Ki67 as a biomarker of proliferation:

An international Ki67 reproducibility study

Torsten O. Nielsen, MD/PhD, FRCPC
Professor of Pathology and Laboratory Medicine
University of British Columbia, Vancouver, Canada

On behalf of the International Ki67 in Breast Cancer Working Group of the BIG-NABCG collaboration

Outline of Presentation

1) Ki67 as a biomarker
2) Variability in visual scoring
3) Improving analytical validity

Breast Cancer Subtyping by IHC

<table>
<thead>
<tr>
<th>ER</th>
<th>PR</th>
<th>HER2</th>
<th>Ki67</th>
<th>CK5/6</th>
<th>EGFR</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td></td>
<td></td>
<td>≤13%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>+</td>
<td>+/-</td>
<td>+/-</td>
<td>&gt;13%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
</tbody>
</table>

Lum A
Lum B
HER2E
Basal-like

Clinical value of Ki67 index

- Prognosis (e.g. among luminal breast cancers)
- Predict drug response (e.g. taxanes)
- Triage need for (or replace) Oncotype test
- Eligibility criterion for clinical trials
- Endpoint for neoadjuvant response
- Intermediate endpoint in adaptive clinical trials – precipitating change of agent

Many and increasing indications for Ki67 measurements!

Ki67

- One of many possible proliferation markers we could be measuring
- Chosen from 1000s of possibilities b/c of excellent IHC performance characteristics
- Expressed in nuclei in G1, S, G2 and on chromosomes in M phase, then rapidly catabolized (no expression in G0)

Some classic references:

Ki67: an antigen with special advantages

Among 1000s of possible proliferation markers, the pattern of Ki67 antigen expression is particularly favourable!

- 345 & 395 kD isoforms both include a unique C-terminal "Ki67 domain": 22AA repeated sixteen times over = an exceptionally specific and sensitive epitope, recognized by MIB-1
- After Ag retrieval, Ki67 Abs (MIB-1) work on FFPE sections– even 60 y.o. specimens
IHC assay for Ki67

Ki67 has more favourable technical characteristics for IHC than almost any other assay

Handling recommendations for ER/HER2 are more than appropriate for Ki67

2011 JNCI publication

Assessment of Ki67 in Breast Cancer: Recommendations from the International Ki67 in Breast Cancer Working Group

Recommendations for:
- Pre-analytical setting
- Analytical setting
- Interpretation & scoring
- Data analysis

“The Ki67 score or index should be expressed as the percentage of positively staining cells among the total number of invasive cells in the area scored.”

“Cut points for prognosis, prediction, and monitoring should only be applied if the results from local practice have been validated against those in studies that have defined the cutoff for the intended use of the Ki67 result.”

Proliferation markers

- Methods studied in breast cancer
  - Radioactive thymidine uptake
  - Flow cytometry S-fraction
  - IHC measures – Ki67, PCNA
  - RNA expression measures – OncotypeDX, PAM50

- All reported as prognostic, perhaps predictive
- ASCO Tumor Marker Guidelines panel has never recommended use of ANY proliferation marker
  … not because they lack clinical validity, but because of poor analytical reproducibility

Analytical validity of Ki67

- Can we deliver reliable Ki67 results on breast cancer?
- To what level of reproducibility can pathologists reliably quantify Ki67 staining?

How should the counting be done?

- Little firm data, across observers …

  • # cells you need to count
  • TMA vs core vs section
  • How to deal with “hotspots”
  • Intra-observer variability
  • Interobserver variability

perceptible “hot spots” of ↑Ki67 were seen in 61% of slides …
**Systematic approach**

Isolate sources of scoring variability across observers

TMA slides, local methods (“Phase 1”)

- TMA make it easy to distribute many cases to many observers, and limit variability in what areas are scored
- Web-based calibration

“Phase 2”

TMA slides, after calibration + standard method

Core biopsies & whole sections

---

**Phase 1**: Can experienced labs deliver consistent Ki67 % on the same cases, using their local visual scoring methods?

- 100 breast cancer cases, 1 mm TMA cores
- Scored visually by labs using own scoring methods
- Three experiments:
  - Intra-observer (repeat scoring of same slide)
  - Inter-lab w/ central staining
  - Inter-lab w/ local staining
- Labs from Canada, France, Italy, UK, and USA
- Universities, major cancer centers, a national reference lab

---

**Intra-observer variability for Ki67 index is small**

Results of Phase 1

Intra-observer consistency was good:

- 6 labs scored same 50 cases 3 times
- Overall ICC = 0.94 (95% CI = 0.93, 0.97)
- Formal counting methods yielded more consistent results over visual estimate.

---

Phase 1 intra-observer variability counting vs estimation approach, on same cases

Phase 1 inter-observer variability (centrally stained TMA slides)

median: 10%  
Overall ICC = 0.71  
(95% CI = 0.47, 0.78)  
median: 28%
Lessons learned from Phase 1

TMA glass slides representing 100 breast cancers (serial sections, centrally-stained, scored by local method)

- Intra-observer consistency good, but interobserver variability problematic
- Cut points not freely transferable – local recalibration against clinical endpoint or reference images is needed

Can reproducibility be improved?

