# CANCER MORTALITY

## ENVIRONMENTAL AND ETHNIC FACTORS

*by*

Dorothy Gaites Wellington
Eleanor J. Macdonald
Patricia F. Wolt

# CANCER MORTALITY

## ENVIRONMENTAL AND ETHNIC FACTORS

*by*

Dorothy Gaites Wellington
Eleanor J. Macdonald
Patricia F. Wolf
Department of Epidemiology
The University of Texas System
Cancer Center
Texas Medical Center
Houston, Texas



ACADEMIC PRESS
NEW YORK SAN FRANCISCO LONDON 1979
A Subsidiary of Harcourt Brace Jovanovich, Publishers

ACADEMIC PRESS, INC.
111 Fifth Avenue, New York, New York 10003

*United Kingdom Edition published by*
ACADEMIC PRESS, INC. (LONDON) LTD.
24/28 Oval Road, London NW1 7DX

**Library of Congress in Cataloging in Publication Data**

Wellington, Dorothy Gaites.
Cancer mortality.

1. Cancer—Mortality. 2. Carcinogenesis. 3. Epidemiology. I. Macdonald, Eleanor J., joint author. II. Wolf, Patricia F., joint author. III. Title. [DNLM: 1. Neoplasms—Mortality—United States. 2. Environmental health—United States. 3. Ethnic groups—United States. QZ200.3 W452c]
RC261.W44 616.9'94'071 79-10560
ISBN 0-12-745850-6

PRINTED IN THE UNITED STATES OF AMERICA

79 80 81 82 9 8 7 6 5 4 3 2 1

# FOREWORD

In Dante Alighieri's search for St. Peter's Gate, geologists' search for oil, and epidemiologists' search for cancer etiology, geographic exploration is a necessary starting point. Stratigraphic exploration of hell under Virgil's expert guidance led Dante to an understanding of the route to salvation. Topographic, stratigraphic, and geomorphic maps of the earth's crust are essentially crude models of the association of local geographic features with the presence of petroleum. These models have on the whole proved to be useful guides in the search for petroleum deposits, even though they not infrequently lead to dry holes and to (pleasant) surprises, such as the recent Chinese and Mexican discoveries.

Systematic search for the many causes of the many cancers is reasonably analogous: crude descriptions of the macroepidemiology of the cancer family suggest strategies for investigations of next level, microepidemiology of individual cases, which in turn lead to laboratory studies of organs, cells, nuclei, and molecules in the search for explanatory mechanisms ("causes").

In the recent generation of epidemiologic science quantum leaps in the technological capability at all levels have set off a new wave of highly refined research that has engendered a new optimism about our ability ultimately to understand cancer processes.

Macdonald, who has spent a lifetime of infinite care developing and refining cancer registries as the foundation of modern microepidemiology, has turned now, with the statistical collaboration of Wellington, to drawing the zipper around the contemporary bag of best available and highest current analytical technology applicable to the macroepidemiology of cancer(s). While several recent studies have updated the classical macroepidemiology of cancer with detailed geographic mapping of mortality rates, by type and site of cancer, various levels of aggregation (in time and space), and have reconnoitered demographic and environmental correlates, Wellington, Macdonald, and Wolf have achieved a synthesis of the available data in a unified, comprehensive model, which step by step addresses the shortcomings of previous analyses.

The multitude of options as to data and analytic method were carefully sifted and tested before selection of an optimum model:

Particular U.S. mortality, demographic, and environmental variates are selected as, overall, of the highest quality available.

States (vis-a-vis counties and SMSAs) are selected as the optimum aggregation

level for segregation of 25 cancer types and sites and for meaningful impact of demographic and environmental correlates thereon.

The 20 years (1950 to 1969) are selected as the optimum time window—a period long enough to generate numbers sufficient for detailed subcategory analysis, late enough to assure uniformly high-quality data, and early enough (though barely) to escape the massive homogenization of U.S. culture that is rapidly obliterating regional variation in possible correlates of cancer. (The "cost" of this 20-year average choice, of course, is the foregone exploration of time trends. Such exploration, however desirable, is severely limited by the small frequencies occurring in short-time intervals.)

