ACCOUNTING FOR ALIGNMENT, UNCERTAINTY AND BIAS IN CHOOSING A SAMPLE SIZE

Michael R. Jiroutek[1] and Keith E. Muller[2]

[1]Research Biostatistician
Bristol-Myers Squibb Pharmaceutical Research Institute
email: michael.jiroutek@bms.com

[2]Associate Professor, Dept. of Biostatistics,
University of North Carolina at Chapel Hill

Connecticut ASA Mini-Conference
March 13, 2004

Talk based largely on:

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 statistical inferences, *Biometrics* **59**, 580-590.

Jiroutek, M. R. and Muller, K. E. (2004) Uncertainty and bias in sample size due to estimating variance when using confidence interval criteria, in review.

OVERVIEW

I. Motivating Example

II. Aligning Sample Size Rule with Study Goals

III. Uncertainty in Chance of Success
Due to Estimating Variance

IV. Bias in Estimated Chance of Success
Due to Truncation

V. Extensions

I. MOTIVATING EXAMPLE

The ***BIG*** question: *Why do so many successful phase IIs lead to disappointing phase IIIs?*

Many factors.

Three problems we can help solve:

1) *Misalignment* between sample size calculation and study objectives.
2) *Uncertainty* in the variance value used for planning.
3) *Bias* due to proceeding only after a significant result.

Focus here on power and power generalization for studies including Confidence Intervals (CIs).

Sample size goals include:

Width (W): CI is as narrow as desired
Validity (V): CI contains true unknown parameter
Rejection (R): of the null hypothesis

Example: Pisano, et al. (2002) *screening* study:

Radiologists read mammograms on film (hardcopy) and computer screen (softcopy).

Is softcopy read faster or slower than hardcopy?

Screening study results suggestive, would like to conduct *target* study.

Choose sample size for target study with CI endpoint.

Board certified radiologists are busy, expensive.

II. ALIGNING SAMPLE SIZE RULE WITH STUDY GOALS

For Pisano, et al. example:

Use screening data to plan target study.

Increase reading time of $<25\%$ acceptable $\Leftrightarrow$ $\log_{10}$ scale CI width of $\delta = 0.125$.

Test $H_0 : \theta = 0$ *vs.* $H_a : \theta \neq 0$ of difference.

$\theta = \delta/2 = 0.0625; \quad \alpha = 0.05.$

$\widehat{\sigma}_s^2 = 0.012$. For now, assume it's population value.

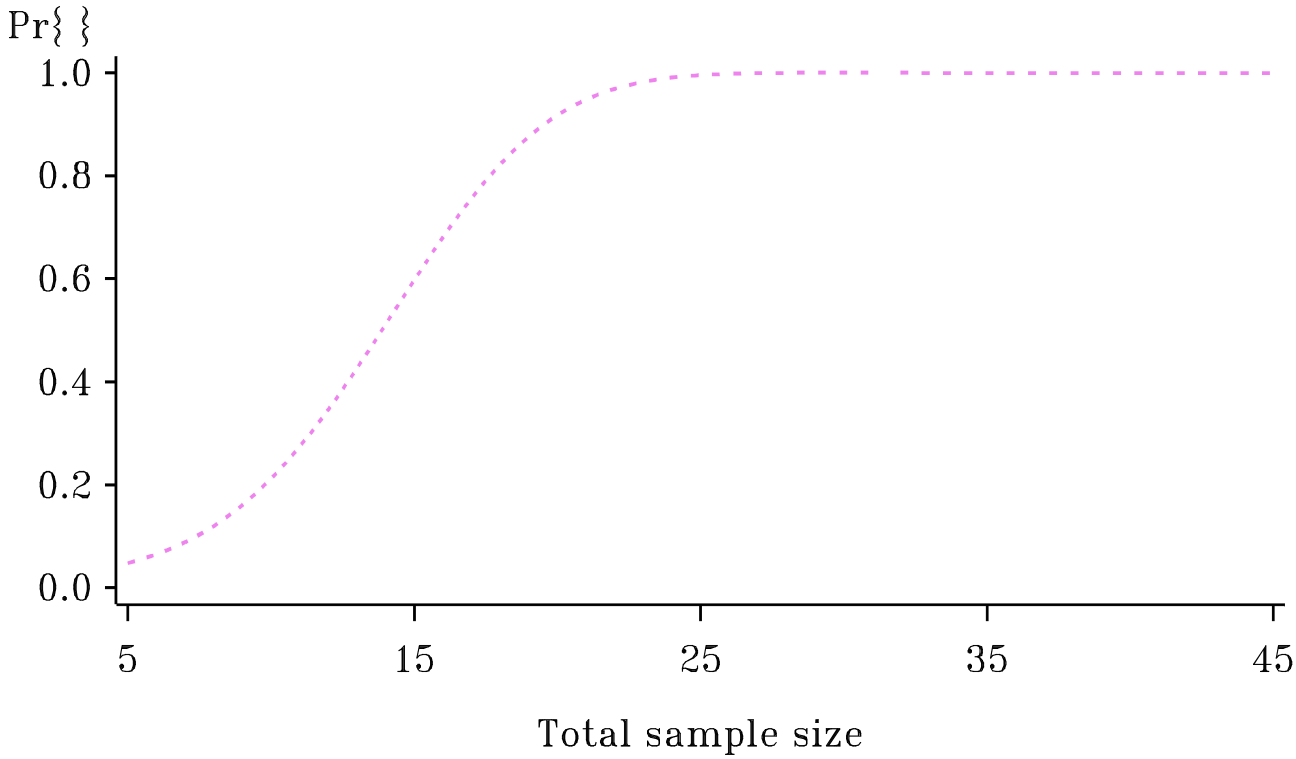

Figure 1. $\Pr\{W|V\}$ curve for target study.