Phase 2: Web-based calibration followed by standardized scoring on glass TMA slides

- 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 trial studies?

Phase 2: Calibration portion

- 9 training + 9 test Web-based TMA images
  - 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)
2 labs w/ best intra-lab reproducibility chosen as reference labs → “gold standard” values for training and test sets

**Calibration criteria for success**

*RMSE < 0.6:* Root mean squared differences between volunteer & reference lab scores among the 9 images < 0.6

*MAXDEV < 1.0:* Maximum absolute difference between volunteer & reference lab scores among the 9 images < 1.0

**Web-based standard images**

Click-tracking app, allowing assessment of differences from reference scorers

“typewriter” pattern, 500 total invasive cancer nuclei

Examples of range of staining levels that should be considered

POSITIVE (red squares) for Ki67 (unmarked & marked side by side)

**Examples of range of staining levels that should be considered**

POSITIVE (red squares) for Ki67 (unmarked & marked side by side)

Calibration criteria for success:

RMSE (Passing: < 0.6) MAXDEV (Passing: < 1.0)

<table>
<thead>
<tr>
<th></th>
<th>Training phase</th>
<th>Test phase</th>
<th>Training phase</th>
<th>Test phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.64</td>
<td>0.41</td>
<td>1.57</td>
<td>0.92</td>
</tr>
<tr>
<td>SD</td>
<td>0.40</td>
<td>0.16</td>
<td>1.21</td>
<td>0.40</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.22</td>
<td>0.17</td>
<td>0.36</td>
<td>0.37</td>
</tr>
<tr>
<td>Maximum</td>
<td>1.46</td>
<td>0.63</td>
<td>3.81</td>
<td>1.53</td>
</tr>
<tr>
<td>Median</td>
<td>0.53</td>
<td>0.38</td>
<td>1.03</td>
<td>0.85</td>
</tr>
</tbody>
</table>

Calibration did improve scoring consistency: improvement between training and test is significant.

*RMSE p = 0.044 for RMSE  MAXDEV p = 0.044 for MAXDEV*

Performance statistics on training and test set digital images
Results from calibration “testing”

- 17 participant labs
- 2 reference labs
  - 12 of 17 reached criteria for a “pass” on the test set
  - Non-passes mainly due to MAXDEV on low Ki67 cases

Examples: Passed testing

Examples: Did not pass testing

Lessons learned from calibration

- Labs were “trainable” – performance did improve using the web-based calibration tool
- Labs differed on what threshold of “brown” they considered positive (so we added example images to standardization instructions, showing what level of staining should be considered positive)

Phase 2: Can a consistent Ki67 index be delivered with a standardized formal visual counting method?

- 50 centrally-stained 1 mm core TMA cases used from phase 1 (none of which were used for calibration)
  - 3 sections from same TMA distributed across three groups
  - 16 labs in Canada, U.S., Europe, Japan (includes 7 who were also in phase 1 = “group 1” in analysis)
  - Labs first completed calibration training/testing on web
  - Applied same scoring method on glass
  - Key-stroke application to obtain count data efficiently (may later identify minimum counts needed)
  - Pre-specified criterion for success: observed ICC > 0.9

“Deliverables”

1. Dissemination of analytically valid and clinically meaningful methods for assessing Ki67
2. Gold-standard set of Web-based calibration cases
3. Cell counting tools, with justification for numbers needed to count

Phase 2 conclusions

- web-based calibration system works
- Ki67 counting method appears practical
- target ICC > 0.9 achieved (at least for visual assessment of same TMA section)
Phase 2: some caveats

- Applies only to centrally stained TMAs
- Applies only to core biopsies
- Applies only to whole sections
- Applies only to added variability from staining
- Clinical validity yet to be confirmed

Success! Next moves...

- Findings to be reported at SABCS 2013
- Apply same scoring to core biopsies
- Whole sections (with hot spot issues) might follow later
- Confirm clinical utility of analytically valid method (esp. prognosis)
- Identify major sources of residual variability in staining method (in context of valid, consistent scoring)

Immunohistochemistry

Steps introducing variability

- Warm ischemia time (diagnostic procedure)
- Cold ischemia time (before fixation)
- Choice of and time in fixative
- Processing and block selection
- Time from cutting to IHC (antigen fade)
- Antigen retrieval (>100°C, protease, none)
- Choice of primary Ab
- Detection system (avidin-biotin vs polymer)
- Chromagen and counterstain
- Interpretation

If we fail, due to variability at core biopsy and staining level:

- Test if automated imaging platforms and algorithms can deliver consistent results, or move to RNA methods.
- Failure would confirm that Ki67 index by IHC should only be used after internal validation for a given clinical context, or as a research tool.