An international Ki67 reproducibility study

Presenter: Torsten O. Nielsen, MD, PhD, FRCPC - University of British Columbia, Canada
Mei-Yin C. Polley, PhD - National Cancer Institute, US
Samuel C.Y. Leung, MSc - University of British Columbia, Canada
Mauro G. Mastropasqua, MD - European Institute of Oncology, Italy
Lila A. Zabaglo, PhD - Breakthrough Breast Cancer Research Centre, UK
John M.S. Bartlett, PhD, FRCPath - Ontario Institute for Cancer Research, Canada
Giuseppe Viale, MD, FRCPath - European Institute of Oncology & University of Milan, Italy
Lisa M. McShane, PhD - National Cancer Institute, US
Daniel F. Hayes, MD - University of Michigan, US
Rebecca A. Enos, RN, MPH, The EMMES Corporation, US
Mitch Dowsett, PhD, BSc - Royal Marsden Hospital, UK

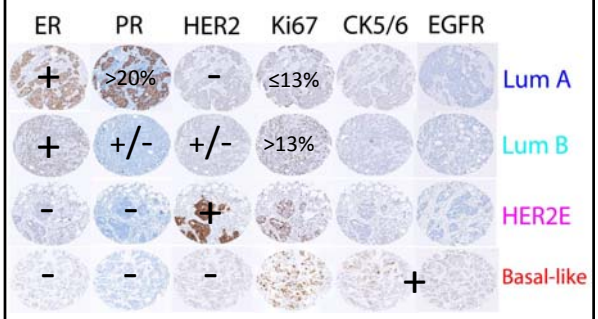
On behalf of the International Ki67 in Breast Cancer Working Group of the BIG-NABCG collaboration

1

Outline of Presentation

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

Breast Cancer Subtyping by IHC



Dowsett M, Nielsen TO et al. Assessment of Ki67 in breast cancer: recommendations from the International Ki67 in Breast Cancer working group. *JNCI*, November 2011

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: an antigen with special advantages

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

- **Unique C-terminal "Ki67 domain" repeated 16 times over: a *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

Mitch Dowsett, Torsten O. Nielsen, Roger A'Hern, John Bartlett, R.Charles Coombes, Jack Cuzick, Matthew Ellis, N.Lynn Henry, Judith C. Hugh, Tracy Lively, Lisa McShane, Soon Paik, Frederique Penault-Llorca, Ljudmila Prudkin, Meredith Regan, Janine Salter, Christos Sotiriou, Ian E. Smith, Giuseppe Viale, Jo Anne Zujewski, Daniel F. Hayes

J Natl Cancer Inst 2011;103:1–9

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."

7

Proliferation markers

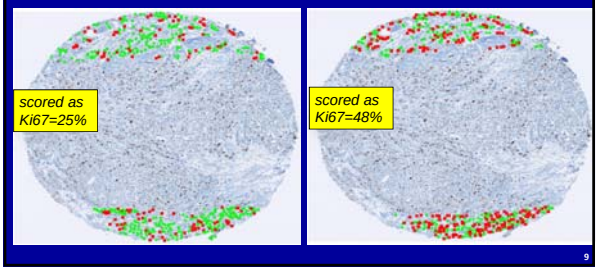
- **Methods studied in breast cancer**
  - Radioactive thymidine updake
  - Flow cytometry S-fraction
  - IHC measures – Ki67, PCNA
  - RNA expression measures – OncotypeDX, PAM50
- **All reported as prognostic, perhaps predictive**
- **ASCO Tumor Marker Guidelines panel has never recommended use of ANY proliferation marker**

  … not because they lack clinical validity, but because of poor analytical reproducibility

8

Analytical validity of Ki67

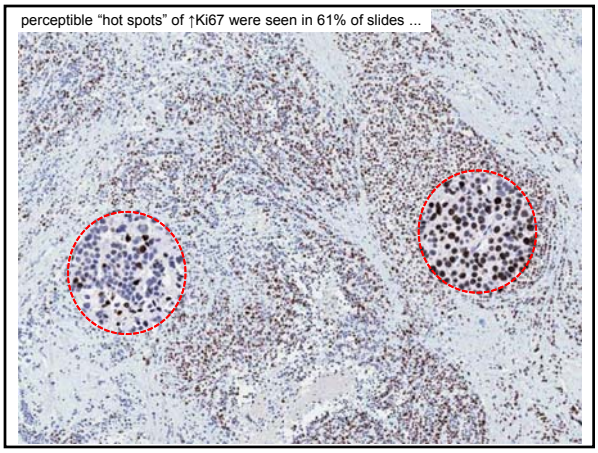
- *Can we deliver reliable Ki67 results on breast cancer?*
- *To what level of reproducibility can pathologists reliably quantify Ki67 staining?*



