Biopharmaceutical Applied Statistics Symposium XIV, Savannah, GA

Statistical Validation of Surrogate Markers

Yongming Qu, PhD

Senior Research Scientist, Biostatistics
Eli Lilly and Company
Indianapolis, Indiana

qu_yongming@lilly.com

Nov 5, 2007

Part I

Which Quantity to be Used?

Outline

- Introduction
  - Surrogate endpoint
  - Surrogate marker
  - Intermediate marker
- Statistical validation quantities
  - Proportion of Treatment Effect (PTE)
  - Likelihood Reduction Factor (LRF)
  - Proportion of Information Gain (PIG)
- Simulations
- An example

Surrogate Endpoint (SE)

- Surrogate endpoint is intended to replace clinical outcome for any therapy
- Reason why validating surrogate endpoint is not feasible
  - Surrogate endpoint needs to be validated
  - To evaluate the surrogate endpoint, large confirmatory clinical trials need to be conducted for both surrogate and clinical endpoints
  - If large confirmatory clinical trials are conducted, the drug efficacy should have been established.
  - No need for surrogate endpoint for this drug
  - The conclusion from this drug cannot be extrapolated to other drugs because different drugs may work through different pathways

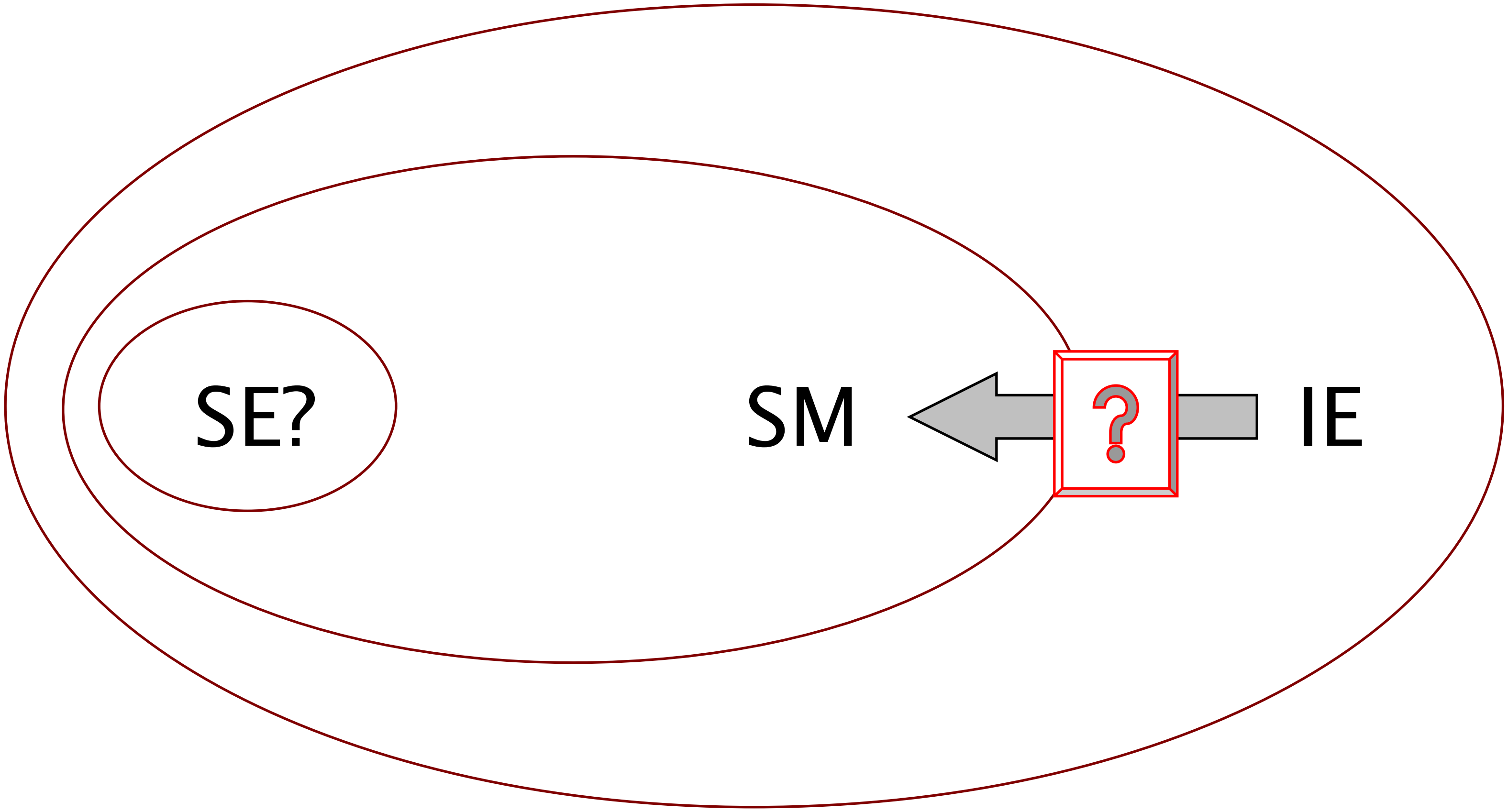
Surrogate Marker (SM)

