Estimation of treatment effects

Constantin T Yiannoutsos, Ph.D.

March 22, 2018

Outline

Data analysis versus good design

Analysis from a trial data and estimation of treatment effects is the penultimate stage of the performance of a clinical study (the last being reporting of the results, which we will discuss in the next lecture).

Analysis of trial data requires a number of statistical methods and models and is considered the most important part of a study's implementation. This is because analysis appears closer to the results of the study.

However, the design of a study is much more important than the analysis of trial data and the latter cannot supplant the former.

In this lecture we discuss analytical approaches having to do with a number of contexts of clinical trials as well as context *within* a single trial (e.g., efficacy versus toxicity considerations).

Dose-finding and pharmacokinetic (PK) studies

Dose-finding studies have the following main outcomes of interest:

- Maximal tolerated dose (MTD)
- Absorption rate
- Elimination rate
- Area under the (drug concentration) curve
- Peak concentration
- Half life
- Correlation between plasma drug levels and side effects
- Proportion of patients who demonstrate evidence of efficacy

PK studies are instrumental in permitting investigators to address all of these outcomes.

A two-compartment PK model

Without going in too much detail about PK studies, we review here the basic two-compartment model.

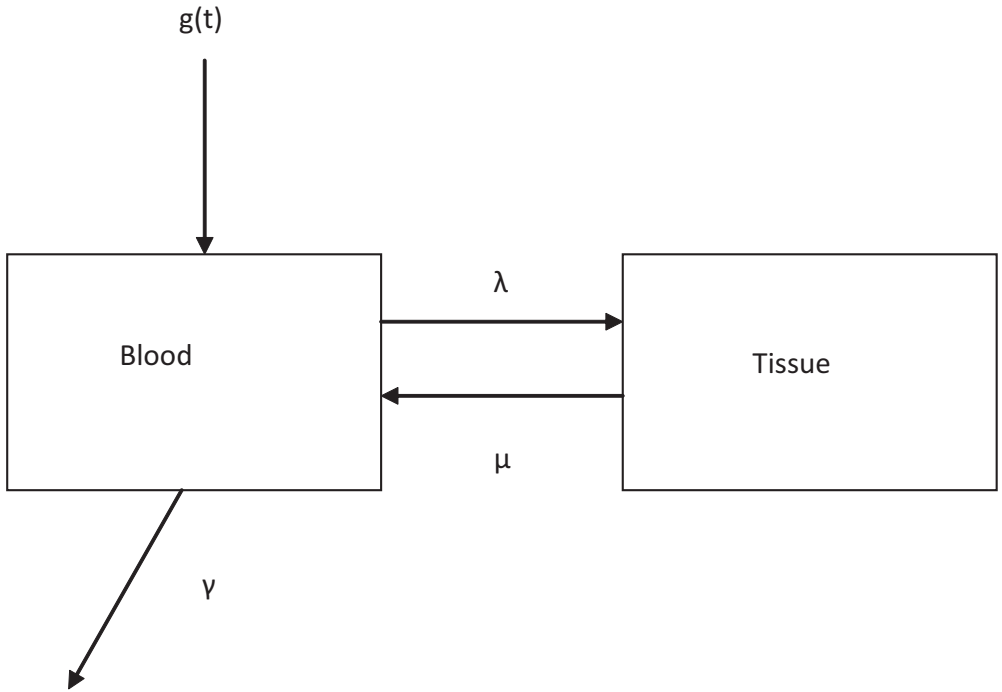

Figure 1: The basic two-compartment PK model

In this model, a drug is infused into compartment X at a rate $g(t)$. The drug is transported from compartment X (e.g., blood) to Y (e.g., tissues) at a rate λ and back to X at a rate μ and is eliminated from X at a rate γ.

Mathematical modeling of the two-compartment PK model

The mathematical analysis of the two-compartment PK model is based on a system of differential (rate) equations such as

$$\frac{dX(t)}{dt} = \underbrace{-(\lambda+\gamma)X(t)}_{\text{levels leaving compartment } X} + \underbrace{\mu Y(t) + g(t)}_{\text{levels returning to } X}$$

The solution to this system of equations is given by the following formulas:

$$X(t) = c_1(t)e^{\xi_1 t} + c_2(t)e^{\xi_2 t}$$

$$Y(t) = c_1(t)\frac{\xi_1+\lambda+\gamma}{\mu}e^{\xi_1 t} + c_2(t)\frac{\xi_2+\lambda+\gamma}{\mu}e^{\xi_2 t}$$

for appropriate ξ_1, ξ_2, and functions $c_1(t)$ and $c_2(t)$.

Mathematics of the two-compartment PK model (continued)

In the special case where the drug is infused at a constant rate $g(t) = g_0$ over time t_0 and the initial concentration in the X and Y compartment is $X(0) = 0$ and $Y(0) = 0$ we obtain two models of the concentration in the two compartments.

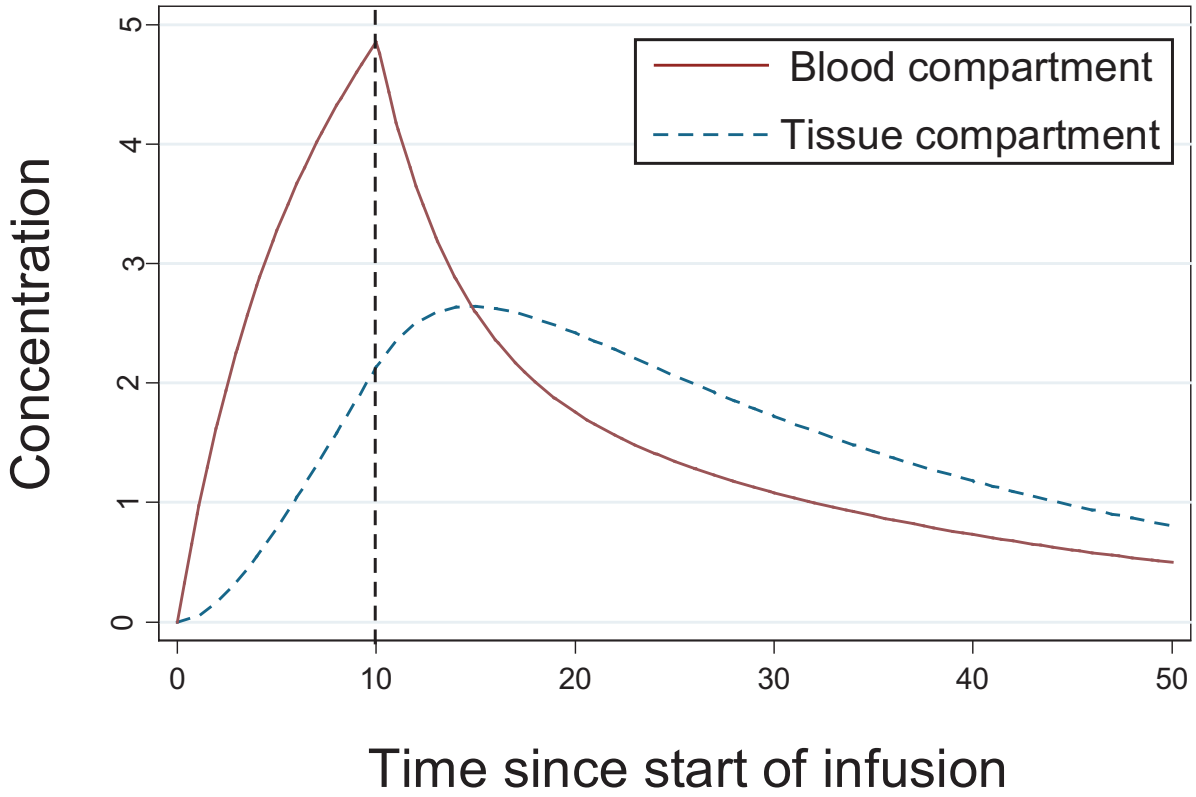

Figure 2: The basic two-compartment PK model

Area under the (drug concentration) curve (AUC)

The area under the concentration curve for compartment X is given as

$$AUC_x = \underbrace{\int_0^{t_0} X(t)dt}_{\text{drug up to } t_0 \text{ in } X} + \underbrace{\int_{t_0}^{\infty} X(t)dt}_{\text{drug after time } t_0 \text{ in } X}$$

So that,

$$AUC_x = \frac{g_0 t_0 + X(0)}{\gamma}$$

Under the special circumstances mentioned earlier, $AUC_x = \frac{g_0 t}{\gamma}$. Similarly,

$$AUC_y = \frac{\lambda g_0 t_0 + \lambda X(0)}{\mu\gamma} = \frac{\lambda}{\mu} AUC_x$$

and under the special circumstances mentioned earlier, $AUC_y = \frac{\lambda g_0 t_0}{\mu\gamma}$.

