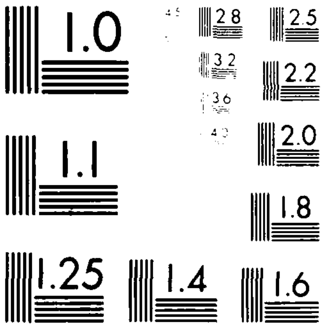
AD-A170 383   USING EMPLOYMENT DATA IN EPIDEMIOLOGY(U) PUBLIC   1/1
RESEARCH INST ALEXANDRIA VA J M JONDROW ET AL. JUN 84
CRM-84-12 N00014-83-C-0725

UNCLASSIFIED   F/G 6/5   NL



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

2

CRM 84-12 / June 1984

AD-A170 383

RESEARCH MEMORANDUM

USING EMPLOYMENT DATA IN EPIDEMIOLOGY

James M. Jondrow
Leon Hunt

N00014-83-C-0725

DTIC
S ELECTE D
JUL 3 0 1986
D

DTIC FILE COPY

CNA
The Public Research Institute
A Division of the Center for Naval Analyses

CRM 84-12 / June 1984

USING EMPLOYMENT DATA IN EPIDEMIOLOGY

James M. Jondrow
Leon Hunt

The Public Research Institute
A Division of the Center for Naval Analyses
2000 North Beauregard Street, Alexandria, Virginia 22311

ABSTRACT

Data have recently become available that can be used to shed light on the determinants of rare but catastrophic diseases such as leukemia. These data are accounting-type records from the unemployment insurance systems in particular states. Because these data cover most workers (whether unemployed or not), they provide the huge samples needed to analyze these diseases. Their combination with health data and their role relative to other data sets are described in this paper.

Accesion For
NTIS CRA&I ☑
DTIC TAB □
Unannounced □
Justification
By per form 50
Distribution/
Availability Codes
Dist | Avail and/or Special
A-1

TABLE OF CONTENTS

INTRODUCTION

One of the responsibilities of the federal government is to protect the populace from a wide variety of threats to health and life. OSHA regulates work-place hazards, EPA regulates ambient hazards, and NIH, NIOSH, and FDA seek to find and inform the public of links between industrial chemicals or food additives and serious diseases such as specific forms of cancer.

Setting exposure standards for hazardous substances requires reliable information on how people respond to varying doses of these substances. A typical dose-response curve for carcinogens is shown in figure 1. On the vertical axis is the percent of a population developing a specific tumor; on the horizontal axis is the cumulative lifetime dose of a specific carcinogen.