9

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



Systematic approach

Isolate sources of scoring variability across observers

TMA slides, local methods ("Phase 1")

TMAs 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

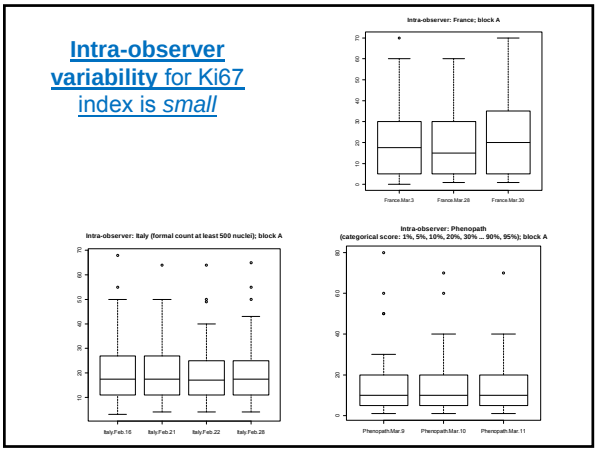
12

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

13

Intra-observer variability for Ki67 index is *small*



Results of Phase 1

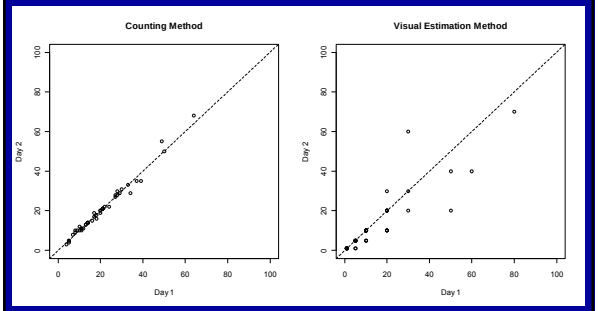
Intra-observer consistency was good:

- 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.

15

Phase 1 intra-observer variability

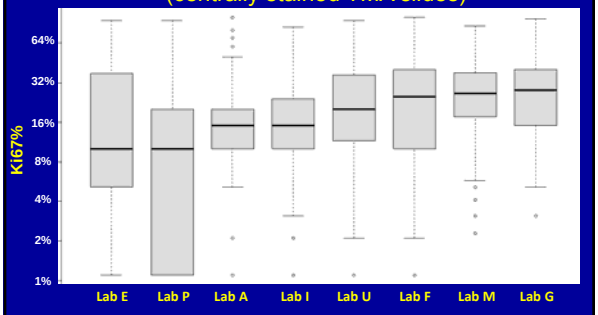
counting vs estimation approach, on same cases



16

Phase 1 inter-observer variability

(centrally stained TMA slides)



median: 10%

Overall ICC = 0.71
(95% CI = 0.47, 0.78)

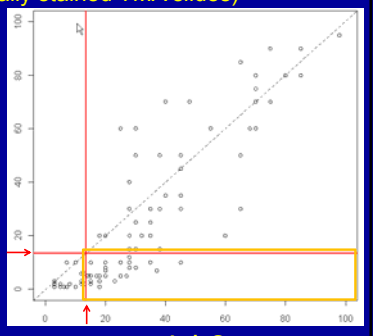
median: 28%

17

Phase 1 inter-observer variability

(centrally stained TMA slides)

At a hypothetical **13.5%** cutoff, there are **32%** cases that Lab G would call high Ki67, but Lab E would call low Ki67.



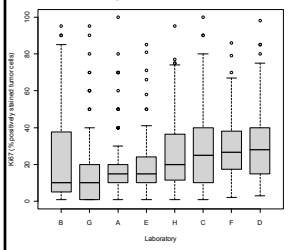
18

Phase 1 Results

100 cases; local scoring methods; 8 labs from Canada, France, Italy, UK, and USA

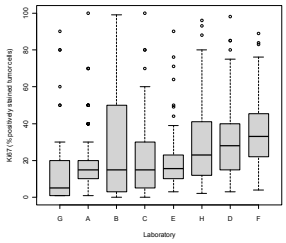
Inter-laboratory reproducibility

Central staining: Median Ki67 10% to 28%



ICC = 0.71, 95% CI = (0.47, 0.78)

Local staining: Median Ki67 5% to 33%



ICC = 0.59, 95% CI = (0.37, 0.68)

Intra-laboratory reproducibility: 6 labs, 50 cases, 3 replicate scorings/laboratory
ICC = 0.94, 95% CI = (0.93, 0.97)

19

Lessons learned from Phase 1

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

20

Lessons learned from Phase 1 (continued)

- Although staining method added some variability, the major source of Ki67 differences was *scoring method*:
  - Estimation vs. Counting
  - Choice of areas to count
  - Invasive Cancer vs. other cells
  - Threshold of brown considered "positive"

21

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?