- Surrogate marker for a drug is a marker which could be used to predict the drug's efficacy or safety
- Example of the usefulness of a surrogate marker
  - Blood glucose is a surrogate marker for Hemoglobin A1c
  - A type-2 diabetes patient took a diabetic drug
  - Clinical studies showed that drug should have an effect on glucose a few hours after taking the drug
  - The patient measured the glucose several times during the first few days of taking the drug
  - If there is not much improvement on glucose, the drug probably does not work for this patient and this patient should switch to a different treatment
  - If there is a clear improvement on glucose, the drug probably works for this patient and the patient should continue taking the drug

Intermediate Endpoint (IE)

- Definition
  - A biomarker associated with or correlated with the clinical endpoint
  - Whether it is a surrogate marker or surrogate endpoint is unknown
- Examples of intermediate endpoint for fracture
  - Bone mineral densities
  - Bone biomarkers
  - Bone images
  - Clinical symptoms (e.g., pain)

Relationship Between SE, SM and IE



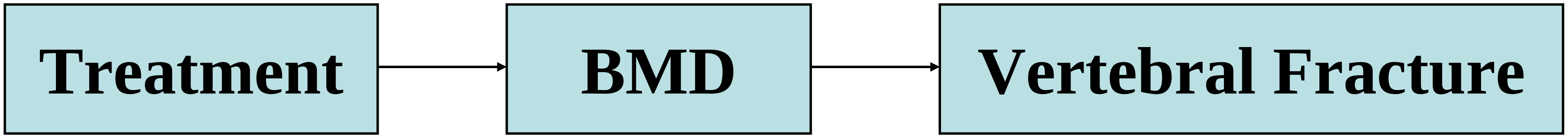
Question: An IE is an SM?

Statistical Validation

- Validation Quantities
  - Proportion of treatment effect (PTE)
  - Likelihood reduction factor (LRF)
  - Proportion of information gain (PIG)
- Notation
  - S = surrogate marker or intermediate marker
  - Z = treatment group
    - 0 = placebo;
    - 1 = active treatment
  - T = clinical outcome
- One of Prentice's key criteria (Prentice, 1989)
  - The distribution of T given S and Z is the same as the distribution of T given S

Example

- S = Intermediate Marker
  - Bone mineral density
- Z = Treatment group
  - Placebo or raloxifene
- T = clinical outcome
  - Vertebral fracture



Treatment → BMD → Vertebral Fracture

PTE and ERO

- Fit two models

  Model 1*:* $T \sim a_0 + a_z Z$

  Model 2*:* $T \sim b_0 + b_z Z + b_x S$

- Proportion of treatment effect (Freedman et al., 1992)

$$\mathrm{PTE} = 1 - a_z^{-1} b_z$$

- Excessive relative odds (ERO) (Sarkar and Qu, 2007) for logistic regression

$$ERO\left(b_z, a_z\right) = \frac{\exp(b_z) - \exp(a_z)}{1 - \exp(a_z)}$$

Drawbacks of PTE and ERO

- Not bounded by [0,1]
- Large variances for the two quantities
- Could be problematic when there is a strong colinearity between S and Z
  - For good surrogate markers, this strong colinearity may be expected

$$T \sim b_0 + b_z Z + b_x\, S$$

Likelihood Reduction Factor (LRF)