Analysis of SA studies

SA studies are concerned with both efficacy and toxicity.

Often the efficacy and toxicity outcomes are expressed in terms of dichotomous (yes/no probabilities). Often, analyzing these data involves the estimation of absolute probabilities (proportions).

Case study: Study of Alimta for thymoma

For example, consider the following two-stage cancer trial of pemetrexed (Alimta) in thymoma, a rare cancer involving the thymus. The study was designed as follows:

- *First stage*
  Eighteen patients were to be accrued at the first stage. If one or more partial or complete response (based on RECIST criteria) were observed, the study would be continued to the second stage.
- *Second stage*
  Nine more patients were to be accrued in the second stage. If four or more responses (defined above) were observed the study would be considered successful.

Design of the thymoma study
Efficacy

The desired response (alternative hypothesis) was $p_A = 0.2$ while, a response below $p_0 = 0.05$ would be considered of no interest. The above design is not optimal in the sense of Simon but has the following characteristics:

- Ensures that the probability of early termination (PET) under the alternative hypothesis (i.e., under the assumption that $p = p_A = 20\%$) is less than 2%.
- Generates power of about 80% (actually, power is 81.6%).
- The exact type I error is $< 5\%$

Design of the thymoma study
Safety

The estimate of the upper and lower limit of the toxicity is based on all evaluable patients ($n = 27$ in this study). The confidence intervals are given at the 90% level, they are two-sided and the exact binomial distribution is used instead of the normal approximation. Given these considerations the 90% confidence interval for various scenarios is as given in the following table:

Number of toxicities [1]	90% CI
0	(0.000, 0.105[2])
1	(0.002, 0.164)
2	(0.013, 0.215)
3	(0.031, 0.263)
4	(0.052, 0.308)
⋮	⋮

[1] Grade 3 or higher toxicities (grade 3: Severe AE, 4: life-threatening, 5: death related to AE)
[2] 95% one-sided upper limit)

Analysis of the thymoma study

Eighteen patients were accrued in the first stage. There were four responses observed (two partial and two complete). The study was continued and nine more patients were accrued with an additional partial response observed among the latter nine patients.

Clearly the study was successful. Now we need to figure out what the estimate and confidence interval of the response rate is.

We note that we cannot simply generate binomial confidence interval based on the final number of patients, but we should account for the fact that an interim analysis was performed.

Analysis of the thymoma study
Estimation of response

To generate confidence intervals, we use the program KSTAGE by Barry Brown.

The program essentially sums up (binomial) probabilities of all possible scenarios that can lead to the current state of affairs (Atkinson & Brown, *Biometrics*, 1985).

The output from this software is as follows:

```
Enter Number of Stages  and  Cumulative Number of  Trials for each Stage:
?
2 18 27

Enter Lo and Hi stopping values starting with stage  1:
 (-1 indicates no stopping)
?
0 -1
```

Note that we entered -1 for the upper limit (of response events) because the study will not stop regardless if the total number of responses required by the design is reached during the first stage.

Analysis of the thymoma study
Estimation of response (continued)

```
Enter  Stage Number  and  Event Number:
 (from which C.I. is calculated)
?
2 5
        K-Stage Design:

Number of Stages =  2

 Stage #    # of Trials    Cum # of Trials    Lo Quit   Hi Quit

    1           18                18            0        -1
    2            9                27

Kstg=  2  Kevt=   5

The 94% Confidence Interval is ( .0630344, .3808469)
```

Thus, the 95% confidence interval for efficacy is between 6.3% and 38%, which excludes 5% (the lower limit of efficacy). Thus, the study is a success!

Analysis of the thymoma study
Estimation of toxicity

There were 8 patients out of 27 that experienced at least one grade-3 toxicity during the study.

Disregarding the sequential nature of the study (which was not based on toxicity criteria anyway), the exact 90% binomial confidence interval for toxicity is 15.7%-47.1%.

On the other hand, no grade-4 or higher toxicities were observed, so the upper bound of the 95% one-sided confidence interval for the rate of grade-4 or higher toxicity is 10.5%.

Case study: The mesothelioma Phase II study

Rusch, Piantadosi and Holmes (*J Thorac Cardiovasc Surg. 1991*), report on a study of mesothelioma, a rare form of lung cancer associated with asbestos exposure. In that study, three approaches, biopsy, limited resection or extrapleural pneumonectomy (EPP) were attempted on 83 patients suffering from mesothelioma.

The complete data are given at

http://www.cancerbiostats.onc.jhmi.edu/Piantadosi_clinicaltrials/Software/Data%2BPrograms.zip.

The survival of patients in the three groups is given in the following table:

	age	sex	ps	hist	wtchg	surg	ptime	prog	stime	dead	X_st	X_d	X_t	X_t0
1	69	1	0	136	1	1	175	1	725	0	1	0	725	0
2	61	1	0	131	2	1	61	1	294	1	1	1	294	0
3	71	1	0	136	1	1	133	1	316	1	1	1	316	0
4	68	1	0	136	1	1	1009	1	1029	0	1	0	1029	0
5	65	0	0	NA	2	1	117	1	545	0	1	0	545	0
6	68	1	1	136	1	1	20	1	122	1	1	1	122	0
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

Case study: The mesothelioma Phase II study
Descriptive summaries

The survival status in the three groups is given in the following table:

	Result		
Group	Dead	Alive	Total
Biopsy	32	5	37
Limited	21	5	26
EPP	15	5	20

We should observe the important fact that, in each group, five patients did not die by the end of the study. Their survival was not observed fully (we know simply that they did not die by the end of the study). These are "censored" observations.

Case study: The mesothelioma Phase II study
Analysis of survival data

Survival data are unique in that not all events (deaths) are observed. We analyse these data by breaking up the time scale in intervals according to observed deaths. The probability of death is given by

$$p_i = \frac{d_i}{n_i}$$

where d_i is the number of deaths observed in that interval and n_i the number of persons that were alive at the start of the interval. The probability of surviving that interval is, therefore, $1 - p_i = 1 - \frac{d_i}{n_i}$. The probability of surviving past the time of the kth failure t_k is

$$\widehat{S(t_k)} = \prod_{i=0}^{k-1} \left(1 - \frac{d_i}{n_i}\right)$$

Case study: The mesothelioma Phase II study
Analysis of survival data (continued)

The results of the analysis are given in the following table:

Event Time t_i	Beg. Total	Number Fail	Number Lost	Failure Probability	Std. Error
4	82	1	0	0.9878	0.0121
6	81	1	0	0.9756	0.0170
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
475	28	1	0	0.3293	0.0519
499	27	0	1	0.3293	0.0519
503	26	1	0	0.3166	0.0514
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
1265	2	1	0	0.0585	0.0502
1338	1	0	1	0.0585	0.0502

Case study: The mesothelioma Phase II study
Analysis of survival data (continued)

From the previous table we see how survival probability estimates are generated.

Starting with 100% probability of survival at time $t = 0$, and excluding the individual with zero survival, we drop to $1 - \frac{1}{82} = 0.9878$ after the first failure at time $t_1 = 4$ days.

