Observer variability in the quantitative assessment of tissue-based biomarkers

Histological Image Analysis Workshop
Pathology Informatics 2011, Pittsburgh, PA

Marios A Gavrielides, Ph.D.

Division of Imaging and Applied Mathematics (DIAM),
Office of Science and Engineering Laboratories
Center for Devices and Radiological Health
U.S. Food and Drug Administration

Brandon Gallas, Ph.D (DIAM), Stephen M Hewitt, M.D., Ph.D., (NCI/NIH)
Overview

• **Introduction**
  - Tissue-based cancer biomarkers
  - Assessment of biomarker expression with immunohistochemistry (IHC)
  - Variability involved in IHC evaluation

• **Observer study**
  - Examining the use of computer-aided assessment
    - HER2 IHC evaluation

• **Related ongoing research at DIAM**
Overview-Cancer biomarkers

• Biological markers (biomarkers) can identify characteristics linked to tumor behavior
  • can lead to improved clinical decisions
  • specific to individual patients

• Uses on all aspects of cancer:
  • diagnosis, staging, prognosis, treatment selection, monitoring treatment
Biomarkers for breast cancer

- HER2/neu (human epidermal growth receptor)
  - Tissue-based biomarker
  - over-expressed in 20-25% of breast cancer patients
  - good responders to adjuvant trastuzumab (Herceptin, Genentech, CA)
  - shown to reduce:
    - risk of recurrence by ~50% and
    - mortality by ~30% (Wolff, Arch Pathol Lab Med, 2007)
Immunohistochemistry (IHC)

- HER2 and other tissue-based biomarkers are assessed with IHC
  - makes it possible to detect antigens in tissue
  - Multi-step procedure resulting in paraffin-embedded stained tissue sections
Immunohistochemistry (IHC)

• Increasing levels of measurement accuracy
  • binary: negative/ positive (i.e. estrogen receptor)
  • semi-quantitative (ordinal): 0, 1+, 2+, 3+ (i.e. HER2)
    • related to clinical follow-up decisions
  • quantitative: continuous scale
    • related to number of receptor
Immunohistochemistry (IHC)

• IHC has become a major part of surgical pathology practice:
  • it can identify a wide number of antigens
  • results can be viewed using only a light microscope
  • slides retain properties for a long time
  • inexpensive
IHC is limited by lack of reproducibility

- **Inter- and intra-laboratory**
  - In > 2000 patients in Canada >40% FN rate in determining ER-positive patients [1,2]

- **Inter- and intra-observer variability**
  - Hsu et. al reported complete agreement in 48% of HER2 cases (22 out of 46, 5 observers) [3]
  - Distinguishing 2+ from 3+ showed agreement in only 13 (59%) of 22 positive cases

Variability in biomarker assessment

- Hinders the clinical utility of biomarkers
  - clinicians must trust the test
- Reduces the statistical power of studies for biomarker discovery
- Reduces the statistical power of clinical trials for drug efficacy
  - increases the size and cost of clinical trials
  - delays the adoption of new targeted therapies
- What are the sources of variability?
Immunohistochemistry: Multiple sources of variability

- Tissue preparation
  - IHC results can be affected greatly by:
    - tissue section thickness
    - choice of fixatives,
    - delay in fixation
    - over-fixation
    - inadequate tissue dehydration prior to paraffin embedding

- Tissue labeling

- Tissue slide interpretation

Immunohistochemistry: Sources of variability

Tissue preparation

Tissue labeling

Tissue slide interpretation

• Choice of antigen retrieval and staining methods major source of inter-laboratory variability

Immunohistochemistry: Sources of variability

- Tissue preparation
- Tissue labeling
- Tissue slide interpretation

Several efforts have sought to:
- develop standardized assay methodologies [1, 2]
- develop objective methods of measurement [3]
- provide external staining standards and proficiency tests [4].

Quality control/standardization in IHC is an ongoing process.

