Can AI Grand-Challenges inform Regulatory Science in Anatomic Pathology?

Roberto Salgado and Francesco Ciompi

Pathology Innovation Collaborative Community

February 28, 2022

Radboudumc

TILS BREAST CANCER
INTERNATIONAL IMMUNO-ONCOLOGY WORKING GROUP

Disclosures

- NO financial conflicts of interest on the topics discussed at this presentation, for neither RS and FC.

Digital medicine

Bridging the chasm between AI and clinical implementation

*Angela Aristidou, Rajesh Jena, Eric J Topol
UCL School of Management, University College London, London E14 5AA, UK (AA); Department of Oncology, University of Cambridge, Cambridge, UK (RJ); Scripps Research Translational Institute, Scripps Research La Jolla, CA, USA (EJT)
a.aristidou@ucl.ac.uk

- Many advances in artificial intelligence (AI) for health care using deep neural networks have been commercialized. But **few** AI tools have been implemented in health systems. Why has this chasm occurred?
- **Transparency, suitability, and adaptability** are key reasons.
- For the information technology (IT) teams, there is the concern that input data are drawn from **outside the health setting** and the algorithm performance, source code, and input data are unavailable to review.
- Many commercial AI applications are in radiology, but **few** are supported by evidence from **published studies**.
- And there are concerns that the algorithms were tested and validated using **retrospective, in-silico** data that may **not** reflect **real-world clinical practice.**
- Regulators reviewing a company's AI data are privy to considerable data, but these data are usually **unavailable** to health system IT teams or clinicians.

A Framework for Testing, Validation and Deployment of Diagnostic Imaging in Anatomic Pathology.

An Internationally Quality Control program on Machine Learning Algorithms to Assess Quantitative Predictive/Prognostic Biomarkers in Breast Cancer such as TILs.

Radboudumc

TILS BREAST CANCER
INTERNATIONAL IMMUNO-ONCOLOGY WORKING GROUP

This helped

Current issue ›

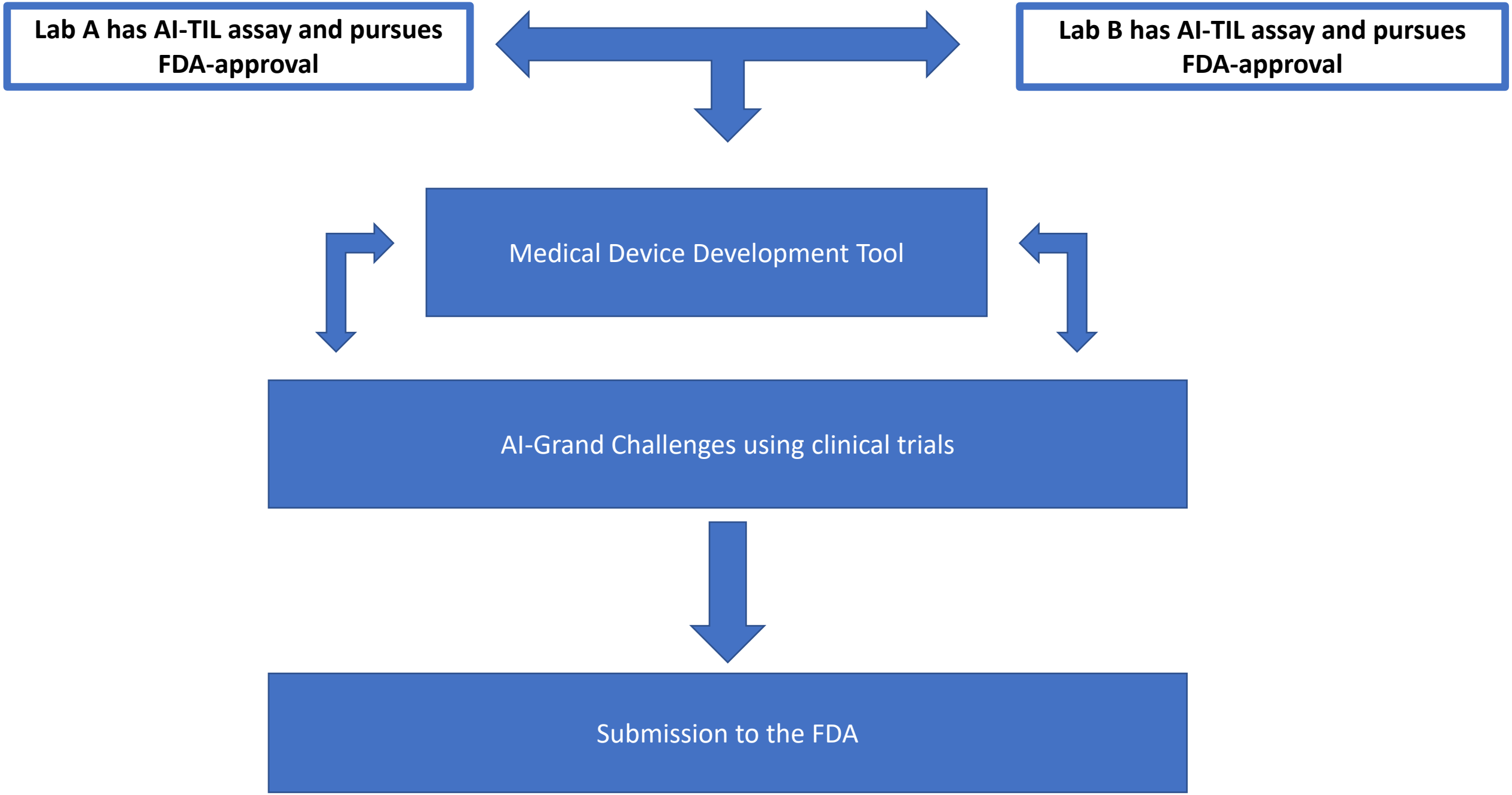
Safe driving cars

Editorial | 23 Feb 2022

FDA fosters innovative approaches in research, resources and collaboration

Brandon D. Gallas, Aldo Badano ... Ed Margerrison

Correspondence | 23 Feb 2022

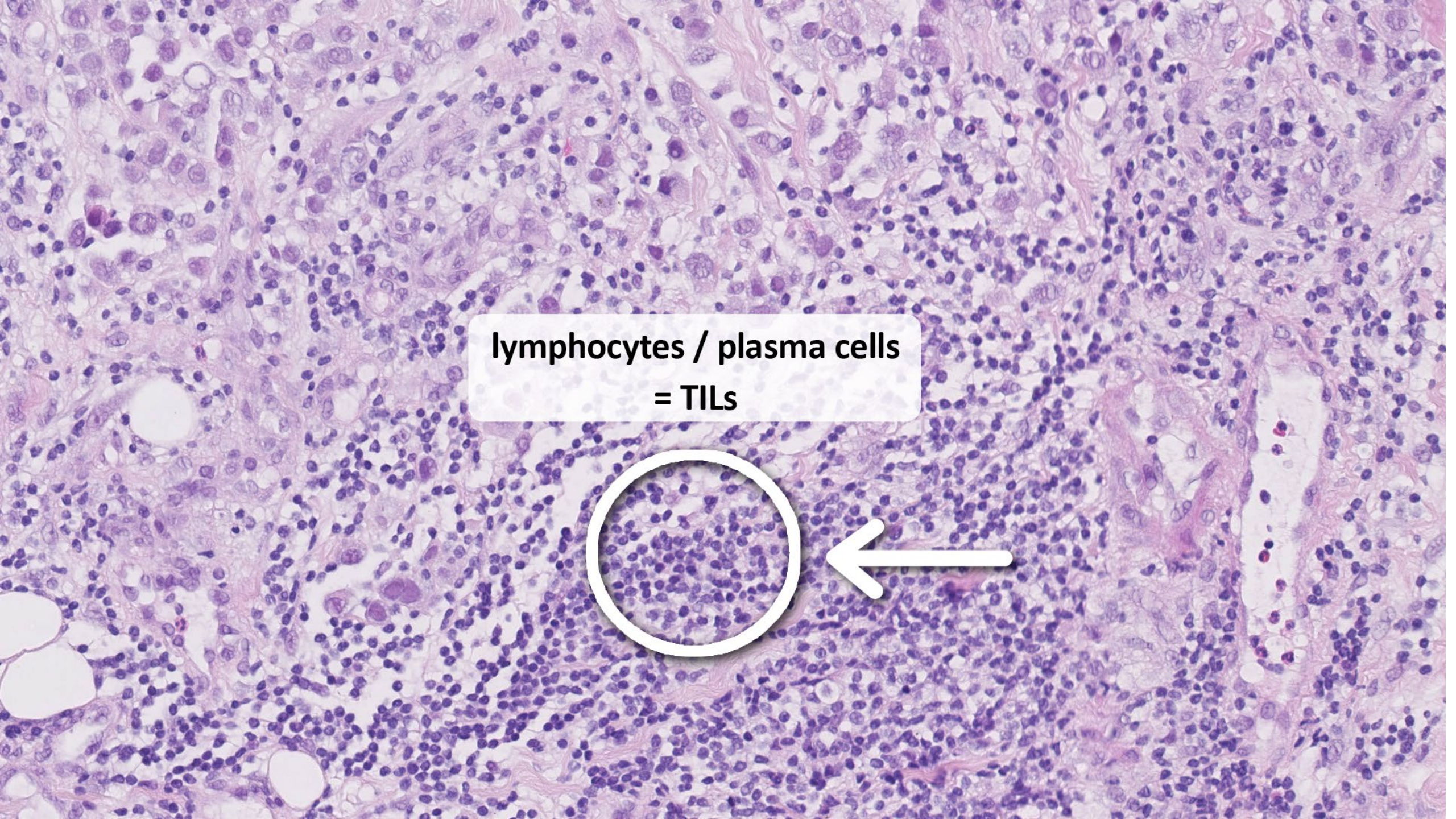

Lab A has AI-TIL assay and pursues FDA-approval
Lab B has AI-TIL assay and pursues FDA-approval
Medical Device Development Tool
AI-Grand Challenges using clinical trials
Submission to the FDA

Why the TILs?

TILs: Tumor Infiltrating Lymphocytes = immune cells

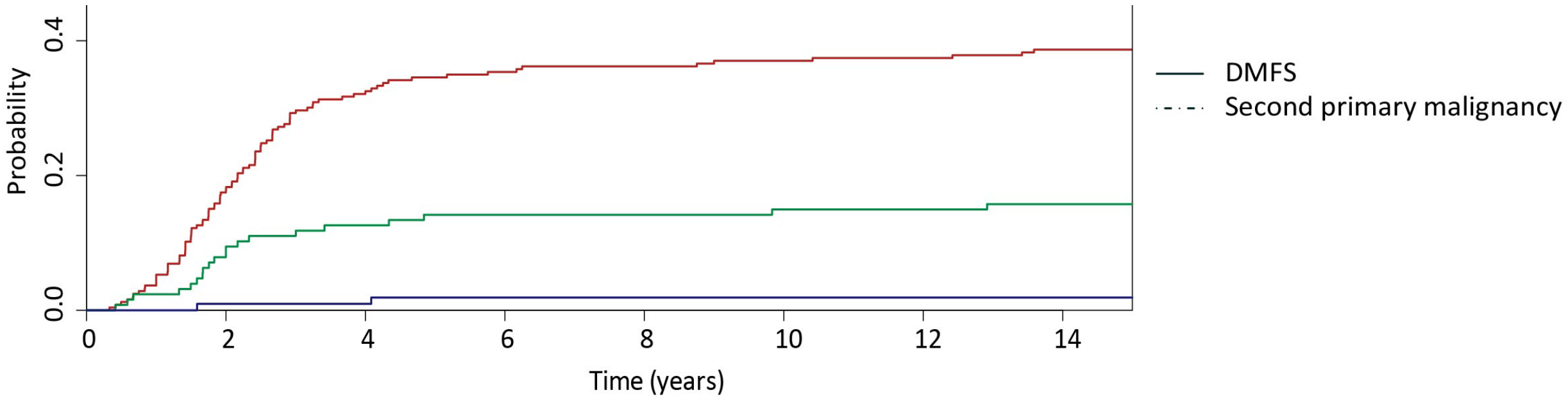
Radboudumc

TILS BREAST CANCER
INTERNATIONAL IMMUNO-ONCOLOGY WORKING GROUP


lymphocytes / plasma cells
= TILs

Low incidence of DMFS in high sTILs group

	DMFS 15 years
sTILs < 30%	39% (36-42)
sTILs 30%-75%	16% (12-19)
sTILs ≥ 75%	1.9% (0-3.4)



TILs in <40 year old TNBC, with 15-year FU, untreated; J Clin Oncol. in press

Relationship Between Tumor-Infiltrating Lymphocytes and Outcomes in the KEYNOTE-119 Study of Pembrolizumab Versus Chemotherapy for Previously Treated, Metastatic Triple-Negative Breast Cancer

