Comparing a binary outcome between groups

Overview:

- Risk Ratio, Relative Risk (RR)
- Chi-square test, Fisher's Exact test
- Risk Difference (RD)
- Number Needed to Treat (NNT)
- Rate Ratio (RR)
- Odds Ratio (OR)

1

Flow-chart for popularly used statistical tests

Q1,Univariate /Mutivariable	Q2, Difference /Correlation	Q3, Paired / related	Q4, Q5 Type of outcome (Normality)	Q6, No. of groups	Q7,sample size	Valid Tests
Univariate	Difference	Independent (un-paired)	Continuous (Normal)	2		Student's t-test
				>2		One-way ANOVA
			Continuous (Non-normal) / Ordered categorical	2		Mann-Whitney U test
				>2		Kruskal-Wallis H test
			Nominal	2	<20	Fisher's exact test
				≥2	≥20	Chi-square test
			Time to Event			Log-Rank test(Kaplan-Meier plot)
		Dependent (paired)	Continuous (Normal)	2		Paied-t test
				>2		Repeated measured ANOVA Mixed effect Regression
			Continuous (Non-normal) / Ordered categorical	2		Wilcoxon signed-rank test
				>2		Friedman test
			Nominal	2		McNemar's test
	Correlation		Continuous (Normal)			Pearson's correlation (r)
			Continuous (Non-normal) /ordered			Spearman's correlation (rs)
			Nominal (2 levels)	2		Spearman/Kappa (Agrreement)
Multivariable		Independent (un-paired)	Continuous (Normal residulas)			Linear Regression
			Continuous (Non-normal residulas)			Linear Regression *
			Ordered categorical			Ordered Logistic Regression
			Nominal (2 levels)			Binary Logistic Regression
			Nominal (>2)			Multinomial Logistic Regression
			Time to Event			Cox Proportional Hazard Regression
		Dependent (paired)	Continuous (Normal residulas)			Linear Mixed Effect Regression
			Continuous (Non-normal residulas)			Linear Mixed Effect Regression *
			Ordered categorical			Generalized Estimation Equation (GEE)
			Nominal (2 levels)			Generalized Estimation Equation (GEE)

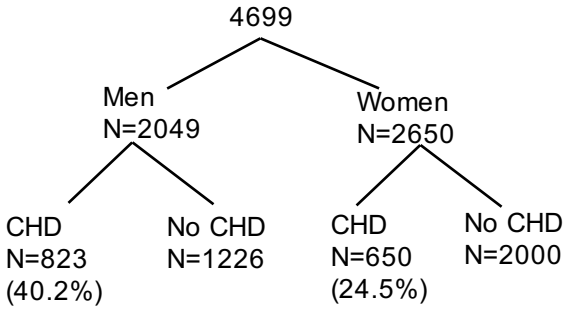
* Transform outcome variables for normalizing residuals

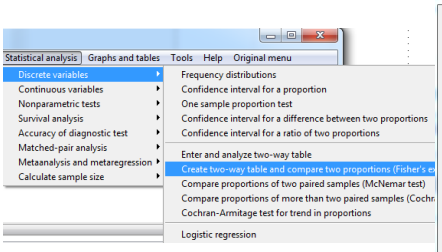
Created based on Publishing Your Medical Research Paper, by Daniel Byrne, Williams and Wilkins (1998)

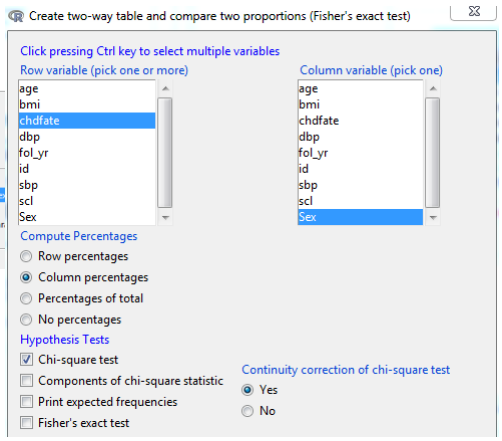
Motivating Example!

