VU university medical center

An efficient alternative to the complete matched-pairs design for assessing non-inferiority of a new diagnostic test

Peter van de Ven (VU University Medical Center)
Joint work with Johannes Berkhof (VUmc)

Workshop Designs for Healthcare (Cambridge) August 15-19, 2011

Outline

- Motivation
- Data structure, statistical model and estimation
- Statistical tests
- Efficiency
- Examples
- Conclusions
- Extensions and future research

Motivation

Comparing performance of diagnostic tests with binary outcome:

- Standard test has positivity rate π_S
- New test has positivity rate π_N

Both diagnostic tests are assumed to be imperfect

Non-inferiority hypothesis:

$$H_0: \pi_N = \delta_0 \pi_S$$
$$H_1: \pi_N > \delta_0 \pi_S$$

δ_0 is a prespecified constant

Example: $\delta_0 = 0.8$ (rejection desired if equivalence: $\delta_1 = 1.0$)

If $\delta_0 = 1$ then McNemar hypothesis for superiority

Motivation

Population	Specificity	Interpretation δ (:= π_N / π_S)
Screening population	Specificity < 100% (false positives occur)	π_N / π_S = relative positivity rate
Screening population	Specificity = 100% for both tests	π_N / π_S = relative sensitivity
Diseased population		π_N / π_S = relative sensitivity

We consider two situations:

- Comparing positivity rates in studies where verification is not possible (such as chlamydia and tuberculosis screening)
- Comparing sensitivities/specificities using biobank samples with documented gold standard and standard test outcome

Motivation

Reasons for non-inferiority testing ($\delta_0 < 1$):

- New test is more patient-friendly (for instance, less invasive)
- New test is easier to perform
- New test leads to faster diagnosis
- New test is anticipated to be cheaper after implementation

Motivation

Standard design: complete matched-pairs design
Both tests administered to all subjects

	T_S+	T_S-
T_N+	*a*	*b*
T_N-	*c*	*m-a-b-c*

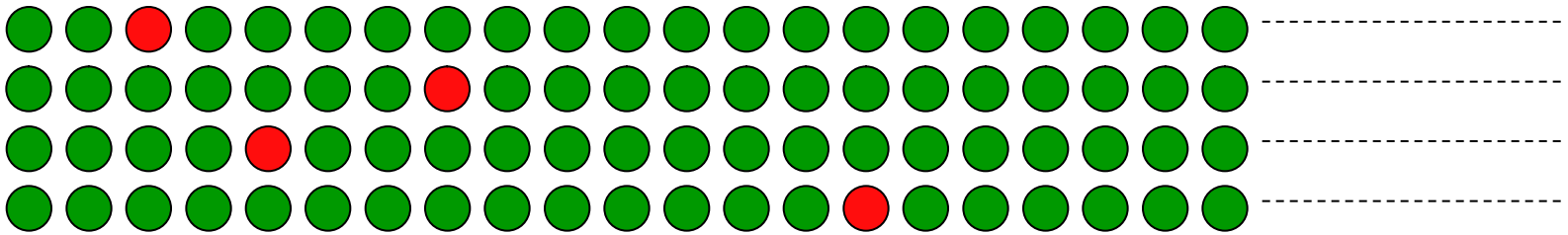
This design may not be optimal if

- positivity rates are low
- new test is expensive or
- results on standard have already been obtained for a large sample

Motivation

Example: biobank with urine samples

non-inferiority testing of new instrument for detection of chlamydia



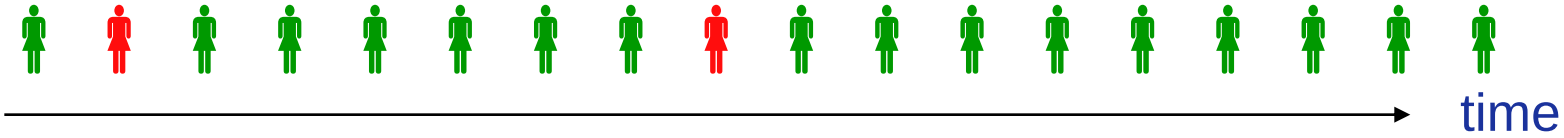
Standard test negative

Standard test positive

How many standard test positives and standard test negatives must be tested with the new method (stratified random sampling)?

Motivation

Example: prospective screening cohort study



Standard test negative

Standard test positive

All standard test positives receive the new test

Standard test negatives receive the new test with probability *p*

Motivation

Two-phase procedure:

1) Standard test given to a *simple random sample* of population

2) New test given to a *stratified random sample* with strata defined by reference test outcomes

	T_S+	T_S-
	n - x	*x*

Stratified random sampling

	y	*m - y*
T_N+	*a*	*b*
T_N-	*y - c*	*m - y - b*

Data structure and estimation

Assumptions:

n and *m* are fixed

$x \sim \text{Binomial}(n, \pi_S)$
$a \mid y \sim \text{Binomial}(y, \alpha)$
$b \mid y \sim \text{Binomial}(m - y, \beta)$

π_S, α, and β are unknown parameters

$y \mid n, m, x$ has distribution that is completely known

T_S+	T_S-
$n - x$	x

Stratified random sampling

	y	$m - y$
T_N+	a	b
T_N-	$y - a$	$m - y - b$

Data structure and estimation

Probability of observing (a,b,y,x):

Standard test in simple random sample

Sampling mechanism

$$P(a, b, y, x \mid n, m, \psi, \pi_S, \alpha, \beta) = \underbrace{P(x \mid n, \pi_S)}\ \underbrace{P(y \mid n, m, \psi, x)}\ \underbrace{P(a \mid \alpha, y)}\ \underbrace{P(b \mid m, \beta, y)}$$