$\geq$ 0.90 target probability $\Rightarrow n = 20$

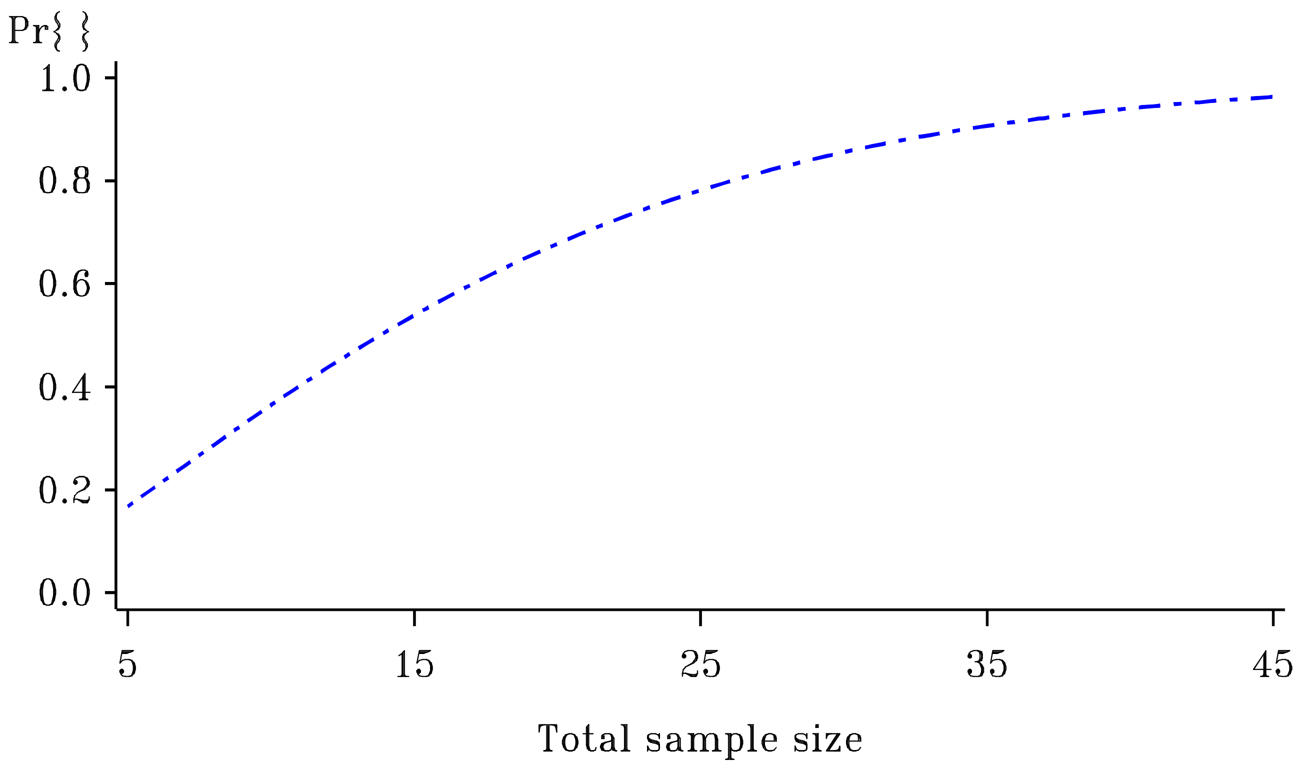

Figure 2. $\Pr\{R\}$ curve for target study.

$\Pr\{R\} \Leftrightarrow$ unconditional power

≥ 0.90 power $\Rightarrow n = 35$

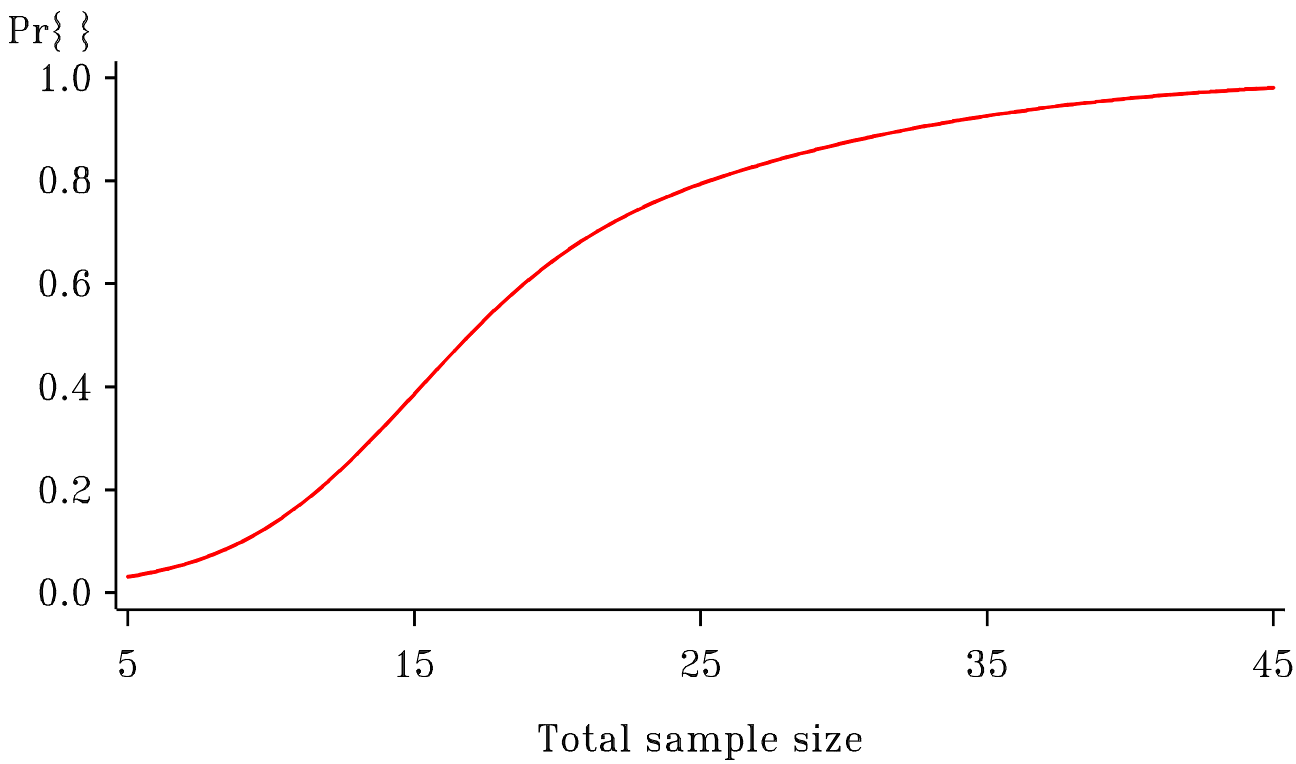

Figure 3. $\Pr\{(W \cap R)|V\}$ curve for target study.