Framingham.csv contains data from Framingham study (prospective cohort study) where people without a coronary heard disease were followed for more than 30 years for cardiovascular events. This week's task is to estimate a relation between an exposure of interest (i.e., being a male) to a risk of having a binary event (i.e., development of a coronary heart disease).

We want to estimate risk of CHD separately for men and women.







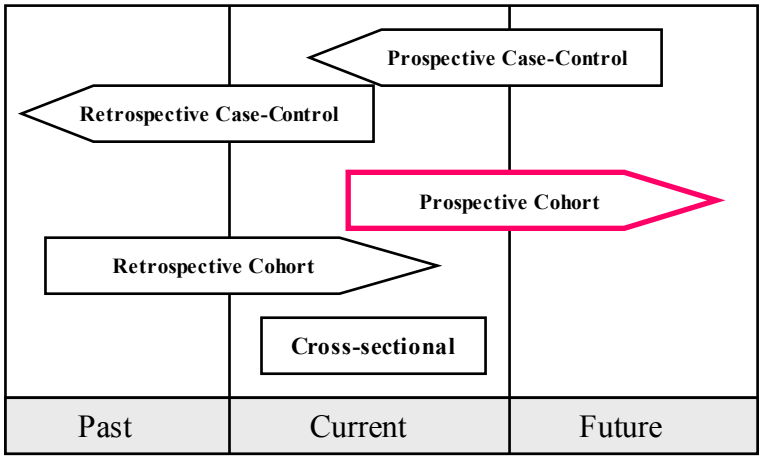

```
> .Table
      Sex
chdfate   Men Women
      0 1226   2000
      1  823    650

> colPercents(.Table)  # Column
Percentages
                Sex
chdfate     Men  Women
  0          59.8   75.5
  1          40.2   24.5
  Total     100.0  100.0
  Count    2049.0 2650.0
```

Proportion of CHD:

Over all:	31.3%
Gender Specific	
Men	40.2%
Women	24.5%

Study Designs



5

Risk ≈
Cumulative Incidence Proportion (%) =

$$\frac{\text{No. of new cases of a disease occurring in a defined period of time}}{\text{Total number of people at risk for the disease (and without the disease) in the population at beginning of period}}$$

In a cohort study, subjects who are followed should be free of outcome at beginning to estimate incidence, and also must have the same risk of developing the outcome.

How do we compare risk of CHD between males and females?

$$\text{Risk ratio (Relative Risk)} = \frac{\text{Proportion of people with CHD among males}}{\text{Proportion of people with CHD among females}}$$

$$= \frac{40.2\%}{24.5\%} = 1.64$$

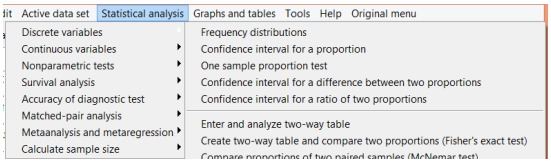
Being a male increases the risk of having CHD by 64% comparing to that of a female.

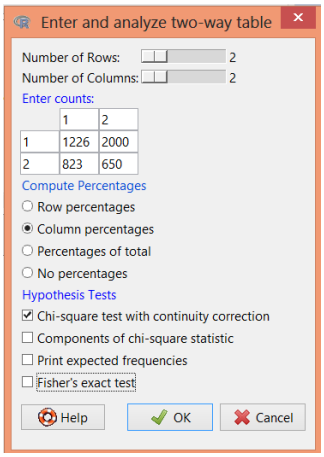
Computing p-value for 2x2 table.

There are 2 ways to compute p-value for 2 x 2 table.

1. Use data from each subject.
2. Enter frequency data.

Computing p-value for 2x2 table: Use frequency data






```
> summary.table
    X1   X2 Chisq.p.value
1 1226 2000      3.07e-30
2  823  650
```

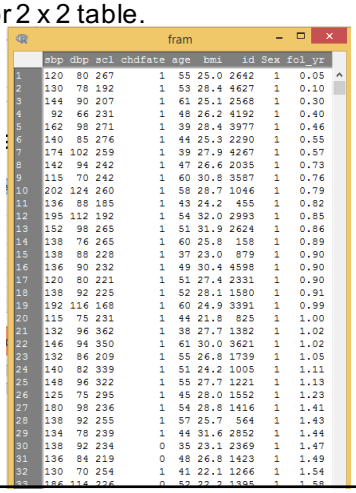
There was a statistically significant difference of cumulative incidence proportion of CHD between males and females. CI of CHD is 823/2049=40.2% for males, and 24.5% for females, which provides Risk Ratio (RR) of 1.64, p<0.001.