The next failure occurs at time $t_2 = 6$ at which point, the probability of survival is $p_2 = \left(1 - \frac{1}{82}\right)\left(1 - \frac{1}{81}\right) = 0.9756$.

By contrast, when an observation is censored (at time $t = 499$ days, the probability of survival through that interval is 100% so there is no difference in the probability for that subject. We see that the probability remains the same as that of the 55th failure at time $t_{55} = 475$ days, i.e., $p_{55} = 0.3293$.

Case study: The mesothelioma Phase II study
The Kaplan-Meier estimator of survival

We plot the estimate of survival over time by drawing a horizontal line between successive failures and a vertical line of length d_i/n_i at each event i. Censored observations are ignored. This produces the so called "Kaplan-Meier" estimate of survival.

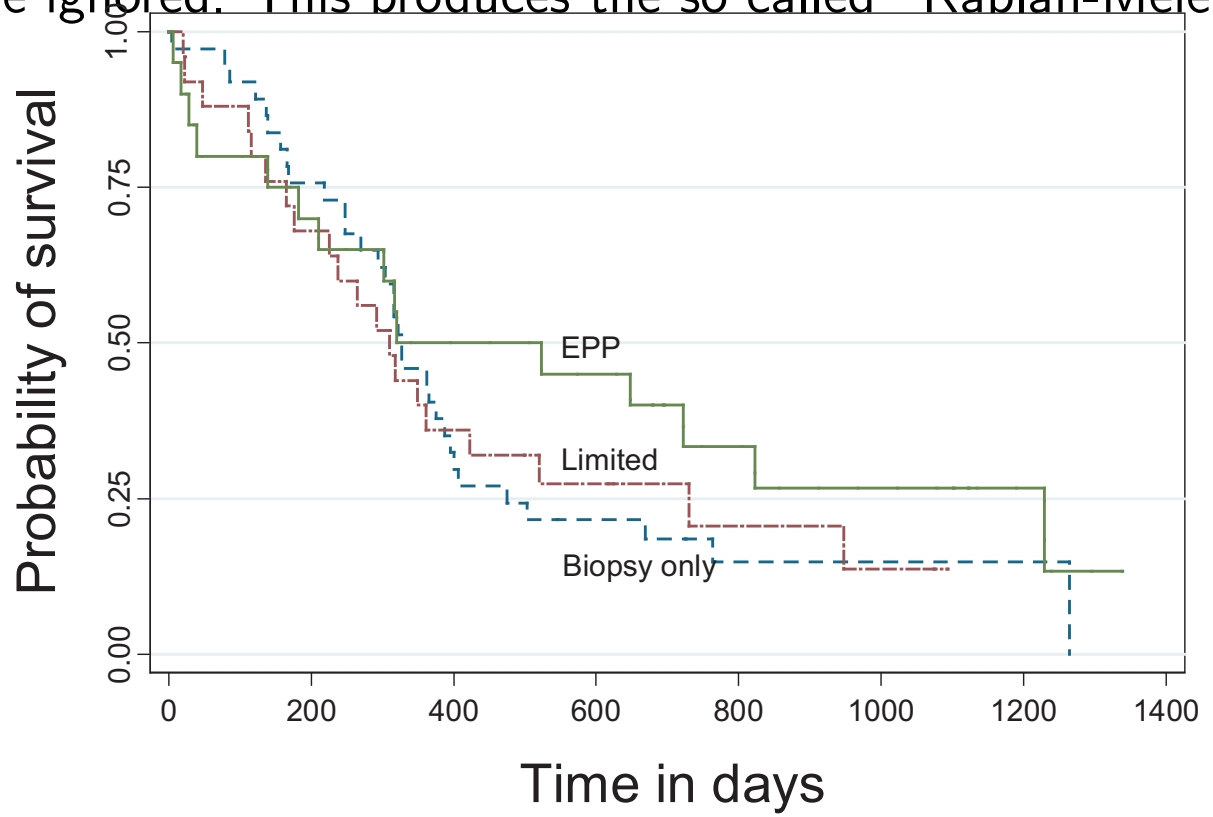

Figure 3: Kaplan-Meier plot of the mesothelioma data

Case study: The mesothelioma Phase II study
Comparisons between groups

Summaries of the survival experience in the three groups are given in the following Table:

surg	time at risk	incidence rate[3]	no. of subjects	Survival time 25%	50%	75%
biopsy	15042	.0021274	37	218	327	475
limited	9678	.0020665	25	165	310	730
EPP	10441	.0014366	20	139	320	1229
total	35161	.0019055	82	168	320	722

For example, the median survival in the three groups is 327, 310 and 320 days respectively.

[3] Equals, number of deaths divided by time at risk. For example, in the biopsy group this is $32/15042 \approx 0.00212$.

Case study: The mesothelioma Phase II study
Comparisons between groups:The log-rank test

To compare the survival in the three groups, we consider the so-called "log-rank" test. This test is based on the inherent ordering of the deaths by the time they occurred. At each death, we can construct a 3×2 table. The table will look as follows for the first failure that occurred at $t = 4$ days in the biopsy group[4]:

	Result		
Group	Dead	Alive	Total
Biopsy	1	36	37
Limited	0	25	25
EPP	0	20	20

After generating these tables, we perform a Mantel-Haenszel test of association between surgical group and survival status. This measures whether, on average, the proportion of deaths falls inordinately on one or more of the three groups.

[4] Note that there is a failure at time $t = 0$ in the limited-resection group that is being ignored)

Case study: The mesothelioma Phase II study
The log-rank test (continued)

Carrying out the log-rank test analysis we obtain the following output:

```
Call:
survdiff(formula = Surv(stime, dead) ~ group, data = mesoth)

                N Observed Expected (O-E)^2/E (O-E)^2/V
group=biopsy   37       32     30.2     0.107     0.195
group=limited  26       21     18.5     0.346     0.483
group=EPP      20       15     19.3     0.969     1.401

Chisq= 1.5  on 2 degrees of freedom, p= 0.479
```

The p value of the log-rank test is 0.479 suggesting that there is no difference in survival among the three surgical procedures.

Resampling methods

A very powerful methodology to generate distributions of various statistics is through resampling. The "bootstrap" as it's called, involves generating repeated analyses by resampling out of the dataset with replacement.

We can run a bootstrap analysis of the previous survival analysis to obtain the distribution. The median and the associated 95% confidence interval, based on the normal distribution, is $\widehat{S(0.5)} = 320$ days and (292–387) days respectively. The bootstrap estimate of the median and the 95% confidence interval is 320 days and (276.1–363.9) days respectively.

The bootstrap, in this case, merely validated a known distributional result. The true power of the bootstrap is that it can generate similar distributions more difficult to calculate (e.g., the difference between two median survivals).

Comparative efficacy trials (Phase III)

While developmental studies such as DF and SA studies use mainly descriptive means to present the treatment effects, comparative trials describe data, quantify possible treatment differences and assess extraneous influence.

The usual approach is a test of statistical significance, i.e., determining to what extent the observed differences are attributable to random variation.