$\geq$ 0.90 target probability $\Rightarrow n = 33$

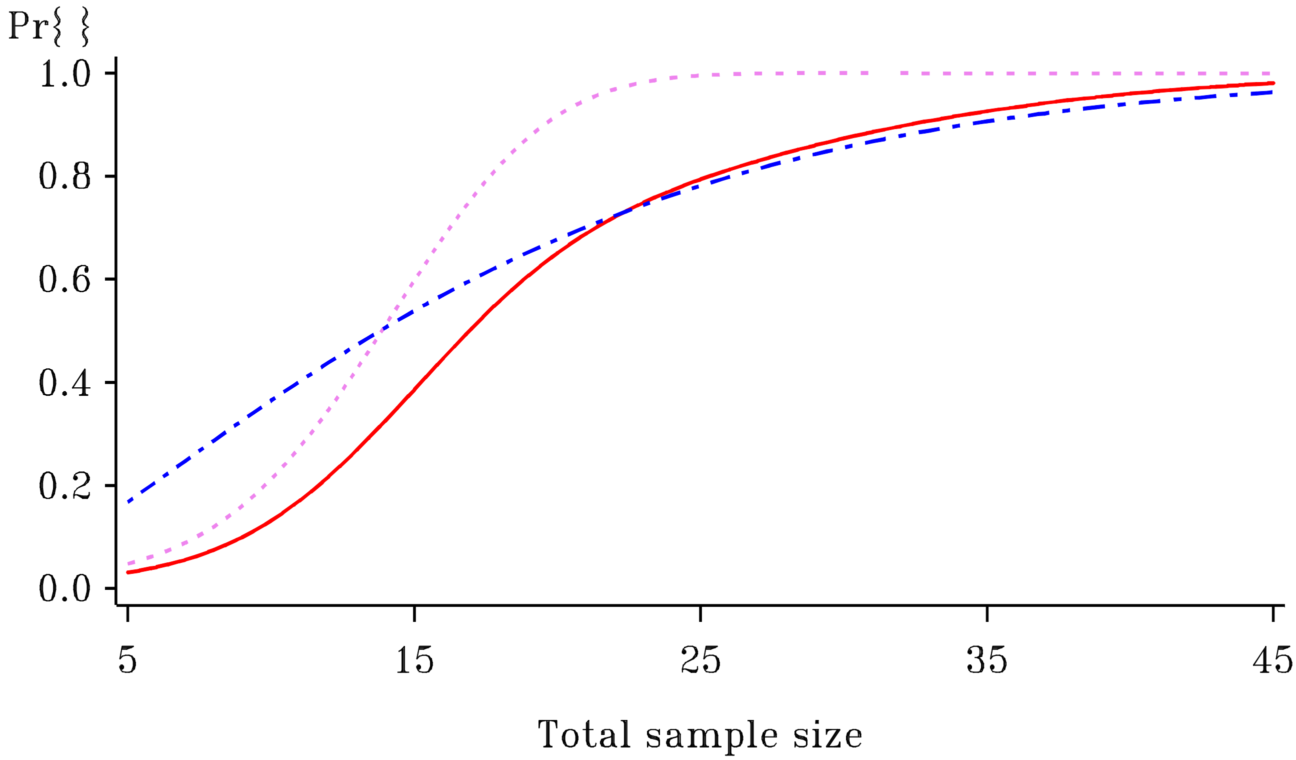

Figure 4. $\Pr\{(W \cap R)|V\}$: solid line, $\Pr\{R\}$: dashed line and $\Pr\{W|V\}$: dotted line curves for target study.

Relative size of CI width to test parameter most important.

Different examples than Pisano, et al. (2002); see Jiroutek (et al., 2003):

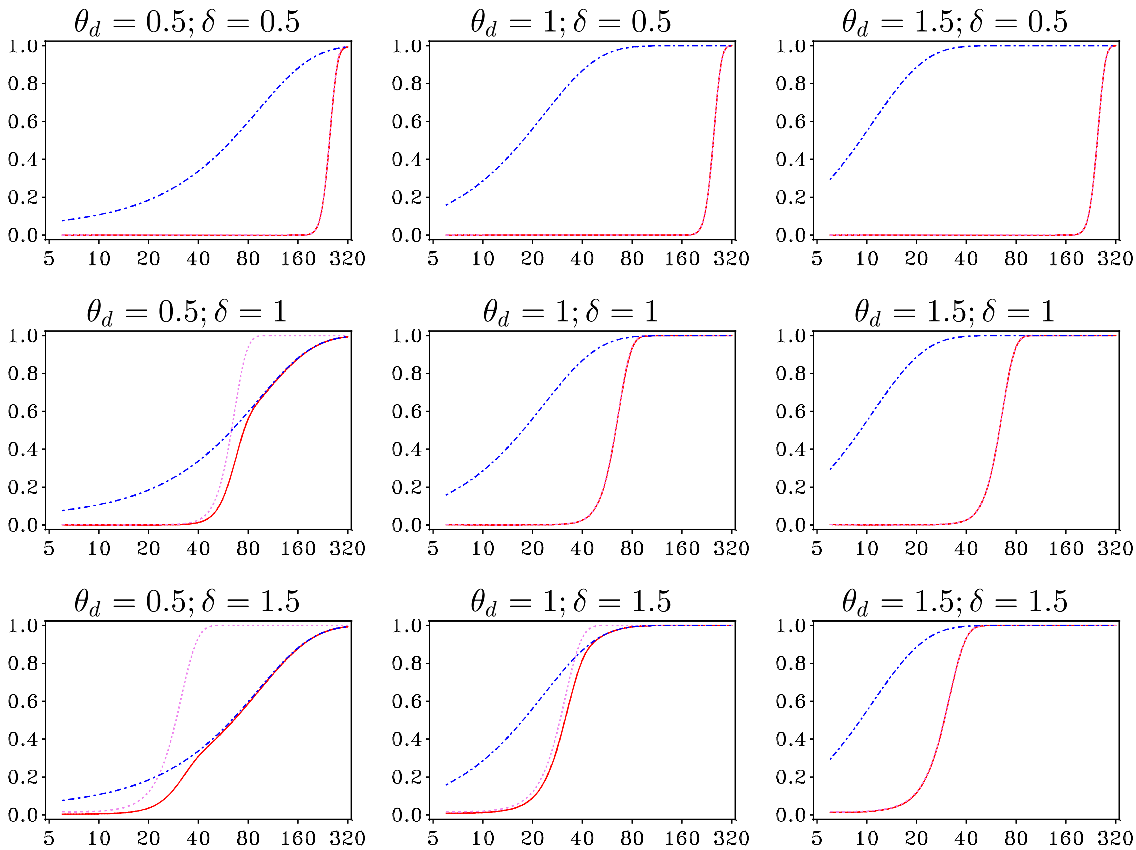

Figure 5. Event probabilities as a function of n with $\log_2$ spacing, $\nu_e = N - r, r = 2, \sigma^2 = 1, \theta_0 = 0$ and $\alpha = 0.05$. $\Pr\{(W \cap R) | V\}$: solid line; $\Pr\{R\}$: dashed line; $\Pr\{W | V\}$: dotted line.