By exhaustive analysis of the components of variation in 23 demographic variables, 6 income, 11 climate, 37 air contamination, 3 radiation, 20 consumption (cigarettes, alcohol, water, milk), and 74 ethnic variables, a "best" explanatory set of four major "factor pools" associated with variation in cancer mortality is teased out (three environmental variables, one income, four consumption, and four ethnic variables). These 12 variables appropriately combined account for 30% (in female trachea, bronchus, and lung specified as primary) to 90% (in male rectum) of the observed variation in cancer mortality. With only slight loss of explanatory power the model can be reduced to six variables, which in turn reduce to two global constellations of urban factors and population density and its concomitants.

Finally, best-fitting multiple regression models with standardized coefficients, developed for uniform application to the 25 site/types, by sex, permit examination of the *net* effects of individual causal variables with others "held constant" and exploration of possible causal pathways. These models are appropriately corrected for the well-known latency, nonnormality, collinearity, and aliasing that have so persistently dogged previous studies.

If, as it may seem to most readers, the substantive findings merely confirm what we already knew or suspected or simply raise more of the same kind of vexing questions for which we have no immediate answer—or even access—it can nevertheless now be said that all that is to be learned from cross-section geographic analysis of available data on cancer mortality and its many possible correlates has been extracted. Further advances will require descent to minute investigation on lower levels of aggregation, for the exploration of which this study may serve epidemiology as Virgil served Dante. "Now let us onward, for the way is long . . . . Long is the journey and the road is rough . . . ."

*Carl E. Hopkins*
*Professor of Public Health*
*School of Public Health*
*University of California*
*Los Angeles, California*

# PREFACE

The study of death records has been a long-time concern of one of the authors (EJM), who published the first definitive paper evaluating the accuracy of cancer mortality records 40 years ago. The regional patterns of cancer mortality rates have been of special interest to the authors since their studies at M. D. Anderson Hospital in the early 1960s revealed the geographic relationship betweeen degree of urbanization and mortality level for some of the major types of cancer. In that project the age-adjusted death rates were calculated for each category of cancer mortality available in the national vital statistics reports by state and by subpopulation group for each year from 1940 through 1959. Time trends were calculated for three periods: 1940–1948, 1949–1959, and 1940–1959. Some of the results were presented at the IX International Cancer Congress in Japan in 1966, and were summarized in an article in *Cancer*, 1967. Not until 1975 were the authors able to direct their attention again to this study. At that time Mason and McKay made available age-adjusted death rates averaged over the period 1950–1969, and it was decided to base further analysis on this later data set, using the results from the earlier period for comparability.

In spite of the problems involved in deriving models based on vital statistics and the available data on economic and sociological variables, preliminary results proved to be reasonable and consistent, and at the same time posed questions that stimulated more detailed and more comprehensive investigation. The goal of this book has been to present in an organized and comprehensive manner the sets of factors whose joint by-state variation best explain the state patterns of cancer mortality and to explore the nonconformation of individual states to the national models. The authors hope that not only will the factor effects that emerge in the models be of interest but that attention may be called to potential etiologies that are implicated through their association with the model effects, such as a nutrition habit that underlies an ethnic effect.

Without the assistance of Alan Romano in managing the data bank and providing computer services, it would not have been possible to accomplish the very large amount of computer analysis with such ease, speed, and flexibility. Paul Callen, Hugh Bray, Jacqueline Wheat, and Lynn Hayward were also helpful in the early stages of the computer programming. Of great importance was the continual availability of sufficient computer time, which enabled unrestricted analysis of the data and immediate investigation into new leads as they arose in the analysis.

The authors are appreciative of those who aided in the preparation of this work: Evelyn B. Heinze, Margaret C. Murphy, Margaret Jansen, John Hanna, Eleanor Hassett, Kay Hermes, Kaye Reed, and Miriam Toloudis.

This project was supported in part by The University of Texas System Cancer Center, USPHS Grant FR-00254, NCI Grant CS-9299, Texas State Department of Health, American Cancer Society, Texas Division, Regional Medical Program Grants RMA-00007 and RMD-00007, Special Project of the Council for Tobacco Research, U.S.A., Inc., T. D. Wellington, Princeton, New Jersey, and The Clayton Fund, Houston, Texas.

