Pitfalls in Assessing Stromal Tumor Infiltrating Lymphocytes (sTILs) in Breast Cancer

Zuzana Kos[1#], Elvire Roblin[2] , Rim S Kim[3], Stefan Michiels[2], Brandon D. Gallas[4], Weijie Chen[4], Koen K. van de Vijver[5,6], Shom Goel[7,8], Sylvia Adams[9], Sandra Demaria[10], Giuseppe Viale[11], Torsten O. Nielsen[12], Sunil Badve[13], W. Fraser Symmans[14], Christos Sotiriou[15], David L. Rimm[16], Stephen Hewitt[17], Carsten Denkert[18], Sibylle Loibl[19], Stephen J. Luen[8,20] , John Bartlett[21], Peter Savas[8,20], Giancarlo Pruneri[22], Deborah A. Dillon[23], Maggie Cheang[24], Andrew Tutt[25], Jacqueline A. Hall[26], Marleen Kok[27], Hugo M. Horlings[5,28], Anant Madabhushi[29], Jeroen van der Laak[30], Francesco Ciompi[30], Anne-Vibeke Laenkholm[31], Enrique Bellolio[32],Tina Gruosso[33], Stephen B. Fox[7,34], Juan Carlos Araya[35], Giuseppe Floris[36], Jan Hudeček[37], Leonie Voorwerk[38], Andrew H. Beck[39], Jen Kerner[39], Denis Larsimont[40], Sabine Declercq[41], Gert van den Eynden[41], Lajos Pusztai[42], Anna Ehinger[43], Wentao Yang[44], Khalid AbdulJabbar[45], Yinyin Yuan[45], Rajendra Singh[46], Crispin Hiley[47], Maise al Bakir[47], Alexander J. Lazar[48], Stephen Naber[49], Stephan Wienert[50], Miluska Castillo[51], Giuseppe Curigliano[52], Maria-Vittoria Dieci[53], Fabrice André[54], Charles Swanton[47], Jorge Reis-Filho[55,56], Joseph Sparano[57], Eva Balslev[58], I-Chun Chen[59], Elisabeth Ida Specht Stovgaard[58], Katherine Pogue-Geile[3], Kim R.M. Blenman[42], Frédérique Penault-Llorca[60], Stuart Schnitt[23], Sunil R. Lakhani[61], Anne Vincent-Salomon[62], Federico Rojo[63,64], Jeremy P. Braybrooke[65], Matthew G. Hanna[55], Teresa Soler[66], Daniel Bethmann[67], Carlos Castaneda[51], Karen Willard-Gallo[68], Ashish Sharma[69], Huang-Chun Lien[70], Susan Fineberg[71], Jeppe Thagaard[72], Laura Comerma[64,73], Paula Gonzalez-Ericsson[74], Edi Brogi[55], Sherene Loi[8,20], Joel Saltz[75], Frederick Klaushen[76], Lee Cooper[77], Mohamed Amgad[78], David A. Moore[79], and Roberto Salgado[20, 41] on behalf of the International Immuno-Oncology Biomarker Working Group[‡]

Author Affiliations

[1] Department of Pathology, BC Cancer - Vancouver, British Columbia, Canada

[2] Service de biostatistique et d'épidémiologie, Gustave Roussy, Université Paris-Saclay, Univ. Paris-Sud, Villejuif, France; INSERM, Unit 1018, Université Paris-Sud, University Paris-Saclay, Villejuif, France

[3] National Surgical Adjuvant Breast and Bowel Project (NSABP)/NRG Oncology, Pittsburgh, PA

[4] Division of Imaging, Diagnostics, and Software Reliability (DIDSR); Office of Science and Engineering Laboratories (OSEL); Center for Devices and Radiological Health (CDRH), Silver Spring, MD, USA

[5] Department of Pathology, The Netherlands Cancer Institute, Amsterdam, The Netherlands

[6] Department of Pathology, Ghent University Hospital, Ghent, Belgium

[7] Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia

[8] Peter MacCallum Department of Oncology, University of Melbourne, Melbourne, Victoria, Australia

