How Causal Heterogeneity Can Influence Statistical Significance in Clinical Trials

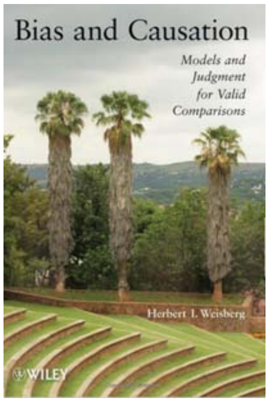
Milo Schield, W. M. Keck Statistical Literacy Project. Minneapolis, MN.

Abstract: Finding that an association is statistically insignificant in a clinical trial has two distinct explanations. (1) The association may be real, but the sample size is too small to distinguish the results from those due to chance. (2) The association is spurious – a coincidence due to chance. Based on Weisberg (2011), this paper argues that there is a third explanation: the association may be real but the mixture of potential outcomes – causal heterogeneity – may give results that are indistinguishable from those due to chance. These potential outcomes involve counterfactuals: outcomes that could have happened but did not. Causal heterogeneity exists when a treatment interacts to give different results with different units of a population. This paper uses a model of causal heterogeneity that is generally accessible. Statistical education should extend introductory statistics courses so they (a) show how potential outcomes affect statistical significance in clinical trials and (b) highlight the importance of causally-heterogeneous subgroups in determining whether a treatment effect is statistically significant.

Keywords: potential outcomes

1. Background

The most common claim in any introductory statistics course is that "association is not causation." There are two types of explanations: confounding and chance. Confounding is used most often. E.g., the Berkeley sex discrimination case (Wikipedia: Simpson's Paradox). See Schield (2006) for other examples. Chance or coincidence is used occasionally. See Schield (2012). This paper presents a third explanation: potential outcomes involving causal heterogeneity.



In discussing causal effects, Weisberg (2011) noted that "Modern statistical methods have achieved great success in many areas of application. However, these technical triumphs depend on the suppression of effect variability. When causal effects can vary substantially across individuals, there is an unavoidable ambiguity in the interpretation of aggregate effects. This ambiguity arises because there are many possible distributions of individual effects that are consistent with any particular aggregate effect."

"If the individual effects are approximately uniform or vary in an essentially random manner (i.e. unrelated to any potentially observable factors), then the population parameter remains relevant to any member of the population. However, suppose there exist potentially identifiable individual characteristics that are related to the causal effect. Then it may become feasible to specify subgroups based on these characteristics for which the overall parameter value is misleading." For more detail, see Weisberg's book (2010).

Consider the potential outcomes in the use of a treatment to kill weeds.

Potential Outcome	Treatment	Control	%
1. Doomed	Die	Die	P1
2. Intended Result	Die	Live	P2
3. Opposite Result	Live	Die	P3
4. Immune	Live	Live	P4

Table 1

The four rows are the four subgroups of interest – the four potential outcomes:

1. *Doomed*: they would have died regardless of whether or not they received the treatment.
2. *Intended-result*: They would have lived in the absence of the treatment but would have died as a result
3. *Opposite-result*: they would have died without the treatment, but would have lived with the treatment. The *opposite result* could be labeled as the *adverse effect* or the *unintended result*.
4. *Immune*: they would have lived regardless of whether or not they received the treatment.

In the absence of any treatment, the population would just have two groups: *lived* and *died*. There would be no counterfactuals. The subdivision of each group depends totally on the impact of the treatment on the subjects.

With the treatment, each of these four outcomes involves a counter-factual component: what did happen with the treatment along with what would have happened without that same treatment.

With a given treatment on a given population, we can only observe two things: (1) which group (*treatment* vs. *control*) the subjects were in and (2) the associated outcome (*die* vs. *survive*).

In the treatment group: Die = pTD= (P1+P2); Live = pTL = (P3+P4). In the control group: Die = pCD = (P1+P3); Live = pCL = (P2+P4). Here we have four equations with four unknowns. Normally, this yields a unique solution for each of the unknowns. But in this case, the equations are not independent, so we cannot solve for the unknowns. See Appendix A. If, on the other hand, the prevalence of any one of the potential outcomes were known, then the prevalence of the other three could be readily deduced. See Appendix B.