Note: $\theta_d = \theta - \theta_0$: parameter of interest

δ : CI width

Alignment Conclusions

• Jiroutek, et al. concluded $\Pr\left\{(W \cap R)\middle|V\right\}$ best aligned sample size with scientific goals.

• New exact small sample results apply to any scalar parameter in General Linear Multivariate Models (GLMM). Includes

– one and two sample t-tests
– paired-data t-test
– planned scalar contrasts in univariate, multivariate or REPM ANOVA

• Unconditional power $\Leftrightarrow$ $\Pr\{R\}$ and $\Pr\left\{W\middle|V\right\}$ are special cases of $\Pr\left\{(W \cap R)\middle|V\right\}$.

III. UNCERTAINTY IN CHANCE OF SUCCESS DUE TO ESTIMATING VARIANCE

Refer to P_t as target probability, (e.g., power, $\Pr\left\{W \middle| V\right\}$, $\Pr\left\{(W \cap R) \middle| V\right\}$).

Ignored in previous results: Variance *estimate* from screening study used.

How to account for using $\widehat{\sigma}^2$ in place of σ^2?

Type I & II error rates, scientifically important difference, and CI width all specified.

How is $\widehat{\sigma}^2$ obtained?

- Guess
- Limited by financial, temporal or other constraints
- Best/most frequent case: Prior data

Use of $\widehat{\sigma}^2$ (not σ^2) from pilot study, other study, literature $\Rightarrow$ random not fixed.

P_t inherits randomness.

Suggests use of confidence bounds for P_t curve.

P_t a smooth, strictly monotone, 1-to-1 function of $\sigma^2 \Rightarrow$ exact CI follows from exact CI for σ^2.

Compute $\left(\widehat{\sigma}^2_{sL}, \widehat{\sigma}^2_{sU}\right)$.

Replace $\widehat{\sigma}^2_s$ in P_t calculation.

Compute $\left(\widehat{P}_{tL}, \widehat{P}_{tU}\right)$.

Pisano, et al. (2002) study (variation, **larger $\boldsymbol{\delta}$**):

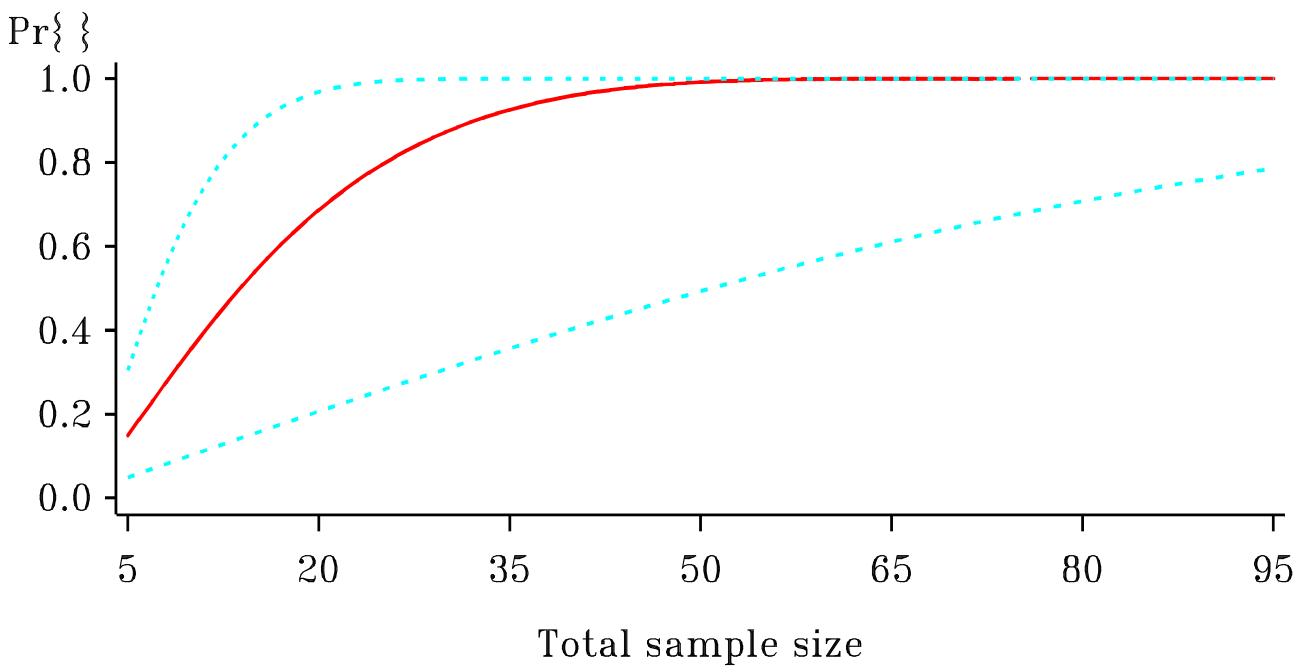

Figure 6. 95% confidence region (dots) for $\Pr\{(W_t \cap R_t)|V_t\}$ (solid) based on $\widehat{\sigma}_s^2 = 0.012; \theta = 0.0625;$ $\boldsymbol{\delta_s = \delta_t = 1.5;}$ $n_s = 8$.

Wide bands due to small n_s.

Confidence region for power (GLUM): Taylor & Muller (1995). Extended to $\Pr\left\{(W \cap R)\middle|V\right\}$ in GLMM by Jiroutek & Muller (2004, in review).

Uncertainty Conclusions

- Screening study sample size more important than target study sample size!
- We believe this explains an important fraction of failures in replicating studies.
- New exact small sample results apply to any scalar parameter in GLMM.

IV. BIAS IN ESTIMATED CHANCE OF SUCCESS DUE TO TRUNCATION

Ignored in previous results: Target study conducted only if screening study successful.

Same in drug discovery process: Ph II (III) trial occurs only after *significant* Ph I (II) result.

Studies with small $\widehat{\sigma}^2$ by chance more likely successful.

Only early studies with sufficiently small variability will lead to later phase studies.