[9] Perlmutter Cancer Center, New York University Medical School, New York, USA
[10] Departments of Radiation Oncology and Pathology and Laboratory Medicine, Weill Cornell Medicine, New York, USA
[11] Department of Pathology, Istituto Europeo di Oncologia, University of Milan, Milan, Italy.
[12] Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, Canada
[13] Department of Pathology and Laboratory Medicine, Indiana University, Indianapolis, USA
[14] Department of Pathology, The University of Texas M.D. Anderson Cancer Center, Houston, TX, USA
[15] Department of Medical Oncology, Institut Jules Bordet, Université Libre de Bruxelles, Brussels.
[16] Department of Pathology, Yale School of Medicine, New Haven, CT, USA
[17] Laboratory of Pathology, National Cancer Institute, NIH, Bethesda MD, USA
[18] Institute of Pathology, Universitätsklinikum Gießen und Marburg GmbH, Standort Marburg and Philipps-Universität Marburg, Marburg, Germany
[19] German Breast Group, Neu-Isenburg, Germany
[20] Division of Research and Cancer Medicine, Peter MacCallum Cancer Centre, University of Melbourne, Victoria, Australia
[21] Ontario Institute for Cancer Research, Toronto, ON Canada; and University of Edinburgh Cancer Research Centre, Edinburgh, UK
[22] Department of Pathology, IRCCS Fondazione Instituto Nazionale Tumori and University of Milan, School of Medicine, Milan, Italy.
[23] Department of Pathology, Brigham and Women's Hospital, Boston, MA Department of Pathology, Dana Farber Cancer Institute, Boston, MA
[24] Institute of Cancer Research Clinical Trials and Statistics Unit, The Institute of Cancer Research, Surrey, UK
[25] Breast Cancer Now Toby Robins Research Centre, The Institute of Cancer Research, London, UK
[26] Vivactiv Ltd, Bellingdon, Bucks, UK
[27] Divisions of Medical Oncology, Molecular Oncology & Immunology, The Netherlands Cancer Institute, Amsterdam, The Netherlands
[28] Division of Molecular Pathology, The Netherlands Cancer Institute, Amsterdam, The Netherlands
[29] Case Western Reserve University; Louis Stokes Cleveland Veterans Health Administration Medical Center
[30] Computational Pathology Group, Department of Pathology, Radboud University Medical Center, Nijmegen, Netherlands
[31] Department of Surgical Pathology Zealand University Hospital, Demark
[32] Departamento de Anatomía Patológica, Universidad de La Frontera. Temuco – Chile
[33] Forbius, Montreal, QC, Canada
[34] Department of Pathology, Peter MacCallum Cancer Centre Department of Pathology
[35] Department of Pathology, Universidad de la Frontera, Temuco, Chile
[36] KU Leuven- Univerisity of Leuven, Department of Imaging and Pathology, Laboratory of Translational Cell & Tissue Research and KU Leuven- University Hospitals Leuven, Department of Pathology, Leuven, Belgium
[37] Department of Research IT, The Netherlands Cancer Institute, Amsterdam, The Netherlands
[38] Division of Molecular Oncology & Immunology, The Netherlands Cancer Institute, Amsterdam, The Netherlands
[39] PathAI, Inc.
[40] Department of Pathology, Jules Bordet Institute, Brussels, Belgium
[41] Department of Pathology, GZA-ZNA, Antwerp, Belgium
[42] Yale Cancer Center Genetics, Genomics and Epigenetics Program, Yale School of Medicine, New Haven, CT , USA

[43] Lund University, Skane University Hospital, Department of Clinical Sciences Lund, Oncology and Pathology, Lund, Sweden
[44] Department of Pathology, Fudan University Shanghai Cancer Centre, Shanghai, China
[45] Centre for Evolution and Cancer; Division of Molecular Pathology, The Institute of Cancer Research, London, UK
[46] Icahn School of Medicine at Mt. Sinai, New York, NY 10029
[47] Cancer Research UK Lung Cancer Centre of Excellence, University College London Cancer Institute, University College London, London, UK
[48] Departments of Pathology, Genomic Medicine, Dermatology, and Translational Molecular Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA
[49] Department of Pathology and Laboratory Medicine , Tufts Medical Center
[50] Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Institute of Pathology, Charitéplatz 1, 10117 Berlin, Germany
[51] Department of Research, Instituto Nacional de Enfermedades Neoplasicas, Lima 15038, Peru
[52] University of Milan, Istituto Europeo di Oncologia, IRCCS; Milan, Italy
[53] Medical Oncology 2, Istituto Oncologico veneto - IRCCS, Padova, Italy; Department of Surgery, Oncology and Gastroenterology, University of Padova, Italy
[54] Department of Medical Oncology, Institut Gustave Roussy, Villejuif, France
[55] Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[56] Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[57] Montefiore Medical Center, Albert Einstein College of Medicine, Bronx, NY, USA
[58] Department of Pathology, Herlev and Gentofte Hospital, Denmark
[59] Department of Oncology, National Taiwan University Cancer Center, Taipei, Taiwan; Department of Oncology, National Taiwan University Hospital, Taipei, Taiwan; Graduate Institute of Oncology, College of Medicine, National Taiwan University, Taipei, Taiwan),
[60] Centre de Lutte contre le cancer - Centre Jean Perrin, Clermont-Ferrand, France
[61] The University of Queensland Centre for Clinical Research and Pathology Queensland, Brisbane, Australia
[62] Institut Curie, Paris Sciences Lettres Université, Inserm U934, Department of Pathology, Paris, France.
[63] Pathology Department, Instituto de Investigación Sanitaria Fundación Jiménez Díaz (IIS-FJD), Madrid, Spain
[64] GEICAM-Spanish Breast Cancer Research Group, Madrid, Spain
[65] Nuffield Department of Population Health, University of Oxford, Oxford and Department of Medical Oncology, University Hospitals Bristol NHS Foundation Trust, Bristol, UK
[66] Department of Pathology, University Hospital of Bellvitge, Oncobell, IDIBELL, L'Hospitalet del Llobregat, Barcelona 08908, Catalonia, Spain
[67] University Hospital Halle (Saale), Institute of Pathology, Halle (Saale), Germany
[68] Molecular Immunology Unit, Institut Jules Bordet, Universitè Libre de Bruxelles, Brussels, Belgium
[69] Department of Biomedical Informatics, Emory University, GA, USA
[70] Department of Pathology, National Taiwan University Hospital, Taipei, Taiwan
[71] Department of Pathology, Montefiore Medical Center and the Albert Einstein College of Medicine, Bronx, NY, United States
[72] DTU Compute, Department of Applied Mathematics, Technical University of Denmark; Visiopharm A/S, Denmark
[73] Pathology Department, Hospital del Mar, Parc de Salut Mar, Barcelona, Spain
[74] Breast Cancer Program, Vanderbilt-Ingram Cancer Center, Vanderbilt University Medical Center, Nashville, Tennessee

[75] Biomedical Informatics Department, Stony Brook University, Stony Brook, USA
[76] Institute of Pathology, Charité Universitätsmedizin Berlin, Berlin, Germany
[77] Department of Pathology, Northwestern University Feinberg School of Medicine, Chicago, IL
[78] Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA
[79] Department of Pathology, UCL Cancer Institute, UCL, London, UK; University College Hospitals NHS Trust, London, UK.
[‡] A full list of members and their affiliations is available at the end of the manuscript

[#]Corresponding author: Zuzana Kos: Zuzana.Kos@bccancer.bc.ca

Abstract

Stromal tumor infiltrating lymphocytes (sTILs) are important prognostic and predictive biomarkers in triple-negative (TNBC) and HER2-positive breast cancer. Incorporating sTILs into clinical practice necessitates reproducible assessment. Previously developed standardized scoring guidelines have been widely embraced by the clinical and research communities. We evaluated sources of variability in sTIL assessment by pathologists in three previous sTIL ring studies. We identify common challenges and evaluate impact of discrepancies on outcome estimates in early TNBC using a newly-developed prognostic tool.