*Dorothy Gaites Wellington*
*Eleanor J. Macdonald*
*Patricia F. Wolf*

# TABLES

# CONTENTS

# Chapter 1
# Description and Purpose of the Study

## I Background and Purpose of the Study

In the United States there are large differences among the states in death rates from different types of cancer and also from other causes of death. In an early study by Macdonald *et al.* (1) the by-state distributions of the age-adjusted death rates for selected primary cancer sites indicated that these differences are largely regional, while states within the same geographic region have similar rates, and even adjacent regions experience closer mortality levels than those farther apart. Large differences among counties within a state do occur, however, as shown in Mason and McKay's publication of cancer death rates by state and by county (2), but frequently the high-risk counties are found in contiguous clusters (3). The very small population base of many counties will produce more volatile estimated rates, especially for the less common cancer primary sites, and when population subsets are involved, such as white males, the chance variation increases. Even an entire state, if it is as sparsely populated as Nevada, will exhibit extremes of high and low death rates not exhibited by the more populous states. The volatility of these estimated rates is modified and their large variance decreased when they are combined into 10- or 20-year averages, as they are in the Macdonald and the Mason and McKay projects cited above.

In recent years it has been recognized that a person's risk of developing cancer is influenced by environmental, consumption, and genetic factors, and by his response to those factors. His risk of dying from cancer also is influenced by the medical facilities available to him. Because of the complex of potentially interactive factors, an individual person is the natural unit for investigating the etiology of each type of cancer mortality, and many studies have been made on comparatively small numbers of individuals with the purpose of associating one or more factors with the incidence of, or mortality from, particular cancer

primaries. The knowledge gained from these studies can be supplemented or reinforced by more generalized population studies. Just as microeconomics, the study of the individual firm or the individual consumer, and macroeconomics, the study of mass flows of prices, income, and spending in large populations, combine to explain economic processes, so a "microepidemiology" that deals with the experience of samples of individuals chosen to represent larger populations can be combined with a "macroepidemiology" that analyzes the experience of the target populations themselves, to delineate etiological factors of disease. The data in the microstudies must be gathered with a high degree of accuracy and under explicit rules of sampling so that inferences to the parent population can be made, and they entail great expense. Statistical estimates derived from the larger data masses used in macrostudies are less vulnerable to the errors in individual records unless those errors result from a strong bias in collection or recording. In the vital statistics of large and complete populations there are massive flows that, like the ocean currents, may be rippled by disturbances, but are usually not diverted from their overall patterns. Since data of this kind are continuously provided by governmental bodies, it behooves researchers to put them to maximum use, as urged in a 1977 editorial in the American Journal of Public Health (4).

The purpose of this study has been to carry out a systematic analysis of state patterns of cancer mortality in all categories of malignant neoplasms for which the data were available and to determine the syndrome of state characteristics associated with a high level of each type of cancer mortality in its population (white). Previously McDonald and Schwing (5) had combined multiple linear regression and ridge regression techniques for considering a large number of factor variables in a model for state variation in death rates, but only one category of mortality was analyzed—deaths from all causes in the total population. Later Breslow and Enstrom (6) used multiple linear regression models to explain the state variation in selected cancer death rates, but with a limited set of factor variables from which to draw the mortality models. Carnow and Meier (7) considered multiple regression inadequate for identifying the factors responsible for differing lung cancer death rates because the intercorrelation among the factor variables and their close relationship with an overall urban or population density factor would lead to "aliasing," i.e., the disguising of one factor's effect under the label of another. Instead they chose a measure of one of the common air pollutants, benzo[*a*]pyrene, as a single index of pollution "to represent the effects of all of the correlated pollution variables combined" (7) and combined it with a per capita measure of cigarette sales in a two-factor multiple regression model of lung cancer mortality. Since the publication of the cancer mortality rates by county as well as by state (2) several studies have employed the epidemiologic technique of geographic correlation between death rate and potential etiologic

factors using either selected counties (8-10) or all 3056 of them (11). In another study, Lave and Seskin (12) derived multiple regression models using data from 117 Standard Metropolitan Statistical Areas (SMSAs).