Since these counterfactual outcomes are generally unobservable, they are best described as "potential outcomes." The particular outcome depends on how the units in the population interact with the treatment. Depending on the treatment, some subjects that would have lived in the control group may either die (intended result) or live (immune). Similarly some subjects that would have died in the control group may either live (opposite effect) or die (doomed) in the treatment group. Since these outcomes are determined by the units of the population responding differently to the same treatment, these units exhibit *causal heterogeneity*. Causal heterogeneity exists when a population has subgroups that interact – deterministically – with a treatment to give different results for each subgroup.

Random assignment can control for confounders – but not for causally-related heterogeneity.

2. Relative Risk

A common measure of association is relative risk: RR.

Equation 1: RR = Fraction of Treatment that died / Fraction of Placebo that Died = (P1+P2) / (P1+P3).

Consider two special cases and the general case:

Equation 2: If P3 = 0, RR = (P1+P2)/P1. So RR>1 if P2 > 0

Equation 3: If P1 = 0, RR = P2/P3. So RR > 1 if P2 > P3

Equation 4: RR = [1 + (P2/P1)] / [1 + (P3/P1)]. So RR > 1 if P2 > P3.

Although P1 has no influence in determining whether RR is bigger or smaller than unity, it can obscure the effects of the treatment. To see this, consider a simple case: no adverse effects. P3 = 0. RR = (P1+P2)/P1.

As the proportion of the "doomed" weeds increases (As P1 increases, assuming P2 > 0), the relative risk of death among the weeds decreases.

3. Example of Causal Heterogeneity

In the following figures, the treatment effect (P2) is fixed at 10% with N = 150: 75 in the treatment group; 75 in the control group. See Appendix B. When the doomed subgroup is 10%, the relative risk is 2.0. Table 5 shows that as the percentage that are doomed increases, the relative risk of dying decreases and approaches unity.

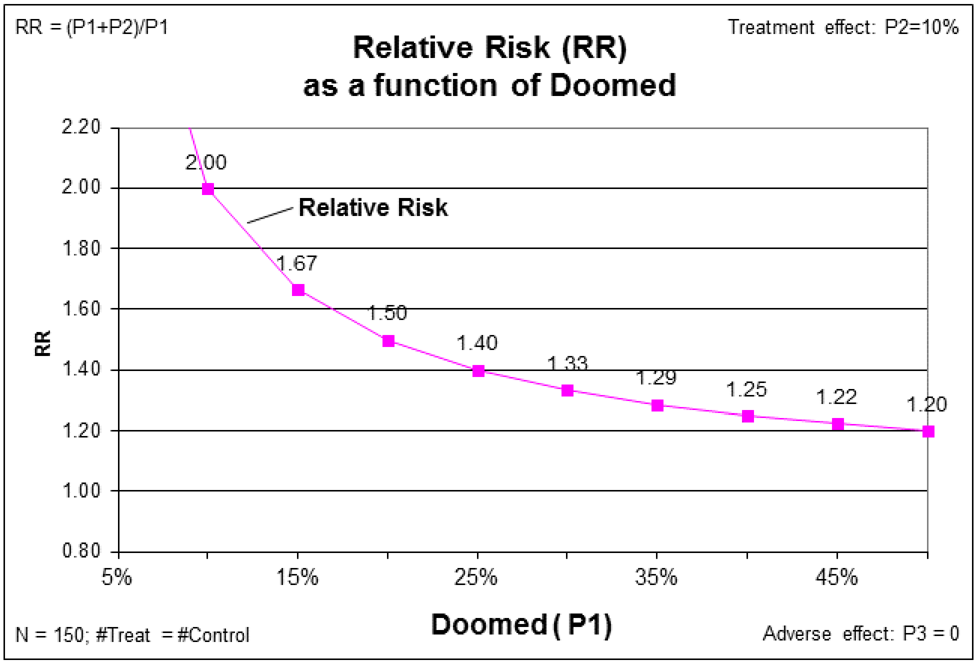


Figure 1

This influence on the association influences the limits of the associated 90% confidence intervals. Figure 2 has the same horizontal and vertical axes as Figure 1 and the same conditions where the treatment effect is fixed at 10% (P2). This graph also shows the lower limit of the 90% confidence interval. It starts at 1.44 on the left and equals unity between 25 and 30%. Appendix C shows how the 90% margin of error is calculated. Table 8 and Table 9 in Appendix D show the data used to generate these figures.