Discordant sTIL assessment is driven by heterogeneity in lymphocyte distribution. Additional factors include: technical slide-related issues; scoring outside the tumor boundary; tumors with minimal assessable stroma; including lymphocytes associated with other structures; and including other inflammatory cells. Small variations in sTIL assessment modestly alter risk estimation in early TNBC but have the potential to affect treatment selection if cutpoints are employed. Scoring and averaging multiple areas as well as use of reference images improve consistency of sTIL evaluation. Moreover, to assist in avoiding the pitfalls identified in this analysis, we developed an educational resource available at www.tilsinbreastcancer.org/pitfalls.

Keywords

Tumor infiltrating lymphocytes; TILs; breast cancer; triple negative breast cancer; HER2 positive breast cancer; prognosis

Introduction

Despite the complexity of the immune system and intricate interplay between tumor and host antitumor immunity, detection of stromal tumor infiltrating lymphocytes (sTILs), as quantified by visual assessment on routine hematoxylin and eosin (H&E)-stained slides, has emerged as a robust prognostic and predictive biomarker in triple-negative and HER2 positive breast cancer.[1–3] Stromal TILs are defined as mononuclear host immune cells (predominantly lymphocytes) present *within* the boundary of a tumor that are located within the stroma *between* carcinoma cells without directly contacting or infiltrating tumor cell nests. Stromal TILs are reported as a percentage, which refers to the percentage of stromal area occupied by mononuclear inflammatory cells over the total stromal area within the tumor (i.e. *not* the percentage of cells in the stroma that are lymphocytes). Intratumoral TILs (iTILs), on the other hand, are defined as lymphocytes within nests of carcinoma *having cell-to-cell contact with no intervening stroma*. Initial studies of TILs in breast cancer evaluated stromal and intratumoral lymphocytes separately and while both correlated with outcome, sTILs were more prevalent, more variable in amount and shown to be more reproducibly assessed.[4–7] As such, recommendations for standardized assessment of TILs in breast cancer by the International Immuno-Oncology Biomarker Working Group (also referred to as TIL-Working Group, or TIL-WG in the manuscript; www.tilsinbreastcancer.org) recommend assessing sTILs whilst strictly adhering to the definition as outlined above.[8]

Stromal TILs are prognostic for disease-free and overall survival in early triple-negative breast cancers treated with standard anthracycline-based adjuvant chemotherapy.[4–6,9,10] High levels of sTILs are associated with improved outcome and increased response to neoadjuvant therapy in both triple-negative and HER2 positive breast cancers.[7,11–14] Recently, experts at the 16th St. Gallen International Breast Cancer Conference

endorsed routine reporting of sTILs in triple-negative breast cancer.[15] Studies involving or evaluating prognosis should now include the evaluation of sTILs.

The expanding role sTILs play in breast cancer research, prognosis and increasingly patient management, is predicated on accurate assessment of sTILs. The pivotal studies cementing the prognostic and predictive role of sTILs have been performed by visual assessment on H&E-stained slides according to published recommendations.[8] In the future, advances in machine learning may open the door to automated sTIL assessment.[16] Until that point, however, the onus for accurate sTIL assessment falls upon the pathologist.

Management of breast cancer is continually evolving. In contrast to the excisional biopsies of previous decades, an initial diagnosis of breast cancer is now routinely rendered on needle biopsy specimens. These small biopsies are particularly susceptible to influence of tumor heterogeneity, limited tumor sampling and technical artifacts such as crushing. Studies assessing concordance of TILs between core needle biopsies and matched surgical specimens (lumpectomy or mastectomy) report higher average TIL counts (4.4%-8.6% higher) in the surgical specimens.[17,18] The difference in TIL scores between biopsies and surgical specimens was found to be reduced when the number of cores was increased,[18] suggesting tumor heterogeneity as a contributing factor. Not specifically addressed was the tissue reaction and inflammatory infiltrate associated with the biopsy procedure itself. No increase in TIL scores within the surgical specimens was seen when surgery was performed within 4 days of the biopsy procedure. Conversely, surgery performed more than 4 days post biopsy was an independent factor correlating with higher TILs in the surgical specimen.[17] This corresponds to the timing of chronic inflammatory infiltrates in wound healing. It should be noted, however, that in most contemporary practice settings the delay between biopsy and surgery is several weeks and per the recommended guidelines, areas of scarring should

be excluded from sTIL assessment. The inflammation associated with wound healing is physically limited closely to the healing area and does not spread extensively into the tumor itself or surrounding stroma. Thus the impact of the biopsy procedure on sTIL levels in the surgical specimen is likely minimal.

