352.
23. Mugglin AS, Carlin BP, Gelfand AE: **Fully model-based approaches for spatially misaligned data.** *Journal of the American Statistical Association* 2000, **95:**877-887.
24. Twigg L, Moon G, Jones K: **Predicting small-area health-related behaviour: a comparison of smoking and drinking indicators.** *Social Science and Medicine* 2000, **50:**1109-1120.
25. Pearce J, Boyle P, Flowerdew R: **Predicting smoking behaviour in census output areas across Scotland.** *Health and Place* 2003, **9:**139-149.
26. Dubin RA: **Spatial autocorrelation: a primer.** *Journal of Housing Economics* 1998, **7:**304-327.
27. Griffith DA: *Spatial autocorrelation and spatial filtering: gaining understanding through theory and scientific visualization* New York, Springer-Verlag; 2003.
28. Fotheringham AS, Brunsdon C, Charlton M: *Geographically weighted regression: the analysis of spatially varying relationships* Chichester, John Wiley & Sons; 2002.
29. Brunsdon C, McClatchey J, Unwin DJ: **Spatial variation in the average rainfall-altitude relationship in Great Britain: an approach using geographically weighted regression.** *International Journal of Climatology* 2001, **21:**455-466.
30. Fotheringham AS, Brunsdon C: **Local forms of spatial analysis.** *Geographical Analysis* 1999, **31:**340-358.
31. Anselin L: **Spatial dependence and spatial structural instability in applied regression analysis.** *Journal of Regional Science* 1990, **30:**185-207.
32. Gelman A, Price PN: **All maps of parameter estimates are misleading.** *Statistics in Medicine* 1999, **18:**3221-3234.
33. Boscoe FP, Pickle LW: **Choosing geographic units for choropleth rate maps, with an emphasis on public health applications.** *Cartography and Geographic Information Science* 2003, **30:**237-248.
34. MacEachren AM, Brewer CA, Pickle LW: **Visualizing georeferenced data: representing reliability of health statistics.** *Environment and Planning A* 1998, **30:**1547-1561.
35. Kafadar K: **Smoothing geographical data, particularly rates of disease.** *Statistics in Medicine* 1996, **15:**2539-2560.
36. Talbot TO, Kulldorff M, Forand SP, Haley VB: **Evaluation of spatial filters to create smoothed maps of health data.** *Statistics in Medicine* 2000, **19:**2399-2408.
37. Ghosh M, Natarajan K, Stroud TWF, Carlin BP: **Generalized linear models for small-area estimation.** *Journal of the American Statistical Association* 1998, **93:**273-282.
38. Mollié A: **Bayesian and empirical Bayes approaches to disease mapping.** In In *Disease mapping and risk assessment for public health* Edited by: Lawson AB. Chichester, John Wiley & Sons; 1999:15-29.
39. Semenciw RM, Le ND, Marrett LD, Robson DL, Turner D, Walter SD: **Methodological issues in the development of the Canadian Cancer Incidence Atlas.** *Statistics in Medicine* 2000, **19:**2437-2449.
40. Toledano MB, Jarup L, Best NG, Wakefield JC, Elliott P: **Spatial variation and temporal trends of testicular cancer in Great Britain.** *British Journal of Cancer* 2001, **84:**1482-1487.
41. Dockerty JD, Sharples KJ, Borman B: **An assessment of spatial clustering of leukaemias and lymphomas among young people in New Zealand.** *Journal of Epidemiology and Community Health* 1999, **53:**154-158.
42. Reynolds P, Von Behren J, Gunier RB, Goldberg DE, Hertz A, Harnly ME: **Childhood cancer and agricultural pesticide use: an ecologic study in California.** *Environmental Health Perspectives* 2002, **110:**319-324.
43. Reynolds P, Von Behren J, Gunier RB, Goldberg DE, Hertz A, Smith D: **Traffic patterns and childhood cancer incidence rates in California, United States.** *Cancer Causes and Control* 2002, **13:**665-673.
44. Reynolds P, Von Behren J, Gunier RB, Goldberg DE, Hertz A, Smith DF: **Childhood cancer incidence rates and hazardous air pollutants in California: an exploratory analysis.** *Environmental Health Perspectives* 2003, **111:**663-668.