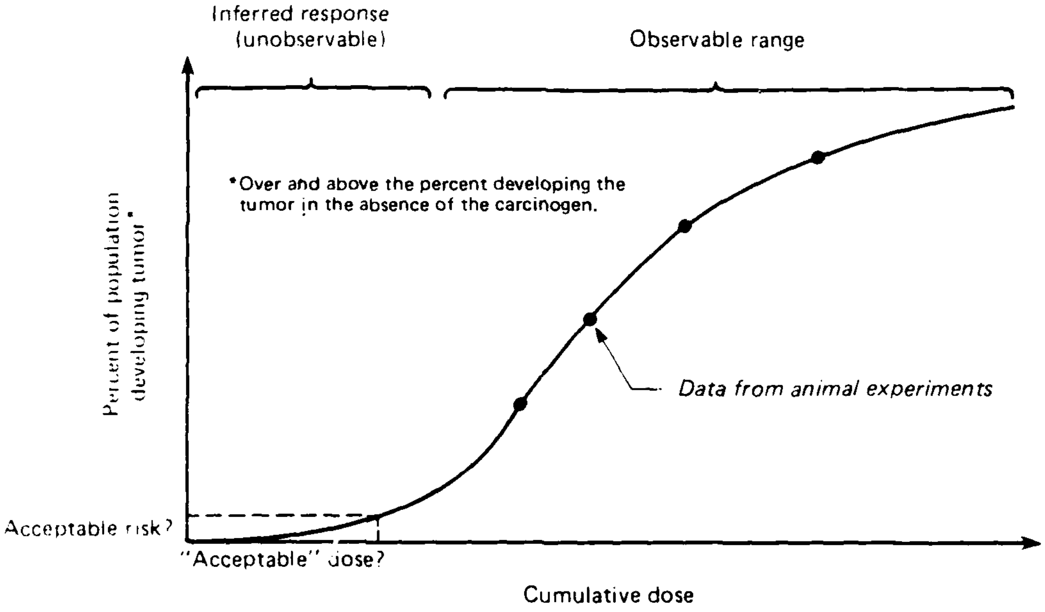


FIG. 1: TYPICAL DOSE RESPONSE CURVE FOR A CARCINOGEN

The exposures of primary interest to the regulator are generally too small to be observed but must be inferred from much higher doses in animal experiments. The degree of "acceptable risk" and the corresponding acceptable level of exposure to the substance are conjectural. Setting standards involves finding some exposure level in this range which is rationally defensible.

THE EMPLOYMENT (OR UI) DATA

The Public Research Institute is developing several large data sets that show promise for estimating dose-response curves at very low exposure levels. These data are based on the administrative or "bookkeeping" records of the unemployment insurance (UI) system in several states. They provide information about individual workers, employers, and UI applicants. As a consequence, complete employment histories can be constructed for individual workers. Though these records are from the UI system, they are not limited to UI claimants; they constitute virtually complete histories of occupational experience in the work force in those states, and--to the degree that occupational histories may be associated with exposures to industrial carcinogens--they provide the most complete record available for study of the potential effects of known or suspected carcinogens.

These data--matched with health records--are a powerful tool for epidemiological studies. They provide the kind of massive data base needed for study of rare but serious diseases. They can distinguish the effect of exposure in an industry (occupational risk) from the effect of residence in an area (ambient risk), although residence must be approximated by length of employment with employers in the area.

Despite their advantages, the UI data do not provide all the information necessary for an ideal epidemiological study. (Probably no such data set exists.) The following paragraphs describe how the UI data can and cannot be used and compare them with other data sets.

HOW THE UI DATA CAN BE USED

We start with an example of how the UI data can be used in epidemiology. The first step is to match the employment data with data on the incidence of the condition to be studied. The most easily accessible sources of health data would be death certificates and tumor registries.

Suppose the issue is whether a particular substance (benzene, for example) increases the incidence of a rare but very serious condition, such as leukemia. Suppose, too, that the substance poses a risk both to those who work with it and those who live near it--that is, there is an occupational risk and a risk to the general population from ambient exposures.

If both occupation and the environment contribute to the risk, disentangling their separate effects is crucial to measuring <u>either</u>. A cross-sectional study (which uses observations on disease incidence and industrial structure by area) cannot make this separation, because it is not clear whether the industrial structure in the area is measuring a risk to those who work in hazardous industries or merely a risk to those who live near them.

The separation can be made, however, with the matched UI and health data. In an equation expressing the relation of disease incidence to exposure, the variable to be explained is the incidence of the condition. Independent variables would include the individual's tenure in the industry and in the vicinity. (We currently do not have life-style variables, such as smoking, though they may be obtainable.)

Statistical estimation of this equation would result in coefficients measuring the increase in the probability of developing the disease as tenure in a particular industry increased by one year, and the increase in disease probability resulting from an extra year in the vicinity.

In reference to the dose-response curve shown in figure 1, tenure is a surrogate for the dose. Tenure also serves, in a statistical sense, to control for lack of information on life-style variables. Omitting life-style variables creates a statistical bias only to the extent that they correlate with other independent variables. Tenure seems less likely to be correlated with life style than is occupation, which is the usual surrogate for dose.

THE ERGODIC ASSUMPTION IN MORTALITY STUDIES OF RARE DISEASES

Most mortality studies compare the observed cases of a disease in a specific population with the expected number of cases of that disease in a comparison population having the same age, race, and sex structure. The expected cases are actually a synthetic estimate (in the statistical sense) based on recorded age-specific, race-specific, and sex-specific rates of the disease in the general population.

Typically, the study population is observed over some period of time, and its "exposure" is measured in terms of age-specific person-years. The fundamental assumption (which is never explicitly stated) in all such mortality studies is that person-years are independent units of risk, regardless of how they are accumulated. For instance, the expected mortality for a 29-year-old white male and a 30-year-old white male, each observed for 1 year, is assumed to be exactly the same as a 2-year observation of a white male beginning at age 29. (This assumption is conceptually equivalent to the ergodic assumption on probability theory: the time average is the same as the ensemble average.) In industries with high turnover (which is now common in many major SICs), the assumption means that the effects of long job tenure cannot be distinguished, since continuity of employment is ignored in accumulating person-years at risk.