Routine use of neoadjuvant therapy is increasingly common in triple-negative and HER2 positive breast cancers. These trends necessitate that sTIL assessment be performed on small biopsy samples and, in the absence of complete pathological response, on postneoadjuvant excision specimens without compromising accuracy. High levels of sTILs in residual tumor post neoadjuvant therapy is associated with improved outcome in TNBC.[19,20] As neoadjuvant samples possess distinct challenges, separate recommendations for assessing TILs in residual disease after neoadjuvant therapy have been published.[21]

Breast cancers show wide variation in morphology, particularly in tumor cellularity and amount of tumor stroma. Two tumors of the same size may exhibit the same absolute numbers of stromal lymphocytes but have a different percentage of sTILs due to the stromal content as a proportion of tumor area. High-grade tumors can show extensive central necrosis with only a thin rim of viable tumor resulting in minimal assessable tumor stroma even in large resection specimens. Other inflammatory cells are not infrequently seen infiltrating tumor stroma, including neutrophils, eosinophils and macrophages, resulting in a more cellular appearance and rendering assessment of stromal TIL density more challenging. Apoptotic cells can mimic lymphocytes. Poor fixation and technical artifacts in cutting and staining are recognized to compromise sTIL assessment. Ill-defined tumor borders and widely separated nests of tumor result in variability in defining what constitutes tumor stroma. Preexisting lymphocytic aggregates surrounding normal ducts and lobules, vessels or ductal carcinoma in situ (DCIS) can also confound assessment. Heterogeneity in sTIL

distribution both within the tumor and at the invasive front versus the central tumor all contribute to variation in pathologist sTIL assessment.

In an effort to identify the sources of variation in assessment of sTILs, we analyzed data and images from three ring studies performed by TIL-WG pathologists specifically evaluating concordance in sTIL evaluation in breast cancer.[22,23] Based on the findings of this analysis we designed an educational resource available via the International Immuno-Oncology Working Group website at www.tilsinbreastcancer.org/pitfalls to assist pathologists in avoiding the different types of pitfalls identified. In addition, we evaluated the impact of sTIL discrepancy on outcome estimation using the data of a pooled analysis of 9 phase III clinical trials.[9]

Results

Identification of cases demonstrating variability using ring studies by the TIL-Working Group

Three ring studies evaluating concordance of sTIL assessment in breast cancer were analyzed (Figure 1). In the first ring study, 32 pathologists evaluated 60 scanned breast cancer core biopsy slides.[22] This international group of pathologists from 11 different countries were all members of the TIL Working Group. Some had a special interest or subspecialty training in breast pathology, while others were general surgical pathologists, illustrating the wide applicability of the approach. The only instructions given to the scoring pathologists were to read and use the TIL assessment guidelines published by the TIL working group.[8] The second ring study was an extension of the first study using a more formalized approach. A subset of 28 of the original 32 pathologists participated and scored 60 different scanned breast cancer core biopsy slides. In this study, each pathologist identified and scored at least three separate 1

mm^2 regions on each slide, representing the range of sTIL variability and averaged the results into a final score. Additionally, reference images representing different sTIL percentages were integrated into the evaluation process (Figure 2).[22] The last ring study was performed by six TIL-WG pathologists who independently scored 100 scanned whole section (excision specimen) breast cancer cases.[23]

In total, results from 220 slides were included for statistical analysis (60 each from ring studies 1 and 2, and 100 from ring study 3). The standard deviation for sTIL scores for each slide is shown in Figure 3. When comparing across studies, ring study 2 shows the least variation in sTIL scores between pathologists. The cases with the 10% greatest standard deviation were identified (Figure 3 – red squares) and the original scanned slides of the cases where reviewed to identify factors contributing to discordant sTIL assessment in these cases. Additionally, in Ring Study 1, a single outlier case in the low sTIL range was also evaluated (Figure 3a – black triangle). From Ring Study 3, three additional cases showing large standard deviation were also included in the scanned slide assessment (Figure 3c – black triangles). Overall, a total of 26 original scanned images were reviewed by ZK (ring studies 1 and 2) and RK (ring study 3) from cases identified as particularly problematic (i.e. showing high variability) in sTIL assessment.

Analysis of scoring variance between pathologists

Table 1 shows the Intraclass Correlation Coefficient (ICC) and concordance rate among pathologists for each of the 3 studies. The ICC is the proportion of total variance (in measurements across patients and laboratories) that is attributable to the biological variability among patients' tumors, while 1 – ICC is the proportion attributable to pathologist variability. The ICC has a range from 0 to 1 with a score of 1 having the maximum agreement. Concordance rates were evaluated comparing different sTIL cutpoints: <1 vs ≥1%; <5 vs ≥5%;

<10 vs ≥10%; <30 vs ≥30%; <75 vs ≥75% for each pathologist by comparing all pairs of pathologists.

The ICC was highest in ring study 2 compared to the other studies. Ring study 2 specifically sought to mitigate effects of sTIL heterogeneity with assessment of 3 separate areas and intra-pathologist scoring bias by necessitating use of standardized percentage sTIL reference images.

Evaluation of sources of variability in the three ring studies

The scanned images of the H&E-stained slides from the most discordant cases in each of the 3 ring studies were evaluated to identify the histological factors contributing to the variation in sTIL assessment. In total 26 original scanned images were reviewed – 7 from ring study 1, 6 from ring study 2 and 13 from ring study 3. Often multiple factors were present in each slide.