LRF (Alonso et al, 2004)

$$\mathrm{LRF}(Z,S{:}Z) = 1 - \exp\{-\mathrm{LRT}(Z,S{:}Z)/n\}$$

where LRT(Z,S:Z) is the likelihood ratio test statistic from models

$$T \sim a_0 + a_Z Z$$

$$T \sim b_0 + b_Z Z + b_s S$$

The LRF is bounded by [0,1], but the maximum possible value may not reach 1

LRF Adjusted (LRF_a)

$$\mathrm{LRF}_a(Z,S:Z) = [\mathrm{LRF}_{\max}]^{-1}\,\mathrm{LRF}\,(Z,S:Z)$$

where $\mathrm{LRF}_{\mathrm{max}}$ is the LRF value for the best-possible fitted model vs. the worst fitted model (e.g., the model only including intercept)

The estimated $\mathrm{LRF}_{\mathrm{max}}$ could be estimated from the fulll model compared to the simplest model

$$\mathrm{LRF}_{\max}=\mathrm{LRF}(Z,S:1)$$

Is LRF (LRF$_a$) a Good Measure?

$$T \sim a_0 + a_Z Z$$

$$T \sim b_0 + b_Z Z + b_s S$$

- LRF (LRF$_a$) is NOT a good measure for the treatment effect by IE
- LRF$_a$ does not reflect the association between *S* and *T*
- LRF$_a$ rather reflects the association between *S* and *T* after adjusting for *Z*

An Artificial Example

$$T = c_0 + c_s S + e$$

$$S = d_0 + d_Z Z + u$$

- The above model meets Prentice's criteria regardless of the variance of *e* and *u*.
- However, LRF$_a$ approaches to 0 as the variance of *u* approaches 0.

A Different Approach

- Instead of comparing

$$T \sim a_0 + a_Z Z$$

$$T \sim b_0 + b_Z Z + b_s S$$

⇒ $\mathrm{LRF}_a(Z,S{:}Z)$

Alonso, et al

- We compare

$$T \sim c_0 + c_S S$$

$$T \sim b_0 + b_Z Z + b_s S$$

⇒ New Quantity

Proportion of Information Gain (PIG)

- Proportion of Information Gain (PIG) (Qu and Case, 2007)

  PIG = LRT(S:1)/LRT(Z,S:1)

- LRT(S:1) = Likelihood Ratio Test statistic comparing models

  $T \sim d_0$

  $T \sim c_0 + c_S S$

- LRT(Z,S:1) = Likelihood Ratio Test statistic comparing models

  $T \sim d_0$

  $T \sim b_0 + b_Z Z + b_s S$

Kullback-Leibler (K-L) Information

- PIG is closed related to K-L Information gain (KLIG)
  - The K-L information gain is LRT/(2n)
- Therefore,

$$\text{PIG} = \text{KLIG}(S{:}1)/\text{KLIG}(Z,S{:}1)$$

Simulation: Setting #1

$$\text{logit}(\Pr(T=1) \mid S, Z) = -S$$

$$S = Z + u, \;\; u \sim N(0, s^2)$$

- Prentice’s criteria are met
- Compare the performance of PTE, LRF_a and PIG for various s^2
- Sample size = 1,000 (n=500 per group)
- 1,000 simulation samples

Simulation Results for Setting #1

s	PTE	LRF_a	PIG
0.01	1.38 (6.66)	0.02 (0.02)	0.98 (0.02)
0.10	1.04 (0.70)	0.06 (0.06)	0.98 (0.02)
1.00	1.02 (0.20)	0.82 (0.05)	1.00 (0.01)
2.00	1.06 (0.34)	0.96 (0.02)	1.00 (0.00)
4.00	1.28 (1.57)	0.99 (0.01)	1.00 (0.00)

Simulation: Setting #2

$$\text{logit}(\Pr(T=1) \mid S, Z) = \gamma_s S + \gamma_z Z$$

$$S = Z + u, \;\; u \sim N(0,1)$$

- Compare the performance of PTE, LRF_a, and PIG for various (γ_s, γ_z)
- Sample size = 1,000 (n=500 per group)
- 1,000 simulation samples

