THE ANALYSIS OF MULTIPLE ENDPOINTS IN CLINICAL TRIALS

Lars Frison, AB Hässle

INTRODUCTION

Treatment comparisons in clinical trials usually involve several endpoints such that conventional significance testing can seriously inflate the overall type I error rate.

For instance in a study comparing blood pressure lowering agents, the outcome could be assessed for both diastolic blood pressure and systolic blood pressure in different positions like supine, sitting and standing, giving six endpoints for that study.

One way to avoid the multiple testing problem is to select a single primary endpoint for formal statistical inference, all the other endpoints would then be given exploratory rather than formal interpretation.

This approach is often useful when analysing clinical trials, but it happens quite often that one is equally interested in several of the variables and want to use them all for the formal statistical inference.

It might for instance be the case that the medical question that one wants an answer to in a study is of a complicated nature and that several variables are needed together to give a satisfactory answer, how do one proceed in such a case?

One method often used is to apply the Bonferroni correction, that is multiply each p-value with the total number of endpoints. But as will be shown later this is conservative when the endpoints are correlated.

Another approach when analysing several normally distributed endpoints is to use Hotelling's $T^2$. However this test is intended to detect any departure from the null hypothesis and hence lacks power to detect any specific types of departure that was considered apriori to be biologically plausible.

In a clinical trial, that case would normally be improvement for all the primary endpoints.

A third approach when there is a need for a single objective probability statement, to answer a medical question, is to derive an appropriate global test statistic.

The relative merits of one such test statistic, applicable to any set of asymptotically normal test statistics, proposed by O'Brien (1984 in Biometrics) will be described later along with a non-parametric test statistic of the same kind.
THE BONFERRONI CORRECTION

The Bonferroni inequality can be used to obtain an adjustment to the smallest p-value for significance tests on k endpoints.

If the k endpoints are independent then is

\[ P(\text{smallest p-value} \leq \alpha) = 1 - (1 - \alpha)^k \approx \alpha^k \], if \( \alpha \) is small.

Hence the Bonferroni correction has each p-value multiplied by k, the number of endpoints.

That is for an overall Type 1 error rate one accepts as statistically significant only those p-values less than \( \alpha/k \).

In practice endpoints are correlated, so that the Bonferroni correction becomes conservative.

With perfect correlation will a test with nominal level \( \alpha \) have a true size \( \alpha/k \).

However it has been shown that as long as the correlations are below 0.5 will the conservatism be quite small.

The main drawback of the Bonferroni correction is that it confines attention to the smallest p-value of k statistics.

Thus five endpoints with p-values of .01 .5 .5 .5 .5 are considered more highly significant than five endpoints all at .02, whereas the latter of course appears to contain more convincing evidence of a treatment difference.

Thus the Bonferroni correction has its greatest power for alternative hypothesis in which only one of k endpoints has a non-zero treatment difference, and furthermore, one does not know in advance which endpoint that will be - a situation unlikely to arise in practice.

For alternative hypothesis, in which several variables depart from zero treatment difference in the same direction, the Bonferroni correction will seriously lack power.
A GLOBAL TEST STATISTIC

Now I will describe a parametric generalized least squares test statistic, that was first proposed by O'Brien in 1984, and extended to cope not only with quantitative endpoints but also with binary and survival endpoints by Pocock, Geller and Tsiatis in 1987.

The main advantage of this procedure is that the power is directed towards departures in which some improvement is demonstrated consistently among the various endpoints.

This global generalized least squares test statistic which provides the best linear unbiased estimates and corresponding optimal tests, uses data in common units via a transformation to relative deviates, by subtracting the overall variable mean from each observation and dividing by the pooled within-group sample standard deviation.

If the correlation matrix $\Sigma$ is known and the vector $z$ follows a multivariate normal distribution, then, under $H_0$, will this statistic follow a standard normal distribution.

$$GLS = \frac{J' \Sigma^{-1} z}{\left( J' \Sigma^{-1} J \right)^{1/2}}$$

The $J'$ is a vector of $(1, \ldots, 1)$, as many as there are endpoints. The weighting factor in the numerator is simply the column sums of $\Sigma^{-1}$.

