Mitotic Counting Reproducibility/Feature Study With Pathologists

Brandon D. Gallas PhD
FDA/CDRH/OSEL/DIDSR
Charles Halsey, DVM, PhD and R. Mark Simpson, DVM, PhD
Molecular Pathology Unit, Laboratory of Cancer Biology and Genetics, NCI

Outline

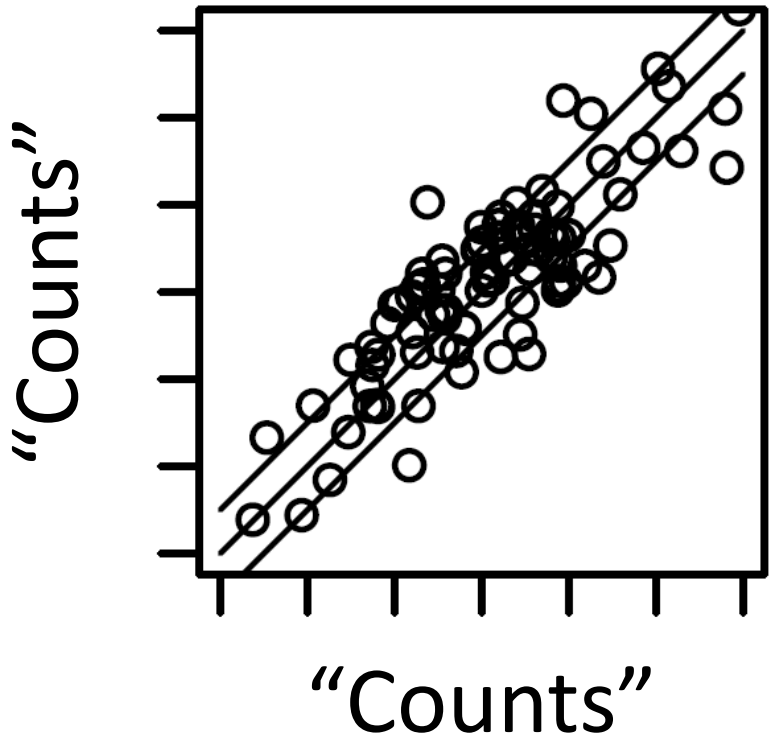
- Motivation
- Study Design and Analysis Concepts
  - Show simulated data
- NIH clinical study
  - Data collection complete
- Reproducibility study
  - Proposal, looking for collaborators and sites

Motivation

- Evaluate new imaging technology
  - Truth not known
  - Reference for comparison determined by readers
    - Intra-reader agreement
    - Inter-reader agreement
  - ***"Reproducibility" Study***
  - Example tasks
    - Impact of viewing mode on mitotic count (digital vs. optical)
    - Impact of stain on mitotic count (H&E vs. PHH3)

- Study designs and analysis methods are lacking
  - Want to account for reader and case variability