Heterogeneity in sTIL distribution

Heterogeneity in sTIL distribution was identified as a major contributing factor in all of the ring studies and as the most prevalent challenge in ring studies 1 and 2 (Table 2; Figure 4). Based on review of the most variable cases, increased sTIL density at the leading edge versus central tumor were contributing factors in 43%, 17 % and 54% of cases in ring studies 1 through 3, respectively (Figure 4a); and marked heterogeneity of sTIL density within the tumor was identified in 29% cases in ring study 1 only (Figure 4b). Whereas in ring studies 1 and 3 pathologists provided a global sTIL assessment based simply on the published scoring recommendations,[8] ring study 2 specifically addressed the issue of sTIL heterogeneity by requiring separate scoring of at least 3 distinct areas of the tumor representing the range of sTIL density. Additionally, matching the tumor area observed with reference percent sTIL images were a necessary part of the evaluation. Our analysis supports

that scoring and averaging multiple areas aids in providing a more consistent result between pathologists. One issue not resolved by this technique is the scenario of a tumor comprised of variably spaced apart clusters of epithelial cells with a dense lymphocytic aggregate associated with each cluster of epithelial nests but sparse infiltrate between the clusters (Figure 4c). This pattern was identified as a contributing factor in 29% of highly discordant cases in ring study 1, 50% of discordant cases in ring study 2 and no cases in ring study 3. There appears to be uncertainty amongst pathologists in this situation as to whether to only include the stroma associated with–but not touching–tumor epithelium (showing high sTIL density) or all stroma within the tumor mass including stroma intervening between spaced apart clusters of malignant epithelium (showing low sTIL density). This uncertainty increases variability in sTIL assessment and would be reduced by strict adherence to the definition of sTILs provided in the introduction. All stroma within a single tumor is to be included within the sTIL assessment. In this situation, both the higher density areas in close proximity to tumor cells and the lower density areas located between epithelial clusters should be included. One notable exception is a tumor with a central hyalinized scar, where the acellular scar tissue should be excluded from sTIL assessment.

Technical factors

Technical factors were the next largest source of discordance (table 3; Figure 5). Poor quality slides with histological artifacts, as can be seen secondary to prolonged ischemic time, poor fixation, issues during processing, embedding or microtomy were identified as a contributing factor for discordance in 85% of the most discordant scanned slides from ring study 3 (Figure 5a). In contrast, this was not deemed a contributing factor in any of the cases from ring studies 1 or 2. These results are highly skewed based on the studies assessed. Ring study 3 used a subset of H&E slides from NSABP-B31, an older completed trial evaluating

benefit of trastuzumab in early HER2 positive breast cancer, which started accrual in February 2000 across multiple centers. These were excision specimens undergoing local community tissue processing. Variable ischemic and fixation times subsequently affected the integrity of stromal connective tissue which is critical in sTIL assessment. Ring studies 1 and 2 used pretherapeutic core biopsies from the neoadjuvant GeparSixto trial, which accrued between Aug 2011 and Dec 2012. Fixation and ischemic time are less likely to have been an issue in these samples, which (i) as biopsy samples are immediately placed in formalin without requirement for serial sectioning and can be processed in a timely fashion and (ii) were procured at a time when the preanalytic variables had become substantially better understood and new recommendations widely adopted. Not to mention, H&E stains fade with passage of time, which itself impacts the ability to produce quality scanned images. In the current era, with awareness and adoption of standardization and monitoring of preanalytical and analytical variables, poor quality H&E slides should no longer be acceptable. Nonetheless, challenges remain and variations in practice can result in poorly processed specimens which are likely to directly and negatively impact sTIL assessment. Crush artifact, which is more commonly seen in core biopsy samples, was seen in 1 case overall in ring study 1 (14%) (Figure 5b).

Out-of-focus scans were identified in 1 case each in ring study 1 (14%) and ring study 2 (17%) (Figure 5c). In clinical practice, particularly as sTILs are poised to impact patient management, an out-of-focus slide should be rescanned before scoring. Notably, this highlights an obstacle to incorporation of whole slide imaging in routine practice. Consistent focus quality remains an issue requiring dedicated support staff for loading, scanning, reviewing and rescanning if necessary.[24]

Including wrong area or cells

Variability in defining the tumor boundary and scoring stroma outside of the tumor boundary appears to have been a contributing factor for variation in 33% of highly discordant cases in ring study 2 and 15% of cases in ring study 3 (Table 4; Figure 6a). The discordant cases also highlighted situations of including lymphocytes associated with DCIS (2 cases ring study (RS)1, 1 case RS2) (Figure 6a), lymphocytes associated with a component of the tumor showing features of an encapsulated papillary carcinoma (1 case RS1) (Figure 6b), and lymphocytes associated with benign terminal duct lobular units (1 case RS1) (Figure 6d). Difficulty distinguishing iTILs from sTILs factored into 2 cases (29%) in ring study 1 and 1 case (17%) in ring study 2 (Figure 7a). Also identified in ring study 1 was 1 case (14%) with prominent stromal neutrophils (Figure 7b) and 1 case (14%) with stromal histiocytes (Figure 7c). It is important to assess slides at a sufficiently high power to be able to differentiate between types of immune cells. Neutrophils, eosinophils, basophils, and histiocytes/macrophages are all excluded from sTIL assessment. Two independent cases in ring study 1 demonstrated misinterpretation of apoptotic cells for lymphocytes (Figure 7d) and artefactual falling apart of tumor cell nests along the edge of a core biopsy mimicking the discohesive appearance of TILs (Figure 7e). Both are previously noted examples of histomorphologic challenges.

Limited stroma within tumor for evaluation

An added factor identified was the presence of minimal stroma in the tumor for assessment (Table 5; Figure 8a). This was identified as a contributing factor in 46% of cases in ring study 3. In a variation, 1 case (14%) in ring study 1 showed extensive tumor necrosis with decreased available stroma for assessment (Figure 8b). Two cases (15%) of mucinous tumors, each with minimal stroma to assess were identified in ring study 3 (Figure 8c).