S. Loi[1]; E. Winer[2]; O. Lipatov[3]; S.-A. Im[4]; A. Gonçalves[5]; J. Cortes[6]; K. S. Lee[7]; P. Schmid[8]; L. Testa[9]; I. Witzel[10]; S. Ohtani[11]; N. Turner[12]; S. Zambelli[13]; N. Harbeck[14]; F. Andre[15]; R. Dent[16]; L. Huang[17]; J. Mejia[17]; V. Karantza[17]; R. Salgado[1]

[1]Peter MacCallum Cancer Centre and University of Melbourne, Melbourne, VIC, Australia; [2]Dana-Farber Cancer Institute, Boston, MA, USA; [3]Republican Clinical Oncology Dispensary, Republic of Bashkortostan, Russian Federation; [4]Seoul National University, Seoul, South Korea; [5]Institut Paoli-Calmettes, Marseille, France; [6]IOB Institute of Oncology, Quiron Group, Madrid, and Vall d'Hebron Institute of Oncology, Barcelona, Spain; [7]National Cancer Center, Goyang, South Korea; [8]Barts Cancer Institute, Centre for Experimental Cancer Medicine, London, United Kingdom; [9]Instituto do Câncer do Estado de São Paulo – Faculdade de Medicina da Universidade do Estado do São Paulo, São Paulo, Brazil; [10]University Medical Center Hamburg, Hamburg, Germany; [11]Hiroshima City Hiroshima Citizens Hospital, Hiroshima, Japan; [12]The Royal Marsden NHS Foundation Trust, London, United Kingdom; [13]San Raffaele Scientific Institute, Milan, Italy; [14]Breast Center, University of Munich, Munich, Germany; [15]Université Paris-Sud, Orsay, and Gustave Roussy, Villejuif, France; [16]University of Toronto, Toronto, ON, Canada; [17]Merck & Co., Inc., Kenilworth, NJ, USA

Tumor marker studies
Levels of evidence

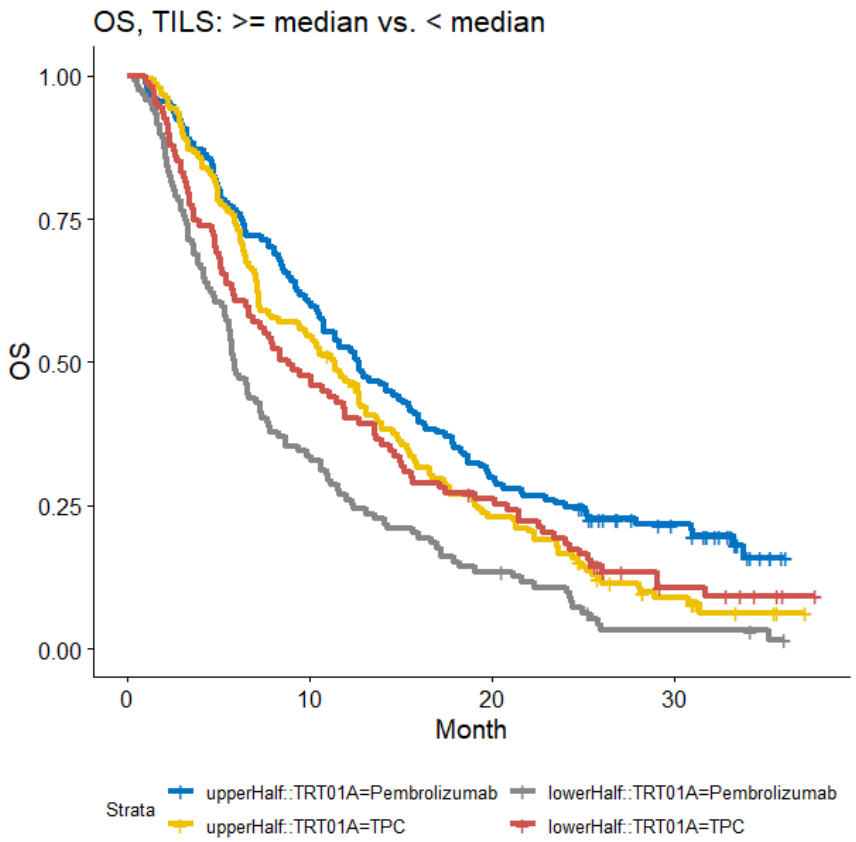
Level IA Prospective randomized controlled trial designed to address the tumor marker utility

Level IB Prospective trial not designed to address tumor marker but design accommodates tumor marker utility

⚠ For a predictive marker the trial must be a R̄ controlled trial

+ ≥ 1 validation study

- For TPC arm, the yellow and red curves represent the TILs≥5% and TILs<5%, with little difference observed
- For Pembro arm, there is separation according to the median TILS cut-off consistent with testing as a continuous measure



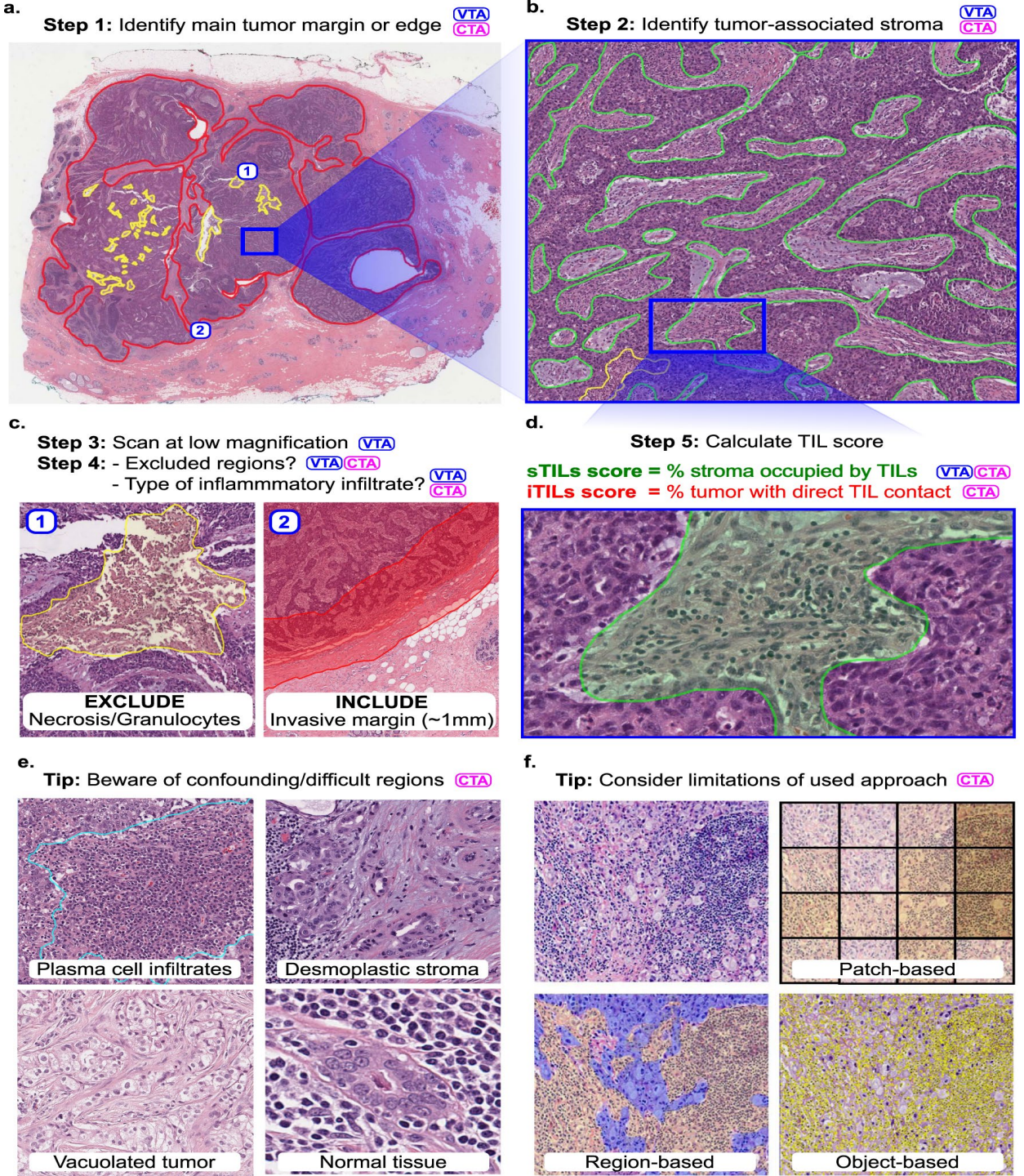
Loi et al., SABC2019

What are the potential issues?

How to score the TILs using visual and computational procedures.

Should a computational method follow the internationally accepted method that has proven clinical validity?

Is there a ground truth? The pathologists or outcome?



NPJ Breast Cancer. 2020 May 12;6:16. doi: 10.1038/s41523-020-0154-2.

Pathologists score TILs as a percentage.

Should a computational pathology method also assess TILs as a %?

Patient Name / ID: DOE, Jane / AQH12CR3-DX-2 21/05/2020 03:22 PM
Gender: Female **Age:** 46 **Dx:** Breast carcinoma, right, primary; Stage IB **Tx:** Not initiated , No NACT
Histology: Invasive ductal carcinoma / NST; Grade 3 **Stain:** H&E, FFPE **Other Markers:** TN (ER-, PR-, Her2-); Ki67 < 25%

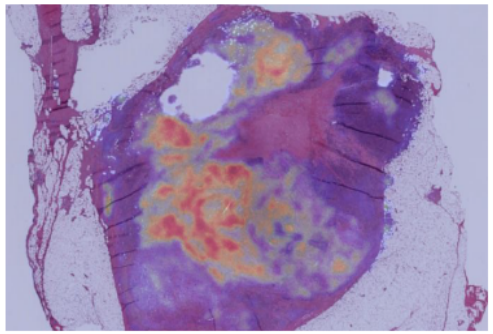
	Global density: Whole-slide score	Local density: 50 μm x 50 μm fields	Local density: 100 μm x 100 μm fields	Local density: 200 μm x 200 μm fields
Stromal TILs	**40.3 %**	54.2 (± 20.1) %	52.1 (± 7.4) %	41.2 (± 5.1) %
Intra-tumoral TILs	**5.6 %**	0.1 (± 3.1) %	2.5 (± 2.1) %	4.9 (± 1.1) %
Invasive margin TILs	**7.8 %**	3.7 (± 4.1) %	6.2 (± 2.6) %	8.2 (± 0.8) %

Tissue delineation confidence: 0.95
TIL classification confidence: 0.86

TIL heatmap: See right; refer to WSI display for detailed tissue delineation, TIL classification, and zoomable heatmap.

Distance from stromal TIL to nearest tumor: 62.1 (± 23.7) μm
Distance from tumor to nearest TIL: 726.9 (± 13.5) μm
Number of TIL clusters per unit area: 1.3 / mm^2
TIL cluster morphology: Brisk, diffuse - moderate heterogeneity
TIL cluster size: 320 (± 129) μm

Multivariable PFS prob.: 0.87 (1 yr) - 0.76 (3 yrs) - 0.67 (5 yrs) - 0.61 (10 yrs)



On visual inspection, what is the quality of computational tissue delineation (tumor, stroma, etc) (circle one):

Very Poor	Poor	Acceptable	Very good	Excellent

On visual inspection, what is the quality of computational TIL localization (circle one):

Very Poor	Poor	Acceptable	Very good	Excellent

Pathologist Comments & Recommendations:

None. Refer to pathology report for detailed histologic comment.

Attending pathologist

NPJ Breast Cancer. 2020 May 12;6:16. doi: 10.1038/s41523-020-0154-2.

Why Clinical Trials?

An AI-assay is an assay like all other assays and the same principles apply.

Evidence category	Definition
Analytical validity	Demonstration that the performance characteristics of the biomarker-based test are acceptable in terms of its sensitivity, specificity, accuracy, precision, and other relevant performance characteristics using a specified technical protocol (which may include specimen collection, handling and storage procedures).
Clinical validity	Demonstration that the biomarker-based test acceptably identifies, measures, or predicts the concept of interest, where “concept” refers to a clinical, biological, physical, or functional state, or experience.
Clinical utility	Demonstration that use of the biomarker-based test will lead to a net improvement in health outcome or provide useful information about diagnosis, treatment, management, or prevention of a disease. Clinical utility includes the range of possible benefits or risks to individuals and populations.

Use of Archived Specimens in Evaluation of Prognostic and Predictive Biomarkers

Richard M. Simon, Soonmyung Paik, Daniel F. Hayes

The development of tumor biomarkers ready for clinical use is complex. We propose a refined system for biomarker study design, conduct, analysis, and evaluation that incorporates a hierarchal level of evidence scale for tumor marker studies, including those using archived specimens. Although fully prospective randomized clinical trials to evaluate the medical utility of a prognostic or predictive biomarker are the gold standard, such trials are costly, so we discuss more efficient indirect "prospective–retrospective" designs using archived specimens. In particular, we propose new guidelines that stipulate that 1) adequate amounts of archived tissue must be available from enough patients from a prospective trial (which for predictive factors should generally be a randomized design) for analyses to have adequate statistical power and for the patients included in the evaluation to be clearly representative of the patients in the trial; 2) the test should be analytically and preanalytically validated for use with archived tissue; 3) the plan for biomarker evaluation should be completely specified in writing before the performance of biomarker assays on archived tissue and should be focused on evaluation of a single completely defined classifier; and 4) the results from archived specimens should be validated using specimens from one or more similar, but separate, studies.

