Multiple Primary Endpoints in Clinical Trials

Michael J. Brown
Robb J. Muirhead
Pfizer Global Research and Development

BASS XI
November 2, 2004: Savannah, GA

Outline

- Two Presentations in one
- Multiple Endpoint Issues (MB)
  - Description
  - Endpoints
  - Measuring Disease
- Composite Endpoint as a solution (MB)
- Statistical Methodology (RM)
  - IUT
  - LRT
  - Size, power, bias, sample size

Multiple Endpoints

- There is concern about an increasing trend towards requiring that confirmatory clinical trials achieve statistical significance on **all** of *p* primary endpoints, where $p>1$.

- Obviously, as *p* increases, it becomes more difficult to achieve success in any given disease setting

- PhARMA / FDA Workshop on Clinical, Statistical and Regulatory Challenges of Multiple Endpoints, October 20-21, 2004, Bethesda, MD

Some Examples

- Migraine
  - Pain-free at 2 hours
  - Nausea at 2 hours
  - Photosensitivity at 2 hours
  - Phonosensitivity at 2 hours
- Alzheimers
  - ADAS-Cog
  - CIBIC+

What this implies

- All endpoints are equally important

  and

- Interchangeable
  - e.g. migraine
    - Study with pain $p<.0001$, nausea $p=.06$ has the same importance as study with pain $p=.06$, nausea $p<.0001$.

Examples with Multiple Endpoints

1. Migraine (4)
2. Alzheimers (2)
3. Acute Pain (3)
4. Lower Back Pain (3)
5. Sleep Disorders (3 or 6)
6. RA (4)
7. OA for symptom modifying (2)
8. Asthma, COPD (2)
9. ED (3)
10. Skin Aging (2)

Examples with Multiple Endpoints (2)

11. Menopausal Symptoms (3)
12. Fracture Healing (2)
13. Acne (4)
14. Male Pattern Baldness (2)
15. Glaucoma (9)
16. Ophthalmology – dry eye (2)
17. Hepatitis B (up to 3)
18. Vaginal Atrophy (3)

Examples with Multiple Endpoints (3)

19. Organ Transplantation (2)
20. Primary Biliary Cirrhosis (PBC) (4)
21. BPH (2)
22. Multiple Sclerosis (2)
23. Epilepsy (3)
24. Vaccines (up to 23)
25. Operable Breast Cancer (with positive auxiliary lymph nodes) (2)
26. Fibromyalgia (2-3)

Multiple Endpoints

- Do we have a good understanding of the statistical properties of the "obvious" testing procedure -- where each endpoint is tested separately?

- Technical problems arise in this testing problem because the null and alternative hypotheses correspond to "non-standard" partitions of the parameter space.

Level of Evidence

- Is it sufficient to argue that multiple endpoints are bad – because there are difficulties in analysis?
- Should ask: What is the evidence that will allow a conclusion of effect in a disease?
- Need to consider evidence on "multiple" levels – not just multiple endpoints

"Primary" and "Secondary"

- Primary Endpoints
  - These endpoints define the disease in the sense that an experimental drug that does not show superiority over placebo for all of these endpoints is not a viable treatment for the disease under study
- Secondary Endpoints
  - These endpoints, although not considered primary, are considered important to prescribing physicians in helping to identify the ideal treatment for each of their patients

Objectives vs. Endpoints

- Objective –
  - The intention of the study (general)
  - The conclusion (hypothesis) you wish to reach (specific)
  - May be primary, secondary, tertiary
- Endpoints –
  - The set of measurements used to address objectives
  - May have one-one mapping, hence primary, secondary, tertiary
  - May meet multiple objectives

Objectives vs. Endpoints

- Is multiplicity because of number of endpoints? Or because of multiple endpoints addressing a single objective?
- Type I / II error rates are functions of conclusions - Easier to associate with an objective.
- Best to evaluate operating characteristics of decision process – more complicated processes are more difficult to evaluate

Measuring the Disease

- Is there a single key measure of the disease?
  - Assess primary objective by requiring a "significant" effect on single endpoint with supporting evidence on other ("secondary") endpoints
- Are there multiple ways to measure, but each is important individually?
  - A drug that has a dramatic effect on only one of the important endpoints should be made available to patients with that symptom. (Drugs could be targeted for different symptoms.)

