Title
POPULATIONS AT RISK TO ENVIRONMENTAL POLLUTION PROSPECTUS: FY 80 AND FY 81

Authors
Selvin, Steve
Sacks, Susan T.
Winkelstein, Warren
et al.

Publication Date
1980

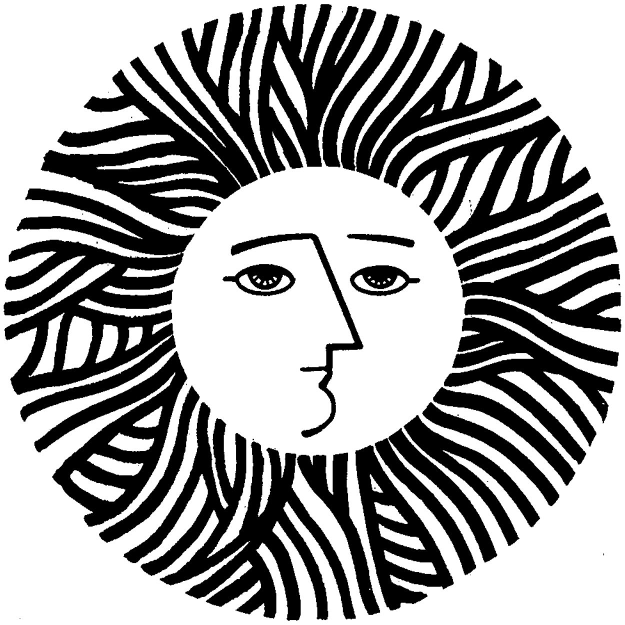
Lawrence Berkeley Laboratory
UNIVERSITY OF CALIFORNIA

ENERGY & ENVIRONMENT DIVISION

POPULATIONS AT RISK TO ENVIRONMENTAL POLLUTION PROSPECTUS: FY 80 and FY 81

Steve Selvin, Susan T. Sacks, Warren Winkelstein, Jr.,
Elizabeth Holly and Deane W. Merrill

January 1980



Prepared for the U.S. Department of Energy under Contract W-7405-ENG-48

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

LBID-164

POPULATIONS AT RISK TO ENVIRONMENTAL POLLUTION
PROSPECTUS: FY 80 and FY 81

Steve Selvin, Susan T. Sacks and Warren Winkelstein, Jr.
School of Public Health
University of California
and
Energy and Environment Division
Lawrence Berkeley Laboratory
Berkeley, California 94720

Elizabeth Holly
School of Public Health
University of California

Deane W. Merrill
Computer Science and Applied Mathematics Department
Lawrence Berkeley Laboratory

January, 1980

The work described in this report was funded by the Office of Health and Environmental Research, Assistant Secretary for Environment of the U.S. Department of Energy under Contract No. W-7405-ENG-48.

INTRODUCTION

The following describes a series of proposed projects to be undertaken by the Populations at Risk to Environmental Pollution (PAREP) project at Lawrence Berkeley Laboratory (LBL). Each project will be described in terms of the data and general analytic strategy to be used and will suggest directions for future study. The variables that comprise the data for these projects were extracted from the PAREP data base, and a complete list is available. Each project will lead to one or more LBL reports or scientific publications. The projects planned for FY-1980 and FY-1981 are:

1. Ecologic Patterns of Disease in the United States
2. Air Quality, Socio-economic Status and Cancer Incidence in the San Francisco Bay Area
3. Consequences of Ecologic Regression
4. The Association of Socio-demographic Measurements and Histologic Type for Nine Cancer Sites
5. Measurement of Disease for County Level Data

The work plan is given in Appendix I and the dictionaries of variables for the PAREP data sets are given in Appendix II. Descriptions of the five projects are given below.

1. ECOLOGIC PATTERNS OF DISEASE IN THE UNITED STATES

For this study, the data to be used will come from a county level file (3,082 U.S. counties) containing variables from three sources: (1) National Center for Health Statistics (mortality data made available by the Council for Environmental Quality, originally tabulated by H. Sauer, University of Missouri; (2) United States Census (1970) (demographic measurements); and (3) Environmental Protection Agency air-quality measurements.

Analytic Methods

Ecologic data have been criticized by various investigators as generating "spurious correlations" or "ecologic fallacies". In an effort to avoid problems encountered with ecologic data, the PAREP staff has developed an analytic strategy for exploring its county level data to describe and better understand patterns of disease in the United States.