Computing p-value for 2x2 table.

There are 2 ways to compute p-value for 2 x 2 table.

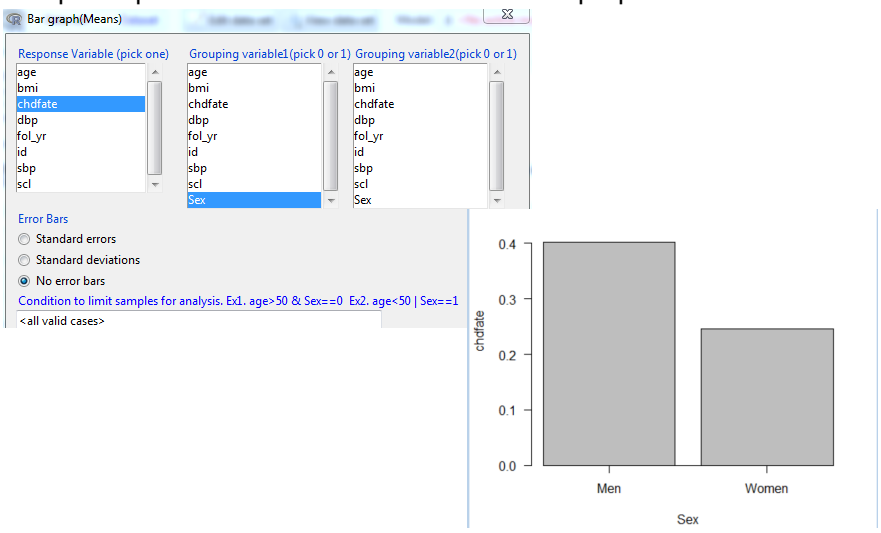
1.Use data from each subject.

2.Enter frequency data.



	sbp	dbp	scl	chdfate	age	bmi	id	Sex	fol_yr
1	120	80	267	1	55	25.0	2642	1	0.05
2	130	78	192	1	53	28.4	4627	1	0.10
3	144	90	207	1	61	25.1	2568	1	0.30
4	92	66	231	1	48	26.2	4192	1	0.40
5	162	98	271	1	39	28.4	3977	1	0.46
6	140	85	276	1	44	25.3	2290	1	0.55
7	174	102	259	1	39	27.9	4267	1	0.57
8	142	94	242	1	47	26.6	2035	1	0.73
9	115	70	242	1	60	30.8	3587	1	0.76
10	202	124	260	1	58	28.7	1046	1	0.79
11	136	88	185	1	43	24.2	455	1	0.82
12	195	112	192	1	54	32.0	2993	1	0.85
13	152	98	265	1	51	31.9	2624	1	0.86
14	138	76	265	1	60	25.8	158	1	0.89
15	138	88	228	1	37	23.0	879	1	0.90
16	136	90	232	1	49	30.4	4598	1	0.90
17	120	80	221	1	51	27.4	2331	1	0.90
18	138	92	225	1	52	28.1	1580	1	0.91
19	192	116	168	1	60	24.9	3391	1	0.99
20	115	75	231	1	44	21.8	825	1	1.00
21	132	96	362	1	38	27.7	1382	1	1.02
22	146	94	350	1	61	30.0	3621	1	1.02
23	132	86	209	1	55	26.8	1739	1	1.05
24	140	82	339	1	51	24.2	1005	1	1.11
25	148	96	322	1	55	27.7	1221	1	1.13
26	125	75	295	1	45	28.0	1552	1	1.23
27	180	98	236	1	54	28.8	1416	1	1.41
28	138	92	255	1	57	25.7	564	1	1.43
29	134	78	239	1	44	31.6	2852	1	1.44
30	138	92	234	0	35	23.1	2369	1	1.47
31	136	84	219	0	48	26.8	1423	1	1.49
32	130	70	254	1	41	22.1	1266	1	1.54

