Reader Studies for Diagnostic Radiopharmaceuticals

Thomas Gwise, Ph.D.
Deputy Director Biometrics V
Office Biostatistics, FDA, CDER
Outline

• Statistician’s Perspective
• Data evaluation of diagnostics for market approval
  – Diagnostics background
  – Approaches
  – Challenges
• Example
A Diagnostic Modality

- **Definition**
  - A method (drug, device, etc.) used to ascertain the presence of absence of a condition.
    - (abstraction, not necessarily diagnose a disease, per se.)

- **Diagnostic (dx) modality fits within patients’ workup**

- **Examples**
  - **Drugs:** Radio-pharmaceuticals, contrast agents, dyes, Allergen skin tests
  - **Others:** Imaging devices, Monitors, In-vitro diagnostics
Regulatory Perspective

*Drugs*

- Safe and Effective
  - 21CFR314.125 (reasons for refusal to approve)
  - 21CFR314.126 (defines adequate, well controlled)

- Labeling medical products
  - 21CFR201 (drug label requirements)
Differences for Diagnostics

• Typically no therapy intended

• “Effective”-- based on value of information

• “Safe”– value of information outweighs:
  – Risk of adverse effects of drug
  – Risk of any radiation dose received
  – *Risk associated with False (+) or False (-)

• *Safety, effectiveness are linked together with the intended use of the product
Diagnostic Guidance Documents

• Drugs
  – 21CFR315.5
  – 21CFR315.6
  – Guidance for Industry, Developing Medical Imaging Drug and Biological Products, Parts 1-3

• Devices
  – 21CFR809.10
  – Guidance for Industry and FDA Staff, Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests
Measurement

- Trueness
  - Do you hit the target?

- Precision
  - Can you do it repeatedly?
Measurement

Bias

True Value

Precision
Imaging vs. IVD

• IVD
  – Trueness- continuous outcomes
    • Compare to reference using regression
    • [pre] Select cut-off, determine sensitivity (Se) & specificity (Sp)
  – “Precision”
    • Sites, operators, platforms

• Imaging
  – Trueness: abstract scale in the reader’s mind- can be defined...
    probability of disease, etc
  – “Precision” – A more difficult concept
    • Within reader, between reader, sites, platforms, within patient
      (movement, etc)
Reader Studies

• Reader makes determination [ (+) or (-) ], ordinal score, nominal classification, p(disease)

• Impact of a product on readers’ performance

  – Consider a physician making diagnoses:
    • Without CT
    • With CT
    • With CT & contrast

• Expect incrementally better dx performance for each added risk
Efficacy.....
One Possible Approach
(rarely done)
Approach One

• Pro:
  – Intended use
  – Impact on Patients

• Con:
  – Large N
  – Long time

• Approach one not popular ==> most diagnostic studies do not get “bias protection” from randomized & blinded assignments
Typical Study Outline

• A set of specimens (images, serum, patients, etc.) is selected
• Specimens are examined with the experimental diagnostic
• Each diagnostic output is compared to that specimen’s true state (using “gold standard”)
• Calculate Dx performance (i.e., Se, Sp)
• Do same with a control Dx, on same specimens (paired design)
• Compare: Experimental VS. Control
General Considerations for Reader Studies
Generally Applicable to Reader Studies:

- **Samples:** representative of population?
  - Not random sample

- **Readers:** representative of population?
  - Not random sample, often experts
  - Other biases associated with readers

- **Spectrum effect can impact Se & Sp, ROC AUC, NPV & PPV, Agreement**
  - Control modality informative
Spectrum Effect

Positives

Negatives
Example: “Amyloid Imager”

- Indication: Radioactive diagnostic for PET imaging of the brain to estimate β-amyloid neuritic plaque density.
- A negative “amyloid imager” scan indicates sparse to no neuritic plaques, and is inconsistent with a neuropathological diagnosis of AD.
- A positive “amyloid imager” scan indicates moderate to frequent amyloid neuritic plaques; neuropathological examination has shown this amount of amyloid neuritic plaque is present in patients with AD, but may also be present in patients with other types of neurologic conditions as well as older people with normal cognition.
- “amyloid imager” is an adjunct to other diagnostic evaluations.
Data

- Autopsy is truth standard
- Difficult to collect data
- Delayed verification
- Negatives – few under truth standard

Healthy

Young Healthy Volunteers

“Specificity Population” (47)

(Most interesting population)

Impaired

Mild Cognitive Impairment

Clearly Impaired

Symptomatic

Brain donations

Correlation study (29)
Pragmatic Approach

• Study Correlation:
  – Is there a relationship between image and deposited β-amyloid?
• Diagnostic Performance
  – Sensitivity & Specificity
• Agreement Study
  – Strengthen support
  – Evaluate precision
## “Amyloid Imager” Results

### Table 8: Number of Positive Amyloid Scan Results within Study Three Subject Groups and Reproducibility of Scan Results Among Readers