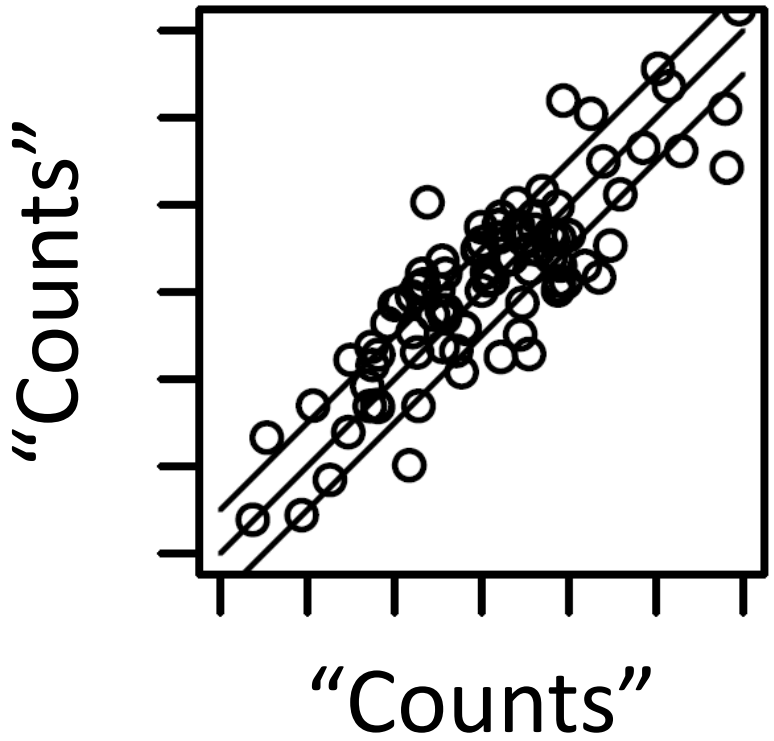
Simulated Data



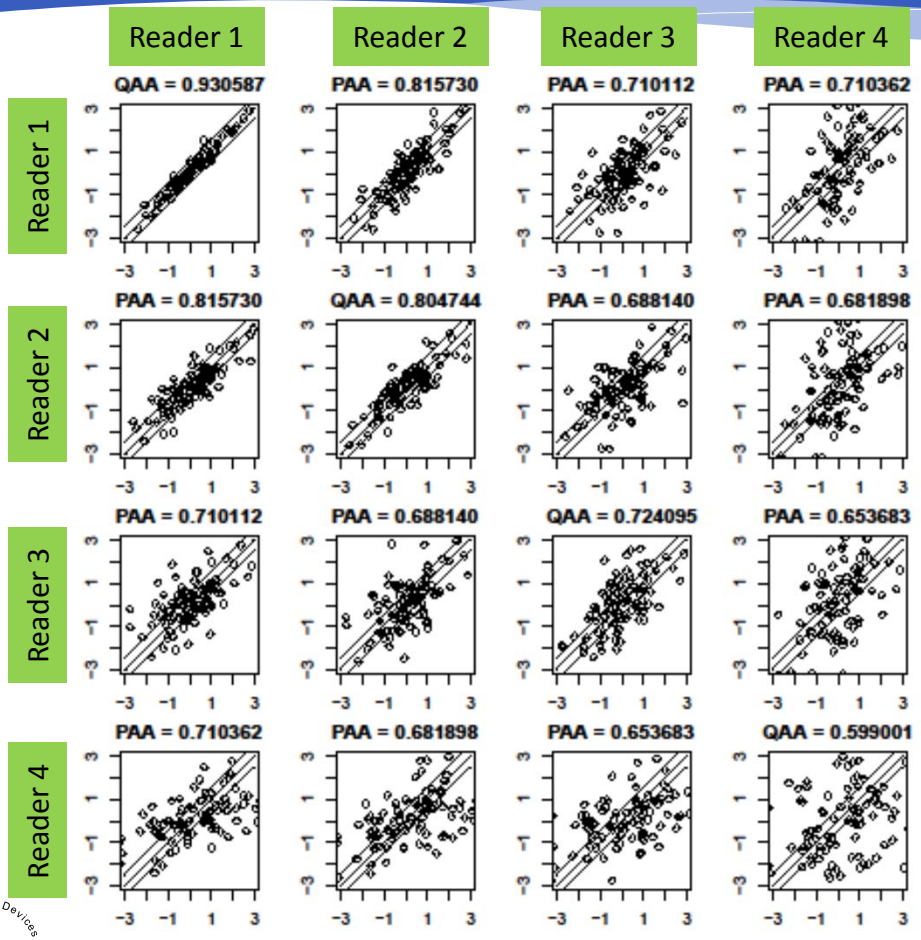
- "Counts" vs. "Counts"
- Scatter plot **shows** agreement

Simulated Data

"Correlation" = 0.804744



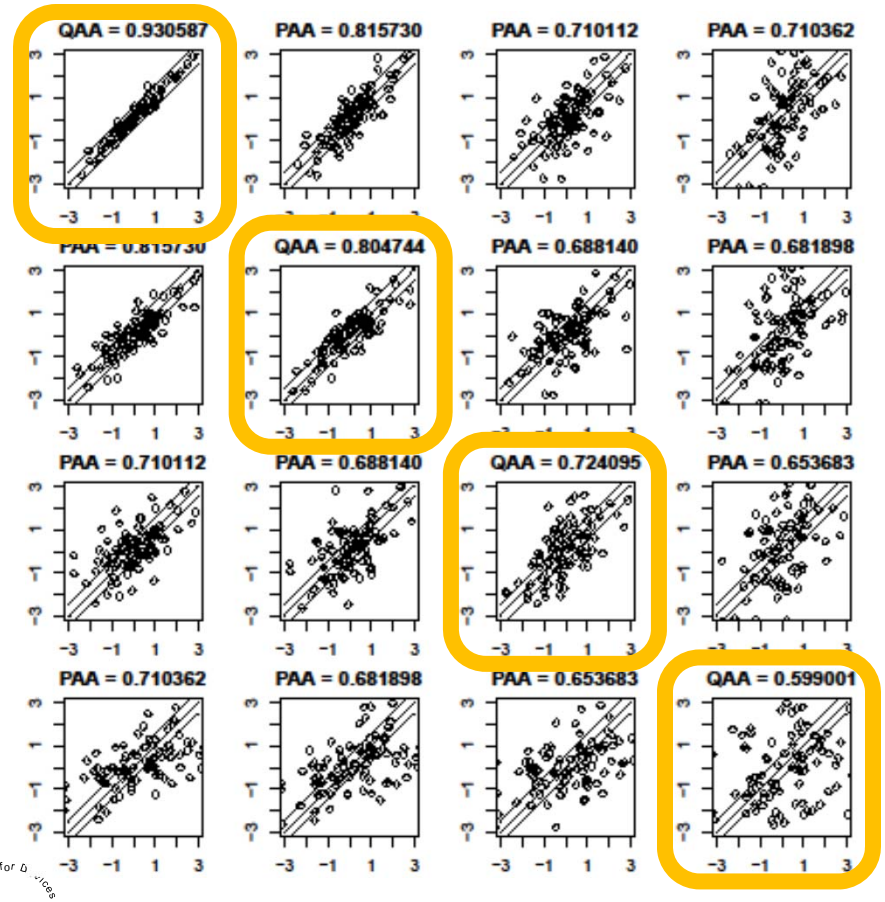
- "Counts" vs. "Counts"
- Scatter plot **shows** agreement
- Summarize agreement
  - Correlation is one of many agreement measures