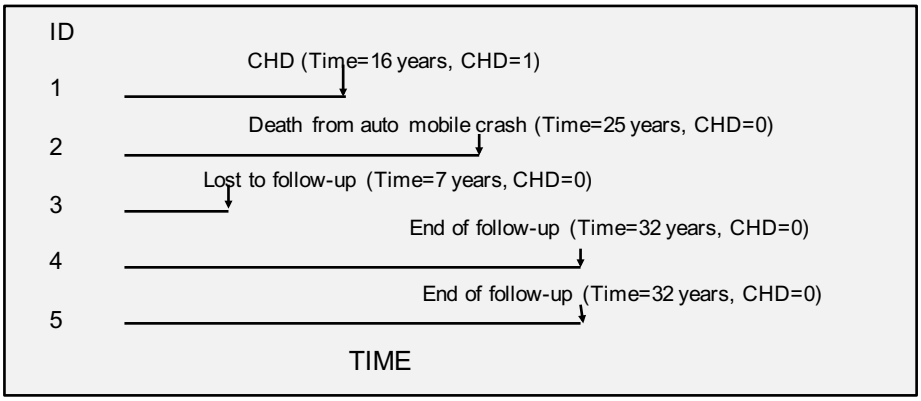
Graphical presentation of cumulative incidence proportion of CHD



Risk ≈
Rate (per person year) =

No. of new cases of a disease occurring in a defined period of time
Total number of follow-up time of people at risk for the disease in the population.

$$\text{Incidence rate of CHD} = \frac{1}{16 + 25 + 7 + 32 + 32} = \frac{1}{112} \text{ per person-year}$$



CHD rate per person-year among males = 823/42259 = 0.402/20.6 =0.0195
This means that 0.0195 person with CHD is observed if we follow a patient for 1 year of time. (0.0195/1 person year) ← Average annual risk of CHD

or equivalently,

If 1000 people are followed for 1 year, then 19.5 people with CHD would be observed with CHD.

CHD rate per person-year among females = 650/61451 = 0.245/23.2 =0.0106

How do we compare risk o CHD between males and females?

$$\text{Rate Ratio} = \frac{\text{Rate of CHD among males}}{\text{Rate of CHD among females}}$$

$$= \frac{0.0195}{0.0106} = 1.83$$

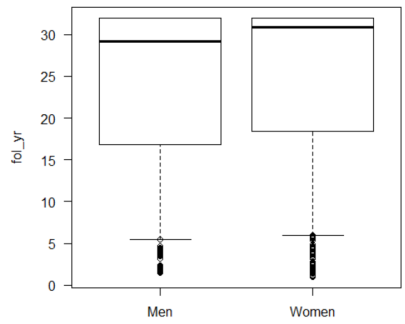
Being a male increases the risk of having CHD by 83% comparing to that of a female.

Risk Ratio = 1.64

Rate Ratio = 1.83

which one is more accurate to estimate the increased risk of CHD for male compared with that of a female?

Mean follow-up time in days by disease class among patients



Among people with No CHF

Men are followed for a shorter period of time than women, thus the risk estimate by using the proportion of CHD in fact under-estimates the additional risk increase of CHD by men.

Under estimation

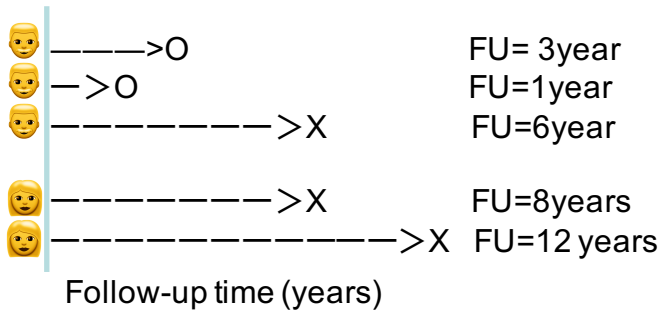
$$\text{Risk ratio} = \frac{\text{Proportion of people with CHD among males}}{\text{Proportion of people with CHD among females}}$$

Thus if follow-up time is meaningfully different between comparison groups, results can be biased when only proportions are compared (i.e., using chi-square test or logistic regression) by failing to take into account varying follow-up time.