New test in T_S+ stratum

New test in T_S- stratum

For stratified random sampling we consider ψ to be the target for the fraction y / m

This implies that y is fixed by x, m and ψ

Data structure and estimation

Unrestricted maximum likelihood estimators

$\tilde{\alpha} = a/y$ $\tilde{\beta} = b/(m-y)$ $\tilde{\pi}_S = x/n$

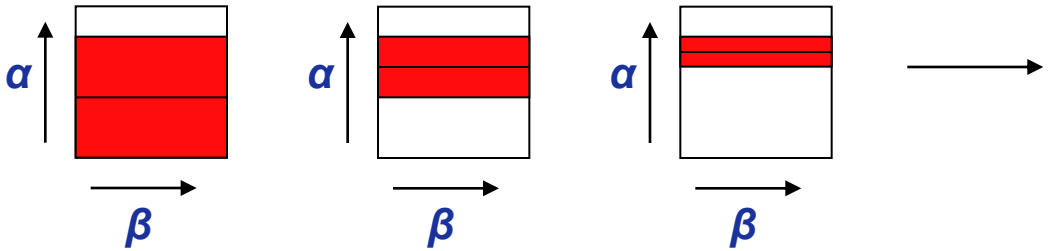
Maximum likelihood estimators under the null: $\pi_N = \delta_0 \pi_S$

Null constraint: $\pi_S = \beta /(\beta + \delta_0 - \alpha)$

MLEs for α and β under the null can be found using a bisection algorithm

Data structure and estimation

- $\partial l_c(\alpha,\beta) / \partial\beta$ is quadratic in β
- Bisection on α using function $g(\alpha) = \max_\beta l_c(\alpha,\beta)$
- Majorization argument shows that
  sign of $\partial g(\alpha)/\partial\beta$ = sign of $\partial l_c(\alpha,\beta^*[\alpha])/\partial\beta$

