FDA U.S. FOOD & DRUG ADMINISTRATION

Exploring Pathologist-Pathologist Agreement as a Baseline for Algorithm-Pathologist Agreement

Brandon D. Gallas

Division of Imaging, Diagnostics, Software Reliability

Office of Science and Engineering Laboratories
Center for Devices and Radiological Health
U.S. Food and Drug Administration

Collaborators

- **Mohamed Amgad, MD**
  - Department of Pathology, Northwestern University
- **Kim Blenman, PhD**
  - Yale School of Medicine
- **Weijie Chen, PhD**
  - FDA/CDRH/OSEL/DIDSR
- **Sarah Dudgeon, MPH**
  - CORE Center for Computational Health Yale-New Haven Hospital
- **Kate Elfer, MPH**
  - FDA/CDRH/OSEL/DIDSR
- **Victor Garcia, MD**
  - FDA/CDRH/OSEL/DIDSR
- **Rajarsi Gupta, MD/PhD**
  - Stony Brook Medicine Dept of Biomedical Informatics
- **Matthew Hanna, MD**
  - Memorial Sloan Kettering Cancer Center
- **Steven Hart, PhD**
  - Department of Health Sciences Research, Mayo Clinic
- **Evangelos Hytopoulos, PhD**
  - iRhythm Technologies Inc
- **Denis Larsimont, MD**
  - Department of Pathology, Institut Jules Bordet
- **Xiaoxian Li, MD/PhD**
  - Emory University School of Medicine
- **Anant Madabhushi, PhD**
  - Case Western Reserve University
- **Hetal Marble, PhD**
  - Massachusetts General Hospital/Harvard Medical School
- **Roberto Salgado, PhD**
  - Division of Research, Peter Mac Callum Cancer Centre, Melbourne, Australia; Department of Pathology, GZA-ZNA Hospitals
- **Joel Saltz, MD/PhD**
  - Stony Brook Medicine Dept of Biomedical Informatics
- **Manasi Sheth, PhD**
  - FDA/CDRH/OPQE/Division of Biostatistics
- **Rajendra Singh, MD**
  - Northwell health and Zucker School of Medicine
- **Evan Szu, PhD**
  - Arrive Bio
- **Darick Tong, MS**
  - Arrive Bio
- **Si Wen, PhD**
  - FDA/CDRH/OSEL/DIDSR
- **Bruce Werness, MD**
  - Arrive Bio

Outline

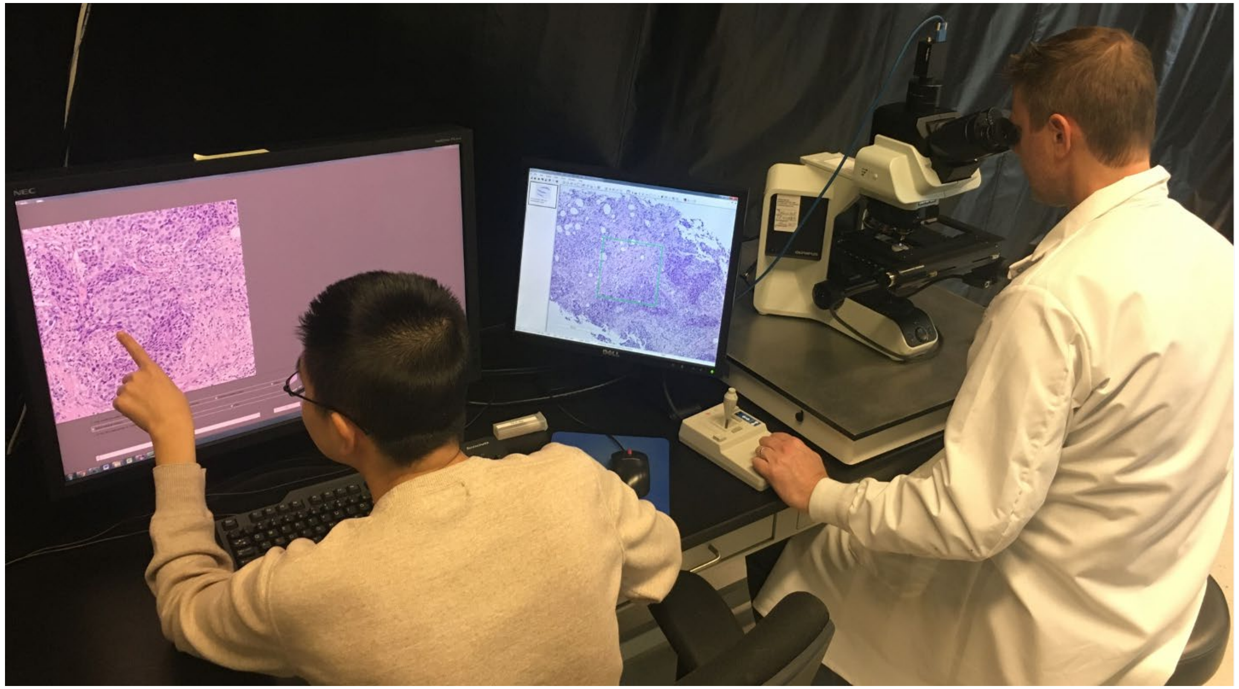
- Clinical Context: Imaging Biomarker
- Initial Analysis of Pilot Study
- Quantitative Agreement
  - Bland-Altman … Limits of Agreement
- Strategy to Use Thresholds
  - Binary Crowd-Expert Agreement for each Expert
  - Then Average over Experts
  - Baseline performance: Expert-Expert Agreement

Clinical Context and Relevance

- Clinical context:
  - Breast cancer
  - Quantitative Pathology Biomarker: Stromal Tumor Infiltrating Lymphocytes (sTILs)

- Clinical relevance of sTILs:
  - Prognostic for survival
  - Expected to inform patient management
  - Expected to reduce use of toxic chemotherapies

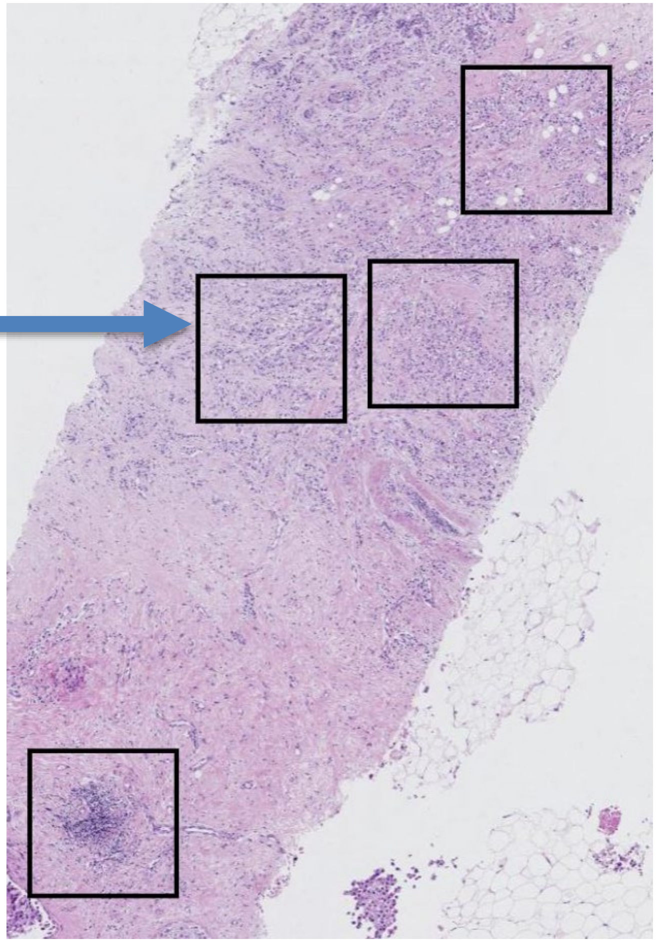
- Biomarker Evaluation by an Algorithm
  - Reduce burden on pathologist
  - Reproducible
  - Quantitative



- Deliverables
  - Reference standard data from pathologists
  - Methods to validate a quantitative algorithm

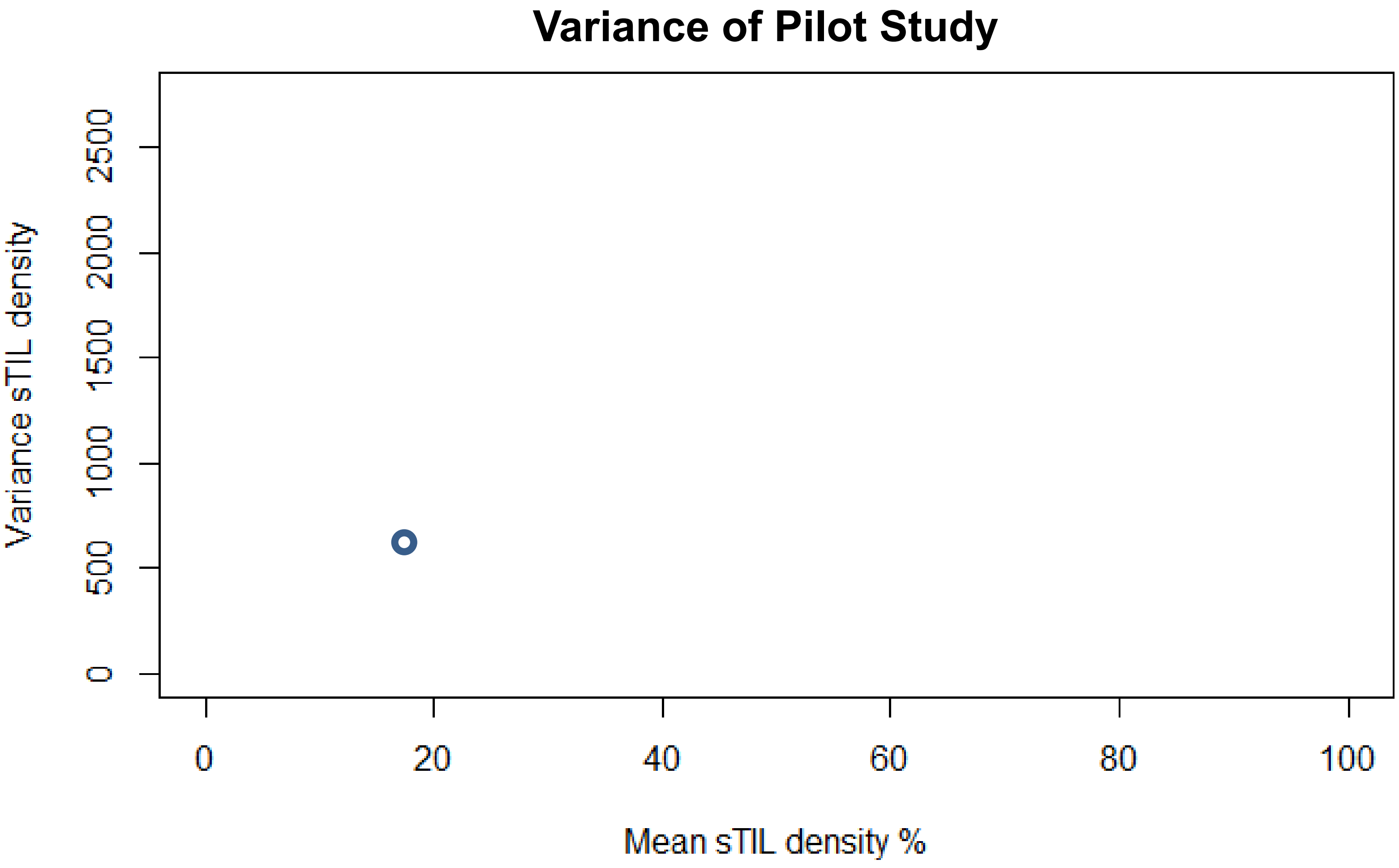
Pilot Study

- Cases:
  - 64 H&E Slides
  - 10 Regions of Interest (ROIs) per Slide
  - Some ROIs are not appropriate for sTIL evaluation
- Evaluation Platforms:
  - 2 digital and 1 microscope
- Readers:
  - 37 readers
  - 7 crowd readers with complete data
  - 7 expert readers are on the collaboration team
- 7,898 Observations
  - 432 observations are from 6 experts that completed "SELECT" subset of 72 ROIs



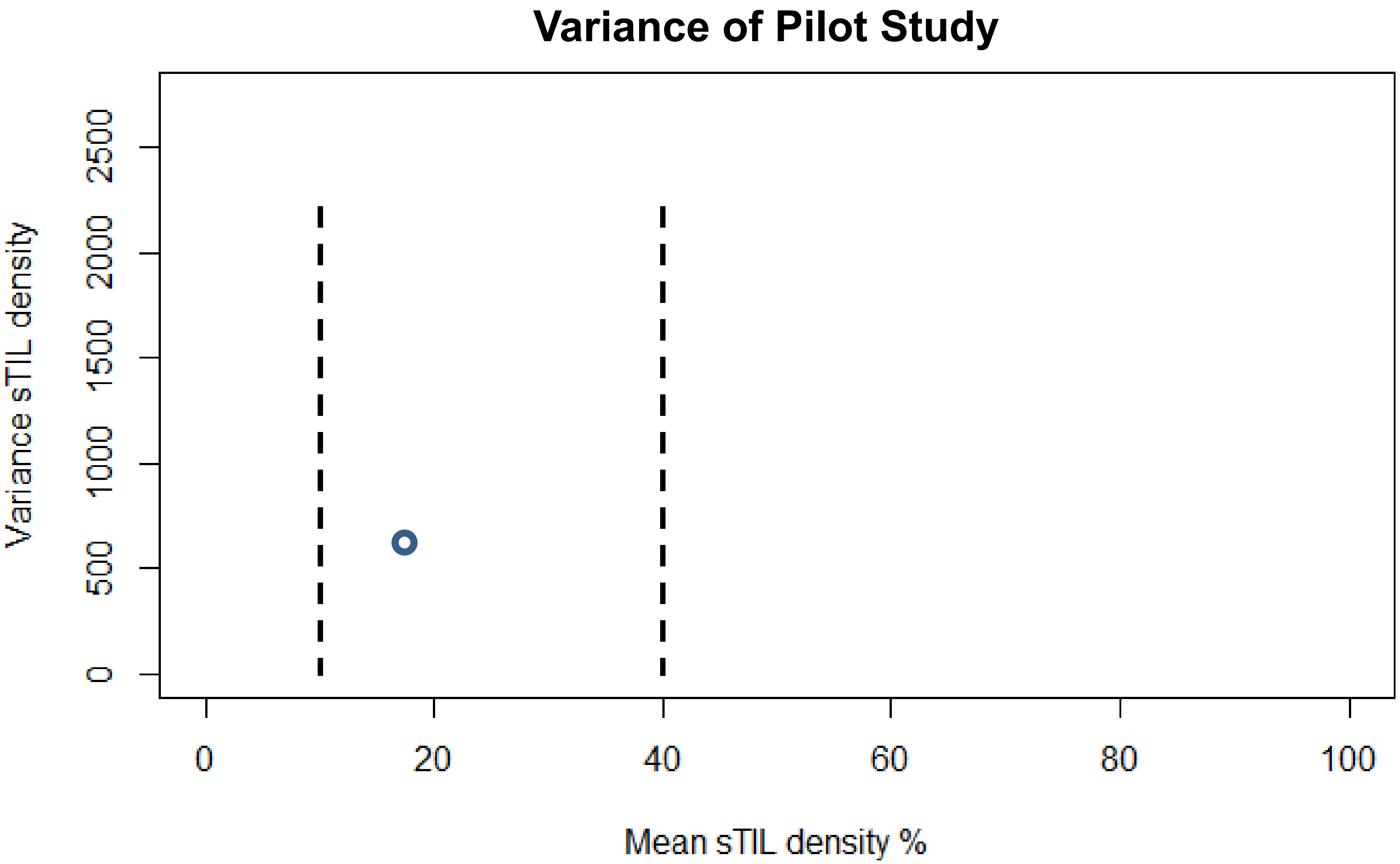
R Data Package
https://github.com/DIDSR/HTT