Thus when follow-up time varies among subjects, we must consider using:

Rate

Rate vs Proportion



FU= 3year
FU=1year
FU=6year
FU=8years
FU=12 years

Men Incidence Proportion 33%
Women Incidence Proportion 100%

Risk Ratio = 1/3

Rate 1/10
Rate 2/20 =1/10

Rate Ratio = 1

Risk Ratio vs Risk Difference vs Number Needed to Treat (NNT)

Blood Pressure Lowering drug
ACE Inhibitor

The New England
Journal of Medicine

VOLUME 342 JANUARY 20, 2000 NUMBER 3

EFFECTS OF AN ANGIOTENSIN-CONVERTING–ENZYME INHIBITOR, RAMIPRIL, ON CARDIOVASCULAR EVENTS IN HIGH-RISK PATIENTS

THE HEART OUTCOMES PREVENTION EVALUATION STUDY INVESTIGATORS*

ABSTRACT

Background Angiotensin-converting–enzyme inhibitors improve the outcome among patients with left ventricular dysfunction, whether or not they have heart failure. We assessed the role of an angiotensin-converting–enzyme inhibitor, ramipril, in patients who were at high risk for cardiovascular events but who did not have left ventricular dysfunction or heart failure.

Methods A total of 9297 high-risk patients (55 years of age or older) who had evidence of vascular disease or diabetes plus one other cardiovascular risk factor and who were not known to have a low ejection fraction or heart failure were randomly assigned to receive ramipril (10 mg once per day orally) or matching placebo for a mean of five years. The primary outcome was a composite of myocardial infarction, stroke, or death from cardiovascular causes.

The trial was a two-by-two factorial study evaluating both ramipril and vitamin E. The effects of vitamin E are reported in a companion paper.

Results A total of 651 patients who were assigned to receive ramipril (14.0 percent) reached the primary end point, as compared with 826 patients who were assigned to receive placebo (17.8 percent) (relative risk, 0.78; 95 percent confidence interval, 0.70 to 0.86; P<0.001). Treatment with ramipril reduced the rates of death from cardiovascular causes (6.1 percent, as compared with 8.1 percent in the placebo group; relative risk, 0.74; P<0.001), myocardial infarction (9.9 percent vs. 12.3 percent; relative risk, 0.80; P<0.001), stroke (3.4 percent vs. 4.9 percent; relative risk, 0.68; P<0.001), death from any cause (10.4 percent vs. 12.2 percent; relative risk, 0.84; P=0.005), revascularization procedures (16.0 percent vs. 18.3 percent; relative risk, 0.85; P=0.002), cardiac arrest (0.8 percent vs. 1.3 percent; relative risk, 0.63; P=0.03), heart failure (9.0 percent vs. 11.5 percent; relative risk, 0.77; P<0.001), and complications related to diabetes (6.4 percent vs. 7.6 percent; relative risk, 0.84; P=0.03).

Conclusions Ramipril significantly reduces the rates of death, myocardial infarction, and stroke in a

20

TABLE 3. INCIDENCE OF THE PRIMARY OUTCOME AND OF DEATHS FROM ANY CAUSE.

OUTCOME	RAMIPRIL GROUP (N=4645)	PLACEBO GROUP (N=4652)	RELATIVE RISK (95% CI)*	Z STATISTIC	P VALUE†
	no. (%)				
Myocardial infarction, stroke, or death from cardiovascular causes‡	651 (14.0)	826 (17.8)	0.78 (0.70–0.86)	−4.87	<0.001
Death from cardiovascular causes§	282 (6.1)	377 (8.1)	0.74 (0.64–0.87)	−3.78	<0.001
Myocardial infarction§	459 (9.9)	570 (12.3)	0.80 (0.70–0.90)	−3.63	<0.001
Stroke§	156 (3.4)	226 (4.9)	0.68 (0.56–0.84)	−3.69	<0.001
Death from noncardiovascular causes	200 (4.3)	192 (4.1)	1.03 (0.85–1.26)	0.33	0.74
Death from any cause	482 (10.4)	569 (12.2)	0.84 (0.75–0.95)	−2.79	0.005

14.0/17.8=

*CI denotes confidence interval.

