

After World War II, evidence began mounting that there was a link between cigarette smoking and pulmonary carcinoma (lung cancer). In the 1950s, two now classic articles were published on the subject. One of these studies was conducted in the United States by Wynder and Graham (“Tobacco Smoking as a Possible Etiologic Factor in Bronchiogenic Cancer,” *Journal of the American Medical Association*, 143(4)). They found records from a large number (684) of patients with proven bronchiogenic carcinoma (a specific form of lung cancer) in hospitals in California, Colorado, Missouri, New Jersey, New York, Ohio, Pennsylvania, and Utah. They personally interviewed 634 of the subjects to identify their smoking habits, occupation, exposure to dust and fumes, alcohol intake, education, and cause of death of parents and siblings. Thirty-three subjects completed mailed questionnaires, and information for the other 17 was obtained from family members or close acquaintances. Of those in the study, the researchers focused on 605 male patients with the same form of lung cancer. Another 1332 hospital patients with similar age and economic distribution (including 780 males) without lung cancer were interviewed by these researchers in St. Louis and by other researchers in Boston, Cleveland, and Hines, Illinois.[Investigation 1.2.1: Smoking & Lung Cancer, ISCAM]

The following two-way table replicates the counts for the 605 male patients with the same form of cancer and for the “control-group” of 780 males.

	none < 1/day	light 1-9/day	mod heavy 10-15/day	heavy 16-20/day	excessive 21-34/day	chain 35+/day
patients	8	14	61	213	187	122
controls	114	90	148	278	90	60

Given the results of the study, do you think we can generalize from the sample to the population? Explain (what is the difference between the sample and the population here?).

	chain smoking	no smoking	
cancer	122	8	130
healthy	60	114	174
	182	122	304

Group A	Group B
expl = smoking status resp = lung cancer	expl = lung cancer resp = smoking status

- If lung cancer is considered a success and no smoking is baseline:

$$RR =$$

$$OR =$$

- If chain smoking is considered a success and healthy is baseline:

$$RR =$$

$$OR =$$

Case-control: identify observational units by response

Cohort: identify observational units by explanatory variable

Cross-classification: identify observational units without respect to either variable

Conclusion1: Can we conclude causation here?

Conclusion2: Can we estimate baseline rates?

Retrospective versus Prospective Studies

Note, many many books *define* retrospective as synonymous with case-control. That is, they define a retrospective study to be one in which the observational units were chosen based on their status of the response variable. I disagree with that definition. As you see in the definition below, retrospective studies are defined based on the when the variables were *measured* (is the explanatory variable measured before or after the response?). I've also given a quote from the Kuiper text where retrospective is defined as any study where historic data are collected (I like this definition less). Regardless of how the data collection is defined, the most important thing for you to focus on is: *what conclusion can be made from how the data were collected?*.

Studies can be classified further as either prospective or retrospective. We define a prospective study as one in which exposure and covariate measurements are made before the cases of illness occur. In a retrospective study these measurements are made after the cases have already occurred... Early writers referred to cohort studies as prospective studies and to case-control studies as retrospective studies because cohort studies usually begin with identification of the exposure status and then measure disease occurrence, whereas case-control studies usually begin by identifying cases and controls and then measure exposure status. The terms prospective and retrospective, however, are more usefully employed to describe the timing of disease occurrence with respect to exposure measurement. For example, case-control studies can be either prospective or retrospective. A prospective case-control study uses exposure measurements taken before disease, whereas a retrospective case-control study uses measurements taken after disease. [**Modern Epidemiology**, 2nd edition, Rothman & Greenland, page 74]

Retrospective cohort studies also exist. In these designs past (medical) records are often used to collect data. As with prospective cohort studies, the objective is still to first establish groups based on an explanatory variable. However since these are past records the response variable can be collected at the same time. [**Stat2Labs**, S. Kuiper, chapter 6, page 24]