The first level of analysis is descriptive (maps, charts, rankings, etc.) and employs a specially developed standardized measure of mortality. Also the county level air quality, a unique part of the PAREP data base, will be similarly described. The simple product-moment correlations among the 53 mortality rates will be computed. The epidemiologic significance of correlation coefficients is discussed in detail in "Correlations of Incidence Rates for Selected Cancers in the Nine Areas of the Third National Cancer Survey," W. Winkelstein, et al. (Am. J. Epid. 105: 407-419). Included is an analysis of mortality rates for "diseases" such as homocide, suicide and accidents because the etiology of these phenomena are understood to some extent and provide insights into the problems and strengths of ecologic data analysis.

The obvious shortcoming of descriptive analysis is that the multivariate aspects of complex relationships are ignored. To account for the simultaneous influence of socio-economic and air quality variables on the incidence of disease, multivariate linear regression will be used. While linear regression equations are useful in adjusting dependent mortality variables, they do not enable us to assess the relative contribution from each of a series of independent variables. The use of ecologic regression as a statistical tool to partition the influences of a dependent variable is the subject of project 3. Patterns of disease will be analyzed only after the linear influences of socio-economic variables and air quality are removed from the data. The study of these "residuals" will provide opportunity to investigate disease relationships in a partial-correlation sense. For example, it is known that gastro-intestinal cancers (stomach, colon and rectum) tend to be correlated across communities. The question arises: After socio-economic and/or air quality factors are statistically removed, what is the strength of the association across communities of gastro-intestinal cancer? The PAREP data will be used to examine a number of such hypotheses.

A second analytic technique, the use of principal components, will be used to determine the major contributors to the variability among counties from both socio-economic and air quality measurements. These components not only serve as a tool for identifying clusters of independent variables, but also, because they reduce a series of complex variables to a set of statistically simpler variables (canocial variables), they will be useful to us in describing the relationships among socio-economic levels, air quality measurements, and mortality rates.

Regardless of the analytic techniques used, however, certain problems will remain. For example, population migration from one county to another certainly influences disease patterns; but while this and similar issues will be identified and discussed, no good measure is available for determining the contribution of such variables to the data.

Future Applications

County level data can be used in a variety of ways in the future.

A few are:

(1) Age-specific mortality rates could be incorporated in the PAREP county level file. Since the very young (e.g., less than one year) and the very old (e.g., greater than 70) are believed to be the most susceptible to environmental insults, such data would enhance our understanding of disease patterns.

(2) Site-specific cancer mortality data could be correlated with the level of background ionizing radiation.

(3) The ratio of male-to-female mortality could be examined and the assumption that excess mortality among males is essentially due to occupational exposures could be further explored.

(4) Matching strategies could be performed (methodologically and analytically). Disease rates among counties could be compared as a means of distinguishing specific ecologic variables that may underlie the etiology of a given disease. The National Institute of Cancer used such methods to assess the influence of the petroleum industry on cancer mortality.

2. AIR QUALITY, SOCIO-ECONOMIC STATUS AND CANCER INCIDENCE IN THE SAN FRANCISCO BAY AREA.

The fundamental goal of this project is to determine whether or the extent to which routinely collected data on air quality has any bearing on epidemiology. The San Francisco Bay Area will serve as a prototype community for this study. Health data for this project will come from the best available cancer incidence data. Air quality data contained in the PAREP data base and derived from air quality monitoring station measurements have already been determined to be the best available.

The analysis of the data will be conducted in two stages. The first stage will consist essentially of descriptive statistics such as tables, maps, charts, and typical summary statistics (e.g., means, medians, etc.). No evaluation of statistical significance will be done in this phase but, rather, an attempt will be made to describe the air quality and disease patterns in the Bay Area on a census-tract level. For the second phase of analysis, statistics will be developed to define and assess the realtionships among the three factors under investigation. A "regression" approach will be used to test for trends in cancer rates with increasing pollution within socio-economic strata (e.g., using Cochran's test for linearity among categorical variables). The fundamental unit in this project is the cancer patient from the Third National Cancer Survey (TNCS) whose socio-economic (SES) status and air-quality exposure are determined by his/her census tract of residence. The validity of using such indirect measures of SES will be investigated in a corollary project (#4). Each patient will be assigned a value to indicate SES level and air-quality exposure. The data will then be classified into a series of tables, the basic table being:

Specific Age Category

	Level of Air Pollutant X				
SES	1 (lowest)	2	3	. . .	c (highest)
1 (lowest)	*	*	*		*
2	*	*	*		*
	.	.	.		.
	.	.	.		.
	.	.	.		.
r (highest)	*	*	*		*

* site-specific cancer incidence rate.

This type of table will be computed for a variety of ambient air contaminants (e.g., TSP, SO_2, NO_2, and O_3) by age categories (36-45, 46-55, 56-65, and 65+). The rows of these tables become the analytic unit; that is, each row represents the response to different levels of air quality for fairly homogeneous age and SES-levels. The dependent variables will be the age-specific incidence rates from cancer (44 sites are recorded in the TNCS data). Analyses will primarily concentrate on the most frequent cancers (e.g., lung, breast, etc.).