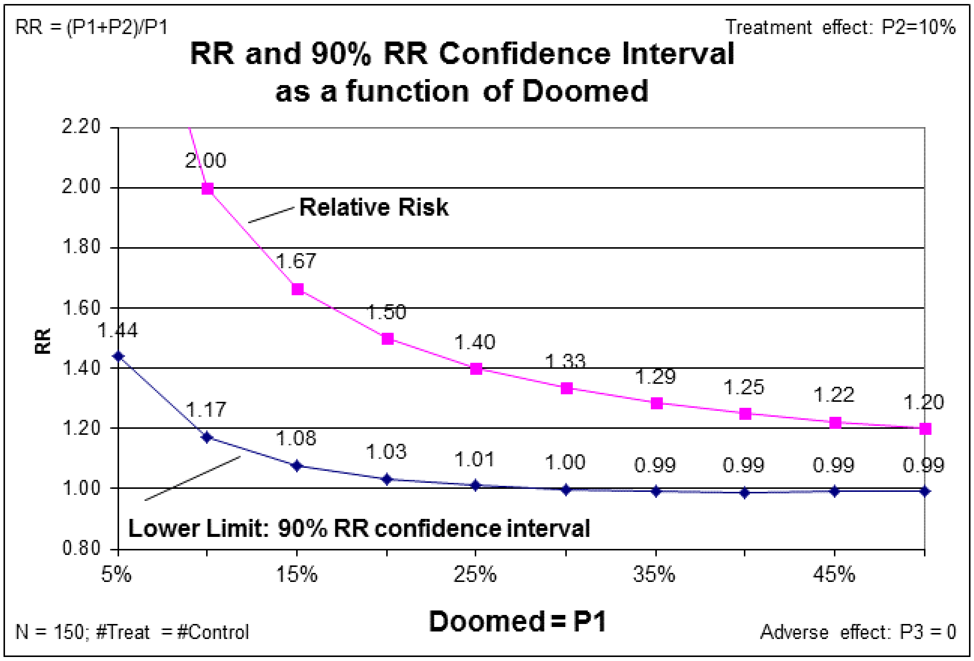


Figure 2

What is the effect of the doomed prevalence (P1) on the statistical significance of a relative risk?

In this paper, “statistical significance” is defined as alpha < 5% where alpha is the probability of getting a critical value or something more extreme due to chance.

On this basis, a relative risk is statistically significant when the lower limit of a 90% confidence interval is higher than unity. Recall that a two-sided 90% confidence interval has 5% below the lower limit. Hence when the lower limit is below unity, the association is not statistically significant. Conversely, when the lower limit is above unity, the relative risk is statistically significant.

The vertical line identifies the doomed prevalence at which the lower limit of the 90% confidence interval equals unity. To the right, the results are not statistically significant; to the left they are. In this case a treatment that is causally effective for 10% of the population is statistically significant when the percentage who are doomed is less than 30%. It is statistically insignificant when the percentage who are doomed is at least 30%. Note that the percentage who are doomed controls whether the treatment effect is statistically significant – or not!