Initial Analysis of Pilot Study



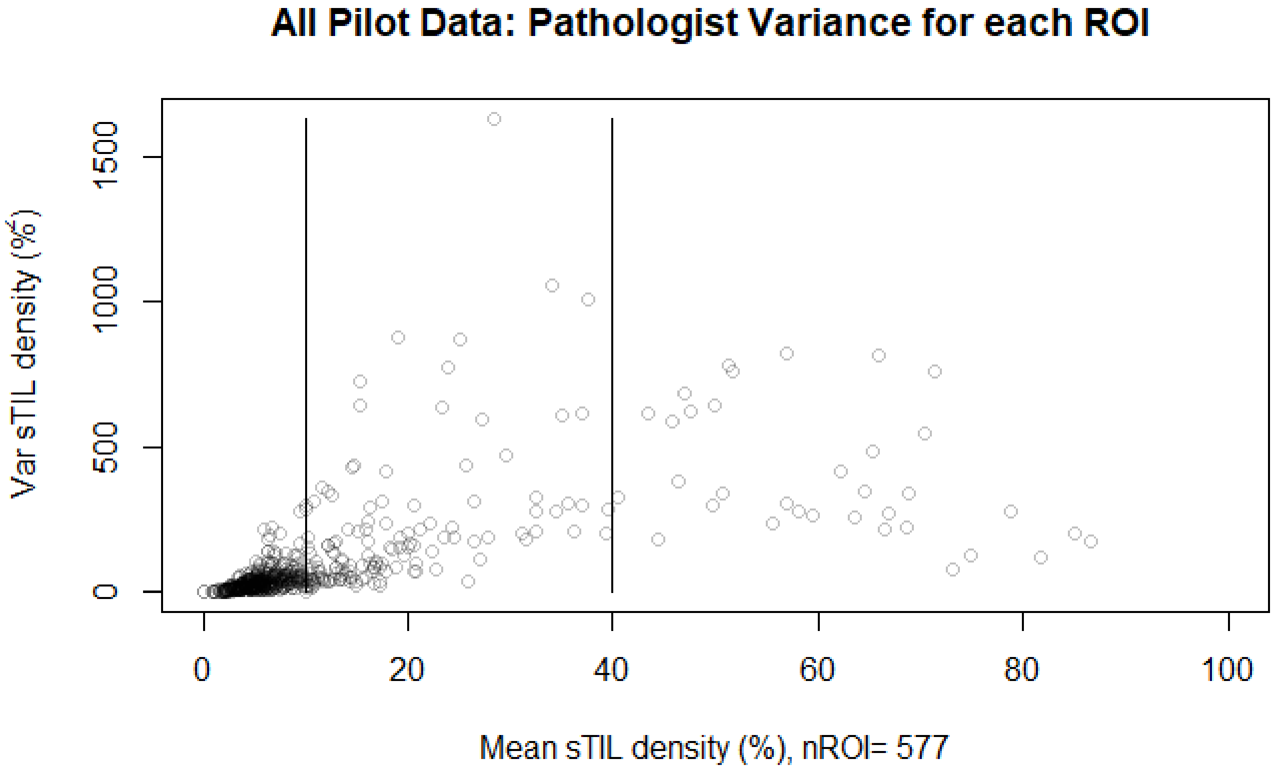
- Mean and Variance are averages over all readers

Initial Analysis of Pilot Study



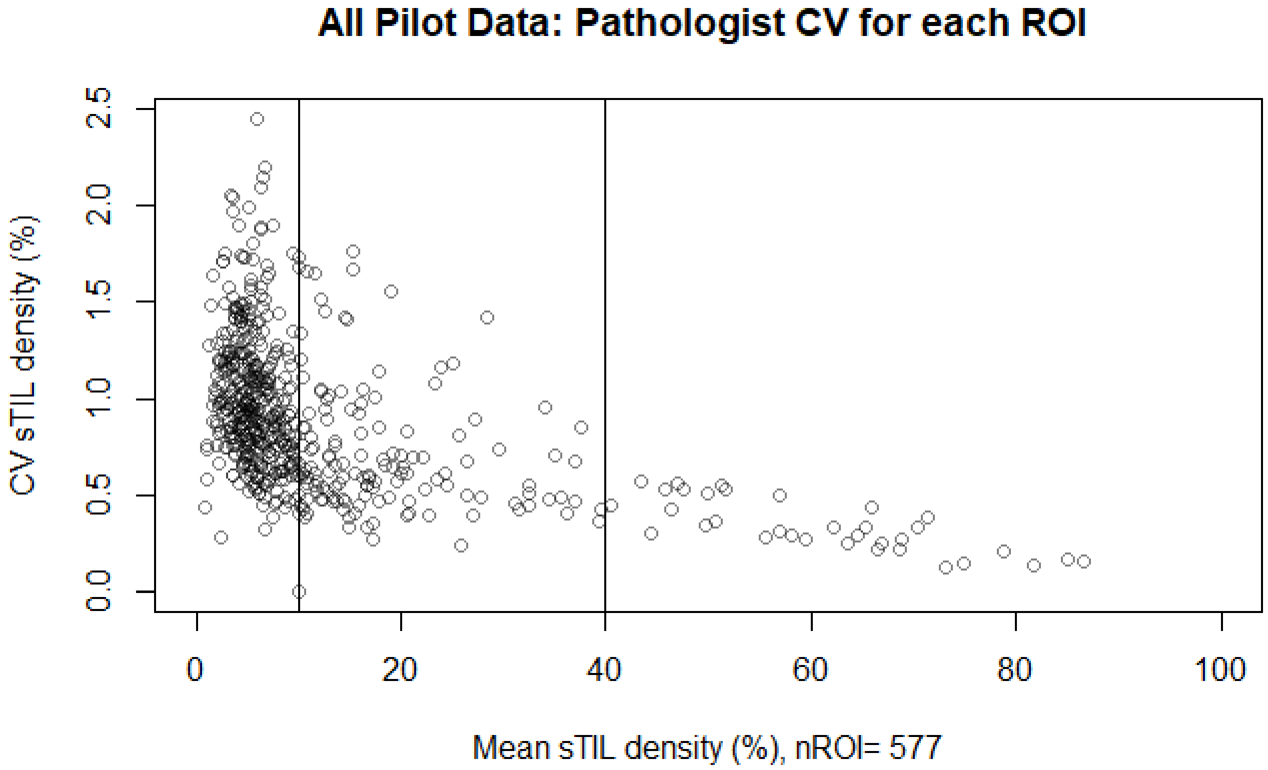
- Mean and Variance are averages over all readers
- Vertical dashed lines represent clinical bins
  - low (≤ 10%)
  - medium (>10% & ≤ 40%)
  - high (>40%)Horizontal

Initial Analysis of Pilot Study



- Means and Variances are averages over all readers
- Vertical lines represent clinical bins
  - low (≤ 10%)
  - medium (>10% & ≤ 40%)
  - high (>40%)
- Variance is increasing with the mean

Initial Analysis of Pilot Study



- Means and Variances are averages over all readers
- Vertical dashed lines represent clinical bins
  - low (≤ 10%)
  - medium (>10% & ≤ 40%)
  - high (>40%)Horizontal
- The variance does not increase with mean in a standard way

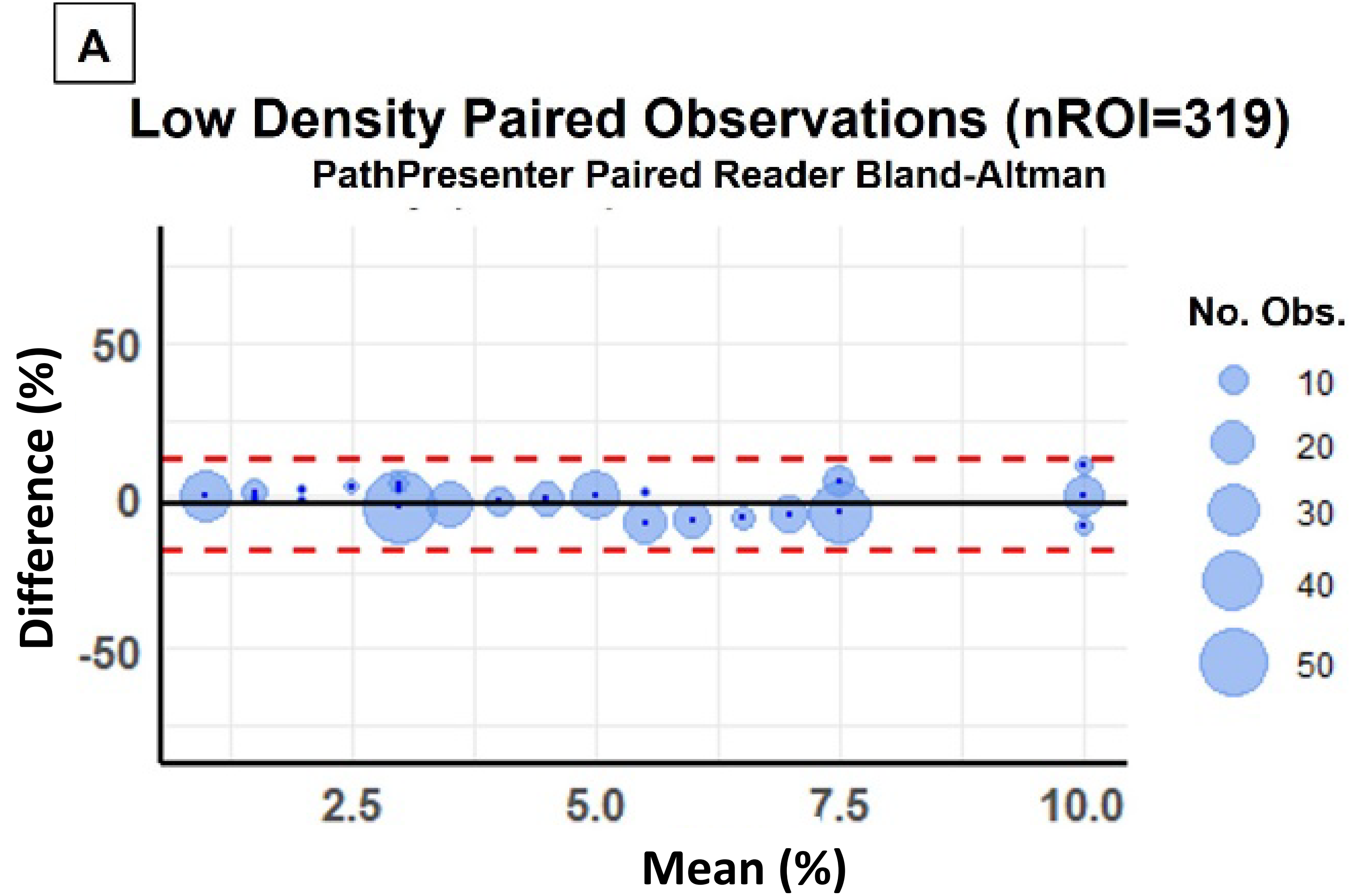
How should we determine …

- If a crowd pathologist is an expert?
- If an AI/ML model is good enough?

- First thought
  - Bland-Altman Plots
  - Limits of Agreement (LOA)

-

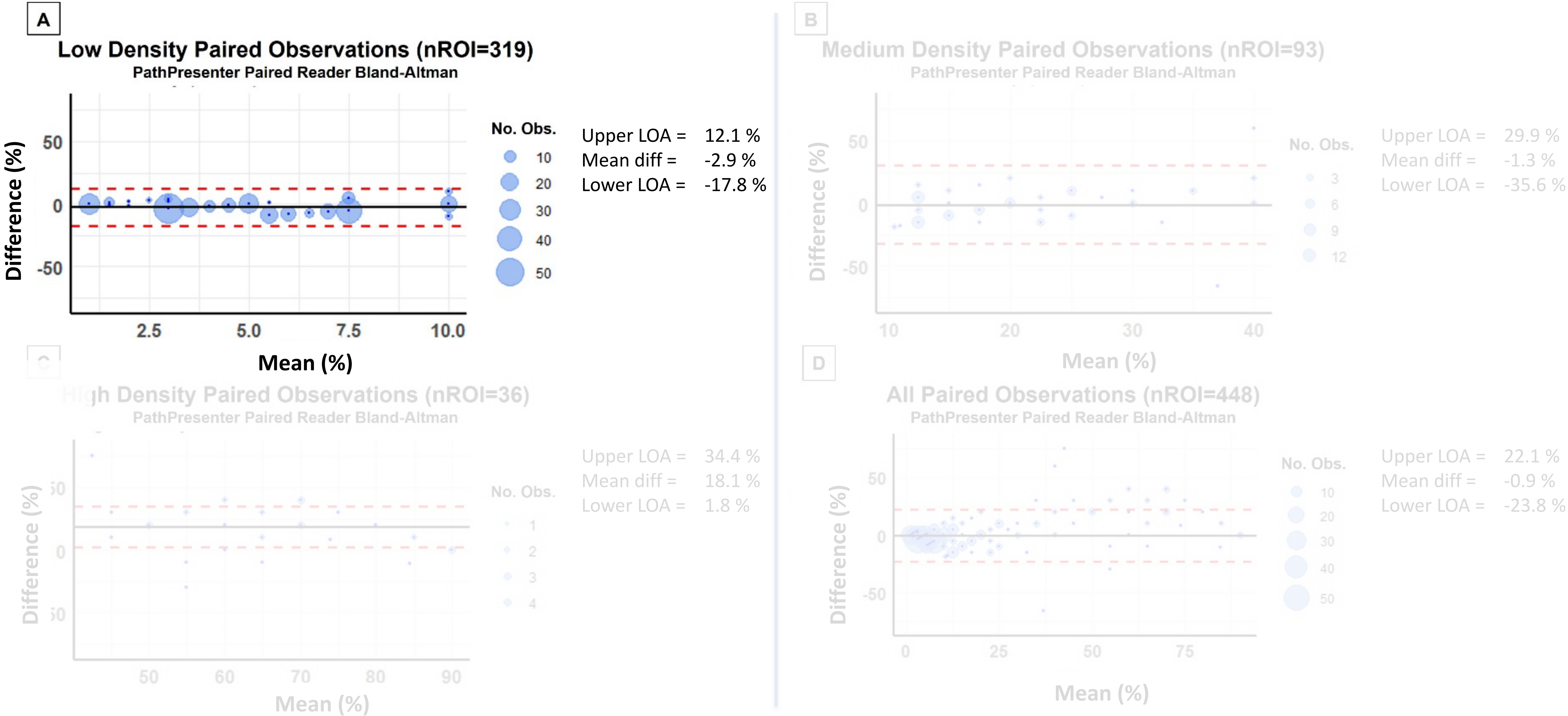
Mean Difference (Bland-Altman) Plots for two pathologists with complete data



Upper LOA =	12.1 %
Mean diff =	-2.9 %
Lower LOA =	-17.8 %

Apologies …
No uncertainty analysis of LOA yet

Mean Difference (Bland-Altman) Plots for two pathologists with complete data



A Low Density Paired Observations (nROI=319)
PathPresenter Paired Reader Bland-Altman