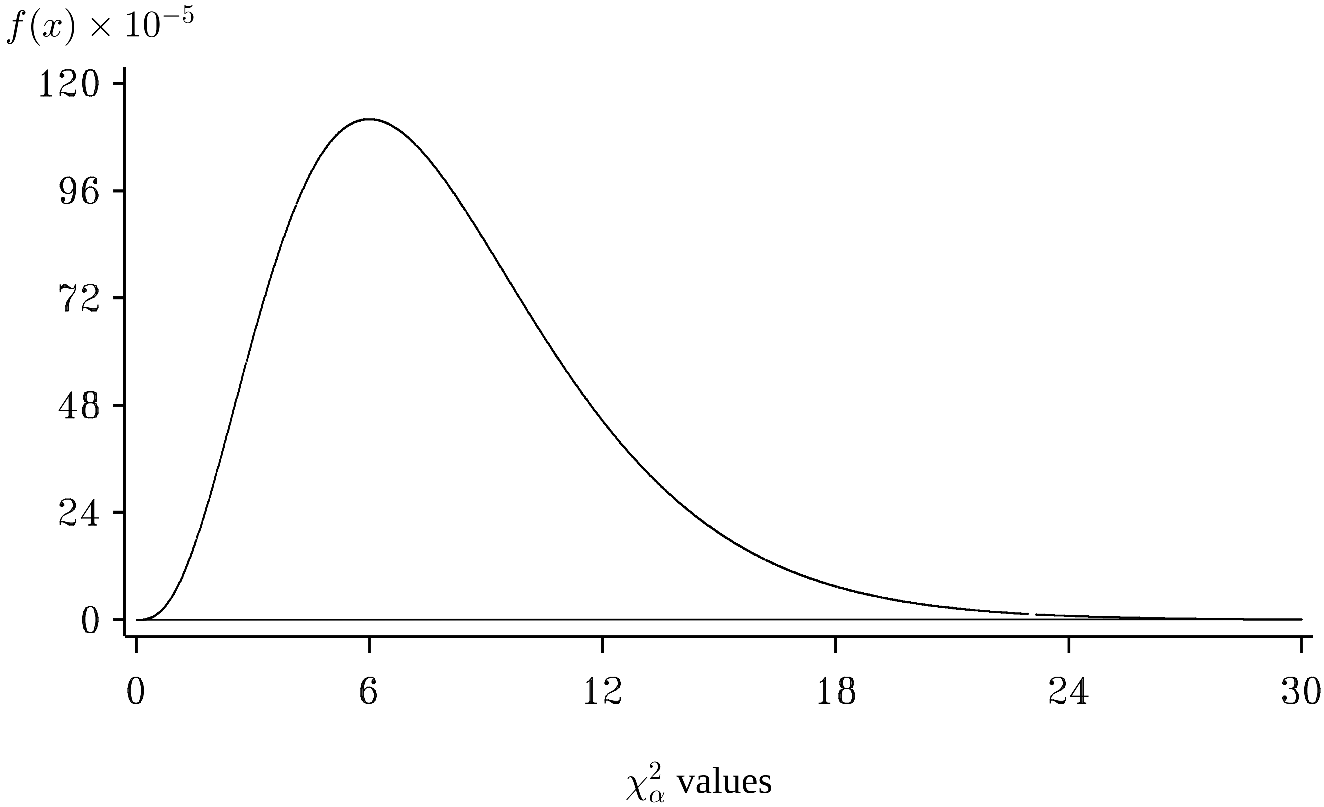

Figure 7. Example distribution of $\widehat{\sigma}_s^2$ (χ^2_α, eight *df*).

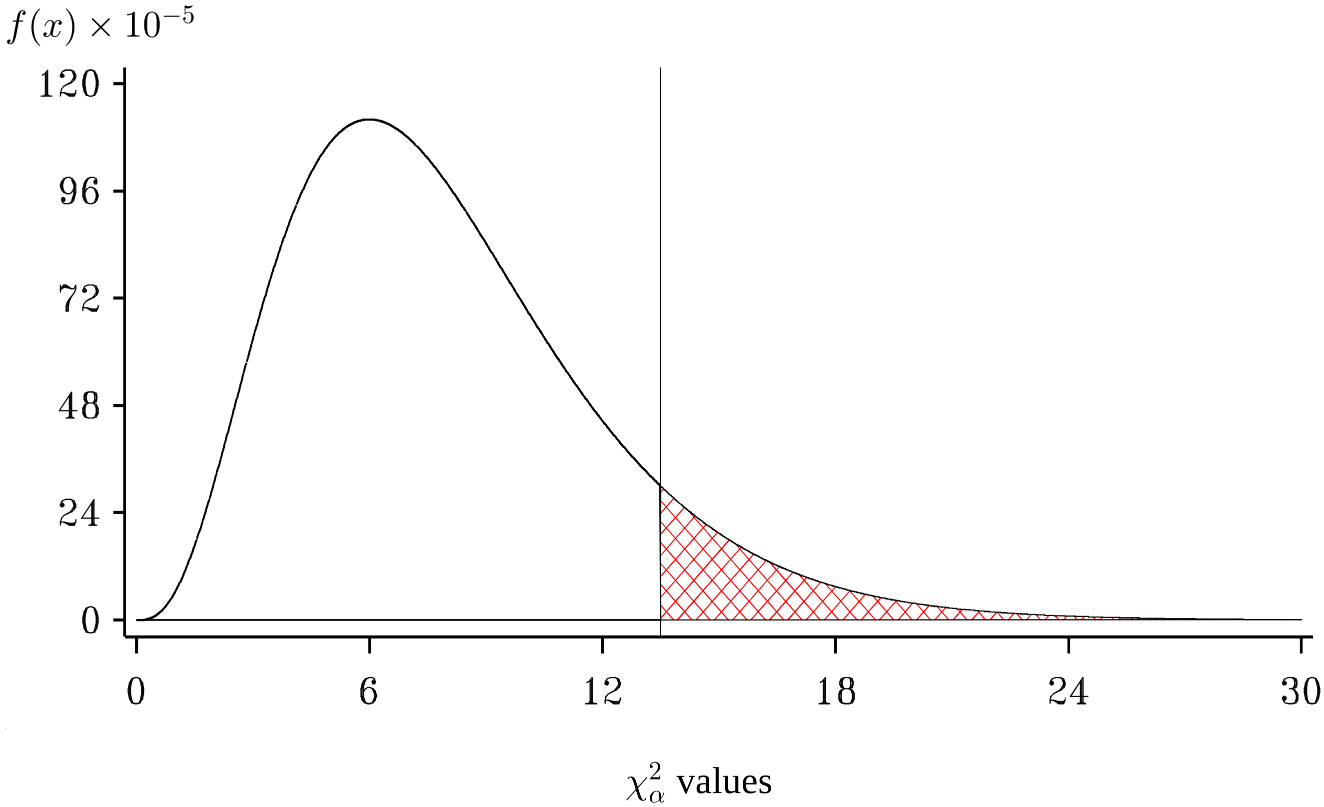

Figure 8. Example distribution of $\widehat{\sigma}_s^2$ (χ^2_α, eight *df*) with truncation point, highlighting failure region.

