LBL-27628
Preprint

Lawrence Berkeley Laboratory
UNIVERSITY OF CALIFORNIA

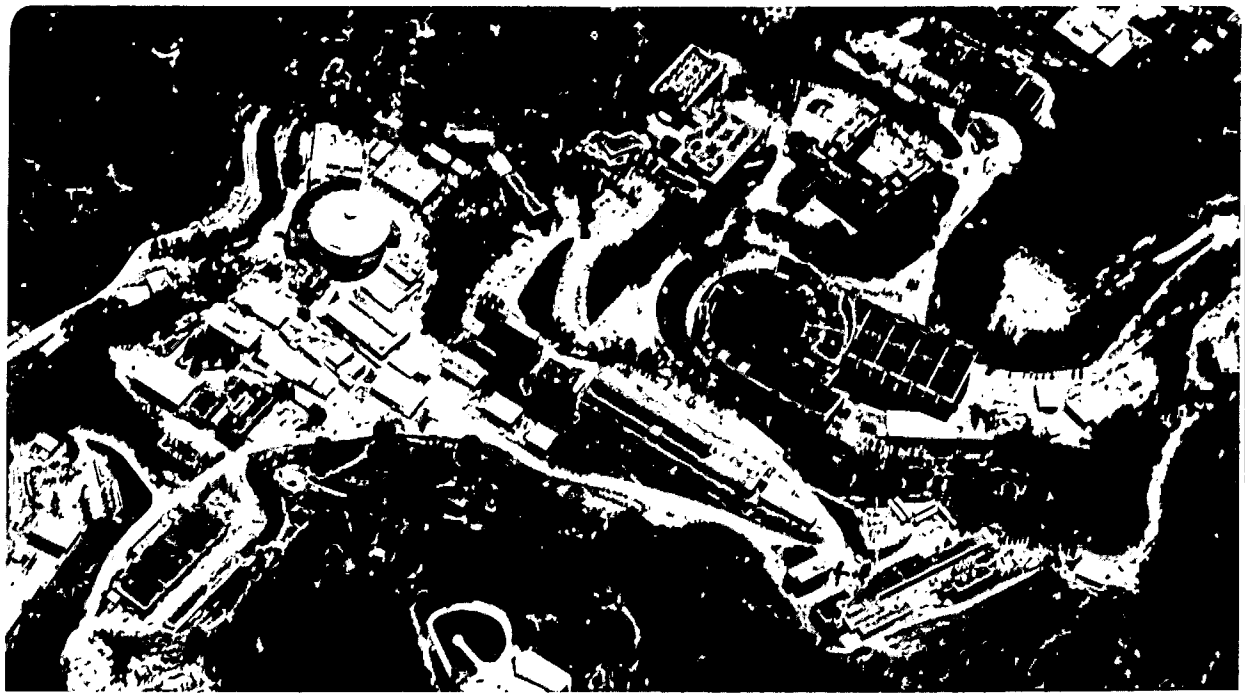
Information and Computing
Sciences Division

Submitted to Archives of Environmental Health

Detection of Excess Disease near an Exposure Point: A Case Study

J. Schulman, S. Selvin, G.M. Shaw, and D. Merrill

August 1989



Prepared for the U.S. Department of Energy under Contract Number DE-AC03-76SF00098.

DISCLAIMER

DETECTION OF EXCESS DISEASE NEAR AN EXPOSURE POINT:

A CASE STUDY

Jane Schulman, Ph.D. [1,3]

Steve Selvin, Ph.D. [2,3]

Gary M. Shaw, Dr. P.H. [1,3]

Deane Merrill, Ph.D. [3]

[1] California Birth Defects Monitoring Program, March of Dimes Birth Defects Foundation, 5900 Hollis Street, Suite A, Emeryville, CA 94608.

[2] Biostatistics, 140 Warren Hall, University of California, Berkeley CA 94720.

[3] Information and Computing Sciences Division, 1 Cyclotron Road, Lawrence Berkeley Laboratory, Berkeley, CA 94720.

Address all correspondence to J. Schulman, California Birth Defects Monitoring Program, 5900 Hollis Street, Suite A, Emeryville, CA 94608, (415) 653-3303, ext. 334.

This work was partially supported by the Director, Office of Energy Research, Office of Health and Environmental Research, Human Health and Assessment Division, of the U.S. Department of Energy under Contract No. DE-AC03-76F00098.

Detection of Excess Disease Near an Exposure Point: A Case Study

Abstract

Many studies have evaluated the likelihood of adverse health effects associated with environmental contamination from point source exposures. Two statistical measures used in these studies are the ratio R of the observed to the expected number of cases occurring in the area containing the point and the average distance $\overline{D}$ between the cases and the point. This paper estimates the probability of detecting an association between disease and exposure when one actually exists (power) for R and $\overline{D}$ for several specific and plausible statistical models. Results are presented in the context of rare diseases such as congenital malformations. The practical implications of using these measures for the evaluation of risk of disease in environmental epidemiologic investigations are discussed.