Upper LOA = 12.1 %
Mean diff = -2.9 %
Lower LOA = -17.8 %

B Medium Density Paired Observations (nROI=93)
PathPresenter Paired Reader Bland-Altman

Upper LOA = 29.9 %
Mean diff = -1.3 %
Lower LOA = -35.6 %

C High Density Paired Observations (nROI=36)
PathPresenter Paired Reader Bland-Altman

Upper LOA = 34.4 %
Mean diff = 18.1 %
Lower LOA = 1.8 %

D All Paired Observations (nROI=448)
PathPresenter Paired Reader Bland-Altman

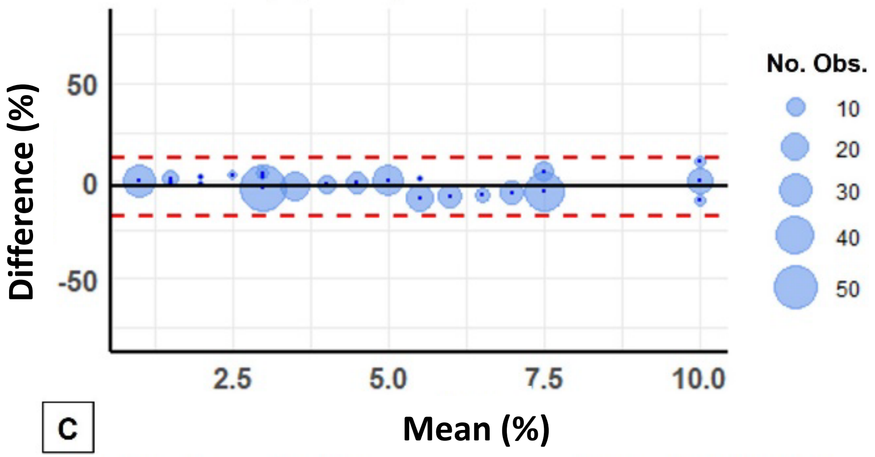
Upper LOA = 22.1 %
Mean diff = -0.9 %
Lower LOA = -23.8 %

Mean Difference (Bland-Altman) Plots for two pathologists with complete data

A

Low Density Paired Observations (nROI=319)

PathPresenter Paired Reader Bland-Altman

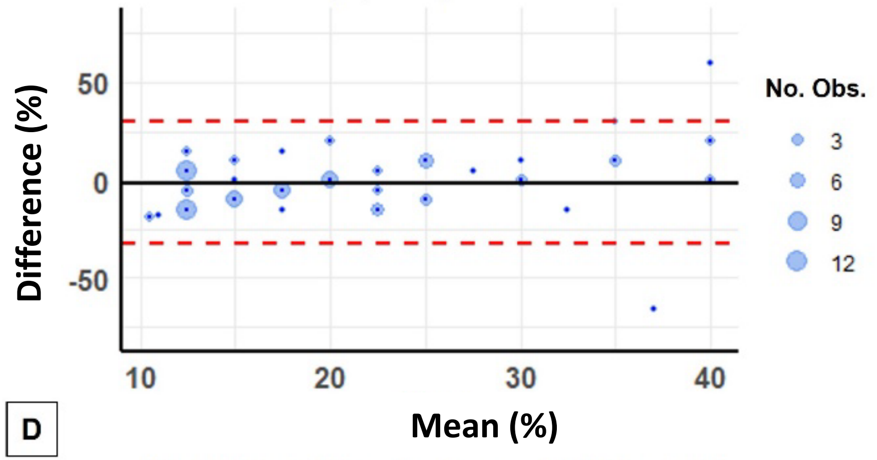


Upper LOA = 12.1 %
Mean diff = -2.9 %
Lower LOA = -17.8 %

B

Medium Density Paired Observations (nROI=93)

PathPresenter Paired Reader Bland-Altman

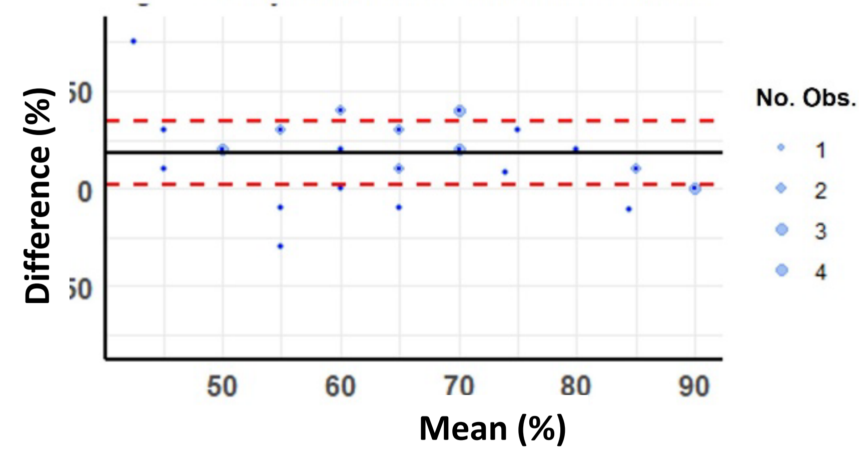


Upper LOA = 29.9 %
Mean diff = -1.3 %
Lower LOA = -35.6 %

C

High Density Paired Observations (nROI=36)

PathPresenter Paired Reader Bland-Altman

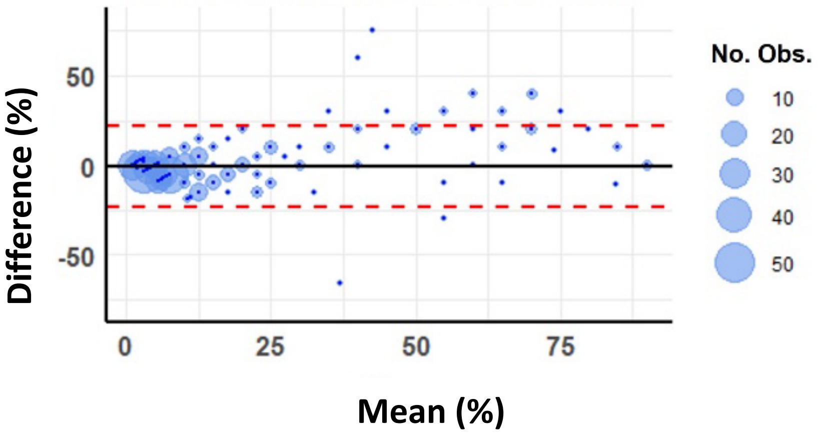


Upper LOA = 34.4 %
Mean diff = 18.1 %
Lower LOA = 1.8 %

D

All Paired Observations (nROI=448)

PathPresenter Paired Reader Bland-Altman



Upper LOA = 22.1 %
Mean diff = -0.9 %
Lower LOA = -23.8 %

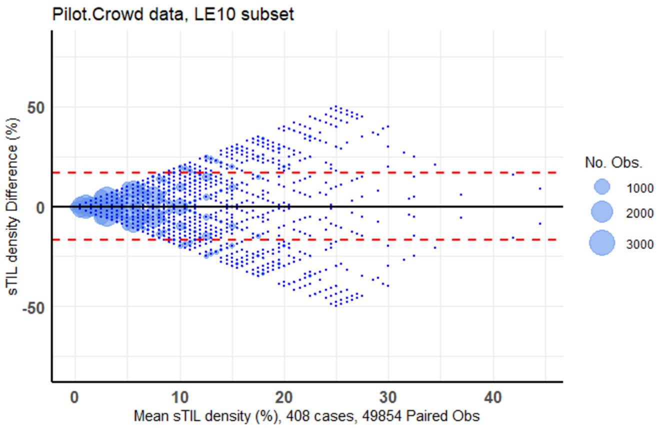
How should we determine …

- If a crowd pathologist is an expert?
- If an AI/ML model is good enough?

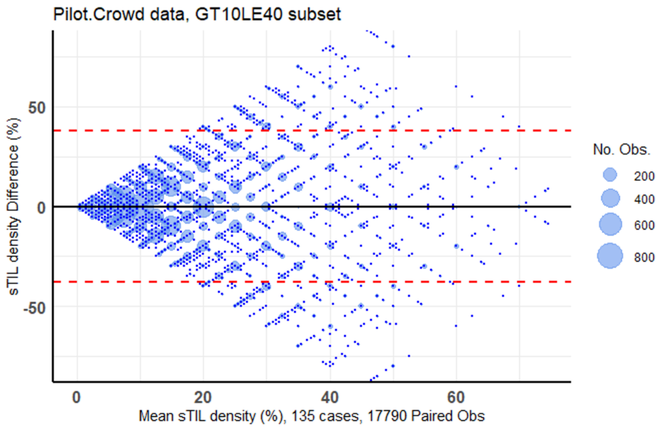
- First thought
  - Bland-Altman Plots
  - Limits of Agreement (LOA)

- Agreement of two pathologists
  - How do we incorporate multiple readers … multiple experts?

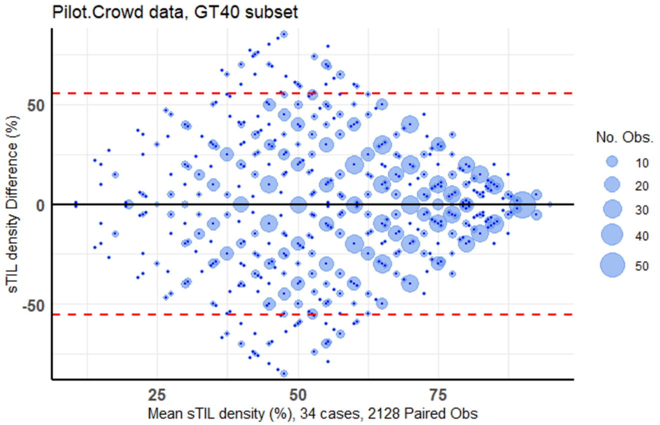
Mean Difference (Bland-Altman) Plots for seven pathologists with complete pilot data



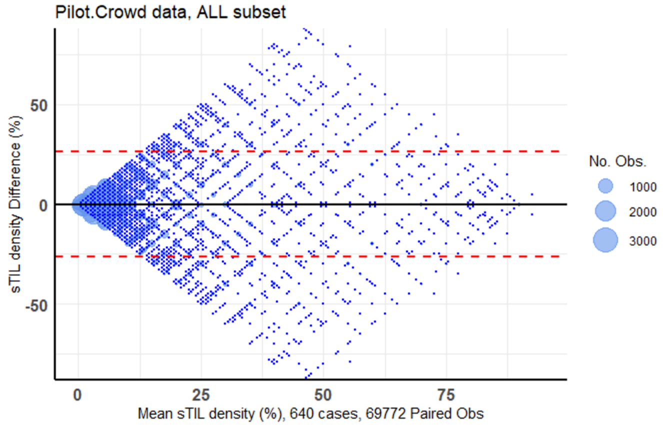
Mean Difference (Bland-Altman) Plots for seven pathologists with complete pilot data



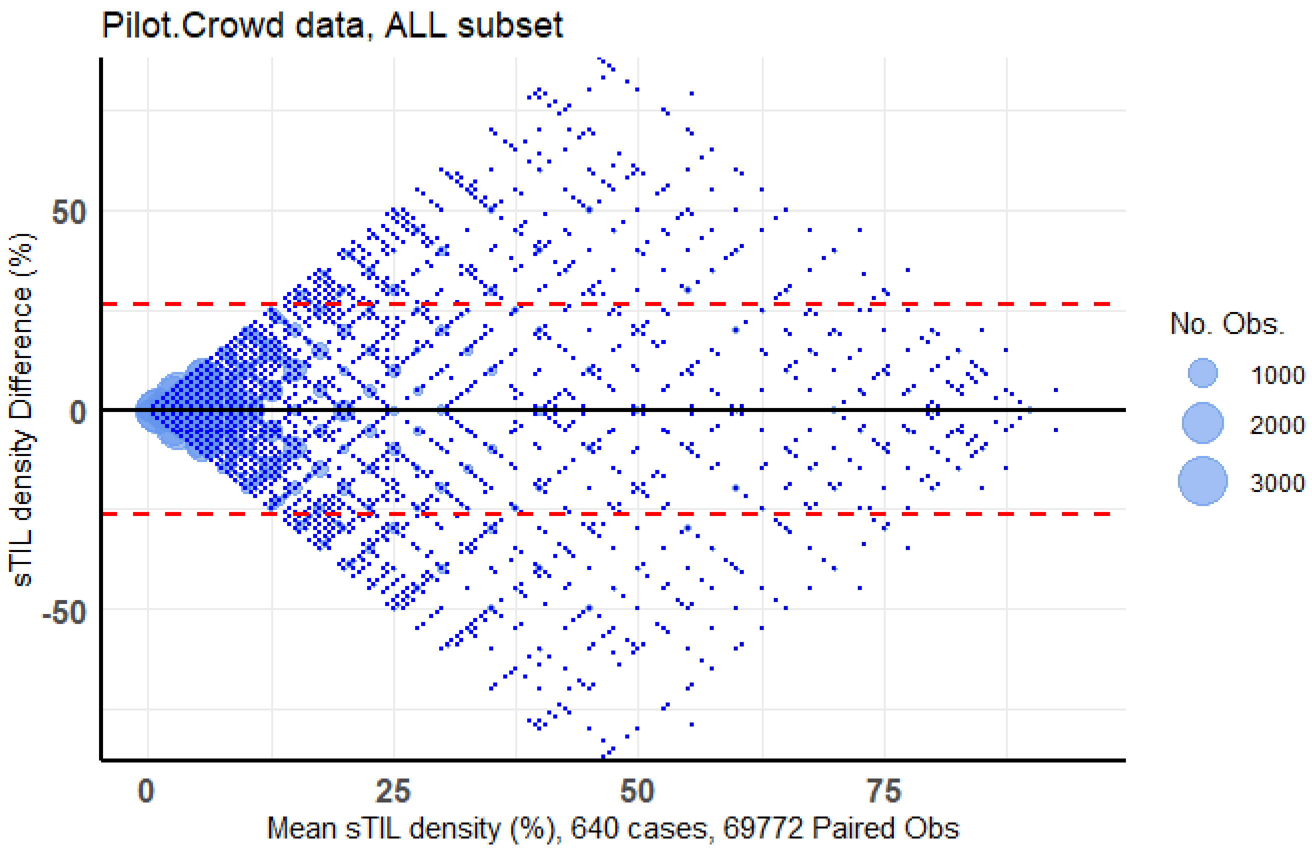
Mean Difference (Bland-Altman) Plots for seven pathologists with complete pilot data