Case study: FAP prevention study

The following is the Familial Adenomatous Polyposis (FAP) dataset (Giardiello et al., *NEJM*, 1993):

id	sex	age	polyp number at month 0	polyp number at month 12	polyp size at month 0	polyp size at month 12	rx
1	0	17	7	–	3.6	–	1
2	0	20	77	–	3.8	–	0
3	1	16	7	4	5.0	1.0	1
4	0	18	5	26	3.4	2.1	0
5	1	22	23	16	3.0	1.2	1
6	0	13	35	40	4.2	4.1	0
7	0	23	11	14	2.2	3.3	1
8	1	34	12	16	2.0	3.0	0
9	1	50	7	11	4.2	2.5	0
10	1	19	318	434	4.8	4.4	0
11	1	17	160	26	5.5	3.5	1
12	0	23	8	7	1.7	0.8	1
13	1	22	20	45	2.5	3.0	0
14	1	30	11	32	2.3	2.7	0
15	1	27	24	80	2.4	2.7	0
16	1	23	34	34	3.0	4.2	1
17	0	22	54	38	4.0	2.9	0
18	1	13	16	–	1.8	–	1
21	1	34	30	57	3.2	3.7	0
22	0	23	10	7	3.0	1.1	1
23	0	22	20	1	4.0	4.0	1
24	1	42	12	8	2.8	1.0	1

Case study: FAP prevention study
Comparisons of month-12 polyp number and size

The baseline (month-0) and month-12 number and size of polyps in each treatment arm is shown in the following Table:

	Treatment arm		
	Treatment 0	Treatment 1	
Time point	Mean (± SD)	Mean (± SD)	p-value[5]
Number of polyps			
Month 0	53.9 (90.2)	28 (44.5)	0.403
Month 12	77.9 (126.7)	13 (10.8)	0.145
Polyp size			
Month 0	3.3 (0.93)	3.2 (1.22)	0.816
Month 12	3.1 (0.73)	1.8 (1.41)	0.022

[1] T test

This suggests that, while there is no significant reduction in the number of polyps in month 12, there might be a reduction of their size due to therapy (active treatment=1, standard treatment=0).

Case study: FAP prevention study
Using differences from baseline

Instead of comparing the month-12 number of polyps or polyp size, we can compare the *difference* between month-12 and month-0 in the number and size of the polyps. The revised analysis is given in the following Table:

	Treatment arm		
	Treatment 0	Treatment 1	
Time point	Mean ($\pm$ SD)	Mean ($\pm$ SD)	p-value
Number of polyps			
Month 12 difference	26.3 (36.9)	-18.7 (43.7)	0.026
Polyp size			
Month 12 difference	-0.2 (0.90)	-1.5 (1.8)	0.053

This analysis shows how much the variability of the measures under comparison has been reduced by removing the biological effect, which is the largest component of the variability. Now, there is both a reduction in the number and, possibly, in the size of the polyps associated with active treatment.

Case study: FAP prevention study
Analysis of covariance (ANCOVA) analysis

Another way to do this analysis is to adjust for the baseline number or size of the polyps. This involves a model for each subject i as follows:

$$Y_{12} = \underbrace{\beta_0}_{\text{intercept}} + \underbrace{\beta_1 Y_{0i}}_{\text{baseline quantity}} + \underbrace{\beta_2 T_i}_{\text{treatment effect}} + \underbrace{\epsilon_i}_{\text{error term}}$$

The results are given in the following table:

Dependent	Model Terms	Parameter Estimate	Standard Error	p-value
Number of polyps	β_0	21.5	14.2	–
	β_1	1.1	0.1	$<$0.0001
	β_2	-43.2	18.9	0.037
Polyp size	β_0	1.13	0.90	–
	β_1	0.21	0.1	0.403
	β_2	-1.29	0.51	0.023

The ANCOVA analysis shows that the size *and* number of polyps are significantly lowered in relation to the active treatment.

Case study: NSCLC lung cancer trial

Lad, Rubinstein, Sadeghi, et al. *J Clin Onc*, 1988) report a randomized trial of CAP (a combination of cytoxan, doxorubicin and platinum chemotherapy as adjuvant treatment to radiotherapy in non-small-cell lung cancer (NSCLC).

The data are given in the following list:

	celltype	karn	t	n	treat	surv	dead	dfs	event	age	race	elig	wtloss	sex	survyear
1	1	2	1	0	1	1046	1	413	1	70	1	0	1	1	2.8657530
2	1	2	2	2	0	342	1	342	0	67	1	0	1	1	0.9369863
3	1	2	2	2	1	54	1	18	1	61	1	0	0	1	0.1479452
4	1	2	2	2	0	303	1	264	1	52	1	0	0	1	0.8301370
5	1	2	1	2	1	295	1	248	1	59	0	0	0	1	0.8082192
6	2	1	3	2	1	88	1	59	1	39	0	0	0	0	0.2410959
7	2	2	2	2	1	241	1	241	0	46	0	0	0	0	0.6602740
8	2	2	1	2	1	567	1	252	1	44	0	0	0	1	1.5534250
9	2	2	2	2	0	286	1	211	1	38	0	0	1	0	0.7835616
10	2	2	2	1	0	265	1	262	1	62	1	0	0	1	0.7260274

Case study: NSCLC lung cancer trial: textitKaplan-Meier analysis

The Kaplan Meier plot is given in the following Figure:

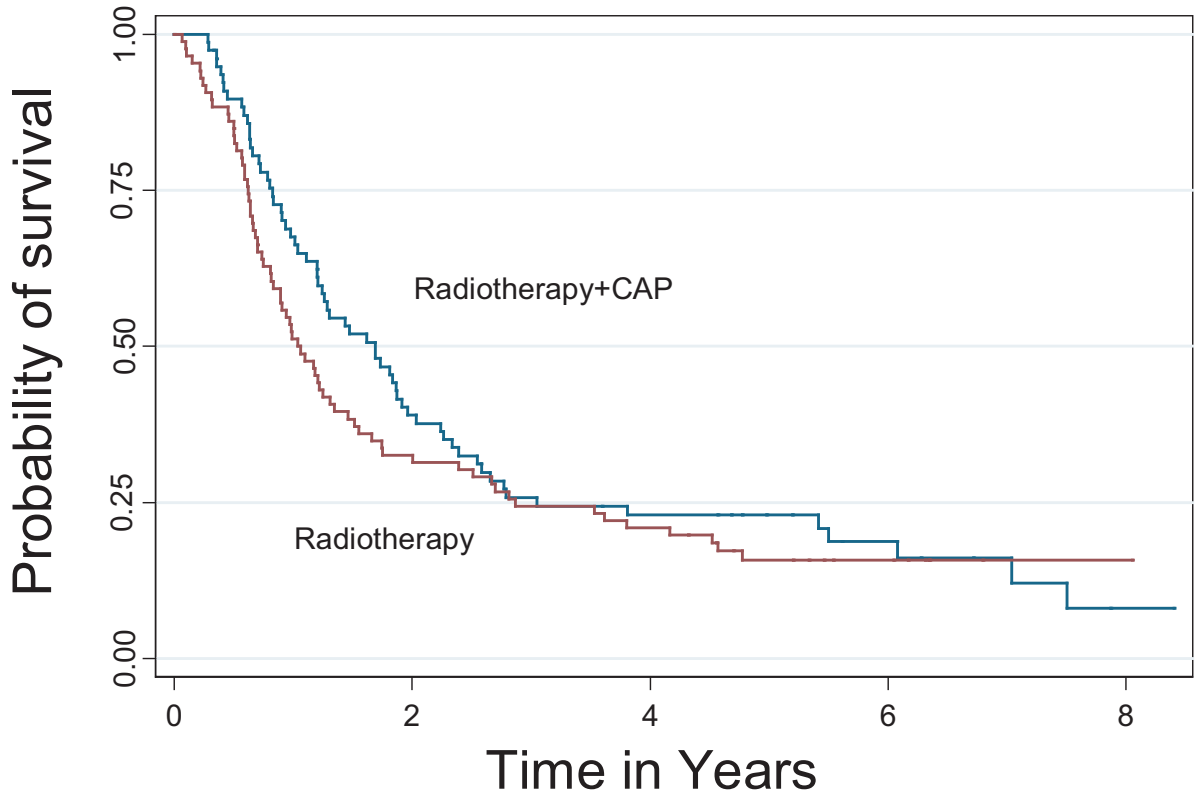

Figure 4: Survival by treatment group in the lung-cancer trial

Case study: Lung cancer trial
The Cox proportional hazards model

An approach to assess the effect of various factors on survival is through the Cox proportional hazards model. This model asserts that the *hazard* of death dependent on a number of predictors $\mathbf{X}$ is given by

$$\lambda(t; \mathbf{X} = \lambda_0(t)e^{\beta \mathbf{X}}$$

in other words, the predictor effect is multiplicative and is constant over time. This model is called proportional because the hazard in various subject subgroup is proportion over time. The implication of this model is that

$$\log \left\{ \frac{\lambda(t)}{\lambda_0(t)} \right\} = \beta_1 X_1 + \beta_2 X_1 + \cdots$$

so this is a linear regression model on the log hazard.