J Natl Cancer Inst 2009;101:1446–1452

Table 1. Elements of tumor marker studies that constitute Levels of Evidence determination*

Category Element	A Prospective	B Prospective using archived samples	C Prospective/ observational	D Retrospective/ observational
Clinical trial	PCT designed to address tumor marker	Prospective trial not designed to address tumor marker, but design accommodates tumor marker utility Accommodation of predictive marker requires PRCT	Prospective observational registry, treatment and follow-up not dictated	No prospective aspect to study
Patients and patient data	Prospectively enrolled, treated, and followed in PCT	Prospectively enrolled, treated, and followed in clinical trial and especially if a predictive utility is considered, a PRCT addressing the treatment of interest	Prospectively enrolled in registry, but treatment and follow-up standard of care	No prospective stipulation of treatment or follow-up; patient [illegible] ted by retrospective [illegible] review
Specimen collection, processing, and archival	Specimens collected, processed, and assayed for specific marker in real time	Specimens collected, processed, and archived prospectively using generic SOPs. Assayed after trial completion	Specimens collected, processed, and archived prospectively using generic SOPs. Assayed after trial completion	Specimens collected, processed and archived with no prospective SOPs
Statistical design and analysis	Study powered to address tumor marker question	Study powered to address therapeutic question and underpowered to address tumor marker question	Study not prospectively powered at all. Retrospective study design confounded by selection of specimens for study	Study not prospectively powered at all. Retrospective study design confounded by selection of specimens for study
		Focused analysis plan for marker question developed before doing assays	Focused analysis plan for marker question developed before doing assays	No focused analysis plan for marker question developed before doing assays
Validation	Result unlikely to be play of chance	[illegible] more likely to be play of chance that A but less likely than C	[illegible] play of chance	Result very likely to be play of chance
	Although preferred, validation not required	Requires one or more validation studies	Requires subsequent validation studies	Requires subsequent validation

* PCT = prospective controlled trial; PRCT = prospective randomized controlled trial; SOPs = standard operating practices.

About Grand Challenges

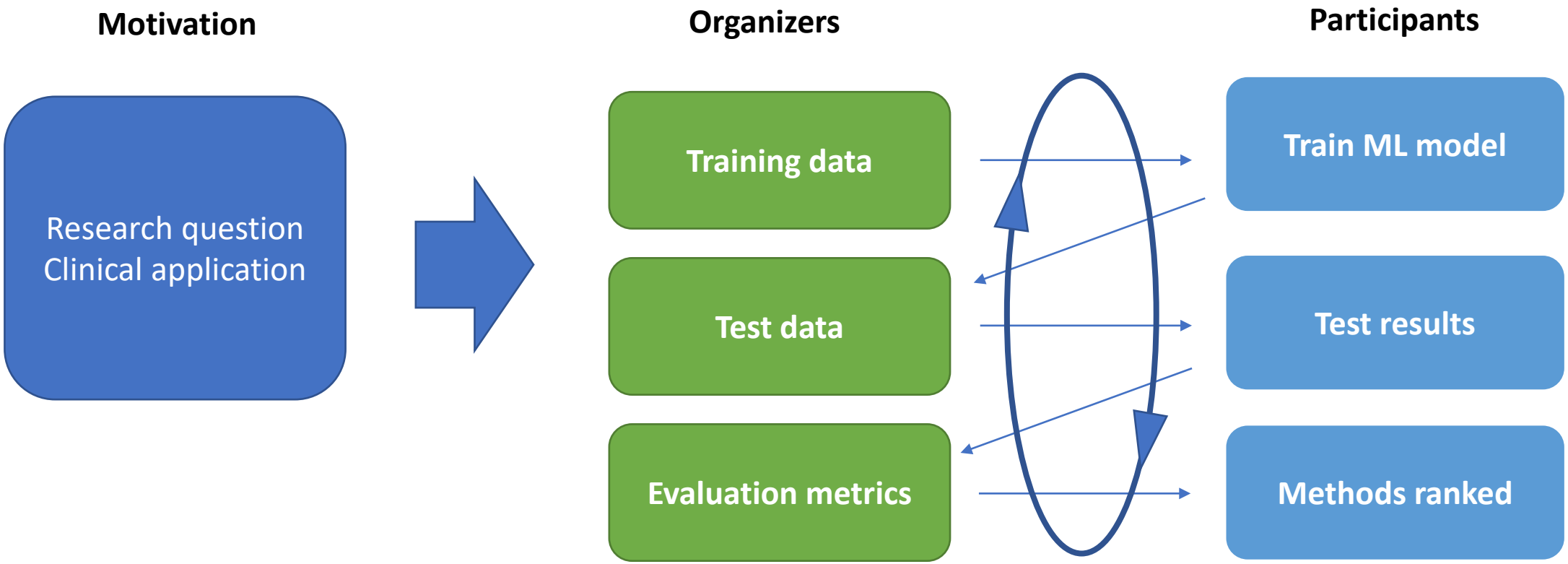
Radboudumc

TILS BREAST CANCER
INTERNATIONAL IMMUNO-ONCOLOGY WORKING GROUP

What is a challenge?

Radboudumc

TILS BREAST CANCER
INTERNATIONAL IMMUNO-ONCOLOGY WORKING GROUP



Typical pattern of a type-1 challenge

Why a challenge?

Radboudumc

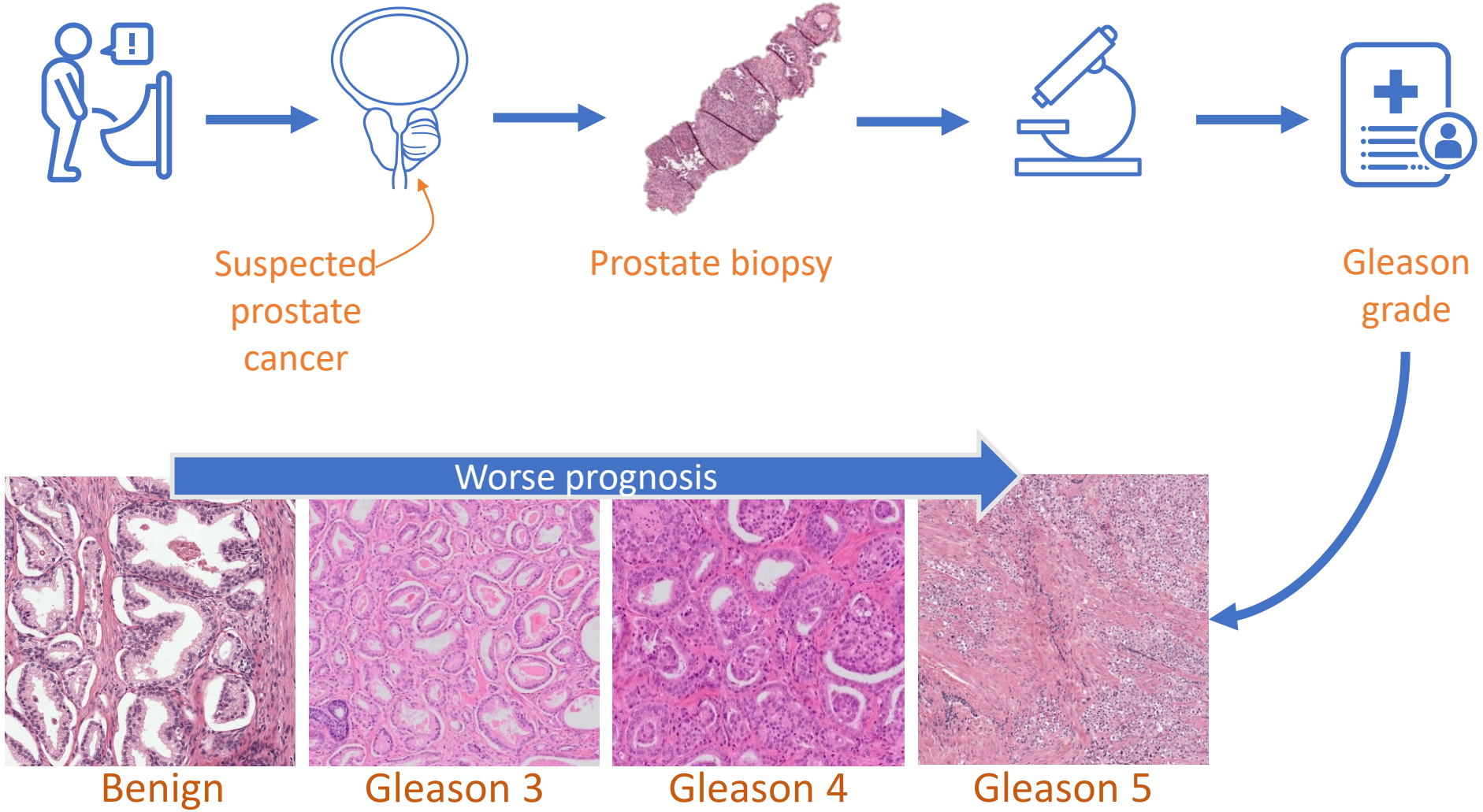
BCRF BREAST CANCER RESEARCH FOUNDATION | TILS BREAST CANCER INTERNATIONAL IMMUNO-ONCOLOGY WORKING GROUP

- **Many papers** are published every year presenting and validating a **"new" algorithm** for solving a particular task in medical image analysis
- For many tasks, **multiple algorithms** are presented
- Obvious question: **which one works best?**
- Hard to say because they are typically tested on separate, locally collected, data sets
- Code is typically not shared
- Data sets are typically not shared
  - This may change slowly because of the demand for open and reproducible science and FAIR data

Challenges can solve this issue because they offer **fair comparison** of **algorithms on the same data**

Slide adapted from a presentation of Bram van Ginneken

The PANDA challenge



Nature Medicine, 2022; Slide courtesy of Wouter Bulten

The PANDA challenge

Radboudumc

BCRF BREAST CANCER RESEARCH FOUNDATION

TILS BREAST CANCER INTERNATIONAL IMMUNO-ONCOLOGY WORKING GROUP

Training data:

10.000 biopsies from Radboud and Karolinska

Test data:

2.000 biopsies with consensus reference standard (internal and external)