Simulation Results for Setting #2

(γ_z, γ_s)	PTE	LRF$_a$	PIG
(0.0, 1.0)	1.03 (0.21)	0.83 (0.05)	1.00 (0.01)
(0.2, 0.8)	0.79 (0.16)	0.74 (0.07)	0.98 (0.02)
(0.5, 0.5)	0.48 (0.10)	0.51 (0.10)	0.88 (0.06)
(0.8, 0.2)	0.20 (0.07)	0.16 (0.09)	0.55 (0.12)
(1.0, 0.0)	0.00 (0.07)	0.02 (0.03)	0.21 (0.11)

Clinical Example

- Data from the Multiple Outcomes of Raloxifene Evaluation (MORE) study
- Duration of the study: 3 years
- Treatments: placebo or raloxifene
- Clinical outcome: prevalent vertebral fracture in 3 years
- A total of 2230 subjects with measurement for
  - vertebral fracture
  - bone mineral density (BMD)
  - bone biomarkers
- <u>Problem of interest</u>: see whether the short-term change in femoral neck BMD and bone markers are acceptable surrogates for the vertebral fracture reduction by raloxifene

Notation

- Z = Treatment group
  - 0 for placebo
  - 1 for raloxifene
- T = New vertebral fracture
- S = Vector for change in
  - femoral neck BMD at 1 year
  - CTX averaged at 6 months and 1 year
  - osteocalcin averaged at 6 months and 1 year
  - BSALP average at 6 months and 1 year

CTX = Urinary type I collagen C-telopeptide excretion, corrected for urinary creatinine excretion
BSALP = Bone-specific alkaline phosphatase

Results from the Clinical Example

Method	Estimate	Standard Error	95% CI*
PTE	0.31	7.39	(-0.03, 1.38)
LRF_a	0.46	0.23	(0.13, 0.95)
PIG	0.79	0.18	(0.36, 0.999)

* 95% confidence interval was calculated by bootstrap method

Summary

- Theory and simulation show PIG better quantifies the IE compared to PTE and LRF
- In the ideal cases where IE is a surrogate endpoint, PIG but not LRF correctly reflects the situation
- The standard error in the estimated PTE is generally large and limits its use in practice

Discussion

- This research focuses on a single study, but it could be potentially extended to multiple studies or meta-analysis
- No threshold value of PIG is given above which the IE could be considered a surrogate
- Research needs to be done to evaluate the performance of PIG and other methods when the analysis models and data-generating models do not match

Part II

The Effect of Measurement Error on Evaluation of Surrogate Markers

Linear Measurement Error Models

$$Y = \beta_0 + \beta_1 X + e,\ e \sim N(0,\ \sigma_e{}^2)$$

$$W = X + u,\ u \sim N(0,\ \sigma_u{}^2)$$

- If no the measurement error, we regress Y on X

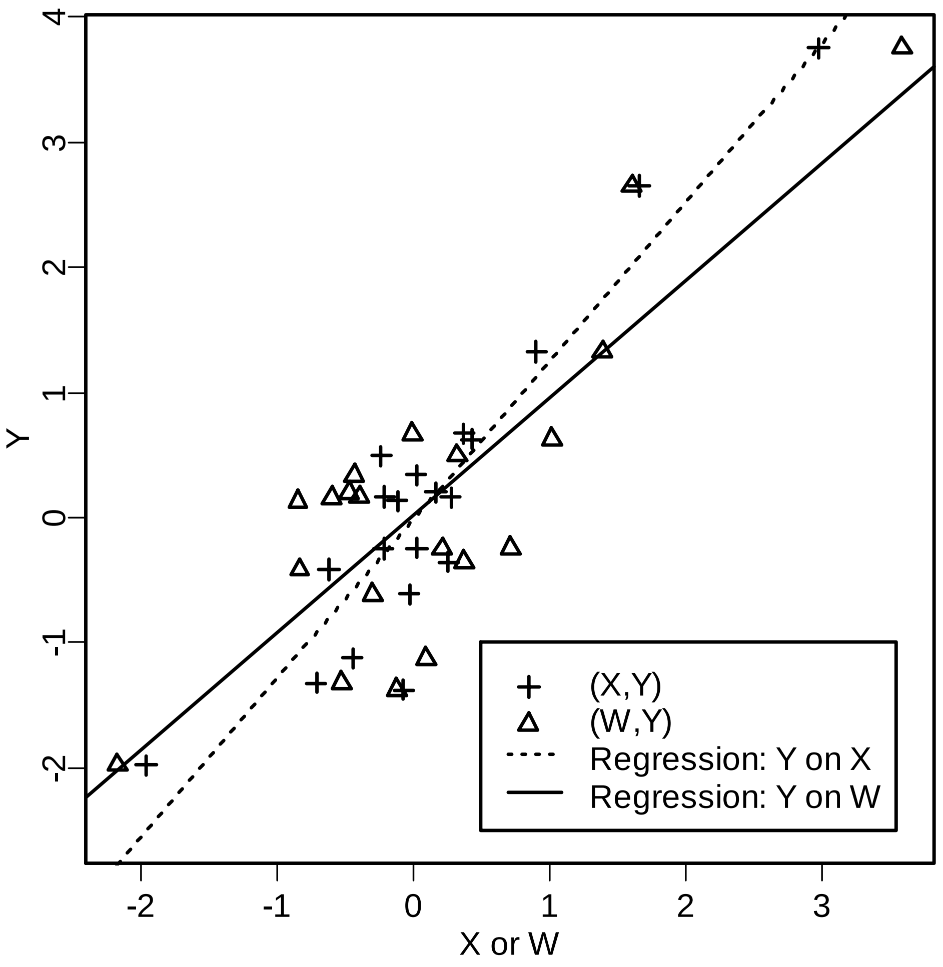
$$b_{1,\text{no ME}} = [\text{Var(X)}]^{-1}\ \text{Cov(X,Y)}\ \ {-}{>}\ (\sigma_x{}^2)^{-1}\sigma_{xy}$$