Case study: Lung cancer trial
Analysis via the Cox proportional hazards model

The analysis of the Cox model is given in the following table:

Factor	Haz. Ratio	Std. Err.	z	$P > \|z\|$	[95% Conf.	Interval]
treat="2"	1.30616	0.233470	1.49	0.135	0.920127	1.854151
cell type="2"	1.31154	0.241337	1.47	0.141	0.914435	1.881098
t="2"	0.91412	0.247078	-0.33	0.740	0.538186	1.552655
t="3"	1.16275	0.362530	0.48	0.629	0.631092	2.142298
n="1"	0.95181	0.356827	-0.13	0.895	0.456500	1.984541
n="2"	1.26627	0.448955	0.67	0.506	0.632026	2.536994
age	1.00383	0.010218	0.38	0.707	0.984000	1.024057
sex	1.06313	0.222285	0.29	0.770	0.705692	1.601625
weight loss	1.10107	0.346521	0.31	0.760	0.594196	2.040323
race	1.28824	0.356660	0.91	0.360	0.748749	2.216457

For example, the hazard among subjects in treatment 2 (radiotherapy) is 30.6% higher than treatment 1 (radiotherapy + CAP), or radiotherapy+CAP reduces the hazard to 76% ($\approx 1/1.31$) regardless of the length of the survival. This is not a statistically significant difference.

A parenthesis
P-values do not measure evidence

The lack of a significant p-value (i.e., $p < 0.05$) for the treatment effect, adjusted for the other factors in the model, may suggest to some that a larger sample size might produce a significant p value.

What many people miss is that p values do not quantify the strength of the evidence (here this is whether radiotherapy plus CAP reduces risk of death compared to radiotherapy alone).

P values simply assess the extent of type-I error.

A parenthesis
P-values don't measure evidence (cont'd)

A related pitfall is to consider factors as "more significant" if they are associated with a lower p value.

For example consider the following two 2×2 tables below:

	A	$\bar{A}$		A	$\bar{A}$
B	1	7		1	6
$\bar{B}$	13	7		13	6

Both column proportions are the same but the Fisher's exact test p values associated with the two tables are 0.033 and 0.026 respectively.

Even though the first table has more information, one might consider the second one as "more significant" based on the p values.

Factorial designs

Factorial designs are used to test the effect of more than one treatment and uses a design that permits the assessment of interaction between them.

A typical 2×2 factorial design comparing the effect of treatment A and treatment B (assuming that A&B can be given in combination) is as follows:

Treatment A	Treatment B		
	No	Yes	Total
No	n	n	$2n$
Yes	n	n	$2n$
Total	$2n$	$2n$	$4n$

Effect estimation in factorial designs

The treatment effects in a typical 2×2 factorial design comparing the effect of treatment A and treatment B (assuming that A&B can be given in combination) is as follows:

Treatment A	Treatment B: No	Treatment B: Yes
No	$\bar{Y}_o$	$\bar{Y}_B$
Yes	$\bar{Y}_A$	$\bar{Y}_{AB}$

Interaction effects

Interaction effect between factors A and B is the modification of the effect of factor A by factor B. This, in the context of two drugs is either

- *Synergism*
  A positive (synergistic or potentiated) interaction between the two drugs (i.e., a larger additive effect than would be expected by adding the individual effects of the two drugs)
- *Antagonism*
  A negative (antagonistic) interaction between two drugs (i.e., a smaller additive effect than would be expected by adding the individual effects of the two drugs)

Efficiency of factorial designs

In the absence of interaction between factor A and B, the estimates of the effect of these two factors can be averaged from the estimates of their treatment effect versus placebo. This is

$$\beta_A = \frac{(\bar{Y}_A - \bar{Y}_0) + (\bar{Y}_{AB} - \bar{Y}_B)}{2}$$

for factor A and,

$$\beta_B = \frac{(\bar{Y}_B - \bar{Y}_0) + (\bar{Y}_{AB} - \bar{Y}_A)}{2}$$

for factor B.

Efficiency of factorial designs (continued)

The efficiency of the factorial design becomes obvious if one considers that, if the variance of the patient response is σ^2 and is the same in all treatment groups, then the variance of β_A (similarly for β_B) is

$$\begin{aligned} \mathrm{var}(\beta_A) &= \mathrm{var}\frac{(\bar{Y}_A - \bar{Y}_0) + (\bar{Y}_{AB} - \bar{Y}_B)}{2} \\ &= \frac{1}{4}\frac{4\sigma^2}{n} \\ &= \frac{\sigma^2}{n} \end{aligned}$$

Now, considering the variance of the treatment effect $\mathrm{var}(\beta'_A) = \mathrm{var}(\bar{Y}_A - \bar{Y}_o) = \frac{2\sigma^2}{n}$, so the variance of the factorial design is equal to a two-arm trial with $2n$ patients.

Testing of interactions

Factorial designs are the only designs where interactions between factor A and B can be measured. The definition of an interaction is that the effect of A is different in the presence versus absence of B.

This can be estimated by comparing with zero

$$\beta_{AB} = (\bar{Y}_A - \bar{Y}_0) + (\bar{Y}_{AB} - \bar{Y}_B)$$

We note that the variance of β_{AB} is

$$\text{var}(\beta_{AB}) = \frac{4\sigma^2}{n}$$

which is four times larger than the variance of the individual effects when there is no interaction. To get the same precision for estimating the interaction effect we need four times the sample size. This shows that estimation of the main and interaction effects cannot be met simultaneously in the same factorial study.

Example: Evaluation of combination therapy for hypertension

We consider the following example[2] Various combinations of two anti-hypertensive drugs, four doses of an ACE inhibitor (drug A) and three doses of a diuretic (drug B) were considered.

The results and sample sizes are given in the following table:

Table 1: Table of the antihypertension study (Hung, 2000)

	Sample size				Mean mmHg in SiDBP			
	A0	A1	A2	A3	A0	A1	A2	A3
B0	75	75	74	48	0	1.4	2.7	4.6
B1	74	75	74	49	1.8	2.8	5.7	8.2
B2	48	50	48	48	2.8	4.5	7.2	10.9

[2] Hung HMJ. Evaluation of a combination drug with multiple doses in unbalanced factorial design clinical trials. *Stat Med*, 2000; **19**: 2079–2087.

Analysis of the hypertension 2×2 factorial design

We consider data simulated from the sample sizes and means listed in Table 1 (with common variance $s = 7.07$ mmHg). We focus here on the 2×2 factorial of the $A0$, $A1$, $B0$ and $B1$ (top left-hand corner of the table). A box plot of the four drug combinations is given in Figure 5.

