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)

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 $\pi_S$
- New test has positivity rate $\pi_N$

Both diagnostic tests are assumed to be imperfect

Non-inferiority hypothesis:

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

$\delta_0$ 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

<table>
<thead>
<tr>
<th>Population</th>
<th>Specificity</th>
<th>Interpretation $\delta (:= \frac{\pi_N}{\pi_S})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Screening population</td>
<td>Specificity &lt; 100% (false positives occur)</td>
<td>$\frac{\pi_N}{\pi_S} =$ relative positivity rate</td>
</tr>
<tr>
<td>Screening population</td>
<td>Specificity = 100% for both tests</td>
<td>$\frac{\pi_N}{\pi_S} =$ relative sensitivity</td>
</tr>
<tr>
<td>Diseased population</td>
<td></td>
<td>$\frac{\pi_N}{\pi_S} =$ relative sensitivity</td>
</tr>
</tbody>
</table>

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

\[
\begin{array}{ccc}
T_{S+} & T_{S-} \\
T_{N+} & a & b \\
T_{N-} & c & m-a-b-c \\
\end{array}
\]

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
Example: biobank with urine samples
non-inferiority testing of new instrument for detection of chlamydia

Motivation

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

\[
\begin{array}{c|c}
T_{S+} & T_{S-} \\
\hline
n - x & x \\
\end{array}
\]

\[
\begin{array}{c|c}
T_{N+} & T_{N-} \\
\hline
a & b \\
\end{array}
\]

\[
\begin{array}{c|c}
y - c & m - y - b \\
\end{array}
\]

Stratified random sampling
Data structure and estimation

Assumptions:

- $n$ and $m$ are fixed
- $x \sim \text{Binomial}(n, \pi_S)$
- $a \mid y \sim \text{Binomial}(y, \alpha)$
- $b \mid y \sim \text{Binomial}(m - y, \beta)$
- $\pi_S, \alpha, \text{and} \beta$ are unknown parameters
- $y \mid n, m, x$ has distribution that is completely known

Stratified random sampling:
Data structure and estimation

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

\[
P(a, b, y, x | n, m, \psi, \pi_S, \alpha, \beta) = P(x | n, \pi_S) P(y | n, m, \psi, x) P(a | \alpha, y) P(b | m, \beta, y)
\]

For stratified random sampling we consider \(\psi\) to be the target for the fraction \(y / m\)

This implies that \(y\) is fixed by \(x, m\) and \(\psi\)
Unrestricted maximum likelihood estimators

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

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

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

MLEs for \( \alpha \) and \( \beta \) under the null can be found using a bisection algorithm
Data structure and estimation

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

- After $k$ iterations the width of the interval for $\alpha$ 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
False rejection rates: $n = 1000$, $m = 250$ (\(\Phi = \text{relative within-pair correlation}\))
Statistical tests

False rejection rates: $n = 500, \ m = 40 \ (\Phi = \text{relative within-pair correlation})$
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 $\pi_S$ 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 \( \pi_S \) is small we set \( n = 500000 \)

For a range of \( \pi_S \) we determined:
- Optimal fraction \( \Psi \) 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 $\pi_S$ 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

\[ \alpha = P(T_{N^+} \mid T_{S^+}) \]
\[ \beta = P(T_{N^+} \mid T_{S^-}) \]
Small $\pi_S$:
A result on the new test in $T_{S^+}$ stratum provides relatively little information about $\alpha$, compared to the information on $\beta$ provided by a result in the $T_{S^-}$ stratum
Large $\pi_S$ and $\delta_0 = 1.0$:
A result on the new test in $T_S$-stratum provides relatively little information about $\beta$, compared to the information on $\alpha$ provided by a result in the $T_S$-stratum
Examples

Positivity rates of screening test for chlamydia

Specificity of PCR tests $\approx 100\%$

Positivity rates in asymptomatic screening population $\approx 3\%$

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

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
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 tests required when:
  – Testing for non-inferiority (small $\pi_S$)
  – Testing for superiority (small $\pi_S$ and large $\pi_S$)
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