INTRODUCTION

Public concern about the possibility of adverse health effects associated with environmental contamination from nuclear power plants and chemical storage or disposal sites has led to a variety of studies to evaluate risk associated with point source exposures.[1-10] Some researchers compare the rate of disease in areas believed to have been contaminated with the rate of disease in unexposed areas.[1-5] Others compare the distribution of the distances between cases and the exposure point to the corresponding distribution in a set of controls.[6-7] Finally, the distribution of the distances between the cases and the exposure point has been evaluated on maps transformed so that the interfering influence of the resident population has been removed.[8-9]

Two statistical measures used in these studies are: the ratio, R, of the observed number of cases to the expected number of cases occurring in the area around the point and the average distance, $\overline{D}$, between the cases and the point. If more cases are observed than are "expected" or if the average distance from cases to the point is smaller than "expected", it is inferred that an association exists between the point source of exposure and the disease.

This paper: 1) explores estimates of the probability of detecting an association between disease and exposure when one actually exists (statistical power) for the ratio R and the distance $\overline{D}$; 2) applies these results to studies of rare diseases such as congenital malformations; and 3) discusses the practical implications of using these measures for the evaluation of the risk of disease in environmental

epidemiologic investigations. The goal is to further understand the issues involved in analyzing the spatial distribution of disease by investigating several specific statistical models.

METHODS

Fundamental to any statistical analysis is a null hypothesis. For spatial analysis it is typically postulated that the spatial distribution of disease is unrelated to a particular exposure point, implying that every person at risk has an equal probability of developing the disease in question. Two possible alternative hypotheses are: 1) the risk of disease is uniformly elevated in the exposed area compared to the unexposed area, and 2) the risk of disease is inversely related to the distance from the exposure point. These two risk patterns will be referred to as dichotomous and continuous, respectively.

Suppose the study area consists of an exposed area proximal to the exposure point and an unexposed area consisting of the remainder of the study area. Denote the size of the population-at-risk in the exposed area by N_e and in the unexposed area by $N_{\bar{e}}$ so that the proportion of the population exposed is $c = N_e/N$, where the total population is $N = N_e + N_{\bar{e}}$. The probability of disease in the exposed area is represented by p_e and in the unexposed area by $p_{\bar{e}}$, so that the relative risk of disease comparing the two areas is then $RR = p_e/p_{\bar{e}}$.

If one assumes a baseline frequency of disease (cases/population), the ratio R can be statistically evaluated by using either the entire study area or the exposed area only. For example, if the boundary of the contaminated area is known, the analysis is focused on the number of cases occurring in the exposed area, using only the baseline frequency of the disease $p_{\bar{e}}$ derived from a relevant unexposed population to calculate the expected number of cases. However, if the boundary of the exposed area is unknown, then it is necessary to consider everyone in the study area as exposed and statistically evaluate the number of cases occurring in the exposed and unexposed areas combined. Denote the R statistic obtained when the exposure boundary is known as R_e (e = exposed area only). Similarly, R_t (t = total area) refers to the R statistic calculated when the exposure boundary is unknown.

To derive expressions for the approximate statistical power for R_e and R_t, the common assumption is made that the observed number of cases follows a Poisson distribution with parameter λ. That is, the risk of disease is constant and small for all individuals at risk. When the risk of disease is incorrectly assumed constant and the sampled data are combined over a series of sub-areas which are heterogeneous for risk of disease, the variance for the estimated parameter is overstated.[11] However, this assumption serves as a first approximation to study the consequences of using ratio statistics to assess risk of disease.

The average distance measure $\bar{D}$ can also be used to evaluate the spatial distribution of disease in two ways, either on a map with standard geopolitical boundaries or on a specially constructed map called a Density Equalized Map Projection (DEMP).[8-9,12] Cases of disease may appear to cluster on a geopolitical map because of variable population density even when all members of the population-at-risk have the same probability of disease. A DEMP is a map which has been transformed so that area is proportional to population; i.e., population density is constant. Therefore, cases are uniformly distributed on a DEMP when the null hypothesis is true, and apparent clustering caused by unequal distribution of the population-at-risk is no longer a factor in the analysis. The average distance statistics obtained from the geopolitical and transformed maps will be represented by $\bar{D}_G$ (G = geopolitical map) and $\bar{D}_T$ (T = transformed map), respectively.

Now, consider the dichotomous alternative hypothesis which assumes that risk of disease is uniformly elevated in the exposed area. To evaluate R_t, the Poisson expectation is $\lambda = N_{\bar{e}}p_{\bar{e}} + (RR)N_e p_{\bar{e}}$ and, under the null hypothesis, $\lambda = N\,p_{\bar{e}}$. Furthermore, to evaluate R_e, $\lambda = (RR)N_e p_{\bar{e}}$ and, if the null hypothesis is true, $\lambda = N_e p_{\bar{e}}$. In evaluating R_t, the approximate power is given by

$$Power = P\left[Z > \frac{z_{1-\alpha} + c\sqrt{Np_{\bar{e}}}(1 - RR)}{\sqrt{1 - c(1 - RR)}}\right]$$

and, in evaluating R_e, the approximate power is expressed as

$$Power = P\left[Z > \frac{z_{1-\alpha} + \sqrt{cNp_{\bar{e}}}(1 - RR)}{\sqrt{RR}}\right],$$

where Z represents a standard normal random variable (mean = 0 and variance = 1) and $z_{1-\alpha}$ is the $100(1-\alpha)^{th}$ percentile of this distribution.