In the all-counties study a multiple regression model relates lung cancer mortality to county demographic and occupational indices that include a percentage urban factor in addition to a redundant percentage rural factor. The SMSA study includes variables measuring both population size and density. The insertion of a variable representing either urbanization or population density into epidemiological mortality models has become almost as standard a procedure as including an income variable in econometric models, but the usage is quite different. The income effect has been thoroughly investigated in economic theory and adjustment for its influence is well understood. On the other hand the urban effect in mortality models has been observed but not well analyzed, and its influence attributed to a number of associated characteristics ranging from measurable and specific to unmeasurable and vague, such as "tension, stress, and unhealthy personal habits" (12). The inclusion of an urban variable in order to adjust for all its associated factors, both known and unknown, essentially begs the question. What is needed is to determine which components of the urban factor, and in what combinations and relative strengths, best explain the state mortality patterns for different types of cancer. When such a factor is adjusted for the overall urban effect of which it is a part, it is thereby partially corrected for itself, weakening or obliterating its potential effect in the model. It is akin to adjusting the factors of beer or wine consumption for the variable comprising total consumption of alcohol.

Further confusion of model effects has arisen when the urban variable has been replaced by a measure of population density with which it is only weakly correlated. The rationale attending the use of a population density variable is even less clearly defined, and the subfactors cited to be associated with it are fewer. The intention of the study presented here is to explain the strong effects that both these generalized variables display in multiple regression models of mortality patterns by *replacing* them with their associated component factors. In this study a very wide range of factor variables were tested in preliminary modeling, and from them were chosen four sets of variables to make up the four variable pools. Each variable pool was used to derive the multiple regression models for every type of cancer mortality *in order to provide a comparability of model effects across all categories of cancer deaths.* In additional analysis these four variable pools, which were constructed specifically for the derivation of cancer mortality models, were also used to derive models of state mortality patterns for all the major categories of death. These served as a type of control, providing the contrast that profiled those factor combinations specifically characteristic of the cancer mortality patterns.

## II Statistical Methodology and Computer Programs

For each mortality pattern multiple regression models were derived by linear least squares methodology and the stability of the coefficient estimates was tested by ridge regression techniques. Subset models were chosen from each of the four variable pools by two computer program packages, which employed different criteria for choosing the "best" set of factor variables to explain each mortality pattern. One was the stepwise multiple regression program 2R from the Biomedical Computer Programs of UCLA (15), referred to henceforth as BMD, and the other the LINCUR regression program (16) developed and described by Daniel and Wood (17), referred to as LIN. In the former, the $F$ value to enter the model was set at 1.5 and the $F$ to remove at 1.0, with the tolerance level set at .01, thereby excluding variables 99% of whose variation could be expressed in linear terms of the variables already entered. The LINCUR choice of the best set of variables was guided by the plot proposed by Mallows (18) of the $C_p$-statistic versus $p$, the number of coefficient estimates. The $C_p$-statistic is an estimate of the standardized "total squared error," the sum of the squared random errors of the dependent variable plus the squared bias at each point due to estimating that point by the derived model rather than by the "true" equation. Only the models with the lowest $C_p$ values were examined, and the patterns of factors entering these models were taken into consideration in the choice. When the chosen LIN model was not the one with the smallest $C_p$ value but rather with the second or third lowest, it is indicated in the text as "second or third LIN." If the model's $C_p$ value was greater than $p$, indicating some bias, it was usually not chosen, and the few cases included are so notated.

When the best subset models were decided upon either by the stepwise procedure of the BMD program or the minimum $C_p$-statistic of the LINCUR program or both, each was run as a full model in the LINCUR program to obtain an analysis of the distribution of its residuals. Included in the LINCUR printout are plots of the cumulative distribution of residuals, of the residuals against estimated mortality rate, and of component effect plus residual against factor value for each independent variable in the model. The component effect of the $i$th factor on the response variable after correction for all the other factor effects, measured at the $j$th observation, is $b_i(x_{ij} - \bar{x}_i)$, where $b_i$ is the estimated coefficient of factor $i$ in the model, $x_{ij}$ is the value of the $i$th factor at the $j$th observation, and $\bar{x}_i$ is the mean value of the $i$th factor. As suggested by Wood (19) these plots were used to estimate the influence of individual observations working through each factor variable in the model, and to detect outlier observations with respect to each model effect. These are similar to the partial residual plots that Larsen and McCleary (20) compare with the usual residual plot against factor variable in that the latter show deviations from linearity while the former show "both the extent of the deviation from linearity and the extent and

the direction of the linearity" (20). Individual states that represented residuals of extreme value were examined for their ranking in factor values and mortality rate to learn why they did not conform to the general model. Besides the outliers in the residual distributions, potential outliers in the influence space were also identified by the LINCUR output of weighted squared standardized differences. The specific influence on the chosen model of any state that constituted an extreme value in the space of model effects was determined by the changes in model effects when that state was excluded from the derivation.