†P values were calculated with use of the log-rank test.

‡In the substudy, 34 of 244 patients (13.9 percent) assigned to take a low dose of ramipril (2.5 mg per day) reached the composite end point, as compared with 31 of 244 assigned to take 10 mg of ramipril per day (12.7 percent) and 41 of 244 assigned to placebo (16.8 percent). The inclusion of the data from the low-dose group did not change the overall results (relative risk of the primary outcome, 0.78; 95 percent confidence interval, 0.70 to 0.86).

§All patients with this outcome are included.

TABLE 3. INCIDENCE OF THE PRIMARY OUTCOME AND OF DEATHS FROM ANY CAUSE.

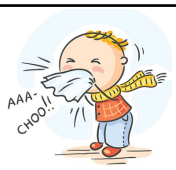
OUTCOME	RAMIPRIL GROUP (N=4645)	PLACEBO GROUP (N=4652)	RELATIVE RISK (95% CI)*	Z STATISTIC	P VALUE†
	no. (%)				
Myocardial infarction, stroke, or death from cardiovascular causes‡	651 (14.0)	826 (17.8)	0.78 (0.70–0.86)	−4.87	<0.001

14.0/17.8=

Relative risk = 14.0/17.8 = 0.78 = 22% Risk reduction
Risk difference = 17.8-14.0=3.8%
(This indicates that 3.8% of pts
did not have the event because of the drug)
1/3.8% =26.3 (In order to prevent 1 person from
getting the event, 26.3patients needed to be
treated with this drug.

results (relative risk of the primary outcome, 0.78; 95 percent confidence interval, 0.70 to 0.86).

§All patients with this outcome are included.



OR

This year's flu vaccine have an efficacy of 18%（RR=0.82）

Prevalence proportion of flu is thought to be 6%,
which means that prevalence proportion of flu is reduced to
6x0.82＝4.92%

Thus, there is only 1.08% of people who could prevent from getting a flu.

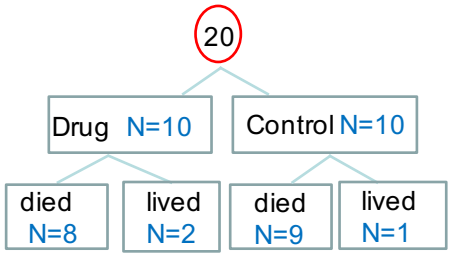
Risk Difference (RD)

Number Needed to Treat（NNT)= 1/(0.06 – 0.049)=92.6
In order to prevent 1 person from having a flu, 92.6 people need to be vaccinated.

NNT = 1/ RD

Risk Ratio vs Odds Ratio

Risk Ratio vs Odds Ratio



	Drug	Placebo	Ratio
Risk	8/10	9/10	$\frac{8/10}{9/10}=0.88$
Odds	8/2	9/1	$\frac{8/2}{9/1}=0.44$

Property of Odds Ratio and Relative Risk

- $0 <= Odds < \infty$
- $0 <= Proportion <= 1$

OR = 1 ⇔ RR=1 No association
(p-value assessed OR=1, thus this is equivalent with assessing RR=1

RR > 1 ⇔ OR > 1
RR < 1 ⇔ OR < 1

- If RR > 1 then OR > RR
  If RR < 1 then OR < RR

26

Comparison between Odds Ratio and Risk Ratio

Risk ratio=2	
Risk in the unexposed group	Corresponding odds ratio
0.001	2.002
0.005	2.010
0.01	2.020
0.05	2.111
0.1	2.25
0.2	3.5
0.3	6.0
0.4	11.0
0.5	∞

The use of odds ratio is not a problem for rare outcomes because OR would be similar to RR. Although for common outcomes, OR can be a lot different from RR, therefore an odds ratio should not be used as RR (used to quantify a risk increase).