The evaluation of the approximate power for both $\bar{D}$ statistics for the dichotomous alternative risk pattern can also be estimated using a normal approximation where, in general,

$$Power = P\left[Z < \frac{E(\bar{D}_0) + z_{1-\alpha}\sqrt{V(\bar{D}_0)} - E(\bar{D}_1)}{\sqrt{V(\bar{D}_1)}}\right].$$

The quantities $E(\bar{D}_0)$, $E(\bar{D}_1)$, $V(\bar{D}_0)$, and $V(\bar{D}_1)$ are the expectations and variances of the particular $\bar{D}$ statistic under the null and alternative hypotheses, respectively. These expectations and variances depend on the shape of the study area under investigation.

A map of a study area can be viewed as a series of polygons (e.g., census tracts) defined by X and Y coordinate values (e.g., latitude/longitude). To obtain the means and variances of $\bar{D}_G$ and $\bar{D}_T$ under the null and alternative hypotheses, first define D^2 as the squared distance between a random point in the study area and the exposure point, denoted by (x_0, y_0). The expectation, μ, and variance, σ^2, of D^2 are functions of the moments of the coordinates X and Y and are given by

$$\mu = E(X'^2) + E(Y'^2) + x_0'^2 + y_0'^2$$

and

$$\sigma^2 = E(X'^4) + E(Y'^4) + 2E(X'^2Y'^2) - 4[x_0'E(X'^3) + y_0'E(Y'^3)] + 4x_0'^2E(X'^2) + 4y_0'^2E(Y'^2)$$
$$- [E(X'^2) + E(Y'^2)]^2 + 8x_0'y_0'E(X'Y') - 4[y_0'E(X'^2Y') + x_0'E(X'Y'^2)],$$

where $X' = X - E(X)$, $Y' = Y - E(Y)$, $x_0' = x_0 - E(X)$, and $y_0' = y_0 - E(Y)$.

The approximate expectations and variances of the $\bar{D}$ statistics can be expressed as functions of μ and σ^2 using a Taylor series approximation or

$$E(\bar{D}) \approx \sqrt{\mu}\left[1 - \frac{\sigma^2}{8\mu^2}\right]$$

and

$$V(\bar{D}) = \frac{\sigma^2}{n4\mu}[1 - \frac{\sigma^2}{16\mu^2}],$$

where n represents the number of cases of disease.

The moments $E(X'^k Y'^l)$ for the particular geographic area under study can be approximated as weighted averages of the centroids of the census tracts (or any small subareas) of the study area; $E(X'^k Y'^l) = \sum w_j (x_j - \bar{x})^k (y_j - \bar{y})^l$ where $\bar{x} = \sum w_j x_j$, $\bar{y} = \sum w_j y_j$ and (x_j, y_j) locates the centroid of the j^{th} census tract. For $\bar{D}_G$, $w_j = RR_j P_j / \sum RR_j P_j$ with P_j = the population in the j^{th} census tract and RR_j = the relative risk in the j^{th} census tract. For $\bar{D}_T$, $w_j = RR_j a_j / \sum RR_j a_j$ with a_j = the area of the j^{th} census tract.

Simulation methods were used to evaluate the power for both distance measures for the continuous alternative hypothesis. Suppose there are n cases of disease in the total study area (exposed and unexposed areas combined). The probability that any particular case falls in the j^{th} census tract was modeled by $p_j = [d_j + \varepsilon^{-1} / \sum [d_j + \varepsilon]^{-1}$ where d_j is the distance between the centroid of the j^{th} tract and the exposure point. This expression for p_j provides probabilities that fall off rapidly as the distance from the exposure point increases, i.e., the likelihood of exposure as a cause for disease decreases sharply as distance from the exposure point increases. The constant ε (in the same units as d_j) was arbitrarily chosen to accommodate the possibility that the exposure point may be located at the centroid of the census tract. Power estimates were based on 2000 simulated values of $\bar{D}$ for each sample size, n.