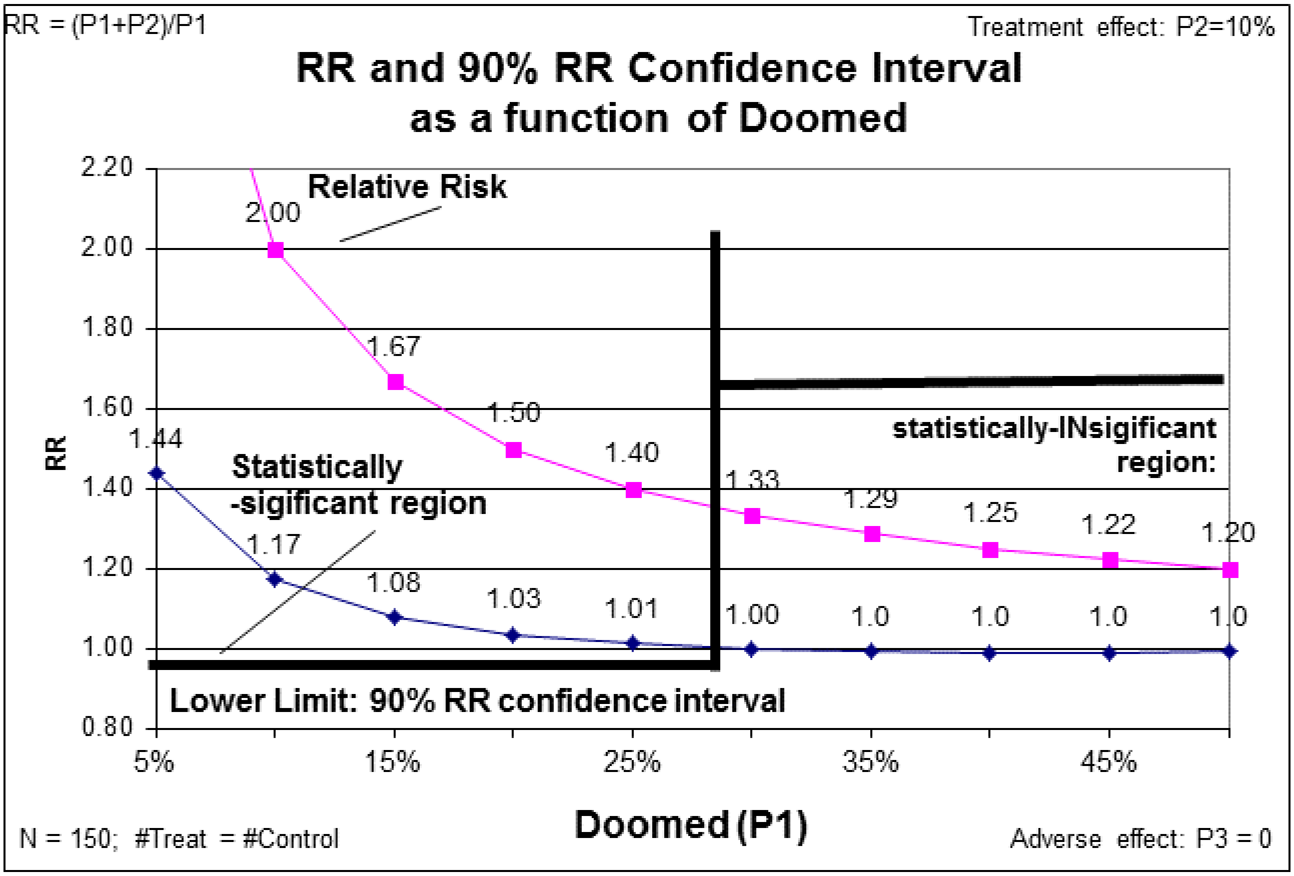


Figure 3

To review. When an association in a clinical trial is not statistically significant, there are three explanations:

1) *No real difference in the averages for the two groups*
2) Difference is real but is not visible because it is *confused with (masked by) chance* in a small sample
3) Difference is real but is not visible because it is *diluted*/masked by *causal heterogeneity*

4. Importance of Causal Heterogeneity

How important is this causal heterogeneity? Causal Heterogeneity is a "Big Deal". Drug companies spend billions per year on clinical trials. Many – if not most – give results that are not statistically significant, or they are rejected because of adverse effects. What if many of these rejected treatments

- were extremely effective for a population subgroup?
- had minimal adverse effects for a subgroup?

