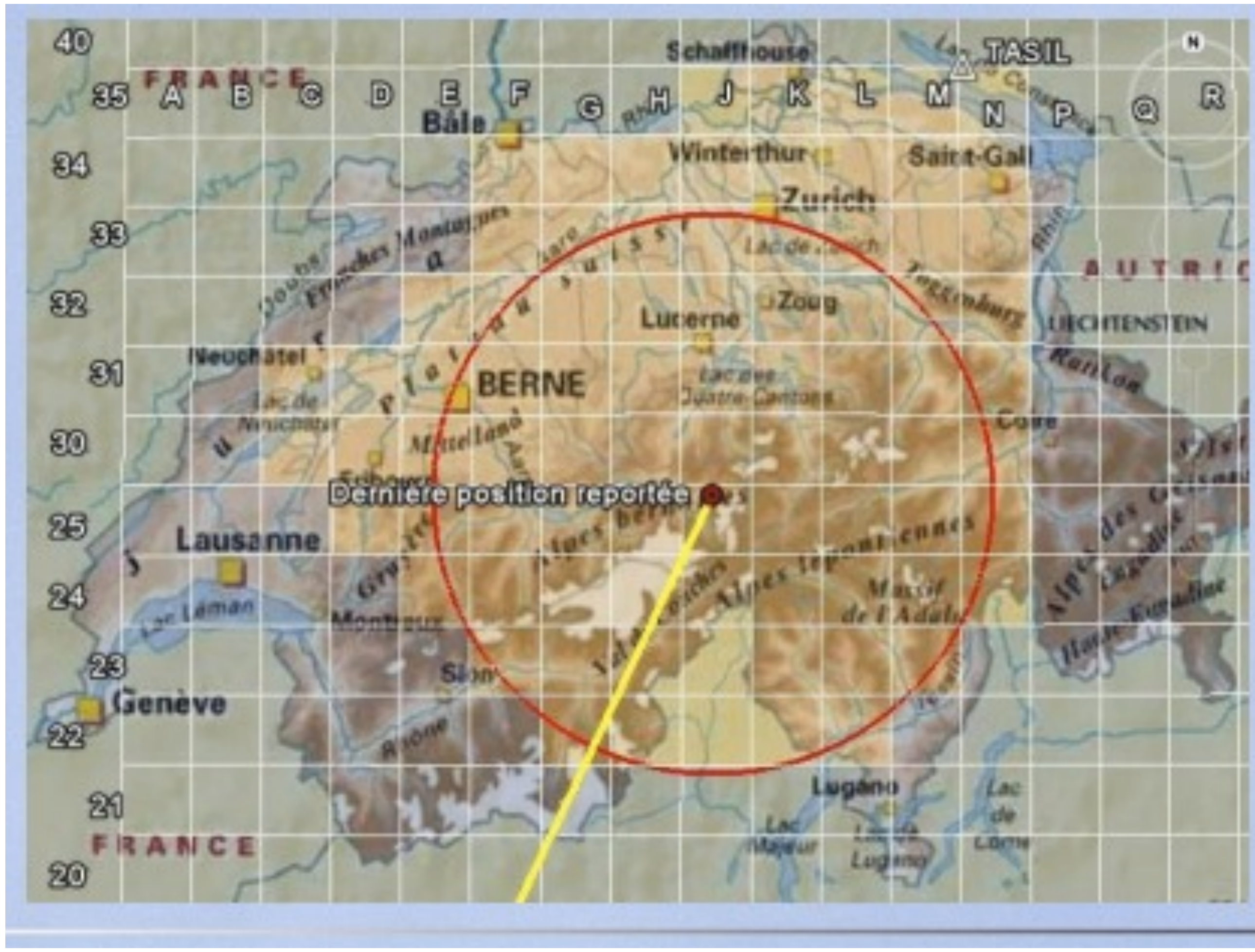

40
35
34
33
32
31
30
25
24
23
22
21
20
A
B
C
D
E
F
G
H
J
K
L
M
N
P
Q
R
FRANCE
Schaffhouse
TASIL
Bâle
Winterthur
Saint-Gall
Zurich
Lac de Zurich
Zoug
Lucerne
Lac des Quatre-Cantons
Neuchâtel
Lac de Neuchâtel
BERNE
Mittelland
Plateau suisse
Franches Montagnes
Doubs
Jura
Toggenburg
AUTRICHE
LIECHTENSTEIN
Rätikon
Coire
Dernière position reportée
Alpes bernoises
Alpes lépontiennes
Massif de l'Adula
Alpes valaisannes
Lausanne
Lac Léman
Montreux
Gruyère
Sion
Rhône
Genève
Lugano
Lac Majeur
Lac de Lugano
Lac de Côme
Alpes des Grisons
Engadine
Haute-Engadine
FRANCE