In the denominator $J' \Sigma^{-1} J$ is the sum of all the cells in $\Sigma$. If we have equicorrelation, that is the correlation is the same between all pairs of variables, then the formula can be written:

$$GLS = \frac{\sum_{i,k} \bar{Z}_{ik}}{\left( \sum_{i,k} \bar{Z}_{ik} \right)^{1/2}}$$

and if for instance $k=5$ and $r=0$, $\bar{Z}$ does only have to be .88 for achieving significance at the 5% level, if $k=5$ and $r=.5$ then $\bar{Z}$ must be at least 1.52.

So the value that $\bar{Z}$ has to achieve decreases with $k$ and increases with $r$, and the combined evidence of several endpoints in the same direction need not be as extreme as for a single endpoint.

When $\Sigma$ is unknown, which of course is the normal case, then it is replaced by $S$, the pooled within-treatments estimate and we use this form (...) of the statistic, where the $z$-vector is replaced by the vector of t-statistics. This statistic follows asymptotically a standard normal distribution, but not a t-distribution.

But it was observed in simulations performed by O'Brien that, when comparing two treatments, a t-distribution with $\sum (n-k)$ df provides an accurate fit even for quite small sample sizes.

Furthermore O'Brien found in his simulations that this test statistic is remarkably robust and achieves optimality in the normal theory setting by utilizing the information contained in the correlation matrix.

The robustness of the individual t-statistics must be assessed in a normal way, and if it is adequate for all of them, then the robustness will be adequate for the global test statistic as well.

If the endpoints consist of binary data we can use the normal approximation to the binomial. $Z_{ij} = \left( \hat{p}_{ij} - \hat{p}_{i \cdot} \right) \sqrt{\frac{\hat{p}_{i \cdot} \hat{q}_{i \cdot}}{n_{ij} \cdot n_{i \cdot} \cdot n_{\cdot j}}} \sim N(0,1)$

$s_{ij} = \text{the proportion of all patients with responses for both the variables i and j.}$

$A_{ij} = \left( \hat{p}_{ij} - \hat{p}_{i \cdot} \hat{p}_{j \cdot} \right) \sqrt{\frac{\hat{p}_{i \cdot} \cdot \hat{q}_{i \cdot} \cdot \hat{p}_{j \cdot} \cdot \hat{q}_{j \cdot}}{n_{ij} \cdot n_{i \cdot} \cdot n_{\cdot j}}} \sim N(0,1)$

And again if the sample sizes are adequate for each univariate normal approximation then it will be adequate also for the global test statistic.
In principle there is no difficulty in estimating correlations leading to an asymptotically valid use of the global test statistic. In the referenced article by Pocock, Geller and Tsiatis are there examples of how to calculate the correlation matrix for several survival endpoints, and for a combination of one survival and one binary endpoint.

When dealing with small data sets with outliers, or with skewed distributions, then a non-parametric version of the global test statistic could be used instead. It is of a rank-sum type and simply consists of performing a two-sample t-test on the sums of the ranks for each patient when ranking the outcomes of each endpoint seperately for the combined sample from the two treatment groups.

EXAMPLE

A pre-clinical study on rats was performed at my company, with the aim of reducing abnormal cell-growth during high dose treatment. And comparing one of our new drugs with a standard therapy.

The results from the univariate tests on the standardized variables and the outcome of the global test statistic are given below, and they are followed by listnings of the programs involved in the analyses.

<table>
<thead>
<tr>
<th>variable</th>
<th>n</th>
<th>mean</th>
<th>sd</th>
<th>n</th>
<th>mean</th>
<th>sd</th>
<th>t-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>drug 1</td>
<td>drug 1</td>
<td>drug 1</td>
<td>drug 2</td>
<td>drug 2</td>
<td>drug 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gastrin</td>
<td>32</td>
<td>-0.229</td>
<td>1.044</td>
<td>33</td>
<td>0.222</td>
<td>0.916</td>
<td>1.855</td>
</tr>
<tr>
<td>hdc</td>
<td>32</td>
<td>-0.237</td>
<td>0.893</td>
<td>33</td>
<td>0.230</td>
<td>1.057</td>
<td>1.918</td>
</tr>
<tr>
<td>ecl_ant</td>
<td>32</td>
<td>-0.199</td>
<td>0.924</td>
<td>33</td>
<td>0.193</td>
<td>1.046</td>
<td>1.597</td>
</tr>
<tr>
<td>ecl_pct</td>
<td>32</td>
<td>-0.215</td>
<td>1.021</td>
<td>33</td>
<td>0.208</td>
<td>0.948</td>
<td>1.732</td>
</tr>
</tbody>
</table>