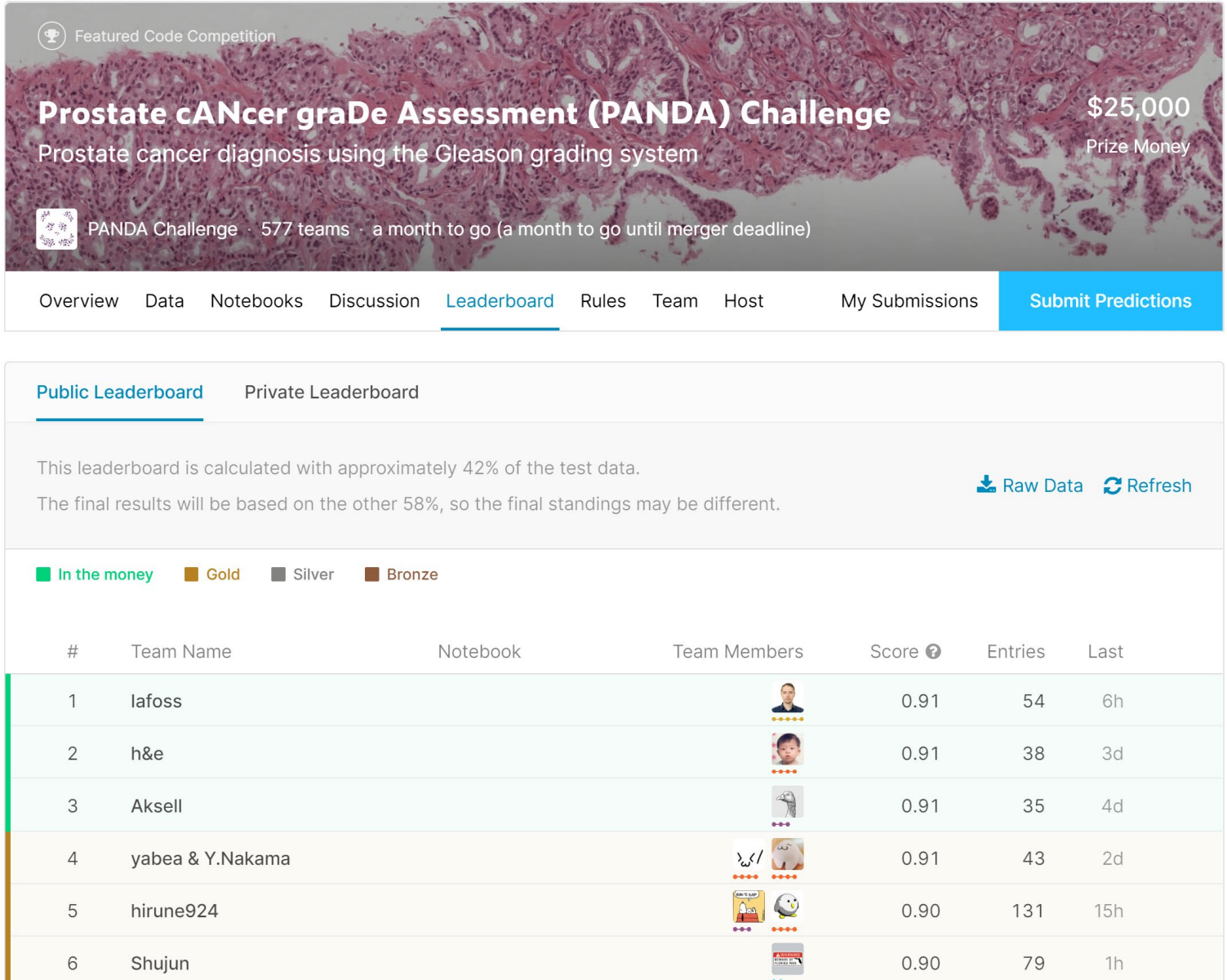
Radboudumc university medical center

Karolinska Institutet

Google Health

kaggle

Nature Medicine, 2022; Slide courtesy of Wouter Bulten



The PANDA challenge

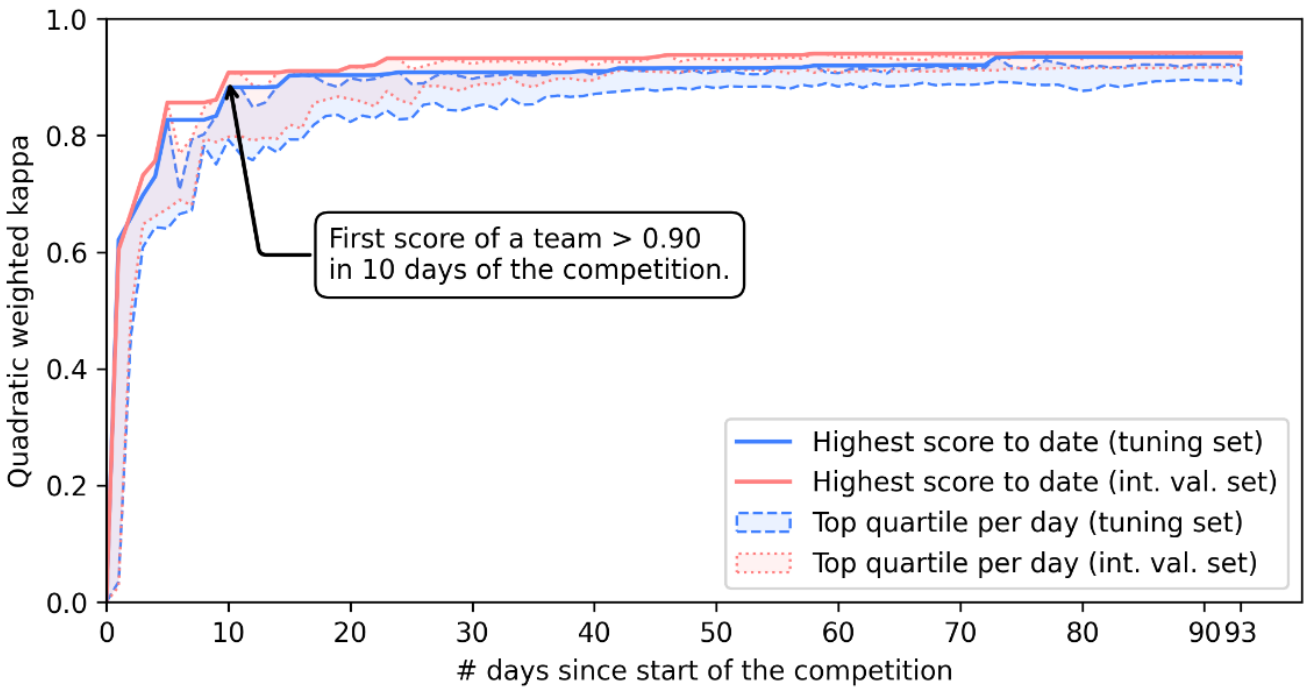
Radboudumc

BCRF BREAST CANCER RESEARCH FOUNDATION

TILS BREAST CANCER INTERNATIONAL IMMUNO-ONCOLOGY WORKING GROUP

Quick progression of solutions

Most of the performance was achieved at the start of the challenge

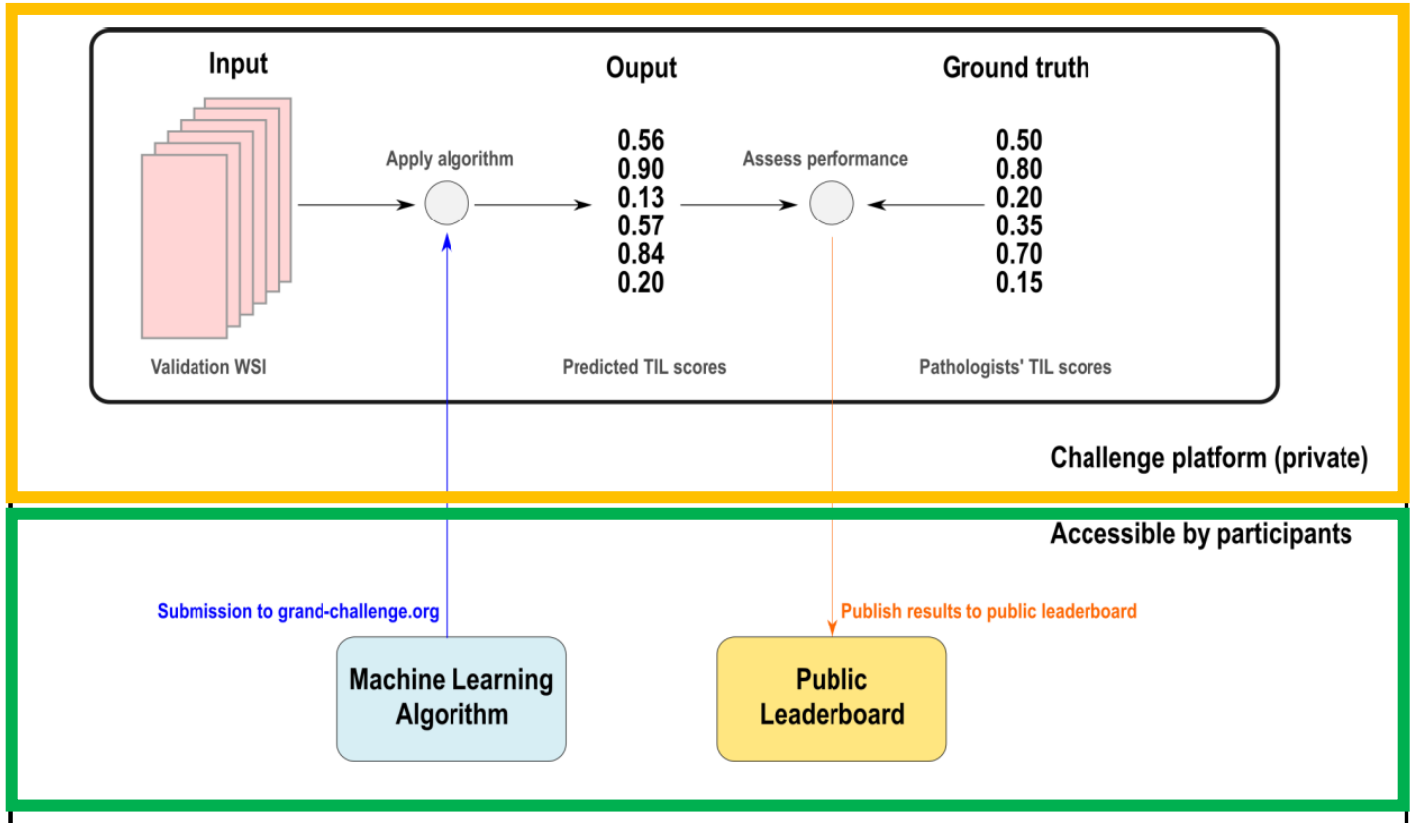


Nature Medicine, 2022; Slide courtesy of Wouter Bulten

Type-2 challenges

A Framework for Testing, Validation and Deployment of Diagnostic Imaging in Anatomic Pathology.

An Internationally Quality Control program on Machine Learning Algorithms to Assess Quantitative Predictive/Prognostic Biomarkers in Breast Cancer such as TILs.



2019-present

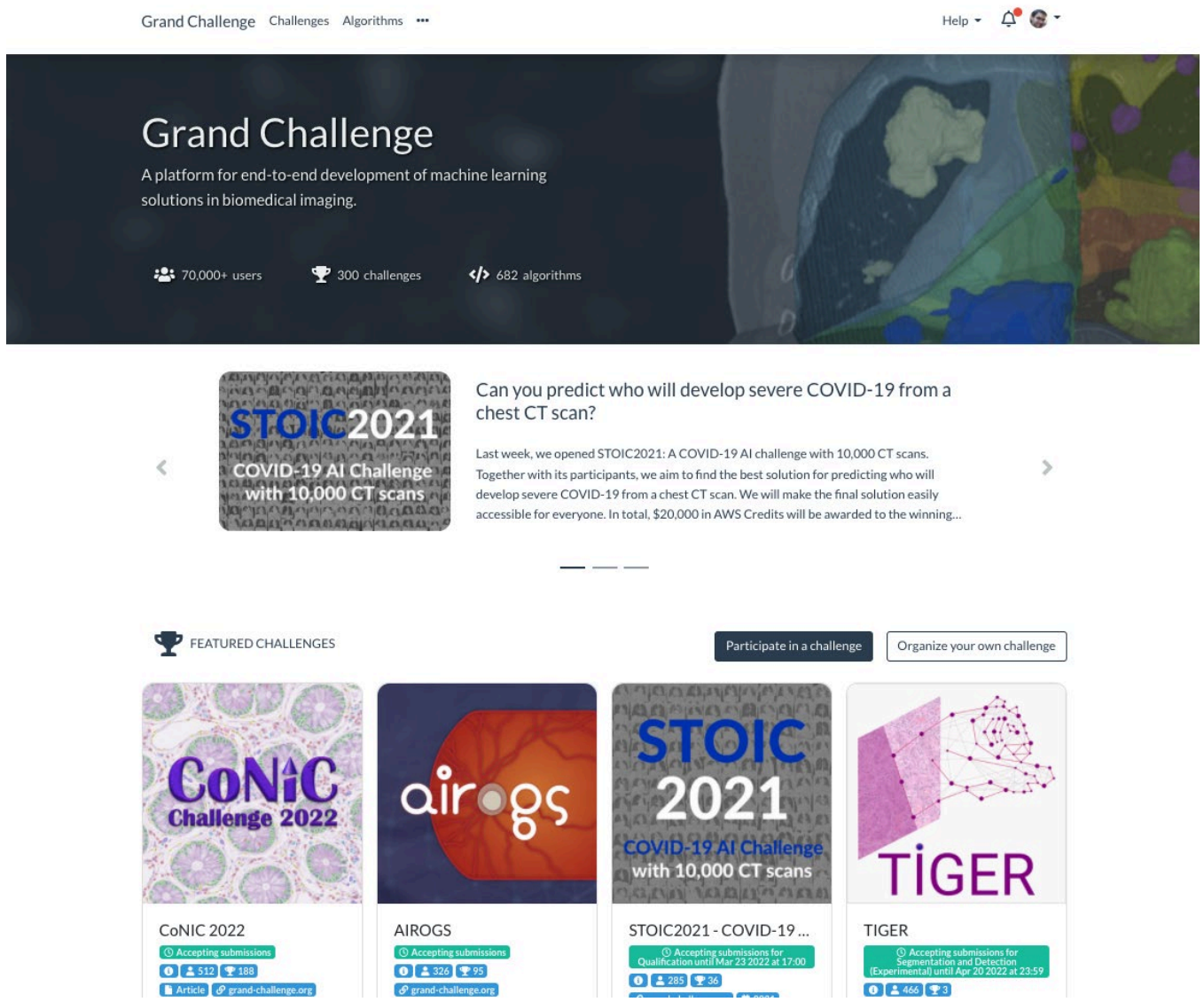
Joint effort:

