A Good N Despite a Bad Start:
A Practical Guide to Easy Internal Pilots

Keith E. Muller

Professor, Health Outcomes and Policy, COM
University of Florida, Gainesville
Keith.Muller@biostat.ufl.edu; web site http://ehpr.ufl.edu/muller

Primary collaborator and coauthor, Chris Coffey.
Others include Kairalla, Taylor, Pasour, Gribbin, and Johnson.

Outline

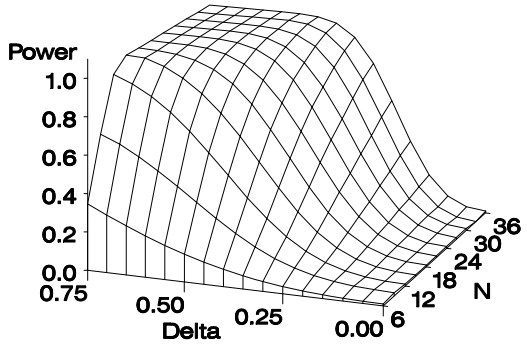
1. Motivation: Good Design and Analysis with Uncertainty About σ^2

Must plan a study with uncertainty about nuisance parameters such as error variance, σ^2.

Want to avoid underpowered study and want to avoid overpowered study (time and cost savings).

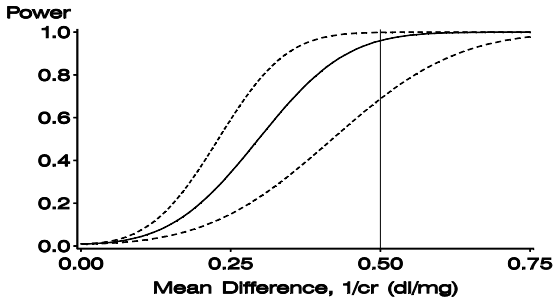
1.1 Gaussian Error Linear Model Power Principles

For a fixed test and predictors, Gaussian linear model power depends *only* on
1) mean differences
2) variance
3) sample size.



t-Test Power as a Function of Mean Difference and N

1.2 Problem: Uncertainty About Nuisance Parameter σ^2



t-Test Power for 3 Variances
(Taylor and Muller 1995, 1996; Muller and Pasour, 1997)

1.3 Design Choices

1 Close your eyes and hope for the best.

2 Traditional external pilot study and design, two studies.

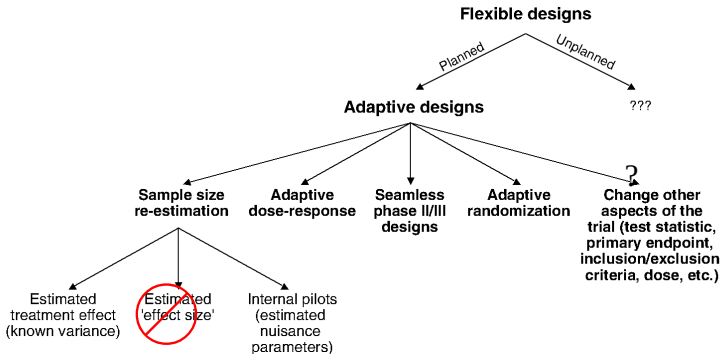
2′ Multistudy design strategies: Muller, Barton, Benignus (1984).

3 Adaptive designs, including internal pilot, group sequential, et al.

How do we provide accurate inference if we use data to adjust the design?

Controversy surrounding some adaptive designs unintentionally implied guilt by association.

Design Choices (Coffey and Kairalla, 2008)



Group Sequential Design Not Controversial

Group sequential includes interim data analysis, which requires adjusted α

Nearly all work for large samples (except Jennision and Turnbull, 1997)

Assumes known σ^2, so not adaptive to it.

Fundamental design goal centered on allowing early stopping.

Opinion: in practice mostly a technique to allow peeking at the data with a nearly powerless test so modest cost to expected sample size.

Internal Pilot Design Not Controversial

No interim data analysis, so no α cost for it (interim data analysis).

Small sample methods well developed for useful range of cases.

Can adjust sample size up or down for σ^2 for small α cost.

Other good choices. Please consult Coffey and Kairalla (2008).