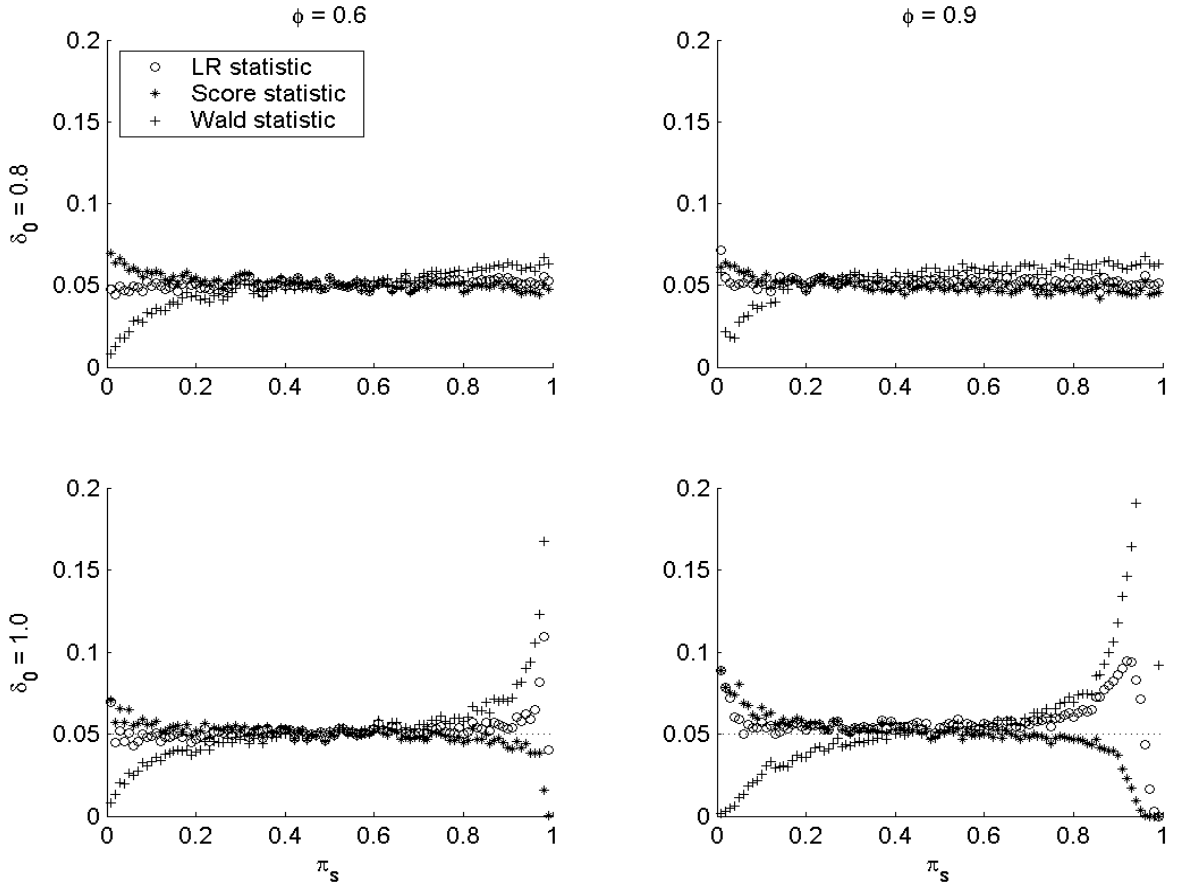


- After k iterations the width of the interval for α is at most $(1/2)^k$

Statistical tests

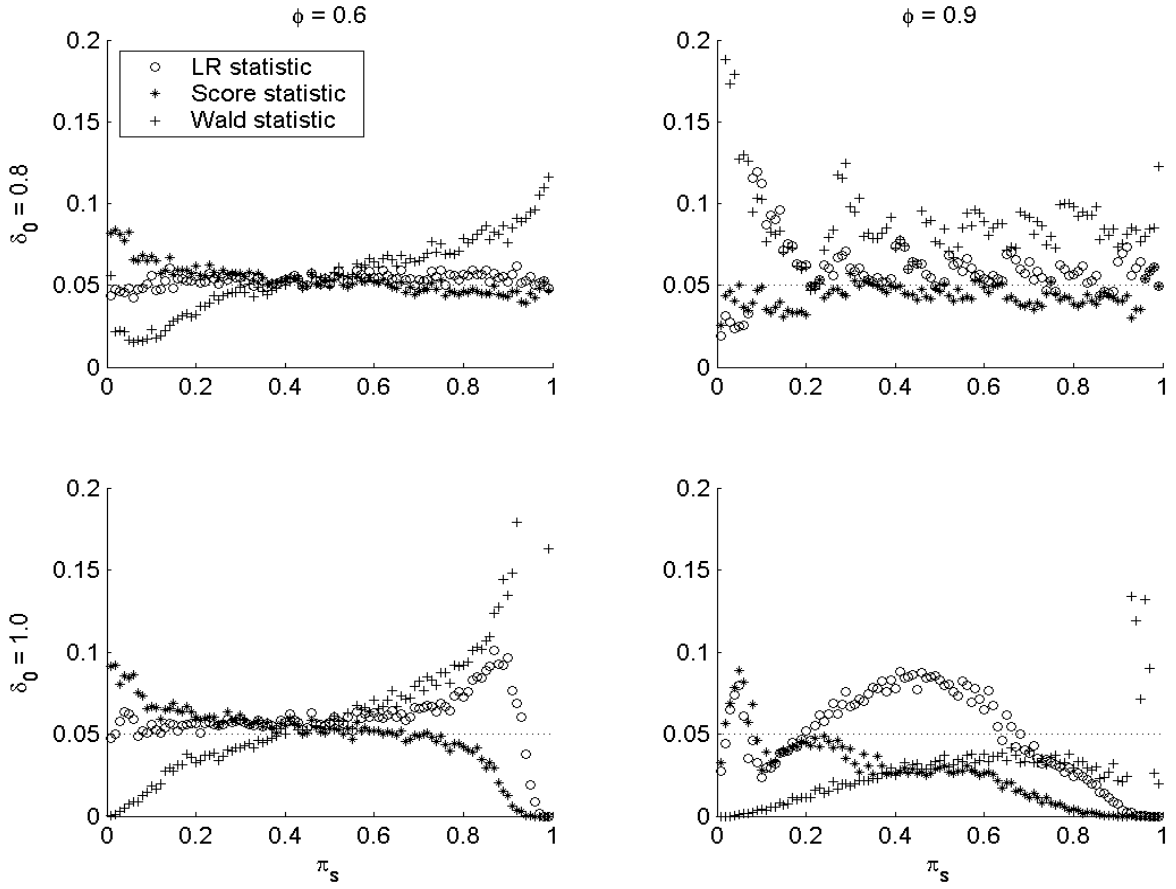
- **Signed version of Likelihood Ratio statistic**
  - Does not depend on sampling procedure
  - Requires calculation of MLEs under the null
- **Score statistic**
  - Sampling procedure needs to be taken into account
  - Requires calculation of MLEs under the null
- **Wald statistic**
  - Sampling procedure needs to be taken into account
  - Only requires unrestricted MLEs
  - Shown to be inferior in matched-pairs designs

Statistical tests



False rejection rates: n = 1000, m = 250 (Φ = relative within-pair correlation)

Statistical tests



False rejection rates: n = 500, m = 40 (Φ = relative within-pair correlation)

Statistical tests

Control of false rejection rates

- $m = 250$ and $\delta_0 = 0.8$: All tests showed good control of the error rate
  Likelihood ratio test preferred when $\pi_S < 0.1$

- $m = 250$ and $\delta_0 = 1.0$: Score test preferred
  Wald and likelihood ratio test become liberal when π_S close to 1

- $m = 40$: only score test showed good control of the error rate, although it is slightly liberal when $\pi_S < 0.1$

Efficiency

Superiority testing

Null hypothesis: $\delta_0 = 1.0$

Alternative hypothesis: $\delta_1 = 1.1$

Correlation: $\Phi = 0.6$

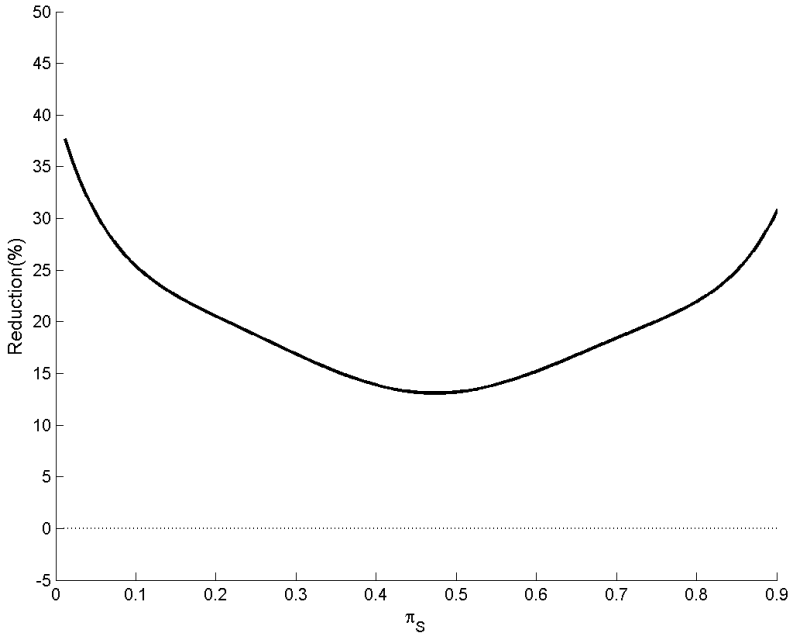
To allow for oversampling of positive samples when π_S is small we set $n = 500000$