To study the power for R_e for the continuous alternative, the values of p_j are summed for all "exposed" areas (i.e., $p = \sum p_j$, where the summation extends over the "exposed" tracts only). The expectation λ takes on the value $\lambda_0 = nc$ when the null hypothesis is true and the value $\lambda_1 = np$ under alternative hypothesis. That is, under the null hypothesis, everyone has the same probability of disease so the expected number of cases in the "exposed" area is equal to the total number of cases in the entire study area (n) multiplied by the proportion of the population residing in the exposed area (c). On the

other hand, when the alternative risk pattern holds, the expected number of cases in the "exposed" area is equal to the total number of cases (n) multiplied by the probability that a case falls in any of the exposed subareas (p). The approximate power is calculated as:

$$Power = P\left[Z > \frac{\lambda_0 - z_{1-\alpha}\sqrt{\lambda_0} - \lambda_1}{\sqrt{\lambda_1}}\right].$$

We chose Santa Clara county, California (Figure 1) to explore these power estimates because it has previously been studied for environmental exposures and disease. Three exposure points, denoted by darkened circles in Figure 1, were identified to illustrate how power depends on the location of the exposure point. Adverse reproductive outcomes associated with point 1 have been studied as a result of a leak of organic solvents into the ground water.[13] Points 2 and 3 were chosen arbitrarily for the purposes of illustration and have not been previously associated with environmental contamination. The three corresponding "exposed" areas, denoted by shading in the figure, were chosen so that the proportion of the population "exposed" did not vary by location of the exposure point. The population-at-risk was defined as the 13,273 white male infants born alive to residents of the county in 1980. Of these 13,273 births, 6.2% were residents in the hypothetically exposed area; i.e., N = 13,273, and c = 0.062. Finally, the baseline prevalence of disease was assumed to be 1 per 1000 livebirths, corresponding to the prevalence of a frequent congenital malformation.

RESULTS

The approximate power to detect elevated risk described by the dichotomous alternative hypothesis is shown in Table 1 for statistical measures R_e, R_t, $\overline{D}_G$, and $\overline{D}_T$ for three exposure points and for four values of the relative risk (RR = 1, 2, 5, and 10). Consistently higher power is obtained using R_e; i.e., knowledge of the exposure boundary increases the power. Conversely, if the exposure boundary is unknown, the chances of detecting an increased risk in the exposed area are small; e.g., the probability of detecting a relative risk of 5 using R_e is 0.82 as compared to 0.27 for R_t.

Both $\bar{D}$ statistics have low power to detect a dichotomous risk pattern (Table 1). Further, although the power of $\bar{D}_T$ is comparable to the power of $\bar{D}_G$ to detect this alternative, the power for $\bar{D}_T$ is somewhat greater than that of $\bar{D}_G$ for areas 1 and 2 and is somewhat less than the power of $\bar{D}_G$ for area 3. In addition, the power of $\bar{D}_G$ varies more by location of the exposure point than does the power of $\bar{D}_T$; e.g., the power of $\bar{D}_G$ to detect a relative risk of 5 varies from 0.20 to 0.54 depending on the location of the exposure point but the power of $\bar{D}_T$ is approximately 0.35 for all three exposure points.

Figures 2 and 3 illustrate the relation between the power to detect the dichotomous alternative and the proportion of the population exposed (c) for R_e, R_l, and $\bar{D}_T$. The power for R_e is higher than the power for R_l for values of $c < 0.5$ but is similar when $c \geq 0.5$; i.e., when the exposed population represents 50% or more of the total population (Figure 2). In addition, even when the boundary of the exposure area is known, the power available to detect small increases in risk is slight when c is less than 0.10. For example, the probability of detecting a relative risk of 2 is approximately 0.30 when 10% of the population is exposed, assuming N = 10,000 and $p_{\bar{e}} = 0.001$. When $\bar{D}_T$ is used to assess risk and the alternative risk pattern is dichotomous, power first increases and then decreases as the proportion of the population exposed increases (Figure 3). The comparison of Figures 2 and 3 shows that R_e has a larger probability of detecting small increases in risk than does $\bar{D}_T$ for all values of c for the dichotomous risk pattern.

The power of R_e, $\bar{D}_G$, and $\bar{D}_T$ to detect the continuous alternative is presented in Table 2 for five sample sizes (n = 5, 10, 20, 30 and 50). This alternative risk pattern corresponds to a relative risk ($RR = \lambda_1/\lambda_0$) of 1.2, 2.5, and 2.0 in the three "exposed" areas, respectively. Similar to the results for the dichotomous alternative, the power to detect the continuous alternative for $\bar{D}_T$ is higher than the power for $\bar{D}_G$ for areas 1 and 2 but not for area 3. However, in contrast to results seen previously, the power for $\bar{D}_T$ is dramatically greater than the power for $\bar{D}_G$ for areas 1 and 2. Also in contrast to the dichotomous alternative, the power for $\bar{D}_T$ is generally greater than the power of R_e for each of the three areas and for the range of sample sizes examined.

Finally, except for area 2, the power to detect this continuous alternative is not very high for any statistic examined. For example, for areas 1 and 3, the probability of rejecting the null hypothesis in favor of this continuous risk pattern ranges from 0.10 to 0.40 when there are 20 cases of disease.

DISCUSSION

The power of two statistical measures to detect an increased risk of disease surrounding an exposure point for two alternative patterns is contrasted. Of course, the choice of which alternative pattern to model should depend on the properties of the alleged exposure. For example, a dichotomous risk pattern might be suitable for studying waterborne contaminants and a continuous risk pattern might be an adequate model for studying airborne contaminants because of the manner in which each of these potential exposures occurs in the environment. A specific alternative hypothesis may only approximate the true exposure pattern in any particular situation and certainly will not apply in all studies; however, plausible statistical models must be constructed to evaluate the performance of a statistical measure.