<table>
<thead>
<tr>
<th>Subject group by cognitive and truth standard (TS, autopsy) status</th>
<th>Positive Scans, n&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Kappa (95% CI)</th>
<th>Percent of Scans with Inter-reader Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>All subjects with a TS, n=59</td>
<td>33</td>
<td>0.75 (0.67, 0.83)</td>
<td>14 (3 of 5 readers agree)</td>
</tr>
<tr>
<td>All subjects without a TS, n=92</td>
<td>33</td>
<td>0.88 (0.82, 0.94)</td>
<td>2 (3 of 5 readers agree)</td>
</tr>
<tr>
<td>AD, n=49 (29 with TS; 20 no TS)</td>
<td>38</td>
<td>0.67 (0.58, 0.76)</td>
<td>10 (3 of 5 readers agree)</td>
</tr>
<tr>
<td>MCI, n=57 (5 with TS; 52 no TS)</td>
<td>17</td>
<td>0.91 (0.83, 0.99)</td>
<td>2 (3 of 5 readers agree)</td>
</tr>
<tr>
<td>Cognitively normal without TS, n=20</td>
<td>4</td>
<td>0.83 (0.69, 0.97)</td>
<td>5 (3 of 5 readers agree)</td>
</tr>
<tr>
<td>Cognitively normal with TS, n=12</td>
<td>1</td>
<td>0.73 (0.55, 0.87)</td>
<td>0 (3 of 5 readers agree)</td>
</tr>
<tr>
<td>Other (non-AD) dementia with TS, n=13</td>
<td>7</td>
<td>0.52 (0.35, 0.69)</td>
<td>23 (3 of 5 readers agree)</td>
</tr>
</tbody>
</table>

<sup>a</sup> Shown is the median number of scans interpreted as positive across the 5 readers for each subgroup of patients listed in the first column.

(Table 8 was taken from typical Label)

Se: 82%; Sp: 95%; K: 0.83; Adequate Correlation

Yang, et al., NEJM 367;10 p885
Conclusion

• Reader studies are challenging

• The Division of Medical Imaging Products will help with difficult study problems
Thank You
References

- Yang, L., Rieves, D., and Ganley, C. Brain Amyloid Imaging — FDA Approval of Florbetapir F18 Injection NEJM 367;10 p885
References

• Shourkri, MM. Asyali, MH, and Donner, A. Sample size requirements for the design of reliability study: review and new results Statistical Methods in Medical Research 2004; 13: 251-271
• Giraudeau, B. and Mary, JY. Planning a reproducibility study: how many subjects and how many replicates per subject for an expected width of the 95 per cent confidence interval of the intraclass correlation coefficient, Statistics in Medicine 2001; 20:3205–3214
<table>
<thead>
<tr>
<th></th>
<th>(N=48)</th>
<th>Intra-Reader</th>
<th>Inter-Reader</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>A</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>C</td>
<td>A v B</td>
<td>A v C</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>B v C</td>
<td></td>
</tr>
<tr>
<td>Alzheimers</td>
<td>1.00</td>
<td>0.73</td>
<td>0.54</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>100%</td>
<td>96%</td>
<td>94%</td>
<td>88%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.24</td>
<td>90%</td>
<td></td>
</tr>
<tr>
<td>Multi-infarct Dementia</td>
<td>-----</td>
<td>100%</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>96%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Infarct</td>
<td>0.65</td>
<td>100%</td>
<td>-0.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td>96%</td>
<td></td>
<td>90%</td>
<td></td>
</tr>
<tr>
<td>Frontal Lobe Dementia</td>
<td>0.48</td>
<td>1.00</td>
<td>0.48</td>
<td></td>
</tr>
<tr>
<td></td>
<td>96%</td>
<td></td>
<td>94%</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>0.45</td>
<td>90%</td>
<td>-0.08</td>
<td></td>
</tr>
<tr>
<td></td>
<td>92%</td>
<td></td>
<td>85%</td>
<td></td>
</tr>
<tr>
<td>No Distinct Pattern</td>
<td>0.38</td>
<td>0.06</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td></td>
<td>69%</td>
<td>33%</td>
<td>54%</td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>0.49</td>
<td>100%</td>
<td>0.11</td>
<td></td>
</tr>
<tr>
<td></td>
<td>81%</td>
<td></td>
<td>79%</td>
<td></td>
</tr>
<tr>
<td>Heterogeneous</td>
<td>0.57</td>
<td>0.30</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td></td>
<td>85%</td>
<td>77%</td>
<td>46%</td>
<td></td>
</tr>
<tr>
<td>Atrophy</td>
<td>0.67</td>
<td>100%</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td></td>
<td>90%</td>
<td></td>
<td>54%</td>
<td></td>
</tr>
</tbody>
</table>

•Data from: Stockbridge, et al. Nuclear Medicine Communications, 2002, 32, 537-554