Could it be that our model of statistical significance and the design of clinical trials is largely responsible for the high cost of new drugs in the US?

Suppose that in their preliminary studies, drug companies were able to identify subgroups for which the treatment was most effective. Assuming this result could be replicated in a clinical trial and be statistically significant, a drug that otherwise might be rejected could go on to provide health benefits for a certain subset of the population.

Causal heterogeneity is certainly a major factor when dealing with human beings. People are complex subjects. That may be one reason why social statistics are so radically different from mathematical and non-social statistics. Winkler (2009, 2011).

5. Recommended Extensions

Future papers should investigate the use of filters (or the selection of subgroups) that reduce the percentage who are doomed (P1) and the percentage who exhibit the unintended (adverse) effects (P3) thus increasing the percentage who exhibit the intended effects (P2). The use of filters fits nicely with Weisberg's (2011) recommendation. "Rather than narrowly focusing on whether or not the treatment "works" in general, we should ask a better question.

- *For whom* (if anyone) is the treatment beneficial and *for whom* is it harmful?
- What individual and circumstantial characteristics are conducive to a positive (or negative) response?

To answer such questions will require a more flexible approach to design and analysis of RCTs."

6. Conclusion

Complex subjects involve four potential outcomes relative to a given treatment. For death: Doomed, Immune, Killed by Treatment, and Saved by Treatment. In clinical trials with complex subjects subject to causal heterogeneity, statistical significance is determined by sample size and by the prevalence of the various potential outcomes. Statistical education should extend introductory statistics courses so that they:

- show how potential outcomes affect statistical-significance in clinical trials
- highlight the importance of causally-heterogeneous subgroups (e.g., doomed) in determining whether a treatment effect is statistically significant.

References:

Schield, M. (2006). Presenting Confounding and Standardization Graphically. STATS Magazine, American Statistical Association. Fall 2006. pp. 14-18. See www.StatLit.org/pdf/2006SchieldSTATS.pdf.

Schield, M. (2012) e-COTS: Big Data Generates Beguiling Coincidences. See www.causeweb.org/ecots/posters/19/

Weisberg, H. I. (2010). *Bias and Causation: Models and Judgment for Valid Comparisons*. John Wiley & Sons.

Weisberg, Herbert I. (2011). Statistics and Clinical Trials: Past, Present and Future. ASA Proceedings of the Section on Statistical Education. [CD-ROM], 1547 – 1561. www.StatLit.org/pdf/2011Weisberg-JSM.pdf.

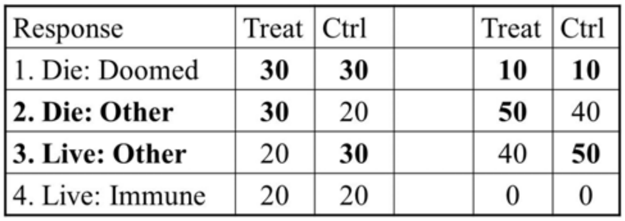
Winkler, Othmar (2009). *Interpreting Economic and Social Data: A Foundation of Descriptive Statistics*. Springer-Verlag Berlin Heidelberg. For details, see www.statlit.org/Winkler.htm

Winkler, Othmar (2011). Interpreting Socio-Economic Data. International Statistical Institute, Dublin. Copy at www.statlit.org/pdf/2011WinklerISI.pdf

Appendix A: Solving for the Sizes of the Four Groups

This model involves four equations with four unknowns. 1) P1+P2=PTdie. 2) P3+P4=PTlive. 3) P1+P3=PCdie. 4) P2+P4=PClive.

A linear system may behave in one of three ways: (a) The system has infinitely many solutions. (b) The system has a single unique solution. (c) The system has no solution. If one equation can be derived from the others, then the equations are not independent. Note that subtracting the 3rd equation from the sum of the first two gives the fourth equation. Thus, these four equations are not independent. Table 2 shows two solutions having the same observable characteristics: PTdied=60; PCdied=50.