45. Reynolds P, Von Behren J, Gunier RB, Goldberg DE, Hertz A: **Residential exposure to traffic in California and childhood cancer.** *Epidemiology* 2004, **15:**6-12.
46. Crosignani P, Tittarelli A, Borgini A, Codazzi T, Rovelli A, Porro E, Contiero P, Bianchi N, Tagliabue G, Fissi R, Rossitto F, Berrino F: **Childhood leukemia and road traffic: A population-based case-control study.** *International Journal of Cancer* 2004, **108:**596-599.
47. Langholz B, Ebi KL, Thomas DC, Peters JM, London SJ: **Traffic density and the risk of childhood leukemia in a Los Angeles case-control study.** *Annals of Epidemiology* 2002, **12:**482-487.
48. Raaschou-Nielsen O, Hertel O, Thomsen BL, Olsen JH: **Air pollution from traffic at the residence of children with cancer.** *American Journal of Epidemiology* 2001, **153:**433-443.
49. Brody JG, Vorhees DJ, Melly SJ, Swedis SR, Drivas PJ, Rudel RA: **Using GIS and historical records to reconstruct residential exposure to large-scale pesticide application.** *Journal of Exposure Analysis and Environmental Epidemiology* 2002, **12:**64-80.
50. Brody JG, Aschengrau A, McKelvey W, Rudel RA, Swartz CH, Kennedy T: **Breast cancer risk and historical exposure to pesticides from wide-area applications assessed with GIS.** *Environmental Health Perspectives* 2004, **112:**889-897.
51. Reynolds P, Hurley SE, Goldberg DE, Yerabati S, Gunier RB, Hertz A, Anton-Culver H, Bernstein L, Deapen D, Horn-Ross PL, Peel D, Pinder R, Ross RK, West D, Wright WE, Ziogas A: **Residential proximity to agricultural pesticide use and incidence of breast cancer in the California Teachers Study cohort.** *Environmental Research* 2004, **96:**206-218.
52. Ralston B: *GIS and Public Data* Clifton Park, Delmar Learning; 2004.
53. Boscoe FP, Miller BA: **Population estimation error and its impact on 1991–1999 cancer rates.** *Professional Geographer* 2004, **56:**516-529.
54. Schenker N, Parker JD: **From single-race reporting to multiple-race reporting: using imputation methods to bridge the transition.** *Statistics in Medicine* 2003, **22:**1571-1587.
55. United States Census Bureau: *Estimates and projections area methodology: county population estimates by age, sex, race, and Hispanic origin for July 1, 2002* Washington DC, United States Census Bureau; 2003.

56. Robinson JG, West KK, Adlakha A: **Coverage of the population in Census 2000: results from demographic analysis.** *Population Research and Policy Review* 2002, **21:**19-38.
57. United States Census Bureau: *Technical summary of A.C.E. Revision II for the Committee on National Statistics* Washington DC, United States Census Bureau; 2003.
58. Krieger N, Quesenberry C Jr, Peng T, Horn-Ross P, Stewart S, Brown S, Swallen K, Guillermo T, Suh D, Alvarez-Martinez L, Ward F: **Social class, race/ethnicity, and incidence of breast, cervix, colon, lung, and prostate cancer among Asian, Black, Hispanic, and White residents of the San Francisco Bay Area, 1988–92 (United States).** *Cancer Causes and Control* 1999, **10:**525-537.
59. Robbins AS, Whittemore AS, Thom DH: **Differences in socioeconomic status and survival among White and Black men with prostate cancer.** *American Journal of Epidemiology* 2000, **151:**409-416.
60. Ghori FY, Gutterman-Litofsky DR, Jamal A, Yeung SCJ, Arem R, Sherman SI: **Socioeconomic factors and the presentation, management, and outcome of patients with differentiated thyroid carcinoma.** *Thyroid* 2002, **12:**1009-1016.
61. Kwok RK, Yankaskas BC: **The use of census data for determining race and education as SES indicators: A validation study.** *Annals of Epidemiology* 2001, **11:**171-177.
62. Pickle LW, Su Y: **Within-state geographic patterns of health insurance coverage and health risk factors in the United States.** *American Journal of Preventive Medicine* 2002, **22:**75-83.
63. Jia H, Muennig P, Borawski E: **Comparison of small-area analysis techniques for estimating county-level outcomes.** *American Journal of Preventive Medicine* 2004, **26:**453-460.