Simulated Data

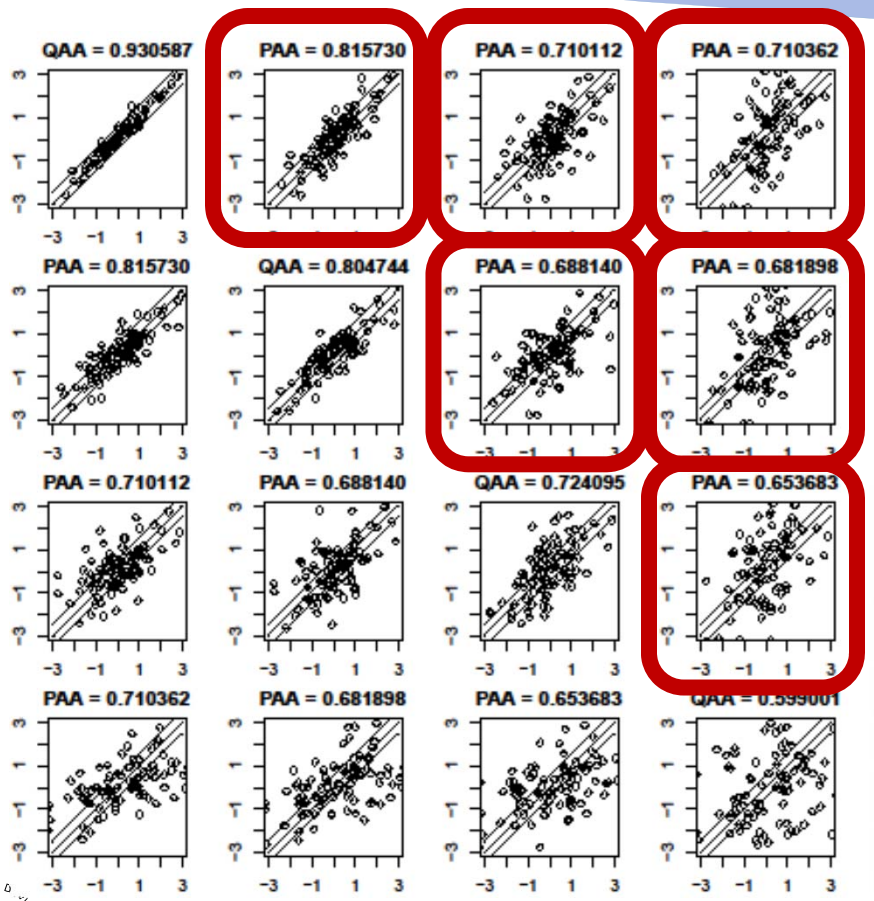
- Multiple readers evaluate Multiple cases
- Variability from cases
- Variability from readers

Simulated Data



- Intra-Reader
- Intra-modality
  - Requires replicated reads

Intra-Modality Agreement		Reference Modality, A				
		Dr. 1	Dr. 2	Dr. 3	…	Dr. 8
Reference Modality, A	Dr. 1	$Q_1^{AA^*}$	P_{12}^{AA}	P_{13}^{AA}	…	P_{18}^{AA}
	Dr. 2	P_{21}^{AA}	$Q_2^{AA^*}$	P_{23}^{AA}	…	P_{28}^{AA}
	Dr. 3	P_{31}^{AA}	P_{32}^{AA}	$Q_3^{AA^*}$	…	P_{38}^{AA}
	…	…	…	…	…	…
	Dr. 8	P_{81}^{AA}	P_{82}^{AA}	P_{83}^{AA}	…	$Q_8^{AA^*}$



Simulated Data

- Inter-Reader
- Intra-modality
  - No replicated reads

Intra-Modality Agreement		Reference Modality, A				
		Dr. 1	Dr. 2	Dr. 3	…	Dr. 8
Reference Modality, A	Dr. 1	$Q_1^{AA^*}$	P_{12}^{AA}	P_{13}^{AA}	…	P_{18}^{AA}
	Dr. 2	P_{21}^{AA}	$Q_2^{AA^*}$	P_{23}^{AA}	…	P_{28}^{AA}
	Dr. 3	P_{31}^{AA}	P_{32}^{AA}	$Q_3^{AA^*}$	…	P_{38}^{AA}
	…	…	…	…	…	…
	Dr. 8	P_{81}^{AA}	P_{82}^{AA}	P_{83}^{AA}	…	$Q_8^{AA^*}$

Study Design & Analysis

- Intra-Reader or Inter-Reader
- Inter-modality
  - No replicated reads
- Examples
  - Optical vs. Digital
  - H&E vs. PHH3

Inter-Modality Agreement		Reference Modality, *A*				
		Dr. 1	Dr. 2	Dr. 3	...	Dr. 8
New Modality, *B*	Dr. 1	Q_1^{AB}	P_{12}^{AB}	P_{13}^{AB}	...	P_{18}^{AB}
	Dr. 2	P_{21}^{AB}	Q_2^{AB}	P_{23}^{AB}	...	P_{28}^{AB}
	Dr. 3	P_{31}^{AB}	P_{32}^{AB}	Q_3^{AB}	...	P_{38}^{AB}
	...	...	...	...	...	...
	Dr. 8	P_{81}^{AB}	P_{82}^{AB}	P_{83}^{AB}	...	Q_8^{AB}