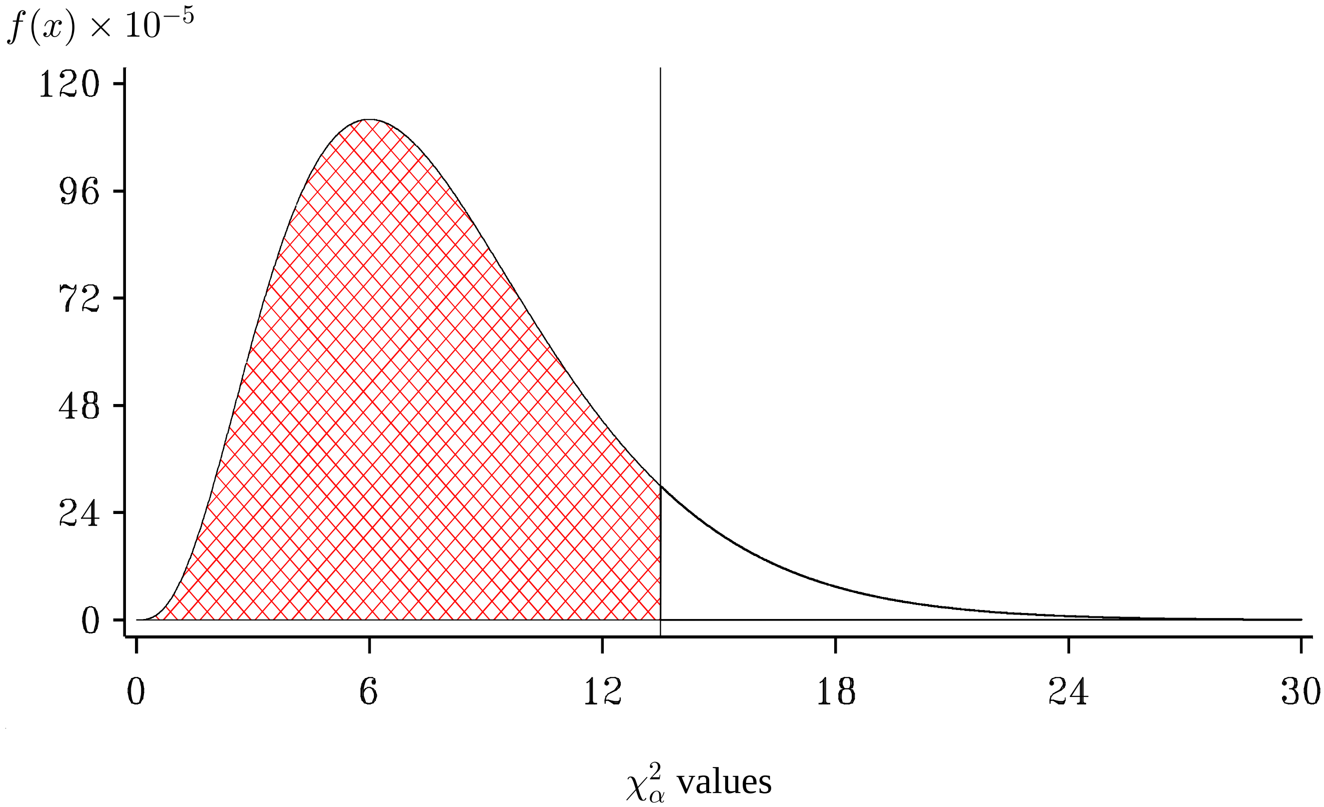

Figure 9. Example distribution of $\widehat{\sigma}^2_s$ (χ^2_α, eight *df*) with truncation point, highlighting success region.