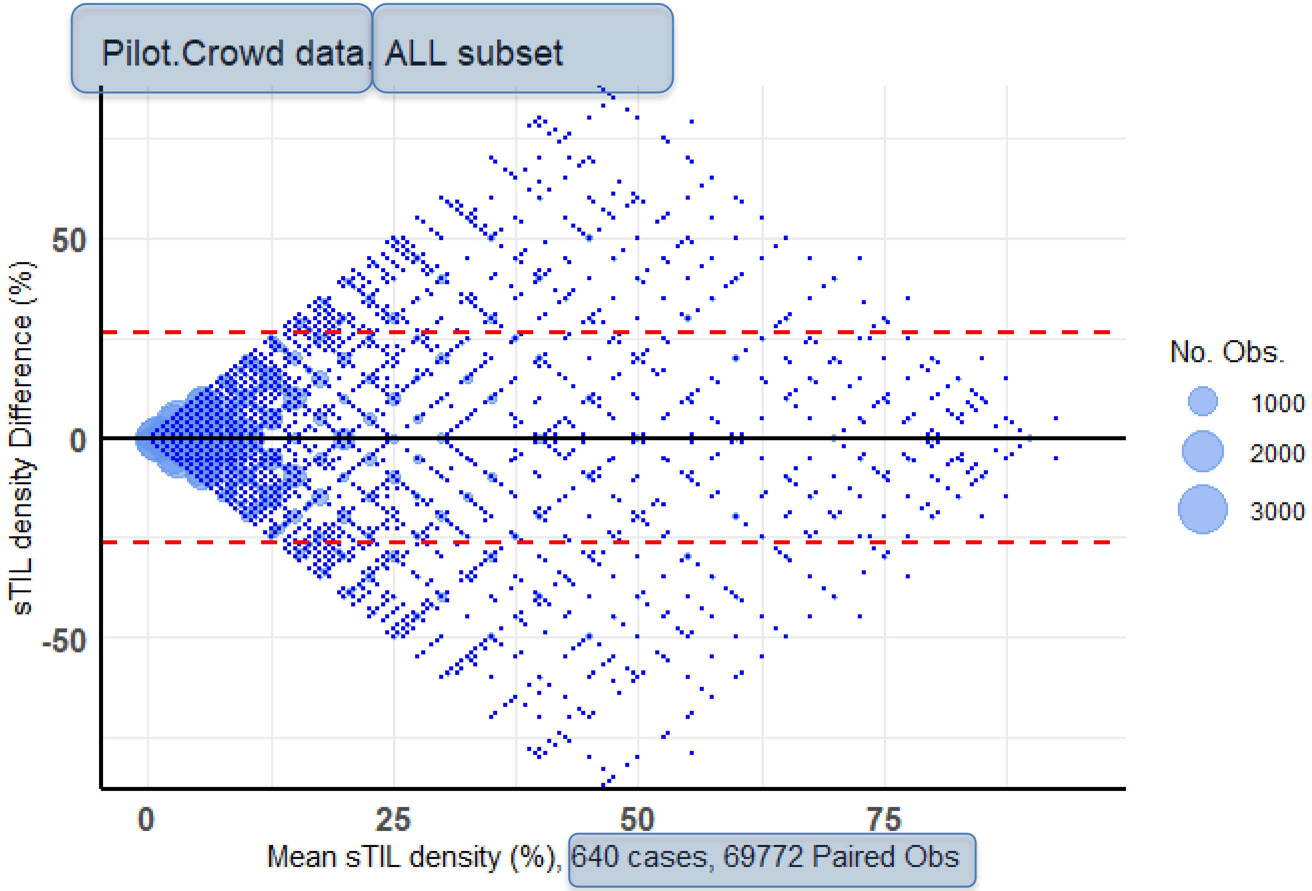
Mean Difference (Bland-Altman) Plots for seven pathologists with complete pilot data



Mean Difference (Bland-Altman) Plots for seven pathologists with complete pilot data

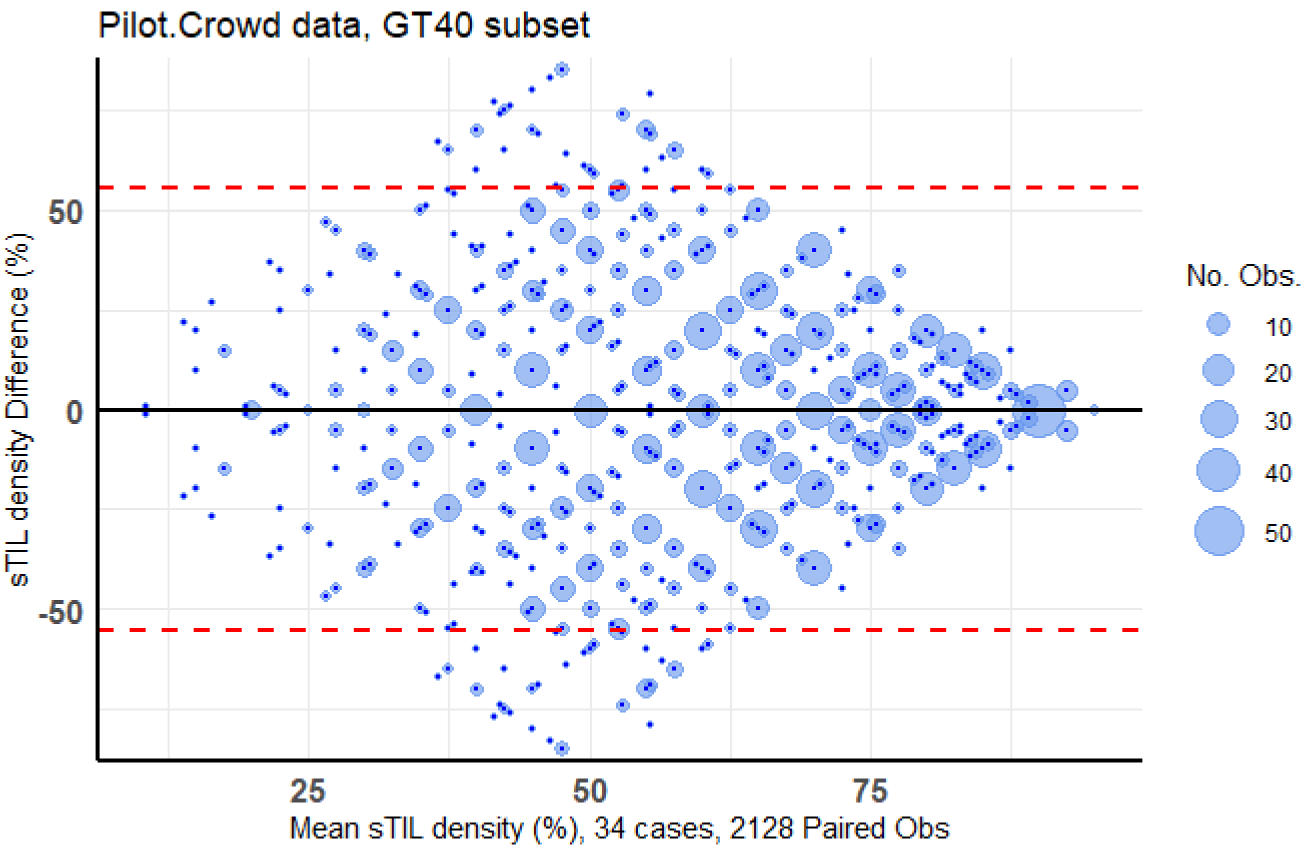


Mean Difference (Bland-Altman) Plots for seven pathologists with complete pilot data

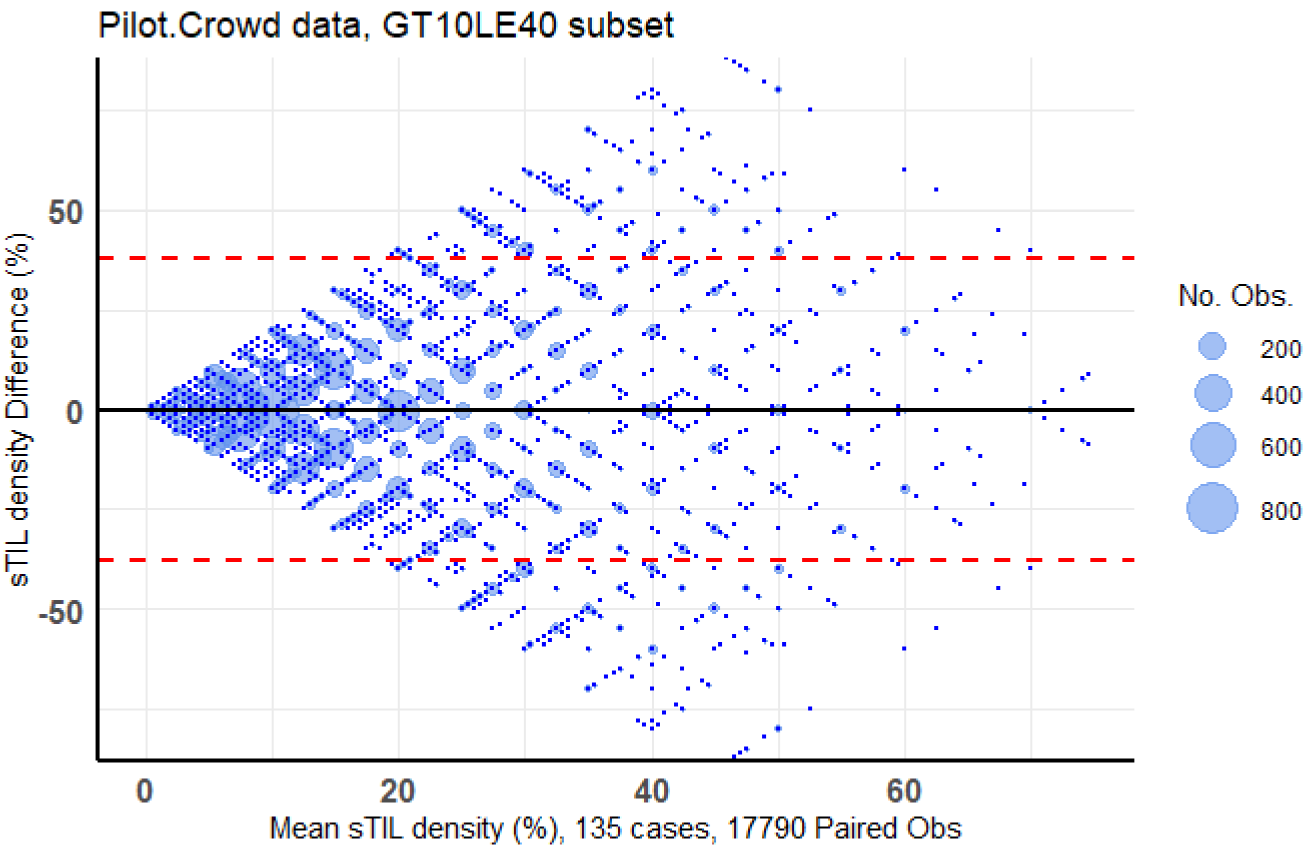


- Plot is symmetric by construction
  - Assume readers are equivalent
  - Difference 12: Reader 1 – Reader 2
  - Difference 21:
  - Reader 2 – Reader 1

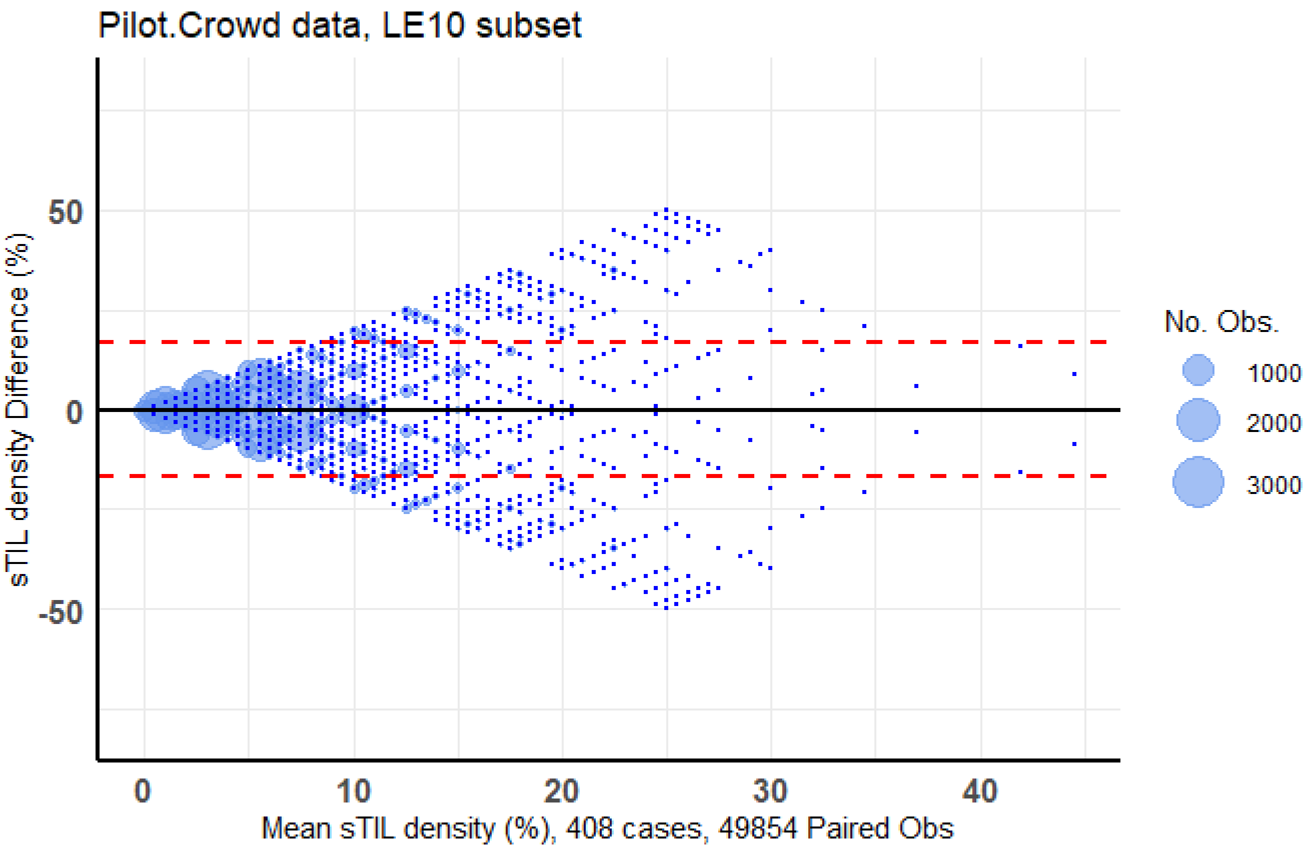
Mean Difference (Bland-Altman) Plots for seven pathologists with complete pilot data



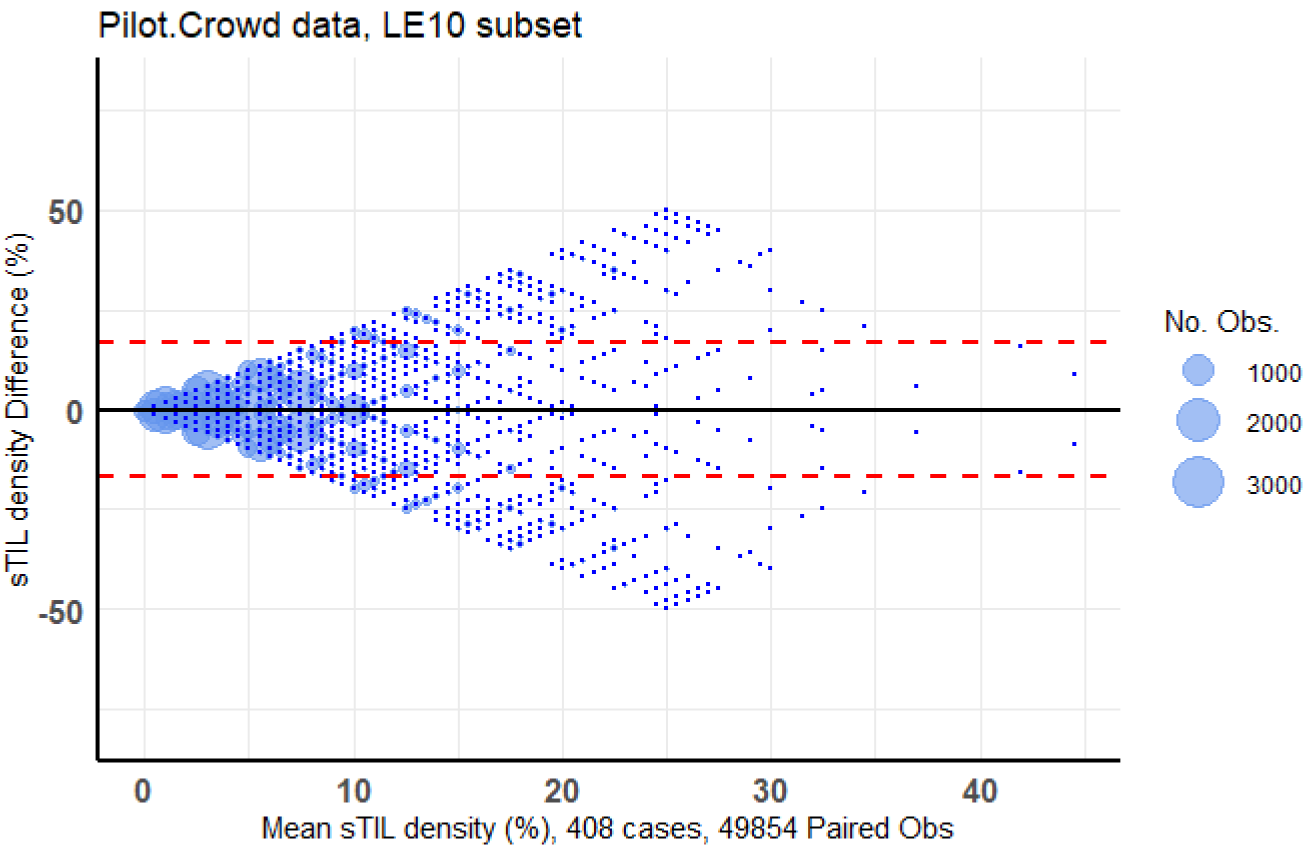
Mean Difference (Bland-Altman) Plots for seven pathologists with complete pilot data