The intercorrelation that is unavoidable among factor variables of the kind used in this study can lead to unstable coefficient estimates and inflated standard errors when least squares procedures are used. Ridge regression analysis, first proposed by Hoerl and Kennard (21, 22) controls the variance resulting from high collinearity by augmenting the diagonal of the normal equations' matrix by small increasing values, producing slightly biased but more stable coefficient estimates. Although variables are not eliminated by this procedure, those whose coefficient estimates decrease almost to zero as the augmentation increases are considered not to maintain their effect in the model in competition against related factors with more stable traces of their coefficient estimates. This technique was applied only to the largest variable pool, examining the plotted trace of the estimated coefficients for each of the variables to determine the relative strength and stability of the model effects chosen for each mortality pattern.

The computer program used for the ridge regression analysis was provided by S. Radhakrishnan of Shell Oil Company.

There are 49 data points in each regression model: the 48 continental states and the District of Columbia. Since these 49 cases constitute the whole population of interest, the statistics in multiple regression, which were developed to enable inference from a sample to the whole population, were used primarily as guidelines to relative judgment about the size or importance of the coefficient of a model factor. Even though, in one sense, the model coefficients are themselves the population coefficients, and thus techniques such as $F$ tests, $C_p$-statistics, and ridge traces of their stability as estimates appear unnecessary, these statistical procedures served as tools in choosing the best subsets of factors for each mortality pattern, in indicating the relative importance of each factor and in focusing on those factors whose effects were the most stable in the face of strong collinearity.

The aim of the statistical methodology was not predictive per se, since the particular conjuncture of circumstances covering those years under study will never reoccur, but rather it was to find which combinations of factors best explained each cancer mortality pattern and which individual factors showed strength or consistency in both the male and the female mortality patterns for a particular cancer primary, or in the mortality patterns that are related by type of cancer primaries, such as those of the digestive organs.

## III The Data

The dependent or response variables are the death rates of white males and of white females for each of the continental 48 states and the District of Columbia. The rates used in deriving the cancer mortality models were Mason and McKay's (2) age-adjusted death rates for each state averaged over the 20-year period 1950-1969. The one exception was the division of the deaths from cancer of the trachea, bronchus, and lung into those specified to be the primary site and those neither specified as primary nor secondary. The rates for these two subcategories are given by Burbank (13) and are averages of the 18-year period 1950-1967. The mortality models for the major categories of death were derived from age-adjusted rates averaged over the 11-year period 1949-1959, which were part of an earlier study at The University of Texas System Cancer Center M. D. Anderson Hospital (1). Also included in the preliminary analysis but not in the final models of this study were age-adjusted cancer death rates averaged over the 9-year period 1940-1948, the 11-year period 1949-1959, and the 20-year period 1940-1959, all of which were done in a previous M. D. Anderson study (1).

Data for the independent or factor variables were gathered from all the available sources, stored in a working data matrix in a CDC 6400 computer, and were continuously on call, enabling the maximum amount of test modeling and immediate feedback in the preliminary stages of determining the most effective pools of variables from which to choose the cancer mortality models. A listing of the basic data and sources from which the factor variables were drawn is given in Table I. A list of the dependent variables, the state age-adjusted death rates, and sources is given in Table II.

Natural logarithms were taken of the values of all the independent variables in order to normalize their generally skewed distributions due to the concentration of a few very high values in a very few states and the completely urban District of Columbia. The natural logarithm values were standardized by subtracting their mean (centering) and dividing by their standard deviation (scaling). Marquardt and Snee (14) point out that centering removes the nonessential ill-conditioning and some of the inflated variance of the coefficient estimates that accompanies collinearity. Strong intercorrelation among the factor variables in this study required this adjustment for mean values. The scaling forced all variances to the value of 1, standardizing the coefficient estimates, and thereby enabling the relative size of each coefficient to reflect the relative importance of each factor within each model. When a principal components analysis was used to combine several factor variables into a single composite factor variable, the calculations were made on the transformed values of the composing variables. The resultant first principal component scores were themselves standardized in order to maintain homogeneity of variances in the independent variables.