Clinical significance of variability in sTIL assessment by pathologists

The online triple-negative breast cancer (TNBC)-prognosis tool (www.tilsinbreastcancer.org) that contains cumulative data of 9 phase III TNBC-trials,[9] was used to analyze the impact of variation in sTIL assessments (using the sTIL-scores of this analysis) on outcome. The impact on outcome of different sTIL levels is represented in Figure 9, showing a prototypical example of a 60-year-old patient with a histological grade 3 triple-negative breast carcinoma, measuring between 2-5 cm (pT2) and showing 30% sTILs. Assuming she is node negative, if a pathologist properly quantifies the percentage of sTILs, the 5-years invasive disease-free survival (iDFS) is estimated at 76%. If the pathologist deviates down 10% in scoring sTILs (i.e. 20% sTILs), the 5-years iDFS decreases to 73%. Conversely, if the pathologist deviates up 10% in scoring sTILs (i.e. 40% sTILs), the 5-years iDFS goes up to 79%. These differences are modest from a purely prognostic viewpoint, although larger variations would lead to more pronounced differences in outcome estimation. If cutpoints are used to decide on therapy, on the other hand, variation in values around the cut point (as reflected in the concordance rates in Table 1 and Supplemental material) may impact clinical management. Additional examples of outcome estimation as a function of sTILs are provided in the Supplemental material.

A new resource for pathologists

To assist pathologists in avoiding the different types of pitfalls in the assessment of sTILs identified in this analysis, we have developed an educational tool available via the International Immuno-Oncology Working Group website at www.tilsinbreastcancer.org/pitfalls. Both conventional pictures of microscopic slides and digitized whole slide images (WSIs) of biopsies and surgical resection specimens of breast and other cancers are available to illustrate the described pitfalls. At this point in time, we have included several examples of each of the pitfalls. In the future we intend to add extra

illustrative examples to make this collection a 'living' library and continuously evolving learning tool for the pathology community. We invite the pathology community to provide examples of challenging cases for TIL evaluation via the website.

Discussion

In the current study we evaluated factors in sTIL assessment which serve to increase the interobserver variability of manual sTILs assessment. The data were analyzed as both continuous and categorical variables. Despite the challenges pathologists face in scoring sTILs, the reported prognostic and predictive value of sTILs remains consistent across multiple datasets analyzed by independent investigators.[9,25] On the individual patient level, however, we have shown that discrepancies in sTILs scoring between pathologists results in different individual outcome estimations, requiring refinements in the paradigm to maximize benefit and minimize risk.

Notable strengths of this study include the evaluation of both core biopsy and excision specimens, which reflect the reality of clinical practice in which sTIL assessment will be performed. Analyzing the concordance rates across various cutpoints allows us to inform regarding reproducibility to aid in educated cut point selection for future trials. If a singular cutpoint is used, variation in values around that cutpoint can result in misassignment. However, in the setting of an understanding of the scoring error, the cutpoint can be adjusted to a range such that below is X, above is Y and between is indeterminate, and based on a strategy of risk management the overall risk is mitigated. The extensive reference images in this manuscript as well as the online education resource with further examples (www.tilsinbreastcancer.org/pitfalls) are a valuable reference guide to the pathology community.

A limitation to consider is the poor quality of many of the slides from the excision specimen sections in ring study 3 that were identified as showing the highest discordance. This skewed the evaluation towards technical factors, which are likely to be less of an issue in contemporary clinical practice, but are of relevance in retrospective analyses from older clinical trials. Nonetheless, if presented with such a case in practice, only intact, morphologically assessable areas should be included in sTIL score. If applicable, one could attempt recutting and staining a new slide or selecting a different block for assessment. This information further bolsters the demands for optimal tissue handling and processing.

Among the sources of variability identified, the greatest challenge appears to be dealing with heterogeneous distribution of sTILs. This issue was partially mitigated in ring study 2 which required assessment and averaging of at least 3 separate areas of tumor. The areas were selected by the pathologist to reflect the range of sTIL density and could be within a single core or across separate cores depending on the case. One may postulate that the increased experience of having participated in ring study 1 accounts for the greater concordance in ring study 2; however, the pathologists in ring study 3 had participated in the previous two ring studies and nonetheless showed lower ICC and concordance rates than ring study 2. Ring study 3 was the only study using whole sections compared to core biopsies in the other two studies. One could consider that the increased area of tumor in an excision specimen could lead to increased discordance.[26] In reality, however, many of the core biopsy cases contained multiple tissue cores per slide with multiple separate fragments of tumor, which likely negated any benefit of smaller tumor area. Although the recommendation to score multiple areas and average them in the setting of a heterogeneous tumor is within the published recommendation guidelines,[8] the software in ring study 2 made this a firm requirement. Similarly, use of reference % sTIL images is recommended in the guideline but

was a mandatory component of ring study 2. We identified these two key recommendations from the scoring guidelines as having a major impact on consistency of results. These two relatively simple steps: scoring multiple areas in heterogeneous tumors and always using reference images (to minimize personal assessment bias to always “score high” or “score low”) [27] substantially improve concordance. This re-enforces the central importance of adhering to recommendations in the scoring guidelines. Once factors of heterogeneity are excluded, taking the time to evaluate slides at a sufficiently high power to distinguish lymphocytes from other immune cells as well as mimics can further improve concordance. Being cognizant of lymphoid aggregates around benign ducts and lobules, vessels and DCIS outside of the tumor will help identify these as unrelated to the invasive carcinoma when present within the tumor boundary where these lymphoid aggregates should be excluded from sTIL assessment.

Demonstration of the reproducibility of sTILs scoring is essential for widespread adoption. The importance of sTILs as a biomarker is being increasingly recognized resulting in recommendations by multiple respected groups. The 2019 St. Gallen Panel recommended that sTILs be routinely characterized in TNBC for their prognostic value. [15,8] As of yet, however, insufficient data exists to recommend sTILs as a test to guide systemic treatment. In addition, the next iteration of the *WHO Classification of Tumours of the Breast* will also include information on sTILs.

