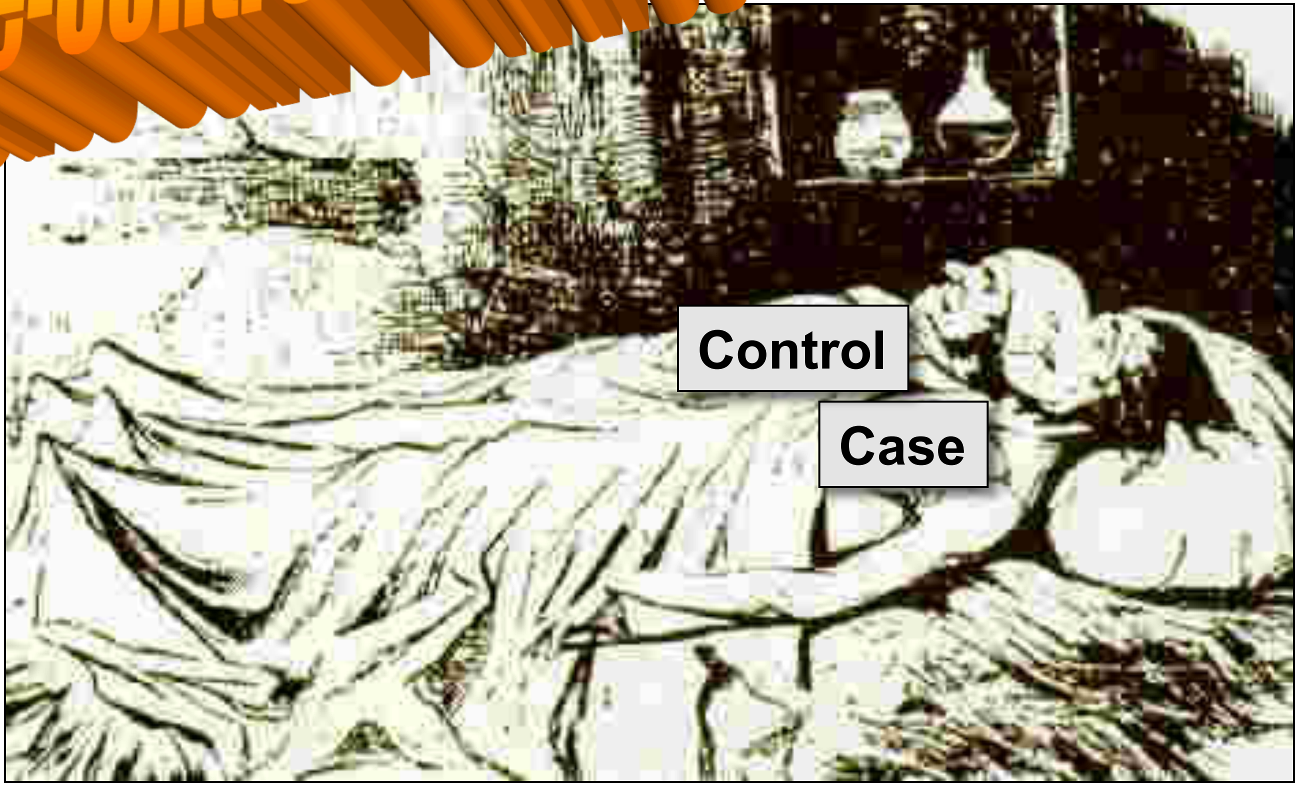
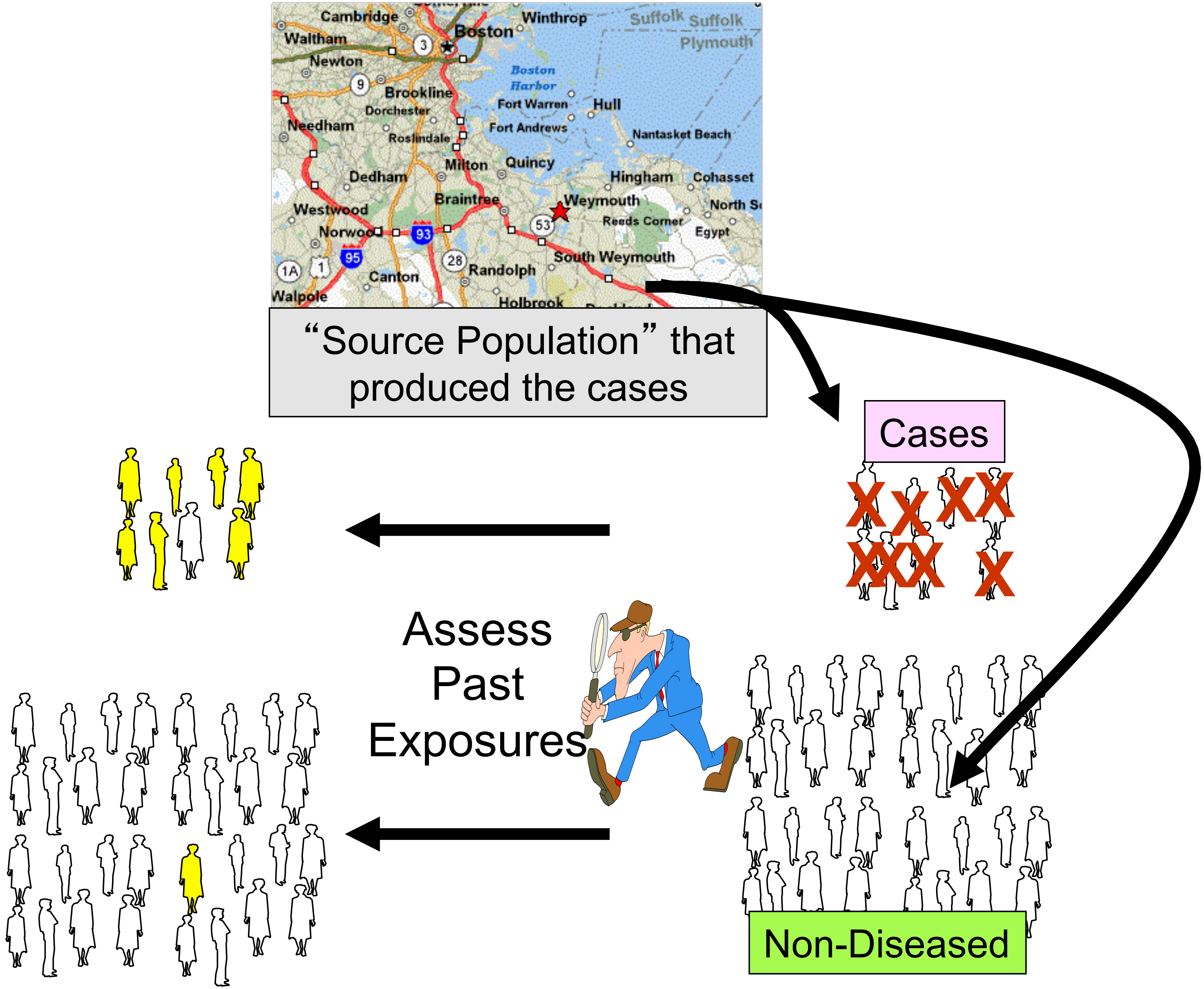
DOYLE