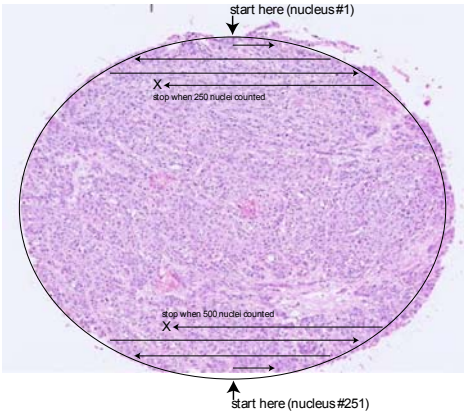
→Can we develop a common reference tool for clinical trial studies?

22

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
- 16 labs from around world
- Continuing with additional labs

23

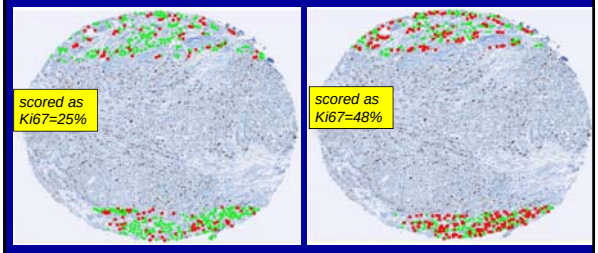


24

Web-based standard images

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

"typewriter" pattern, 500 total invasive cancer nuclei



25

Labs w/ highest inter- and intra-lab reproducibility chosen as reference labs. Their average $\log_2$ transformed Ki67 scores = 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

and

MAXDEV < 1.0:

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

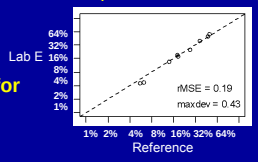
26

Results from calibration "testing"

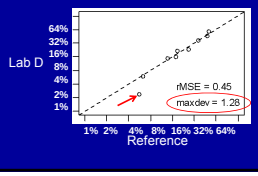
16 participant labs
2 reference labs

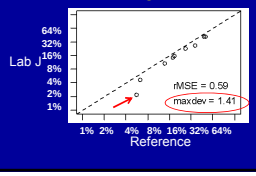
- **9 of 16 reached criteria for a "pass" on the test set**

Example: Passed testing



Examples: Did not pass testing





27

Performance statistics on training and test set digital images

	RMSE (Passing: < 0.6)		MAXDEV (Passing: < 1.0)	
	Training phase (1st attempt)	Test phase	Training phase (1st attempt)	Test phase
Mean	**0.64**	**0.41**	**1.57**	**0.92**
SD	0.40	0.16	1.21	0.40
Minimum	0.22	0.17	0.36	0.37
Maximum	1.46	0.63	3.81	1.53
Median	0.53	0.38	1.03	0.85

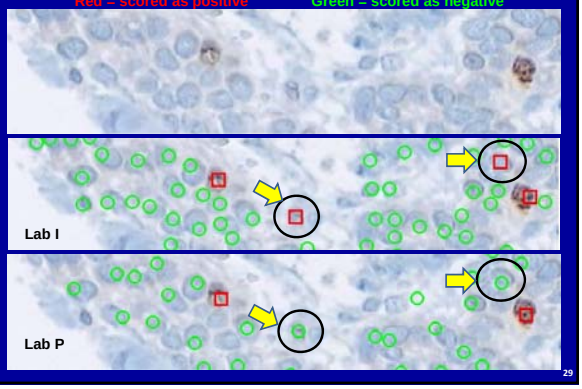
Calibration did improve scoring consistency: differences between training and test were significant. p = 0.044 for RMSE p = 0.044 for MAXDEV

(n.b. preliminary result; definitive statistical analyses pending)

