6. Statistical Analysis Using Baseline Measurements

6.1 Baseline Data in Clinical Trials

In most clinical trials data is collected on the characteristics of patients in addition to the outcome measures. As well as recording demographic data such as age and sex, information will be collected regarding the clinical status of the patient at the time of entry into the trial, which could include values of the trial outcome measures at entry into the trial. For example in a trial comparing treatments for osteoarthritis of the knee, one might record information regarding pain, physical impairment or psychological distress, on entry into the trial . Such data may be required to confirm that patients satisfy the inclusion criteria for the trial. It is also used to describe the characteristic of patients entering the trial. Standard practices would be to present a table summarizing the characteristics for each treatment group.

Data collected prior to randomisation are called *baseline* data. This data can also be used in the estimation and testing hypotheses regarding the treatment effect. As we shall see, for just one outcome measure, there are several ways in which this can be done. If these are all carried out and the investigator allowed to choose on the basis of the results, it is likely that the most favourable will be presented. Alternatively, all could will be presented, which could be a problem, if they give conflicting results. Either way, this could distort the published report and would be a source of *statistical analysis* bias. To prevent this the choice of

analysis should not be based on the results of the analyses of the trial, but need to be documented in advance in a statistical analysis plan. To do this we require criteria to make the decision in advance as to which method of analysis should be used .

Ex 6.1 The FAP Trial Data

FAP is a genetic defect that predisposes those affected to develop large numbers of polyps in the colon that are prone to become malignant. In this trial patients with FAP were randomly allocated to receive a drug therapy (sulindac) or a placebo.

Patient ID	Treatment Group	Polyp Size	
		Baseline (X)	12 Months (Y)
1	sulindac	5.0	1.0
2	placebo	3.4	2.1
3	sulindac	3.0	1.2
4	placebo	4.2	4.1
5	sulindac	2.2	3.3
6	placebo	2.0	3.0
7	placebo	4.2	2.5
8	placebo	4.8	4.4
9	sulindac	5.5	3.5
10	sulindac	1.7	0.8
11	placebo	2.5	3.0
12	placebo	2.3	2.7
13	placebo	2.4	2.7
14	sulindac	3.0	4.2
15	placebo	4.0	2.9
16	placebo	3.2	3.7
17	sulindac	3.0	1.1
18	sulindac	4.0	0.4
19	sulindac	2.8	1.0

Piantadosi S. Clinical Trials: A methodological Perspective p302 , Wiley 1997

6.2 Possible Treatment Effect Estimators

(i) Unadjusted

Suppose the random variable Y_i represents the continuous outcome for the i^{th} patient in either the new treatment group (T) or the control group (C), and suppose

$$Y_i = \mu_U + \varepsilon_i \qquad \text{for } i \in C$$

$$Y_i = \mu_U + \tau_U + \varepsilon_i \qquad \text{for } i \in T$$

with ε a random variable with $E[\varepsilon_i \mid i \in T] = E[\varepsilon_i \mid i \in C] = 0$.

$$\tau_U = E[Y_i \mid i \in T] - E[Y_i \mid i \in C] ,$$

which can be estimated by

$$\hat{\tau}_U = \overline{Y}_T - \overline{Y}_C$$

(ii) Changes Scores

Suppose X_i is the value of outcome measure Y_i recorded at baseline. Medical researchers sometimes calculate the change from baseline, $C_i = Y_i - X_i$, which is call the *change score*.

Treatments are compared using C_i instead of Y_i .

$$C_i = Y_i - X_i = \mu_C + \varepsilon_i' \qquad \text{for } i \in C$$

$$C_i = Y_i - X_i = \mu_C + \tau_C + \varepsilon_i' \qquad \text{for } i \in T$$

with ε' a random variable with $E[\varepsilon_i' \mid i \in T] = E[\varepsilon_i' \mid i \in C] = 0$.

$$\tau_C = E[C_i \mid i \in T] - E[C_i \mid i \in C] ,$$

which can be estimated by

$$\hat{\tau}_C = \bar{C}_T - \bar{C}_C = \left(\bar{Y}_T - \bar{X}_T\right) - \left(\bar{Y}_C - \bar{X}_C\right)$$

where $\bar{C}_T$ and $\bar{C}_C$ are the sample means of the change score for each group.

For both these methods of analysis statistical inference can be based on the two-sample t-test and the associated confidence interval.