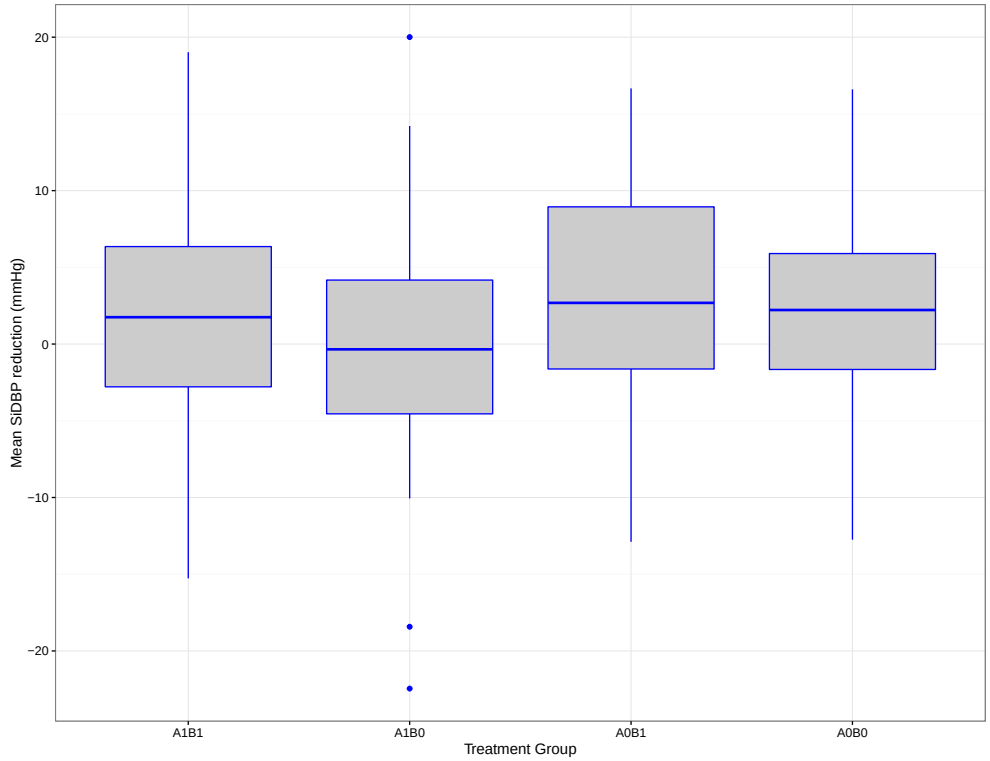

Figure 5: Box plot of the four treatment combinations in the 2×2 factorial design

Looking for interactions

The power of the 2×2 factorial design is in the fact that it combines both groups with the effect $B1$ (i.e., $A1B1$, $A0B1$) and the two groups without the effect $B1$ (i.e., $A1B0$, $A0B0$). This works only when there is no interaction effect. We can assess the presence of interaction graphically:

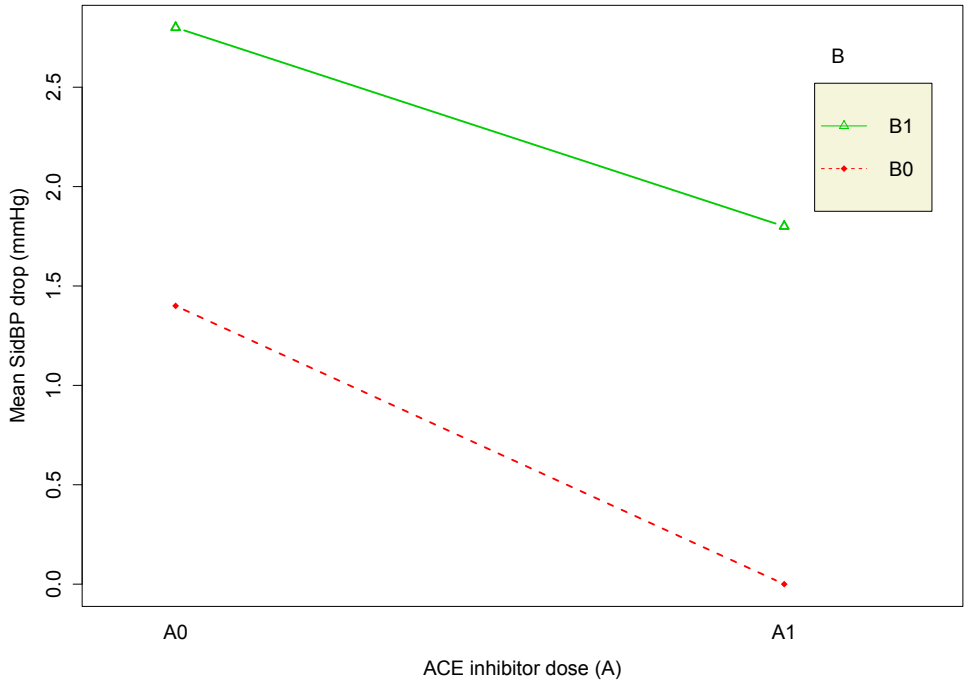

Figure 6: Interaction plot of A versus B

It does not appear that a significant interaction effect is present.

Analysis of the 2×2 factorial design

We can run the 2×2 factorial design without interaction as follows:

```
Call:
lm(formula = y ~ A + B, data = hungab)

Residuals:
     Min       1Q   Median       3Q      Max
-22.5523  -4.4318   0.2277   4.6367  19.9064

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.3003     0.7063   1.841   0.0666 .
AA1          -1.2007     0.8164  -1.471   0.1425
BB1           1.5993     0.8164   1.959   0.0511 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 7.059 on 296 degrees of freedom
Multiple R-squared:  0.01993,Adjusted R-squared:  0.01331
F-statistic:  3.01 on 2 and 296 DF,  p-value: 0.05082
```

Analysis of the 2×2 factorial design with interaction

We can run the 2×2 factorial design with interaction as follows:

```
Call:
lm(formula = y ~ A * B, data = hungab)

Residuals:
    Min      1Q  Median      3Q     Max
-22.453  -4.433   0.128   4.658  20.006

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.4000     0.8164   1.715   0.0874 .
AA1          -1.4000     1.1545  -1.213   0.2262
BB1           1.4000     1.1545   1.213   0.2262
AA1:BB1       0.4000     1.6355   0.245   0.8070
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 7.07 on 295 degrees of freedom
Multiple R-squared:  0.02013,Adjusted R-squared:  0.01016
F-statistic:  2.02 on 3 and 295 DF,  p-value: 0.1112
```

Analysis of the four-way treatment comparison

We can run this a four-way comparison where each treatment combination is considered separately.

```
Call:
lm(formula = y ~ AB, data = hungab)

Residuals:
    Min      1Q  Median      3Q     Max
-22.453  -4.433   0.128   4.658  20.006

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.8000     0.8219   2.190   0.0293 *
ABA1B0       -1.8000     1.1584  -1.554   0.1213
ABA0B1        1.0000     1.1584   0.863   0.3887
ABA0B0       -0.4000     1.1584  -0.345   0.7301
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 7.07 on 295 degrees of freedom
Multiple R-squared:  0.02013,Adjusted R-squared:  0.01016
F-statistic:  2.02 on 3 and 295 DF,  p-value: 0.1112
```

Comments

1. Note that the effect of B (the diuretic) is almost statistically significant at the 5% level and statistically significant at the 10% level in the no-interaction model.
2. The interaction model is not statistically significant. Neither is the straight four-way treatment model.

Equivalence of the interaction and four-way model

The interaction model is *identical* to the four-way comparison model (note the overall F statistics in the two models). The p-values associated with the various effects are not the same, since the two models are parametrized differently: The interaction model is

$$E(y_i) = \beta_0 + \beta_1 X_A + \beta_2 X_B + \gamma X_A X_B = \begin{cases} E(y_{A_0B_0}) = \beta_0 \\ E(y_{A_1B_0}) = \beta_0 + \beta_1 \\ E(y_{A_0B_1}) = \beta_0 + \beta_2 \\ E(y_{A_1B_1}) = \beta_0 + \beta_1 + \beta_2 + \gamma \end{cases}$$

where $X_A = 1$ is equivalent to $A1$, $X_B = 1$ with $B1$ and, respectively, zero values denote $A0$ and $B0$. The four-way design is

$$E(y_{ij}) = \zeta + \eta + \theta + \kappa = \begin{cases} E(y_{A_0B_0}) = \zeta + \kappa \\ E(y_{A_1B_0}) = \zeta + \eta \\ E(y_{A_0B_1}) = \zeta + \theta \\ E(y_{A_1B_1}) = \zeta \end{cases}$$

so that (see previous output), $\beta_0 = \zeta + \kappa$, $\beta_0 + \beta_1 = \zeta + \eta$, etc.