Internal Pilot Steps for Univariate Gaussian Linear Model

1. Plan

1.1 Choose *design*, test, target test size α_t, power P_t, and means defining scientifically important effect.

1.2 Use σ_0^2 to pick n_0 target total and $n_1 = \pi \cdot n_0$ sizes.

1.3 Choose method for final $\widehat{\sigma}^2$ and decision rule. Use GLUMIP.

2. Conduct internal pilot

2.1 Collect n_1 observations, compute $\widehat{\sigma}_1^2$.

2.2 Power analysis finds $N_2 \geq 0$ observations to achieve P_t.

3. Complete study **3.1** Collect N_2 observations and **3.2** conduct (adjusted) analysis.

1.4 I Discourage Full Blinding

I recommend mean blinding but not mean and variance blinding.

Mean blinding: use design knowledge to compute model residual $\widehat{\sigma}_1^2$ but remain blind to $\widehat{\boldsymbol{\beta}}$ ("noprint" option).

Total blinding: ignore design knowledge to compute model residual $\widehat{\sigma}_1^2$. Recommended by some (Gould, et al.).

Why should internal pilots be treated differently than group sequential?

Many references in the bibliography.

Could start with Waksman (2007) to illustrate many issues.

2. Internal Pilots for a Gaussian Linear Model

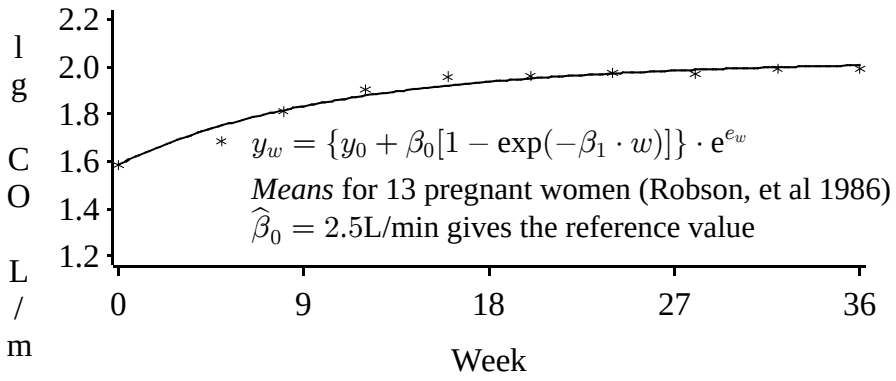
Step 1.1 Plan a Fixed Sample Size Study

Example Obstetrician Dr. Kirk Conrad plans to compare Cardiac Output (L/min) in 4 groups of women during pregnancy. Will record baseline (revised submission of P01 in review).

Fitting model as baseline + one compartment model in log space, corresponding to 7.4 maximum $= 4.9$ baseline $+ 2.5$ L/min.

50% change from baseline!

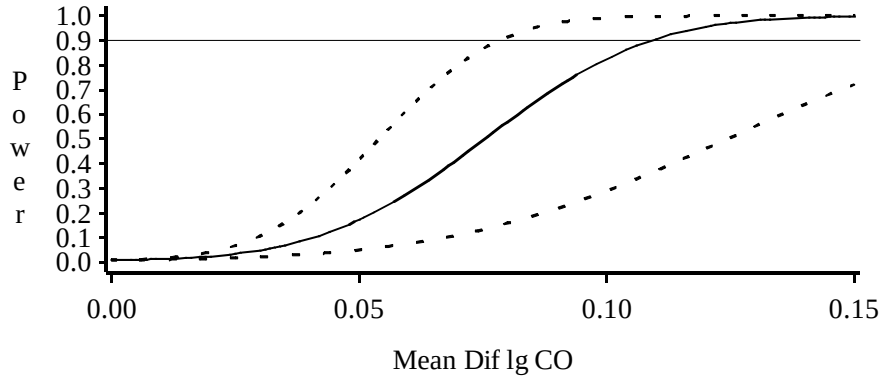
Fit a nonlinear model via transforming both sides with logarithm.