Measuring the Disease

- Are multiple measures required to characterize disease?
  - Assess primary objective by requiring a "significant" effect on two or more endpoints
  - Use a composite (is this a single measure?)
  - Corollary: A patient with one symptom but not the others does not have the disease

What is the right question?

Example

- Insomnia is a disease that has a number of symptoms associated with it, but not all patients have all of them
  - Look for benefit in onset of sleep
  - Look for benefit in longer, continuous sleep
  - Effect on either would be important

Composite Endpoints: Solution?

- Composite endpoint – a single measure of effect from a combined set of different variables

- Common in time to event analyses
  - CV: First event of MI, Stroke, CABG, Hospitalization, Death
  - Diabetic Nephropathy: Decreased Renal Function, End Stage Renal Disease, Death
  - Oncology: Progression or Death

Composite Endpoints: Solution?

- Rheumatoid Arthritis – ACR20 Response
  - 20% improvement in tender joint count
  - 20% improvement in swollen joint count
  - Plus 20% improvement in 3 out of 5 of:
    - Patient pain assessment
    - Patient global assessment
    - Physician global assessment
    - Patient self-assessed disability
    - Acute phase reactant

Composite Endpoints: Components

- How to interpret components?
  - Significant in one and weak in others
  - None significant, but all in right direction
  - Should you analyze components individually?
- Question may be:
  - Does the drug do something? vs. What does the drug do?
  - Public health needs vs. labeling and informing the prescriber
- Number of components may impact interpretation

Composite Endpoints: Components

- How to weight different components?
  - Death in time to event
    - Use life years as weighting for event (up-weight death)
    - Death (all cause) is not sensitive
    - Death is a competing risk but may be important or not (do not expect impact)
  - ACR20 has built in weighting – is that reflected in component analysis?

Composite Endpoints: Components

- Is the composite a measure of the disease (individual components do not fully measure the disease) or is it for convenience of analysis?
  - Sparse events
  - Competing risk
  - Multiplicity
- Are the events surrogates for other events or surrogates for something else?
  - CV events are an outcome of underlying disease
  - Diabetic Nephropathy increasing severity of disease

Clinical Need vs. Statistical Method

- Align the statistical approach with the medical/clinical requirements for a "win"
- Statistical underpinnings but a clinical problem
- Clarity of definitions and consensus regarding the clinical trial structure for a "win" is a strong motivation for why we are here

- Robert T. O'Neill, Director, Office of Biostatistics CDER, FDA, PhARMA /FDA Workshop Oct 20-21, 2004

Summary

- Issues in the use of Multiple Endpoints are multi-faceted - The Discussion needs to focus on the following questions:
  - What set of measures are necessary to characterize a disease and the impact of intervention on that disease?
  - How should the measures be used to establish evidence of effect? Single primary? Multiple primary? Composite?
  - What is the best statistical methodology for showing effect?

Multiple Primary Endpoints: A Model

- Joint work with Morris L. Eaton (University of Minnesota)
- Suppose we have n_1 subjects on drug and n_2 subjects on placebo
- Suppose there are *p* primary endpoints, assumed to have a *p*-variate normal distribution.
- Thus we have:

$$X_1^{(D)}, \ldots, X_{n_1}^{(D)} \sim iid\ N_p\left(\mu^{(D)}, \Sigma\right)$$

$$X_1^{(P)}, \ldots, X_{n_2}^{(P)} \sim iid\ N_p\left(\mu^{(P)}, \Sigma\right)$$

- Let $\Delta = \mu^{(D)} - \mu^{(P)}.$
- To show efficacy on all *p* endpoints, we need to be able to conclude that $\Delta_i > 0$ for all $i = 1, \ldots, p$. This will then be the alternative hypothesis.