Figure 6.1 STATA Output for FAP trial

(i) Unadjusted Analysis

```
Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
Sulindac |       9    1.833333    .4711098    1.413329    .7469522    2.919714
 Placebo |      10    3.110000    .2306753    .7294595    2.588176    3.631824
---------+--------------------------------------------------------------------
    diff |           -1.276667    .5245527                .130485    2.422848
------------------------------------------------------------------------------
    diff = mean(Sulindac) - mean(Placebo)                         t =  -2.4338
Ho: diff = 0                     Satterthwaite's degrees of freedom =  11.6981

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0160         Pr(|T| > |t|) = 0.0320          Pr(T > t) =0.9840
```

(ii) Change Score Analysis

```
Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
Sulindac |       9   -1.522222    .5969314    1.790794   -2.898749   -.1456959
 Placebo |      10   -.1900000    .2857544    .9036346   -.8364213    .4564213
---------+--------------------------------------------------------------------
    diff |           -1.332222    .6618026               -.1160118    2.780456
------------------------------------------------------------------------------
    diff = mean(Sulindac) - mean(Placebo)                         t =  -2.0130
Ho: diff = 0                     Satterthwaite's degrees of freedom =  11.5476

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0340         Pr(|T| > |t|) = 0.0680          Pr(T > t) =0.9660
```

Satterthwaite's Test is an unequal variance "t-test" This was discussed on page 33-34 Sect 2.4

Note: Adjusted degrees of freedom have been used in both analyses. This method, called the Satterthwaite test and mention in 2.2, adjusts inference for variance of the two arms being unequal.

(iii) Analysis Adjusted by Baseline Variables

Suppose that X is a baseline variable. Suppose for some values θ we define an adjusted treatment effect as

$$A_i = Y_i - \theta X_i = \mu_A + \varepsilon_i'' \qquad \text{for } i \in C$$

$$A_i = Y_i - \theta X_i = \mu_A + \tau(\theta) + \varepsilon_i'' \qquad \text{for } i \in T$$

with ε_i'' a random variable with $E\left[\varepsilon_i'' \mid i \in T\right] = E\left[\varepsilon_i'' \mid i \in C\right] = 0$.

Taking expectations,

$$\tau(\theta) = E\left[Y_i - \theta X_i \mid i \in T\right] - E\left[Y_i - \theta X_i \mid i \in C\right],$$

which can be estimated by

$$\hat{\tau}(\theta) = \bar{A}_T - \bar{A}_C = \left(\bar{Y}_T - \theta \bar{X}_T\right) - \left(\bar{Y}_C - \theta \bar{X}_C\right)$$

- For simplicity it will be assumed that X is a single variable, but it could be a vector of covariates.

- Baseline variables can be binary or continuous. If X is a binary variable, it is generally convenient to code it with 0 and 1. A variable coded in this way is sometimes called an *indicator* or *dummy* variable.

6.3 Comparison of Adjusted and Unadjusted Analyses

In a randomised controlled trial $E\left[\hat{\tau}(\theta)\right]=\tau(\theta)$ is independent of θ. Hence the expected values of unadjusted, change and an adjusted estimate of the treatment effect are all equal, that is

$$E\left[\hat{\tau}_U\right]=E\left[\hat{\tau}_C\right]=E\left[\hat{\tau}(\theta)\right]$$

Proof

Considering $\hat{\tau}(\theta)=\left(\bar{Y}_T-\theta\bar{X}_T\right)-\left(\bar{Y}_C-\theta\bar{X}_C\right)$.

$$E\left[\hat{\tau}(\theta)\right]=E\left[\bar{Y}_T\right]-E\left[\bar{Y}_C\right]-\theta\left(E\left[\bar{X}_T\right]-E\left[\bar{X}_C\right]\right).$$

Randomisation means that $E\left[\bar{X}_T\right]=E\left[\bar{X}_C\right]$.

Therefore $E\left[\hat{\tau}(\theta)\right]=E\left[\bar{Y}_T\right]-E\left[\bar{Y}_C\right]$, which is independent of θ.

Values of θ equal to 0 and 1 correspond to the treatment effect in an unadjusted $(\hat{\tau}_U)$, and change $(\hat{\tau}_C)$ giving the required result ■

Suppose $\sigma_X^2,\sigma_Y^2,\sigma_{XY}$ are the variances and covariance of X and Y with $\lambda=\sqrt{\frac{1}{n_T}+\frac{1}{n_C}}$. The treatment effect $\hat{\tau}_A$ has a minimum variance when $\theta=\beta$ equal to $Var\left[\hat{\tau}(\beta)\right]=\lambda^2\sigma_Y^2\left(1-\rho^2\right)$ where β is the regression coefficient of *Y* on *X* and ρ is the correlation between *X* and *Y* conditional on treatment group.