The power of both $\bar{D}$ statistics to detect the dichotomous alternative is lower than the power of R_e. This occurs primarily because information regarding actual distance to the exposure point is not relevant. Therefore, if risk is adequately described in a dichotomous manner, then it is preferable to use R_e rather than $\bar{D}_G$ or $\bar{D}_T$. In addition, R_e is traditional, simple to calculate, and easily interpreted. However, when risk truly decreases as a function of distance from the exposure point, it is preferable in terms of power to use $\bar{D}_T$ to evaluate the risk rather than the R_e statistic. The contrast of R_e and $\bar{D}_T$ illustrates the general statistical principle that the efficacy of a statistical summary depends on the question being investigated; i.e., the alternative hypothesis.

The power of both $\bar{D}$ statistics to detect a uniformly increased risk in the area surrounding the exposure point appears to increase as the exposure point moves away from the population centroid of the study area. On the other hand the power for both R statistics is unaffected by the location of the exposure point. Regardless of the alternative risk pattern examined, the power of $\bar{D}_T$ is higher than the power of $\bar{D}_G$ when the exposure point is located near the centroid of the study area. However, when

the exposure point is located in a remote portion of the study area the power of $\bar{D}_G$ surpasses the power of $\bar{D}_T$. Also, if $\bar{D}_T$ is used to evaluate a dichotomous risk model in Santa Clara county and approximately 20-30% of the population is exposed, then the power of $\bar{D}_T$ is maximized. In practice, however, it seems unlikely that the exposure point will be located far from the population centroid and that 20-30% of the population will be exposed, making the power to detect an increased risk of this type using $\bar{D}_T$ low. Nevertheless, it is important to note that the likelihood of detecting excess disease depends on the location of the exposure point and on the ratio of exposed to non-exposed individuals.

Our results illustrate the importance of having information about the exposure boundary in these investigations. If this information is not available and R_t is used to evaluate risk, then the chances of detecting increased risk of disease relative to a fixed exposure point are dramatically reduced, especially if only a small percentage of the population is exposed. If 30% or more of the population is exposed, the necessity for having this information becomes less important in terms of power. However, it is unlikely that environmental exposures will affect such a large proportion of the population in most epidemiologic investigations .

In summary, this investigation demonstrates that the likelihood of being able to detect excess disease associated with a point source of exposure using the ratio or distance statistics depends on the 1) particular alternative risk pattern being investigated, 2) location of the exposure point relative to the population distribution in the study area, 3) amount of information available concerning the alleged exposure, 4) proportion of the population which is exposed, and 5) magnitude of the disease risk associated with the exposure. Although not addressed in this case study, power also depends on the baseline frequency of the disease in question and the size of the population-at-risk; these issues were not addressed because we were specifically concerned with the study of rare diseases such as congenital malformations in relatively large populations.

REFERENCES

1. Johnson, C.J. 1981. Cancer incidence in an area contaminated with radionuclides near a nuclear installation. *Ambio 10:*176-182.

2. Kaldor, J.; Harris, J.A.; Glazer, E.; et al. 1984. Statistical association between cancer incidence and major-cause mortality, and estimated residential exposure to air emissions from petroleum and chemical plants. *Env Health Persp 54:*319-332.

3. Matanoski, G.M.; Landau, E.; Tonascia, J; et al. 1981. Cancer mortality in an industrial area of Baltimore. *Environmental Research 25:*8-28.

4. Ozonoff, D.; Colten, M.E.; Cupples, A.; et al. 1987. Health problems reported by residents of a neighborhood contaminated by a hazardous waste facility. *Am J Ind Med 11:*581-597.

5. Baker, D.B.; Greenland, S.; Mendlein, J. and Harmon, P. 1988. A health study of two communities near the Stringfellow waste disposal site. *Arch Env Health 43:*325-334.

6. Lyon, J.L.; Klauber, M.R.; Graff, W. and Chiu, G. 1981. Cancer clustering around point sources of pollution: assessment by a case-control methodology. *Environmental Research 25:*29-34.

7. Stinnett, S.S.; Buffler, P.A. and Eifler, C.W. 1981. A case-control method for assessing environmental risks from multiple industrial point sources. *Environmental Research 25:*62-74.

8. Schulman, J.; Selvin, S. and Merrill, D.W. 1988. Density equalized map projections: a method for analyzing clustering around a fixed point. *Stat Med 7:*491-505.

9. Selvin, S.; Shaw, G.M.; Schulman, J. and Merrill, D.W. 1987. Spatial distribution of disease: three case studies. *JNCI 79:*417-423.

10. Harjulehto, T.; Aro, T.; Rita, H.; Rytomaa, T.; and Saxen, L. 1989. The accident at Chernobyl and outcome of pregnancy in Finland. *BMJ 298:*995-997.