UI data provide a unique opportunity to test this crucial assumption for low-frequency, long-latency conditions like cancers.

IDEAL DATA FOR EPIDEMIOLOGICAL STUDY

The ideal data for epidemiological studies have the following characteristics:

- They include a full population, as opposed to a sample, of those at high risk of the condition of interest.
- Observations on exposure and on time are for individuals.
- The precise dose of the hazard can be measured for each individual.
- Precise information on morbidity and mortality is available for each individual.
- A very large sample is available so that rare conditions (a few cases per 100,000 per year) are observable.
- Information about such life-style factors as as smoking, diet, and general health is available for each individual.
- Results of the epidemiological studies are available immediately (or at least within a year or two).
- There are no selection biases; that is, individuals are not selected for analysis because they are at exceptional risk of the disease.

THE NEED FOR COMPROMISE WITH THE IDEAL

In practice, no data sets have all these characteristics. The reason is well illustrated by imagining the cost of combining the requirements for large sample size and for detailed monitoring of individual exposure. Thus each epidemiological study is a compromise between the requirements of the ideal data and the data that are available.

The nature of the best compromise will depend on the nature of the hazard and the type of damage it does. For example, if the hazard is a type of air pollution that causes illness but not fatalities, and if the illness is a common response, then sample size is not crucial and can be sacrificed to obtain data on life style and perhaps exact exposure. On the other hand, suppose the illness is rare, but lethal--leukemia, for example. Then virtually nothing can be learned in small samples (say 10,000 observations), and large sample size becomes crucial.

ALTERNATIVE DATA SOURCES

Aside from the UI data, there are a variety of other data sets that have been used or can be used for epidemiological studies. They fit into these categories:

- Long-period historical studies of individuals with records of occupation and life style. A good example is the Dorn data set, which has followed 300,000 World War I veterans.

- Cross-sectional information on disease incidence, occupational structure, etc., by SMSA.

- Retrospective samples of those with a particular condition, as well as with a control group of those without the condition.

- National health samples with detailed information on a relatively small sample of individuals not concentrated geographically.

- Prospective samples created by monitoring a group of people from now into the future.

THE UI DATA AND THE IDEAL

The UI data provide rich information, but they are by no means ideal. Table 1 compares the strengths and weaknesses of the UI data with those of other data options.

As can be seen in the table, each data set has "holes." The point is not to evaluate the data sets on the basis of the number of pluses and minuses for each one, since that would involve the incorrect assumption that each advantage is equally important. The point is to show that none of the data sets are without major flaws and to identify their niches.

For example, prospective studies offer an almost unique ability to take account of life style and to monitor dosage closely, but they are prohibitively expensive (and hence cannot generate massive sample sizes). Further, the results will only be available in a distant future, far removed from present decisions on policy.

TABLE 1

COMPARISON OF DATA SETS

	UI data	Long period	Health	Retrospective	Cross section	Prospective
Individual data	X	X	X	X		X
Full population for a given geographical area	X					X
Precise dose						X
Morbidity/mortality			X		X	X
Huge sample	X	X				X
Life-style information		X	X	X	X	X
Results available early	X	X	X	X	X	
No selection biases	X	X	X		X	X
Results pertain to current occupational and ambient conditions	X		X	X	X	
Tenure information	X			X		X
Expensive	X	X	X	X	X	

The UI data appear to fill a particularly useful niche. They can pull in the full population from a geographical area; they can provide the huge sample necessary to spot rare disorders, and at reasonable cost; and they can provide some separation of occupational and ambient hazards (though they do not have information on residence, only workplace). On the other hand, they cannot provide life-style data or precise dose or response information. Thus, their value lies in the opportunity they offer to use tenure by industry and area to isolate the effect of cumulative occupational and ambient hazards as causes of rare but disastrous conditions such as the cancers. Since the mandate for much federal regulation concerns such hazards and such conditions, it appears that these data can be quite valuable indeed.

END

DTIC

8 — 86