Proof

Again consider $\hat{\tau}(\theta)=\left(\bar{Y}_T-\theta\bar{X}_T\right)-\left(\bar{Y}_C-\theta\bar{X}_C\right)$

$$Var\left[\hat{\tau}(\theta)\right]=Var\left[\bar{Y}_T-\bar{Y}_C-\theta\left(\bar{X}_T-\bar{X}_C\right)\right]$$

$$=Var\left[\bar{Y}_T-\bar{Y}_C\right]+Var\left[\theta\left(\bar{X}_T-\bar{X}_C\right)\right]-2Cov\left[\bar{Y}_T-\bar{Y}_C,\theta\left(\bar{X}_T-\bar{X}_C\right)\right]$$

$$=Var\left[\bar{Y}_T-\bar{Y}_C\right]+\theta^2.Var\left[\bar{X}_T-\bar{X}_C\right]-2\theta.Cov\left[\bar{Y}_T-\bar{Y}_C,\bar{X}_T-\bar{X}_C\right]$$ **[1]**

Considering the first term

$$Var\left[\bar{Y}_T-\bar{Y}_C\right]=Var\left[\bar{Y}_T\right]+Var\left[\bar{Y}_C\right]-2Cov\left[\bar{Y}_T,\bar{Y}_C\right].$$

Since treatment groups are independent, $Cov\left[\bar{Y}_T,\bar{Y}_C\right]=0$.

Therefore $Var\left[\bar{Y}_T-\bar{Y}_C\right]=Var\left[\bar{Y}_T\right]+Var\left[\bar{Y}_C\right]$.

Since observations are independent, $Var\left[\bar{Y}_T\right]=\frac{\sigma_Y^2}{n_T}$ and

$Var\left[\bar{Y}_C\right]=\frac{\sigma_Y^2}{n_C}$.

Therefore $Var\left[\bar{Y}_T-\bar{Y}_C\right]=\lambda^2\sigma_Y^2$ where $\lambda=\sqrt{\frac{1}{n_T}+\frac{1}{n_C}}$.

Similarly $Var\left[\bar{X}_T - \bar{X}_C\right] = \lambda^2\sigma_X^2$ and $Cov\left[\bar{Y}_T - \bar{Y}_C, \bar{X}_T - \bar{X}_C\right] = \lambda^2\sigma_{XY}$.

Substitution into **[1]** gives $Var\left[\hat{\tau}(\theta)\right] = \lambda^2\left(\sigma_y^2 + \theta^2\sigma_X^2 - 2\theta\sigma_{XY}\right)$ **[2]**

A minima can be found by differentiation with respect to θ.

$$\frac{\partial}{\partial\theta}Var\left[\hat{\tau}(\theta)\right] = \lambda^2\left(2\theta\sigma_X^2 - 2\sigma_{XY}\right).$$

This equals zero when $\theta = \sigma_{XY}/\sigma_X^2$, which is the coefficient for regression of *Y* on *X* within each treatment group.

The second derivative $\frac{\partial^2}{\partial\theta^2}Var\left[\hat{\tau}(\theta)\right] = 2\lambda^2\sigma_X^2$.

As this is positive, it follows that this is a minimum. The treatment effect estimate with minimum variance is therefore

$$Var\left[\hat{\tau}(\beta)\right] = \lambda^2\left(\sigma_Y^2 + \beta^2\sigma_X^2 - 2\beta\sigma_{XY}\right) = \lambda^2\sigma_Y^2\left(1 - \frac{\sigma_{XY}^2}{\sigma_X^2\sigma_Y^2}\right)$$

Since $\sigma_{XY}/\sqrt{\sigma_X^2\sigma_y^2} = \rho$, $Var\left[\hat{\tau}(\beta)\right] = \lambda^2\sigma_Y^2\left(1-\rho^2\right)$ as required ■

Estimation of β and $\tau(\beta)$

The treatment effect $\tau(\beta)$ can be estimated by fitting a linear model, which is a generalization of linear regression. The general form of a linear model with *k* covariates is

$$Y_i = \mu + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i.$$

where the *X*'s are the *k* covariates and the β's the corresponding *k* coefficients. The random variable ε_i is usually assumed to be $N\left[0, \sigma_{\varepsilon^*}^2\right]$. If one of the *X*'s is an indicator variable, the coefficient β is the difference in the mean value of *Y* for *X* =1 as compared to *X*=0 , adjusted for other *X*'s.