grand-challenge.org

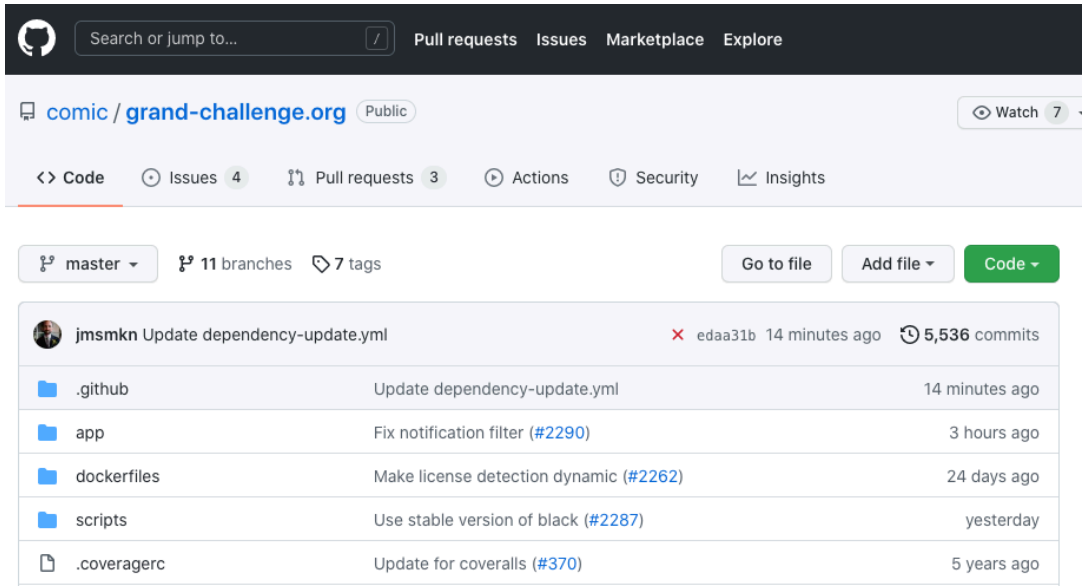
Radboudumc

BCRF BREAST CANCER RESEARCH FOUNDATION

TILS BREAST CANCER INTERNATIONAL IMMUNO-ONCOLOGY WORKING GROUP



- web-based
- open-source
- 70,000+ users
- 300 challenges
- 682 algorithms
- Archives
- Reader studies
- Web-based viewers + annotations



tiger.grand-challenge.org

Radboudumc

BCRF BREAST CANCER RESEARCH FOUNDATION

TILS BREAST CANCER INTERNATIONAL IMMUNO-ONCOLOGY WORKING GROUP

TiGER

Tumor **I**nfiltratin**G** lymphocytes in breast canc**ER**

Info | Forum | Teams | Submit | Leaderboards | Admin | Join

- Home
- Contact
- Videos
- Data
- Code
- Rules
- Evaluation
- Timeline
- Prizes

Welcome to TIGER

TIGER is the first challenge on fully automated assessment of tumor-infiltrating lymphocytes (TILs) in H&E breast cancer slides. It is organized by the Diagnostic Image Analysis Group (DIAG) of the Radboud University Medical Center (Radboudumc) in Nijmegen (The Netherlands), in close collaboration with the International Immuno-Oncology Biomarker working Group (www.tilsinbreastcancer.org).

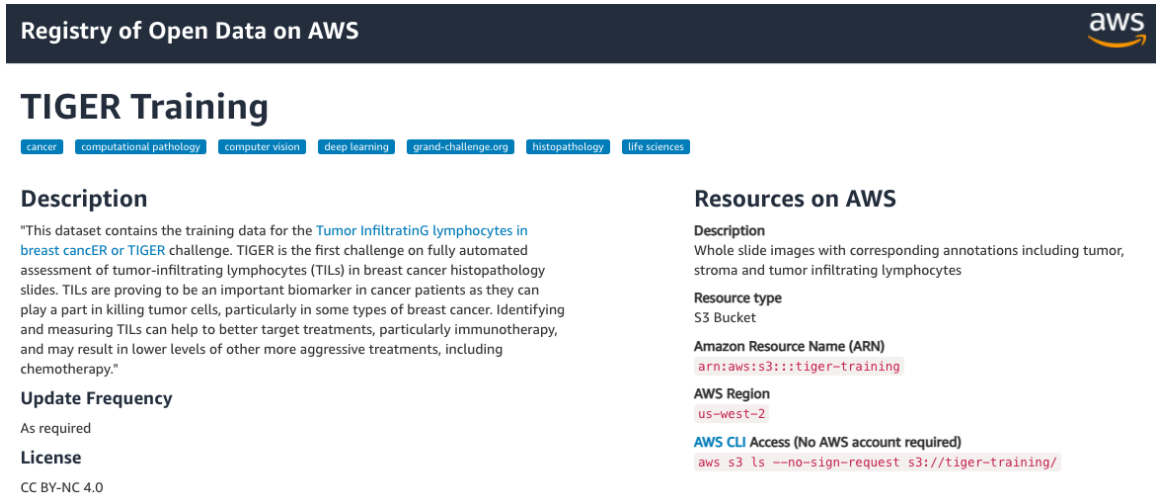
The goal of this challenge is to evaluate new computer algorithms for the automated assessment of tumor-infiltrating lymphocytes (TILs) in Her2 positive and Triple Negative breast cancer (BC) histopathology slides. In recent years, several studies have shown the predictive and prognostic value of visually scored TILs in BC as well as in other cancer types, making TILs a powerful biomarker that can potentially be used in the clinic. With TIGER, we aim at developing computer

Publicly available training data

Triple-negative breast cancer and Her2+ BC

Training (390 WSI, 1800 ROIs)

- Manual annotations of tissue and TILs
- Visual TILs scores
- Manual annotations of “tumor bulk”
- Publicly available under CC BY-NC 4.0



Registry of Open Data on AWS

TIGER Training

cancer | computational pathology | computer vision | deep learning | grand-challenge.org | histopathology | life sciences

Description

"This dataset contains the training data for the Tumor InfiltratinG lymphocytes in breast cancER or TIGER challenge. TIGER is the first challenge on fully automated assessment of tumor-infiltrating lymphocytes (TILs) in breast cancer histopathology slides. TILs are proving to be an important biomarker in cancer patients as they can play a part in killing tumor cells, particularly in some types of breast cancer. Identifying and measuring TILs can help to better target treatments, particularly immunotherapy, and may result in lower levels of other more aggressive treatments, including chemotherapy."

Update Frequency

As required

License

CC BY-NC 4.0

Resources on AWS

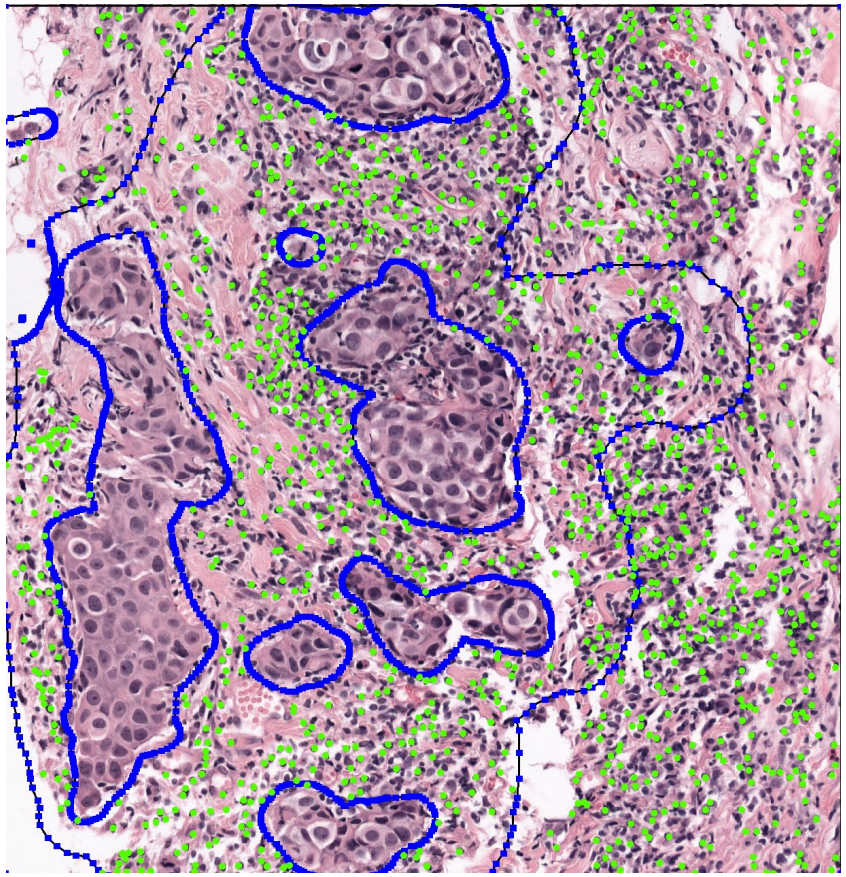
Description
Whole slide images with corresponding annotations including tumor, stroma and tumor infiltrating lymphocytes

Resource type
S3 Bucket

Amazon Resource Name (ARN)
`arn:aws:s3:::tiger-training`

AWS Region
`us-west-2`

AWS CLI Access (No AWS account required)
`aws s3 ls --no-sign-request s3://tiger-training/`



https://registry.opendata.aws/tiger/

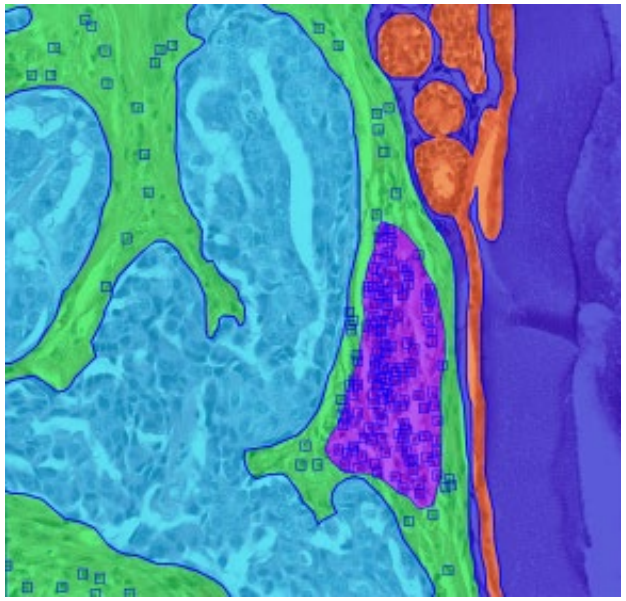
Training data: how did we build it?

Radboudumc

BCRF BREAST CANCER RESEARCH FOUNDATION

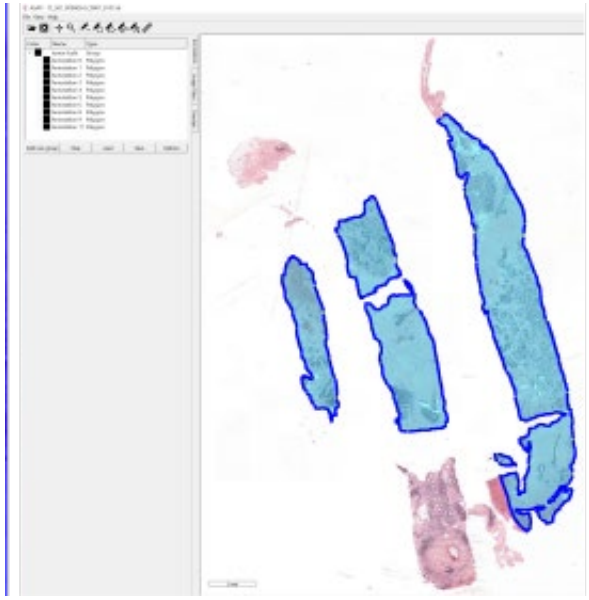
TILS BREAST CANCER INTERNATIONAL IMMUNO-ONCOLOGY WORKING GROUP

WSIROIS



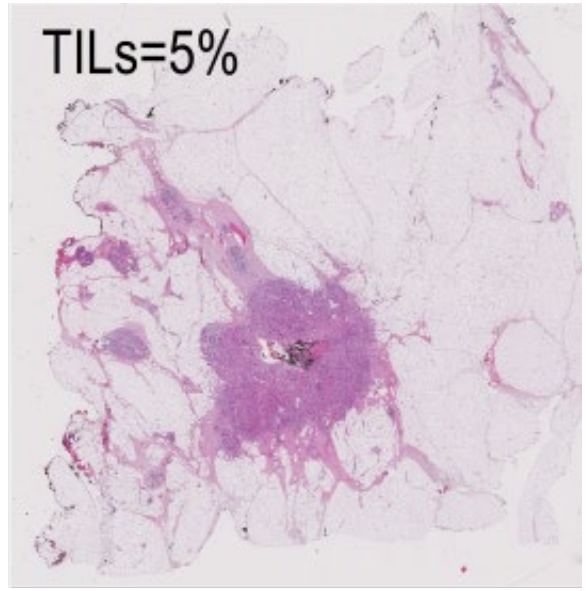
- 5 breast pathologists
- Web-based annotations via GC
- 3 pre-selected ROIs/slide
- Tissue and TILs annotated
- Independent annotations
- Consensus for uncertain annotations
- Merging with BCSS and NuCLS projects