Immunohistochemistry: Procedure

- Tissue preparation
- Tissue labeling
- Tissue slide interpretation

• The interpretation of IHC staining by pathologists is one of the most important sources of variability in the assessment of biomarkers
Biomarker staining interpretation in pathology

• Generally: Human perception varies among individuals
  • Can be affected by training, experience, physical differences, fatigue
  • Observer variability is known to exist in other fields like Radiology
Biomarker staining interpretation in pathology

- Observer variability specific to pathology:
  - Evaluating different regions of a slide (tumor heterogeneity)
  - Using different cut-offs to determine stained cell positivity
  - Using different approaches to combine region scores into a single slide score
  - Use of different microscopes, illumination sources, reading conditions
  - Subjective criteria/guidelines for assessment of biomarker expression
Interpretation of Her-2/neu using IHC: membranous staining

- Evaluation based on color stain assessment:
  - membrane staining completeness
  - membrane staining intensity

Subjective criteria

<table>
<thead>
<tr>
<th>SCORE</th>
<th>STAINING PATTERN</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>No staining observed or membrane staining &lt;10% of tumor cells</td>
</tr>
<tr>
<td>1+</td>
<td>Faint partial membrane staining in &gt;10%</td>
</tr>
<tr>
<td>2+</td>
<td>Weak to moderate complete membrane staining in &gt;10% of cells</td>
</tr>
<tr>
<td>3+</td>
<td>Moderate to strong complete membrane staining in &gt;30% of tumor cells</td>
</tr>
</tbody>
</table>
IHC interpretation in pathology: Computer-aided assessment

• *Computer-aided assessment* of IHC could make the task more objective, quantitative, reproducible, automated, more efficient
  
  • by-product of digital pathology
  • Enabled by technological advances in whole slide imaging (WSI)
    • WSI scanning of ~1 min/slide
    • autofocus algorithms, z-axis focus
    • High resolution
    • Automated feeding of multiple slides for high throughput

Year
1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009
Number of Publications
0
50
100
150
200

Courtesy of Berkman Sahiner, FDA
Computer-aided diagnosis

- In 2008, radiologists used CAD in 74% of screening exams*
  - Medicare part B physician/supplier procedure summary master files
    - 5,827,326 screening mammograms
    - 4,305,595 with CAD

Computer-aided assessment of IHC

- Commercial software available
  - FDA cleared for IHC assessment
- Still under-examined area
  - Interaction of pathologists with computer aids

- We have developed a computer aid for the assessment of HER2
- Examined benefit when used by observers
Development of an automated method for the quantitative assessment of HER2

- Automated extraction of continuous measures of:
  - membrane staining intensity and
  - membrane staining completeness
- Measures can be provided to observer or used to classify slide

Materials

• 77 breast cancer tissue slides stained for HER2
  • HER2 scores: 26 (1+), 27 (2+), 24 (3+)
  • Truth from archives of the Department of Pathology, University of California, Irvine

• Whole slide scanning
  • Aperio Scanscope T2 system
  • Stephen Hewitt, NCI

• Multiple images (~5) containing invasive cancer cells were extracted from each slide
Algorithm Overview

**Color Pixel Classifier**

- Nuclei segmentation
- Cell membrane detection
- Feature extraction

Using color features of training pixels

**Continuous measures of HER2 staining**

**Slide classification**

(1+, 2+, 3+)

Categories:
- nuclei, membrane, background
Algorithm Overview

- Color Pixel Classifier
- Nuclei segmentation
  - Cell membrane detection
  - Feature extraction
- Continuous measures of HER2 staining
- Incorporating watershed transform and shape information
  - Slide classification (1+, 2+, 3+)
Algorithm Overview

- Color Pixel Classifier
- Nuclei segmentation
- Cell membrane detection
- Feature extraction
- Continuous measures of HER2 staining

Slide classification (1+, 2+, 3+)

Ellipse fit model
Assume an ellipse model on the membrane

Two features are extracted for each detected membrane:

**Membrane closing (completeness):** the percentage of ellipse pixels overlapping with membrane pixels

**Membrane Intensity:** the average staining intensity (average membrane color pixel classifier output) of the fitted ellipse pixels.
Examples

Yellow ellipses $\rightarrow$ membrane completeness $< 0.5$
Red ellipses $\rightarrow$ membrane completeness $> 0.5$

1+

2+

3+
Algorithm Overview

Color Pixel Classifier

Cell segmentation

Cell membrane detection

Feature extraction

Continuous measures of HER2 staining

membrane staining intensity
membrane staining completeness
Averaged across all nuclei of each image