When considering the analysis of data from a randomised trial it is notational clearer to separate the matrix of covariates into an indicator variable I_i equal to 1 for treatment *T* and 0 for the control *C*, and a matrix *X* of other covariates. The model is then written as

$$Y_i = \mu_A + \tau I_i + \beta X_i + \varepsilon_i$$

and the treatment effect is the coefficient of the indicator variable I_i. Statistical inference is simply the test of whether the coefficient of the indicator variable I_i differs from zero, that is $H_0 : \tau = 0$ *vs* $H_1 : \tau \neq 0$. The matrix of coefficients, β, for other variable is generally of less interest and often not given in published reports of trials.

Fig 6.2 STATA Output for FAP trial

Linear Model Analysis Adjusting for Baseline Polyp Size

```
      Source |       SS       df       MS              Number of obs =      19
-------------+------------------------------           F(  2,    16) =    3.48
       Model |  8.63531123     2  4.31765562           Prob > F      =  0.0556
    Residual |  19.8541618    16  1.24088511           R-squared     =  0.3031
-------------+------------------------------           Adj R-squared =  0.2160
       Total |  28.4894731    18   1.5827485           Root MSE      =   1.114

------------------------------------------------------------------------------
     size12 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
      size0 |   .2087081   .243071     0.86   0.403    -.3065794    .7239956
  treatment |  -1.288262   .5120029   -2.52   0.023    -2.373659   -.2028639
      _cons |   2.421263   .8760753    2.76   0.014     .5640666    4.27846
------------------------------------------------------------------------------
```

Fitted Model

size12 = 0·209×size0 − 1·288×treatment + 2·421

Table 6.2 summarizes the treatment effect and inference for all three analyses. The null hypothesis of no treatment effect would not have been rejected at a 5% level if the change score analysis had been carried out.

Table 6.2 Summary of treatment effect estimates for the FAP trial

	Treatment Effect	*SE*	*95% Lower*	*C.I Upper*	*p-value*
Unadjusted $\hat{\tau}_U$	-1.28	0.52	-2.42	-0.13	0.032
Change $\hat{\tau}_C$	-1.33	0.66	-2.78	0.12	0.068
Linear Model $\hat{\tau}(\beta)$	-1.29	0.51	-2.37	-0.20	0.023

From page 80

Note that the SE for the linear model treatment is the smallest, which is what is expected from the result on page 83.

Comparison of standard errors of Unadjusted, Change Score and Linear Model Analyses

From above $Var\left[\hat{\tau}(\beta)\right] = \lambda^2\sigma_Y^2\left(1-\rho^2\right)$, which is a quadratic in ρ.

From [2] $Var\left[\hat{\tau}_C\right] = \lambda^2\left(\sigma_Y^2+\sigma_X^2-2\sigma_{XY}\right) = \lambda^2\left(\sigma_Y^2+\sigma_X^2-2\sigma_Y\sigma_X\rho\right)$,

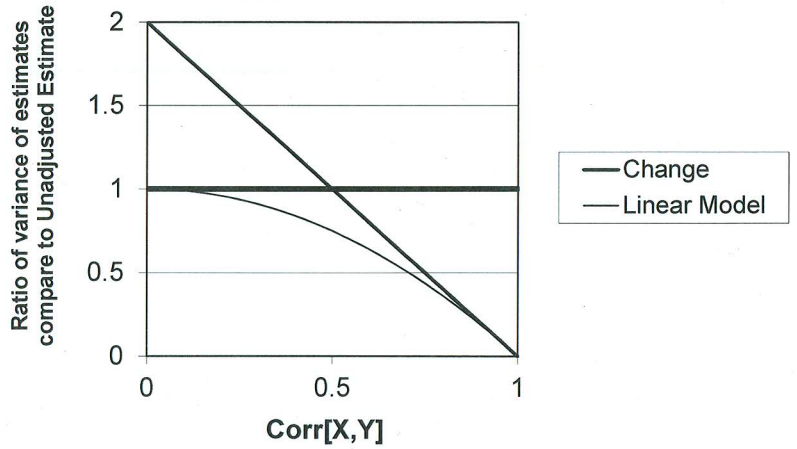
which is a linear function of ρ.