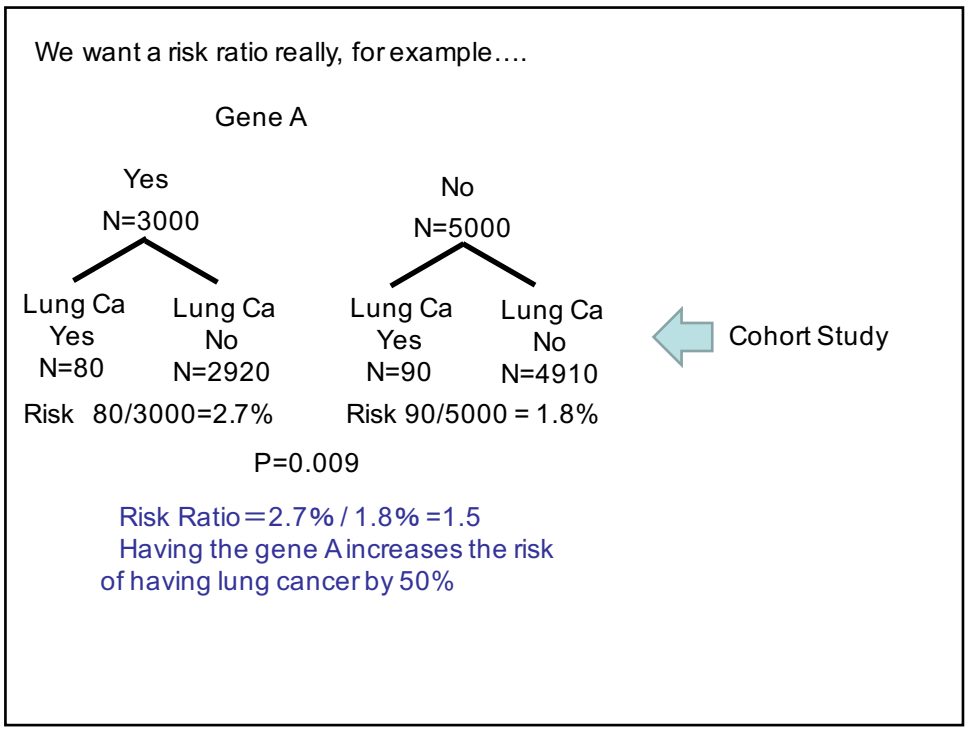
27

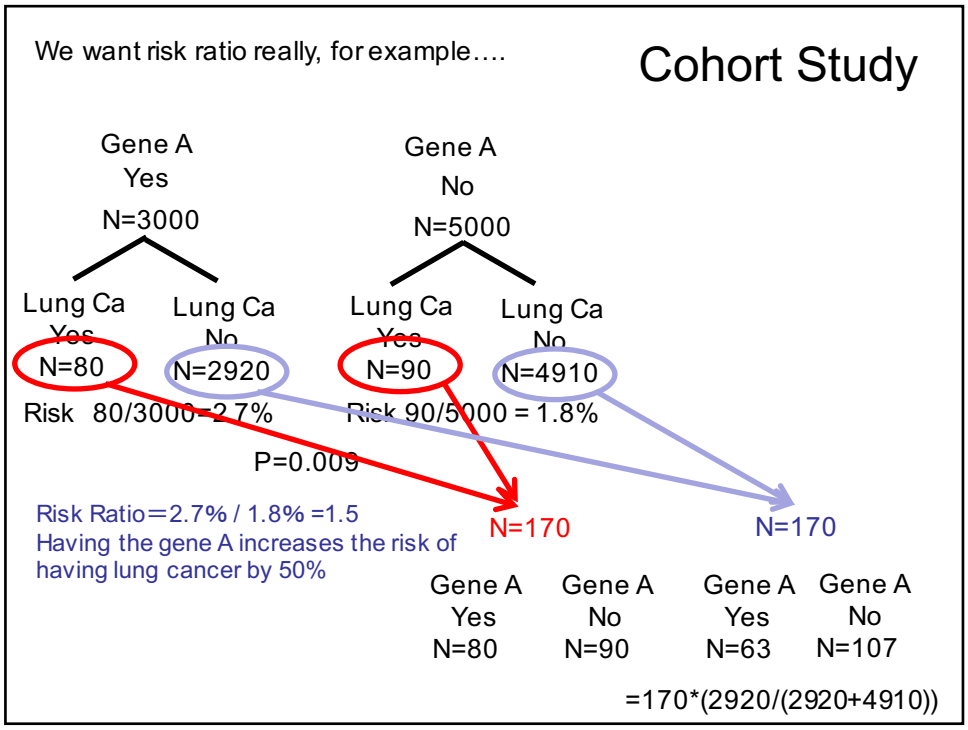
Question:
Then, do we need to use OR not RR?