Case-Control Studies

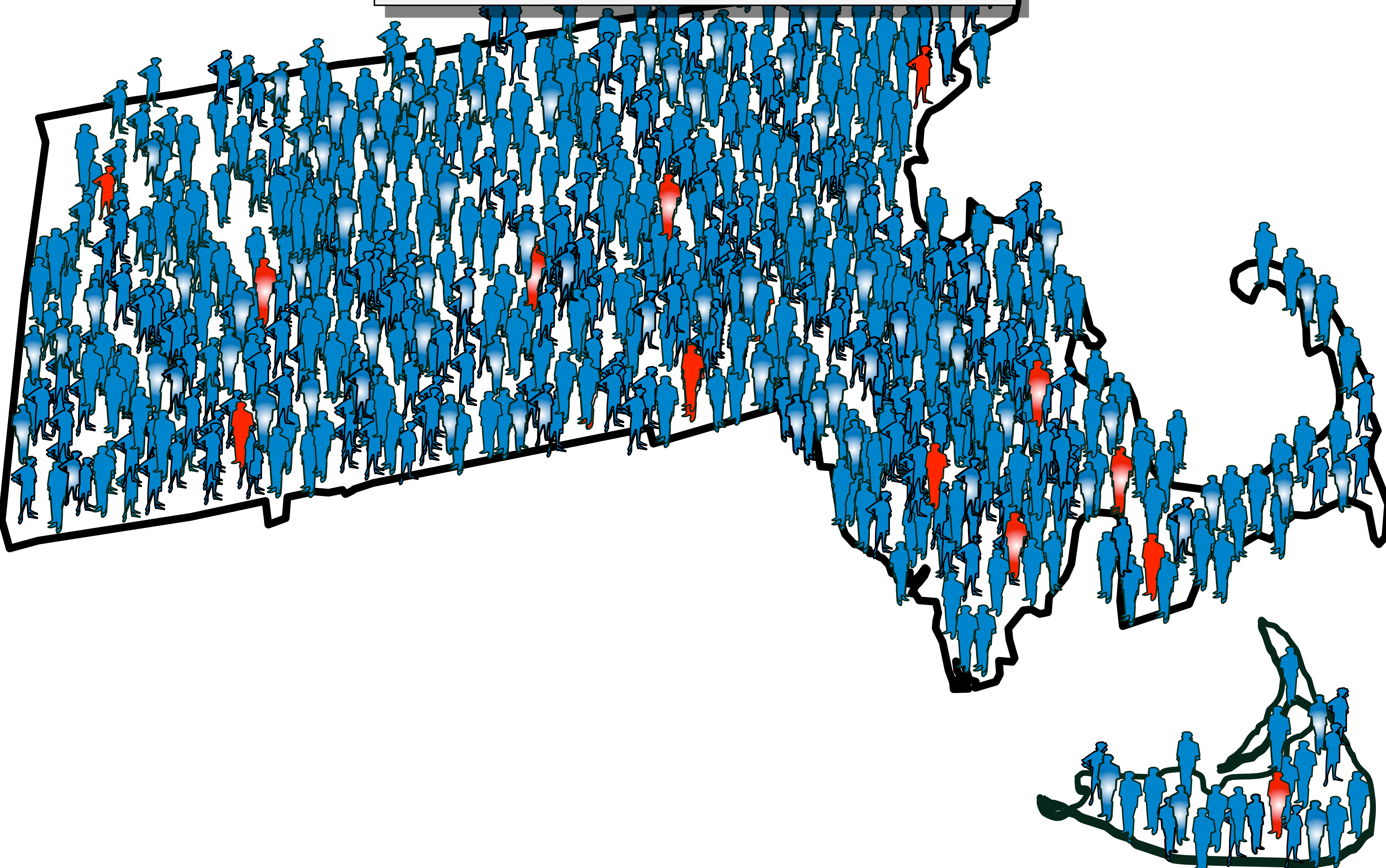


In the 1700s patients in charity hospitals sometimes slept two or more to a bed, regardless of diagnosis.