WSIBULK



- 3 resident pathologists
- Web-based annotations via GC
- Coarse annotations of tumor regions
- Intersect with AI-based tissue mask

WSITILS



- 1 pathologist
- Web-based annotations via GC
- Single TILs score per slide
- Comments on potential pitfalls

Evaluation

Radboudumc

BCRF BREAST CANCER RESEARCH FOUNDATION

TILS BREAST CANCER INTERNATIONAL IMMUNO-ONCOLOGY WORKING GROUP

We will have two leaderboards, to assess:

1. <u>"Computer vision performance"</u>
   - Tissue segmentation (Dice stroma segm. and Dice tumor segm.)
   - Lymphocyte detection (FROC analysis)
   - Algorithms ranked on a combination of these performance
   - **Test data: 64 WSIs with 279 manually annotated ROIs**

2. <u>"Prognostic value"</u>
   - Prediction of cancer recurrence
   - Concordance index of multivariate Cox regression model
   - **Test data: 907 cases from phase-3 trial and clinical practice**

Test sets and evaluation

	Leaderboard 1	Leaderboard 2	Challenge phase
Experimental test set	• 26 WSIs • 130 ROIs	• 200 WSIs • 200 patients	• during TIGER • multiple runs
Final test set	• 38 WSIs • 149 ROIs	• 707 WSIs • 707 patients	• At the end of TIGER • one run

Goal of TIGER

Radboudumc
BCRF BREAST CANCER RESEARCH FOUNDATION
TILS BREAST CANCER INTERNATIONAL IMMUNO-ONCOLOGY WORKING GROUP

- Develop open-source AI algorithms for automated TILs assessment
  - Source code of awarded algorithms will be released.
  - AWS-award of 13K US in credits.
- Boost research and development on AI for automated TILs assessment
  - Training data publicly released under CC license
- Validate developed algorithms in a fair using a secure platform
  - Platform remains open for future benchmarking
- Identify top algorithms for future research
  - Algorithms on grand-challenge as base for potential collaborations
  - Correlation between AI and pathologists
  - Role of automated TILs in prognosis and treatment response

Code and processing pipelines

Radboudumc

BCRF BREAST CANCER RESEARCH FOUNDATION

TILS BREAST CANCER INTERNATIONAL IMMUNO-ONCOLOGY WORKING GROUP

DIAGNijmegen / **pathology-whole-slide-data** Public

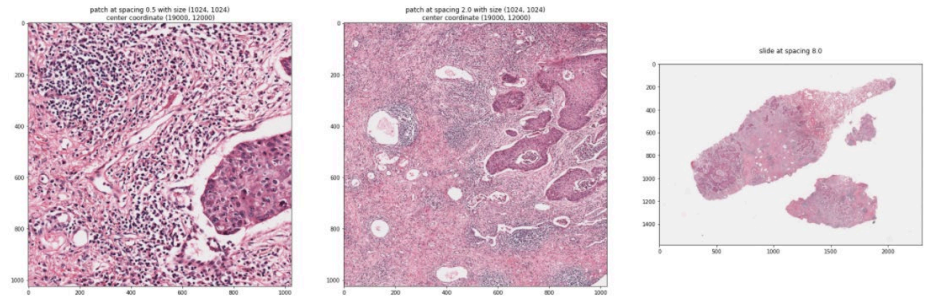
Pin | Watch

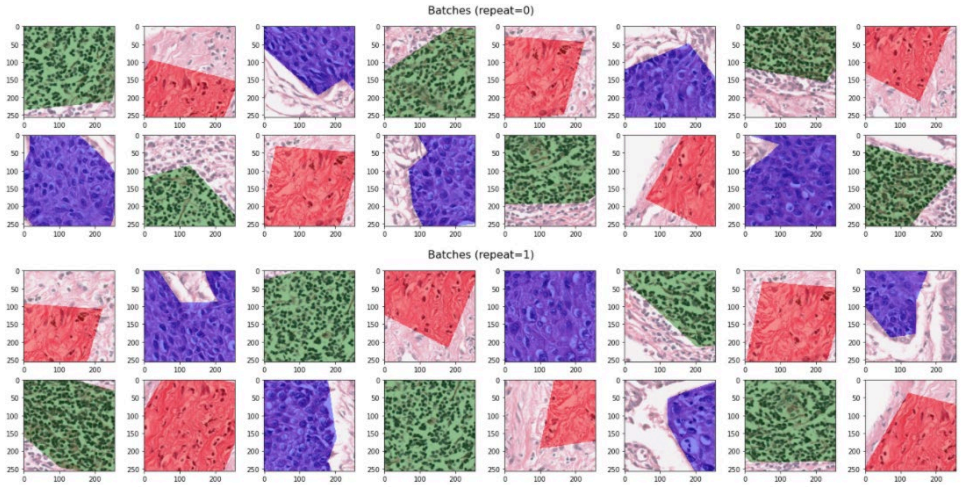
Code | Issues 4 | Pull requests | Actions | Projects | Wiki | Security | Insights

main | 2 branches | 0 tags | Go to file | Add file | Code

martvanrijthoven write_point_set2 — d967849 yesterday — 194 commits

.github/workflows	installatiosn for docs	5 months ago
docs	docstrings, typing and maskparsing optimization	4 months ago
notebooks	hooknet tiger example	last month
tests	Finished AlbumentationsCallback. See test for specifications how to ...	12 days ago
tutorials	Update readme.md	23 days ago
wholeslidedata	write_point_set2	yesterday
.gitignore	Ignore .idea and tif files	12 days ago

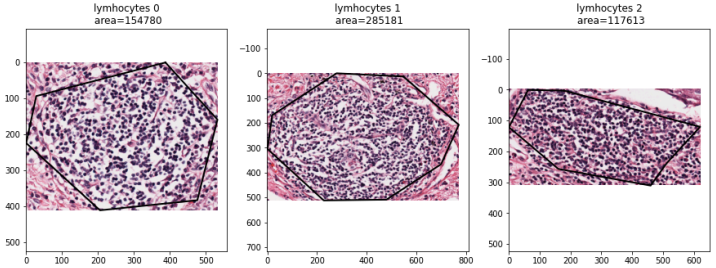




pip install wholeslidedata

```
In [9]: WholeSlideImage('/home/mart/Radboudumc/data/breast/AQ_S02_P000174_C0001_L03_A01.tif', backend='asap').spacings

Out[9]: [0.2430938093312698,
         0.4861876186625396,
         0.9723752373250792,
         1.9447504746501585,
         3.8898287602042423,
         7.780967836694896,
         15.567182502511177,
         31.15537457741085,
         62.394950654330664]
```



Code and processing pipelines

DIAGNijmegen / pathology-tiger-algorithm-example Public

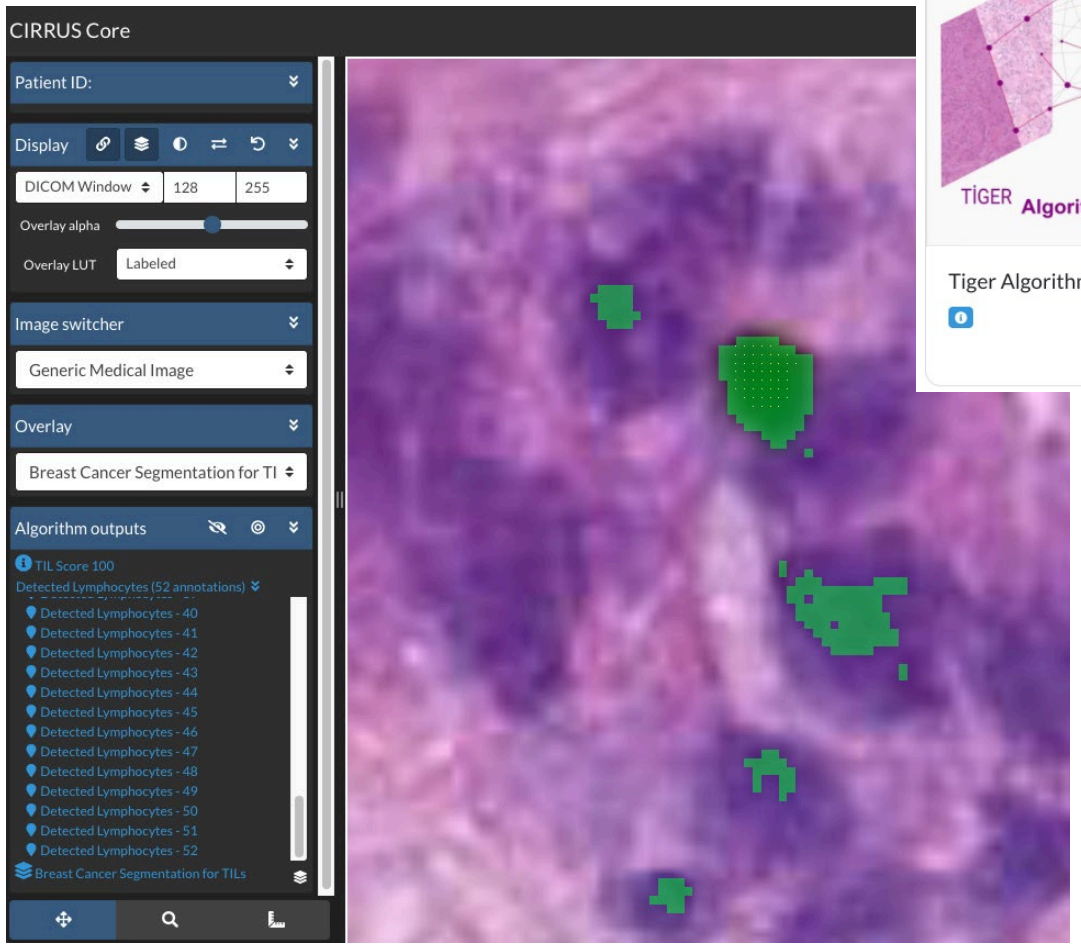
Pin | Unwatch

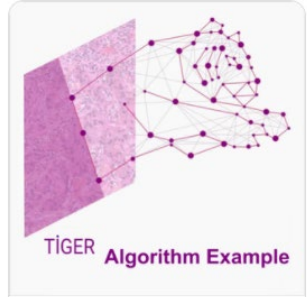
Code | cast tile to uint8 | Issues | Pull requests | Projects | Wiki | Security | Insights | Settings

main | 1 branch | 0 tags | Go to file | Add file | Code

martvanrijthoven cast tile to uint8 b84531d 2 days ago 26 commits

testinput	change test to testinput	last month
tigeralgorithmexample	cast tile to uint8	2 days ago
.gitignore	add test script	last month
Dockerfile	add test script	last month
LICENSE	Initial commit	2 months ago
README.md	Update README.md	last month
Tiger - algorithm example.png	add tiger png	last month





Tiger Algorithm Example

THE CATALINA CHALLENGE (CollAborative Til vALidatIoN chAllenge)

original report

Tumor-Infiltrating Lymphocytes and Prognosis: A Pooled Individual Patient Analysis of Early-Stage Triple-Negative Breast Cancers

Sherene Loi, MD[1]; Damien Drubay, PhD[2,3]; Sylvia Adams, MD[4]; Giancarlo Pruneri, MD[5]; Prudence A. Francis, MD[1]; Magali Lacroix-Triki, MD[2]; Heikki Joensuu, MD[7]; Maria Vittoria Dieci, MD[8,9]; Sunil Badve, MD[10]; Sandra Demaria, MD[11]; Robert Gray, PhD[12]; Elisabetta Munzone, MD[13]; Jerome Lemonnier, PhD[6]; Christos Sotiriou, MD[14]; Martine J. Piccart, MD[14]; Pirkko-Liisa Kellokumpu-Lehtinen, MD[15]; Andrea Vingiani, MD[16]; Kathryn Gray, PhD[12]; Fabrice Andre, MD[2,3]; Carsten Denkert, MD[17]; Roberto Salgado, MD[1,18]; and Stefan Michiels, PhD[2,3]

PMID: 30650045 PMCID: PMC7010425 DOI: 10.1200/JCO.18.01010

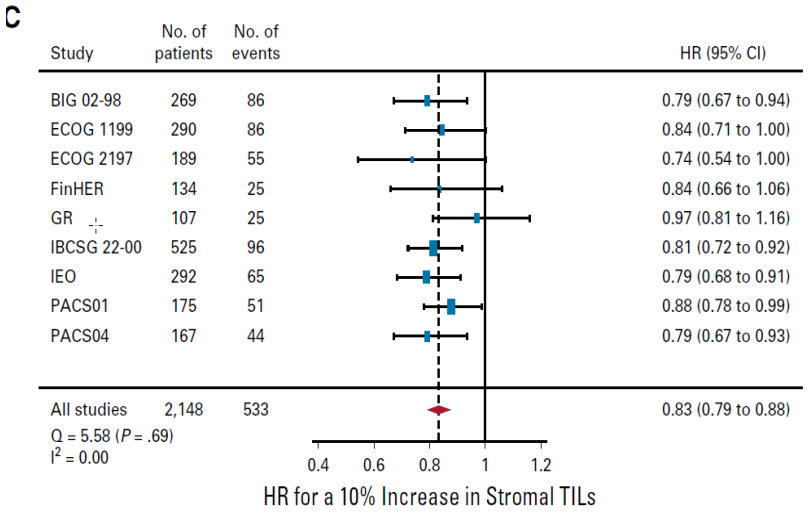
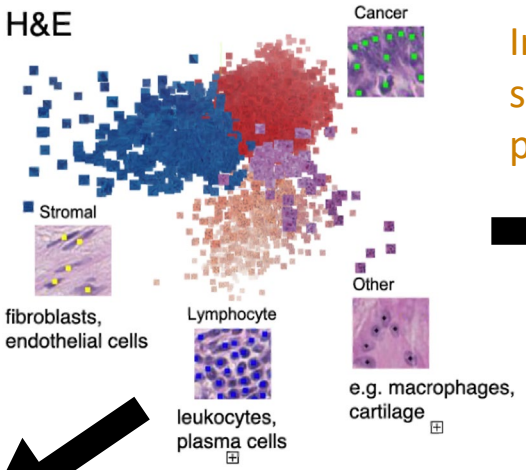


Table 1. Comparison of intraclass correlation coefficient and pair-wise observer concordance rate for 3 ring studies.