BEA

A METHOD OF ESTIMATING COMPARATIVE RATES FROM CLINICAL DATA. APPLICATIONS TO CANCER OF THE LUNG, BREAST, AND CERVIX[1]

JEROME CORNFIELD, *National Cancer Institute, National Institutes of Health, U. S. Public Health Service, Bethesda, Md.*

A frequent problem in epidemiological research is the attempt to determine whether the probability of having or incurring a stated disease, such as cancer of the lung, during a specified interval of time is related to the possession of a certain characteristic, such as smoking. In principle, such a question offers no difficulty. One selects representative groups of persons having and not having the characteristic and determines the percentage in each group who have or develop the disease during this time period. This yields a true rate. The difference in the magnitudes of the rates for those possessing and lacking the characteristic indicates the strength of the association. If it were true, for example, that a very large percentage of cigarette smokers eventually contracted lung cancer, this would suggest the possibility that tobacco is a strong carcinogen.

An investigation that involves selecting representative groups of those having and not having a characteristic is expensive and time consuming, however, and is rarely if ever used. Actual practice in the field is to take two groups presumed to be representative of persons who do and do not have the disease and determine the percentage in each group who have the characteristic. Thus rather than determine the percentage of smokers and nonsmokers who have cancer of the lung, one determines the percentage of persons with and without cancer of the lung who are smokers. This yields, not a true rate, but rather what is usually referred to as a relative frequency. Relative frequencies can be computed with comparative ease from hospital or other clinical records, and in consequence most investigations based on clinical records yield nothing but relative frequencies. The difference in the magnitudes of the relative frequencies does not indicate the strength of the association, however. Even if it were true that there were many more smokers among those with lung cancer than among those without it, this would not by itself suggest whether tobacco was a weak or a strong carcinogen. We are consequently interested in whether it is possible to deduce the rates from knowledge of the relative frequencies.

[1] Received for publication February 23, 1951.

A GENERAL METHOD

To fix our ideas we may illustrate how the general problem can be attacked with some data recently published by Schrek, Baker, Ballard, and Dolgoff (*1*). They report that 77 percent of the white males studied, aged 40–49, with cancer of the lung, smoked 10 or more cigarettes per day, while only 58 percent of a group of white males, aged 40–49, presumed to be representative of the non-lung-cancer population, smoked that much. Can we estimate from these data the frequency with which cancer of the lung occurs among smokers and nonsmokers?

Denote by p_1 (=0.77) the proportion of smokers among those with cancer of the lung, by p_2 (=0.58) the proportion of smokers among those without cancer of the lung, and by X the proportion of the general population that has cancer of the lung during a specified period of time. We may then summarize the relevant information for the general population in a two-by-two table showing the proportion of the population falling in each of the four possible categories.

Characteristic	Having cancer of the lung	Not having cancer of the lung
Smokers	p_1X	$p_2(1-X)$
Nonsmokers	$(1-p_1)X$	$(1-p_2)(1-X)$
Total	X	$1-X$

One can now compute that the percentage of the general population that smokes is $p_2+X(p_1-p_2)$, that the proportion of smokers having cancer of the lung is:

