Power in Three-Arm Generic Drug Trials of Equivalence

Alan Davis

Young Kim

SCT 2010

The views expressed herein are solely those of the author and do not necessarily reflect the official policy, position or opinions of PharmaNet Development Group, Inc. and its affiliates

Agenda

- Motivation
- Introduction
- Generic Drug Trials
- Sample Size and Power
- Approaches to Estimating Sample Size
- Example
- Conclusions
- Acknowledgements
- References

Introduction

- Generic drug
  - Definition – A drug product that is comparable to a brand/reference listed drug product in dosage form, strength, route of administration, quality and performance characteristics, and intended use
  - Generic Pharmaceutical Association (GphA), 2007
    - Generics represent 69% of the total prescriptions dispensed in the United States but only 16% of all dollars spent on prescription drugs (IMS Health)
    - 10,072 of the 12,751 drugs listed in the FDA's Orange Book have generic counterparts (FDA)
    - The generic industry is growing at more than 7.8%, a pace that is faster than the world pharmaceutical market (IMS Health)
  - NDA (Brand Name) vs. ANDA (Generic) – similar except animal studies, clinical studies and bioavailability (brand name) vs. bioequivalence (generic)
    - 1984 Hatch-Waxman Act
    - Patent protection – 17 years after patent is issued (~12 years of marketing protection)
  - For some forms of dosage, the active ingredient is not measured in the blood or excreted in a way that can be used to compare the generic and brand name products

Generic Drug Trials

- Goal – regulatory approval….show that the generic (test) and brand name (reference) products are equivalent (21 CFR Part 320)
  - Bioequivalence – uptake, excretion, maximum concentration and half-life
  - FDA guidance on bioequivalence (2001)
  - Sometimes clinical response is used to infer that the test and reference are bioequivalent
- Trial to establish bioequivalence in this way may resemble a Phase III clinical trial; clinical response may be a change from baseline, a success/fail evaluation, or other measure
- Showing equivalence requires that the difference between test and reference are smaller than an amount Δ; not clinically important
- A less restrictive criterion would be to show that the test product is “non-inferior” by Δ
- An added assurance is often required; a significant difference between active and an ineffective treatment, it possesses “assay sensitivity”. Requires the inclusion of a third, placebo, arm into the study

Generic Drug Trials

- Example – dichotomous response: success vs. failure
  - Topical medications have a dichotomous end point, i.e., the treatment results in either success or a failure for the subject
  - Treatment effect is summarized as the percentage of subjects, π, whose treatment is a success, as opposed to a failure
  - The test and reference treatments may be considered equivalent when the observed success rates, p_T and p_R, are within a clinically important range Δ
  - A test product is not substantially inferior if $\pi_T \geq \pi_R - \Delta$
  - If the difference $p_T - p_R$ is greater than -20% with high degree of certainty, the test is considered non-inferior to the reference, not ruling out the possibility that the test is in fact superior

Sample Size and Power

- Determined by the objectives for the trial – primary endpoint
- Research historic information on the primary endpoint
  - Typically, an average change from baseline or percentage change
    - Historical, journal articles, statistical review of brand name
  - If unknown, select values large enough to be considered clinically important
    - Example: a 20% increase in treatment success
  - Determine region of each test resulting in satisfaction of trial (rejection region, region of equivalence)
  - Power is now the probability that the trial will be a success, i.e., all tests will fall into a region which satisfies the trial
  - Compute power by summing conditional probability over region of rejection, or by repeated simulation of trials and observing the proportion that all tests are satisfied
  - Protect alpha (Type I error) against false rejection
    - Lauzon and Caffo – if all pair-wise comparisons of k independent groups are being evaluated for equivalence, then simply scaling the nominal Type I error rate down by (k - 1) is sufficient to maintain the family-wise error rate at the desired value or less
    - Less conservative than Bonferroni method