Model (cont.)

- Let $\bar{X}^{(D)}, \bar{X}^{(P)}, S^{(D)}, S^{(P)}$ be the sample mean vectors and sample covariance matrices.
- Put $S = (n_1 - 1)S^{(D)} + (n_2 - 1)S^{(P)}$.
- Finally, let

$$Y = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \left(\bar{X}^{(D)} - \bar{X}^{(P)} \right), \quad \mu = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \Delta.$$

Model (cont)

- Then

$$Y \sim N_p(\mu, \Sigma) \text{ and } S \sim W_p(n, \Sigma) \text{ (Wishart)}$$

with $n = n_1 + n_2 - 2.$

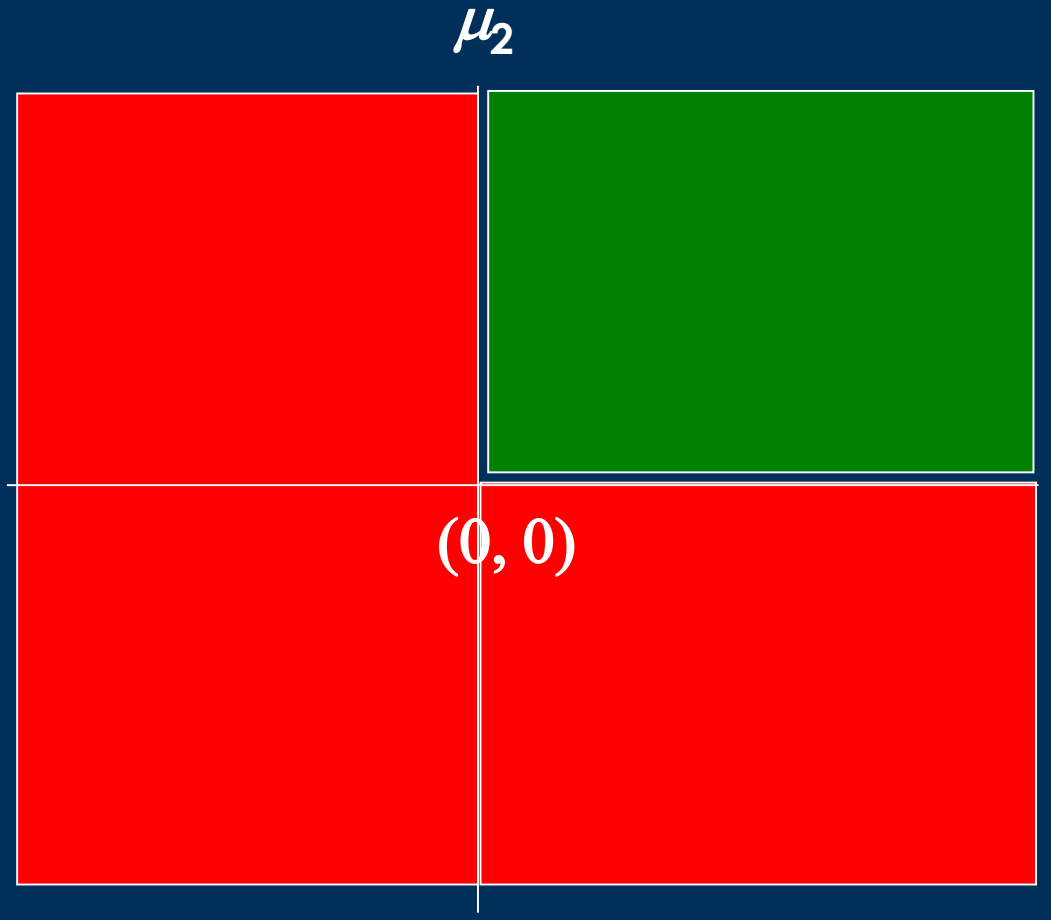
- The alternative hypothesis of interest is then

$$\mu_i > 0 \text{ for all } i = 1, \ldots, p.$$

- A natural null hypothesis is then that

$$\mu_i \leq 0 \text{ for at least one } i.$$

p =2: Null & Alternative μ Parameter Spaces



This is not the whole story! It is **not** the complete parameter space, which also involves the covariance

The Testing Problem

- To summarize, we observe a random vector Y and a random matrix S, where

$$Y \sim N_p(\mu, \Sigma) \text{ and } S \sim W_p(n, \Sigma) \text{ (Wishart)}$$

with both μ and Σ unknown.