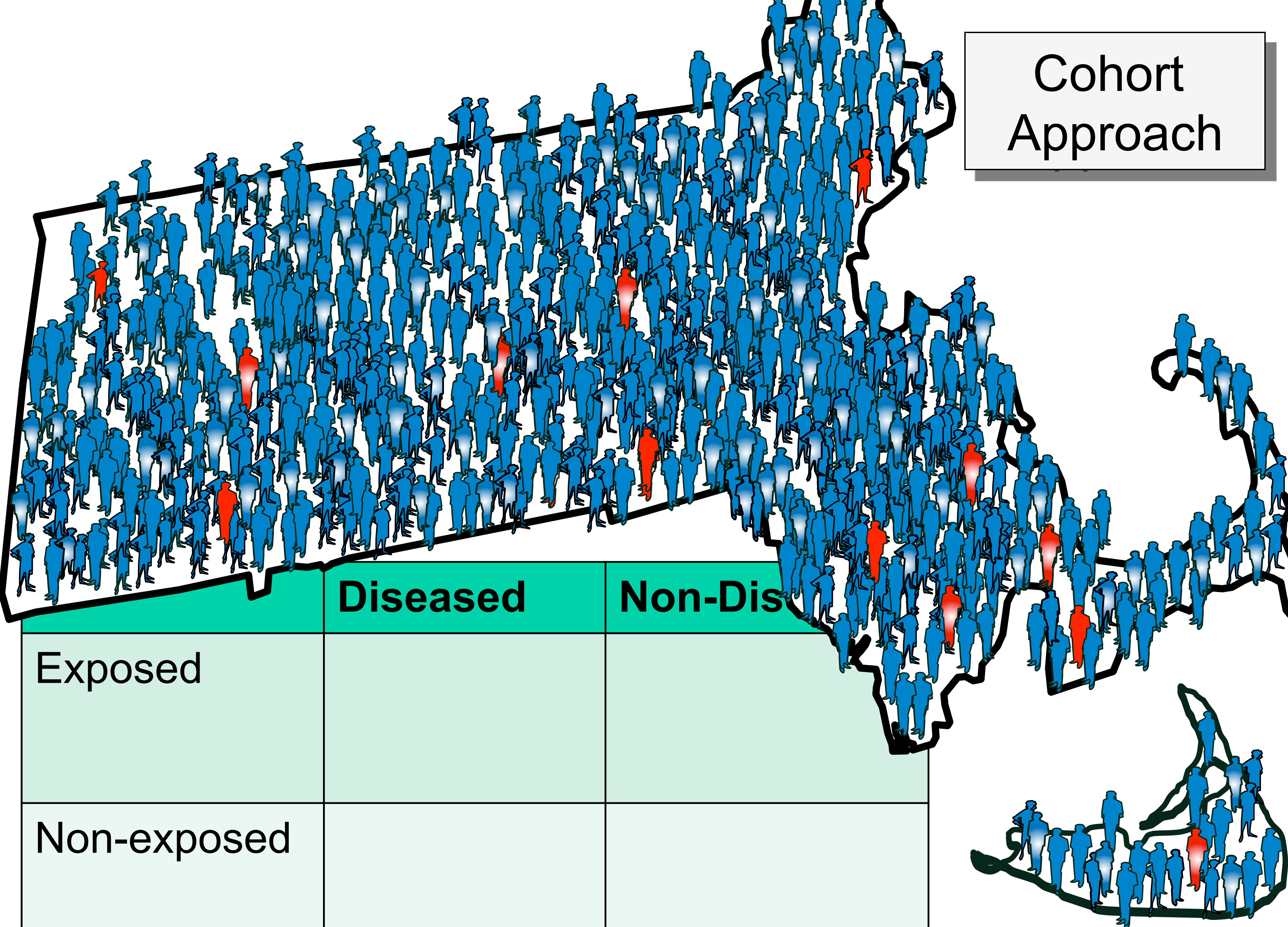
This depicts a patient who finds himself lying with a corpse (definitely a "case").


Cambridge
Waltham
Newton
Boston
Winthrop
Suffolk
Suffolk
Plymouth
Boston Harbor
Brookline
Dorchester
Fort Warren
Hull
Needham
Fort Andrews
Roslindale
Nantasket Beach
Milton
Quincy
Dedham
Hingham
Cohasset
Braintree
Weymouth
Westwood
Reeds Corner
Egypt
Norwood
South Weymouth
Canton
Randolph
Walpole
Holbrook
"Source Population" that produced the cases
Cases
Assess Past Exposures
Non-Diseased