To further test this methodology and to check the consistency of the results of the San Francisco study, we anticipate extending the data derived from this study to the eight other TNCS areas and to mortality data from Los Angeles county (LA county has tracted mortality records).

3. CONSEQUENCES OF ECOLOGIC REGRESSION

Employing regression techniques on aggregated data is simple to execute but difficult to interpret. In a broad sense, the "ecologic fallacy" refers to the difficulty of interpreting regression coefficients calculated from aggregated data. This project will explore the issue termed "ecologic fallacy" and attempt to rigorously define the sources of bias incurred by applying linear models directly to aggregated data.

First, we will use Monte Carlo simulation techniques to compare ecologic regression (macro-model) with usual regression (micro-model). Artificial data for which relationships are known (e.g., linear) and associated statistical errors will be generated. These microdata will be analyzed by the usual techniques to provide baseline estimates and then aggregated according to a series of models. For example, the aggregations could be formed by ranking one of the independent variables and then dividing the data into a series of clusters. This model simulates the situation where geographic units such as counties are fairly homogeneous with respect to specific variables such as income. The results from these simulation experiments will both identify and quantify the "fallacies" that arise from direct application of regression techniques to aggregated data.

The next step will be to rigorously identify sources of bias, algebraically. Once the algebraic relationships are known, biases that arise in applied regression situations can be addressed. For example, with this information one could describe the degree and sources of bias in the estimates of regression coefficients associated with county median income in an ecologic regression. This project, essentially methodologic, complements projects 1 and 2 and represents a key element to understanding standard techniques applied to a type of data that is often readily available. The proper method of interpretation of regression equations applied to ecologic data is often ignored and, to our knowledge, has not been investigated.

4. THE ASSOCIATION OF SOCIO-DEMOGRAPHIC MEASUREMENTS AND HISTOLOGIC TYPE FOR NINE CANCER SITES

This project is designed to correlate certain socio-demographic characteristics of a population with the incidence of cancers of specific histologic types and sites selected for their relatively high frequency of occurrence. Again, data from the Third National Cancer Survey (TNCS) will be used. The cancer sites to be studied are: lung, thyroid, brain, cervix, uterus, corpus uteri, ovary, testis, as well as lymph, spleen, liver and neoplastic growth associated with Hodgkin's disease.

Two sources of data on education and income will be compared -- one derived from each TNCS patient's census tract of residence in the San Francisco Standard Metropolitan Statistical Area (SMSA) (indirect data) and the other, actual education and income levels for these patients, as recorded for a ten percent sample of the TNCS population (direct data). The data on individual patients will be extracted from a sample survey conducted by the National Cancer Institute (1970 data). Census tract information for the SMSA will be available from the PAREP project (project 2) and will be linked with the TNCS data. The comparison of these two measures of education and income will allow us to determine whether census tract information can be validly substituted for actual data gathered by survey.

Individuals will be classified into several categories (to be determined) by their actual levels of education and income (direct data). The same individuals will also be classified by the median educational and income levels of their tract of residence (as measured by 1970 U.S. census). The lack of agreement between these two methods of classification is a fundamental element for employing the census-derived characteristics to the study of disease (sometimes called indirect studies).

If census tract measures prove to be valid, we will examine the association between socio-demographic variables and histologic types of the nine site-specific cancers using census-tract data on median education and income for each of the nine TNCS areas (Atlanta, Birmingham, Dallas-Fort Worth, Detroit, Colorado, Iowa, Pittsburgh and San Francisco) and all areas combined. This analysis will be specific for race (black and white), age, sex, as well as histologic type. The usual biostatistical/epidemiologic techniques will be used (t-tests, analyses of variance, chi-square, etc.).

Third, seasonality by month of diagnosis and month of birth will be examined in relation to cancer in general and histologic type in particular. Hodgkin's disease is of special concern since the etiology of this neoplastic disease has been hypothesized to be infectious. Although the date of diagnosis is a poor indicator of the date of onset, it is the best available indicator of the beginning of a malignant neoplasm. Appropriate adjustment will be made for the uneven distribution of births throughout the year and for the variable number of days in the months. Usual chi-square "lack of fit" statistics will be used, and a ranking of cancer sites by degree of observed seasonality will be produced.

5. MEASUREMENT OF DISEASE FOR COUNTY LEVEL DATA

Important to any statistical analysis is choosing and measuring the dependent variable. In epidemiologic investigations, the rate of disease is a typical measurement. For county or any other geographically based data, the rate, or the age-adjusted rate, of disease is not an ideal measure for two reasons. First, for most rates of disease, distributions are extremely asymmetric (clustering occurs at small values). Secondly, a rate, similar to a percentage, does not reflect the numbers of individuals that were aggregated to estimate the risk of disease.