Mean Difference (Bland-Altman) Plots for seven pathologists with complete pilot data



Mean Difference (Bland-Altman) Plots
for seven pathologists with complete pilot data



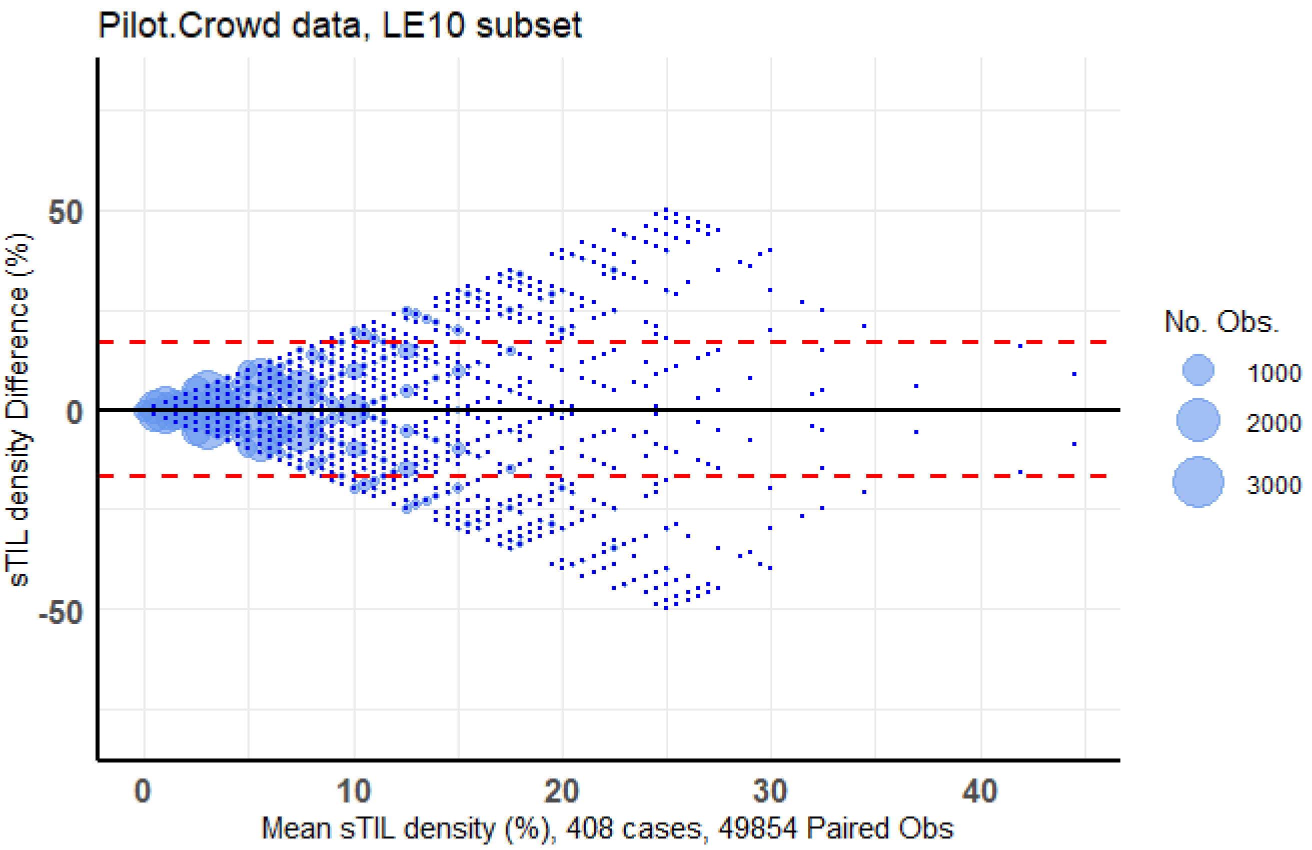
Two readers

- Upper LOA = 12.1 %
- Mean diff = -2.9 %
- Lower LOA = -17.8 %

Seven readers, MRMC analysis

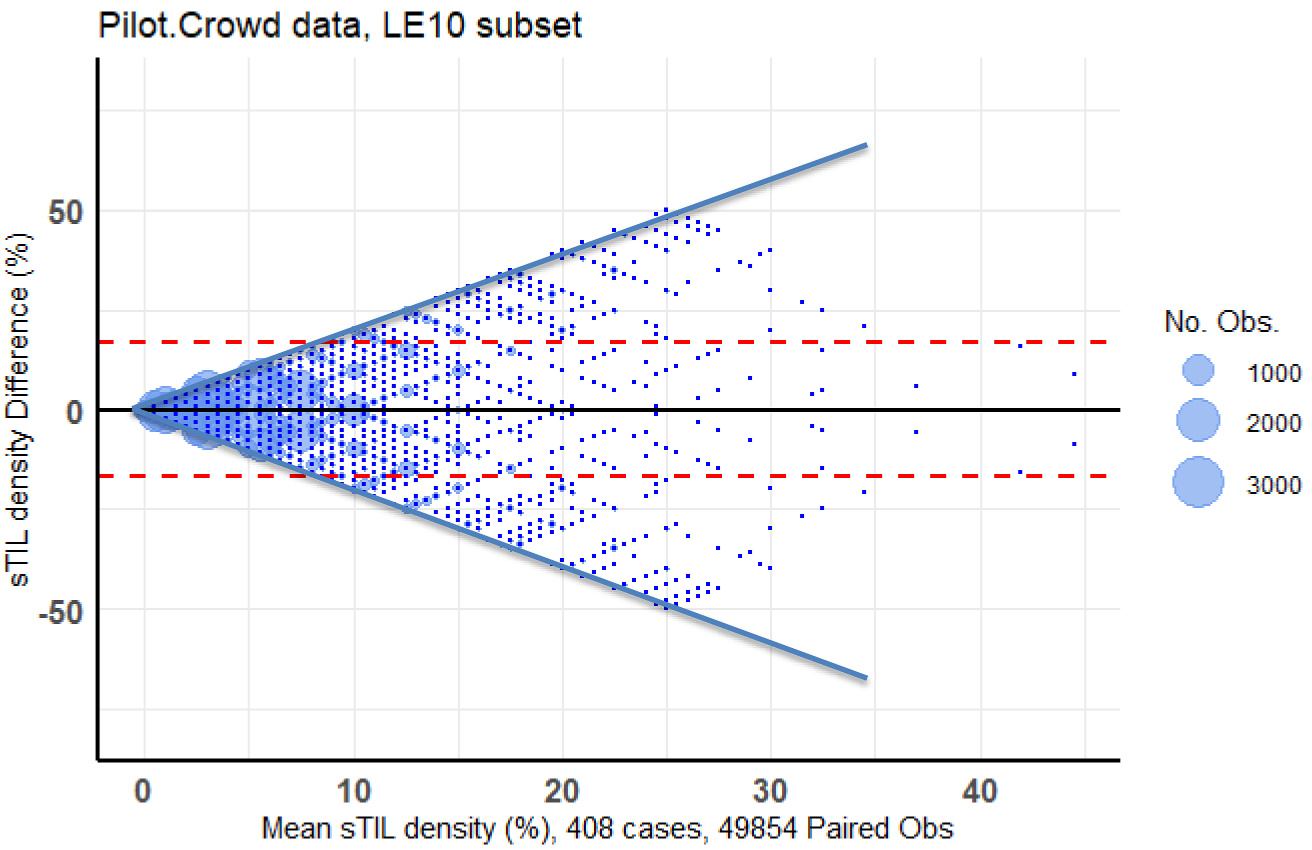
- Upper LOA = 17.8 %
- Mean diff = 0 (By construction)
- Lower LOA = -17.8 %

Mean Difference (Bland-Altman) Plots for seven pathologists with complete pilot data



- Differences not independent
  - Multiple readers, Multiple Cases
  - Fully-crossed data
- Differences not identically distributed
  - Differences increase with the mean

Mean Difference (Bland-Altman) Plots for seven pathologists with complete pilot data



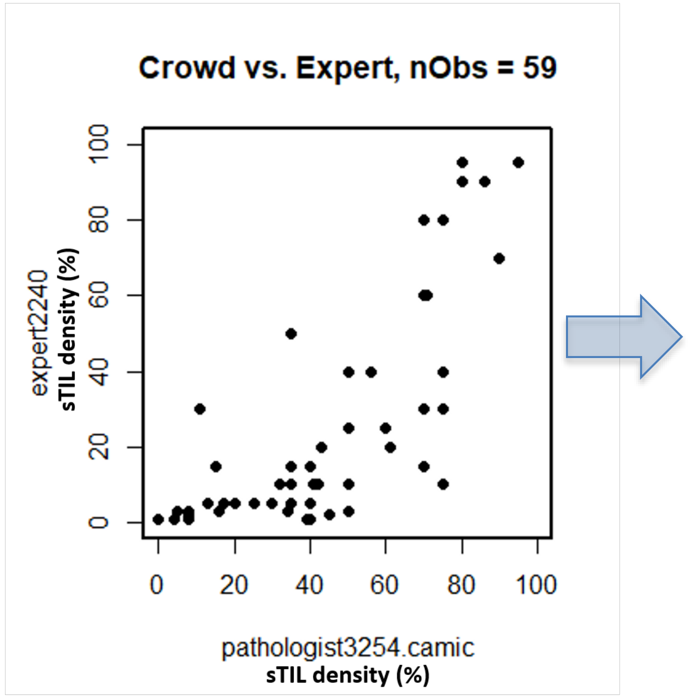
- Differences not independent
  - Multiple readers, Multiple Cases
  - Fully-crossed data
- Differences not identically distributed
  - Differences increase with the mean
- Differences not normally distributed
  - Cone of maximum possible difference

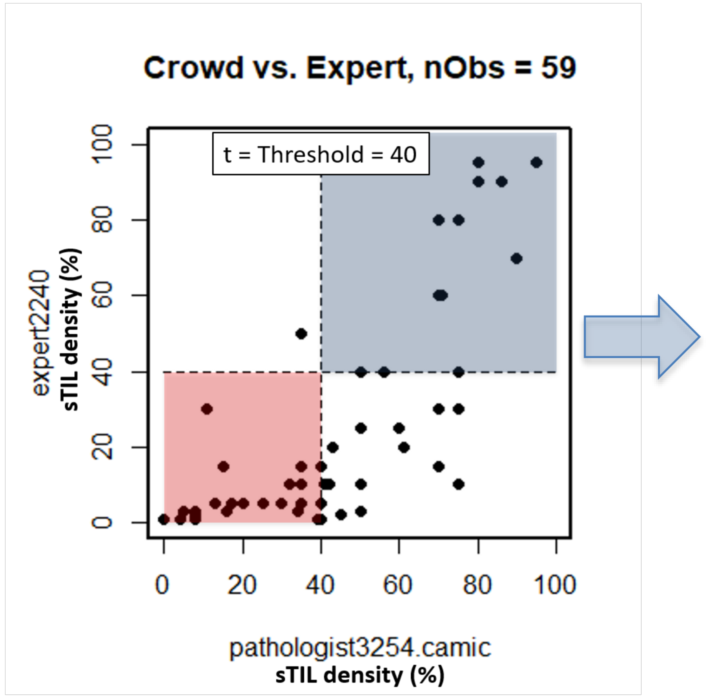
How should we determine …

- If a pathologist is an expert?
- If an AI/ML model is good enough?

- First thought
  - Bland-Altman Plots
  - Limits of Agreement (LOA)

- Assumptions not satisfied … Good for exploratory analysis
  - What next?





Crowd-Expert Agreement

threshold	expert	crowd
40	expert2240	pathologist3254.camic

		crowd		Row Total	Fraction Agree	Standard Error
		≤ t	> t			
Expert	> t	1	10	11	0.91	0.0867
	≤ t	30	18	48	0.63	0.0699