The Problem of Studying a Rare Disease

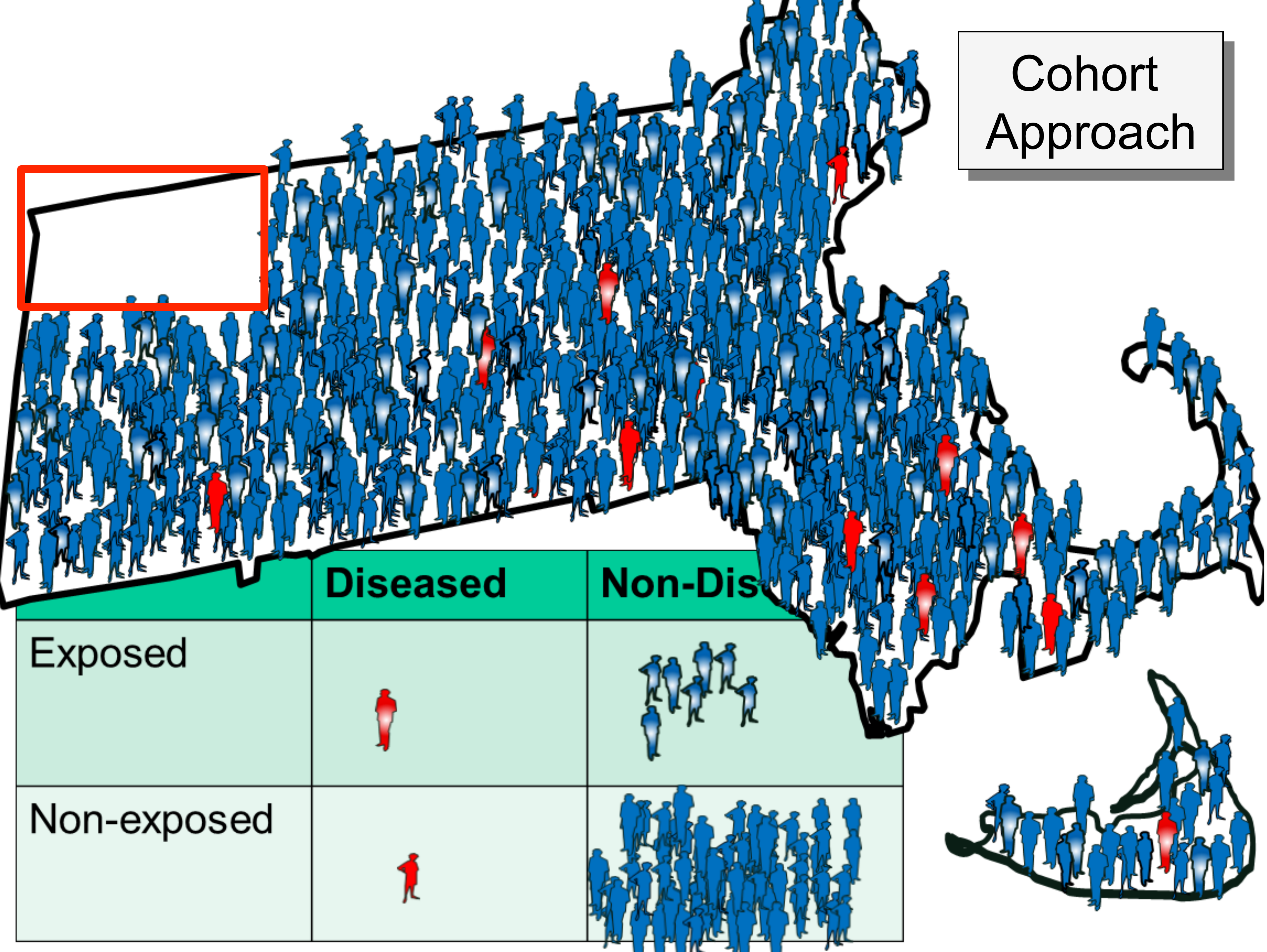


Cohort Approach



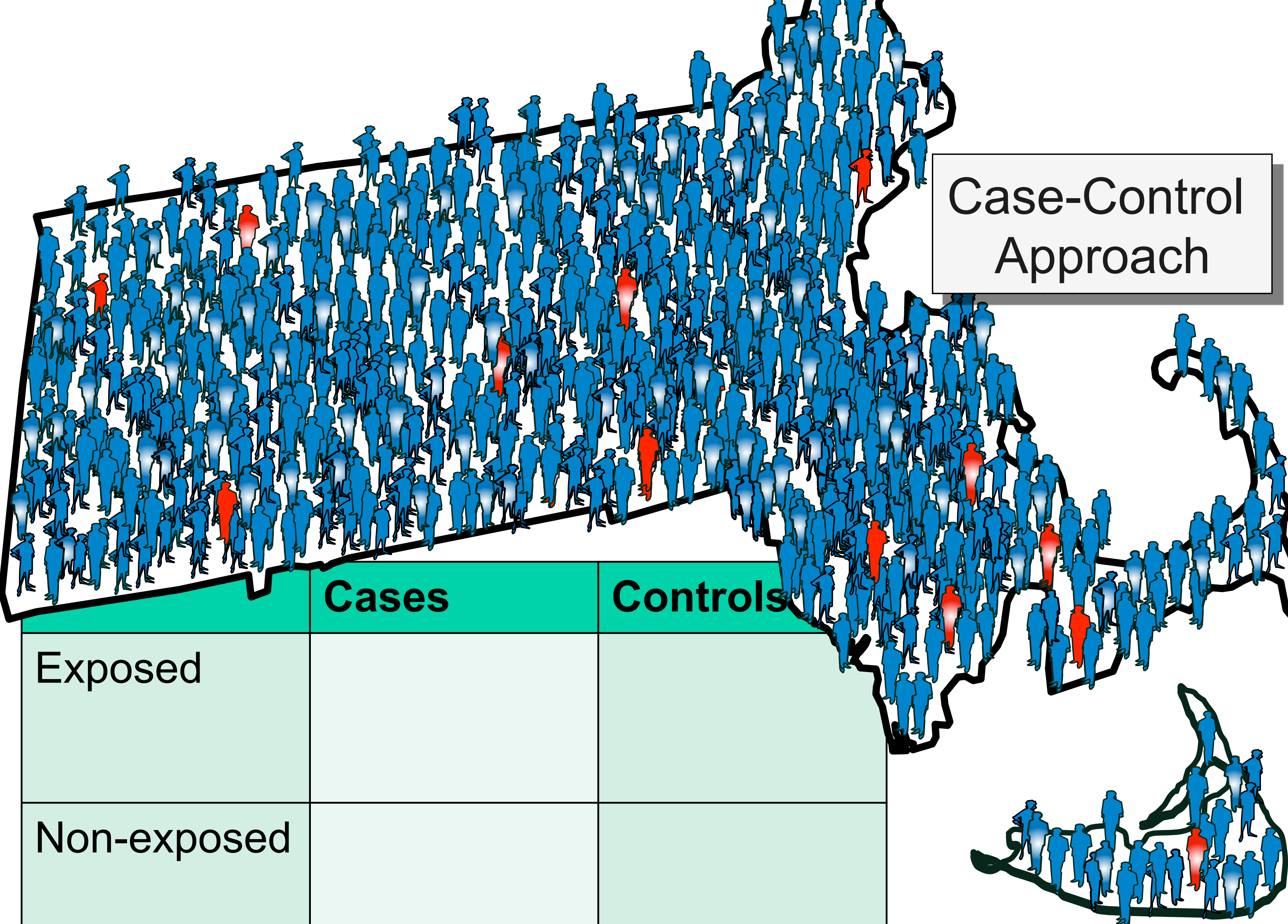
	Diseased	Non-Dis
Exposed		
Non-exposed		

Cohort Approach



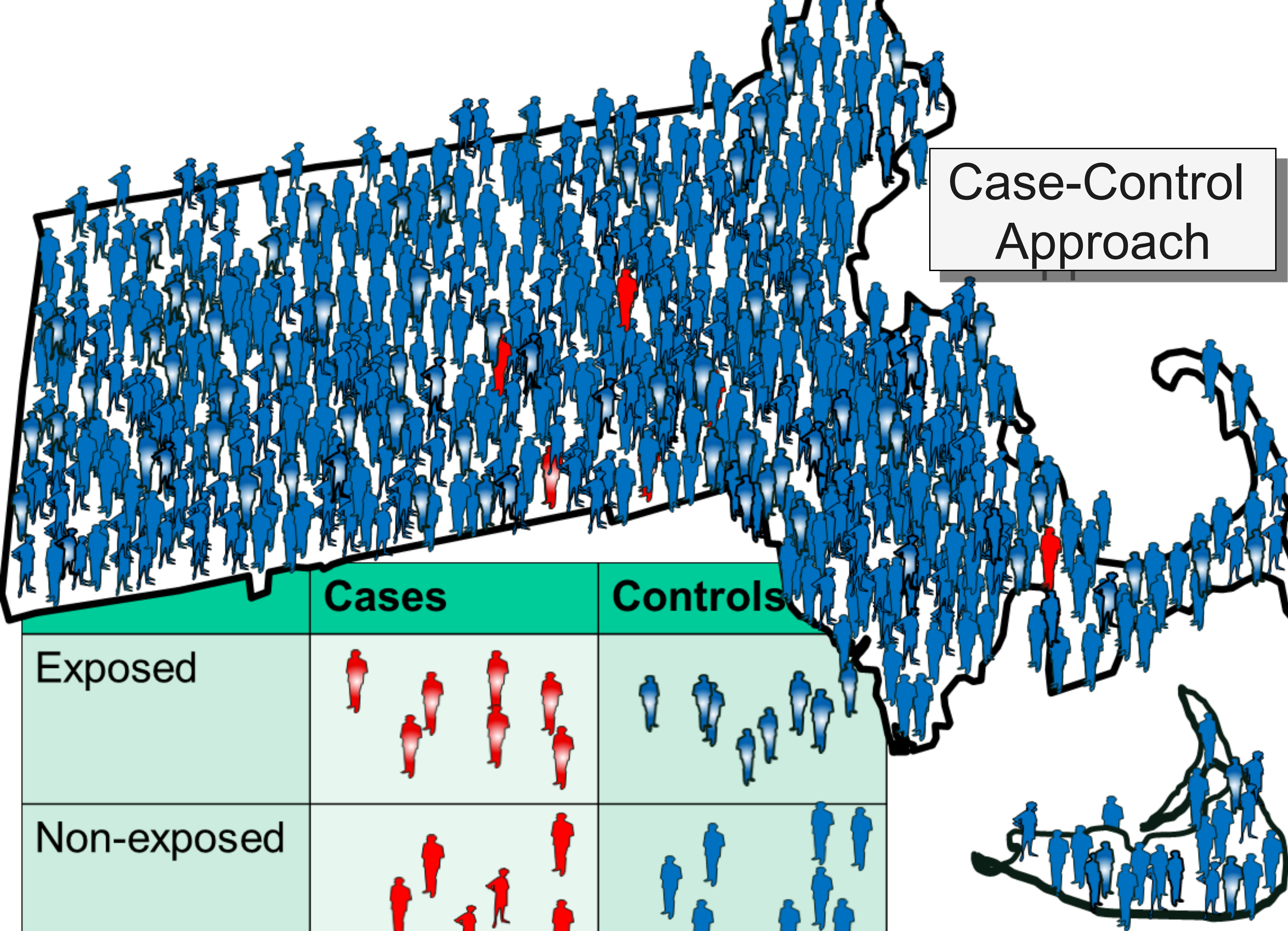
	Diseased	Non-Dis
Exposed		
Non-exposed		

Case-Control Approach



	Cases	Controls
Exposed		
Non-exposed		

Case-Control Approach



	Cases	Controls
Exposed		
Non-exposed		

Importance of a Comparison Group (Controls)

In order to quantify the association it is essential to know the exposure distribution in the population that produced the cases.

1940s: Dr. Alton Ochsner

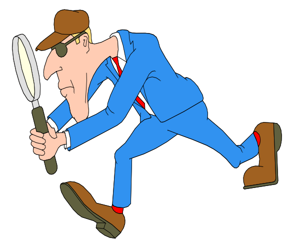
… noted that most of the patients he operated on for lung cancer were smokers. The link between smoking and lung cancer was not known, but Dr. Ochsner hypothesized that cigarettes were the cause.

What if 70% of his lung cancer patients were smokers?