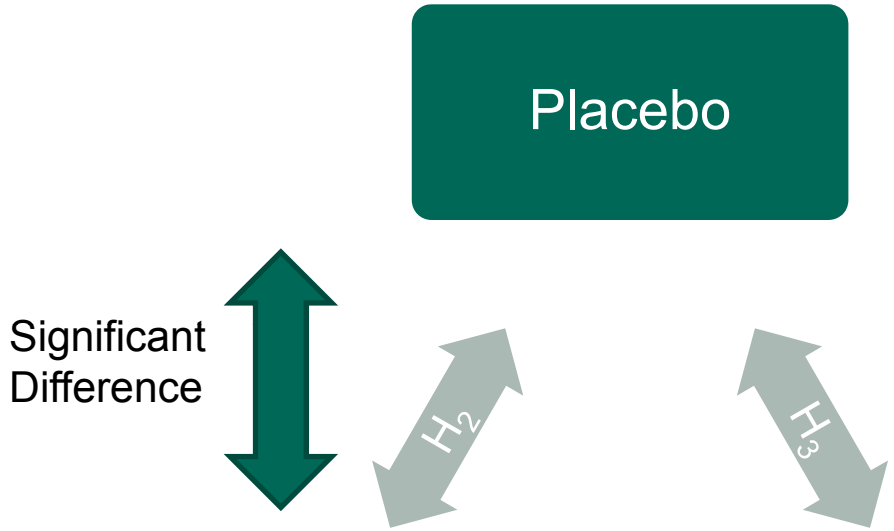
Sample Size and Power

- Calculation is similar to a standard (non-generic) clinical trial
  - Two treatment groups (e.g., study drug vs. placebo), even though there could be more
- Typical values
  - Type 1 error = 0.05
  - Power = 90%
- Statistical methods
  - TOST – equivalence in means (Schuirmann, 1987; Phillips, 1990)
  - Interval estimation – difference in proportions (Newcombe, 1988)
- Calculated sample size is used for all treatment groups
  - Except the third (placebo) arm; ½ the sample size
- Use software (SAS, nQuery, others)
- But, is this sample size, assuming two treatment groups for the primary endpoint, adequate when we have three groups (test, reference, placebo)?

Example: A Trial of a Topical Generic Medication

- For a topical medication, since medication is not systemically absorbed, bioequivalence of a generic equivalent to an established medication must be inferred from clinical effects rather than systemic uptake and release of active ingredient
- This means that subjects are assigned at random to groups having the following treatment:
  - Group 1: Receives the test medication, the generic equivalent
  - Group 2: Receives the standard, or reference medication
  - Group 3: A placebo control, needed to show the study has the precision necessary to distinguish between effective and ineffective treatments
- To be successful in establishing bioequivalence, the sponsor of the trial needs to demonstrate that subjects in Group 1 and Group 2 are close together in terms of treatment effect, and both groups show superior treatment effect to subjects in Group 3

Desired Result of Clinical Trial for Equivalence/Superiority



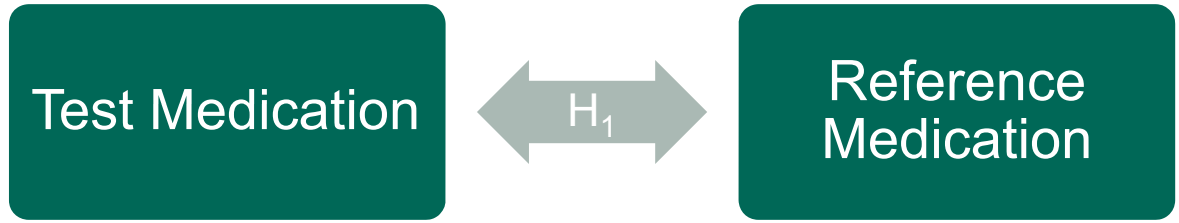
Three necessary Hypotheses:

H_1: Inferiority

($\pi_T - \pi_R \leq -\Delta$)

H_2: Equality of Test vs. Placebo ($\pi_T = \pi_P$)

H_3: Equality of Reference vs. Placebo ($\pi_R = \pi_P$)



Example: Binomial Inference

- For topical medication, outcome could be a reduction in lesion count, symptoms, or a combination of variables. These may be summarized as a simple yes/no result or as percent success in each group.
- Outcome: Success or Failure of Treatment, dependent on π_i, the probability that intervention or treatment is successful in group *i*.
- Let Δ = an interval within which outcome differences will be considered acceptable. For the following, assume $\Delta = 0.2$.
- Lower bound L of the confidence interval of difference between groups will be used to reject the hypotheses:
  - For H_1, that $L_{(T-R)} > -0.2$ (non-inferiority)
  - For H_2, that $L_{(T-P)} > 0.0$
  - For H_3, that $L_{(R-P)} > 0.0$
- Use the Yates' Continuity Corrected Confidence interval for the difference in two independent proportions (Fleiss, 1981).