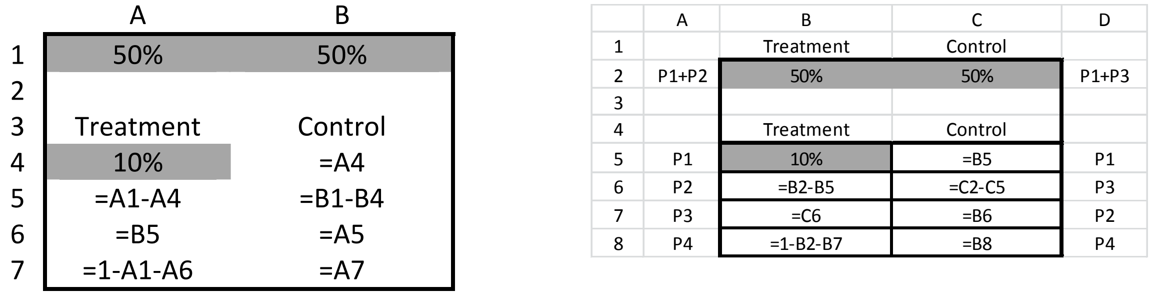


Response	Treat	Ctrl		Treat	Ctrl
1. Die: Doomed	**30**	**30**		**10**	**10**
2. Die: Other	**30**	20		**50**	40
3. Live: Other	20	**30**		40	**50**
4. Live: Immune	20	20		0	0

Table 2

Appendix B: Subgroup Distributions

Given just the died in the treatment group (P1+P2) and the died in the control group (P1+P3) is inadequate to determine P1, P2, P3 and P4. But given one additional unknown is sufficient to solve for the values as shown in Table 3.

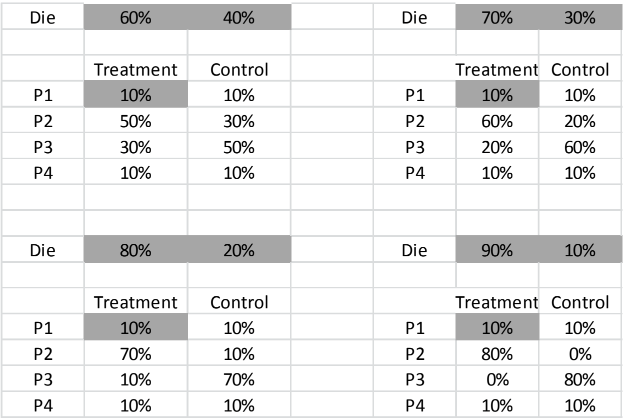


	A	B
1	50%	50%
2		
3	Treatment	Control
4	10%	=A4
5	=A1-A4	=B1-B4
6	=B5	=A5
7	=1-A1-A6	=A7

	A	B	C	D
1		Treatment	Control	
2	P1+P2	50%	50%	P1+P3
3				
4		Treatment	Control	
5	P1	10%	=B5	P1
6	P2	=B2-B5	=C2-C5	P3
7	P3	=C6	=B6	P2
8	P4	=1-B2-B7	=B8	P4

Table 3

When the percentage that died is the same in both the treatment and control groups, then P2=P3. If death is the intended outcome for the treatment, then we expect a higher percentage that died in the treatment group than in the control group. Table 4 illustrates various percentages that died in the treatment and control groups when P1 = 10%.



Die	60%	40%		Die	70%	30%
	Treatment	Control			Treatment	Control
P1	10%	10%		P1	10%	10%
P2	50%	30%		P2	60%	20%
P3	30%	50%		P3	20%	60%
P4	10%	10%		P4	10%	10%
Die	80%	20%		Die	90%	10%
	Treatment	Control			Treatment	Control
P1	10%	10%		P1	10%	10%
P2	70%	10%		P2	80%	0%
P3	10%	70%		P3	0%	80%
P4	10%	10%		P4	10%	10%

Table 4

Appendix C: Margin of Error for Relative Risks

Margin of error for relative risks is seldom – if ever – presented in an introductory statistics textbook. First, it is not readily derived. Secondly, introductory textbooks seldom present relative risk as a measure of association. But an association involving two binary variables is typically measured using relative risk, so having access to the margin-of-error formula is valuable.