- The null and alternative hypotheses are

$$H_0 : \mu_i \leq 0 \text{ for at least one } i; \quad \text{i.e., } \min_i \mu_i \leq 0$$

$$H_A : \mu_i > 0 \text{ for all } i = 1, \ldots, p; \quad \text{i.e., } \min_i \mu_i > 0$$

The Intersection Union Test (IUT)

- The "standard" procedure, where each coordinate of the parameter vector μ is tested separately at the same level α is an intersection-union test (IUT).
- Let S_p^+ be the set of all *p*x*p* positive definite matrices.
- The full parameter space is then

$$\Theta = \left\{ (\mu, \Sigma) \middle| \ \mu \in R^p, \Sigma \in S_p^+ \right\}.$$

The IUT (cont)

- Let

$$\Theta_i = \left\{ (\mu, \Sigma) \,\middle|\, \mu_i \leq 0, \Sigma \in S_p^+ \right\}, \text{ for } i = 1, \ldots, p.$$

- Then the null and alternative hypotheses are

$$H_0 : (\mu, \Sigma) \in \bigcup_{i=1}^{p} \Theta_i$$

$$H_A : (\mu, \Sigma) \in \bigcap_{i=1}^{p} \Theta_i^c$$

IUT (cont)

- A one-sided test of level α for testing Θ_i versus Θ_i^c has the rejection region $T_i \geq c_\alpha$ where

$$T_i = \frac{Y_i}{\sqrt{s_{ii}/n}}$$

and c_α is the upper α point of the t_n distribution.

- The test that rejects H_0 if and only if

$$T_i \geq c_\alpha \text{ for all } i = 1, \ldots, p$$

is an IUT. (The rejection region is the intersection of all the individual rejection regions.)

IUT (cont)

- From now on, we assume $0 < \alpha < 1/2$, so $c_\alpha > 0$.
- Let

$$T = \min_{1 \le i \le p} T_i.$$

- The IUT with size α rejects H_0 if $T \ge c_\alpha$.
- This is sometimes called the "min test".

The Likelihood Ratio Test (LRT)

Result 1: The LRT is identical to the IUT.

Steps involved in showing this:

- The likelihood function is proportional to

$$L(\mu, \Sigma) = |\Sigma|^{-(n+1)/2} \exp\left[-\frac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu) - \frac{1}{2}\operatorname{tr}\Sigma^{-1}S \right]$$

- For fixed μ, the matrix

$$\hat{\Sigma} = \frac{1}{n+1}\left(S + (y-\mu)(y-\mu)'\right)$$

  maximizes *L*.

The LRT (cont)

- Now, $L(\mu, \hat{\Sigma})$ is proportional to

$$L^*(\mu) = \left[1 + (y-\mu)' S^{-1} (y-\mu)\right]^{-(n+1)/2}$$

- So, for testing H_0 versus H_A the LRT rejects for small enough values of

$$\Lambda = \frac{\sup_{H_0} L^*(\mu)}{\sup_{\mu \in R^p} L^*(\mu)}$$

The LRT (cont)

- The denominator here is equal to 1, so the LRT rejects H_0 for large enough values of

$$D = \inf_{H_0} (y-\mu)' S^{-1} (y-\mu)$$

- But it can be shown that

$$D^{1/2} = \frac{T}{\sqrt{n}}$$

- Thus rejecting H_0 for large *D* is equivalent to rejecting for large *T,* and this is the IUT.

What now?