The Difference In Two Independent Proportions

- An estimate of the difference between Test and Reference is $p_T - p_R = s_T/n_T - s_R/n_R$,

where

p_T = Success Rate of Test,

p_R = Success Rate of Reference.

The standard error of the estimate is then

$$se = \text{sqrt}\ (p_T*(1-p_T)/n_T + p_R*(1-p_R)/n_R),$$

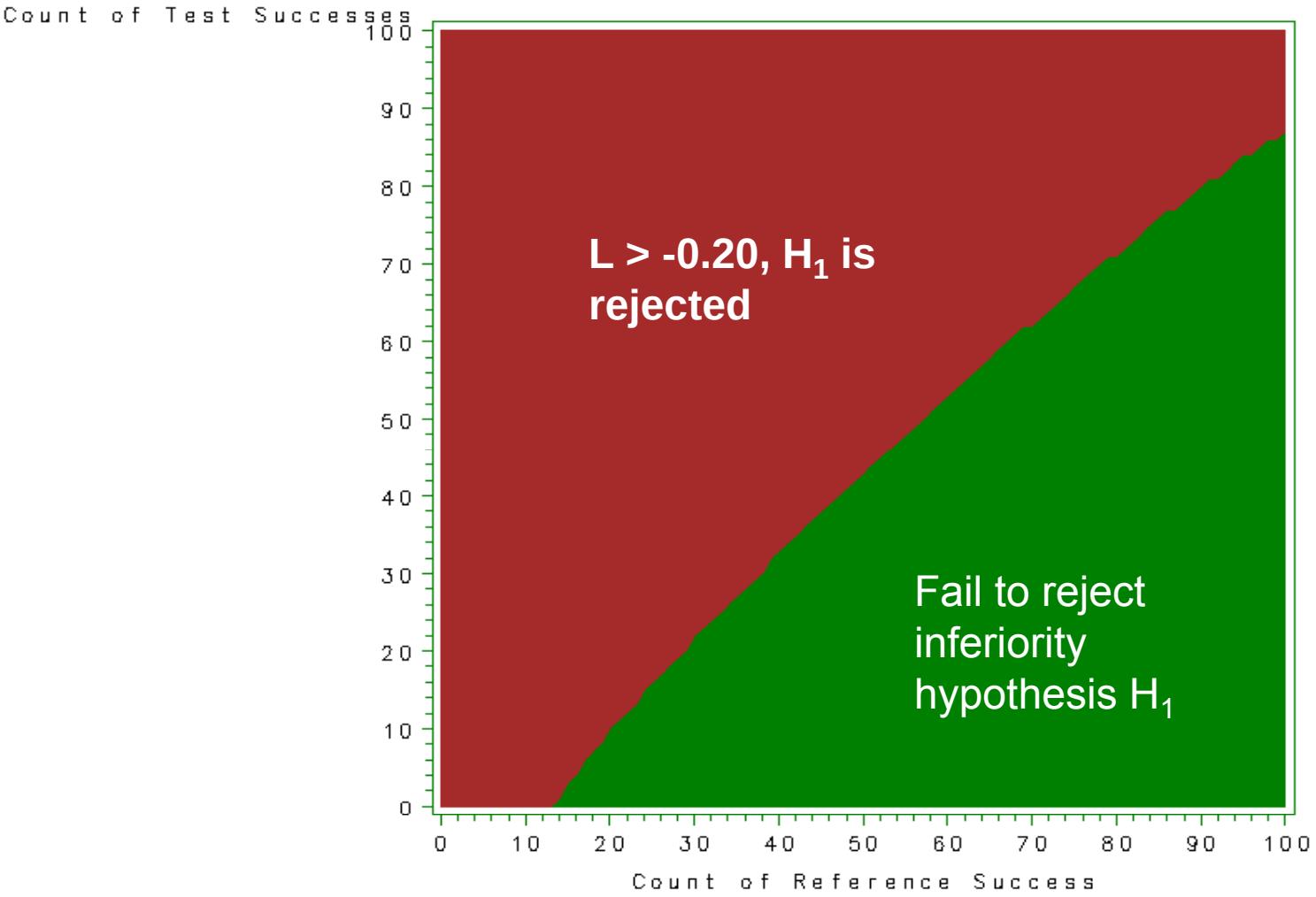
and

$$L = (p_T - p_R) - 1.645*se - (1/n_T + 1/n_R)/2$$

Is the lower 95% one-sided confidence bound of the difference.