Slide classification
(1+, 2+, 3+)
Algorithm Overview

- Color Pixel Classifier
- Cell segmentation
- Cell membrane detection
- Feature extraction

Continuous measures of HER2 staining

Provided to observer

Slide classification (1+, 2+, 3+)
Results

• 13 slides (out of 77) were used for:
  • pixel classification training
  • development of nuclei segmentation algorithm
• Remaining 64 slides (22 1+, 22 2+ and 20 3+)
  • K-fold cross validation (slide classification)

▪ Performance assessment metric:
  ▪ percent correct agreement with pathologist scores
  ▪ Overall: 83% (82% 1+, 78% 2+, 88% 3+)

- Could this computer aid provide benefit to pathologists?
IHC assessment of HER2 with unaided and computer-aided digital microscopy: observer study

• Goal: examine whether computer-aided information can benefit pathologist performance
• Reader study performed in our Display lab:
  • 14 readers
    • 7 pathologists (range in experience) from FDA, NCI/NIH
    • 7 novices (DIAM scientists, no experience in pathology)
• Two reading modes:
  • Unaided assessment of HER2 expression
  • Computer-aided assessment of HER2 expression
HER2 Assessment: unaided

ROI 1/500
HER2 Assessment: computer-aided

ROI 1/500

Computer Analysis

Membrane Intensity vs. Membrane Completeness

Score | Rating
--- | ---
3 | 97
IHC assessment of HER2 with unaided and computer-aided digital microscopy: observer study

- 3-step Training
  - Presentation about study, HER2, guidelines for scoring, instructions for using computer-aided
    - Use the whole range! (not all 3s are equal)
  - Scoring session with feedback (30 training images)
    - If score differed significantly from score of expert, score again
  - Practice session with unaided and computer-aided (30 training images)
IHC assessment of HER2 with unaided and computer-aided digital microscopy: reader study

• Main study
  • 335 images from 64 HER2 slides,
  • Randomized: case order, reading mode
  • Read in 2 sessions, > 1 month apart
  • 241 read in 2 reading modes (inter-observer variability)
  • 94 (47 aided, 47 unaided) read in same reading mode (intra-observer variability)
  • All images read on same calibrated monitor, same reading conditions
  • Reading time: 1 ½ - 2 ½ hours/session
IHC assessment of HER2 with unaided and computer-aided digital microscopy: observer study

- **Analysis:**
  - No truth available for HER2 expression
    - Quantify observer variability with Agreement Analysis
  - Intra-reader agreement
    - How well does a reader agree with self?
  - Inter-reader agreement within a group
    - How well do like readers agree?
  - Inter-reader agreement across groups
    - How well do readers from two groups agree?
Agreement metrics

• **Intra-class coefficient (ICC)**
  • data are pooled to estimate the mean and variance
  • describes how strongly units in the same group resemble each other > **group agreement (inter-observer)**

• **Kendall’s tau**
  • Consider a pair of readers ranking a pair of cases:
    • $C = \# \text{ of Concordances: two readers rank a pair of cases in the same order}$
    • $D = \# \text{ of Discordances: two readers rank a pair of cases in the opposite order}$
    • $T_1: \text{ Tie for reader 1 only}$
    • $T_2: \text{ Tie for reader 2 only}$
  • suitable for both continuous and categorical data
  • Takes ranking ties into account
    • $> \text{ pair-wise agreement (inter and intra-observer)}$

\[
K_\tau = \frac{C - D}{\sqrt{(C + D + T_1)(C + D + T_2)}}
\]
IHC assessment of HER2 with unaided and computer-aided digital microscopy: observer study

<table>
<thead>
<tr>
<th>Observer group</th>
<th>Inter-observer (group) agreement from continuous data (Intraclass correlation coefficient)</th>
<th>Inter-observer (group) agreement from categorical data (Intraclass correlation coefficient)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unaided</td>
<td>Computer-aided</td>
</tr>
<tr>
<td>Overall</td>
<td>0.81 (0.78-0.84)</td>
<td>0.92 (0.91-0.93)</td>
</tr>
<tr>
<td>Pathologists</td>
<td>0.80 (0.76-0.83)</td>
<td>0.91 (0.89-0.92)</td>
</tr>
<tr>
<td>Novices</td>
<td>0.83 (0.80-0.86)</td>
<td>0.93 (0.92-0.94)</td>
</tr>
</tbody>
</table>
### IHC assessment of HER2 with unaided and computer-aided digital microscopy: reader study