11. Stuart A., Ord J.K. *Kendall's Advanced Theory of Statistics*. New York: Oxford University Press, 1987;1163-164 (*5th* ed., Vol. 1).

12. Selvin, S.; Merrill, D.W.; Schulman, J.; Sacks, S.; Bedell, L. and Wong, L. 1988. Transformations of maps to investigate clusters of disease. *Soc Sci Med 26:*215-221.

13. Swan, S.H.; Shaw, G.M.; Harris, J.; and Neutra, R. 1989. Congenital cardiac anomalies in relation to water contamination, Santa Clara County, California, 1981-1983. *AJE 129:*885-893.

Table 1

Approximate power of R and $\bar{D}^*$: Dichotomous Alternative Hypothesis

			area 1		area 2		area 3	
RR	R_e	R_i	$\bar{D}_G$	$\bar{D}_T$	$\bar{D}_G$	$\bar{D}_T$	$\bar{D}_G$	$\bar{D}_T$
1	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
2	0.31	0.09	0.10	0.10	0.08	0.10	0.15	0.11
5	0.82	0.27	0.28	0.36	0.20	0.31	0.54	0.38
10	0.98	0.65	0.63	0.73	0.43	0.65	0.88	0.76

*$p_{\bar{i}}$ = 0.001; c = 0.062; N = 13273

Table 2

Approximate power of R and $\bar{D}$: Continuous Alternative

	n	R_e	$\bar{D}_G$	$\bar{D}_T$
Area 1: (RR = 1.2)				
	5	0.08	0.05	0.14
	10	0.09	0.06	0.20
	20	0.10	0.10	0.26
	30	0.11	0.12	0.30
	50	0.12	0.15	0.42
Area 2: (RR = 2.5)				
	5	0.30	0.00	0.59
	10	0.38	0.02	0.85
	20	0.49	0.15	0.99
	30	0.59	0.32	1.00
	50	0.72	0.65	1.00
Area 3: (RR = 2.0)				
	5	0.22	0.18	0.19
	10	0.27	0.29	0.25
	20	0.35	0.40	0.36
	30	0.42	0.49	0.44
	50	0.53	0.64	0.54