Previous investigators have not produced satisfactory methods for comparing disease rates in subpopulations such as counties. Maps of cancer mortality developed by the National Cancer Institute employ standardized direct age-adjusted rates using variance estimates from Chiang's work (Standard Error of Age-Adjusted Rates -- Public Health Service publication). This approach presents two problems: (1) comparing counties with zero deaths is not possible since the variance in this case is itself equal to zero and (2) the large amount of computing associated with the variance of the age-adjusted rate often results in prohibitive costs. The general mortality atlas to be produced by the National Center for Health Statistics employs percentiles -- essentially a rank procedure that lacks the sensitivity of actual rates, or function of rates. This procedure also ignores sampling variation; that is, if a single person dies of any cause in a county with a small population (e.g. Alpine, California = 160 males), the mortality rate is enormous and will be ranked at or near the top (e.g. Alpine, California = 625/100,000) when, in fact, the mortality experience of the county is not typical.

The PAREP project will test and describe several approaches based on age-adjusted rates (obtained by direct and indirect methods) to produce a variable that measures mortality and also takes into account county size. These standardized rates will also be defined when no deaths occur in a specific county. Such an easily interpreted and easily calculated measure of disease is necessary for describing disease patterns for both univariate or multivariate analyses.

APPENDIX I - WORK PLAN

A tentative plan for the work to be accomplished by the PAREP project for FY-80 and FY-81.

First-Quarter (FY-80)

1) Publication of an LBL technical report entitled "Measurement of Disease for County Level Data."

2) Draft of an LBL report on the ecologic mortality patterns in the United States as they relate to 15 specific socio-economic measures.

Second-quarter (FY-80)

1) LBL report entitled "Ecologic Mortality Patterns of 10 Specific Diseases."

2) Draft of an LBL report on the creation and data of the PAREP air-quality data base.

Third-quarter (FY-80)

1) Various presentations at the annual meetings of the Society for Epidemiologic Research.

2) LBL technical report on the influence of seasonality and birth month on the incidence of 40 sites of cancer.

Fourth-quarter (FY-80)

1) Technical report and/or draft of a publication entitled "Correlations of Mortality Rates for Selected Cancer in the United States."

2) Draft of an LBL report on the relationships of air quality in the San Francisco Bay Area and selected cancer incidence.

First-quarter (FY-81)

1) Publication of an LBL report and/or paper entitled "Cancer Incidence for Selected Sites and Air Quality in the San Francisco Bay Area."

2) Preliminary draft on U.S. mortality patterns of leukemia.

Second-quarter (FY-81)

1) Develop a work plan for studying "Epidemiologic Patterns of United States Leukemia Mortality" and prepare LBL report.

2) Preliminary draft of an LBL report focussing on the issues and consequences that arise from regression techniques applied to aggregated data.

Third-quarter (FY-81)

1) LBL report and/or publication entitled "Consequences of Ecologic Regression."

2) Various presentations at the meetings of the Society for Epidemiologic Research.

Fourth-quarter (FY-81)

Several projects could form reports/publications for the last quarter but cannot be described specifically at the present time. They include:

1) The issues surrounding the use of matching strategies when applied to aggregated data (e.g., county level data).

2) An investigation of the use of census tract measures as representative of the individuals who live in those tracts.

3) The epidemiology of a number of selected histologic sites of cancer with particular focus on the influences of socio-economic status.

4) A further investigation of leukemia mortality with special reference to "high-risk" industries (e.g., nuclear and coal power generation or nuclear and chemical waste disposal).

APPENDIX II - DICTIONARIES OF VARIABLES

Three dictionaries of variables for the principal PAREP data sets for the FY-80 projects are described below. Each data set is maintained in two forms--SS (binary and CODATA). The SPSS binary file is a quick and inexpensive format for usual statistical analysis (e.g., regression, analysis of variance, correlations, etc.). The second form is a specialized LBL structure that allows automatic interface with many of the powerful computer software tools developed by the Computer Science and Mathematics division (e.g., CARTE, CHART, and SEEDIS). Also the CODATA format allows for easy access by specialized FORTRAN programs.

These dictionaries can be obtained by contacting the authors at the School of Public Health, University of California, Berkeley, CA, 94720.

This report was done with support from the Department of Energy. Any conclusions or opinions expressed in this report represent solely those of the author(s) and not necessarily those of The Regents of the University of California, the Lawrence Berkeley Laboratory or the Department of Energy.

Reference to a company or product name does not imply approval or recommendation of the product by the University of California or the U.S. Department of Energy to the exclusion of others that may be suitable.

TECHNICAL INFORMATION DEPARTMENT
LAWRENCE BERKELEY LABORATORY
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720