- Ignoring the measurement error, we regress Y on W

$$b_{1,\text{naive}} = [\text{Var(W)}]^{-1}\ \text{Cov(W,Y)}\ \ {-}{>}\ (\sigma_x{}^2 + \sigma_u{}^2)^{-1}\sigma_{xy}$$

- Therefore, the naïve estimator is biased
- Reliability ratio: $(\sigma_x{}^2 + \sigma_u{}^2)^{-1}\ \sigma_x{}^2$ (Fuller, 1987)

Effect of Measurement Error on Linear Regression Coefficients



How to Correct the Bias?

- Linear or polynomial regression
  - Methods of moments (Fuller, 1987)
  - Regression calibration (Carroll, et al. 2006)
- Logistic regression
  - Regression calibration (Carroll, et al. 2006)
- General nonlinear model
  - Simulation extrapolation – an approximate method (Cook and Stefanski, 1994)

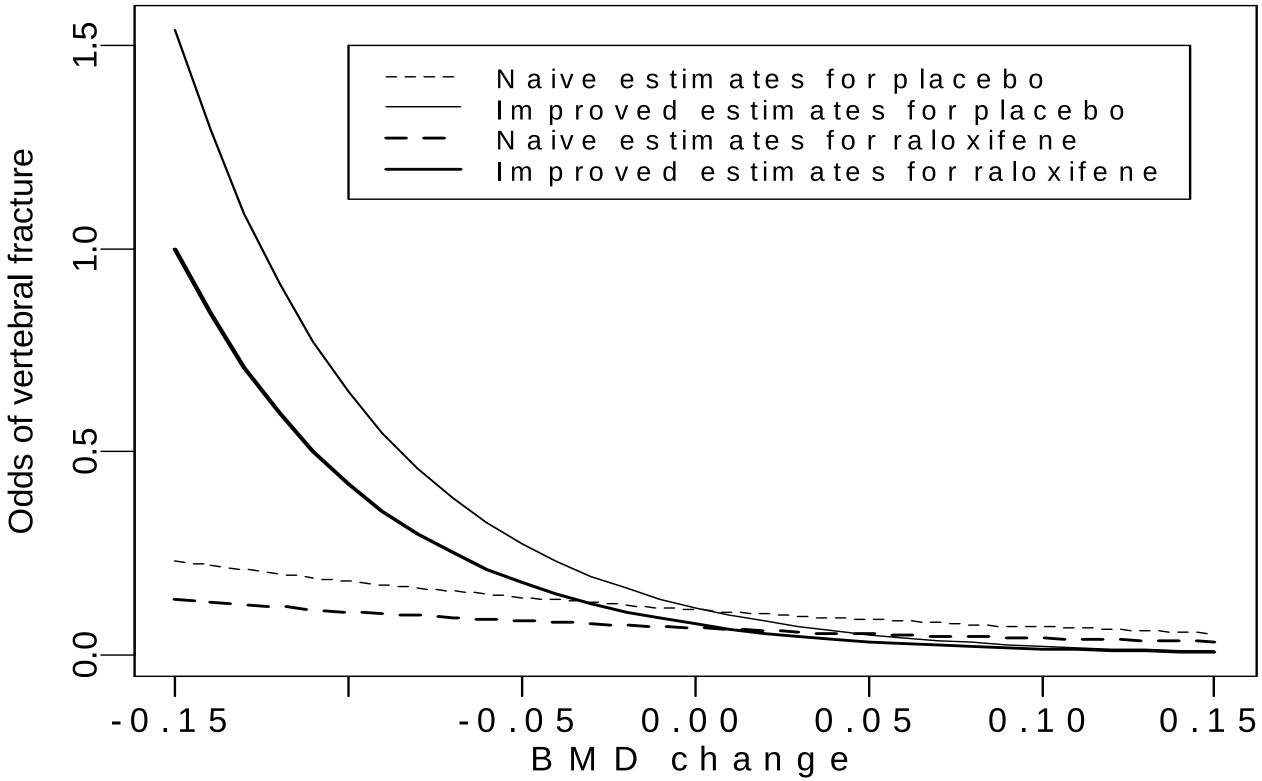
An Example

- MORE Study
- Is change in femoral neck BMD a good marker for vertebral fracture?
- We use ERO (similar to PTE) to illustrate the problem

$$\boxed{\begin{aligned}\log\left(\frac{P(T=1|S,Z)}{1-P(T=1|S,Z)}\right)&=\beta_0+\beta_x S+\beta_z Z\\ \log\left(\frac{P(T=1|Z)}{1-P(T=1|Z)}\right)&=\alpha_0+\alpha_z Z\end{aligned}}\Rightarrow ERO(b_z,a_z)=\frac{\exp(b_z)-\exp(a_z)}{1-\exp(a_z)}$$

- Regression calibration was used to correct the bias
  - Regression calibration is to find the distribution of T given the variables measured with error

Effect of Measurement Error



$$\log\left(\frac{P(T=1|S,Z)}{1-P(T=1|S,Z)}\right)=\beta_0+\beta_x S+\beta_z Z$$

Reliability ratio = 30%

Effect of Measurement Error on ERO

$$ERO(b_z, a_z) = \frac{\exp(b_z) - \exp(a_z)}{1 - \exp(a_z)}$$

Table. Estimates for ERO

	Point Estimate	95% CI*
Naïve	0.052	0.011-0.115