Figure 1

Santa Clara County, California

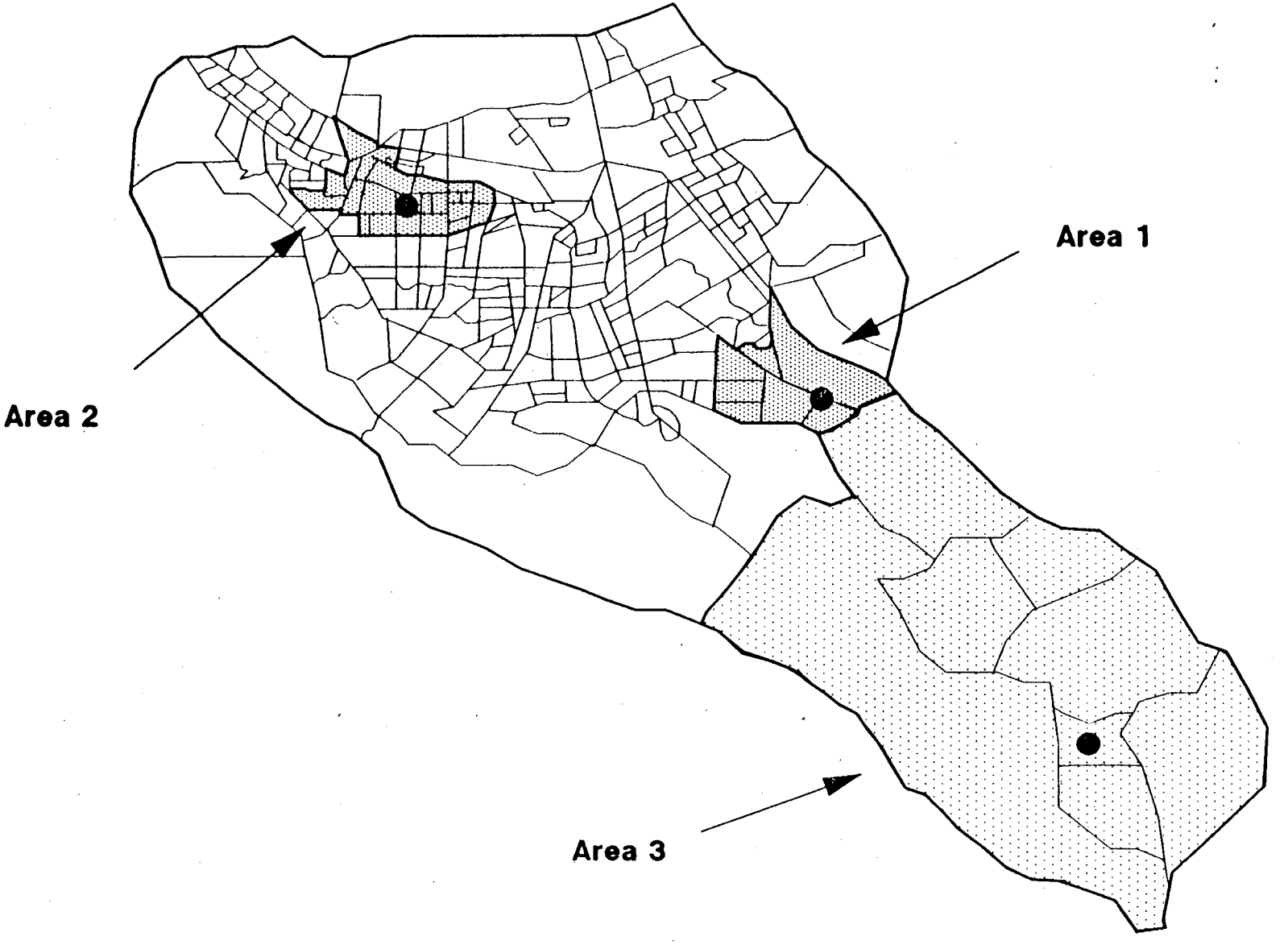
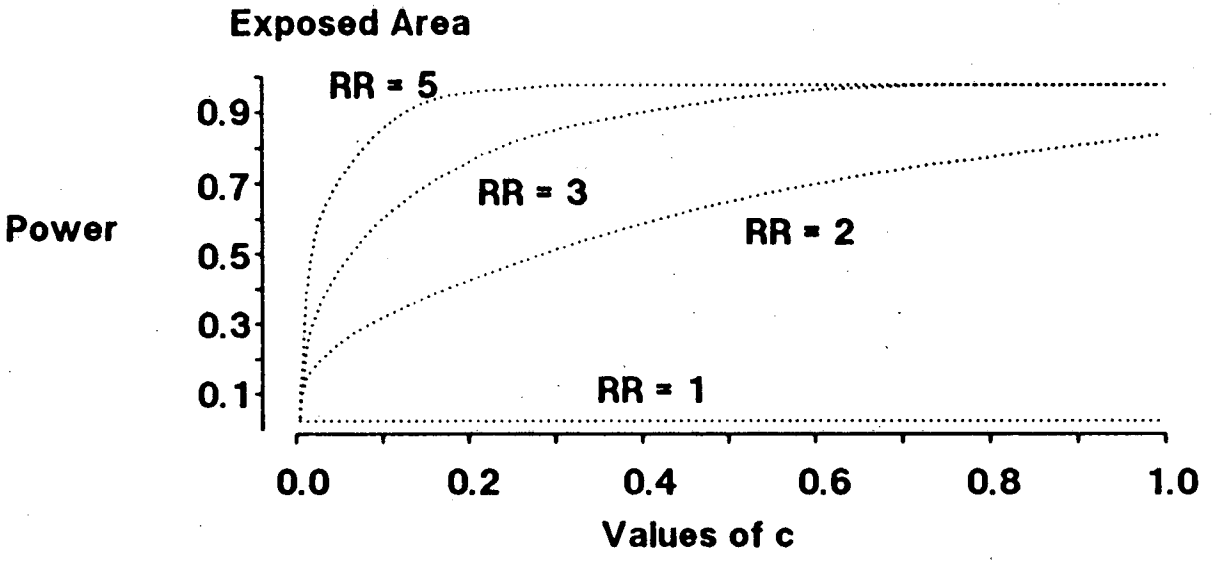


Figure 2
Power for R_e and R_t by c and RR
Dichotomous Alternative
$p_{\bar{e}}$ = .001; N = 10000