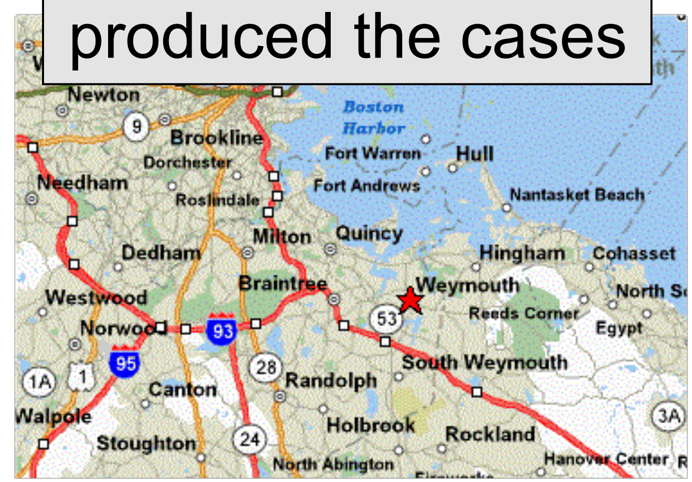
Would he be justified in concluding that smoking was the cause?

Lung Cancer Cases

70% Smoked



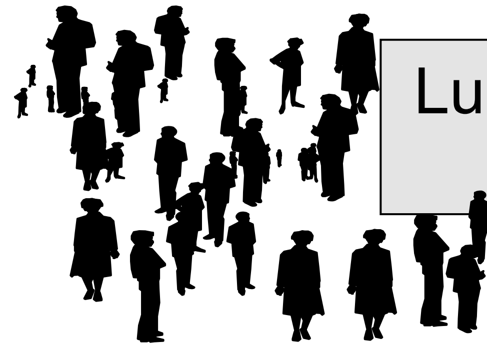
Population that produced the cases



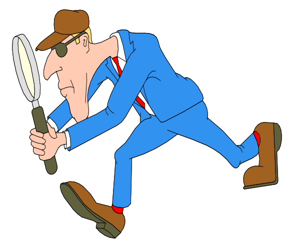
70% Smoked

But… what if….?

90% Smoked

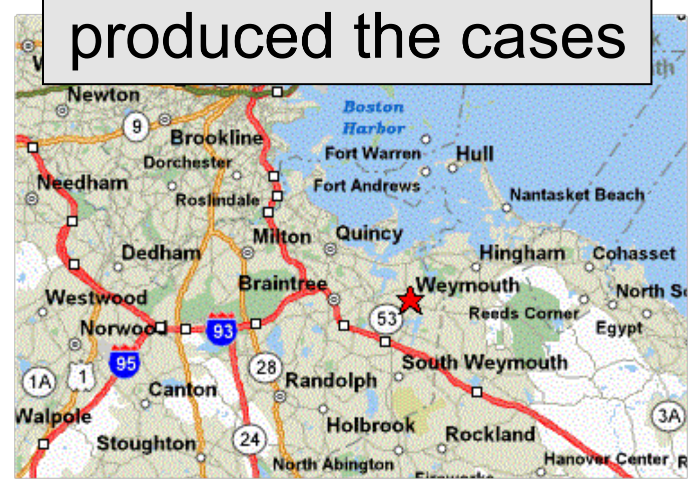


Lung Cancer Cases



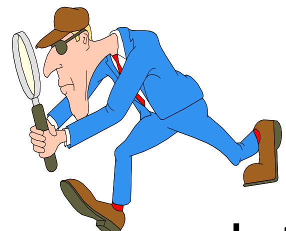
20% Smoked

Population that produced the cases



Lung Cancer Cases

70% Smoked



In a case series you can determine the proportion of the cases that had a particular exposure, …

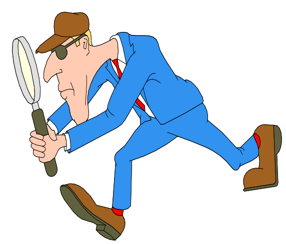
?

… but whether there is an association isn’t clear until you know what proportion of the population that produced the cases has that same exposure.

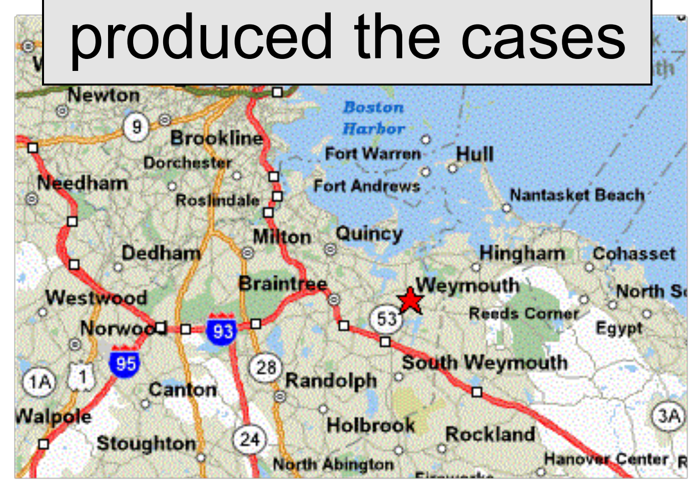
This approach is very <u>efficient</u> if you are investigating a rare disease, because you begin by collecting enough "cases".

Patients with Oropharyngeal Cancer

Odds of regular tooth brushing
90/10



Population that produced the cases



Odds of regular tooth brushing
194/6

Odds Ratio = 5.4

Role of "Controls"

Provide the distribution of exposure in the source population from which the cases arose, thus indicating the relative size of exposed and unexposed denominators for measures of disease association

Hypothetical Study (“The Teacher’s Health Study”)

- 89,949 women aged 34-59; 1,439 breast cancer cases were identified over 8 years of follow-up.

- Blood was drawn on all 89,949 at beginning of follow-up and frozen.

- After the study was begun investigators wanted to examine the association between exposure to DDT and risk of breast cancer.

New Hypothesis to Test: Pesticide exposure increases the risk of breast cancer.

Null hypothesis?