Intra-Modality Agreement		Reference Modality, A				
		Dr. 1	Dr. 2	Dr. 3	...	Dr. 8
Reference Modality, *A*	Dr. 1	$Q_1^{AA^*}$	P_{12}^{AA}	P_{13}^{AA}	...	P_{18}^{AA}
	Dr. 2	P_{21}^{AA}	$Q_2^{AA^*}$	P_{23}^{AA}	...	P_{28}^{AA}
	Dr. 3	P_{31}^{AA}	P_{32}^{AA}	$Q_3^{AA^*}$	...	P_{38}^{AA}
	...	...	...	...	...	...
	Dr. 8	P_{81}^{AA}	P_{82}^{AA}	P_{83}^{AA}	...	$Q_8^{AA^*}$

Study Design & Analysis

- Is New Device Performance inter-modality agreement worse than Baseline Performance intra-modality agreement?
- Is New Device Performance intra-modality agreement worse than Baseline Performance intra-modality agreement?

Inter-Modality Agreement		Reference Modality, A				
		Dr. 1	Dr. 2	Dr. 3	...	Dr. 8
New Modality, B	Dr. 1	Q_1^{AB}	P_{12}^{AB}	P_{13}^{AB}	...	P_{18}^{AB}
	Dr. 2	P_{21}^{AB}	Q_2^{AB}	P_{23}^{AB}	...	P_{28}^{AB}
	Dr. 3	P_{31}^{AB}	P_{32}^{AB}	Q_3^{AB}	...	P_{38}^{AB}
	...	...	...	...	...	...
	Dr. 8	P_{81}^{AB}	P_{82}^{AB}	P_{83}^{AB}	...	Q_8^{AB}

Intra-Modality Agreement		Reference Modality, A				
		Dr. 1	Dr. 2	Dr. 3	...	Dr. 8
Reference Modality, A	Dr. 1	$Q_1^{AA^*}$	P_{12}^{AA}	P_{13}^{AA}	...	P_{18}^{AA}
	Dr. 2	P_{21}^{AA}	$Q_2^{AA^*}$	P_{23}^{AA}	...	P_{28}^{AA}
	Dr. 3	P_{31}^{AA}	P_{32}^{AA}	$Q_3^{AA^*}$	...	P_{38}^{AA}
	...	...	...	...	...	...
	Dr. 8	P_{81}^{AA}	P_{82}^{AA}	P_{83}^{AA}	...	$Q_8^{AA^*}$

requires replicated readings

Study Design & Analysis

- Intra-reader agreement
  - Summarize with average over all readers
- Inter-reader agreement
  - Summarize with average over all pairs of readers

Two options to produce study endpoints to answer study question

Inter-Modality Agreement		Reference Modality, A				
		Dr. 1	Dr. 2	Dr. 3	...	Dr. 8
New Modality, B	Dr. 1	Q_1^{AB}	P_{12}^{AB}	P_{13}^{AB}	...	P_{18}^{AB}
	Dr. 2	P_{21}^{AB}	Q_2^{AB}	P_{23}^{AB}	...	P_{28}^{AB}
	Dr. 3	P_{31}^{AB}	P_{32}^{AB}	Q_3^{AB}	...	P_{38}^{AB}
	...	...	...	...	...	...
	Dr. 8	P_{81}^{AB}	P_{82}^{AB}	P_{83}^{AB}	...	Q_8^{AB}

Intra-Modality Agreement		Reference Modality, A				
		Dr. 1	Dr. 2	Dr. 3	...	Dr. 8
Reference Modality, A	Dr. 1	$Q_1^{AA'}$	P_{12}^{AA}	P_{13}^{AA}	...	P_{18}^{AA}
	Dr. 2	P_{21}^{AA}	$Q_2^{AA'}$	P_{23}^{AA}	...	P_{28}^{AA}
	Dr. 3	P_{31}^{AA}	P_{32}^{AA}	$Q_3^{AA'}$	...	P_{38}^{AA}
	...	...	...	...	...	...
	Dr. 8	P_{81}^{AA}	P_{82}^{AA}	P_{83}^{AA}	...	$Q_8^{AA'}$

requires replicated readings

Study Design & Analysis

- Intra-reader agreement
  - Summarize with average over all readers
- Inter-reader agreement
  - Summarize with average over all pairs of readers

Two options to produce study endpoints to answer study question.

- Which best?
- Can do both.

- Pros and Cons for each. Still learning practical and statistical efficiencies.
- Intra-reader requires replicated reading.
- Inter-reader agreement more variable, but averaging over more observations (pairs of readers).

Study Design & Analysis

- What about error bars for study endpoints?
- **MRMC** analysis: account for variability from
  - **M**ultiple **R**eaders
  - **M**ultiple **C**ases
- Gallas, Anam, Chen, Wunderlich, Zhang (2016), "MRMC analysis of agreement studies." *In Proc. SPIE,* ***9787, pp. 97870F-97870F-12.***
  - *U-statistics*
  - *Agreement by concordance*
  - *Novel, builds on methods for area under the ROC curve*
  - *Peer reviewed manuscript in development*

Happy to discuss later.