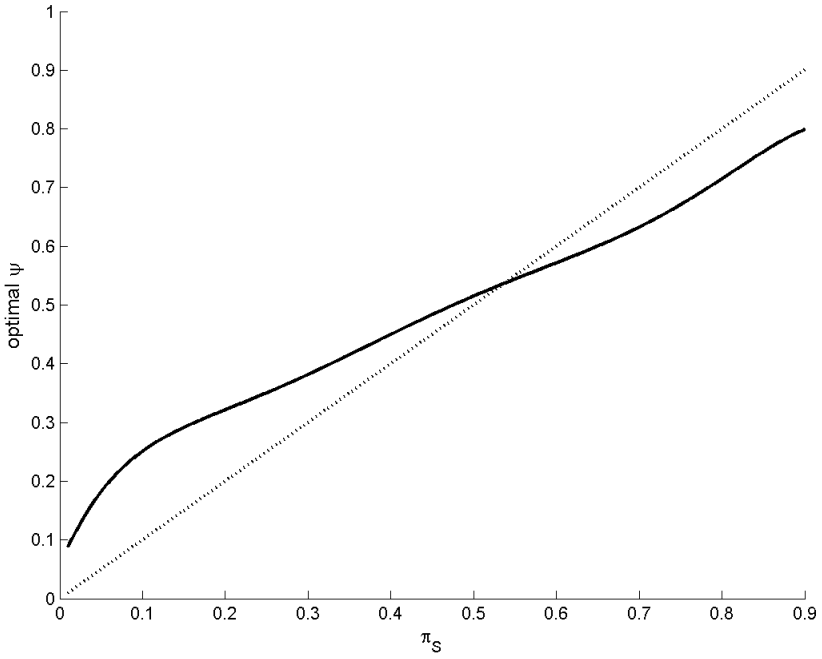
For a range of π_S we determined:

- Optimal fraction Ψ of standard test positives to be sampled
- Minimal m required for achieving 80% power

Efficiency

Reduction in sample size (*m*) under optimal non-proportional stratified random sampling compared to number in complete matched-pairs design