Stromal TIL-assessment by pathologists is now recognized as an analytically and clinically validated biomarker. There is Level 1B evidence that high levels of sTILs are associated with improved outcome and an enhanced response to neoadjuvant therapy in triple-negative and HER2-positive breast cancers, [7,11–14,28] and are prognostic for disease-free and overall survival in early triple-negative breast cancers treated with standard

anthracycline-based adjuvant chemotherapy.[4,6,9] Clinical utility [likelihood of improved outcomes from use of the biomarker test compared to not using the test][29] remains to be defined. A recent retrospective study demonstrated that patients with Stage I TNBC with ≥30% sTILs had excellent survival outcomes (5-year overall survival rate of 98% [95%CI: 95% to 100%]) in the absence of chemotherapy,[30] paving the way for future randomized trials of chemotherapy de-escalation in early TNBC.

Clinical utility for sTILs is also likely to come from cancer immunotherapy, a rapidly emerging field aimed at augmenting the power of a patient's own immune system to recognize and destroy cancer cells. The immune system is able to impart selective pressure on cancer cells resulting in immune-evading clones. Stromal TILs can identify tumors amenable to immunotherapies targeting immunosuppression.[31] Checkpoint inhibitors of programmed cell death protein 1 (PD-1) and programmed death-ligand 1 (PD-L1) are promising therapeutic interventions, however predicting tumor response to these agents remains challenging.[32] There is increasing hesitation about the utility of the current predictive biomarker PD-L1 expression by IHC. The utility of PD-L1 IHC is undermined by the well-characterized geographic and temporal heterogeneity and dynamic expression on tumor or tumor-infiltrating immune cells.[33] Technical differences, variable expression and variation in screening thresholds for PD-L1 expression across assays pose additional limitations. Studies have shown that although pathologists can score PD-L1 on tumor cells with high concordance, even with training they are not concordant in scoring PD-L1 on immune cells.[34–36] There is emerging data that sTILs as assessed by the consensus-method defined by the TIL Working Group are predictive for response to checkpoint-inhibition in metastatic triple-negative and HER2-positive breast cancer. [37,38] The response rate is linear with increasing sTILs related to a higher response rate.[38] Further investigations are ongoing.

As we look to the future, automated sTIL assessment holds the promise of adding complementarity to the current pathological evaluation of breast cancers. A heterogeneous pattern of lymphocyte infiltration may be better addressed with computational pathology methods.[39,40] Further, there is some evidence that the spatial distribution of TILs may provide additional prognostic information.[41] One study reported improved prognosis and response to chemotherapy in TNBC with a diffuse, homogeneous lymphocyte distribution versus a heterogeneous distribution.[42] This requires further evaluation. Lymphocytes are particularly well-suited to image analysis, as it is easier to recognize these small blue dark cells against a stromal background than, for example, to distinguishing malignant cells from normal epithelium. There is a surge in the development of machine learning methods for TIL assessment.[43] The histopathologic diagnostic responsibility will continue to reside with the pathologist. Image analysis and computation pathology, which are proven to be faster and more reproducible, are adjuncts that aid the pathologist but do not replace the function of histopathologic interpretation. Until these tools are available, the well-educated and well-trained pathologist is the best approach. Rigorous training, evaluation and practice are well documented to result in improved intra- and inter-pathologist reproducibility. It is hoped that by highlighting the specific pitfalls in sTIL assessment in this manuscript – the forewarned pathologist is the forearmed pathologist. Ongoing efforts to ensure reliable and reproducible reporting of sTILs are a key step in their smooth progression into the routine clinical management of breast cancer.

Methods

Identification of cases demonstrating variability using ring studies by the TIL-Working Group

We identified 3 ring studies evaluating concordance of sTIL assessment in breast cancer performed by TIL-WG pathologists, for which we could obtain individual pathologist data and images. [22,23] The ring studies were performed on clinical trials material. All participating patients gave written informed consent to sample collection and the use of these samples for translational biomarker research, as approved by the Ethics Commission of the Charité Universitätsmedizin Berlin. All relevant ethical regulations have been complied with for this study. In ring study 1, 32 pathologists evaluated 60 scanned breast cancer core biopsy slides.[22] Scores were missing for 5 slides; the missing values were replaced by the mean of the 31 remaining scores. Ring study 2 was an extension of the first study. A subset of 28 of the original 32 pathologists participated and scored 60 different scanned breast cancer core biopsy slides.[22] Ring study 3 was performed by six TIL-WG pathologists who independently scored 100 scanned whole slide breast cancer cases.[23] In total, 220 slides were included. For each individual slide, the variability (standard deviation) among pathologists was measured from individual sTILs scores. The slides with the highest 10% standard deviation were identified for evaluation.

Statistical analysis of scoring variance between pathologists

The R software environment was used for statistical computing and graphics (version 3.5.0). Scoring variance among pathologists was analyzed using the Intraclass Correlation Coefficient (ICC). ICC estimates and their 95% confidence intervals were calculated based on individual-pathologist rating (rather than average of pathologists), absolute-agreement (i.e., if different pathologists assign the same score to the same patient), 2-way random-effects model (i.e., both pathologists and patients are treated as random samples from their respective populations).[44] To compute ICC, we used the “aov” function to fit the data with a two-way random effect ANOVA model (readers and cases). We followed Fleiss and Shrout’s

method to approximate the ICC confidence intervals.[45] We created custom code for the concordance analysis. Concordance rates for all pairs of pathologists were calculated at several sTIL density cutpoints: <1 vs ≥1%; <5 vs ≥5%; <10 vs ≥10%; <30 vs ≥30%; <75 vs ≥75%. Specifically, each concordance was the percent agreement from the 2x2 table created from each cutpoint and pair of readers. The analyses were performed and confirmed independently by two separate groups (RE & SM; Gustave Roussy) and (BDG & WC; FDA). Details of the concordance analysis are presented in Supplementary Tables 1-3.