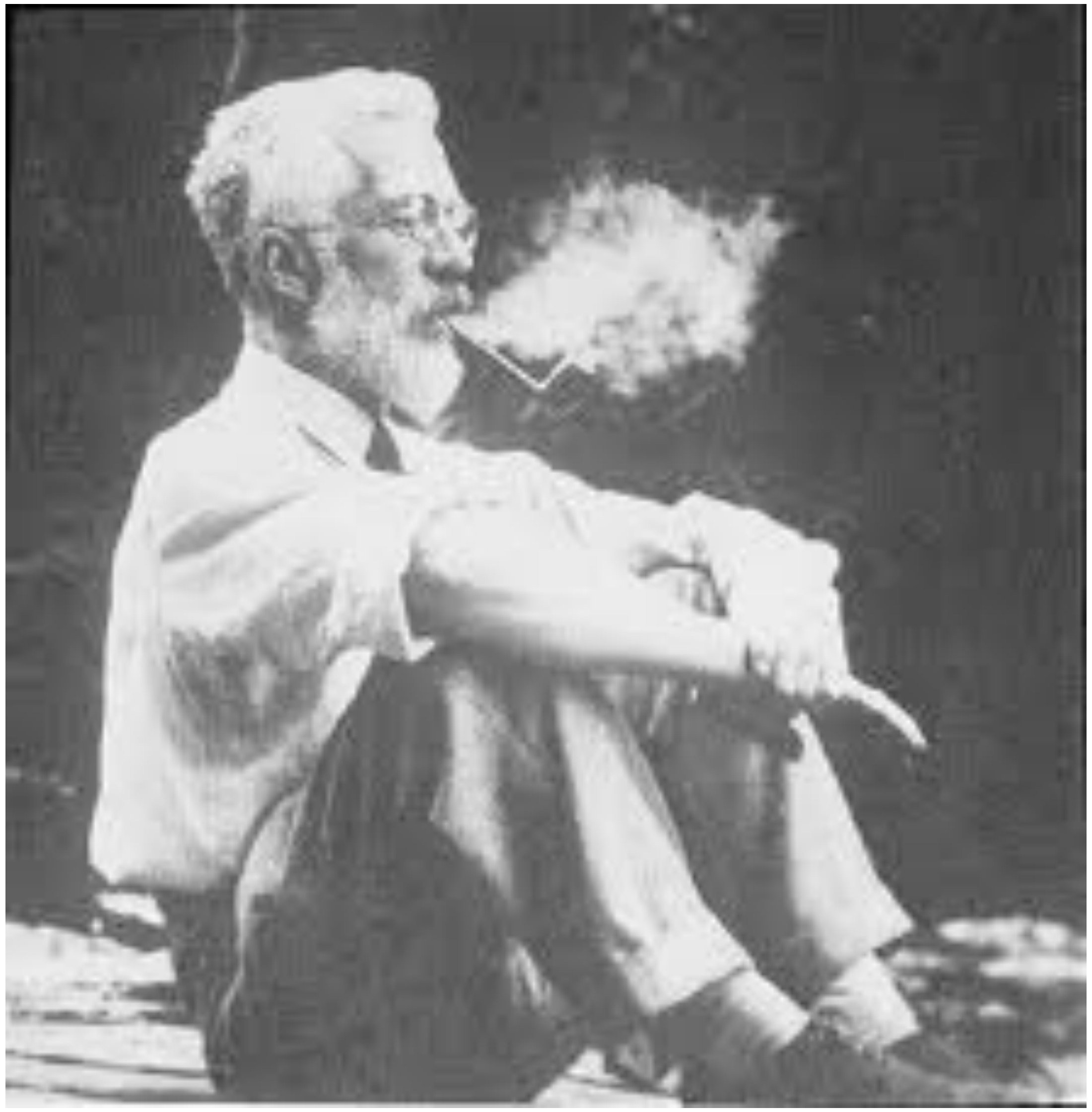
$$(1)\quad p_1X/[(p_2+X(p_1-p_2)].$$

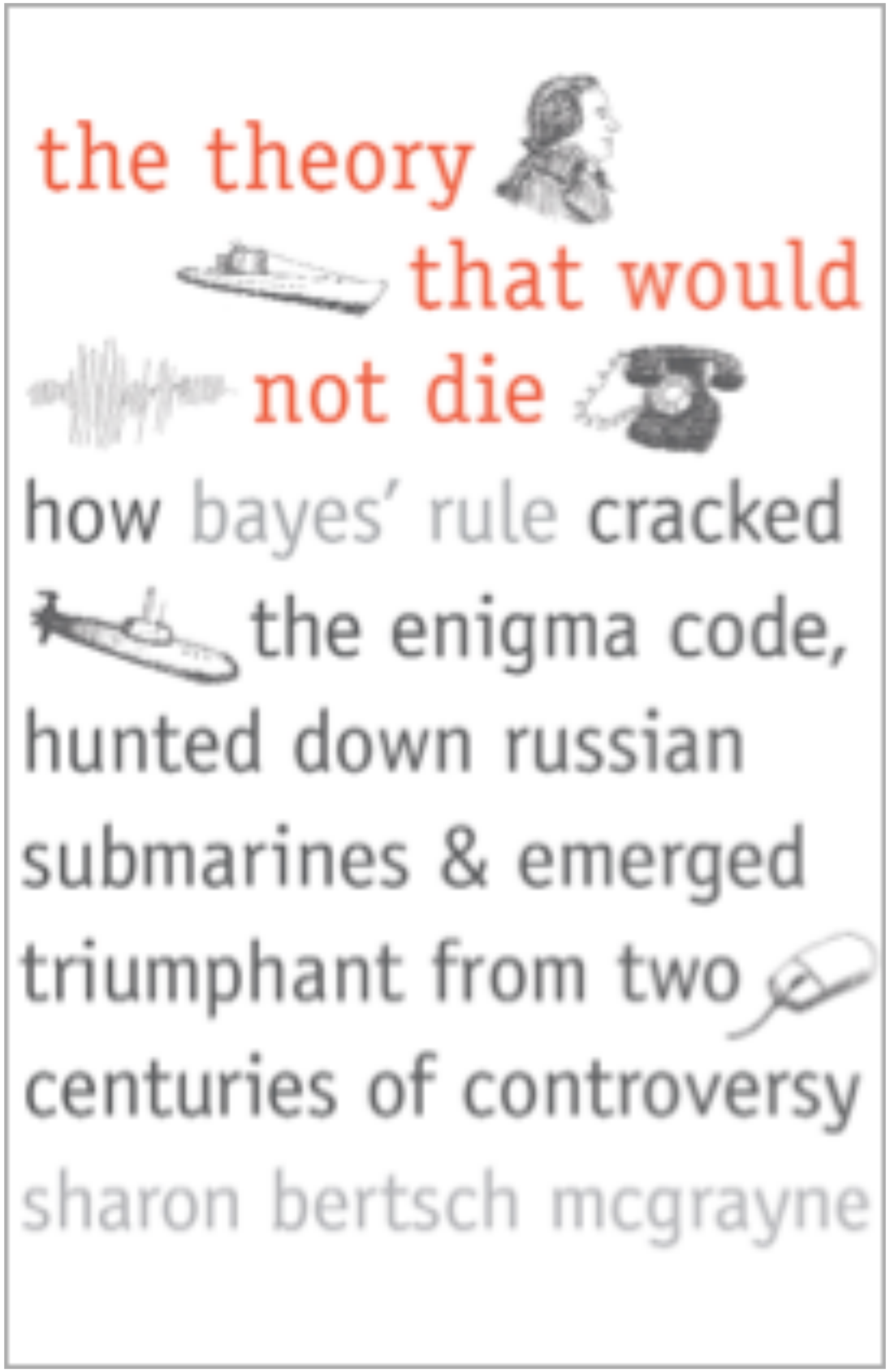
Similarly, the proportion of nonsmokers having cancer of the lung is

$$(2)\quad (1-p_1)X/[(1-p_2)-X(p_1-p_2)].$$

Formulas (1) and (2) yield the true rates we seek.

Given the appropriate data, formulas (1) and (2) are easy to compute. They are somewhat cumbersome algebraically, however. The following approximation to the true rates, therefore, seems useful. If the proportion of the general population having cancer of the lung, X, is small relative to both the proportion of the control group smoking and not smoking, p_2 and $1-p_2$, the contribution of the term $X(p_1-p_2)$ to the denominator of formulas (1) and (2) is trivial and may be neglected. In that case the approximate rate of cancer of the lung among smokers becomes $\frac{p_1X}{p_2}$ and the corresponding rate for nonsmokers $\frac{(1-p_1)X}{1-p_2}$. Whenever p_1-p_2. is greater than zero, p_1/p_2 is greater than unity. We may conclude from the approximation, therefore, that whenever a greater proportion of the diseased than of the control group possess a characteristic, the incidence of the disease is always higher among those possessing the characteristic. This is the intuition on which the procedures used in such clinical studies





the theory that would not die

how bayes' rule cracked the enigma code, hunted down russian submarines & emerged triumphant from two centuries of controversy

sharon bertsch mcgrayne