The correlation matrix are given below, with its inverse to the right:

<table>
<thead>
<tr>
<th>gas</th>
<th>hdc</th>
<th>ecl1</th>
<th>ecl2</th>
<th>gas</th>
<th>hdc</th>
<th>ecl1</th>
<th>ecl2</th>
</tr>
</thead>
<tbody>
<tr>
<td>gas</td>
<td>1</td>
<td>.32</td>
<td>.66</td>
<td>.39</td>
<td>1.80</td>
<td>0.08</td>
<td>-1.14</td>
</tr>
<tr>
<td>hdc</td>
<td>1</td>
<td>.48</td>
<td>.50</td>
<td>.58</td>
<td>1.63</td>
<td>-0.47</td>
<td>-0.74</td>
</tr>
<tr>
<td>ecl_ant</td>
<td>1</td>
<td>.49</td>
<td>1.63</td>
<td>0.49</td>
<td>2.16</td>
<td>-0.34</td>
<td></td>
</tr>
<tr>
<td>ecl_pct</td>
<td>1</td>
<td>1.67</td>
<td>1.67</td>
<td>1.67</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Global test statistic for positively correlated variables

<table>
<thead>
<tr>
<th>test-statistic</th>
<th>df</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.328</td>
<td>57</td>
<td>0.023</td>
</tr>
</tbody>
</table>
PROGRAM 1:

* Weighting of results for positively correlated variables *;

options nonumber nodate ps=60 ls=79; title;
filename meta 'overall.dat';

* getting the input data set *

data x;
   infile meta;
   input lab $ drug pat res;
proc sort; by pat lab;

%let data=x;
%let var=lab;
%let drug=drug;
%let pat=pat;
%let val=res;
%include 'hots_sas:global_weight.sas';

---

PROGRAM 2:

******************************************************************************************
* Filename:       GLOBAL_WEIGHT.SAS                                      *
* Title:          Calculation of a global test statistic.                *
* Author:         Lars Frison                                         *
* Date:           890911                                              *
*                  ******************************************************************************************
* Abstract:       The procedure performs a weighting of the results for   *
*                  positively correlated variables (for example all symptom  *
*                  variables), and calculates a global test statistic.    *
*                  ******************************************************************************************
* Parameters:     DATA.......input data set (individual patient data for each var.)*
*                  VAR........variable containing the names of the variables tested *
*                  DRUG.......the variable containing the two groups for which the *
*                  global test statistic is to be calculated                *
*                  PAT.......patient numbers                              *
*                  VAL.......values for different variables                *
*                  ******************************************************************************************
*                  The analysis of multiple endpoints in clinical trials.  *
*                  Biometrics 43, 487-498.                       *
*                  ******************************************************************************************

options nonumber nodate ps=60 ls=79 dquote; title;
libname in '[ ]';
*===== getting the input data set and calculating the number of variables =====*;

data base;
  set &data;
  var=&var; rand=&drug; pat=&pat; val=&val; x=1; drop drug;
proc sort; by x var pat;

data x0; set base; by x var pat;
  if n=1 then nr=0;
  if last.var then nr+1;
  retain nr; keep x nr;
proc sort; by x nr;

data x01; set base; by x var pat;
  drop x; proc sort; by rand var pat;

data x1; set x01; by rand;
  if n=1 then drug=0;
  if first.rand then drug+1;
  retain drug;
proc sort; by nr pat var;

*========== transposing and deleting patients with missing values =========*
  *and calculating the standardized normal deviates *;
proc transpose out=x2; var val; by nr pat;

* deleting patients with missing values *
data x3; set x2;
  array cool col1-col150;
  do i=1 to nr;
    if cool[i]=. then delete;
  end;
  keep nr pat cool col1 col2 col3 col4;

data xy4; merge x3(in=q) x1; by nr pat; if q; drop rand;
proc sort; by var drug;

proc means noprint; var val; by var;
  output out=dev mean=mean std=std;
data x4; merge xy4 dev; by var; val=(val-mean)/std; drop mean std;
data p1; set x4; by var; if first.var; keep var;