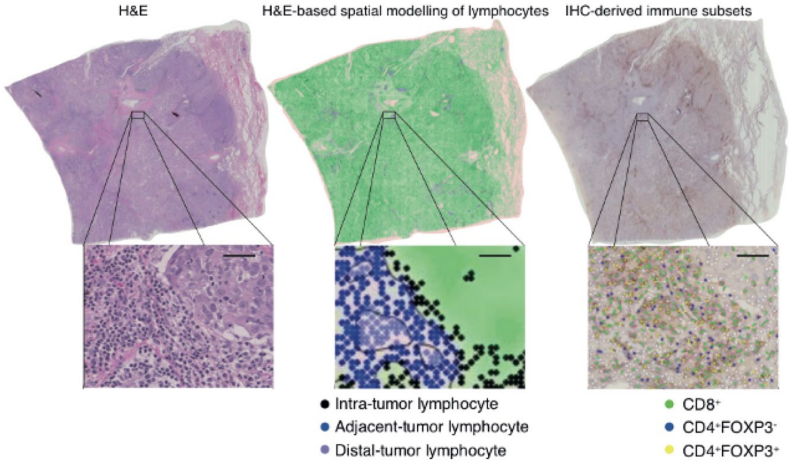
	Ring study 1	Ring study 2	Ring study 3
ICC	0.7 (0.62–0.78)	0.89 (0.85–0.92)	0.76 (0.69–0.83)
Concordance rates[a]			
TILs <1 vs ≥1%	0.94 (±0.08)	0.94 (±0.04)	0.91 (±0.06)
TILs <5 vs ≥5%	0.83 (±0.09)	0.89 (±0.05)	0.84 (±0.1)
TILs <10 vs ≥10%	0.77 (±0.08)	0.86 (±0.05)	0.79 (±0.06)
TILs <30 vs ≥30%	0.81 (±0.08)	0.93 (±0.03)	0.87 (±0.04)
TILs <75 vs ≥75%	0.90 (±0.06)	0.92 (±0.03)	0.94 (±0.03)

ICC intraclass correlation coefficient, TILs tumor-infiltrating lymphocytes.
[a]The concordance of all pairs of pathologists was calculated for five different TIL-groups. The values in the table are the sample mean and sample standard deviation of these concordance rates for all pairs of pathologists in each study.

npj Breast Cancer (2020)6:17 ; https://doi.org/10.1038/s41523-020-0156-0



Immune spatial profiling



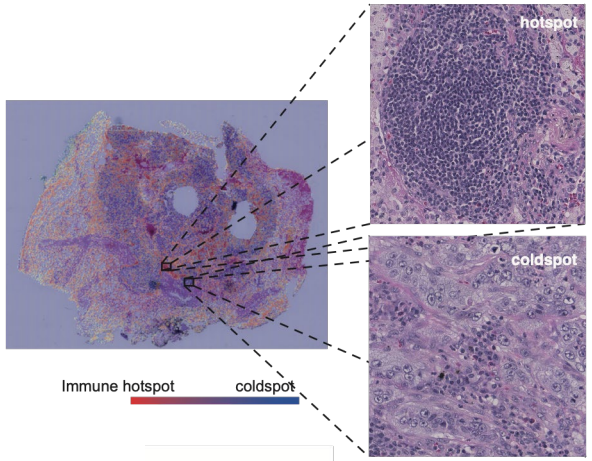
2 **ATL/stromal cell ratio**: proxy of pathology sTIL [AbdulJabbar 2020]

3 **ITLR**: intra lymphocyte cells/ tumor cell ratio [Yuan 2015]

1

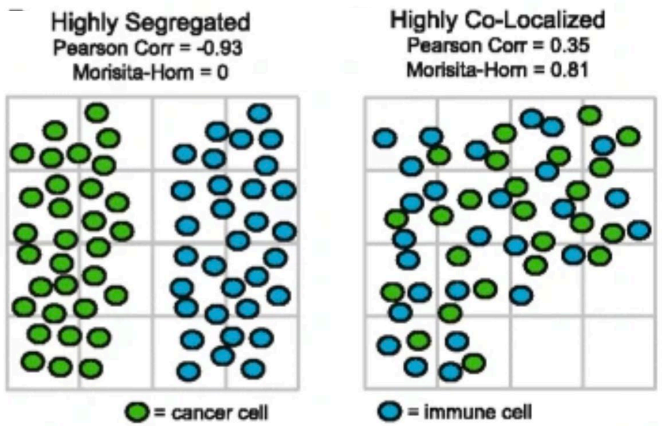
%Lymphocyte: baseline and no spatial info.
*identified lymphocytes/(sum of all identified cells) * 100*
[AbdulJabbar 2020, and several publications]

Immune clustering



4 **Fraction of immune hotspot**: [Nawaz 2015]

Immune co-localisation



5 **Immune-cancer spatial co-localisation**: Morisita index [Maley 2015]

TIL-ML project

Immune metrics generated using the single-cell identification pipeline

Why a risk-assessment strategy for biomarker-assesment using AI in clinical trials?

PERSONAL VIEW | VOLUME 15, ISSUE 4, E184-E193, APRIL 01, 2014

A risk-management approach for effective integration of biomarkers in clinical trials: perspectives of an NCI, NCRI, and EORTC working group

Dr Jacqueline Anne Hall, PhD • Roberto Salgado, MD • Tracy Lively, PhD • Prof Fred Sweep, PhD • Anna Schuh, MD

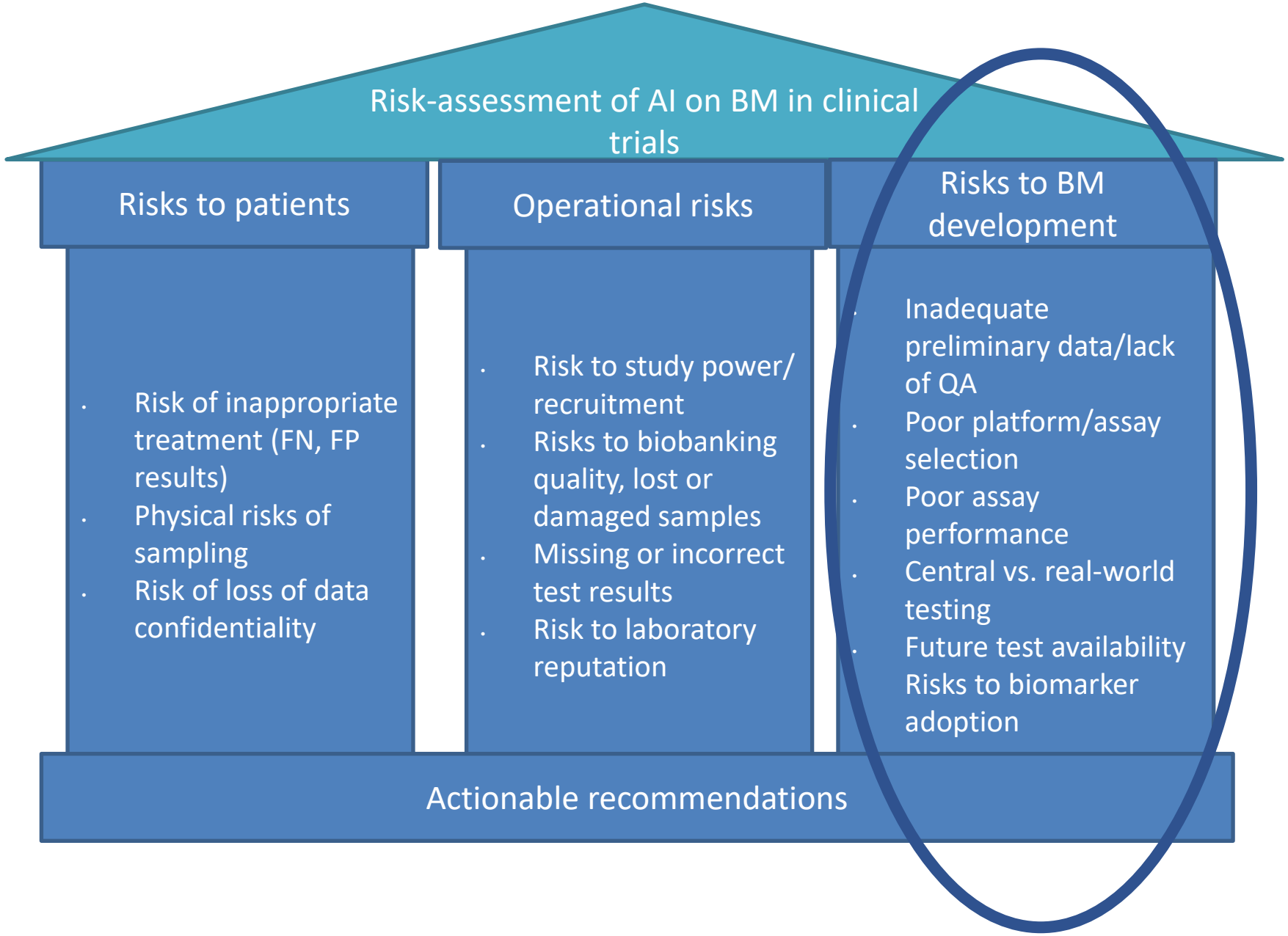
Radboudumc

TILS BREAST CANCER
INTERNATIONAL IMMUNO-ONCOLOGY WORKING GROUP

What is risk?

- **Risk:** combination of the probability of occurrence of harm and the severity of that harm (ISO/IEC Guide 51)

- **Risk assessment:** overall process comprising a risk analysis and a risk evaluation (ISO/IEC Guide 51)



The three core pillars of risk-assessment of designing and executing clinical trials including biomarker assessment using AI-Tools

Item	Risk	Risk mitigation strategies	TRIPOD
Specimen type and collection procedure	Different tissue preparation and pre-analytic factors introduce artefacts and noise (i.e. *batch effects*) and limit generalizability and reproducibility.	1. Standardize and report the following pre-analytic factors: - Tissue of origin - Preparation (FFPE, frozen, tissue microarray, cytology, etc) - Staining type (H&E, IHC, ISH, etc). - Staining procedure (reagents, vendors, concentrations, etc). 2. In a multi-centric, distributed setting, standardize tissue preparation and shipping.	4a
Scanning / Digitization procedure	Variable scanning parameters can limit the applicability of CP models in different settings.	1. Clear reporting of scanning procedure, including: - Scanner type and model - Scanning magnification and other settings - Visual inspection of physical slides (eg. wiping off “marker” ink) 2. Visual assessment after scanning (eg. illumination, staining or stitching artifacts).	4a
Whole-Slide Image standards	Non-standard formats and opaque image preprocessing procedures limit interoperability and broad applicability of CP models.	1. Consider using standard WSI image formats. If not applicable, provide details on accessing WSI data and details on image compression, magnification levels, etc. 2. Consider the use of a DICOM standard for interoperability. 3. Describe any post-scanning color management and image processing.	4a

Comment

How current assay approval policies are leading to unintended imprecision medicine

*Roberto Salgado, Andrew M Bellizzi, David Rimm, John M S Bartlett, Torsten Nielsen, Moch Holger, Anne-Vibeke Laenkholm, Cecily Quinn, Gábor Cserni, Isabela W Cunha, Isabel Alvarado-Cabrero, Ian Cree

