S2-07
An international Ki67 reproducibility study: Harmonizing scoring methodology

Dr. Nielsen: Consulting Fees, Bioclassifier LLC; Nanostring Technologies.
Dr. Polley: Nothing to disclose.
On Behalf of the international Ki67 in Breast Cancer Working Group of BIG-NABCG
Background

- Proliferation matters: prognostic in ER+ breast Ca
- Ki67: widely used IHC assay. Value for subtyping, prognosis, prediction, neoadjuvant endpoint ...
- ASCO does not recommend Ki67 for clinical practice due to concerns about analytical reproducibility.

Lessons learned from Phase 1

**Local scoring methods on 100 centrally stained TMA serial sections**


- **Intra**-observer consistency was good
  - ICC = 0.94 (95% CI 0.93, 0.97)
- **Inter**-observer reproducibility was not satisfactory
  - ICC = 0.71 (95% CI 0.47, 0.78)
- **Scoring method** was a major source of Ki67 differences.
- Formal counting methods yielded more consistent results.
Can reproducibility be improved?

Phase 2:

→ Can Ki67 scorers be “trained” in a common visual scoring method, that might be taken forward to clinical use?

→ Can we develop a common reference tool for clinical studies?

(a) Web-based calibration, followed by
(b) Standardized scoring on glass TMA slides

Phase 2 (a): Calibration

• 9 training + 9 test Web-based TMA images
  o Centrally-stained, representing the range of Ki67 scores

• Practical scoring method with good internal consistency chosen

• Simple instructions with visual examples of “positive” and “negative” nuclei

• 19 labs from around world (including 2 ref labs)

• Pre-specified criteria for level of agreement with reference labs.
Web-based standard images

Click-tracking app to mark counted nuclei
"typewriter" pattern, 500 total invasive cancer nuclei

Lab A scored as Ki67=25%
Lab B scored as Ki67=48%

red = scored as positive
green = scored as negative

Different interpretations of “brown”

Red = scored as positive
Green = scored as negative

Lab E
Lab G
Calibration: results

- Labs differed on what threshold of “brown” they considered positive.
  - Added example images to instructions, showing what staining should be considered positive/negative
- Scorers repeated training set analyses until met preset criteria for variability and outlier scores
- Test set on new cases: 12 of 17 labs passed preset stringent agreement criteria on first attempt

Conclusion: Labs were “trainable” – performance showed trend for improvement following use of the web-based calibration tool

Phase 2 (b): Can a consistent Ki67 index be delivered by calibrated, trained observers with a standardized formal visual counting method?

- 50 centrally-stained 1mm TMA cases from phase 1
- 3 sections from same TMA block distributed
- 16 labs in Canada, U.S., Europe, Japan
- Labs first completed calibration on web images
- Applied same scoring method on glass
- Key-stroke application captured count data
Time to count 500 tumor nuclei, Phase 2

N = 735 counted cores in which a full 500 cells were counted

Median = 5.6 minutes per core

Phase 2 (b): Criteria for success

ICC = Intraclass correlation coefficient

Proportion of total variance (in measurements across patients and laboratories) that is attributable to the biological variability among patients’ tumors.

Range of 0–1

1 = highest agreement

Prespecified Goal: Achieve ICC consistent with a true value of 0.9 and significantly >0.7
## Results from Phase 1:

<table>
<thead>
<tr>
<th>Phase, Method</th>
<th>ICC (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICCs reflecting labs’ scoring different sections of same TMA block</td>
<td></td>
</tr>
<tr>
<td>Phase 1 Local staining, Local scoring method (8 labs)</td>
<td>0.59 (0.37, 0.68)</td>
</tr>
<tr>
<td>Phase 1 Central staining, Local scoring method (8 labs)</td>
<td>0.71 (0.47, 0.78)</td>
</tr>
</tbody>
</table>

---

## Phase 2: major and significant improvement in ICC

<table>
<thead>
<tr>
<th>Phase, Method</th>
<th>ICC (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICCs reflecting labs’ scoring different sections of same TMA block</td>
<td></td>
</tr>
<tr>
<td>Phase 1 Local staining, Local scoring method (8 labs)</td>
<td>0.59 (0.37, 0.68)</td>
</tr>
<tr>
<td>Phase 1 Central staining, Local scoring method (8 labs)</td>
<td>0.71 (0.47, 0.78)</td>
</tr>
<tr>
<td>Phase 2 Central staining, Standardized scoring method (16 labs)</td>
<td>0.92 (0.88, 0.95)</td>
</tr>
</tbody>
</table>
### Phase 1: Labs scored different sections of same TMA block