Agreement with Expert 1

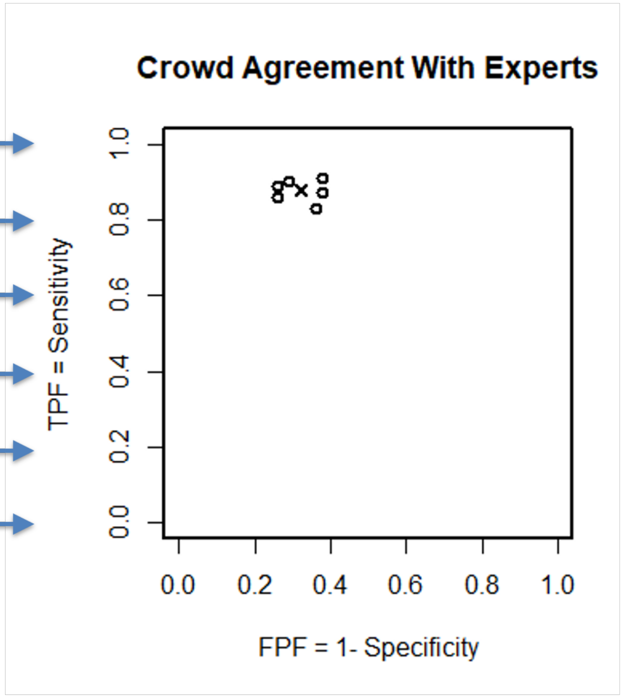
Agreement with Expert 2

Agreement with Expert 3

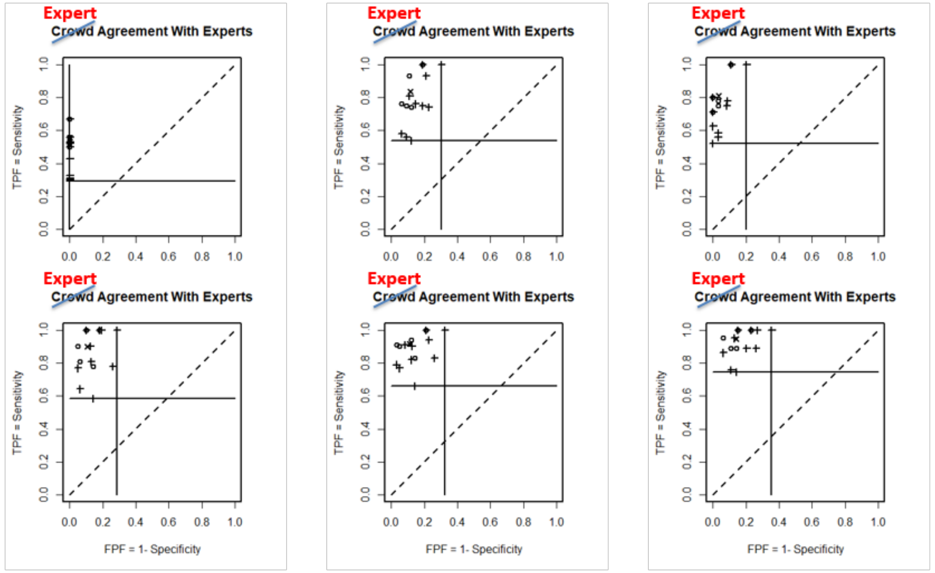
Agreement with Expert 4

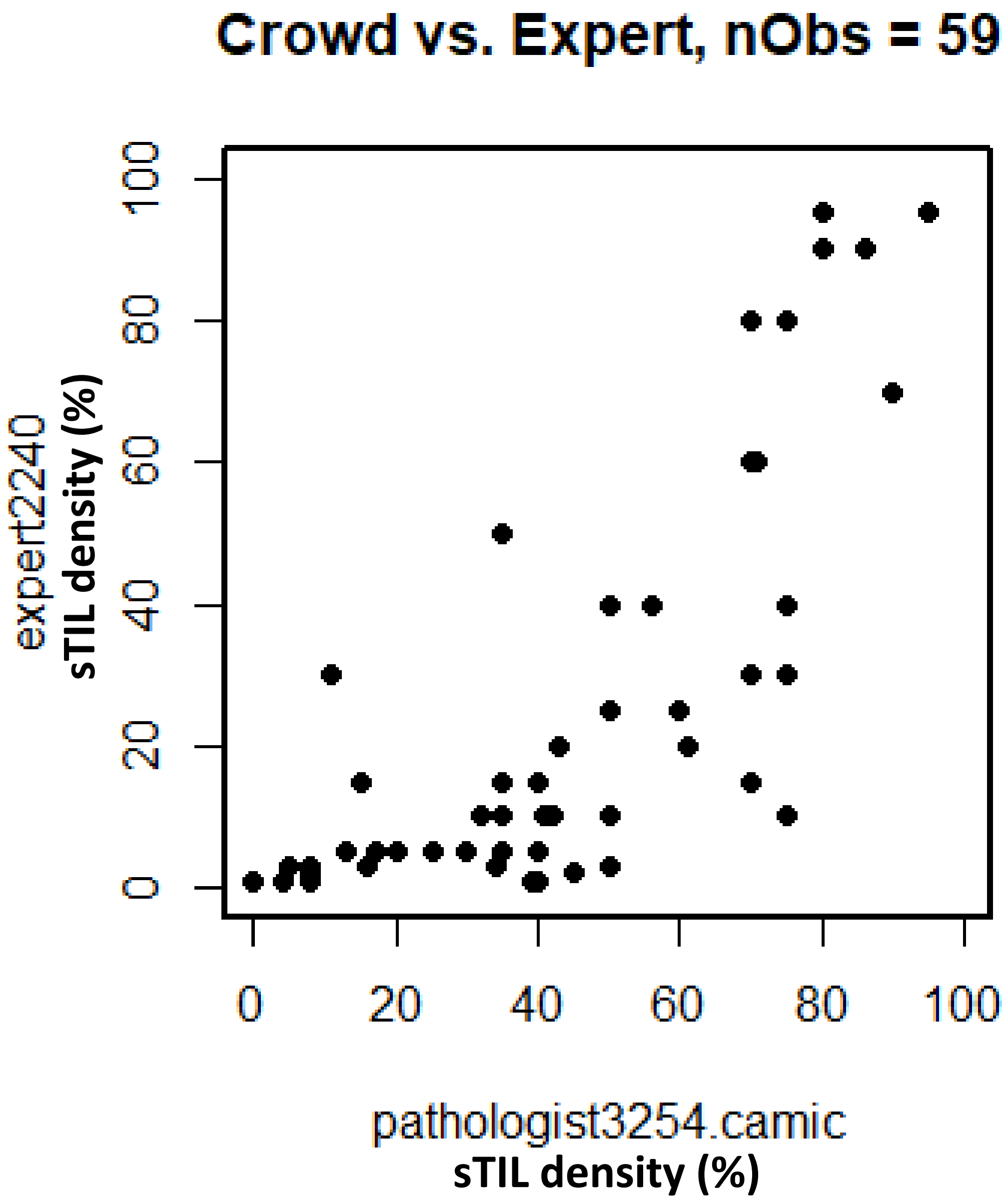
Agreement with Expert 5

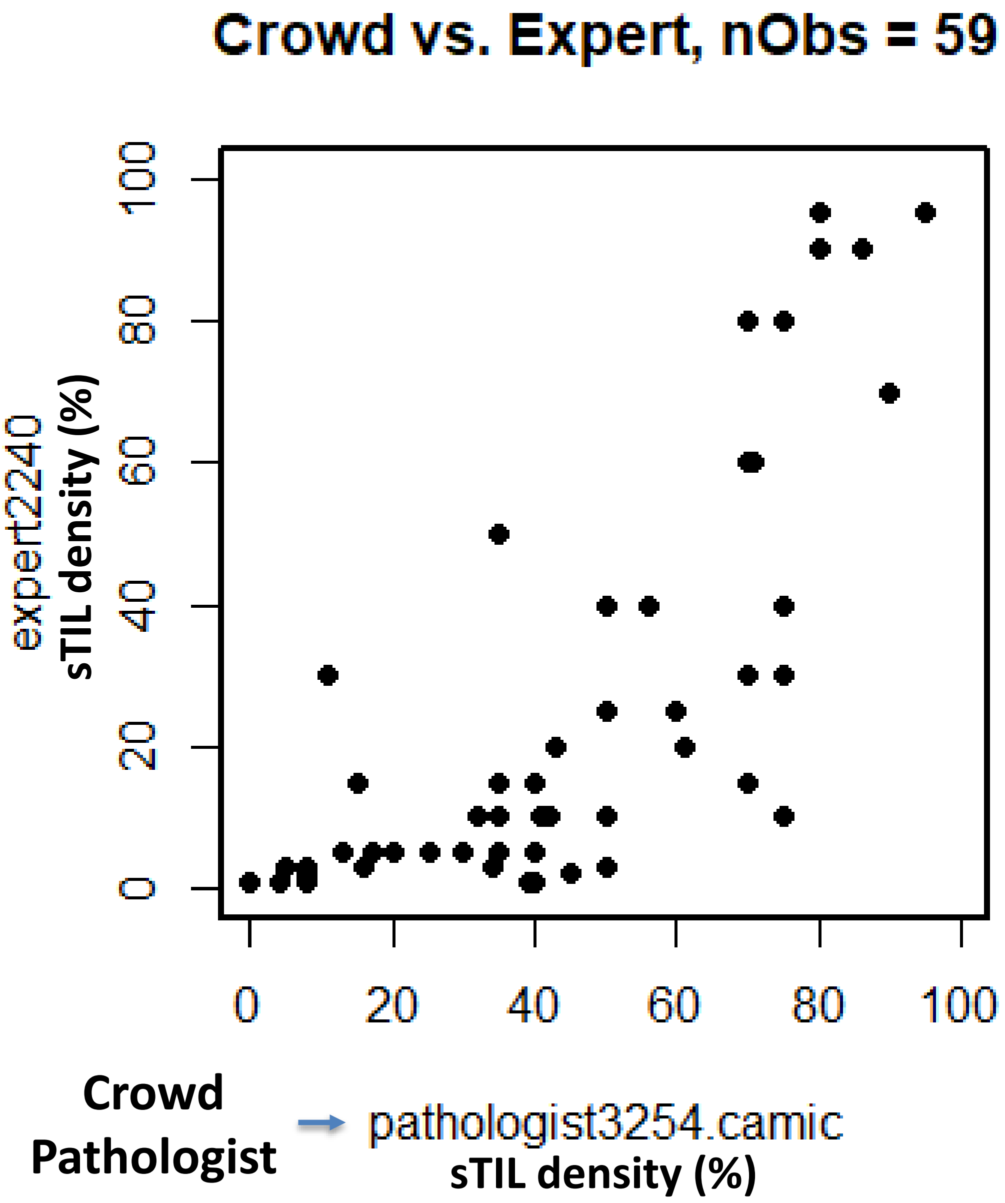
Agreement with Expert 6



Expert-Expert Agreement

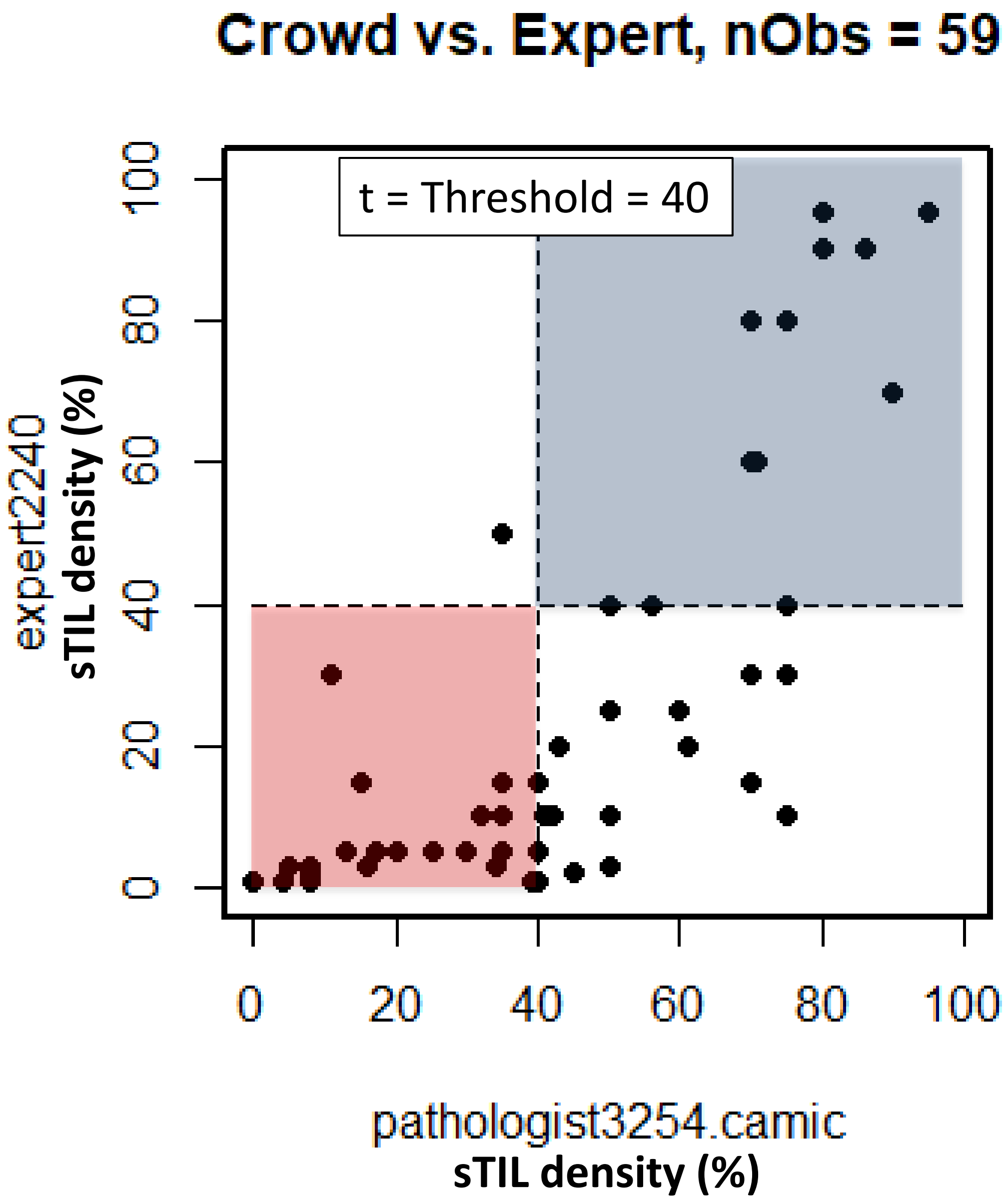



Crowd vs. Expert, nObs = 59
expert2240
sTIL density (%)
0
20
40
60
80
100
0
20
40
60
80
100
pathologist3254.camic
sTIL density (%)

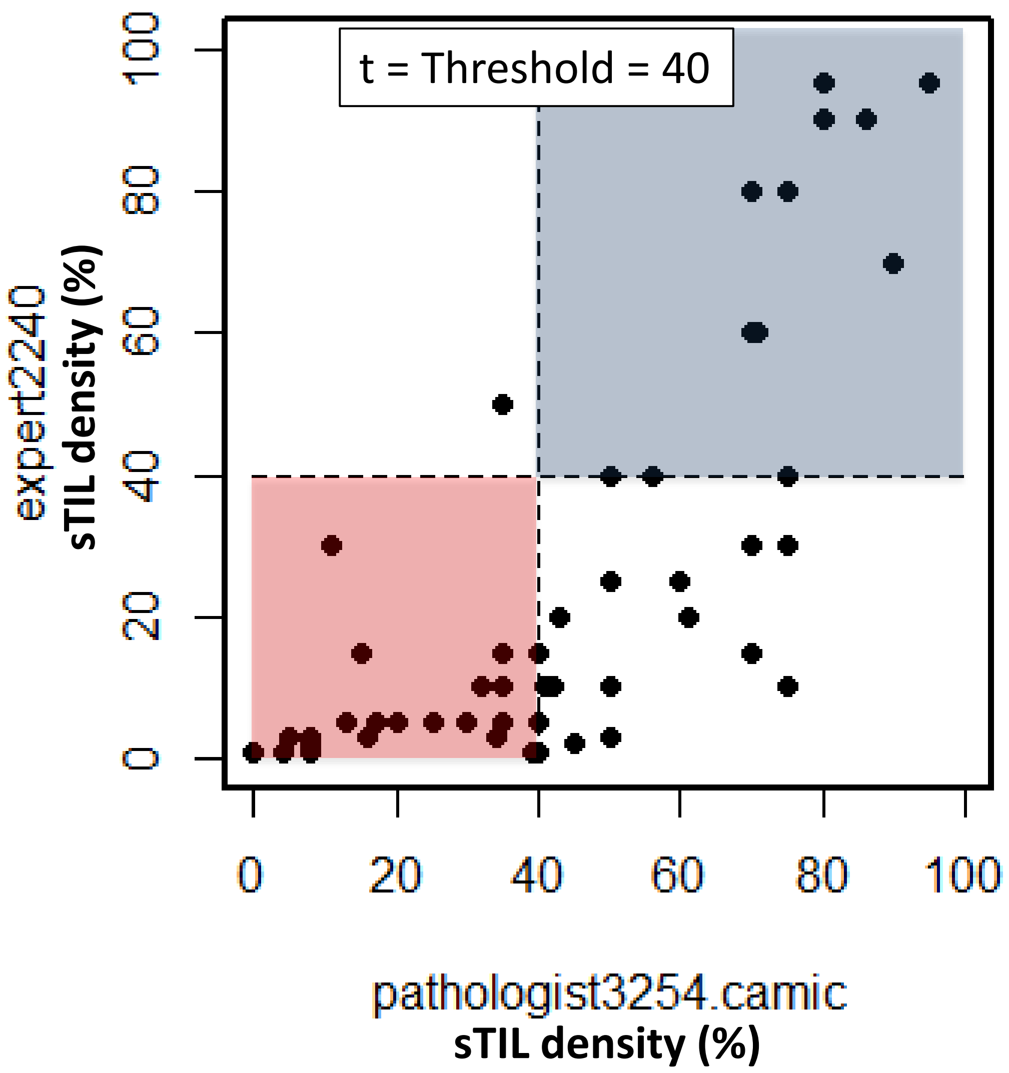


Crowd Pathologist →

- Crowd pathologist
  - Typical data
  - Substitute "AI Model"
- SELECT data
  - 72 cases, some labeled not evaluable
- Not clustered around diagonal
- Not normally distributed
- Not IID