When L > -0.20, H_1 is rejected, i.e., we conclude that the test is not inferior to the reference.

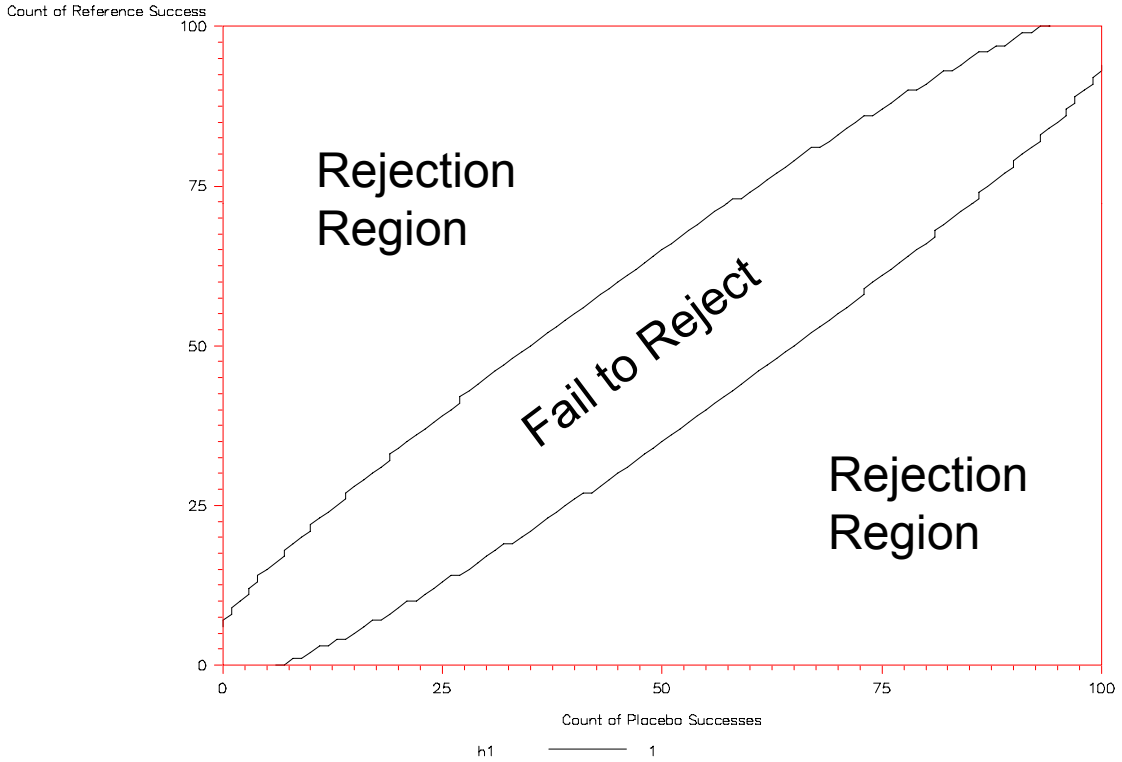
Region to Reject Inferiority, nR = nT = 100