Crossover designs

In contrast to the designs where participants are treated with a single or combination treatment *concurrently*, in crossover designs, treatments are administered *sequentially*. The main advantage of this study is that treatment effects can be compared within the same subjects (thus eliminating within-subject or biological variability).

Crossover designs are different in objective and scope from trials that give treatments sequentially (e.g., $A \rightarrow B \rightarrow C$ versus $A \rightarrow B$) but assess the incremental effect of a treatment (here treatment C) or from factorial designs where patients are administered a combination of treatments *simultaneously*.

Efficiency of crossover studies

To see why crossover designs are efficient, we consider that each subject is its own control. Thus, a potentially large component of treatment effect variability is removed from the estimation.

To see this, consider the variance of the difference in treatment effects A and B, $\hat{\Delta}_{AB}$, noted here as $\bar{Y}_A$ and $\bar{Y}_B$ respectively, will be:

$$\begin{aligned} \text{var}(\hat{\Delta}_{AB}) &= \frac{\sigma^2}{n} + \frac{\sigma^2}{n} - 2\text{cov}(\bar{Y}_A, \bar{Y}_B) \\ &= 2\frac{\sigma^2}{n}(1 - \rho_{AB}) \end{aligned}$$

We note that, in comparative studies, $\rho_{AB} = 0$ because the groups receiving the treatments A and B are independent.

So $\text{var}(\hat{\Delta}_{AB}) = 2\sigma^2/n$. However, if $\rho_{AB} > 0$, as it is usually expected when treatment is administered to the same patient, the variance of the effect difference will be smaller in a crossover than a comparative study.

Example of cross-over trial design: Pronethalol for angina pectoris.

The drug pronethalol was tested in angina pectoris. A cross-over design was used where patients were randomized to receive placebo or pronethalol. The number of angina episodes while receiving one or the other treatment were counted. The data can be stored in long format (i.e. one row per patient) or a wide format (i.e., one row per patient, where each column stores the response to pronethalol or placebo. Here, the data (stored in text file `angina.txt` are in the wide format:

```
> # 5. Example of a Cross-over trial for angina pectoris
> angina <- read.table("C:/Clinical trials/R labs/lab1/data/angina.txt",
+                      header=TRUE, quote="\"")
> head(angina)
  Patient Placebo Pronethalol Difference
1       1      71          29         42
2       2     323         348         25
3       3       8           1          7
4       4      14           7          7
5       5      23          16          7
6       6      34          25          9
```

Exploratory data analysis

We can check the normality assumption for the differences of angina attacks or for the actual measurements on the two groups.

```
> # (a) Check normality assumption for the difference in attacks of angina
> library(lattice)
> pdf("hist_diff.pdf",height = 5.5,width = 5.5)
> histogram( ~ Difference,data = angina)
> dev.off()
RStudioGD
        2
```

Histogram of number of angina episodes

This produces the following pictorial representation:

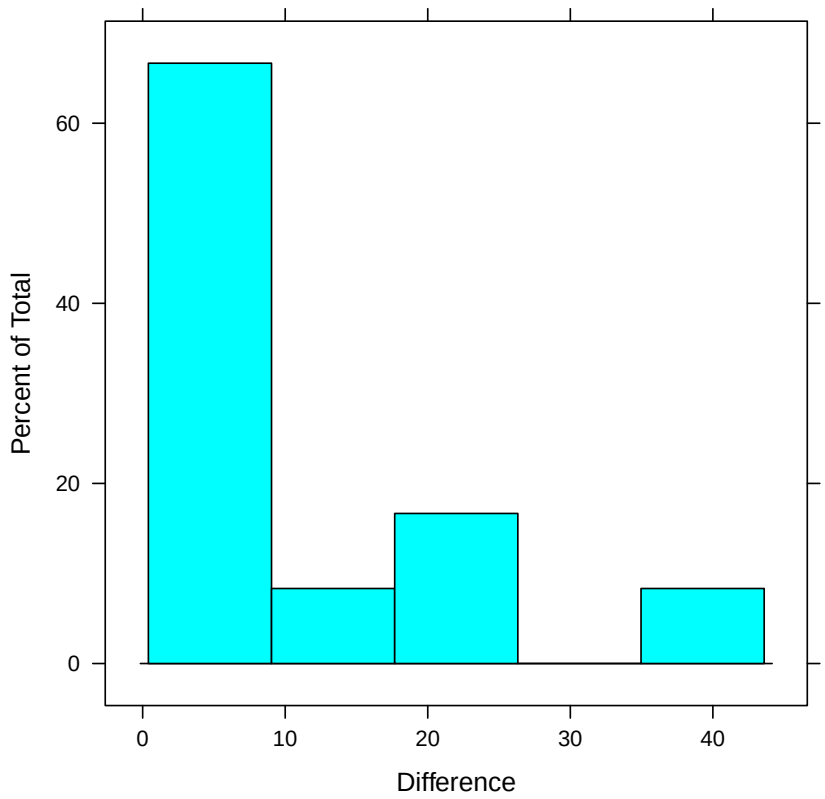

Figure 7: Histograms for the difference in angina pectoris episodes between pronethanol and placebo groups.

Data manipulation

We turn the data in the long format (multiple lines per patient) as follows:

	attacks	trt
1	29	Pronethalol
2	348	Pronethalol
3	1	Pronethalol
4	7	Pronethalol
5	16	Pronethalol
6	25	Pronethalol
7	65	Pronethalol
8	41	Pronethalol
9	0	Pronethalol
.	.	.
.	.	.
.	.	.
21	2	Placebo
22	3	Placebo
23	17	Placebo
24	7	Placebo

Histogram of raw numbers of angina episodes by treatment group

The previous data set allows a by treatment histogram of the number of episodes:

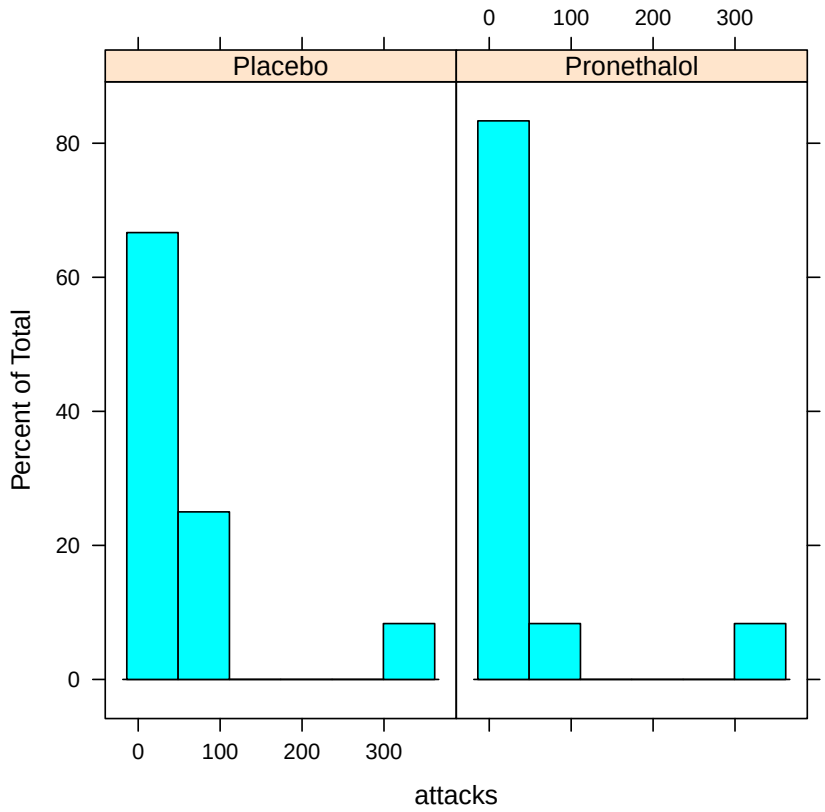

Figure 8: Histograms of the number of angina pectoris episodes between pronethanol and placebo groups.