Answer: Don't have to use OR

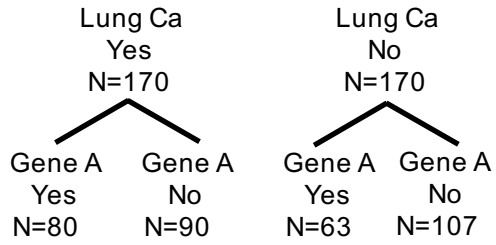
Except for ….

Case-control study


We want a risk ratio really, for example….
Gene A
Yes
N=3000
No
N=5000
Lung Ca Yes N=80
Lung Ca No N=2920
Lung Ca Yes N=90
Lung Ca No N=4910
Cohort Study
Risk 80/3000=2.7%
Risk 90/5000 = 1.8%
P=0.009
Risk Ratio=2.7% / 1.8% =1.5
Having the gene A increases the risk of having lung cancer by 50%


We want risk ratio really, for example….
Cohort Study
Gene A Yes N=3000
Gene A No N=5000
Lung Ca Yes N=80
Lung Ca No N=2920
Lung Ca Yes N=90
Lung Ca No N=4910
Risk 80/3000=2.7%
Risk 90/5000 = 1.8%
P=0.009
Risk Ratio=2.7% / 1.8% =1.5
Having the gene A increases the risk of having lung cancer by 50%
N=170
N=170
Gene A Yes N=80
Gene A No N=90
Gene A Yes N=63
Gene A No N=107
=170*(2920/(2920+4910))

Case-Control Study



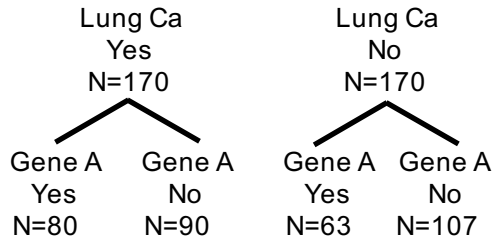
A person with Gene A has a risk of having lung ca
80/ (80+63)= 56%

A person without Gene A has a risk of having lung ca
90/ (90+107)= 46%

Risk Ratio = 56%/46%=1.21
Having Gene A increases the risk of having lung ca by 21%

Isn't this the same as 1.5 On the previous slide???

Case-Control Study



~~A person with Gene A has a risk of having lung ca~~
~~80/ (80+63)= 56%~~

~~A person without Gene A has a risk of having lung ca~~
~~90/ (90+107)= 46%~~

~~Risk Ratio = 56%/46%=1.21~~
~~Having Gene A increases the risk of having lung ca by 21%~~

Instead, we use OR in a case-control study.

Odds Ratio = 1.27/0.84 = 1.51
This is in fact, very similar to Risk Ratio on Slide 36!!!!!!!!!!

In the previous page of the case-control study,

Cumulative incidence (risk) of LungCancer among gene A is 56%?
Cumulative incidence (risk) of LungCancer among no gene A is 46%?

Why do you think risk of Lung Cancer became so large in case-control study?

Do you see some thing wrong here??????

In case-control study, cases are OVER SAMPLEed, thus we cannot estimate cumulative incidence proportion (risk).

33

Case-control study <- Good study design for a rare disease

When event prevalence is less than 0.1 among control group, Odds ratio approximates Risk ratio pretty well. Thus, in a case-control study, or a study with rare outcome, it is OK to use OR to estimates RR.

Although, odds ratios are used all most always where outcome is binary even in a cohort study where relative risk can be obtained. Why? Logistic regression is a lot more widely used than relative risk regression. Logistic regression computes OR not RR.

Then people interpret an odds ratio as if it were a relative risk regardless of study design (common mistake by literature). And this is not OK in many cases.

Computing OR in R