INFERENCES ABOUT POPULATION ATTRIBUTABLE RISK FROM CROSS-SECTIONAL STUDIES

Joseph L. Fleiss

March 26, 1979

PREPARED UNDER THE AUSPICES OF
SIAM INSTITUTE FOR MATHEMATICS AND SOCIETY

DIVISIONS OF BIOSTATISTICS AND EPIDEMIOLOGY, &
DEPARTMENT OF MATHEMATICAL STATISTICS
COLUMBIA UNIVERSITY
NEW YORK, NEW YORK

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

Prepared For
THE U.S. DEPARTMENT OF ENERGY
UNDER CONTRACT NO. EY-76-S-02-2874
DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or 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 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. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.
DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.
INFERENCES ABOUT POPULATION ATTRIBUTABLE RISK FROM CROSS-SECTIONAL STUDIES

by

Joseph L. Fleiss
Division of Biostatistics
Columbia University School of Public Health
600 West 168 Street
New York, N.Y. 10032

March 26, 1979

STUDY ON STATISTICS AND ENVIRONMENTAL FACTORS IN HEALTH

PREPARED UNDER SUPPORT TO SIMS FROM

DEPARTMENT OF ENERGY (DOE)
ENVIRONMENTAL PROTECTION AGENCY (EPA)
NATIONAL SCIENCE FOUNDATION (NSF)
ROCKEFELLER FOUNDATION
SLOAN FOUNDATION

DIVISION OF BIOSTATISTICS
SCHOOL OF PUBLIC HEALTH
COLUMBIA UNIVERSITY
NEW YORK, NEW YORK

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED
The population attributable risk (PAR) is defined as the fraction of all cases of a disease in a population due to exposure to a given risk factor. Letting $P_e$ denote the proportion of the population exposed to the risk factor and $R$ the relative risk, PAR may be defined (see 1) as

$$\text{PAR} = \frac{P_e(R-1)}{P_e(R-1)+1}.$$

Both Levin (1) and more recent students of PAR (2,3,4) have considered its estimation from case-control studies, when it may be assumed that the odds ratio closely approximates $R$ and that the rate of exposure in the control group closely approximates $P_e$. Under these assumptions, Walter (3) derived an expression for the approximate standard error of $\ln(1-\text{PAR})$, where $\text{PAR}$ denotes the estimated PAR and $\ln$ denotes the natural logarithm.

Little attention seems to have been paid to the problem of drawing inferences about PAR when the data are collected in cross-sectional surveys (such as those conducted by the National Center for Health Statistics) or as part of routine registration (such as the recording of number of live births and numbers of infant deaths in a city). Walter (5) considered this problem, and derived a complicated expression for the standard error of $\text{PAR}$. Below, a relatively simple expression for the standard error of $\ln(1-\text{PAR})$ is presented.

Suppose the results of a cross-sectional study are tabulated as follows:

<table>
<thead>
<tr>
<th>Outcome Condition</th>
<th>Present</th>
<th>Absent</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>a</td>
<td>b</td>
<td>a + b</td>
</tr>
<tr>
<td>Not Exposed</td>
<td>c</td>
<td>d</td>
<td>c + d</td>
</tr>
<tr>
<td>Total</td>
<td>a+c</td>
<td>b+d</td>
<td>n</td>
</tr>
</tbody>
</table>
Because the estimate of $R$ is $a(c+d)/c(a+b)$ and the estimate of $P_e$ is $(a+b)/n$, the estimate of $\text{PAR}$ for this kind of study becomes, after simplification,

$$\hat{\text{PAR}} = \frac{ad - bc}{(a+c)(c+d)}.$$

Using standard methods appropriate for large samples (6), it may be shown that $\ln(1 - \hat{\text{PAR}})$ is approximately normally distributed with mean $\ln(1 - \hat{\text{PAR}})$ and standard error

$$s = \text{s.e.} (\ln(1 - \hat{\text{PAR}})) = \sqrt{\frac{b + (a+d)\hat{\text{PAR}}}{n}}.$$

This formula is related to Walter's (see, e.g., formula 2 in (7)) by

$$s = \text{s.e.}(\hat{\text{PAR}})/(1 - \hat{\text{PAR}}).$$

An approximate 95% confidence interval for $\text{PAR}$ is $(\hat{\text{PAR}}_L, \hat{\text{PAR}}_U)$, where

$$\hat{\text{PAR}}_L = 1 - \exp(\ln(1 - \hat{\text{PAR}}) + 1.96s)$$

and

$$\hat{\text{PAR}}_U = 1 - \exp(\ln(1 - \hat{\text{PAR}}) - 1.96s).$$

As an example, consider the following data from Eric County, N.Y. on levels of suspended particulates and mortality over a three year period (1959-1961) from cancer of the prostate (8). The individuals were all white males aged 50-69 years residing in areas with a median family income between approximately $6,000 and $12,000 per year.

<table>
<thead>
<tr>
<th>Mean Daily Level of Suspended Particulates</th>
<th>Died due to Prostatic Cancer</th>
<th>Alive, or Died of Other Causes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$&gt;_100$ $\mu$g/m$^3$</td>
<td>9</td>
<td>8,456</td>
<td>8,465</td>
</tr>
<tr>
<td>$&lt;=$100 $\mu$g/m$^3$</td>
<td>22</td>
<td>34,897</td>
<td>34,919</td>
</tr>
<tr>
<td>Total</td>
<td>31</td>
<td>43,353</td>
<td>43,384</td>
</tr>
</tbody>
</table>
The estimated PAR is
\[ \hat{\text{PAR}} = \frac{9 \times 34,897 - 8,456 \times 22}{31 \times 34,919} = 0.118, \]
indicating that about 12% of deaths due to prostatic cancer in the study group may have been due to residence in areas with high levels of suspended particulates.

The estimated standard error of \( \ln(1-\text{PAR}) \) is
\[ \hat{s} = \sqrt{\frac{8,456 + (9 + 34,897) \times 0.118}{43,353 \times 22}} = 0.115. \]

The lower 95% confidence bound on PAR is therefore
\[ \text{PAR}_L = 1 - \exp(-0.126 + 1.96 \times 0.115) \]
\[ = 1 - 1.105 = 0.105, \]
and the upper 95% bound is
\[ \text{PAR}_U = 1 - \exp(-0.126 - 1.96 \times 0.115) \]
\[ = 1 - 0.704 = 0.296. \]

The interval includes the value 0, which is consistent with the absence of a statistically significant association between level of suspended particulates (as dichotomized here) and death due to prostatic cancer. Also consistent with the data, however, is the possibility that as many as 30% of all deaths due to prostatic cancer could have been eliminated if daily levels of suspended particulates had been reduced below 100 \( \mu \text{g/m}^3 \).
REFERENCES