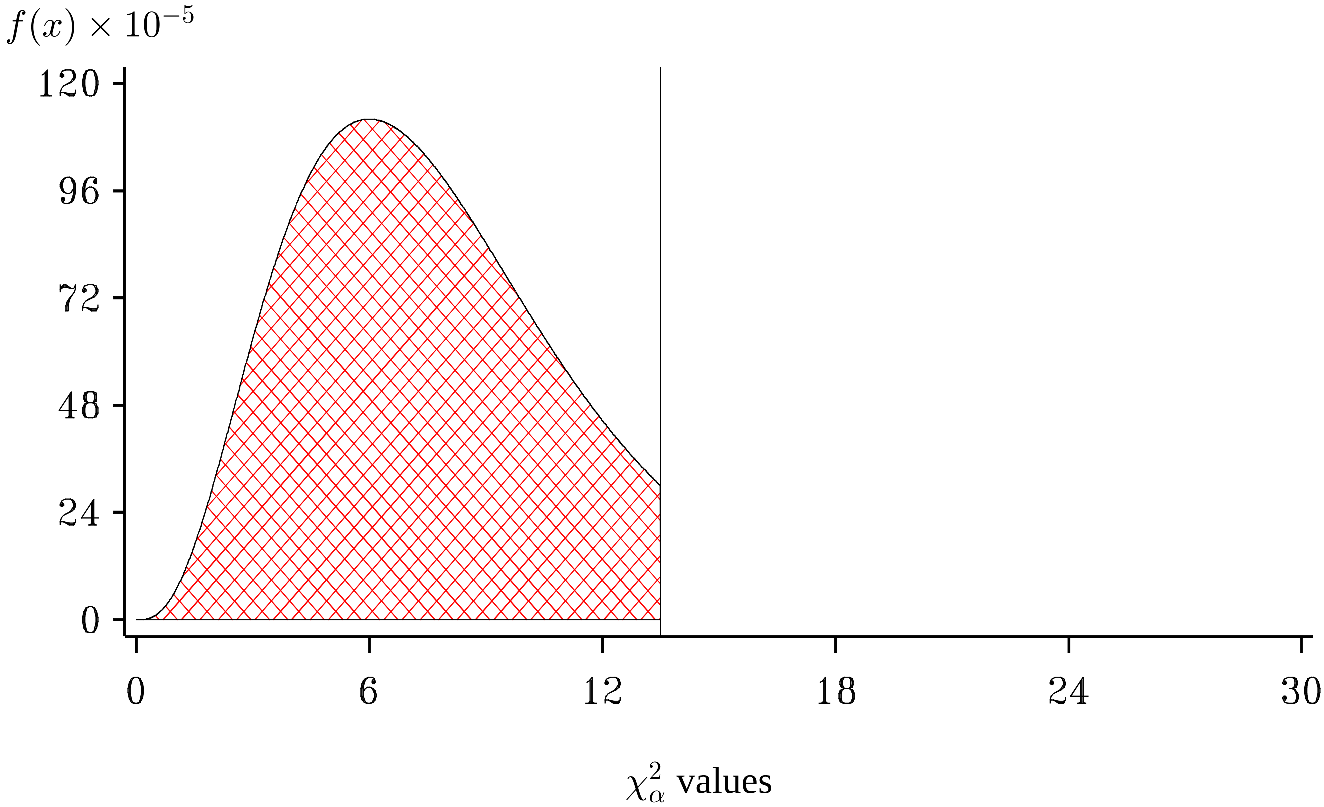

Figure 10. Example of “success truncated” distribution of $\widehat{\sigma}^2_s$ (χ^2_α, eight *df*).

Distribution of sufficiently small $\widehat{\sigma}^2$ different than that of all $\widehat{\sigma}^2$.

"Success truncation" describes this effect on PDF (CDF) of $\widehat{\sigma}_s^2$.

Under normality, $\widehat{\sigma}_s^2$ a truncated, scaled χ^2.

Truncation occurs as a result of observing only $\widehat{\sigma}_s^2$ that achieve pre-specified criteria.

Muller & Pasour (1997) derived exact expression for truncated CDF of $\widehat{\sigma}_s^2$ for power.

Jiroutek and Muller (2004, in review) extended to Pr$\left\{(W \cap R) \middle| V\right\}$, while considering better aligned truncation.

Impact on P_t?

For power, success truncation occurs when screening study hypothesis test significant.

For $\Pr\left\{(W \cap R)\middle|V\right\}$, success truncation occurs when screening study hypothesis test significant *and* CI width achieved.

Estimated P_t computed with $\widehat{\sigma}_s^2$ (truncated or not).

Exact CI for estimated probability criterion based on truncated $\widehat{\sigma}_s^2$: replace untruncated $\widehat{\sigma}_s^2$ bounds with appropriate truncated values.

Remaining inputs fixed constants, may or may not coincide with screening study values.

Recall, **Figure 6** for variation of Pisano, et al. (2002) study:

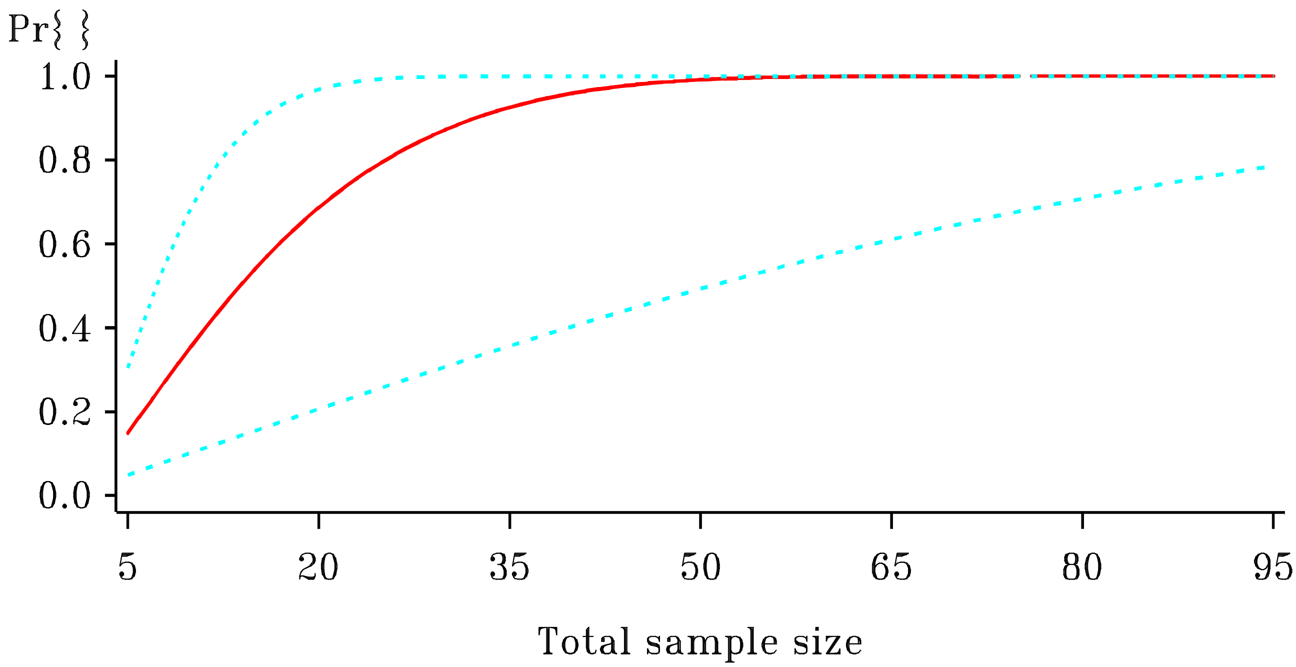

Figure 11. 95% confidence region (dots) for $\Pr\{(W_t \cap R_t)|V_t\}$ (solid) based on $\widehat{\sigma}_s^2 = 0.012$; $\beta = 0.0625$; $\boldsymbol{\delta_s = \delta_t = 1.5}$; $n_s = 8$.