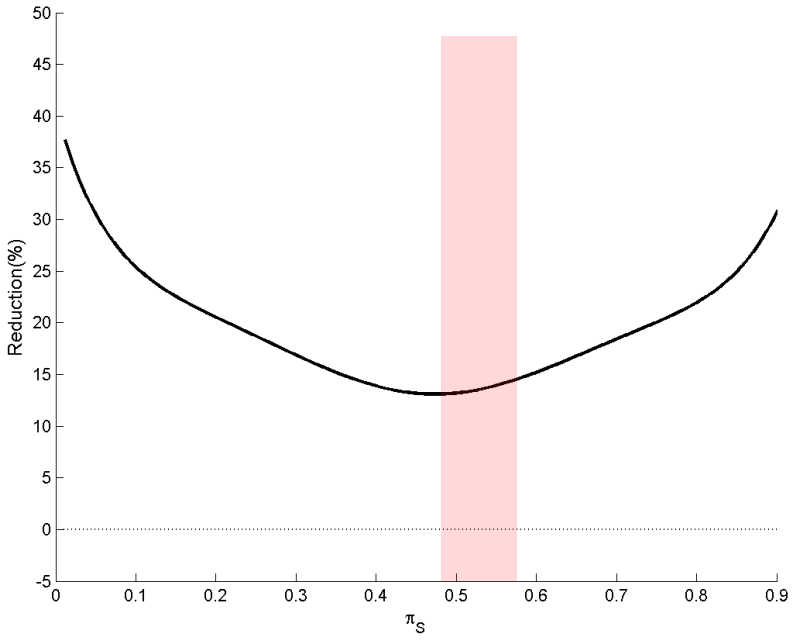


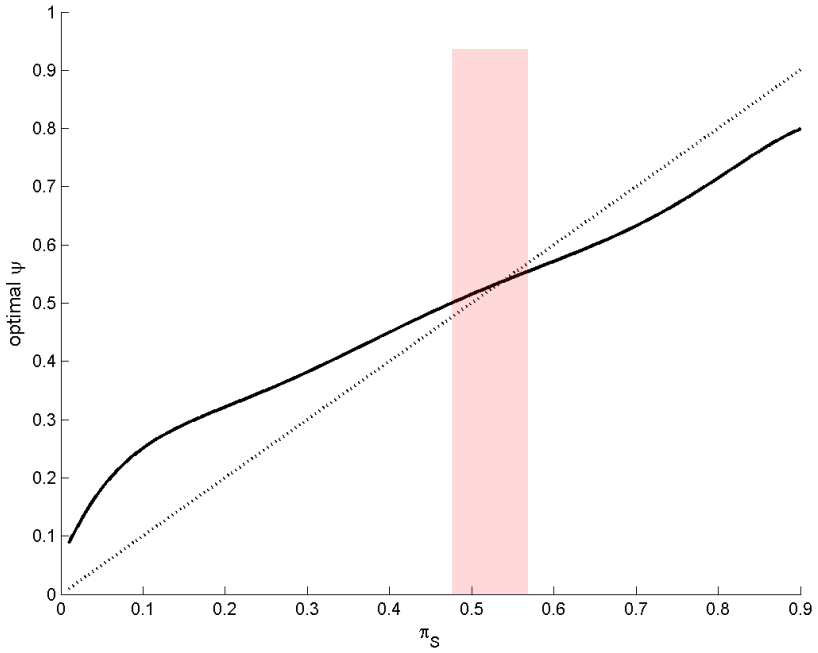


Efficiency

Reduction when sampling is proportional

Caused by using *n* standard tests results instead of *m*

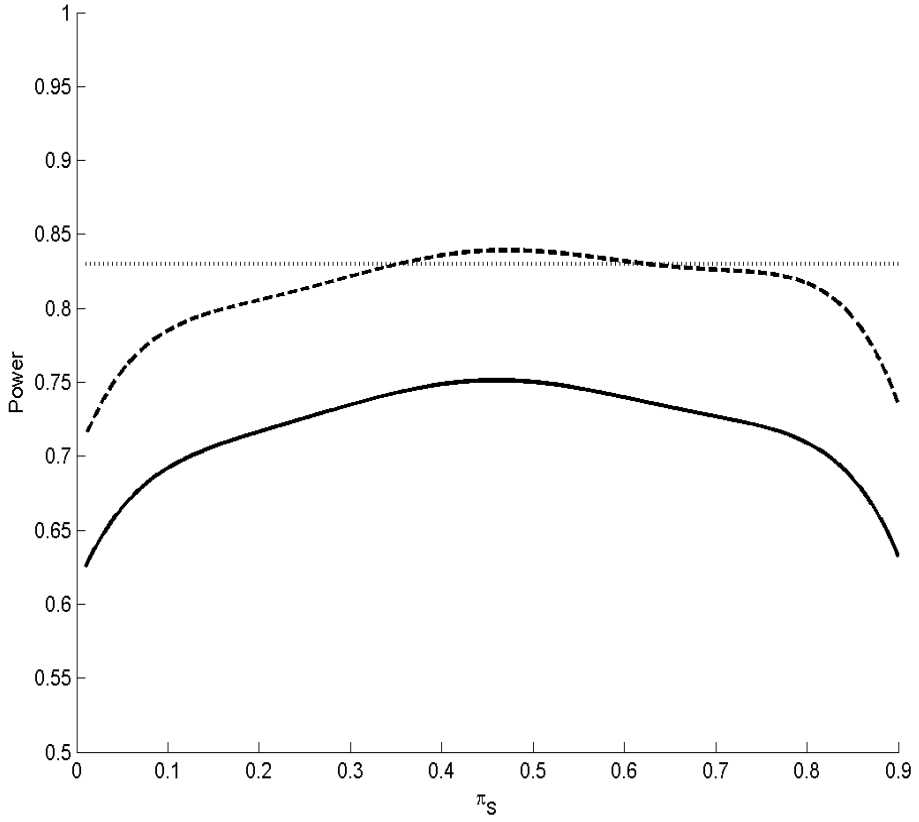




Efficiency

m determined to have 80% power under optimal non-proportional stratified random sampling

- Optimal non-proportional stratified random sampling (dotted line)
- Simple random sampling of *m* out of *n* (dashed line)
- Complete matched-pairs design with *m* (solid line)



Efficiency

Non-inferiority testing

Null hypothesis: $\delta_0 = 0.8$

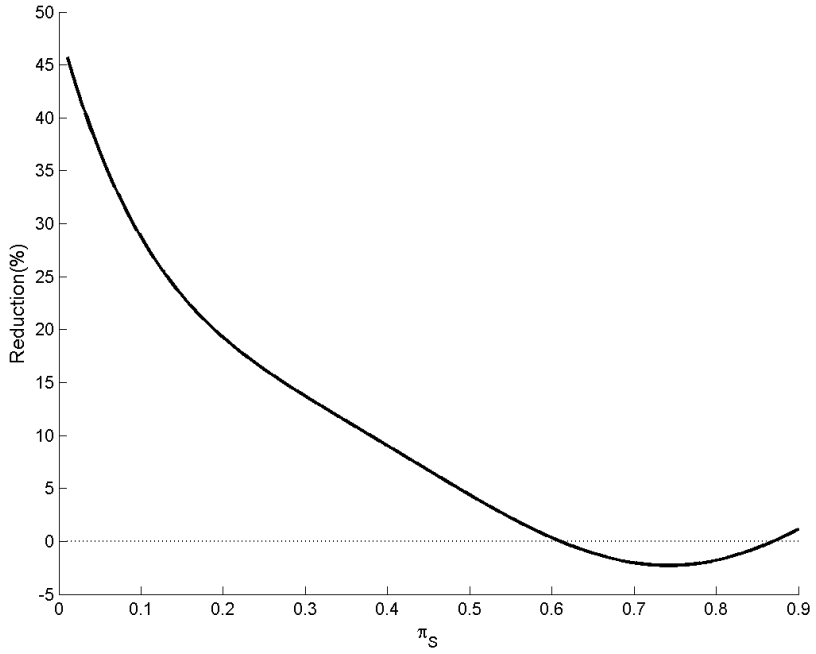
Alternative hypothesis: $\delta_1 = 1.0$