NIH Clinical Study

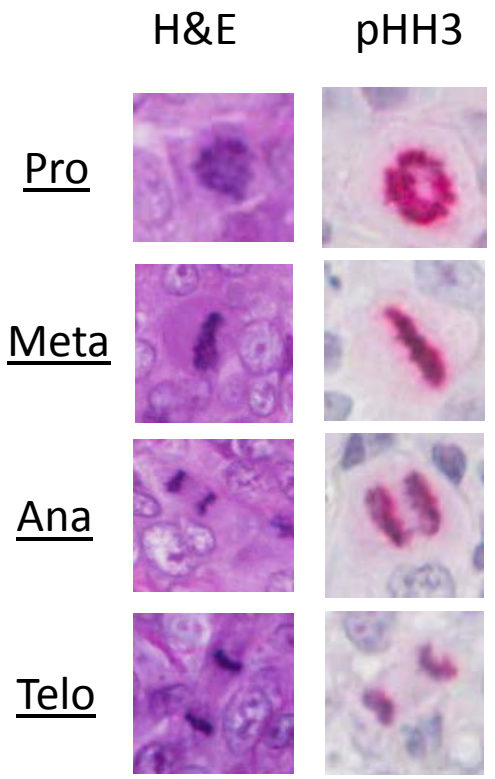
- PI's from NIH
  - Mark Simpson, DVM, PhD
  - Charles Halsey, DVM, PhD

- Collaborators from FDA: Study Design and Analysis
  - Brandon Gallas, PhD
  - Weijie Chen, PhD
  - Zhiwei Zhang, PhD

NIH Clinical Study

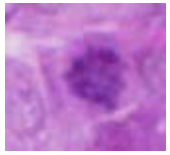
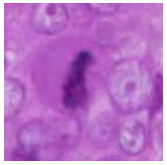
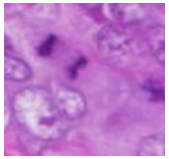
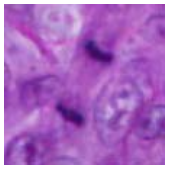
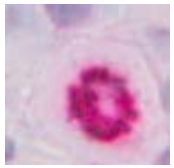
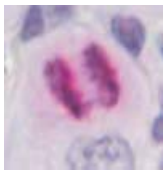
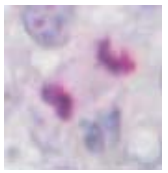
Study Design

- Feature and task:
  - Mitotic Counting
- 12 readers:
  - veterinary pathologists at 4 sites
  - NCI, 2 academic centers, one regional reference lab
- 113 cases:
  - 1 case = 1 slide
  - specimens from canines with oral melanoma
  - H&E vs. Phosphohistone H3 (PHH3)



NIH Clinical Study

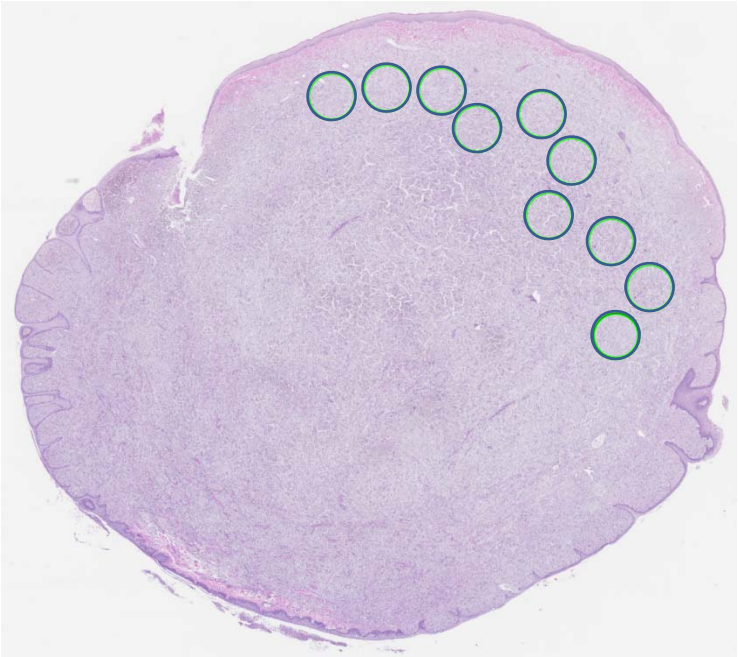
H&E vs. Phosphohistone H3 (PHH3)

	Pro	Meta	Ana	Telo
H&E				
pHH3				

NIH Clinical Study

Study Design

- Counts collected according to standard of care
  - Microscope mode: each pathologist selects 10 evaluation FOVs per case and then counts at 40x
  - Digital mode: simulate standard of care
  - FOV locations saved for each pathologist in digital mode



NIH Clinical Study

Aims

- Compare prognostic ability of mitotic count from
  - H&E Optical, 40x
  - H&E Digital, 20x
  - H&E Digital, 40x
  - PHH3 Optical, 40x
  - PHH3 Digital, 20x
  - PHH3 Digital, 40x

6 modalities = 6 viewing modes

Ultimately, we did not study all modalities and comparisons.