64. Cantor KP, Lynch CF, Hildesheim ME, Dosemeci M, Lubin J, Alavanja M, Craun G: **Drinking water source and chlorination byproducts. I. Risk of bladder cancer.** *Epidemiology* 1998, **9:**21-28.
65. Hildesheim ME, Cantor KP, Lynch CF, Dosemeci M, Lubin J, Alavanja M, Craun G: **Drinking water source and chlorination byproducts. II. Risk of colon and rectal cancers.** *Epidemiology* 1998, **9:**29-35.
66. King WD, Marrett LD: **Case-control study of bladder cancer and chlorination by-products in treated water (Ontario, Canada).** *Cancer Causes and Control* 1996, **7:**596-604.
67. Ward MH, Mark SD, Cantor KP, Weisenburger DD, Correa-Villasenor A, Zahm SH: **Drinking water nitrate and the risk of non-Hodgkin's lymphoma.** *Epidemiology* 1996, **7:**465-471.
68. Ward MH, Cantor KP, Riley D, Merkle S, Lynch CF: **Nitrate in public water supplies and risk of bladder cancer.** *Epidemiology* 2003, **14:**183-190.
69. Lyman GH, Lyman CG, Johnson W: **Association of leukemia with radium groundwater contamination.** *Journal of the American Medical Association* 1985, **254:**621-626.
70. Bean JA, Isacson P, Hahne RM, Kohler J: **Drinking water and cancer incidence in Iowa. II. Radioactivity in drinking water.** *American Journal of Epidemiology* 1982, **116:**924-932.
71. Ferreccio C, Gonzalez C, Milosavjlevic V, Marshall G, Sancha AM, Smith AH: **Lung cancer and arsenic concentrations in drinking water in Chile.** *Epidemiology* 2000, **11:**673-679.
72. Chiou HY, Hsueh YM, Liaw KF, Horng SF, Chiang MH, Pu YS, Lin JS, Huang CH, Chen CJ: **Incidence of internal cancers and ingested inorganic arsenic: a seven-year follow-up study in Taiwan.** *Cancer Research* 1995, **55:**1296-1300.
73. Cohn P, Savrin J, Fagliano J: **Mapping of volatile organic chemicals in New Jersey water systems.** *Journal of Exposure Analysis and Environmental Epidemiology* 1999, **9:**171-180.
74. Gallagher MD, Nuckols JR, Stallones L, Savitz DA: **Exposure to trihalomethanes and adverse pregnancy outcomes.** *Epidemiology* 1998, **9:**484-489.
75. Johnson CJ, Kross BC: **Continuing importance of nitrate contamination of groundwater and wells in rural areas.** *American Journal of Industrial Medicine* 1990, **18:**449-456.
76. Environmental Protection Agency: *National survey of pesticides in drinking water wells. Phase I report (EPA 570/9-90-015)* Washington DC, Environmental Protection Agency; 1990.
77. Environmental Protection Agency: *Another look: national survey of pesticides in drinking water wells. Phase 2 report (EPA 570/9-91-020)* Washington DC, Environmental Protection Agency; 1992.
78. Rosenbaum AS, Axelrad DA, Woodruff TJ, Wei YH, Ligocki MP, Cohen JP: **National estimates of outdoor air toxics concentrations.** *Journal of the Air and Waste Management Association* 1999, **49:**1138-1152.
79. Pope CA III, Burnett RT, Thun MJ, Calle EE, Krewski D, Ito K, Thurston GD: **Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution.** *Journal of the American Medical Association* 2002, **287:**1132-1141.
80. Nafstad P, Haheim LL, Oftedal B, Gram F, Holme I, Hjermann I, Leren P: **Lung cancer and air pollution: a 27 year follow up of 16 209 Norwegian men.** *Thorax* 2003, **58:**1071-1076.
81. Cohen AJ: **Outdoor air pollution and lung cancer.** *Environmental Health Perspectives* 2000, **108:**743-750.
82. Ward MH, Prince JR, Stewart PA, Zahm SH: **Determining the probability of pesticide exposures among migrant farmworkers: results from a feasibility study.** *American Journal of Industrial Medicine* 2001, **40:**538-553.
83. USDA Economic Research Service: *Quantities of pesticides used by farmers in 1964. Agricultural Economic Report No. AER-131. AER-131* Washington DC, USDA Economic Research Service; 1968.