For $\mu_{0,A} = \log(\beta_0)$ *test* $H_0 : \mu_{0,A} = \mu_{0,B} = \mu_{0,C} = \mu_{0,D}$
Assuming $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}$ with $\boldsymbol{e} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$

Design Feature	Description
$\boldsymbol{X} = \boldsymbol{I}_4 \otimes \boldsymbol{1}_m$	cell mean design with m replicates
$\boldsymbol{\beta}_{\text{plan}} = [\,0.92\ 0.92\ 0.92\ 1.01\,]'$	Scientific important diff $\approx 10\%$
$H_0 : \boldsymbol{C\beta} = \boldsymbol{\theta}_0$	Test by $\boldsymbol{C} = [\,\boldsymbol{1}_3\ -\boldsymbol{I}_3\,], \boldsymbol{\theta}_0 = \boldsymbol{0}$
α_t	target type I $= 0.05/$# primary outcomes
P_t	target power $= 0.90$
σ_0^2 here a SWAG	$(0.04)^2$ log(L/min) $\Rightarrow n_0 = 20$

Specifying $\{\boldsymbol{X},\boldsymbol{C}, \boldsymbol{\beta}, \boldsymbol{\theta}_0, \alpha_t, \sigma_0^2\}$ and P_t implies $n_0 = 24$



Step 1.2 Choose IP, *Guess* $\sigma_0^2 \Rightarrow n_0$ *Target Total and* $n_1 = \pi \cdot n_0$.

Yikes! If do interim power analysis at $n_1 = 12$ for $\pi = 0.50$ will have 3 women per group. What is your lowest choice?

Remember, we will fit a separate pharmacokinetic model to 7 values (from echo cardiograms) of Cardiac Output, and test mean $\log(\widehat{\beta}_0)$.

Consider recommending $n_1 = 16$ (4 per group) for $\pi = 0.67$ (so 3 per group at interim power analysis).

A beautiful part of an internal pilot: we have done nothing different so far! However, I have far less stress about the SWAG σ_0^2.

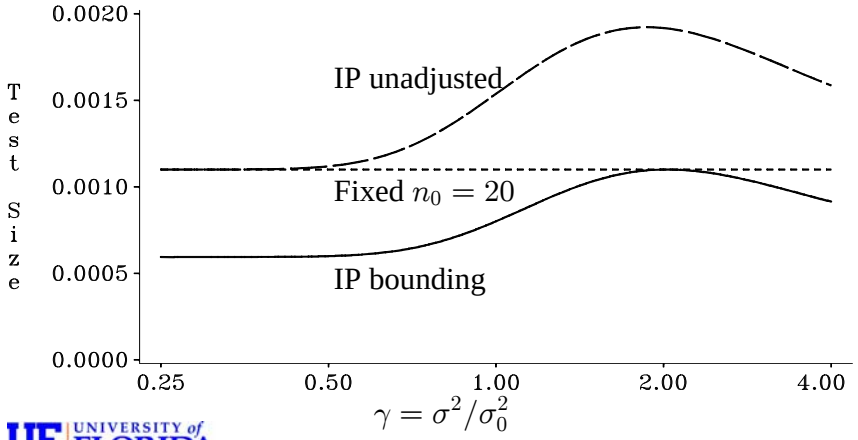
Step 1.3 Choose IP Methods: Choice for Final $\widehat{\sigma}^2$ and Decision Rule.

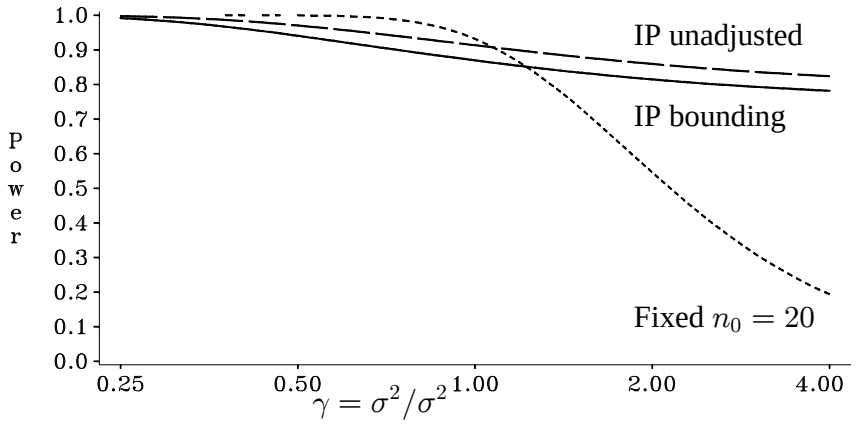
Choice # 1: how will σ^2 be estimated?

Using $\widehat{\sigma}^2_+$ from the total sample can inflate type I error rate.