28

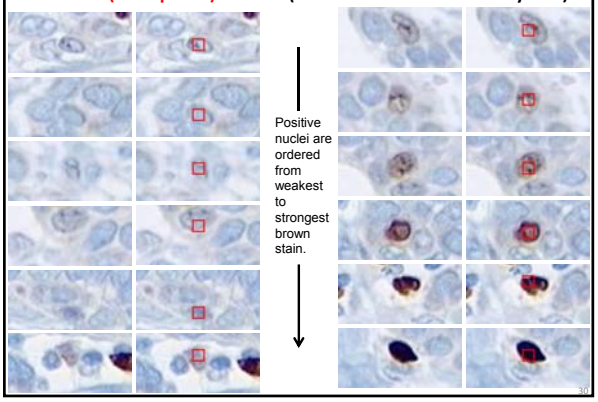
Different interpretations of "brown"

Red = scored as positive Green = scored as negative



29

Examples of range of staining levels that should be considered POSITIVE (red squares) for Ki67 (unmarked & marked side by side)



30

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.
- Added example images to standardization instructions, showing what level of staining should be considered positive

31

Phase 2: Can a consistent Ki67 index be delivered with a standard, formal visual counting method on TMA glass slides?

- 50 centrally-stained 1 mm core TMA cases
- Labs first complete calibration test
- Similar scoring, but on glass
- Key-stroke application to gather count data, to later identify minimum counts needed
- Pre-specified criterion for success: ICC > 0.9
- NIH to report results from 16 labs - May 2013…

32

"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

33

If we are successful…

- Apply same scoring to core biopsies
  - ➢Whole sections (with hot spot issues) would follow later
- Confirm clinical utility of analytically valid method (i.e. prognosis, prediction, neoadjuvant endpoints)
- Define levels of expected residual variability in "best practice" method

34

If we fail…

- 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.

35

BIG-NABCG Ki67 Working Group (all phases)

Centre Jean Perrin, France:
Frédérique Penault-Llorca, MD, PhD
Ines Raoelfils, MD

Baylor College of Medicine, US
Carolina Gutierrez, MD, FCAP
C. Kent Osborne, MD

Dana-Farber Cancer Institute, US:
Meredith Regan, ScD

Denmark: Region Sjaelland
Anne-Vibeke Lænkholm, MD

European Institute of Oncology, Italy
Mauro Mastropasqua, MD
(and University of Milan):
Giuseppe Viale, MD, FRCPath

Fred Hutchinson Cancer Research Center, US:
Ming-Gang Lin, MD
Peggy Porter, MD

Heart of England NHS (Birmingham):
Jane Starczynski, PhD

Imperial College London, UK:
Charles Coombes, FRCP

Institute of Cancer Research, UK / Royal Marsden Hospital:
Roger A'Hern, MSc
Mitch Dowsett, PhD, BSc
Janine Salter, PhD
Ian Smith, FRCP, FRCPE
Lila Zabaglo, PhD

Leads: L McShane, D Hayes, M Dowsett, T Nielsen

Indiana University Simon Cancer Ctr
Sunil Badve, MBBS, MD, FRCPath

Japan:
Takuya Moriya, MD, PhD (Kawasaki)
Takashi Sakatani, MD, PhD (Jichi)
Dr. Yasuyo Ohi (Sagara)

Jules Bordet Institute, Belgium:
Denis Larsimont, MD, PhD
Roberto Salgado, MD, PhD
Christos Sotiriou, MD, PhD

Lund University, Sweden:
Signe Borgquist, MD, PhD
Dorthe Grabau, MD, PhD

Massachusetts General Hosp., US:
Jose Baselga, MD, PhD

McMaster University, Canada
Anita Bane, PhD, MB, BAO, BCh

MD Anderson Cancer Center, US:
Fraser Symmans, MD

Montefiore / Albert Einstein, US:
Susan Fineberg, MD

NSABP:
Soonmyung Paik, MD

Ontario Institute for Cancer Research, Canada:
John Bartlett, BSc, PhD, FRCPath
Mary Anne Quintayo, BSc

PhenoPath, US:
Allen Gown, MD
Patricia Kandalaft, MD

University of Alberta, Canada:
Judith Hugh, MD

University of British Columbia, Canada:
Torsten Nielsen, MD, PhD, FRCPC
Doris Gao, MD
Samuel Leung, MSc
Erika Mehl, BMLSc

University of Edinburgh, UK
Tammy Piper

University of Michigan, US
Daniel Hayes, MD
Lynn Henry, MD, PhD

Vall d'Hebron University Hospital, Spain:
Ludmila Prudkin, MD

Washington University, US:
Matthew Ellis MB, BChir, PhD

Wolfson Institute of Preventive Medicine, UK:
Jack Cuzick, PhD

National Cancer Institute, US:
Tracy Lively, PhD
Lisa McShane, PhD
Mei-Yin Polley, PhD
Jo Anne Zujewski, MD
EMMES: Rebecca Enos, RN, MPH

36