<table>
<thead>
<tr>
<th>Reader group</th>
<th>Inter-reader (pairwise) agreement using continuous data ((Kendall's \tau_b))</th>
<th>Inter-reader (pairwise) agreement using categorical data ((Kendall's \tau_b))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unaided</td>
<td>Computer-aided</td>
</tr>
<tr>
<td>Overall</td>
<td>0.62 (0.56-0.67)</td>
<td>0.76 (0.71-0.81)</td>
</tr>
<tr>
<td>Pathologists</td>
<td>0.61 (0.53-0.67)</td>
<td>0.75 (0.66-0.81)</td>
</tr>
<tr>
<td>Novices</td>
<td>0.63 (0.57-0.69)</td>
<td>0.78 (0.72-0.83)</td>
</tr>
<tr>
<td>Pathol. vs. Novices</td>
<td>0.62 (0.57-0.66)</td>
<td>0.76 (0.71-0.81)</td>
</tr>
</tbody>
</table>
## IHC assessment of HER2 with unaided and computer-aided digital microscopy: observer study

<table>
<thead>
<tr>
<th>Reader group</th>
<th>Intra-reader agreement using <em>continuous</em> data <em>(Kendall’s τₖ)</em></th>
<th>Intra-reader agreement using <em>categorical</em> data <em>(Kendall’s τₖ)</em></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unaided</td>
<td>Computer-aided</td>
</tr>
<tr>
<td>Overall</td>
<td>0.71 (0.64-0.76)</td>
<td>0.81 (0.75-0.85)</td>
</tr>
<tr>
<td>Pathologists</td>
<td>0.72 (0.64-0.79)</td>
<td>0.80 (0.72-0.86)</td>
</tr>
<tr>
<td>Novices</td>
<td>0.70 (0.60-0.77)</td>
<td>0.81 (0.73-0.87)</td>
</tr>
</tbody>
</table>
IHC assessment of HER2 with unaided and computer-aided digital microscopy: reader study

- Inter- and intra-observer agreement was improved
- Novices performed comparably to pathologists
  - some tasks in pathology can be done by non-physicians if properly trained
- Score variability was much smaller for 3+ cases
  - possible role of such computer-aided as “triage” software
- Potential for computer-aided assessment to improve pathologist performance

Discussion/Future work

- **On computer aid:**
  - Algorithm based on manually selected training pixels
    - Difficult to retrain on different datasets (different image properties)
    - Same issue exists with available “tunable” software
  - Have developed alternative method using color content
    - Allows practical, supervised re-training of the algorithm for slides with different color properties

Brad Keller, Weijie Chen, Marios A Gavrielides, “Quantitative assessment and classification of tissue-based biomarker expression using color content analysis”, Archives of Pathology and Laboratory Medicine, (accepted July 2011)
Discussion/Future work

- Study focused only on digital microscopy
- New study compares assessment with optical and digital microscopy
  - Multiple biomarkers for breast cancer: Ki67, ER/PR, HER2
  - Analysis of whole section slides and WSIs (ROI in previous study)
  - Analysis of tissue microarray (TMA) slides and images
  - Analysis of IHC interpretation when using different staining antibodies
- Other projects focusing on:
  - color reproducibility/management
  - Inter-scanner variability
Ongoing related projects

• Will provide data and experience toward answering regulatory questions regarding the performance evaluation of digital pathology devices/software
  • Technical evaluation of systems
  • Diagnostic performance of pathologists
• Provide guidance to developers
In summary

• Have presented ongoing research
  • quantitative assessment of HER2
  • effect of computer aids on observer variability

• New technologies such as WSI and computer aided assessment in tissue imaging have potential in improving diagnostic efficacy in pathology
  • Still need research to determine the role and limitations of such technologies
Acknowledgement

• Office of Women’s Health for their support
• Collaborators at NCI/NIH (Dr. Stephen Hewitt)
• Collaborators at OIVD
• All the readers that participated in our reader studies (FDA/OIVD, NCI/NIH)

• Comments/questions/suggestions:
  • marios.gavrielides@fda.hhs.gov