- So the IUT of size α and the LRT of size α are identical.
- The test itself does **no**t involve the correlations between the endpoints (but its properties do).
- What's known, or can be proved, about the test?

Properties of the Test

1. Its size is $\alpha -$ that is, the maximum Type I error probability is α. (Under quite general conditions this is true for IUTs, so no multiplicity adjustment is needed with IUTs.)

2. It may be conservative. The intended level may be quite a bit smaller than α. For example, if all $\mu_i = 0$ and $\Sigma = I_p$, the probability of a Type I error is α^p, which is less than α.
   - But the correlations also play an important role that is often overlooked. For example, when p=2 and the correlation is 1, the Type I error probability is α.

Properties of the Test (2)

- More on size: The size α is achieved in the null parameter space when Σ is fixed, one coordinate of μ is zero, and the remaining coordinates of μ are $+\infty$.
- Suppose $p = 2$. The Type I error probability reaches the intended significance level when either (1) $\Delta_1 = 0$ and $\Delta_2 = +\infty$, or (2) $\Delta_1 = +\infty$ and $\Delta_2 = 0$. If either (1) or (2) hold, the treatment has no effect on one endpoint and an infinitely large effect on the other.

Properties of the Test (3)

3. The test is **biased**, which means that there are parameter values in the ***alternative space*** for which the probability of rejecting the null hypothesis (the power) is smaller than α. (Recall that when all $\mu_i = 0$ and $\Sigma = I_p$, the probability of rejecting the null hypothesis is α. This implies, since the power function is continuous in the parameters, that there are points close to 0 in the alternative space for which the power is less than α. This may not be a serious problem – many tests in common use are biased.)

Properties of the Test (4)

What can we say about statistical issues such as:

- The *p*-value of the test?
- The power function of the test?
- Sample sizes needed to achieve a specified power?

The *p*-value

- The test which rejects if $T \geq c_\alpha$, where

$$T = \min_{1 \leq i \leq p} T_i,$$

  is both the IUT and LRT of size α.
- Suppose the value $T = t_0$ is observed.
- The *p*-value is then

$$p = \sup_{H_0} P\{T \geq t_0 \mid \mu, \Sigma\}.$$

p-value (cont)

- The *p*-value is just the upper tail probability of a *t* distribution, and so is easily calculated.

Result 2:

$$p = P\left\{T_1 \geq t_0 \,\middle|\, \mu_1 = 0, \Sigma = I_p\right\}$$
$$= P\left\{t_n^* \geq t_0\right\}$$

where t_n^* is a random variable with a t_n distribution.

Power

- For any (μ, Σ), the power function is

$$\pi(\mu, \Sigma) = P\{T \geq c_\alpha \mid \mu, \Sigma\}.$$

- Thus the power appears to depend on

$$p + \frac{1}{2} p(p+1)$$

parameters.

$p = 2 \Rightarrow 5$ parameters in μ and Σ

Power (cont)

- But, because the test is invariant under positive scale changes of each coordinate,

$$\pi(\mu,\Sigma)=\pi(\xi,R)$$

where *R* is the correlation matrix and

$$\xi=\left(\mu_1/\sqrt{\sigma_{11}},\ \cdots\ ,\mu_p/\sqrt{\sigma_{pp}}\right)'$$

- Thus the power depends "only" on

$$p+\frac{1}{2}p(p-1)=\frac{1}{2}p(p+1)\ \text{parameters}$$

$$p=2\Rightarrow 3\ \text{parameters}$$

Power (cont)

- Marginally, each T_i has a non-central *t* distribution with *n* degrees of freedom and non-centrality parameter

$$\xi_i = \mu_i \Big/ \sqrt{\sigma_{ii}}.$$

Result 3: If the covariance matrix Σ is diagonal, then

$$\pi(\mu, \Sigma) = \pi(\xi, I_p) = \prod_{i=1}^{p} P\{T_i \geq c_\alpha \mid \xi_i\}.$$

Power (cont)

- In the multiple endpoint setting, it is probably reasonable to assume that the elements of Σ are non-negative – i.e., the correlations between endpoints are non-negative.
- In this case it is possible to obtain a lower bound for the power function.