- Inter-reader and Intra-reader agreement analyses
  - Inter-modality and Intra-modality

NIH Clinical Study

Study Design

- Split-plot study design
  - Reduce the total number of reads
  - Reduce workload of each reader

  Each case read by each reader 3 times instead of 6
  Washout time was minimum 1 week
  - Allow for slides to be shipped

Complicated.
Happy to discuss later.

- Obuchowski, N.; Gallas, B. D. & Hillis, S. L. (2012), 'Multi-Reader ROC Studies with Split-Plot Designs: A Comparison of Statistical Methods.' *Acad Radiol,* ***19, (12), 1508-1517.***

NIH Clinical Study

Primary Aim

- Compare prognostic ability of mitotic count
  - H&E slides on a microscope
  - PHH3 slides on a microscope
  - Truth is survival data with some censoring
  - Prognostic ability
    = concordance between counts and survival
  - Test non-inferiority with possible superiority

NIH Clinical Study

Preliminary Results

- Sorry. Nothing to share yet.
- Data checks still in progress.

Proposed Reproducibility Study

- PI: Brandon Gallas
- NIH collaborators
  - Mark Simpson, DVM, MD
  - Charles Halsey, DVM, MD
- YOU!?

Proposed Reproducibility Study

Synopsis

- Complement NIH clinical study:
  - mitotic counting
  - H&E vs. PHH3
  - Optical vs. Digital

- Inter-reader and Intra-reader agreement analyses
  - Inter-modality and Intra-modality

- Study Goal:
  Eliminating location variability yields faster and more precise data

Proposed Reproducibility Study

Synopsis

- Difference from NIH clinical study
  - All pathologists evaluate *same pre-selected FOVs*
  - FOVs are smaller to fit on a screen
  - Equivalent in area to 10 traditional microscope FOVs
  - Use a reticle on the microscope to get same area as screen
  - Enabled by **eeDAP**:
    evaluation environment for digital and analog pathology

Proposed Reproducibility Study

Status

- Protocol under development
- NIH clinical study
  - Source for slides and images
  - Inform study design: modalities, intra- vs. inter-reader
  - Inform study size: readers and cases
  - Results provide basis for comparison
- Collaborators/readers/sites being recruited
  - Hope to involve WSI WG members
  - Need feedback on study protocol
  - Loan and install eeDAP system at your site
  - Find a balance between workload and statistical power (more readers = smaller workload)