	GOOD OUTCOME		
TREATED	YES	No	ALL
YES	A	B	A+B
NO	C	D	C+D
ALL	A+C	B+D	N

Table 5

PDieTreatment = A/(A+B). PDieControl = C/(C+D)

RR = PTreat/PControl. LnRR = Log_e(RR)

Var[Ln(RR)] = [(B/A)/(A+B)] + [(D/C)/(C+D)]. Assume equal-sized groups: (A+B) = (C+D) = N/2

Sources: http://en.wikipedia.org/wiki/Relative_risk and www.bioterrorism.slu.edu/bt/products/bio_epi/scripts/mod12.pdf

Var[Ln(RR)] = [(B/A) + (D/C)]/(N/2) so Std. Error = Sqrt{Var[Ln(RR)]} = Sqrt{(2/N)[(B/A) + (D/C)]}

Confidence Level = 90% to get two 5% tails. Zcutoff = NORMINV(0.95, 0, 1) = 1.64

90% Margin of Error = Zcutoff * Std. Error

Limits 90% LnRR CI: LnRR ± 90%_Margin_of_Error

Limits 90% RR CI: [Exp (LnRRLow), Exp(LnRRHigh)]

Table 6 involves a group of 150 subjects; half in the treatment group and half in the control group. If P2 = 10% and P1 = 30%, then the percentage that died in the treatment group (P1+P2) is 40%: 30 out of 75. If the relative risk is 1.333, then the percentage that died in the control group (P1+P3) is 30%: 22.5 out of 75

	Died		
Exposed	Yes	No	All
Yes	30.0	45.0	75
No	22.5	52.5	75
All	52.5	97.5	150

Table 6

Variance = (2/150)*[(45/30) + (52.5/22.5)]
= (1.5+2.33)/75 = 3.83/75 = 0.051.
Std Error = sqrt(0.051) = 0.23.
90% Margin of Error = 1.28 * 0.23 = 0.29.
Upper Ln(RR)=0.29+0.29=0.58. RR= exp(.58)=1.78
Lower LN(RR) = 0.29–0.29=0.00. RR = exp(0) = 1.00

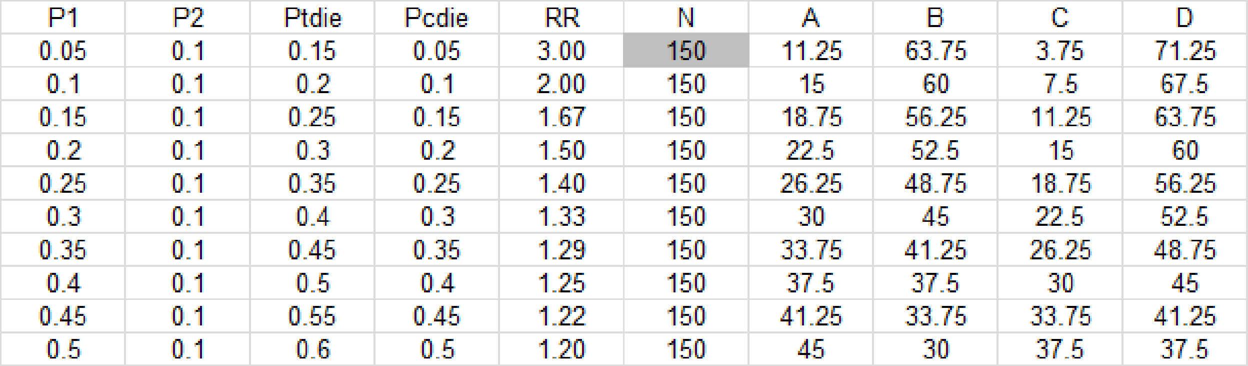
RR	1.33
Ln(RR)	0.29
Var Ln(RR)	0.05
StdErr Ln(RR)	0.23
ConfLevel	0.90
Z	1.28
MrgErr	0.29
Upper LN(RR)	0.58
Lower LN(RR)	0.00
Upper CI RR	1.78
Lower CI RR	1.00

Table 7

Table 7 illustrates the calculation of the limits of the 90% confidence interval for the relative risk in this case.