The amount of inflation varies with the parameter $\gamma = \sigma^2/\sigma_0^2$

Following plots are type I error rate and power as a function of $\gamma = \sigma^2/\sigma_0^2$ for $\alpha_t = 0.0011$, $P_t = 0.90$, $n_1 = 10$, $n_0 = 20$, $n_{+,\text{min}} = 10$, $n_{+,\text{max}} = \infty$





Choices include $\widehat{\sigma}^2_1$ from first sample only
$\widehat{\sigma}^2_\perp$, information added by second sample only
$\widehat{\sigma}^2_w$ a weighted value to make unbiased
$\widehat{\sigma}^2_+$, total sample with correction in small samples

I recommend a *bounding method*: use $\widehat{\sigma}^2_+$ and critical value for $\alpha_* \leq \alpha_t$ for α_t = nominal type I error rate.
(Coffey and Muller 2001; Coffey, Kairalla, and Muller 2007)

Free SAS/IML code (GLUMIP version 2.0) at *http://www.jstatsoft.org/v28/i07* will give you α_*

Step 1.3 Finding $\alpha_* \leq \alpha_t$ for Our Design

Adjusted Alpha $\alpha_* = 0.0338$ (20 lines of IML code)

```
PROC IML WORKSIZE=8000 SYMSIZE=8000;
   %INCLUDE "&PROGPATH.\GLUMIP20.IML" / NOSOURCE2;
   ALPHAT  = .05;  POWERT  = .90;
   ESSENCEX = I(4);   C  = J(3,1,-1)||I(3);
   BETA_PLN = {0.92 , 0.92 , 0.92 , 1.01};
   SIGMA0  = ({.04})**2;
   N1 = {12};   NPLUSMIN = N1;   NPLUSMAX = 36;
   RUN FINDADJ;
 PRINT _FINDADJ[COLNAME=_FINDADJNM];
```


Step 2. Conduct Internal Pilot

Step 2.1 collect n_1 observations, compute $\widehat{\sigma}_1^2$

Step 2.2 Power analysis finds $N_2 \geq 0$ observations needed to achieve P_t

Use standard power software, so very convenient.

The approach does create an "alignment error" with actual power, but modest. Improvements may be developed in future research.

Even here simulations a big, ugly, and very slow bear.

Step 3.2. *Complete the Study*

Step 3.1 Collect N_2 observations

Step 3.2 Conduct analysis with (adjusted) α_* if using bounding, using standard software.

Most other methods require some fiddling with variance estimates to piece together test statistics, but just simple programming (but be careful:)

REVIEW: Internal Pilot Steps for Univariate Gaussian Linear Model

1. Plan Choose design and analysis as usual.
Choose internal pilot size features.
Use GLUMIP for bounding method and refining design.

2. Internal pilot Collect n_1 observations, compute $\widehat{\sigma}_1^2$
Power analysis $\Rightarrow$ $N_2 \geq 0$ observations

3. Complete study Collect N_2 observations
Conduct analysis with *adjusted* α

3. Complications and Future Work

1. Glossed over many issues around confidence intervals, stepdown tests. Answers depend on variance estimator. Does α_* suffice?

2. IP for other nuisance parameters? Some large sample approximations.

3. Random predictors, such as fractions in a blocking variable?

4. Non inferiority in small samples?! Caveat emptor, in small samples.

5. IP for large samples? Easy, should typically do it at no α cost. Common complaint centers on logistical barriers, funding. However, N_+max most of fix, as in group sequential.

6. Research in Review and in Progress
   a) IP for some repeated measures; required theory and software coming
   b) small sample IP with 2 stage group sequential

Summary Review

1. Motivation: Good Design and Analysis with Uncertainty About σ^2

2. Internal Pilots for a Gaussian Linear Model:
   EASY TO DO.
   Not controversial.
   Small sample valid methods and free software available for many cases.
   Gives power insurance and cost protection.

3. Complications and Variations

I close by recommending the bibliography and online **scholarly** searches.

Banish Power Uncertainty:
USE AN INTERNAL PILOT DESIGN!

Brief bibliography for IP and power for Gaussian linear models; please see references in sources and use CIS.

Free SAS/IML code (GLUMIP version 2.0) for IP design and analysis at *http://www.jstatsoft.org/v28/i07*

****web site *http://ehpr.ufl.edu/muller* links to most of the following Muller articles and power software***

General references for IP and power.

Coffey C.S. and Kairalla J.A. (2008) Adaptive clinical trials: progress and challenges. *Drugs in R&D*, **9**, 229-242.