Proposed Reproducibility Study

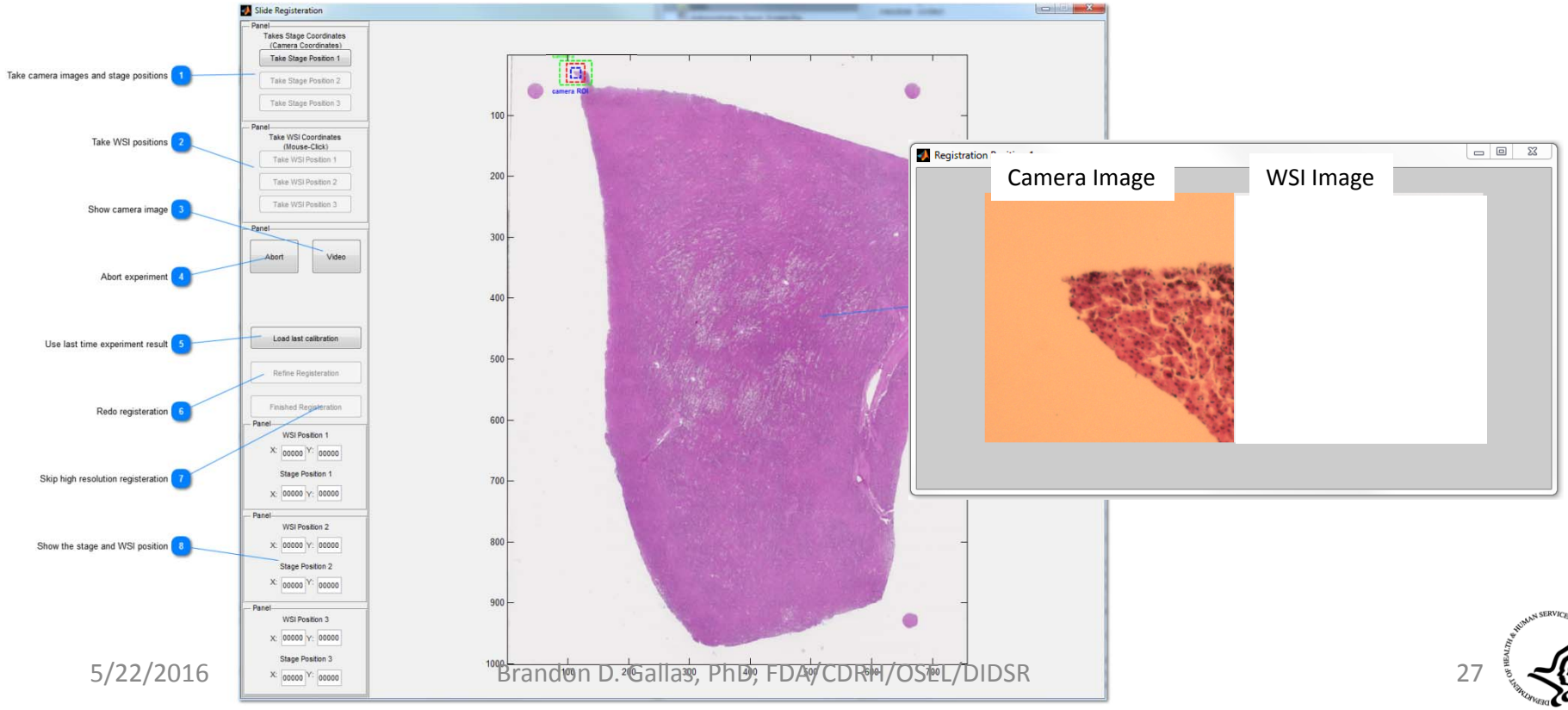
eeDAP

- evaluation environment for digital and analog pathology

- Camera
- Microscope
- Moving stage with multiple slides
- Stage controller
- Joystick for stage control
- Computer and monitor not shown

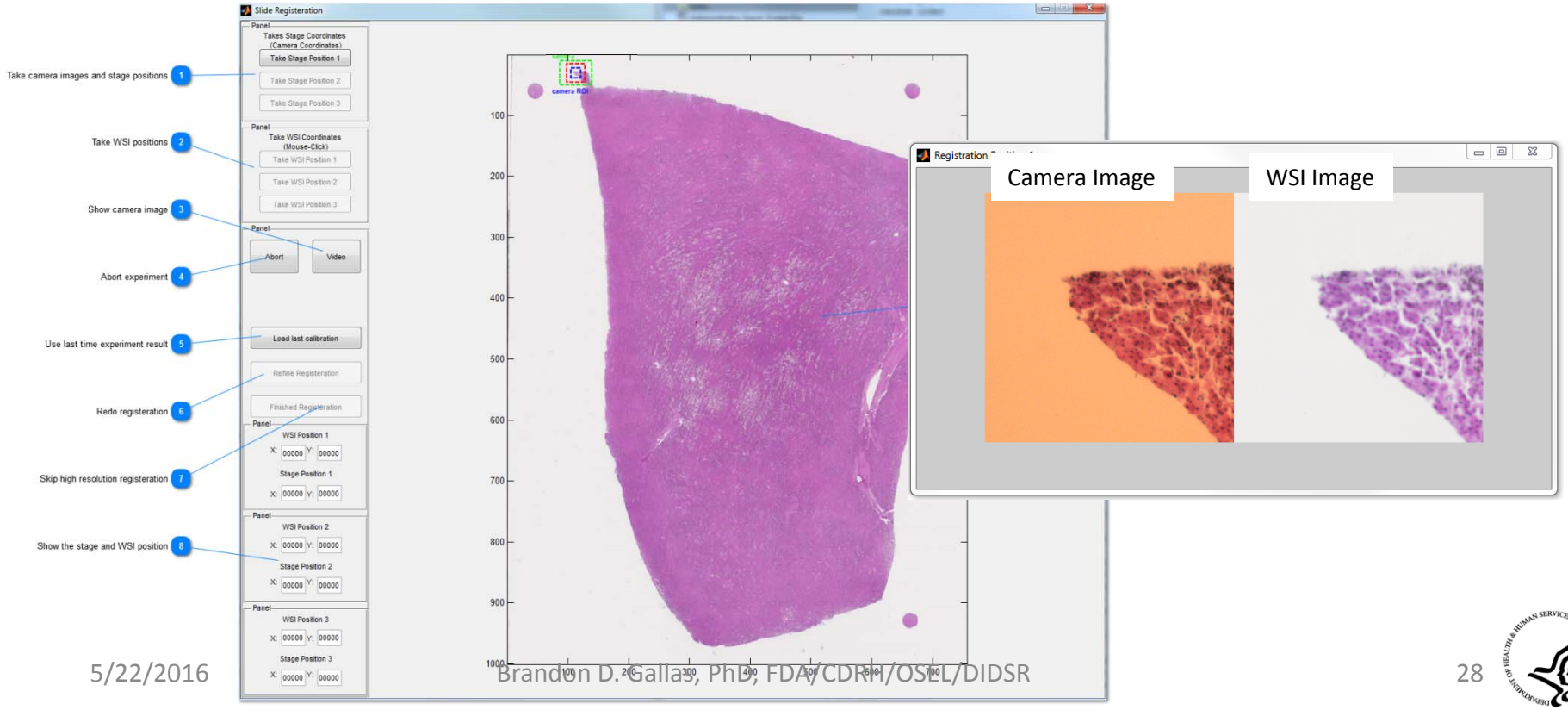
Proposed Reproducibility Study

eeDAP: Register WSI to Glass Slide (Stage)