Testing Superiority: H_2 and H_3

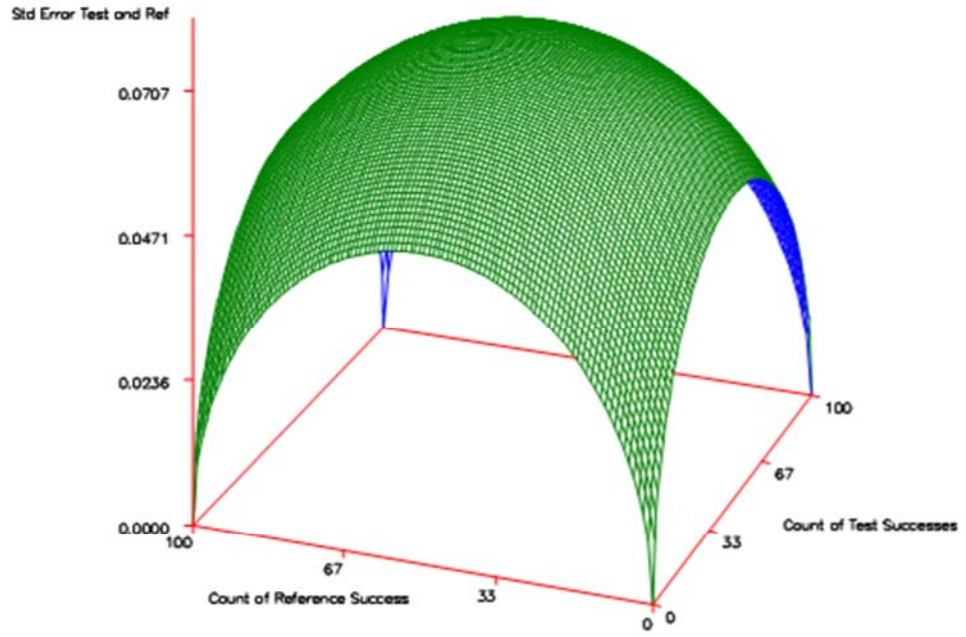
- If superiority over a Placebo group is to be demonstrated, two-sided confidence intervals may be requested by the Regulatory Agency. We have chosen the two-sided 95% confidence interval as the criterion.
- If $p_V = s_V/n_V$ represents the success rate for Placebo (Vehicle), then
  - $se_1 = \text{sqrt}\ (p_T*(1 - p_T)/n_T + p_V*(1 - p_V)/n_V)$
  - $se_2 = \text{sqrt}\ (p_R*(1 - p_R)/n_R + p_V*(1 - p_V)/n_V)$.
- Where
  - se_1 is the standard error of the difference between Test and Placebo, and
  - se_2 is the standard error of the difference between Reference and Placebo.
- Superiority of test and reference over placebo requires that the confidence interval of the difference each exceed zero, i.e. for 95%,
  - $L_{TV} = (p_T - p_V) - 1.96*se_1 - (1/n_T + 1/n_V)/2 > 0$
  - $L_{RV} = (p_R - p_V) - 1.96*se_2 - (1/n_R + 1/n_V)/2 > 0$.

Regions to reject $\pi_R = \pi_P$ for n=0, 1, ..100 successes



Standard Error of Binomial Estimator “se”

For results consistent with 50% success, Standard error is largest = power is smallest.



Simulate Trial Data Results based on π_i and Δ

- Assume that Test is non-inferior and actually $\pi_1 = \pi_2$.
- Assume Placebo less than active doses by Δ or more: $\pi_3 \leq \pi_i - \Delta$, i =(1,2).
- Generate enough trials to accurately determine success rate of trial. Using 2500 trials per set gives a standard deviation of power less than or equal to 0.5%.
- Record percent of trials where H_1, H_2, and H_3 are rejected.
- Tables 1 and 2 show a range of success in rejecting individual and multiple hypotheses where $\Delta = 0.2$ and $\pi_3 = \pi_i - 0.2$, i =(1,2).
- For sample size 100, power is adequate for H_1 or H_2 when $\pi_1 = \pi_2 = .35$ and $\pi_3 = .15$ but not adequate for rejecting all three hypotheses (Table 1, line 2).
- For sample size 160 (Table 2), power is adequate to reject H_1, H_2 and H_3 even at values around $\pi_1 = \pi_2 = .5$.

Table 1. For n=100 per group, power is often inadequate for rejection of all 3 hypotheses.