Jennison, C and Turnbull, BW (2006) Adaptive and nonadaptive group sequential tests *Biometrika*, **93**, 1-21.

Jennison, C and Turnbull, B W. (1997) Distribution theory of group sequential t chi sq and F-tests for general linear models, *Sequential Analysis*, **16**, 295-317.

Muller K.E. and Fetterman B.A. (2002) *Regression and ANOVA: An Integrated Approach Using SAS® Software*. Cary, NC: SAS Institute.

Muller K.E. and Stewart P.W. (2006) *Linear Model Theory; Univariate, Multivariate, and Mixed Models*. NY Wiley.

Power and IP articles with collaborators

Kairalla J.A., Muller K.E., and Coffey C.S. (2010) Combining an internal pilot with an interim analysis for single degree of freedom tests, *Communications in Statistics, Theory and Methods*, in press.

Johnson J.L., Muller K.E., Slaughter J.C., Gurka M.J., Gribbin M.J., and Simpson S.L. (2008) POWERLIB: SAS/IML software for computing power in multivariate linear models. *Journal of Statistical Software*, **30**(5) 1-27. http://www.jstatsoft.org/v30/i05.

Kairalla J.A., Coffey C.S., and Muller K.E. (2008) GLUMIP 2.0: SAS/IML software for planning internal pilots. *Journal of Statistical Software*, **28**(7), 1-32, http://www.jstatsoft.org/v28/i07.

Gurka M.J., Coffey C.S. and Muller K.E. (2007) Internal pilots for a class of linear mixed models with Gaussian and compound symmetric data, *Statistics in Medicine*, **26**, 4083–4099.

Muller K.E., Edwards L.J., Simpson, S.L. and Taylor D.J. (2007) Statistical tests with accurate size and power for balanced linear mixed models, *Statistics in Medicine*, **26**, 3639–3660.

Coffey C.S., Kairalla J.A. and Muller K.E. (2007) Practical methods for bounding type I error rates with an internal pilot design, *Communications in Statistics - Theory and Methods*, **36**, 2143 - 2157.

Jiroutek M.R., Muller K.E., Kupper L.L. and Stewart P.W. (2003) A new method for choosing sample size for confidence interval-based inferences, *Biometrics*, **59**, 580-590.

Coffey C.S. and Muller K.E. (2003) Properties of internal pilots with the univariate approach to repeated measures, *Statistics in Medicine*, **22**, 2469-2485.

Glueck DH and Muller KE (2003) Adjusting power for a baseline covariate in a linear model *Stat in Med* **22** 2535-2551.

Taylor D.J., Kupper L.L. and Muller K.E. (2002) Improved approximate confidence intervals for the mean of a lognormal distribution, *Statistics in Medicine*, **21**, 1443-1459.

Coffey C.S. and Muller K.E. (2001) Controlling test size while gaining the benefits of an internal pilot design, *Biometrics*, **57**, 625-631.

Coffey C.S. and Muller K.E. (2000) Some distributions and their implications for an internal pilot study with a univariate linear model, *Communications in Statistics - Theory and Methods*, **29**, 2677-2691.

Coffey C.S. and Muller K.E. (2000) Properties of doubly-truncated gamma variables, *Communications in Statistics - Theory and Methods*, **29**, 851-857.

Coffey C.S. and Muller K.E. (1999) Exact test size and power of a Gaussian error linear model for an internal pilot study, *Statistics in Medicine*, **18**, 1199-1214.

********Articles by others about blinding in internal pilots***********

Gould, A. Lawrence (1997) Issues in blinded sample size re-estimation, *Comm. in Stat: Sim and Comp*, **26**, 1229-1239

Gould, A. L. and Shih, W. J. (2005) Comment on ``On the inappropriateness of an EM algorithm based procedure for blinded sample size re-estimation" (2002V21 p165-176) *Statistics in Medicine*, **24**, 147-154.

Friede, Tim and Kieser, Meinhard (2003) Blinded sample size reassessment in non-inferiority and equivalence trials, *Statistics in Medicine*, **22**, 995-1007.

Friede, Tim and Kieser, Meinhard (2002) On the inappropriateness of an EM algorithm based procedure for blinded sample size re-estimation, *Statistics in Medicine*, **21**, 165-176

Kirby, S. McBride, S. Puvanarajan L. (2003) An example of an unblinded, third-party interim analysis for sample size re-estimation, *Drug Information Journal*, **37**, 17–320.