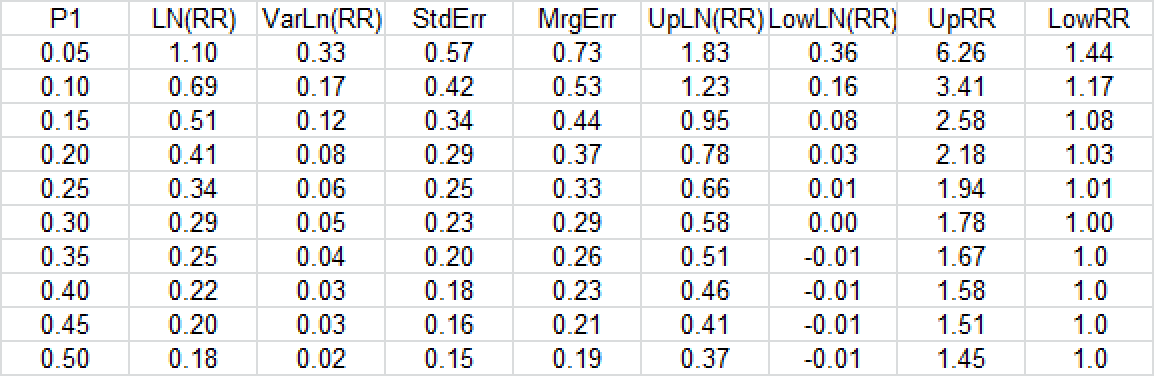
Appendix D: Data Used to Generate the Figures

Table 8 and Table 9 show the data used to generate the figures in this paper.



P1	P2	Ptdie	Pcdie	RR	N	A	B	C	D
0.05	0.1	0.15	0.05	3.00	150	11.25	63.75	3.75	71.25
0.1	0.1	0.2	0.1	2.00	150	15	60	7.5	67.5
0.15	0.1	0.25	0.15	1.67	150	18.75	56.25	11.25	63.75
0.2	0.1	0.3	0.2	1.50	150	22.5	52.5	15	60
0.25	0.1	0.35	0.25	1.40	150	26.25	48.75	18.75	56.25
0.3	0.1	0.4	0.3	1.33	150	30	45	22.5	52.5
0.35	0.1	0.45	0.35	1.29	150	33.75	41.25	26.25	48.75
0.4	0.1	0.5	0.4	1.25	150	37.5	37.5	30	45
0.45	0.1	0.55	0.45	1.22	150	41.25	33.75	33.75	41.25
0.5	0.1	0.6	0.5	1.20	150	45	30	37.5	37.5

Table 8



P1	LN(RR)	VarLn(RR)	StdErr	MrgErr	UpLN(RR)	LowLN(RR)	UpRR	LowRR
0.05	1.10	0.33	0.57	0.73	1.83	0.36	6.26	1.44
0.10	0.69	0.17	0.42	0.53	1.23	0.16	3.41	1.17
0.15	0.51	0.12	0.34	0.44	0.95	0.08	2.58	1.08
0.20	0.41	0.08	0.29	0.37	0.78	0.03	2.18	1.03
0.25	0.34	0.06	0.25	0.33	0.66	0.01	1.94	1.01
0.30	0.29	0.05	0.23	0.29	0.58	0.00	1.78	1.00
0.35	0.25	0.04	0.20	0.26	0.51	-0.01	1.67	1.0
0.40	0.22	0.03	0.18	0.23	0.46	-0.01	1.58	1.0
0.45	0.20	0.03	0.16	0.21	0.41	-0.01	1.51	1.0
0.50	0.18	0.02	0.15	0.19	0.37	-0.01	1.45	1.0

Table 9