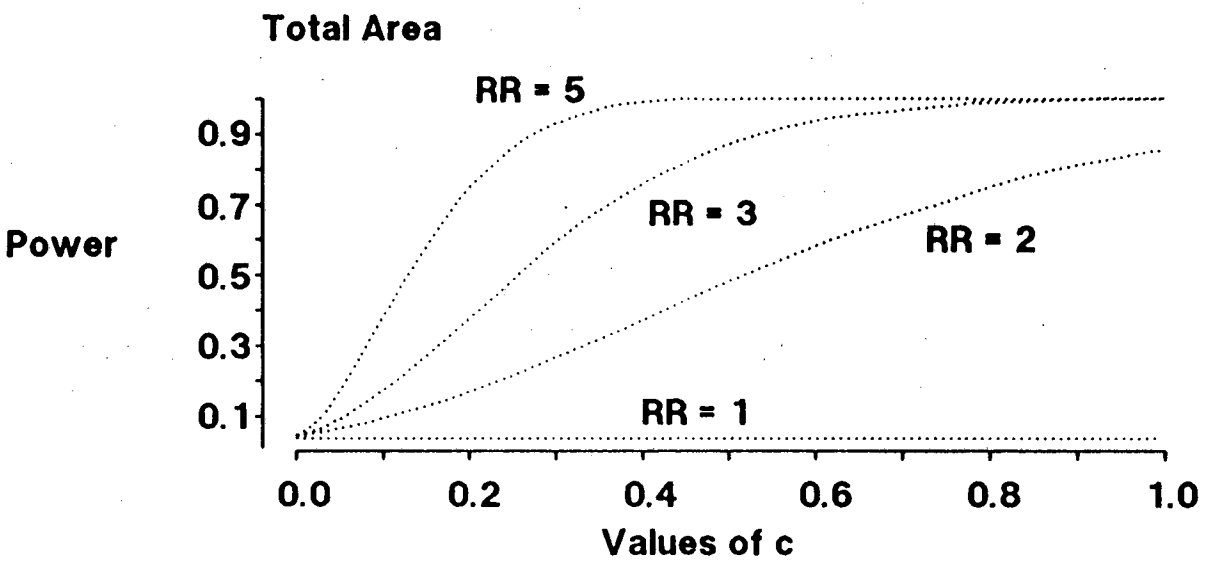
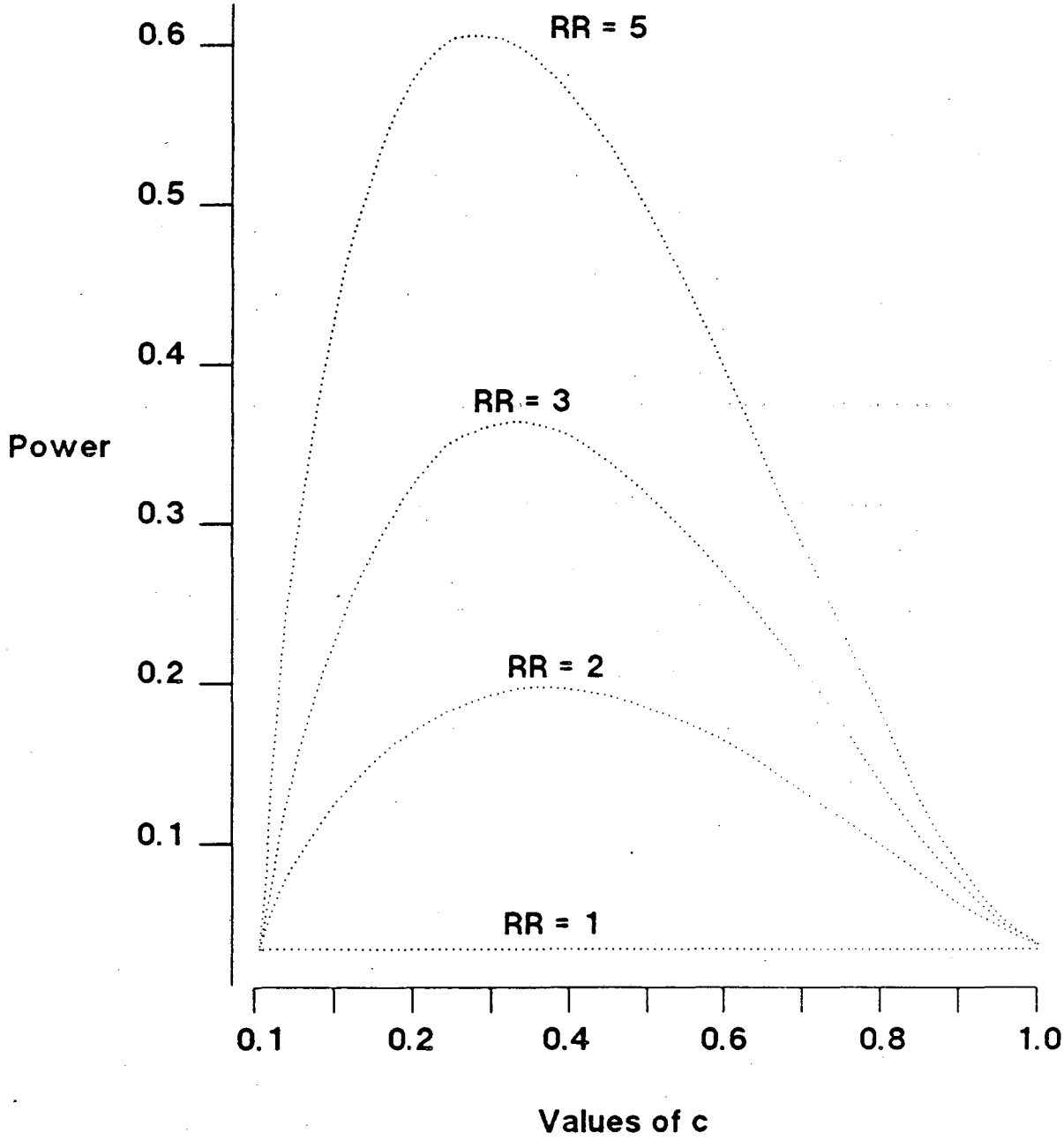


Figure 3
Power for $\bar{D}_T$ by c and RR
Exposure Point at Centroid of County
Dichotomous Alternative
Sample Size = 10



LAWRENCE BERKELEY LABORATORY
TECHNICAL INFORMATION DEPARTMENT
1 CYCLOTRON ROAD
BERKELEY, CALIFORNIA 94720