If I had analyzed *all* samples…

…this is what I would have found:

		Breast Cancer: Yes	Breast Cancer: No	
DDT Exposure	Yes	360	13,276	13,636
	No	1,079	75,234	76,313
		1,439	88,510	89,949

(Hypothetical): RR = (360/13,636) / (1,079/76,313) = 1.87

(95% CI: 1.66-2.10; $p<0.00001$)

Analyzing all 89,949 blood samples @ \$20 each will cost \$1,798,980.

		Breast Cancer		
		Yes	No	
DDT Exposure	Yes			
	No			
		1,439	88,510	89,949

1,439 have breast cancer but to analyze all 89,949 blood samples @ $20 each will cost $1,798,980.

I only have $90,000.
What should I do?

The Critical Information

DDT Exposure	Breast Cancer: Yes	Breast Cancer: No	
Yes	360	13,276	13,636
No	1,079	75,234	76,313
	1,439	88,510	89,949

Another way of thinking about the critical information is that the 1st column tells us the exposure distribution in cancer patients & the denominators tell us the exposure distribution in the overall population from which the cases arose.

RR Computed from Exposure Distributions

DDT Exposure	Breast Cancer: Yes	Breast Cancer: No	
Yes	360	13,276	13,636
No	1,079	75,234	76,313
	1,439	88,510	89,949

$$RR = \frac{(360/13{,}636)}{(1{,}079/76{,}313)}$$

This can be rearranged algebraically to:

$$= \frac{360}{1{,}079} \times \frac{76{,}313}{13{,}636} = \boxed{\frac{(360/1{,}079)}{(13{,}636 / 76{,}313)}}$$

Distributions of exposure in disease cases and in source population.

DDT Exposure	Breast Cancer: Yes	Breast Cancer: No	
Yes	360	13,276	13,636
No	1,079	75,234	76,313
	1,439	88,510	89,949
Exposure Distr. Ratios:	0.3336	0.1765	0.1787

Since it is a rare disease, the exposure distribution in the non-diseased subjects is about the same as that in the population that produced the cases.

Breast Cancer

DDT Exposure	Yes	No	
Yes	360	432	13,636
No	1,079	2,446	76,313
	1,439	2,878	89,949
Exposure Distr. Ratios:	0.3336	0.1765	0.1787

If so, I can just *take a* ***sample*** *of the non-diseased subjects* in order to estimate the exposure distribution in the population that produced the cases.

A Sample of Non-diseased Subjects

DDT Exposure	Breast Cancer: Yes	Breast Cancer: No
Yes	360	432
No	1,079	2,446
	1,439	2,878

Odds of Exposure: (360/1079) vs. (432/2446)

Odds Ratio = (360/1079) / (432/2446) = 1.89

(95% CI: 1.61-2.21; p<0.00001)

Cost = $86,340

A "Nested" Case-Control Study

DDT Exposure	Breast Cancer: Yes	Breast Cancer: No	
Yes	360	13,276	13,636
No	1,079	75,234	76,313
	1,439	88,510	89,949

We started with a large cohort study.

DDT Exposure	Yes	No
Yes	360	432
No	1,079	2,446
	1,439	2,878

But we sampled from it using a case-control approach.

A case-control study '*nested*' within a cohort study.

In estimating exposure status of the population, what would happen if I were more likely to select a non-diseased person who had the exposure?

DDT Exposure	Breast Cancer: Yes	Breast Cancer: No
Yes	360	
No	1,079	
	1,439	2,878

DDT Exposure	Breast Cancer: Yes	Breast Cancer: No
Yes	360	500 (432)
No	1,079	2,378 (2,446)
	1,439	2,878

Odds Ratio = 1.59. <u>Selection bias</u> caused an underestimate of the association.

Traditional View of Case-Control

		Breast Cancer	
		Yes	No
DDT Exposure	Yes	**360**	**432**
	No	**1,079**	**2,446**
		1,439	**2,878**

$$\frac{360/1079}{432/2446}$$

Odds Ratio = 1.89

In the traditional view we are comparing the odds of exposure in people with the disease to the odds of exposure in people who do not have disease.

Modern View of a Case-Control Study

		Breast Cancer	
		Yes	No
DDT Exposure	Yes	360	432
	No	1,079	2,446
		1,439	2,878

$$\frac{360/432}{1079/2446} = 1.89$$

These are equivalent.

$$\frac{360/1079}{432/2446} = 1.89$$

Traditional View

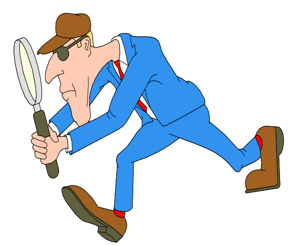
This is equivalent to asking: how do the *odds of disease in exposed subjects* compare to the *odds of disease in unexposed subjects*?

Summary: Definition of Case-Control Study

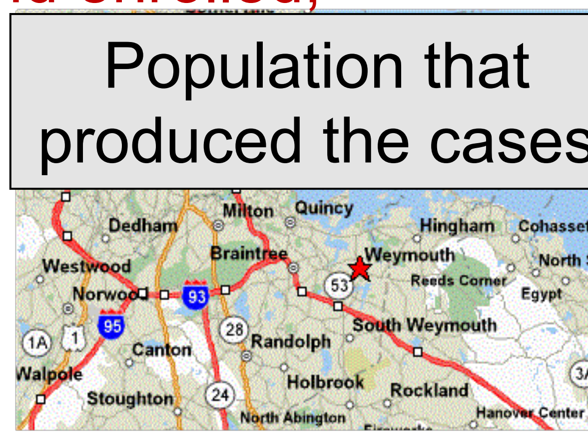
DDT
Exposure

Breast
Cancer
Cases

A method of sampling a population in which cases of disease are identified and enrolled,



Population that
produced the cases



DDT
Exposure

… and a sample of the population that produced the cases is identified and enrolled.