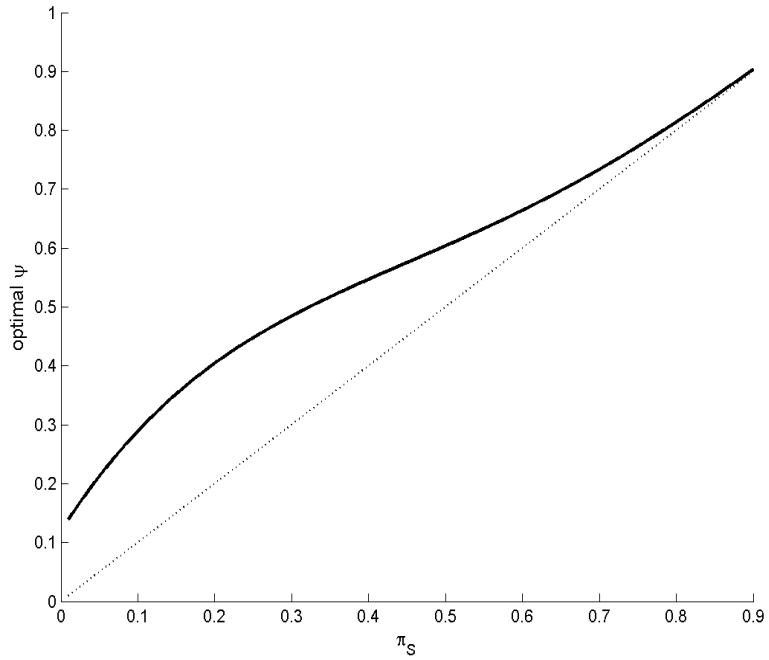
Correlation: $\Phi = 0.6$

To allow for oversampling of positive samples when π_S is small we set $n = 500000$

Efficiency

Reduction in sample size (*m*) under optimal non-proportional stratified random sampling compared to number in complete matched-pairs design

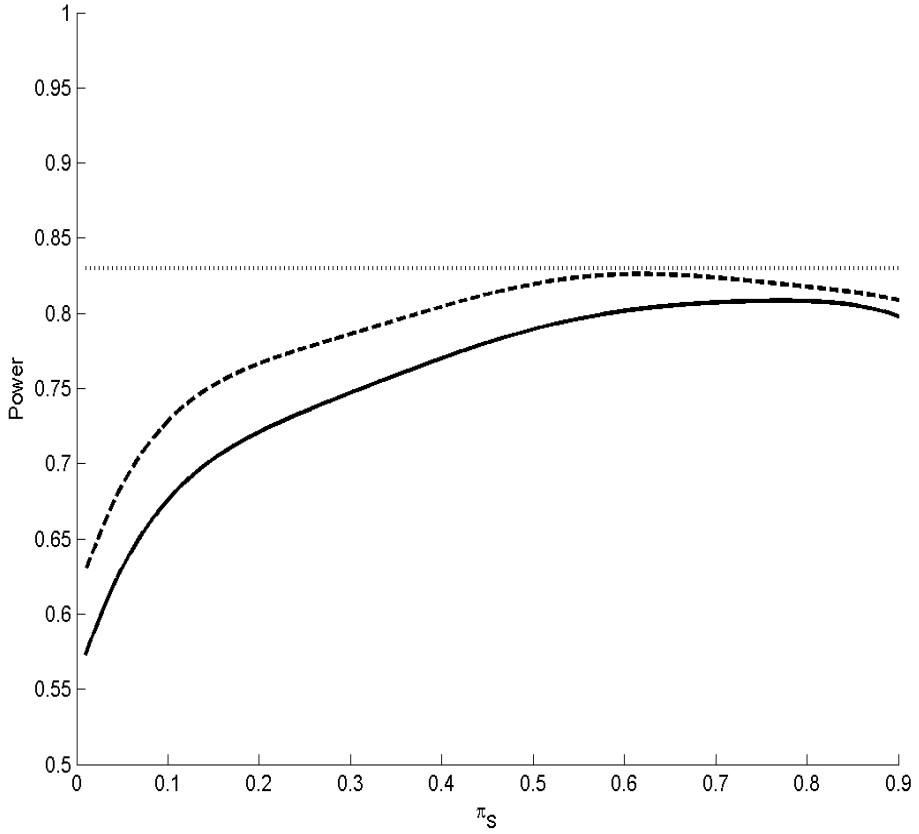




Efficiency

m determined to have 80% power under optimal non-proportional stratified random sampling

- Optimal non-proportional stratified random sampling (dotted line)
- Simple random sampling of *m* out of *n* (dashed line)
- Complete matched-pairs design with *m* (solid line)



Efficiency

Superiority testing

Increase in power under optimal non-proportional stratified sampling compared to simple random sampling when

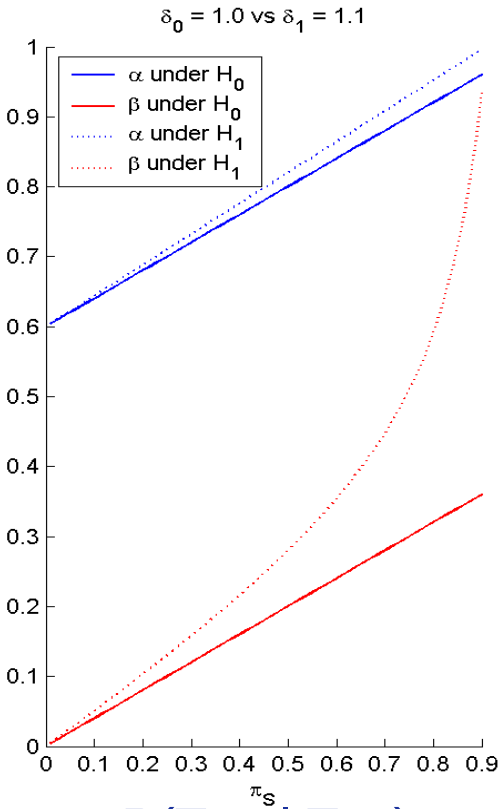
- $\pi_S \leq 0.2$ (oversampling of standard test positives)
- $\pi_S \geq 0.8$ (oversampling of standard test negatives)