84. Gunier RB, Harnly ME, Reynolds P, Hertz A, Von Behren J: **Agricultural pesticide use in California: pesticide prioritization, use densities, and population distributions for a childhood cancer study.** *Environmental Health Perspectives* 2001, **109:**1071-1078.
85. Clary T, Ritz B: **Pancreatic cancer mortality and organochlorine pesticide exposure in California, 1989–1996.** *American Journal of Industrial Medicine* 2003, **43:**306-313.
86. Ward MH, Nuckols JR, Weigel SJ, Maxwell SK, Cantor KP, Miller RS: **Identifying populations potentially exposed to agricultural pesticides using remote sensing and a Geographic Information System.** *Environmental Health Perspectives* 2000, **108:**5-12.
87. Rull RP, Ritz B: **Historical pesticide exposure in California using pesticide use reports and land-use surveys: an assessment of misclassification error and bias.** *Environmental Health Perspectives* 2003, **111:**1582-1589.
88. Neumann CM: **Improving the U.S. EPA Toxic Release Inventory database for environmental health research.** *Journal of Toxicology and Environmental Health B* 1998, **1:**259-270.
89. Neumann CM, Forman DL, Rothlein JE: **Hazard screening of chemical releases and environmental equity analysis of populations proximate to toxic release inventory facilities in Oregon.** *Environmental Health Perspectives* 1998, **106:**217-226.
90. Argo J: **Retrospective exposure assessment with emission inventories: a new approach to an old problem.** *Environmetrics* 1998, **9:**505-518.
91. Johnson KC, Mao Y, Argo J, Dubois S, Semenciw R, Lava J: **The National Enhanced Cancer Surveillance System: a case control approach to environment-related cancer surveillance in Canada.** *Environmetrics* 1998, **9:**495-504.
92. Johnson KC, Pan S, Fry R, Mao Y: **Residential proximity to industrial plants and non-Hodgkin lymphoma.** *Epidemiology* 2003, **14:**687-693.
93. Linos A, Blair A, Gibson RW, Everett G, Van Lier S, Cantor KP, Schuman L, Burmeister L: **Leukemia and non-Hodgkin's lymphoma and residential proximity to industrial plants.** *Archives of Environmental Health* 1991, **46:**70-74.
94. Shore DL, Sandler DP, Davey FR, McIntyre OR, Bloomfield CD: **Acute leukemia and residential proximity to potential sources of environmental pollutants.** *Archives of Environmental Health* 1993, **48:**414-420.
95. Amler RW, Lybarger JA: **Research program for neurotoxic disorders and other adverse health outcomes at hazardous chemical sites in the United States of America.** *Environmental Research* 1993, **61:**279-284.
96. White E, Aldrich TE: **Geographic studies of pediatric cancer near hazardous waste sites.** *Archives of Environmental Health* 1999, **54:**390-397.
97. Goldberg MS, al Homsi N, Goulet L, Riberdy H: **Incidence of cancer among persons living near a municipal solid waste landfill site in Montreal, Quebec.** *Archives of Environmental Health* 1995, **50:**416-424.
98. Comba P, Ascoli V, Belli S, Benedetti M, Gatti L, Ricci P, Tieghi A: **Risk of soft tissue sarcomas and residence in the neighbourhood of an incinerator of industrial wastes.** *Occupational and Environmental Medicine* 2003, **60:**680-683.
99. Orloff KG, Dearwent S, Metcalf S, Kathman S, Turner W: **Human exposure to polychlorinated biphenyls in a residential com-**

munity. *Archives of Environmental Contamination and Toxicology* 2003, **44:**125-131.
100. Perlin SA, Wong D, Sexton K: **Residential proximity to industrial sources of air pollution: interrelationships among race, poverty, and age.** *Journal of the Air and Waste Management Association* 2001, **51:**406-421.
101. Ward MH, Giglierano J, Nuckols JR, Wolter C, Colt JS, Camann D, Hartge P: **Proximity to crops and residential exposure to agricultural pesticides in Iowa.** *Proceedings, EUROHEIS/SAHSU Conference* 2003:140-145.
102. Falke SR: **Environmental data: finding it, sharing it, and using it.** *Journal of Urban Technology* 2002, **9:**111-124.