Crowd vs. Expert, nObs = 59
t = Threshold = 40
expert2240
sTIL density (%)
0
20
40
60
80
100
0
20
40
60
80
100
pathologist3254.camic
sTIL density (%)

Crowd vs. Expert, nObs = 59



		crowd	
		≤ t	> t
Expert	> t	1	10
	≤ t	30	18

Crowd-Expert Agreement

threshold	expert	crowd
40	expert2240	pathologist3254.camic

		crowd		Row Total	Fraction Agree	Standard Error
		≤ t	> t			
Expert	> t	1	10	11	0.91	0.0867
	≤ t	30	18	48	0.63	0.0699

Crowd-Expert Agreement

threshold	expert	crowd
40	expert2240	pathologist3254.camic

		crowd		Row Total	Fraction Agree	Standard Error
		≤ t	> t			
Expert	> t	1	10	11	0.91	0.0867
	≤ t	30	18	48	0.63	0.0699

- TPF = Fraction Agree "> t"
- FPF = Fraction Agree "≤ t"

- TPF and FPF

  understood to be

- Crowd-Expert Agreement

- Compare Crowd to all Experts

Agreement with Expert 1

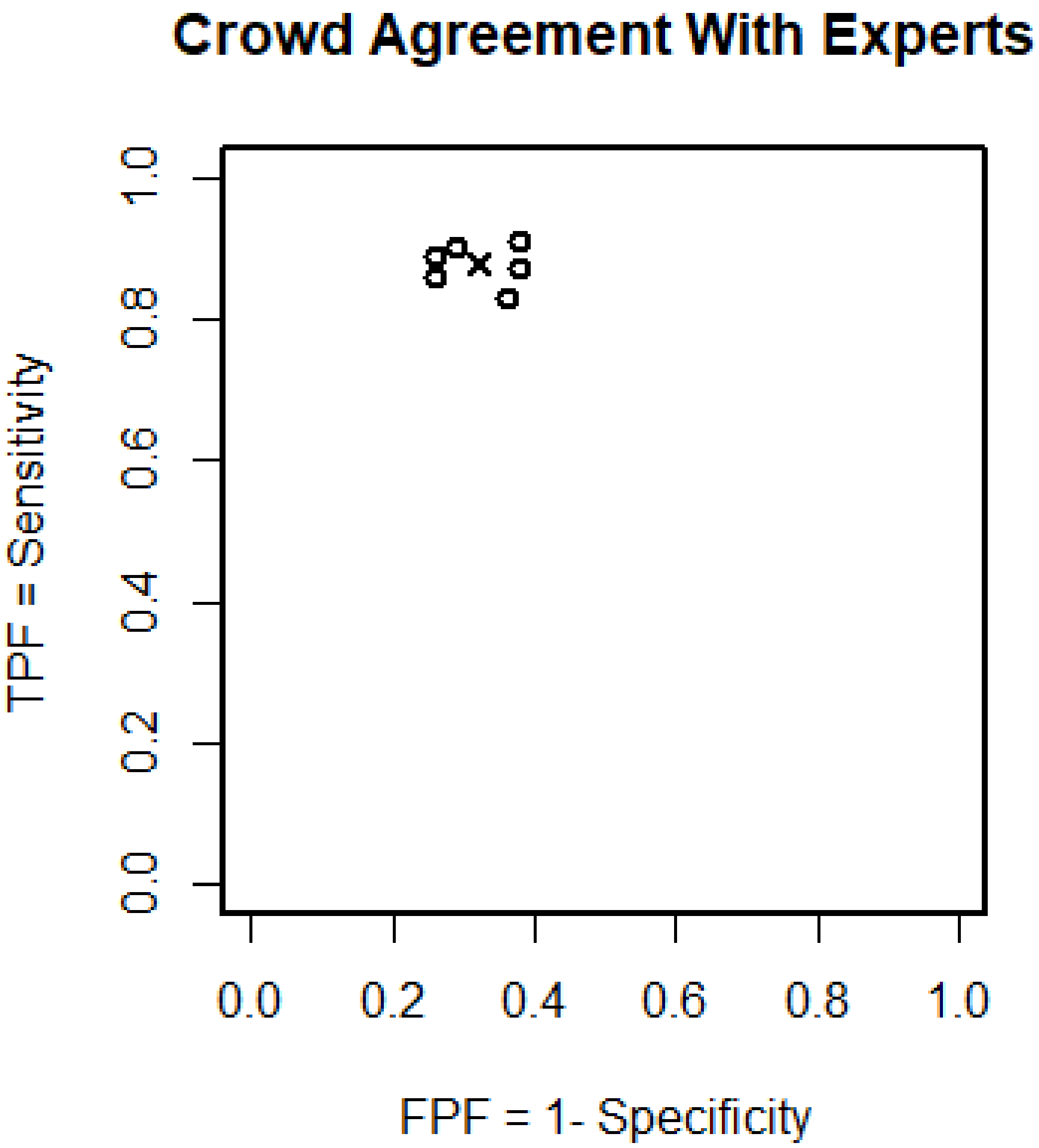
Agreement with Expert 2

Agreement with Expert 3

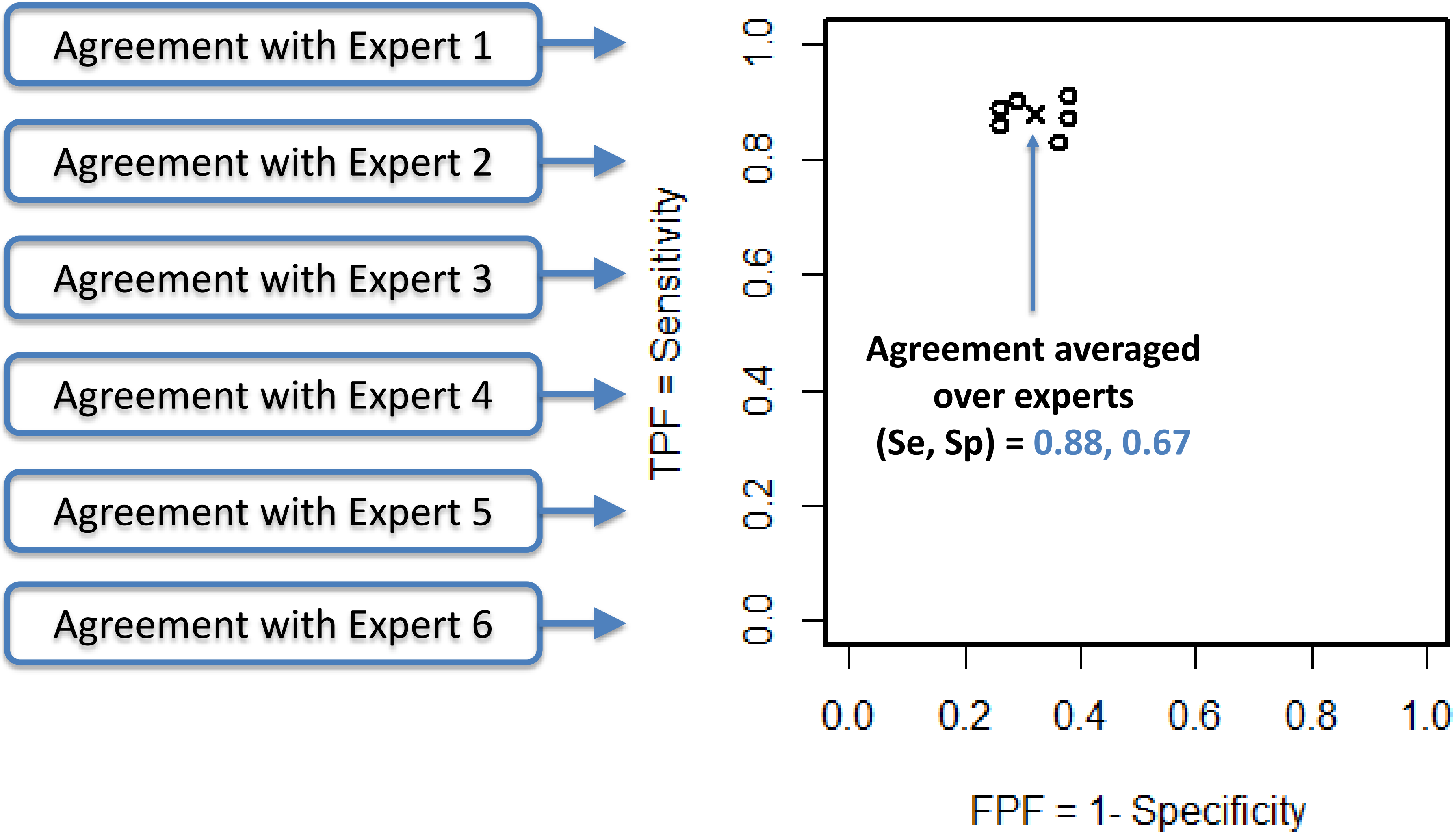
Agreement with Expert 4

Agreement with Expert 5

Agreement with Expert 6



Crowd Agreement With Experts



Agreement with Expert 1

Agreement with Expert 2

Agreement with Expert 3

Agreement with Expert 4

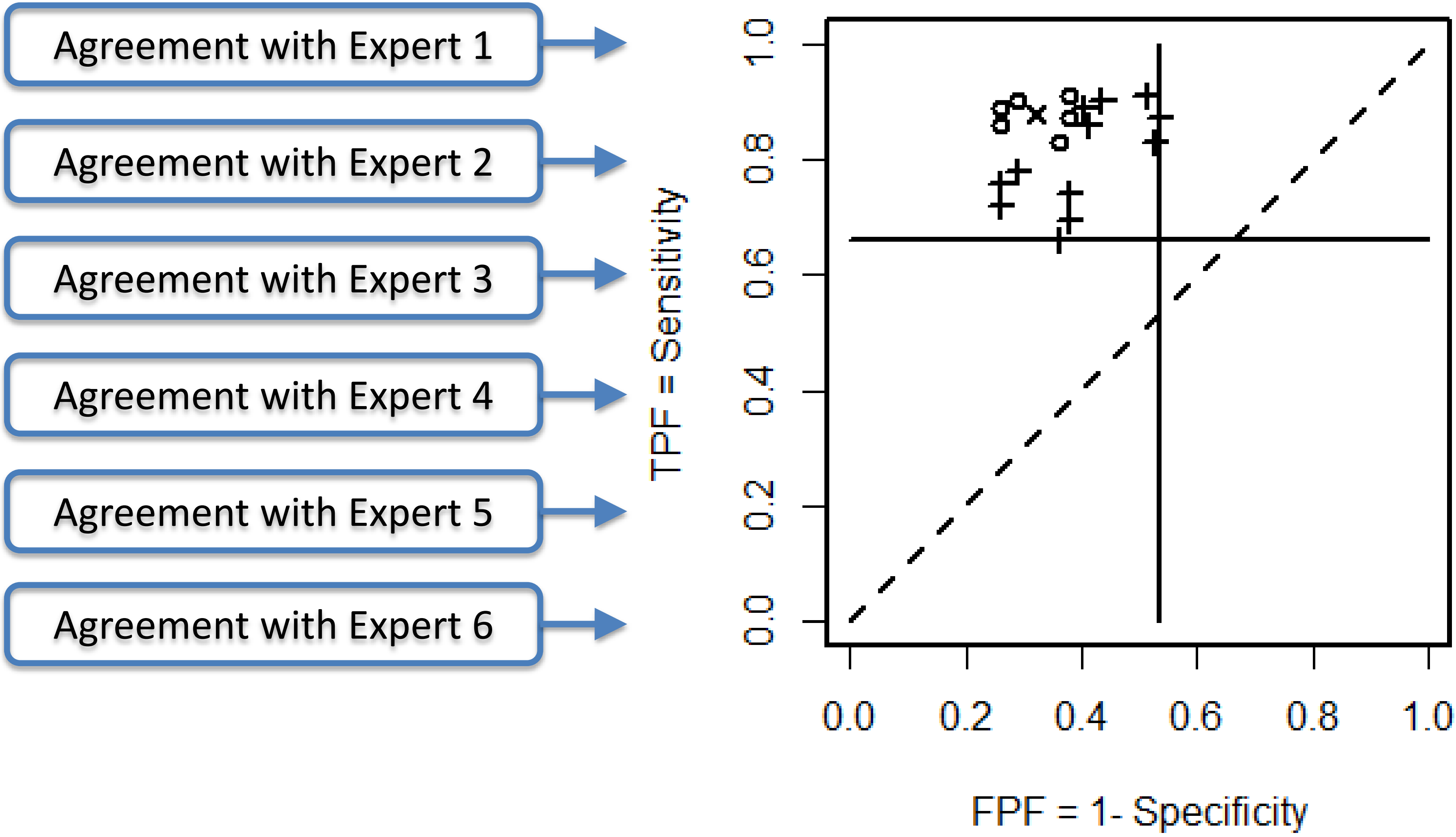
Agreement with Expert 5

Agreement with Expert 6

Agreement averaged over experts
(Se, Sp) = 0.88, 0.67

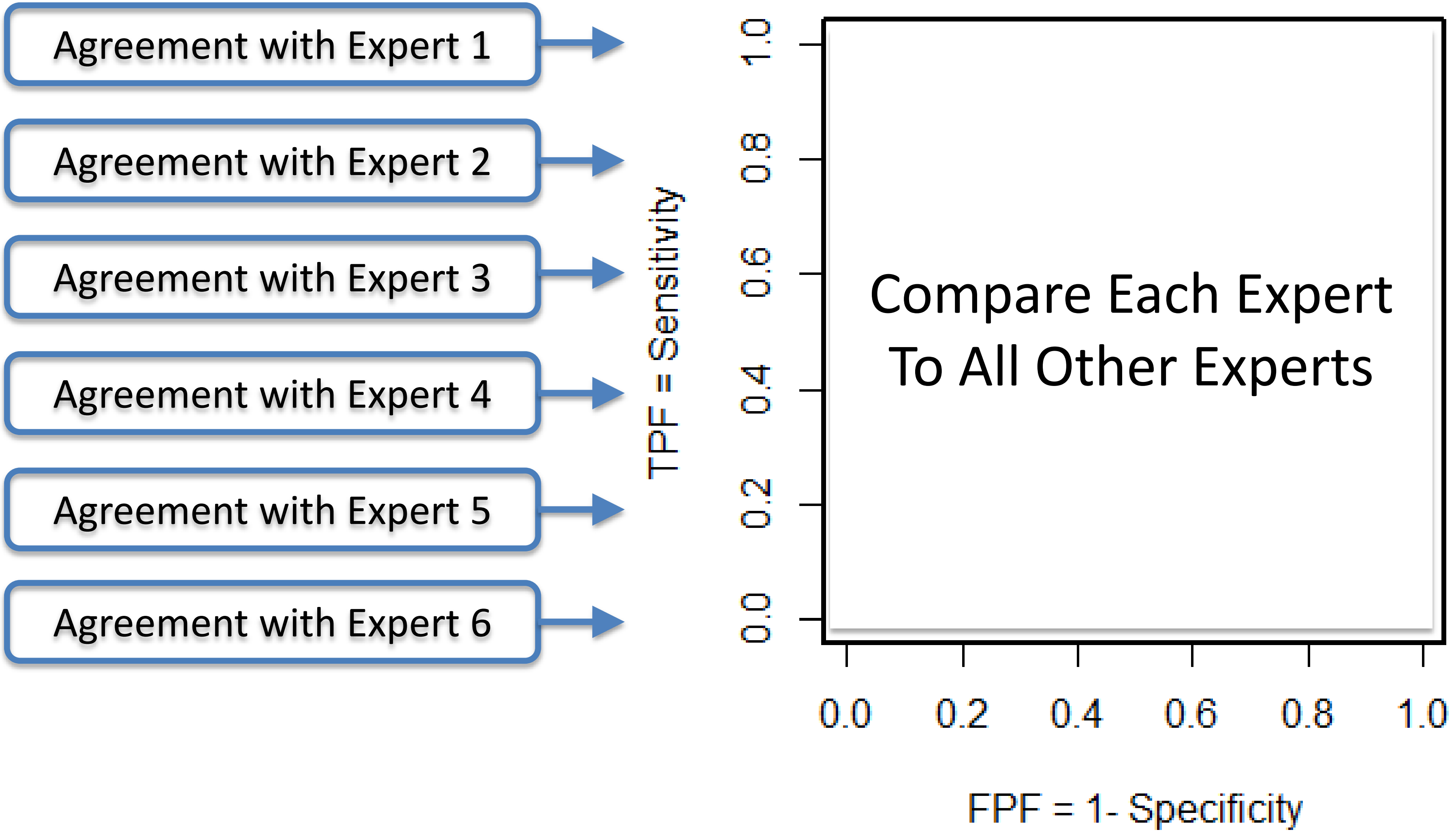
- Circles:
  - Agreement with each expert
- NEXT:
  - Uncertainty given each expert