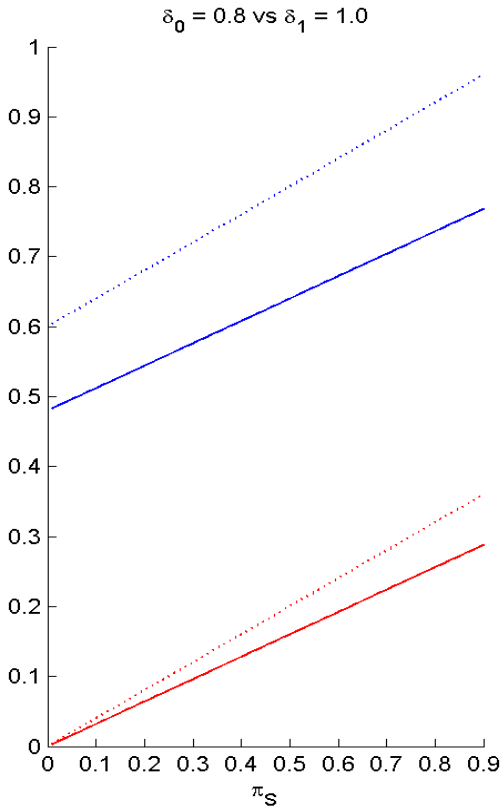
Non-inferiority testing

Increase in power under optimal non-proportional stratified sampling compared to simple random sampling when

- $\pi_S \leq 0.4$ (oversampling of standard test positives)