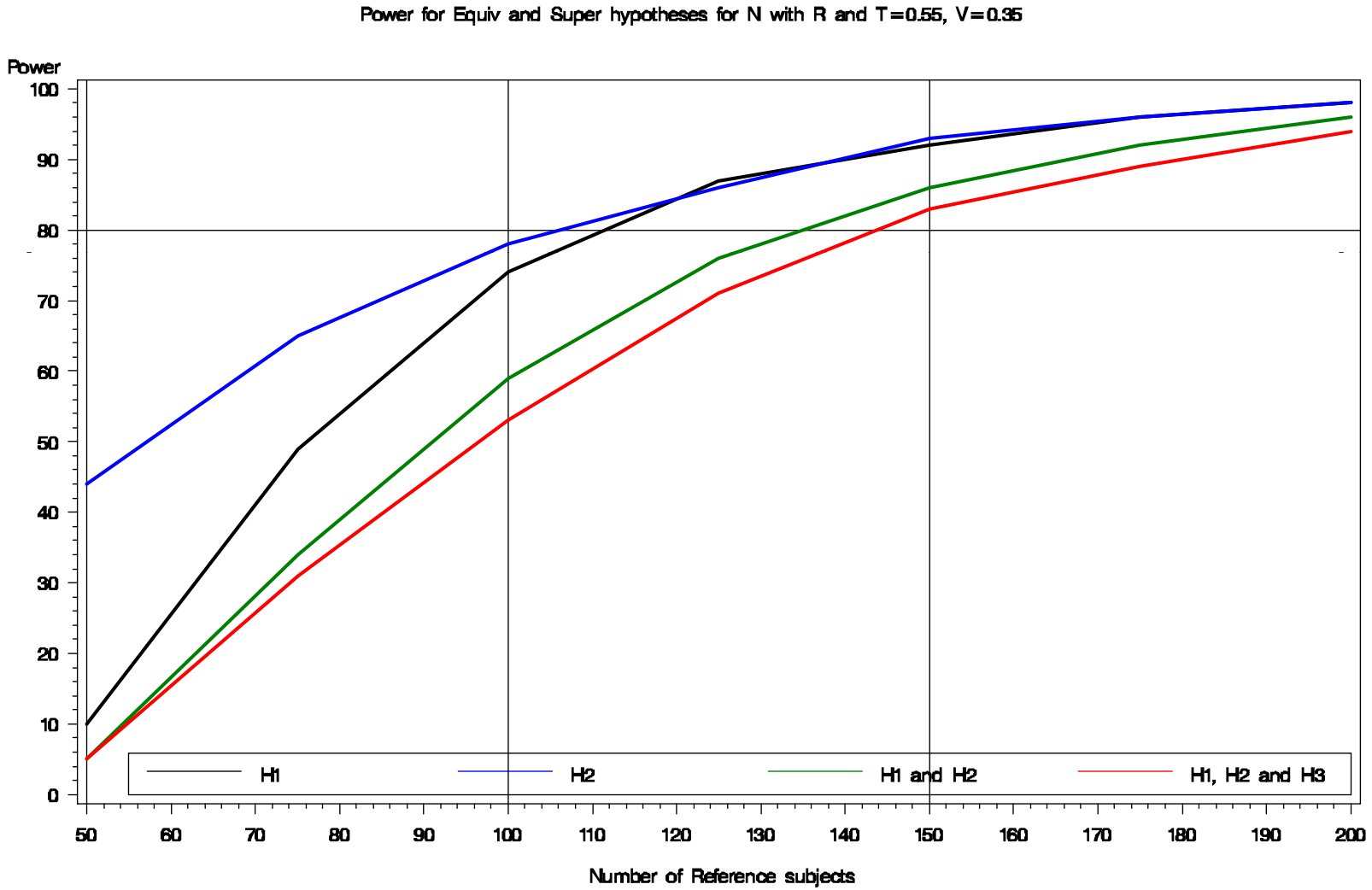
Test, Reference Success Rate (%)	Placebo Success Rate (%)	Reject H_1 (%)	Reject H_2 or H_3 (%)	Reject H_1 and H_2 (%)	Reject H_1 and H_2 and H_3 (%)
25	5	94	98	93	91
35	15	89	90	83	76
45	25	87	83	76	66
55	35	87	78	71	60
65	45	89	78	73	62
75	55	94	83	81	70
85	65	98	91	90	84
95	75	100	99	99	98

Table 2. Increasing the sample size to 160 allows rejection of all hypotheses with good probability.

Test, Reference Success Rate (%)	Placebo Success Rate (%)	Reject H_1 (%)	Reject H_2 or H_3 (%)	Reject H_1 and H_2 (%)	Reject H_1 and H_2 and H_3 (%)
25	5	99	100	99	99
35	15	98	98	97	96
45	25	97	96	94	92
55	35	97	94	92	87
65	45	98	94	93	88
75	55	99	97	96	94
85	65	100	99	99	98
95	75	100	100	100	100

Power to Reject Individual and Joint Hypotheses

Calculated at n=50, 75, 100, 125, 150 and 200.



Power Using 2:2:1 Patient Allocation

- Here, power is adequate (> 80%) only for the extreme low or high success rates with the smallest standard deviation.

Number of Reference or Test Subjects=120, Number of Placebo Subjects=60.