Checking the normality assumption

We can also check the normality of the distribution of the differences. The Q-Q plot of the differences in the number of angina episodes is given in the following Figure:

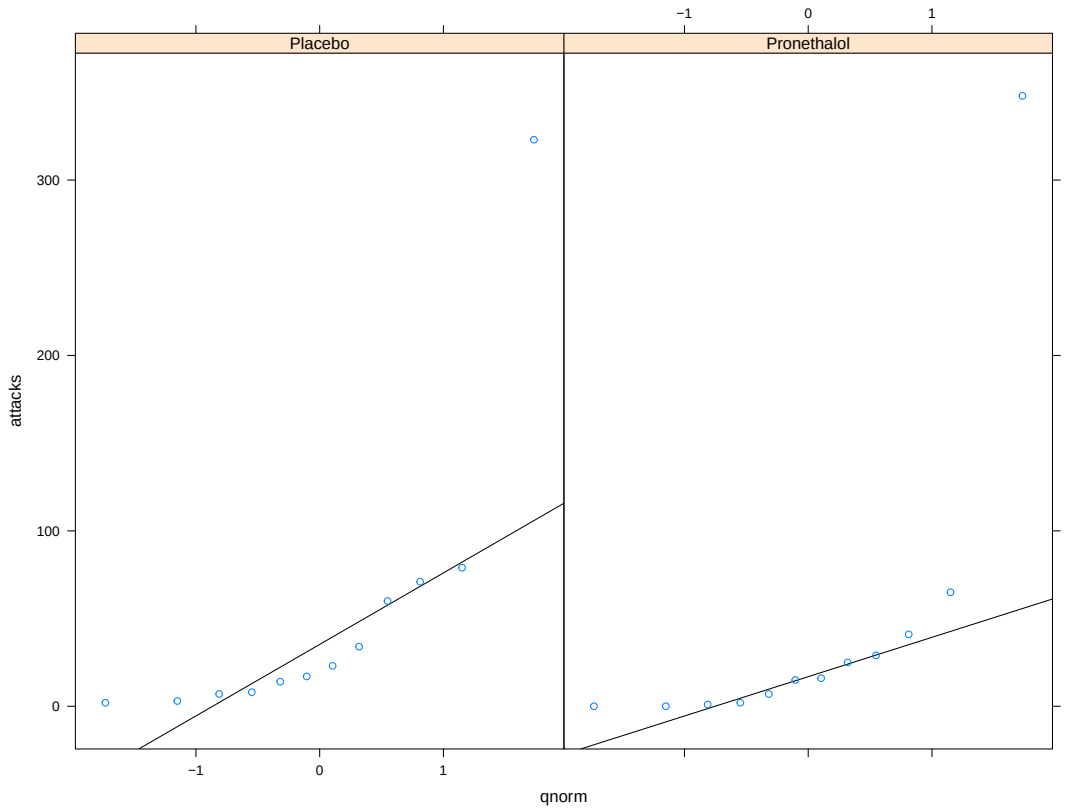

Figure 9: Q-Q plot of the number of angina pectoris episodes between pronethanol and placebo groups.

More checks

... and the Q-Q plot of the differences as well:

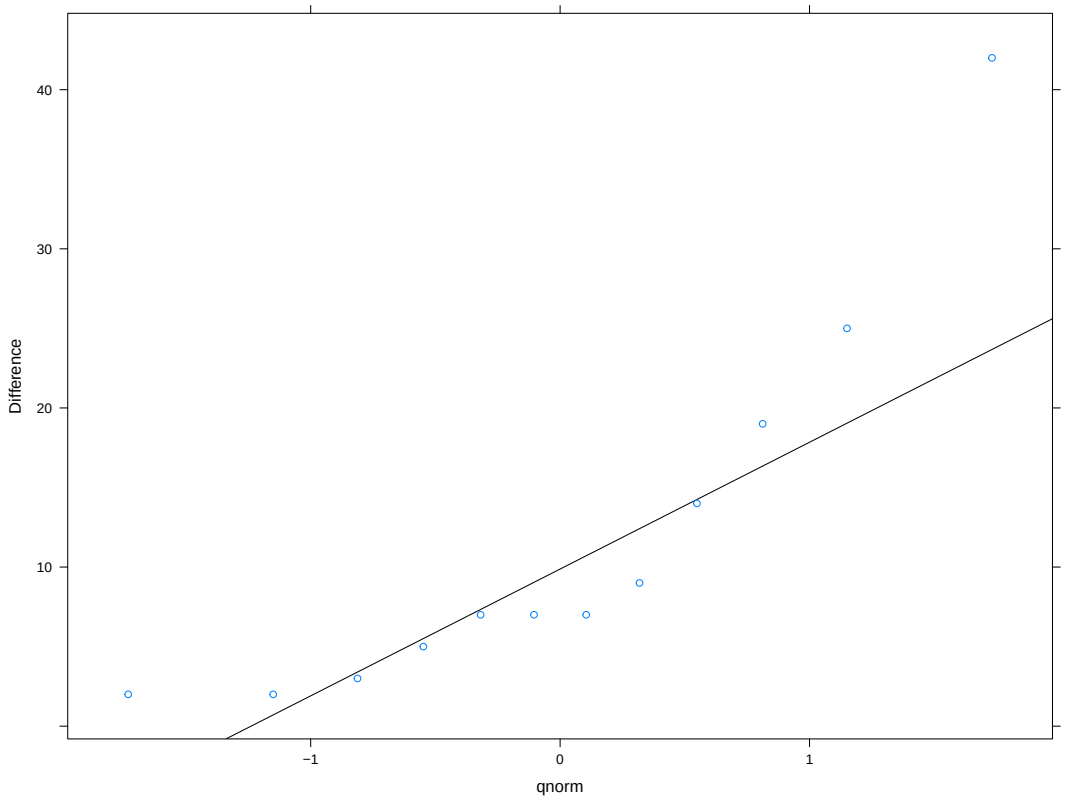

Figure 10: Q-Q plot of the differences in the number of angina pectoris episodes between pronethanol and placebo groups.

The distribution is far from normal so we go on with nonparametric tests.

Data analysis
The wrong analysis

First, we ignore the study design and assume that the data come from two independent populations. To do so, we have to use the data frame which includes two rows per patient. Then, we perform a two-sample Wilcoxon test

```
Wilcoxon rank sum test with continuity correction

data:  attacks by trt
W = 88, p-value = 0.3705
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(x = c(71L, 323L, 8L, 14L, 23L, 34L, 79L,  :
  cannot compute exact p-value with ties
```

This test suggests that there is no significant difference between the two groups with respect to the number of attacks of angina.

Data analysis
The right analysis

Let's see what happens if we take into account the matched study design. We are to use the paired version of Wilcoxon test (i.e. a nonparametric equivalent of Student's t tests)

```
Wilcoxon signed rank test with continuity correction

data:  angina$Placebo and angina$Pronethalol
V = 67, p-value = 0.03066
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(angina$Placebo, angina$Pronethalol, paired = T) :
  cannot compute exact p-value with ties
```

Now the treatment effect becomes significant!

Other analyses

The spectrum of analysis of data generated from clinical trials is extensive. Some additional analyses used are:

- *Longitudinal analyses*
  These are analyses involving repeated measurements on the same subjects.
- *Time-dependent covariates*
  While many predictors are fixed at baseline and are assumed to have a constant effect over time, time-updated factors attempt to model factors that change over time.
- *Measurement error*
  While most analyses assume that all factors are measured exactly, a number of analyses have been introduced that allow for some error in the measurement of covariate predictors.

Other analyses (continued)

- *Random versus fixed effects*
  Most statistical models assume that the effects of predictors are fixed (non-random).

  For example, in assessing the effect of institution in a multi-center study, a fixed-effect analysis considers the institutions participating as fixed, while a random-effect analysis considers these as a random sample from all possible institutions.

  This has also been applied in longitudinal models that increase the flexibility of the statistical model by allowing different slopes or intercepts for each subject thus more realistically modeling response.