### Phase 2: 3 groups, single section scored within each group

<table>
<thead>
<tr>
<th>Phase, Method</th>
<th>ICC (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ICCs reflecting labs’ scoring different sections of same TMA block</strong></td>
<td></td>
</tr>
<tr>
<td>Phase 1 Local staining, Local scoring method (8 labs)</td>
<td>0.59 (0.37, 0.68)</td>
</tr>
<tr>
<td>Phase 1 Central staining, Local scoring method (8 labs)</td>
<td>0.71 (0.47, 0.78)</td>
</tr>
<tr>
<td>Phase 2 Central staining, Standardized scoring method (16 labs)</td>
<td>0.92 (0.88, 0.95)</td>
</tr>
<tr>
<td><strong>ICCs reflecting labs’ scoring exact same section</strong></td>
<td></td>
</tr>
<tr>
<td>Phase 2 Group 1 labs (7 labs)</td>
<td>0.92 (0.79, 0.95)</td>
</tr>
<tr>
<td>Phase 2 Group 2 labs (4 labs)</td>
<td>0.96 (0.78, 0.97)</td>
</tr>
<tr>
<td>Phase 2 Group 3 labs (5 labs)</td>
<td>0.95 (0.77, 0.97)</td>
</tr>
<tr>
<td>Phase 2 Combined Groups (16 labs)</td>
<td>0.94 (0.90, 0.97)</td>
</tr>
</tbody>
</table>

This presentation is the intellectual property of the author/presenter. Contact presenter@email.ubc.ca for permission to reprint and/or distribute.
Spaghetti plots: Visualizing variability in individual cases

Phase 1 Study (n=37)

Local scoring method, each lab scored different section

(7 labs common to both phases 1 and 2)

Spaghetti plots: Ki67 of 10-20% (7 labs common to both phases)

Phase 1 Study (n=37)

Local scoring method, each lab scored different section

37 cases scored by ≥ 1 lab as 10-20%.
0 of the 37 scored by all labs as 10-20%.

Phase 2 Study (n=25)

Standardized scoring method, each lab scored exact same section

25 cases scored by ≥ 1 lab as 10-20%.
0 of the 25 scored by all 7 labs as 10-20%.
1 case, scored by 5 of the 7 labs, was scored by all 5 labs as 10-20%.
**Spaghetti plots: Ki67 of 10-20% (7 labs common to both phases)**

**Phase 1 Study (n=37)**

Local scoring method, each lab scored different section.

37 cases scored by ≥1 lab as 10-20%.
0 of the 37 scored by all labs as 10-20%.

**Phase 2 Study (n=25)**

Standardized scoring method, each lab scored exact same section.

25 cases scored by ≥1 lab as 10-20%.
0 of the 25 scored by all 7 labs as 10-20%.
1 case, scored by 5 of the 7 labs, was scored by all 5 labs as 10-20%.

---

**Kappas for well-known Ki67 cutoffs, based on phase 2 data (exploratory analysis)**

<table>
<thead>
<tr>
<th>Ki67 cutoff</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤ 2.7% vs. &gt; 2.7% (relevant to PEPI score = 0)</td>
<td>0.78</td>
</tr>
<tr>
<td>≤ 5% (low) vs. &gt; 5% (high)</td>
<td>0.91</td>
</tr>
<tr>
<td>≤ 10% (low) vs. &gt; 10% (high)</td>
<td>0.75</td>
</tr>
<tr>
<td>≤ 13.25% (low) vs. &gt; 13.25% (high) (cutoff from Cheang et al.)</td>
<td>0.74</td>
</tr>
<tr>
<td>&lt; 14% vs. ≥ 14% (St. Gallen 2011 Luminal A vs. Luminal B/HER2- N0)</td>
<td>0.73</td>
</tr>
<tr>
<td>≤ 20% (low) vs. &gt; 20% (high)</td>
<td>0.70</td>
</tr>
</tbody>
</table>

---

This presentation is the intellectual property of the author/presenter. Contact the author for permission to reprint and/or distribute.
**Variability at a 14% cutoff between labs with highest & lowest Ki67 median scores**

Comparing the 2 most discrepant labs ...

At a < 14% vs. > 14% cutoff (2011 St. Gallen), there are 8/46 (17%) cases that Lab C would call high Ki67, but Lab E would call low Ki67.

---

**Phase 2: Conclusions**

- Web-based calibration system is useful tool
- Ki67 counting method appears practical
- All ICCs significantly > 0.70 as targeted in study design
- Observed ICC ≥ 0.90 for both within- and between-section scoring
Phase 2: Some caveats

- Clinically important discrepancies persisted.
- Applies only to centrally stained TMAs
  - core biopsies (subject of Phase 3)
  - whole sections
  - added variability from staining
- Clinical validity of this scoring method still needs to be confirmed.