Test Success Rate (%)	Placebo Success Rate (%)	Reject H_1 (%)	Reject H_2 or H_3 (%)	Reject H_1 and H_2 (%)	Reject H_1 and H_2 and H_3 (%)
25	5	97	97	94	92
35	15	94	83	80	72
45	25	92	74	71	60
55	35	92	68	65	54
65	45	94	68	65	55
75	55	97	72	71	60
85	65	100	79	79	70
95	75	100	93	93	90

Using 2:2:1 Allocation, Varying Placebo Success Rates

- When rate differences are large enough, even mid-range power can be high, with $n_P=n_T/2$.

Number of Reference Subjects=120, Number of Placebo Subjects=60.

Test Success Rate (%)	Placebo Success Rate (%)	Reject H_1 (%)	Reject H_2 or H_3 (%)	Reject H_1 and H_2 (%)	Reject H_1 and H_2 and H_3 (%)
55	35	92	68	65	54
55	30	92	88	82	76
55	25	92	98	90	88
55	20	92	100	92	91
55	15	92	100	92	92

Conclusions

1. Power in a generics clinical trial can be estimated similarly to other clinical trials.
2. Typically, a third (placebo) arm is included.
3. In selecting a sample size, the biostatistician should consider all the requirements imposed for the study to successfully demonstrate the efficacy or safety of the treatment being tested. It may be a regulation that several goals must be achieved. These goals may each be reached by rejecting hypotheses which may be interdependent in ways that may be unknown.
4. In this example, the hypotheses to be rejected have a degree of dependence in that although the treatment groups have independent outcomes, the statistics comparing the differences are correlated.
5. Repeated simulations of trial outcomes based on hypothesized group response parameters is a useful way to approach this problem.
6. If a smaller placebo group is desirable, accurate foresight in the population outcome parameters can aid in sample size selection.
7. The methods employed in the example may also be useful in applications where equivalence in safety is of importance.

Acknowledgments and Contact information

The authors would like to thank PharmaNet Development, Incorporated, for providing the materials and experiences that inspire this presentation.

PharmaNet Development Group, Inc.

1000 CentreGreen Way

Suite 300

Cary, NC 27513

Alan Davis:
adavis@pharmanet.com

Young Kim:
ykim@pharmanet.com

References

- 21 CFR § 320.24 “Types of evidence to measure bioavailability or establish bioequivalence”
- Statistical Approaches to Establishing Bioequivalence, FDA Guidance for Industry, 2001, CDER
- Schuirmann, D.J., “A comparison of the two one-sided tests procedure and power approach for assessing the equivalence of average bioavailability”, Journal of Pharmacokinetics and Biopharmaceutics, 15(1987), pp. 657-680
- Phillips, K.E., “Power of the two one-sided tests procedure in bioequivalence”, Journal of Pharmacokinetics and Biopharmaceutics, 18(1990), pp. 137-143
- Newcombe, R.G., “Interval estimation for the difference between independent proportions: comparison of eleven methods”, Statistics in Medicine, 17(1988), pp. 873-890
- Blackwelder, W.C., Journal of Dental Research 83, C113-C115, 2004.
- International Conference on Harmonisation (2000). Guidance E10: choice of control group and related issues in clinical trials, http://www.ifpma.org/ich1.html
- European Medicines Agency, “Guideline on the Choice of the Non-Inferiority Margin”, 2005.
- “Non-inferiority trials: the ‘at least as good as’ criterion”, L. L. Laster and M.F. Johnson, Statistics in Medicine 2003: 22:187-200

References

- Jacob Cohen, Statistical Poser Analysis in the Behavioral Sciences, 1988, Lawrence Erlbaum Associates
- Carolyn Lauzon and Brian Caffo, “Easy Multiplicity Control in Equivalence Testing Using Two One-Sided Tests”, The American Statistician, May 2009 Vol 63, No. 2
- Fleiss, Joseph L., page 23, Statistical Methods for Rates and Proportions, 1981, John Wiley & Sons.
- Food and Drug Administration Reference P04-021, “Mupirocin Calcium Cream 0.2%)”
- nQuery Advisor® 6.0, Copyright © 1995-2005 Janet D. Elashoff
- SAS Institute, Inc. (1990) SAS/GRAPH and SAS/STAT Software: Reference, Cary, NC: SAS Institute, Inc.