Assuming $\sigma_X^2 = \sigma_Y^2$, $Var\left[\hat{\tau}_C\right] = 2\lambda^2\sigma_Y^2(1-\rho)$. ← Straight line as a function of ρ

The unadjusted standard error is simply $Var\left[\hat{\tau}_U\right] = \lambda^2\sigma_Y^2$

Hence $\dfrac{Var\left[\hat{\tau}(\beta)\right]}{Var\left[\hat{\tau}_U\right]} = 1-\rho^2$ and $\dfrac{Var\left[\hat{\tau}_C\right]}{Var\left[\hat{\tau}_U\right]} = 2(1-\rho)$.

Figure 6.3 comparison of Change Score and Linear models with the unadjusted analysis assuming $\sigma_X^2 = \sigma_Y^2$.



Summary Analyses using Baseline Data

All three estimates of the treatment effect defined in (6.2) are unbiased, but an estimate of the treatment effect based on a linear model has smaller expected variance, where baseline covariates correlate with the outcome measure and it does not matter whether this correlation is positive or negative.

Reducing the variance of the treatment effect estimate is important as this increases the precision of the estimate, thereby giving greater power for a given sample size. As a consequence, if a baseline variable is thought to predict outcome, an analysis adjusting for this variable is recommended. Where an outcome measure is recorded at baseline, then it is usually a strong predictor of outcome, and the variable should be used as a covariate.

To prevent the analysis bias, a single set of baseline covariates should be selected prior to starting analysis. This should be recorded in the statistical analysis plan of the trial. This choice will therefore need to be based on prior knowledge or reasoning as to what variables are likely to predict outcome irrespective of which treatment is received.

6.4 A Flawed Analysis using *Within Group Change from Baseline*

A statistical analysis sometime seen in the medical literature is to carry out a separate paired t-test on each treatment groups. Treatments are then compared by using the results of the separate statistical tests. If improvement in one group is statistically significant but not the other, it is concluded that one treatment is more effective than the other. This analysis is illustrated below with the FAP data

Figure 6.4 STATA Output for paired t-test analysis of each treatment

Results for: polyp.mtw (treat = 0)

```
Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
  size12 |      10        3.11    .2306753    .7294595    2.588176    3.631824
   size0 |      10         3.3    .3076795     .972968    2.603981    3.996019
---------+--------------------------------------------------------------------
    diff |      10        -.19    .2857544    .9036346   -.8364213    .4564213
------------------------------------------------------------------------------
     mean(diff) = mean(size12 - size0)                             t =  -0.6649
 Ho: mean(diff) = 0                         degrees of freedom =        9

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.2614         Pr(|T| > |t|) = 0.5228          Pr(T > t) = 0.7386
```

Results for: polyp.mtw (treat = 1)

```
Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
  size12 |       9    1.833333    .4711098    1.413329    .7469522    2.919714
   size0 |       9    3.355556    .4163703    1.249111    2.395404    4.315707
---------+--------------------------------------------------------------------
    diff |       9   -1.522222    .5969314    1.790794   -2.898749   -.1456959
------------------------------------------------------------------------------
     mean(diff) = mean(size12 - size0)                             t =  -2.5501
 Ho: mean(diff) = 0                         degrees of freedom =        8

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.0171         Pr(|T| > |t|) = 0.0342          Pr(T > t) = 0.9829
```

Why the Analysis using Within-Group Changes is Flawed

The main reason why this method is flawed is because the two p-values relate to two separate hypotheses test and so do not directly test the benefit of one treatment as compared to the other, that is they do not compare the two potential outcomes.

Use of this type of analysis also suggests other misunderstandings.

- Failure to reject the null hypothesis, for a treatment does not imply that there is no change. The absolute change within each treatment groups could be the same but unequal variances may affect the probability of rejecting the null hypothesis for one treatment and not another.

- Tests of within group change are often statistically significant, but change within a treatment group may not be due to treatment. It may occur because the condition naturally resolves. They may tell us more about the natural history of the condition than the benefit of receiving treatment one treatment as compared to another.

Unfortunately, clinical researchers often carry out this type of analysis, when the statistical analysis directly comparing the two treatments is not statistically significant. This is done in the desperate search for a statistically significant result to report.