Result 4: When all correlations are non-negative,

$$\pi(\mu,\Sigma) \geq \pi(\xi, I_p) = \prod_{i=1}^{p} P\{T_i \geq c_\alpha \mid \xi_i\}.$$

Sample size

- The calculation of this lower bound

$$\pi(\mu,\Sigma)\geq\prod_{i=1}^{p}P\left\{T_i\geq c_\alpha\,|\,\xi_i\right\}.$$

for the power function requires specification of the non-centrality parameters

$$\xi_i=\sqrt{\frac{n_1n_2}{n_1+n_2}}\,\frac{\Delta_i}{\sqrt{\sigma_{ii}}},\quad i=1,\ldots,p$$

Sample size (cont)

- Suppose $n_1 = n_2 = m,$ and all $\dfrac{\Delta_i}{\sqrt{\sigma_{ii}}} = \gamma.$
- Then all the ξ_i's are equal to $\beta = \sqrt{\dfrac{m}{2}}\,\gamma.$
- The lower bound result is then

$$\pi(\mu, \Sigma) \geq \left(P\left\{T_1 \geq c_\alpha \mid \beta\right\}\right)^p,$$

where T_1 has a non-central *t* distribution with 2*m*-2 degrees of freedom and non-centrality $\beta.$

Sample size (cont)

- Setting e.g.

$$\left(P\left\{T_1 \geq c_\alpha \mid \beta\right\}\right)^p = 0.8$$

and solving for *m* yields a sample size necessary to ensure that the power is at least 0.8

- This would, of course, have to be done numerically – but seems straightforward.

Final Comment about Power and Bias

- Take $\Sigma = I_p$. The equation $\pi(\beta e, I_p) = \alpha$ implies

$$P\{T_1 \geq c_\alpha \mid \beta\} = \alpha^{1/p},$$

where T_1 has a non-central *t* distribution with 2*m*-2 degrees of freedom and non-centrality β.

- For example, if $m = 26$, $\alpha = .05$, and $p = 4$, then (approx) $\beta = 1.9$. Thus in the alternative parameter space with $\Sigma = I_p$ and all $\beta_i = 1.9$, the power of the test is .05. In the (unlikely) event that this parameter configuration were deemed clinically meaningful, this would be rather unsettling…..

Summary

- In testing multiple endpoints, the usual test consists of testing each endpoint separately using one-sided *t* tests at level α, and to conclude that the drug is efficacious only if each endpoint is statistically significant; that is, only if

$$T_i = \frac{Y_i}{s_{ii}/\sqrt{n}} \geq c_\alpha \quad \text{for all} \quad i = 1, \ldots, p.$$

- This is equivalent to concluding efficacy only if

$$T = \min_{1 \leq i \leq p} T_i \geq c_\alpha.$$

Summary (2)

- This test is both an IUT and the LRT of size α; that is, the maximum probability of a Type I error is α.
- The test may be conservative, depending on the parameter configuration in the null space.
- The test is biased; that is, there are values of the parameters in the alternative space for which the probability of rejecting the null hypothesis is less than α.

Summary (3)

- A simple expression for the *p*-value is available:

$$p = \sup_{H_0} P\left\{T \geq t_0 \,\middle|\, \mu, \Sigma\right\} = P\left\{t_n^* \geq t_0\right\}.$$

- A simple lower bound for the power function is available in terms of non-central *t* tail probabilities:

$$\pi(\mu, \Sigma) \geq \prod_{i=1}^{p} P\left\{T_i \geq c_\alpha \,\middle|\, \xi_i\right\}.$$

- This lower bound can be used to help determine sample sizes.

Final Thoughts

- The problem of testing multiple endpoints becomes even more complicated when the endpoints are:
  - Discrete; e.g. binary (as in the case of migraine)
  - Some are discrete and some are continuous

- How should such situations be modeled, so that the power function (which answers questions about level, size, bias, power) can be calculated?