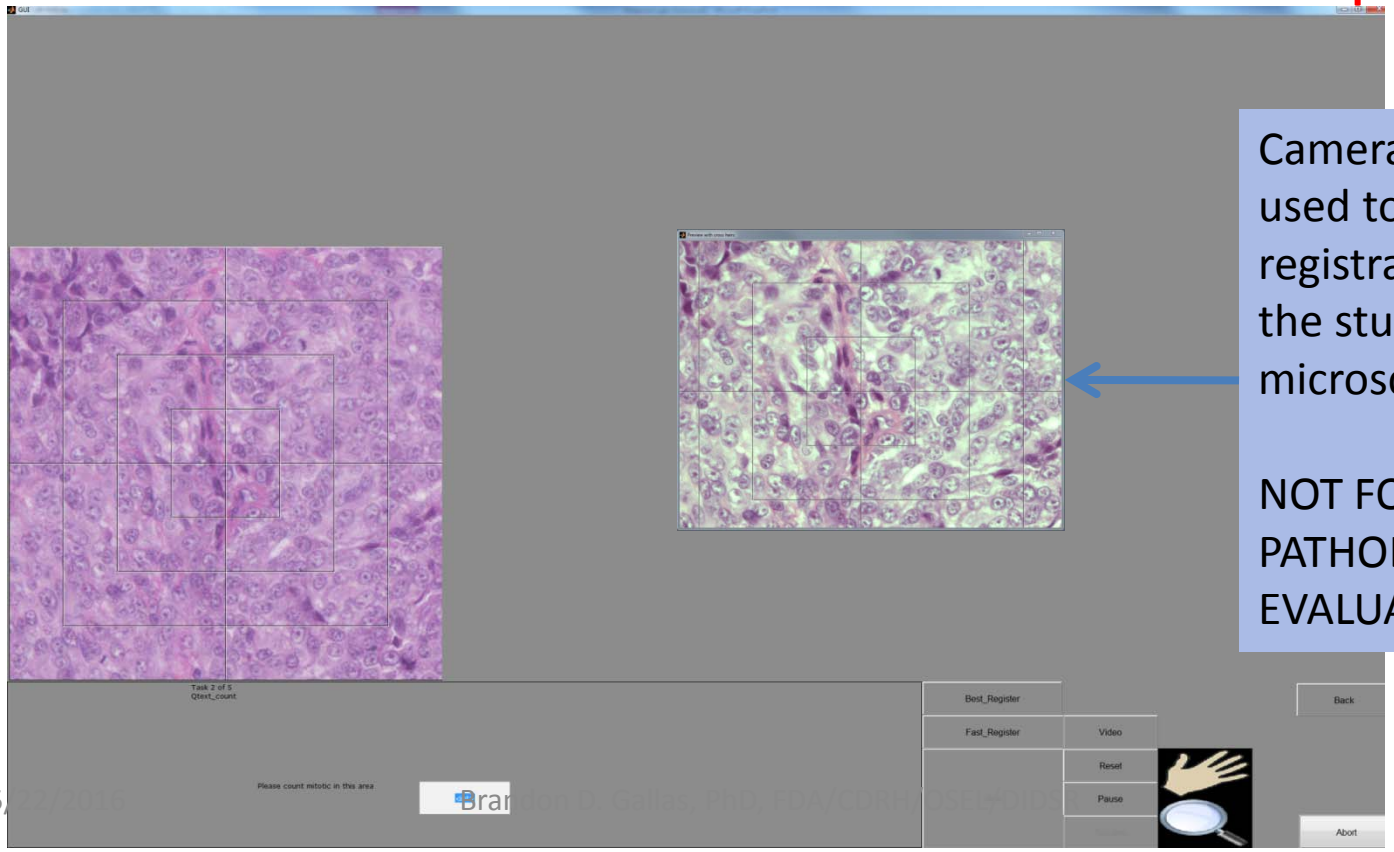
Proposed Reproducibility Study

eeDAP: Register WSI to Glass Slide (Stage)



Proposed Reproducibility Study

eeDAP: Collect Data in WSI mode or Microscope mode



Camera image used to verify registration during the study on microscope.

NOT FOR PATHOLOGIST EVALUATION

Parting Messages

Talk summary:

- Presented study design and analysis methods
- Generalize to new readers and new cases
  = Account for reader and case variability

- Outlined NIH study

- Proposed complementary eeDAP study
  - *Looking for collaborators/sites/readers*
  - *Offer to loan and install an eeDAP system*

Parting Messages

- eeDAP available at https://github.com/DIDSR/eeDAP

- Reference
  - Gallas, B. D.; Gavrielides, M. A.; Conway, C.; Ivansky, A.; Keay, T.; Cheng, W.-C.; Hipp, J. & Hewitt, S. M. (2014), 'Evaluation Environment for Digital and Analog Pathology (eeDAP): a platform for validation studies.' *J Med Img,* ***1, (3), 037501.***

Parting Messages

Studies of reproducibility can be a bridge/link between …

- Technical performance
- Clinical performance

- For example: Track human performance on clinically related task as we vary IQ parameter.

Reductionist approach …

- Pathologist expertise, case complexity, and clinical context are confounders to technology evaluation and system optimization.

- Limit case variability by defining a narrow task
  - Detect, count, characterize a single feature

- Limit case variability by defining a narrow study population