Lancet Oncol 2020

Published Online October 21, 2020 https://doi.org/10.1016/S1470-2045(20)30592-1

***Panel:* Solutions to improve the current assay approval pathway**

- Industry should be mandated to do concordance studies with other similar assays or standardised controls before a drug is approved
- Industry should support, in concert with all stakeholders, relabelling or revising approved companion diagnostics if evidence exists that the labelling might lead to uncertainty in the identification of patients for treatments
- Industry should support, in concert with all stakeholders, relabelling or revising of the companion diagnostics if equivalent clinical validity has been shown with other biomarkers or standards, providing access to clinical trial tissues to validate other assays
- Industry, when considering the incorporation of assays in their trials, should communicate and share assay information when using an assay that identifies the same molecule (eg, epitope, antigen, DNA, RNA) as in other competitive trials—eg, method information related to the binding sites of the antibodies used in the companion diagnostic assay should be made public, even if this information is commercially sensitive
- Pathways for regulatory acceptance of other assays that are equivalent, but less expensive and easier to implement in daily practice, should be developed by governments and regulatory agencies, ideally before a drug is labelled together with a companion diagnostic
- Early engagement by all stakeholders in external quality control schemes to allow rapid development of guidelines and quality standards is essential, preferably before an assay is approved by the regulatory agencies
- Clinical practice guidelines developed by professional organisations like the American Society of Clinical Oncology and the European Society for Medical Oncology should endorse not just a companion diagnostic assay used in the trial, but any rigorously and technically validated equivalent laboratory assays that can define essentially the same population as the companion diagnostic
- Regulators should require data confirmation of the analytical validity of the companion diagnostic in the distributed setting in which it would be applied, at a level of rigor similar to that required to show efficacy of the drug in question

Solutions to improve the current diagnostic AI/ML algorithm approval narrative (Work in process)
Industry and academia should (?) be mandated to perform concordance studies with state-of-the-art algorithms or <u>standardized controls</u> before an algorithm is submitted.
Industry should support, in concert with all stakeholders, relabeling or revising approved computational diagnostic assays if there is evidence that the existing labeling may lead to uncertainty in the identification of patients for treatments.
Industry should support, in concert with all stakeholders, relabeling or revising of computational diagnostic assays if equivalent clinical validity has been demonstrated with other biomarkers or standards, providing access to clinical trial datasets for validation.
Industry, when considering the incorporation of AI/ML algorithms in their trials, should communicate and share pertinent details when using an algorithm that performs similar tasks (e.g., similar clinical endpoint, same molecular targets, etc) as in other competitive trials. Methodological information related to the algorithm design (e.g. neural network architecture in the case of deep learning), hyperparameters, as well as details on the datasets used for algorithm training, should be made public, even if this information is commercially sensitive.
Pathways for regulatory acceptance of other algorithms that are equivalent but require less computational resources and/or are easier to implement in daily practice, should be supported by governments and regulatory agencies ideally before an algorithm is labeled together with or as a companion diagnostic**.**
Early engagement by all stakeholders in External Quality Control Schemes to allow rapid development of guidelines and quality standards is essential, preferably before an algorithm is approved by the regulatory agencies.
Clinical practice guidelines developed by professional organizations like ASCO, ESMO, etc…should endorse not just the companion diagnostic assay used in a trial of interest, but any rigorously analytically validated equivalent computational diagnostic assays that can define the same population as the companion diagnostic.
Regulators should require data confirming the analytical validity of the algorithm in the distributed setting in which it would be applied.

The High Throughput Truthing (HTT) Project

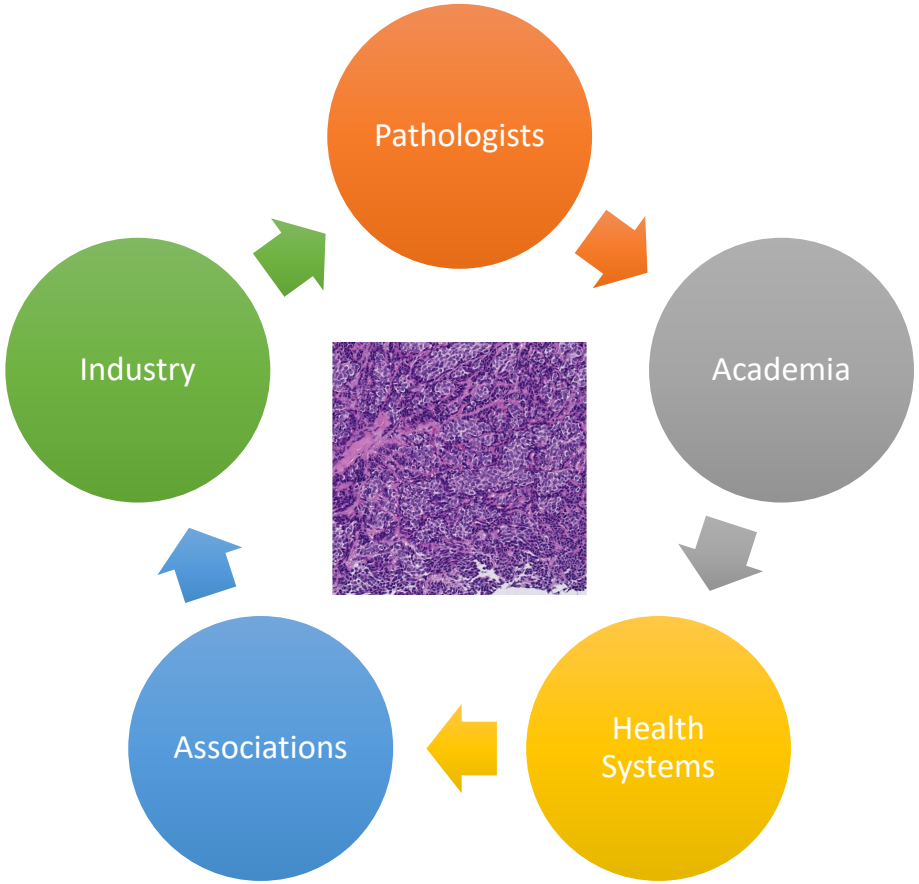
P.I. Brandon Gallas
U.S. FDA – CDRH – OSEL - DIDSR
https://ncihub.org/groups/eedapstudies

- **Goal:** Create a **dataset** of **pathologist annotations** for **validating** AI/ML models

- **Context**: TIL assessment to support
  - Clinical practice
  - Clinical trials

- Multi-stakeholder effort to elicit best practices

- Medical Device Development Tool (MDDT) | FDA



Why focus on the TIL assessment?

- Anticipate an influx of artificial intelligence and machine learning algorithms to assess TILs
- Community challenges on the computational TIL assessment already underway
  - TIGER Challenge (https://tiger.grand-challenge.org/)
  - CATALINA challenge (https://www.tilsinbreastcancer.org/tils-grand-challenge/)
- **Want to understand methods to quantify the uncertainty in reference standards being used by ML algorithms so we can better understand their performance and applications.**

HTT Pilot Study

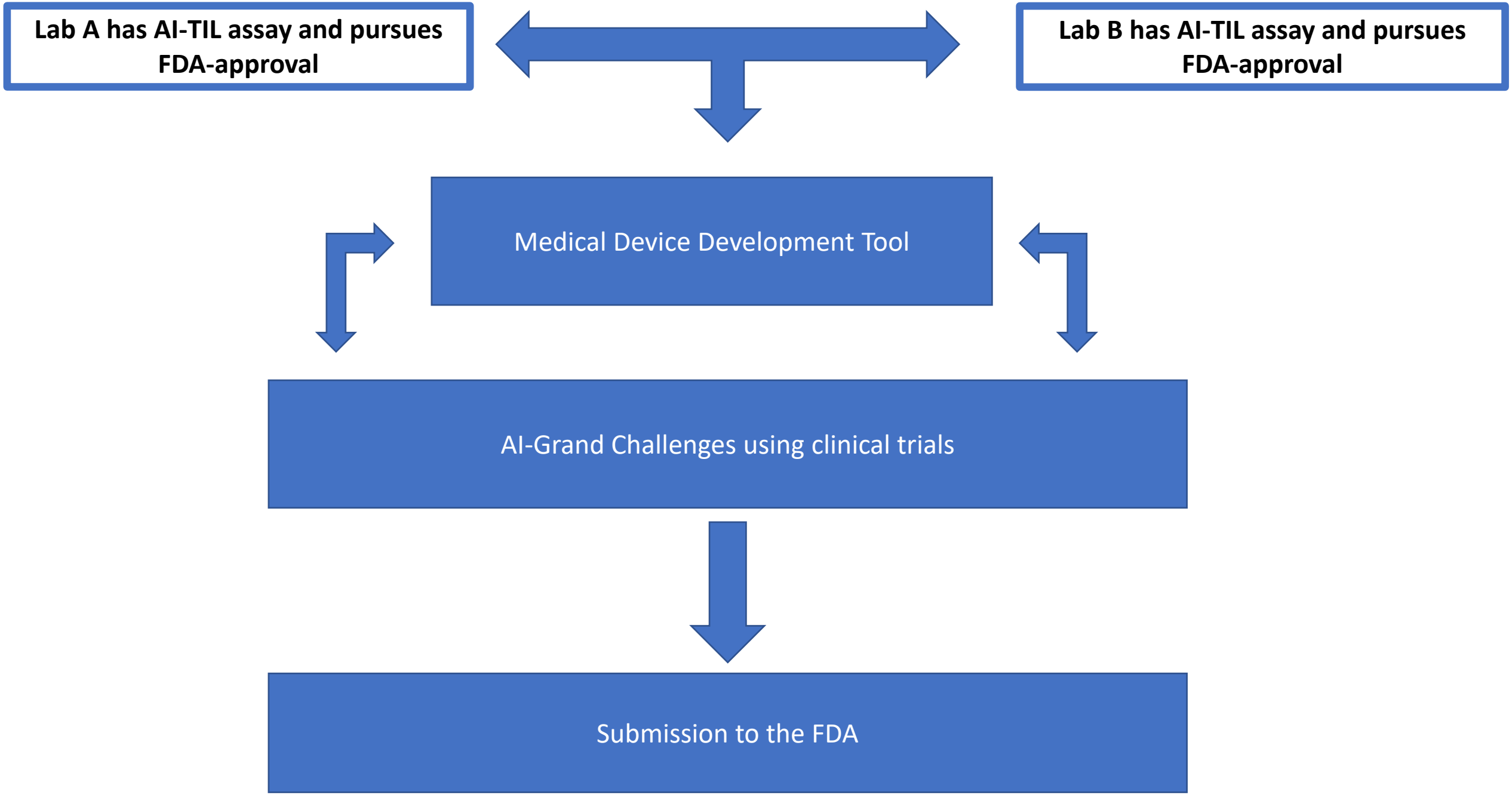
- Data collection in accordance with the TILs Working Group guidelines
  - Is the ROI evaluable for sTILs?
  - Percent of tumor-associated stroma
  - Estimated sTIL density
- Data collected using both light microscopy and digital annotation platforms
- Lessons being applied to a pivotal study currently under development
  - Pilot Study had higher than desired variance in collected data
  - Used an expert panel to create new training materials for the pivotal study
  - Developing statistical methods to analyze the nested and correlated data

A Framework for Testing, Validation and Deployment of Diagnostic Imaging in Anatomic Pathology.

An Internationally Quality Control program on Machine Learning Algorithms to Assess Quantitative Predictive/Prognostic Biomarkers in Breast Cancer such as TILs.

Radboudumc

TILS BREAST CANCER
INTERNATIONAL IMMUNO-ONCOLOGY WORKING GROUP


Lab A has AI-TIL assay and pursues FDA-approval
Lab B has AI-TIL assay and pursues FDA-approval
Medical Device Development Tool
AI-Grand Challenges using clinical trials
Submission to the FDA

Next steps

In parallel ongoing initiatives

- **Finalize the analysis of both private and public challenges.**
- **Present the data publicly at this forum.**
- **Progress in the MDDT-development on TILs.**
- **Publish a “Best Practices Manuscript” (ongoing).**
- **Develop New Challenges (options will be presented to the Trial groups)**

Thank you

Francesco, Brandon, James, Victor, Kate, Sunil, Mohamed, Sarah, Lee, Joe, Kim, Yinyin, Khalid, Sherene, Stefan, Carsten, Balazs, Johan, David, Lajos, Sybille, Jeannette, Anant, German, David, Damien, Jeroen, Bill, Torsten, John, Peter, Stephen,...>700 active people in the TILs-WG.

Radboudumc

TILS BREAST CANCER
INTERNATIONAL IMMUNO-ONCOLOGY WORKING GROUP

TIGER team

Contributors

- Laura Comerma Blesa
- Dieter Peeters
- Anna-Vibeke Laenkholm
- Harry Haynes
- Elisabeth Ida Specht Stovgaard
- Valerie Dechering
- Cyril de Kock
- Lee Cooper
- Mohamed Amgad
- Stephan Michiel
- Damien Drubay

TIGER is supported by:

TIGER is sponsored by AWS
Part of a general sponsorship of AWS and grand-challenge.org

Winning solutions will receive in total 13,000$ in the form of AWS credits

Radboudumc

BCRF BREAST CANCER RESEARCH FOUNDATION

TILS BREAST CANCER
INTERNATIONAL IMMUNO-ONCOLOGY WORKING GROUP

Pathology
Innovation
Collaborative
Community

PIcc

The Alliance for Digital Pathology

A collaborative community with FDA participation