Crowd Agreement With Experts



- Circles:
  - Agreement with each expert
- "+"
  - Lower bound of the 95% confidence interval
- Lines
  - Minimum (over experts) of the lower bound of the 95% confidence intervals

Expert
~~Crowd~~ Agreement With Experts



Agreement with Expert 1

Agreement with Expert 2

Agreement with Expert 3

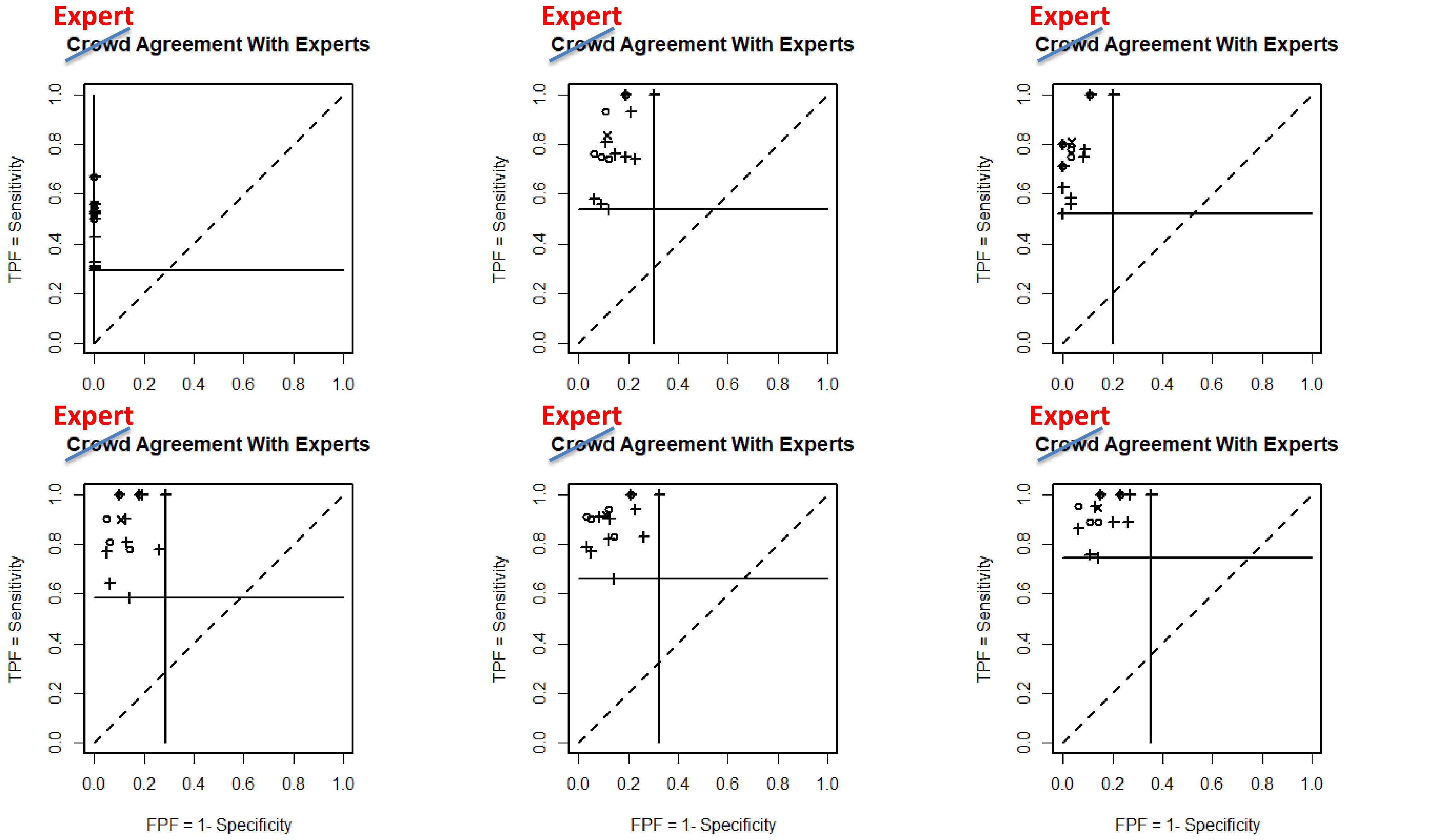
Agreement with Expert 4

Agreement with Expert 5

Agreement with Expert 6

Compare Each Expert To All Other Experts

- Circles:
  - Agreement with each expert
- "+"
  - Lower bound of the 95% confidence interval
- Lines
  - Minimum (over experts) of the lower bound of the 95% confidence intervals


Expert
~~Crowd~~ Agreement With Experts
TPF = Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
Expert
~~Crowd~~ Agreement With Experts
TPF = Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
Expert
~~Crowd~~ Agreement With Experts
TPF = Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
Expert
~~Crowd~~ Agreement With Experts
TPF = Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
FPF = 1- Specificity
Expert
~~Crowd~~ Agreement With Experts
TPF = Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
FPF = 1- Specificity
Expert
~~Crowd~~ Agreement With Experts
TPF = Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
FPF = 1- Specificity

How should we determine …

- If a pathologist is an expert?
- If an AI/ML model is good enough?

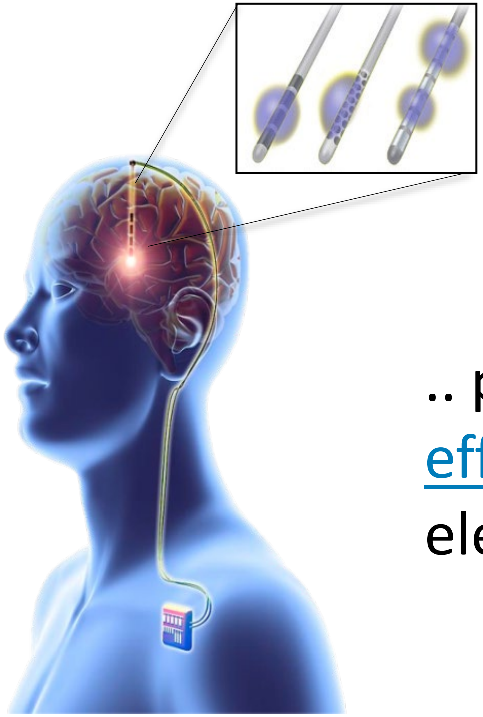
- Current Strategy
  - Compare crowd reader to all experts
  - Compare each expert to all other experts
  - Establish criteria for a crowd-expert agreement
  - Develop Multi-Expert Multi-Case (MEMC) analysis methods

Summary

- Clinical Context: Imaging Biomarker
- Initial Analysis of Pilot Study
- Quantitative Agreement
  - Bland-Altman … Limits of Agreement
- Strategy to Use Thresholds
  - Binary Crowd-Expert Agreement for each Expert
  - Then Average over Experts
  - Baseline performance: Expert-Expert Agreement
  - **Develop Multi-Expert Multi-Case (MEMC) Analysis Methods**

Conclusions

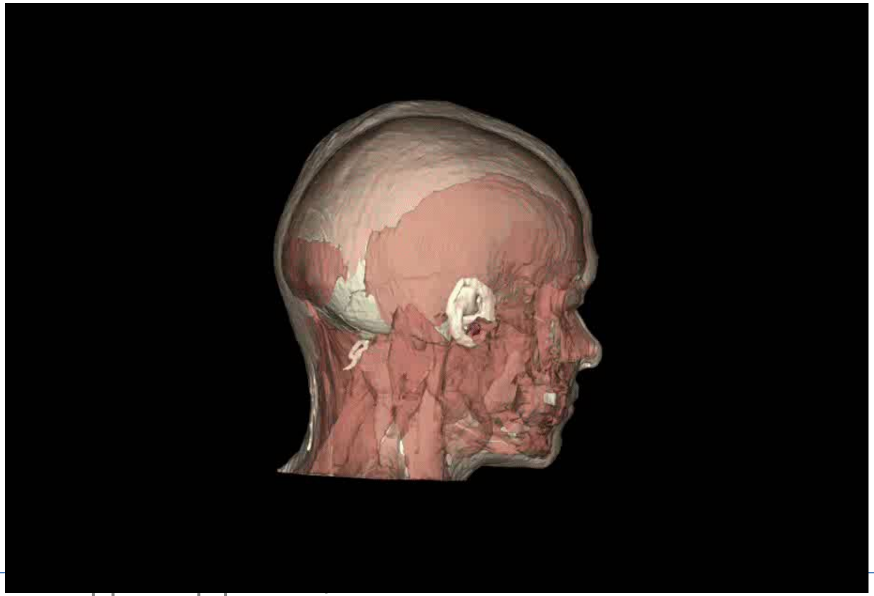
- Analyzing objective estimates of quantitative values from humans is hard!
  - I object to referring to the estimates are “subjective”
  - Not based on or influenced by personal feelings, tastes, or opinions
  - They are noisy
- Data from humans violate assumptions for Limits of Agreement
  - Not normally distributed
  - Not independent and identically distributed
- Strategies that treat the data as ordinal can sidestep assumptions
  - Add calibration thresholds
  - Explore calibration thresholds



CDRH Mission

.. protect and promote the health of the public by ensuring the safety and effectiveness of **medical devices** and the safety of radiation-emitting electronic products…

We facilitate medical device innovation by advancing regulatory science, providing industry with predictable, consistent, transparent, and efficient regulatory pathways, and assuring consumer confidence in devices marketed in the U.S.



CDRH in Perspective

1900
EMPLOYEES

18k
Medical Device Manufacturers

183k
Medical Devices
On the U.S. Market

22k/year
Premarket Submissions
includes supplements and amendments

570k
Proprietary Brands

25k
Medical Device Facilities Worldwide

1.4 MILLION/year
Reports on medical device adverse events and malfunctions

Office of Science and Engineering Laboratories (OSEL)

- Conduct laboratory-based regulatory research to facilitate development and innovation of safe and effective medical devices and radiation emitting products
- Provide scientific and engineering expertise, data, and analyses to support regulatory processes
- Collaborate with colleagues in academia, industry, government, and standards development organizations to develop, translate, and disseminate science and engineering-based information regarding regulated products
- https://www.fda.gov/about-fda/cdrh-offices/office-science-and-engineering-laboratories

OSEL in Perspective

183
FEDERAL EMPLOYEES
Up to 180 visiting scientists

140 Projects
In 27 Laboratories and Program Areas

400/year
Peer reviewed presentations, articles, and other public disclosures

2,500k/year

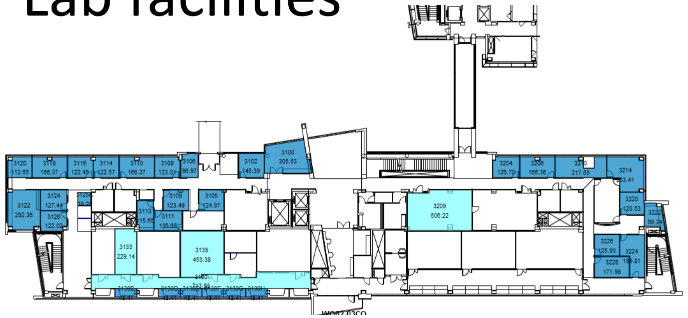
Premarket Regulatory consults

75
Standards and conformity assessment committees

70%
Staff with post graduate degree

55,000 ft^2

Lab facilities



Division of Imaging, Diagnostics and Software Reliability (DIDSR)

- Develop least burdensome approaches for regulatory evaluation of imaging and big-data devices
  - Efficient clinical trials accounting for reader variability, simulation tools, in silico phantoms and imaging trials, addressing issues related to imperfect / missing reference standards, and limited data for training/testing of machine classifiers
- Develop measures of technical effectiveness of imaging and big-data technologies
  - Phantoms, laboratory measurements, computational models

DIDSR in Perspective

35

FEDERAL EMPLOYEES

14 Fellows/Students

3 Open Staff Positions

145/year

Peer reviewed articles, code and presentations

4 Program Areas

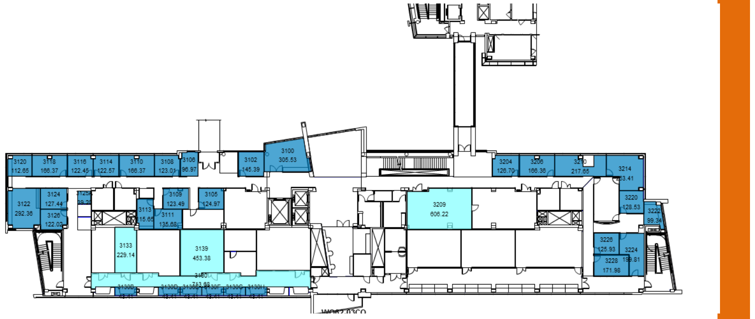
- AI/ML
- Medical Imaging and Diagnostics
- Digital Pathology
- Mixed Reality (AR/VR/XR)

550/year

Premarket Regulatory consults

~15,000 ft^2

DIDSR Lab and facilities


3120 112.65
3118 166.37
3116 122.45
3114 122.57
3110 166.37
3108 123.01
3106 96.97
3102 145.39
3100 305.53
3204 126.70
3206 166.36
3210 217.65
3214 253.41
3122 292.36
3124 127.44
3126 122.02
3109 123.49
3105 124.97
3113 115.65
3111 135.68
3209 606.22
3220 128.53
3222 99.34
3133 229.14
3139 453.38
3226 125.93
3224 199.81
3228 171.98
WO62 03CO