Evaluating Diagnostic Accuracy in the Face of Multiple Reference Standards

Christiana A Naaktgeboren, MPH; Joris AH de Groot, PhD; Maarten van Smeden, MSc; Karel GM Moons, PhD; and Johannes B Reitsma, MD, PhD

July 16, 2013
Differential Verification (DV)

- Is the use of different reference standards (RS) in different groups of participants in a diagnostic study\(^1\)
- Can bias accuracy estimates\(^2\)
- May occur in \(\frac{1}{4}\) of all DTA studies\(^3\)
- Difficult to score QUADAS item, partially due to poor reporting

1) de Groot (2011) BMJ
2) Rutjes et al (CMAJ) 2006
Verification Patterns

Single test as Reference Standard?
- Reference standard applied in all patients?
  - Complete Verification
  - Partial Verification
  - Composite Reference Standard
    - Panel Diagnosis
    - Latent Class Analysis
  - Differential Verification
- Same component tests in all patients?
# Examples of DV

<table>
<thead>
<tr>
<th>Target Condition</th>
<th>Index Test</th>
<th>Preferred RS</th>
<th>Inferior RS(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast Cancer</td>
<td>1 Mammography Biopsy Follow-up</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Congenital Heart</td>
<td>Pulse oxometry (blood O2 levels)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Depression</td>
<td>Diagnostic prediction rule</td>
<td>In-person interview</td>
<td>Phone interview</td>
</tr>
</tbody>
</table>

3) Zuithoff et al (2009) Fam Practice
Complete Index Test  Dependent DV

A

Participants

Index test

Choice of reference standard depends on:

Index test results

index
+-

Preferred reference

+ b
- 0

Inferior reference

+ 0 0
- g h

Index

+ a
- 0

Naive Analyses are based on these 2x2 tables:

Disease Outcome

+ a b
- g h

Key

TRUTH

Preferred Reference

+ a b
- c d

Inferior Reference

+ e f
- g h

Missingness of results

- no missings by design
- missings possible
- all missings by design
All other types of DV

Choice of reference standard depends on:

Variable(s) other than the index test alone

Preferred reference

Inferior reference

Disease Outcome

Index

Preferred Reference

Inferior Reference

Key

Missingness of results

- no missings by design
- missings possible
- all missings by design
Differential Verification Patterns

Choice of reference standard depends on:

- Index test results
  - Index test
    - Preferred reference
      - Index +
        - Disease Outcome +
          - a
          - b
        - Disease Outcome -
          - g
          - h
      - Index -
        - Disease Outcome +
          - a
          - b
        - Disease Outcome -
          - g
          - h
  - Index -
    - Disease Outcome +
      - a
      - b
    - Disease Outcome -
      - g
      - h

- Variable(s) other than the index test alone
  - Index +
    - Disease Outcome +
      - a
      - b
    - Disease Outcome -
      - e
      - f
  - Index -
    - Disease Outcome +
      - c
      - d
    - Disease Outcome -
      - g
      - h

Key:
- | Preferred Reference | Inferior Reference |
- | + | + |
- | a | e |
- | b | f |
- | c | g |
- | d | h |

Missingness of results:
- Green: no missings by design
- Blue: missings possible
- White: all missings by design

Naive Analyses are based on these 2x2 tables:
Differential Verification Bias

**Pap Smear has imperfect accuracy** (Sens=.7, Spec=1)

<table>
<thead>
<tr>
<th>VIA</th>
<th>Colposcopy + Biopsy</th>
<th>Pap smear</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+ 150</td>
<td>+ 0</td>
</tr>
<tr>
<td></td>
<td>- 150</td>
<td>- 0</td>
</tr>
<tr>
<td>-</td>
<td>+ 0</td>
<td>+ 70</td>
</tr>
<tr>
<td></td>
<td>- 0</td>
<td>- 630</td>
</tr>
</tbody>
</table>

Estimated accuracy of VIA:
- Sens = 150/(150+70) = 0.68
- Spec = 630/(630+150) = 0.81

**Pap Smear has perfect accuracy** (Sens=1, Spec=1)

<table>
<thead>
<tr>
<th>VIA</th>
<th>Colposcopy + Biopsy</th>
<th>Pap smear</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+ 150</td>
<td>+ 0</td>
</tr>
<tr>
<td></td>
<td>- 150</td>
<td>- 0</td>
</tr>
<tr>
<td>-</td>
<td>+ 0</td>
<td>+ 100</td>
</tr>
<tr>
<td></td>
<td>- 0</td>
<td>- 600</td>
</tr>
</tbody>
</table>

Estimated accuracy of VIA:
- Sens = 150/(150+100) = 0.60
- Spec = 600/(600+150) = 0.80
Correcting for DV bias

• Must know the verification pattern
• Bayesian model\(^1\)
  • takes into account the verification pattern and imperfectness of one or both RSs
• Perform both RSs in a (random) subgroup
  • improves accuracy estimates of the inferior RS

1) de Groot et al (2011) Epidemiology
Reporting Recommendations

• Flow charts of DV pattern
• Reason(s) why patients received which RS
• Estimate or educated guess on the accuracy of the inferior RS
• Separate 2x2 tables for each RS
Reporting Example

Patients suspected of DVT (n=1100)

Excluded (n=90)

Patients included (n=1010)
Judging the risk of bias

- Choice of RS completely dependent on index test results?
- PPV and NPV are clinically interpretable
- Accuracy of inferior RS?
- % of diagnosed by inferior RS?
- Follow-up:
  - Detects all cases at time of index test, but no new cases?
  - Detects same type of cases?
- Higher accuracy of the inferior RS, lower risk of bias
- If a negligible %, risk of bias is low
- If yes to both, risk of bias is low
Key Points

• DV is a common situation in DTA studies in which different groups of patients receive different RSs

• DV may introduce bias

• Reporting recommendations
  • verification pattern
  • 2x2 per RS
  • accuracy estimate of alternative RS

• How to judge risk of bias:
  • accuracy of the inferior RS
  • verification pattern
Evaluating Diagnostic Accuracy in the Face of Multiple Reference Standards

Christiana A. Naaktgeboren, MPH; Joris A.H. de Groot, PhD; Maarten van Smeden, MSc; Karel G.M. Moons, PhD; and Johannes B. Reitsma, MD, PhD

A universal challenge in studies that quantify the accuracy of diagnostic tests is establishing whether each participant has the disease of interest. Ideally, the same preferred reference standard would be used for all participants; however, for practical or ethical reasons, alternative reference standards that are often less accurate are frequently used instead. The use of different reference standards across participants in a single study is known as differential verification.

Differential verification can cause severely biased accuracy estimates of the test or model being studied. Many variations of differential verification exist, but not all introduce the same risk of bias. A risk-of-bias assessment requires detailed information about which participants receive which reference standards and an estimate of the accuracy of the alternative reference standard. This article classifies types of differential verification and explores how they can lead to bias. It also provides guidance on how to report results and assess the risk of bias when differential verification occurs and highlights potential ways to correct for the bias.

Ann Intern Med. 2013;159: * * * FILL THIS IN * * *

For author affiliations, see end of text.
References

1) de Groot (2011) BMJ
6) Zuithoff et al (2009) Fam Practice
7) Gupta (2011) Clin Infect Dis
8) de Groot et al (2011) Epidemiology