*======== calculating some descriptive statistics for the two groups ======*
proc means data=x4 noprint maxdec=2; var val; by var drug;
  output out=x5 n=n mean=mean std=std;
data x6; set x5; by var drug;
  mean1=lag(mean); std1=lag(std); mean2=mean; std2=std;
  n1=lag(n); n2=n;
  drop mean std var n drug;
data x7; set x6;
  x=n/2;
  if mod(2*x,2)=0;
proc sort; by x;
*======== calculating the t-statistics and stripping the glm output =========*

proc printto unit='new' new; /* will direct output to the file for001.dat. */
proc glm data=x4;
  class drug; by var;
  model val=drug / ss3;
proc printto new; /* will direct output back to standard output. */
%include 'hots_sas:glm_global1.sas';

*=== computing correlation matrices and other data to be used by proc iml ===*
* data sets to be used by proc iml *
data x8; set in.temp; t=sqrt(t); x=_n_
data x9; set in.temp; t=sqrt(t);
data x10; set x4; keep nr drug pat val;
proc sort; by drug pat;
proc transpose out=x11; by nr drug pat;
* calculation of the correlation matrix *
proc corr noprint outp=x12; by nr drug; var col1 col2 col3 col4;
data x13; set x12; if left( type )='N'; if drug=1 then n1=col1; else n2=col1;
data x14; set x13; if drug=1; keep n1;
data x15; set x13; if drug=2; keep n2;
data x16; set x12; if 3< _n_<nr+4; drop drug _type_ _name_ nr;
data x17; set x12; if nr+6<_n_<nr+11; drop drug _type_ _name_ nr;
data xnr; set x0; drop x;

*=== using the interactive matrix language to get the global test-statistic ===*
proc iml;
  use x16; read all into s1;
  use x17; read all into s2;
  use x14; read all into n1;
  use x15; read all into n2;
  s=((n1-1)/(n1+n2-2))*s1+((n2-1)/(n1+n2-2))*s2;
  use x9;
  read all into t;
  use xnr;
  read all into k;
i=shape(1,k,1);
y=(i*inv(s)*t)/sqrt(i*inv(s)*i);
create out from y (colname=c);
append from y;
quit;
*--- getting the results in shape for presentation and printing the results ---*

```sas
data x18; set x14; x=1;
data x19; set x15; x=1;
data x20; set out; y=coll; x=1;
data x21; merge x18 x19 x20 x0; by x;
df=n1+n2-2*nr;
pval=2*(1-probt(y,df));
label y='test-statistic' df='df' pval='p-value';
x 'delete [' for 001.dat.*';
data p2; set pl; x=_n_; proc sort; by x;
data x22; merge p2 x7 x8; by x;
  label var='variable' mean1='mean/drug 1' mean2='mean/drug 2' n2='n drug 2'
  std1='sd/drug 1' std2='sd/drug 2' t='t-value' nl='n/drug 1';
proc print noobs label split='/' var var nl mean1 std1 n2 mean2 std2 t;
format mean1 mean2 std1 std2 t 8.3;
title 'Global test statistic for positively correlated variables';
proc print noobs label data=x21; var y df pval; format y pval 8.3; title;
```

---

**PROGRAM 3**:

```sas
OPTIONS LS=79 PS=60 NODATE NONUMBER;
LIBNAME IN 'DBLWF:MBL.FATPROCS';
FILENAME GLM 'FOR001.DAT';
```

**DATA IN.TEMP (KEEP=T)**;

```sas
INFILE GLM;
LENGTH ROWTEXT $ 80;
LENGTH TESTTEXT $ 8;
INPUT MARKER $ 1 ROWTEXT & @;
TESTTEXT=SUBSTR (ROWTEXT, 1, 8);
IF (TESTTEXT EQ 'MODEL ') THEN INPUT DF SSM MSM T;
IF T NE .;
```

---

**CONCLUSIONS**

There is no general optimal strategy for the use of significance testing, when analyzing multiple endpoints in clinical trials.

Which approach one should use depends on what kind of hypothesis one is interested in.
If, as often is the case as far as clinical trials are concerned, the alternative hypothesis one is most interested in, is that the differences between the treatments are of the same magnitude and in the same direction, then a global test statistic like the one I have described might be the best alternative.

**REFERENCES**

---