If Pisano, et al. screening study significant:

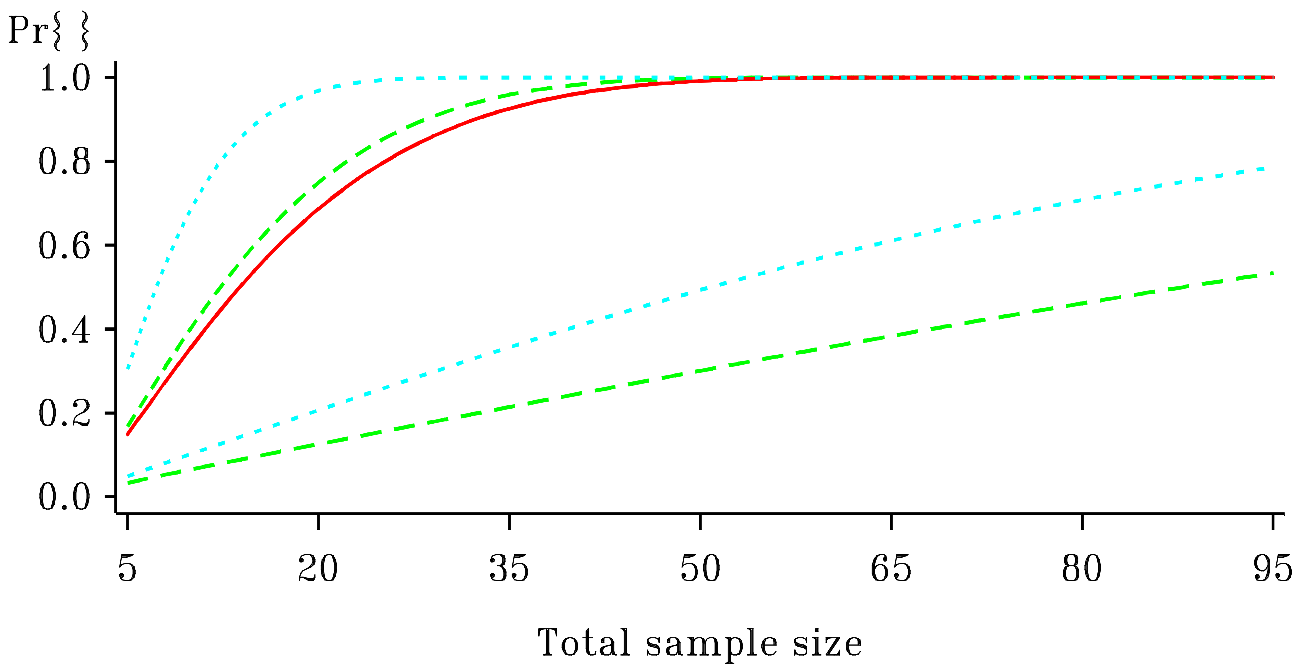

Figure 12. 95% success truncation (dashes) and no-truncation (dots) confidence regions for $\Pr\{(W_t \cap R_t)|V_t\}$ (solid) based on $\widehat{\sigma}_s^2 = 0.012$; $\beta = 0.0625$; $\boldsymbol{\delta_s = \delta_t = 1.5}$; $n_s = 8$.

Bias occurs if **success truncation ignored $\Rightarrow$ optimistic bias** and **sample size too small.**

Wide bands due to small n_s.

Bias Conclusions

• New exact small sample results account for success truncation in analysis of any scalar parameter in GLMM.

• Ignoring success truncation causes optimistic bias when computing sample size.

• Correcting sample size eliminates bias, should lead to more successes.

• We believe this explains another important fraction of failures in replicating studies.

• In non-GLMM, if using (asymptotically) Gaussian test, above results may apply.

• “failure truncation” creates **pessimistic bias** and **sample size too big.**

V. EXTENSIONS

Work in progress:

User-friendly freeware for $\Pr\left\{(W \cap R)\middle|V\right\}$ (Figure 5). Uncertainty, bias extensions to follow.

Internal Pilot Designs (interim power analysis).

Important unanswered questions:

Group sequential designs.

Binomial data. More complex due to dependence between mean and variance.

Exponential data.

REFERENCES

Beal, S. L. (1989) Sample size determination for confidence intervals on the population mean and on the difference between two population means, *Biometrics*, **45**, 969-977.

Bristol, D. R. (1989) Sample sizes for constructing confidence intervals and testing hypotheses, *Statistics in Medicine*, **8**, 803-811.

Coffey CS, Muller KE. Properties of doubly-truncated gamma variables. *Communications in Statistics - Theory & Methods* 2000; **29**:851-857.

Gatsonis, C. and Sampson, A. R. (1989) Multiple correlation: exact power and sample size calculations, *Psychological Bulletin*, **106**(3), 516-524.

Glueck, D. H. (1995) Power for a generalization of the GLMM with fixed and random predictors, Ph.D. dissertation, Department of Biostatistics, University of North Carolina, Chapel Hill.

Grieve, A. P. (1991) Confidence intervals and sample sizes, *Biometrics*, **47**, 1597-1603.

Hsu, J. C. (1989). Sample size computation for designing multiple comparison experiments. *Computational Statistics & Data Analysis* **7**, 79-91.

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 statistical inferences, *Biometrics* **59**, 580-590.

Jiroutek, M. R. and Muller, K. E. (2004). Uncertainty and bias in sample size due to estimating variance when using confidence interval criteria, in review.

Kupper, L. L. and Hafner, K. B. (1989) How appropriate are popular sample size formulas?, *American Statistician*, **43**(2), 101-105.

Lehmann, E. L. (1959) Testing Statistical Hypotheses. Wiley; New York.

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician* **55**, 187-193.

Leventhal, L. and Huynh, C. (1996). Directional decisions for two-tailed tests: power, error rates, and sample size. *Psychological Methods* **1**(3), 278-292.

Muller, K. E., LaVange, L. M., Ramey, S. L and Ramey, C. T. (1992) Power calculations for general linear multivariate models including repeated measures applications, *Journal of the American Statistical Association*, **87**(420), 1209-1226.

Muller, K. E. and Pasour, V. B. (1997). Bias in linear model power and sample size due to estimating variance. *Communications in Statistics - Theory & Methods* **26**(4), 839-851.

Pisano, E. D., Cole, E. B., Kistner, E. O., Muller, K. E., Hemminger, B. M., Brown, M., Johnston, R. E., Kuzmiak, C., Braeuning, M. P., Freimanis, R., Soo, M. S., Baker, J. and Walsh, R. (2002). Interpretation of digital mammograms: a comparison of speed and accuracy of softcopy versus printed film display. *Radiology* **223**, 483-488.

Sampson, A. R. (1974) A tale of two regressions, *Journal of the American Statistical Association*, **69**(347), 682-689.

Taylor, D. J. and Muller, K. E. (1995). Computing confidence bounds for power and sample size of the general linear univariate model. *American Statistician* **49**(1), 43-47.

Taylor, D. J. and Muller, K. E. (1996). Bias in linear model power and sample size calculation due to estimating noncentrality. *Communications in Statistics: Theory & Methods* **25**, 1595-1610.