Efficiency

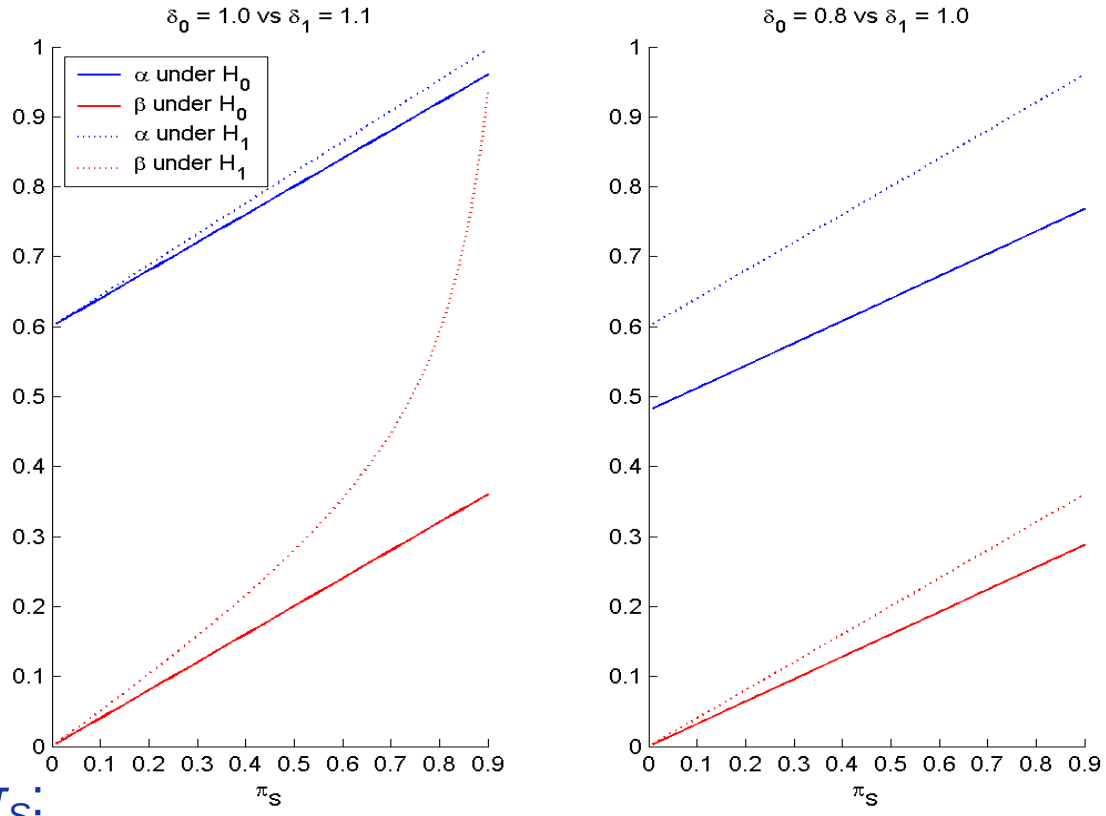




$\alpha = P(T_N+ \mid T_S+)$

$\beta = P(T_N+ \mid T_S-)$

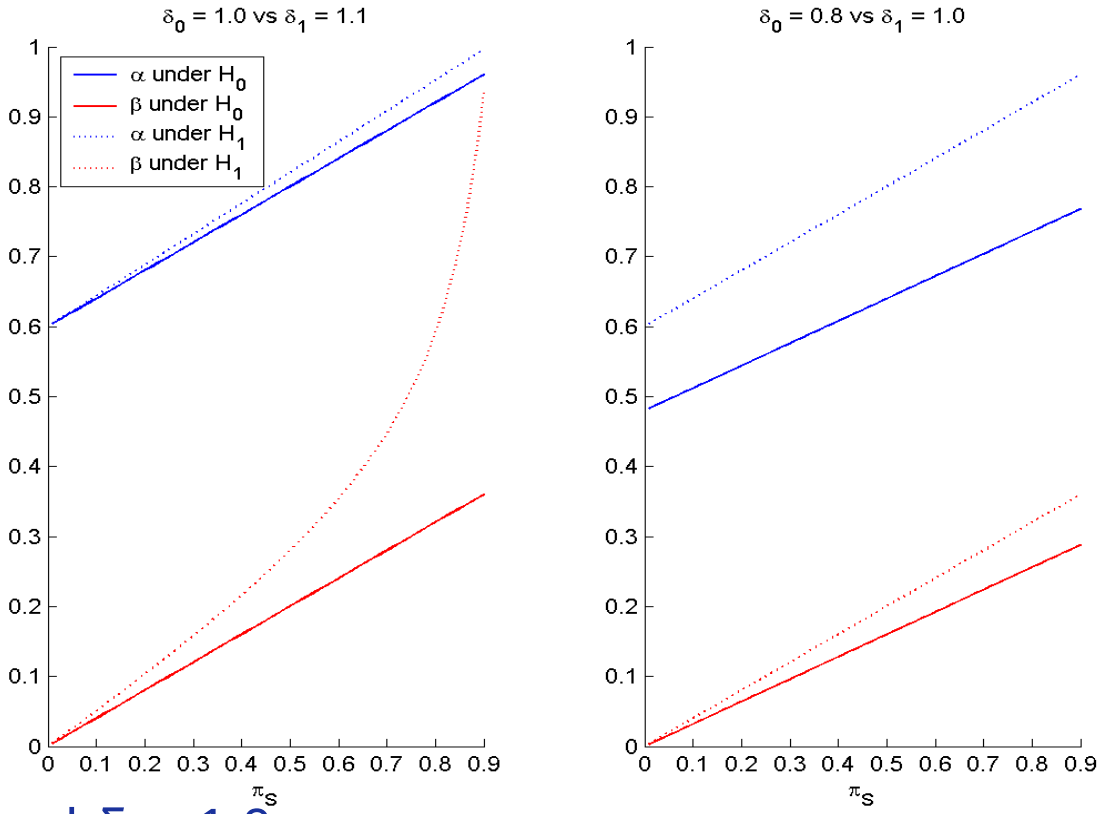
Efficiency



Small π_S:

A result on the new test in T_S+ stratum provides relatively little information about α, compared to the information on β provided by a result in the T_S- stratum

Efficiency



Large π_S and $\delta_0 = 1.0$:
A result on the new test in T_S- stratum provides relatively little information about β, compared to the information on α provided by a result in the T_S- stratum

Examples

Positivity rates of screening test for chlamydia
(Morré et al., *J. Clin. Microbiol.,* 1999: 37; 3092 - 3096)

Specificity of PCR tests ≈ 100%

Positivity rates in asymptomatic screening population ≈ 3%

Within-pair correlation of LCx (Abbot) and COBAS (Roche):
$\rho = 0.8$, $\Phi = 0.86$

Comparing positivity rates of two tests in a complete matched-pairs design, assuming $\delta_0 = 0.9$, $\pi_S = 0.03$ and $\Phi = 0.85$ requires 5816 samples to achieve 80% power

Examples

Positivity rates of screening test for chlamydia (continued)

Suppose that a biobank containing 50000 urine samples with documented outcome on the standard test becomes available

Only 3000 samples are required when non-proportional stratified random sampling is used ($\Psi = 0.14$)

Reduction in m compared to complete matched-pairs design: 48%

Examples

Specificity of candidate HPV assays for detection of an underlying lesion

(Meyer et al., *Int. J. Cancer*, 2009, 124:516-520)

Specificity of Hybrid Capture 2 test (FDA approved standard test) for underlying lesion is $\pi_S = 0.93$

Testing superiority with $\delta_0 = 1.0$ under alternative $\delta_1 = 1.05$

McNemar test requires 145 HPV women without a lesion to achieve 80% power

Examples

Specificity of candidate HPV assays for detection of underlying lesion (continued)

Stratified random sampling of 145 HPV woman without a lesion from a cohort of 1000 gives 91% power (score test)

To achieve 80% power, we need to sample 110 women ($\Psi = 0.85$)

Reduction in *m* compared to complete matched-pairs design: 25%

Conclusions

- We derived a likelihood ratio test, score test and Wald test statistic for non-inferiority testing for studies where standard test outcome can be used as a stratification tool
- Non-inferiority was defined in terms of relative risk
- Score test performed best in small samples (provided that $\pi_S \geq 0.1$)
  This is in line with results reported in Tang et al. (*Statistics in Medicine*, 2003: 22: 1217-1233)
- Non-proportional stratified sampling leads to a substantial reduction in the number of new test required when
  - Testing for non-inferiority (small π_S)
  - Testing for superiority (small π_S and large π_S)

Extensions and future research

- Derive exact tests for small m ($m < 40$)

- Derive more powerful tests by defining more than two strata based on ancillary information, such as an underlying continuous score on the standard test, and choosing the optimal:
  - Strata
  - Stratification weights