The distribution of past exposures is then compared in order to estimate the risk ratio

How many cigarette smokers did it take to produce 256 people with bladder cancer?

Incidence?

Bladder Cancer

Cigarette Use	Case	Control	
Yes	256 (a)	202 (b)	?
No	43 (c)	85 (d)	?
	299	287	586

Odds of a case being a smoker = 256/43

Odds of a control being a smoker = 202/85

$$\text{Odds Ratio} = \frac{a/c}{b/d} = \frac{ad}{bc} = \frac{256/43}{202/85} = 2.5$$

OR as an Estimate of RR

If the disease is uncommon, then …

…the exposure status of non-diseased people provides an estimate of the exposure of the population that produced the cases.

… the odds ratio provides a legitimate estimate of the risk ratio.

The Odds Ratio is Interpreted Like a Risk Ratio

Example:

Smoking and bladder cancer; Odds Ratio = 2.5:

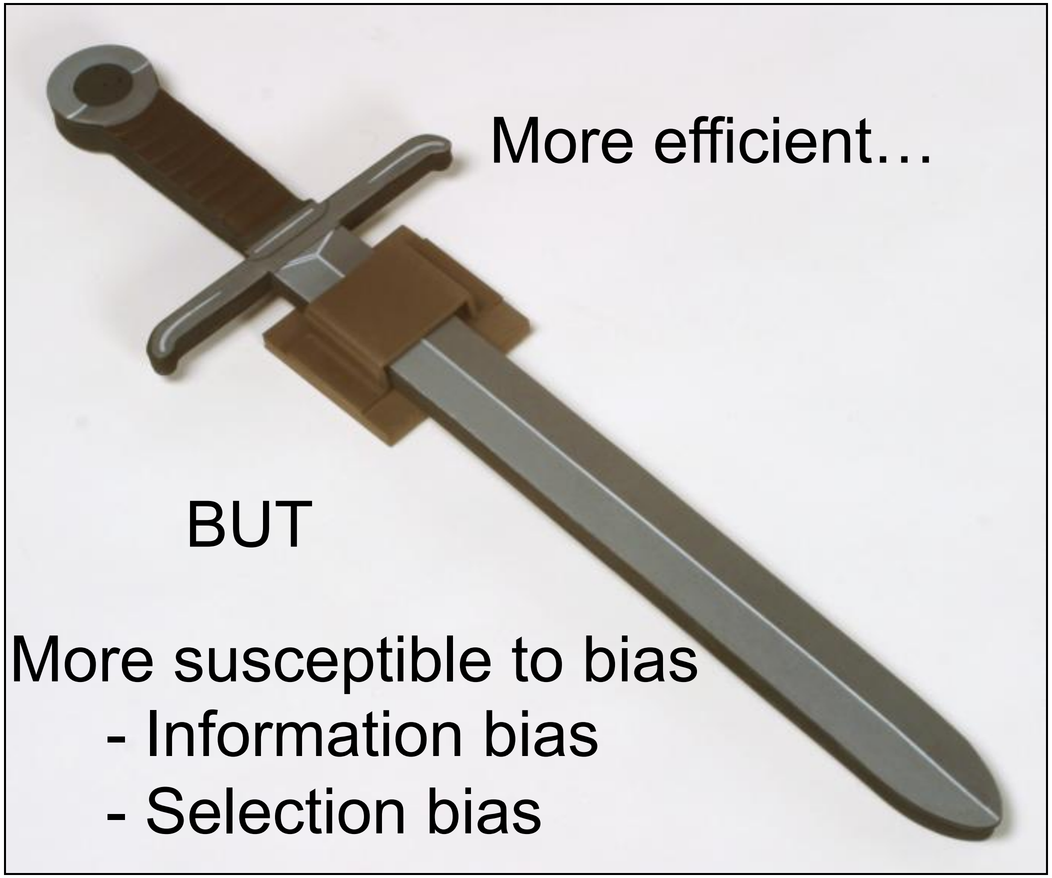
"In this study individuals who smoked had 2.5 times the risk of bladder cancer compared to those who did not smoke."

When Are Case-Control Studies Advantageous?

The Major Advantage:
Very efficient for studying rare diseases and diseases with long latency.

- Rare Disease
- Long Latent Period
- Exposure Data Hard to Get
- Dynamic Population
- (Little Known of Disease)

Since Exposure & Disease Have Already Occurred….



Preventing Misclassification

Measures that increase the accuracy of the data reduce both non-differential and differential misclassification.

- Look for clear, specific definitions of outcomes & exposures.
- Assess accuracy of data. Was information of the same quality in all groups?
- Is it possible to verify the data?

Finding Cases

1) Hospital/Clinic-based cases (easier to find)

2) Population-based cases, e.g. state cancer registry
(not biased by factors drawing a patient to a particular hospital)

- Newly diagnosed cases?
  - ➢ They provide better information (less misclassification and less recall bias)

- A random sample of available cases or focused subset?
  - ➢ It's easier to generalize with a random sample, but… it is essential to get accurate data (validity).

Finding the Comparison Group ("Controls")

- Selected by a similar mechanism, i.e., they "would" have been in the case group if they had been diseased.
- Should be comparable to the cases with respect to other (confounding) factors that affect the outcome.

1) **Hospital controls**
   - Easy to find; more cooperative
   - Similar tendency to remember past exposures
   - Similar in other factors (ethnic/socioeconomic)
   - **BUT** they aren't healthy & may have diseases with the same risk factors. In a study of smoking & lung cancer "non-cancer" controls might be hospitalized for heart disease, emphysema, bronchitis.

2) **Controls from the general population**
   - Random digit dialing, residence lists, drivers' license records

3) **Special control groups** (neighbors, relatives) provide control for heredity and environment