Improved	0.203	0.041-0.456

The 95% CI was calculated by bootstrap methods.

Summary and Discussion

- It is important to consider measurement error if the magnitude of measurement error is large
  - Use reliability ratio $(\sigma_x^2 + \sigma_u^2)^{-1} \sigma_x^2$
  - Generally, there is no need to consider measurement error if reliability ratio > 70%
- Research is ongoing to incorporate the measurement error for PIG

Thank You!

Questions?

References

Alonso, A. et al. (2004). Prentice's approach and the meta-analytic paradigm: A reflection on the role of statistics in evaluatiion of surrogate endpoints. *Biometrics*, 60:724-728.

Carroll R. J., Ruppert D., Stefanski L. A., Craineceanu C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective* (2nd edn). Chapman & Hall, CRC Press: London, Boca Raton, FL, 2006.

Cook, R. J. and Stefanski, L. A. (1994) Simulation-extrapolation in parametric measurement error models. *J. Am. Statist. Ass.*, 89, 1314–1328.

Freedman L, Graubard B (1992). Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine.*

Fuller, W. A. (1987). *Measurement Error Models*. Wiley, New York.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics* 22, 79–86.

Prentice RL (1989). Surrogate endpoints in clinical trials. *Statistics in Medicine.*

Sarkar S. and Qu Y. (2007). Quantifying the proportion of treatment effect explained by surrogate markers in the presence of measurement error. *Statistics in Medicine*. 26(9):1955-63.

Qu Y. and Case M. (2007). Quantifying the Effect of the Surrogate Marker by Information Gain. *Biometrics*. 63: 958 – 963