Evaluation of sources of variability in the three ring studies

Slides for ring study 1 and 2 were Whole Slide Images (WSI) and were viewed using a virtual microscope program (CognitionMaster Professional Suite; VMscope GmbH). Each slide identified as showing the top 10% discordance, as well as specifically chosen cases (1 outlier low sTIL case in ring study 1 and 3 additional high discordance cases from ring study 3) were examined in order to identify potential confounding factors for routine sTIL assessment.

Clinical significance of variability in sTIL assessment by pathologists

The impact of variation in sTILs on outcome estimation was evaluated using the online triple-negative breast cancer (TNBC)-prognosis tool (www.tilsinbreastcancer.org) that contains cumulative data of 9 phase III TNBC-trials. The sTIL scores of this analysis were used as the ground truth. Specifically, different patient profiles were defined based on standard clinicopathological factors: age, tumor size, number of positive nodes, tumor histological grade and treatment. For a specific patient profile and a value of sTIL, the tool was used to calculate the 5-year invasive disease-free survival (iDFS). The iDFS is defined as the date of first invasive recurrence, or second primary or death from any cause.

Data Availability

The histology images supporting figure 2 and figures 4-8, are publicly available in the figshare repository, as part of this record: https://doi.org/10.6084/m9.figshare.11907768.[46] Data supporting figure 3, tables 1-5 and supplementary tables 1-3 are not publicly available in order to protect patient privacy. These datasets can be accessed on request from Dr. Roberto Salgado, upon the completion of a Data Usage Agreement, according to policies from the German Breast Group and NSABP, as described in the data record above. Figure 9 and supplementary figures 1-8, were generated using the publicly available prognosis tool at www.tilsinbreastcancer.org/, which utilises datasets from a pooled analysis of 9 phase 3 breast cancer trials, including BIG 02-98, ECOG 1199, ECOG 2197, FinHER, GR, IBCSG 22-00, IEO, PACS01 and PACS04 (**https://doi.org/10.1200/JCO.18.01010**). This paper is intended to serve as a practical reference for practicing pathologists. The ring studies were simply a means to identify representative cases that are particularly challenging to score in order to provide reference images and guidance on how to deal with these cases.

Code Availability

The code is available from the corresponding author by request.

International Immuno-Oncology Biomarker Working Group Members

Aini Hyytiäinen[80], Akira I. Hida[81], Alastair Thompson[82], Alex Lefevre[83], Allen Gown[84], Amy Lo[85], Ana Sapino[86], Andre Moreira[87], Andrea Richardson[88], Andrea Vingiani[89], Andrew M. Bellizzi[90], Angel Guerrero[91], Anita Grigoriadis[92], Anna Garrido Castro[93], Ashley Cimino-Mathews[94], Ashok Srinivasan[3], Balazs Acs[95], Baljit Singh[96], Benjamin Calhoun[97], Benjamin Haibe-Kans[98], Benjamin Solomon[8], Bibhusal Thapa[99], Brad H. Nelson[100], Carmen Ballesteroes-

458 Merino[101], Carmen Criscitiello[102], Carolien Boeckx[83], Cecile Colpaert[103], Cecily

459 Quinn[104], Chakra S. Chennubhotla[105], Cinzia Solinas[106], Damien Drubay[107], Dhanusha

460 Sabanathan[108], Dieter Peeters[109], Dimitrios Zardavas[110], Doris Höflmayer[111], Douglas B.

461 Johnson[112], E. Aubrey Thompson[113], Edith Perez[114], Ehab A. ElGabry[115], Elizabeth F.

462 Blackley[8], Emily Reisenbichler[16], Ewa Chmielik[116], Fabien Gaire[117], Fang-I Lu[118], Farid

463 Azmoudeh-Ardalan[119], Franklin Peale[120], Fred R. Hirsch[121], Gabriela Acosta-Haab[122], Gelareh

464 Farshid[123], Glenn Broeckx[124], Harmut Koeppen[85], Harry R. Haynes[125], Heather

465 McArthur[126], Heikki Joensuu[127], Helena Olofsson[128], Ian Cree[129], Iris Nederlof[130], Isabel

466 Frahm[131], Iva Brcic[132], Jack Chan[133], James Ziai[85], Jane Brock[23], Jelle Weseling[5], Jennifer

467 Giltnane[85], Jerome Lemonnier[134], Jiping Zha[135], Joana Ribeiro[136], Jochen K. Lennerz[137], Jodi M.

468 Carter[138], Johan Hartman[139], Johannes Hainfellner[140], John Le Quesne[141], Jonathon W.

469 Juco[142], Jose van den Berg[5], Joselyn Sanchez[51], Joël Cucherousset[143], Julien Adam[144], Justin

470 M. Balko[145], Kai Saeger[146], Kalliopi Siziopikou[147], Karolina Sikorska[148], Karsten Weber[19], Keith

471 E. Steele[135], Kenneth Emancipator[142], Khalid El Bairi[149], Kimberly H. Allison[150], Konstanty

472 Korski[117], Laurence Buisseret[68], Leming Shi[151], Loes F. S. Kooreman[152], Luciana

473 Molinero[120], M. Valeria Estrada[153], Maartje Van Seijen[130], Magali Lacroix-Triki[154], Manu M.

474 Sebastian[155], Marcelo L. Balancin[156], Marie-Christine Mathieu[157], Mark van de

475 Vijver[158], Marlon C. Rebelatto[135], Martine Piccart[159], Matthew P. Goetz[114], Matthias

476 Preusser[140], Mehrnoush Khojasteh[160], Melinda E. Sanders[161], Meredith M. Regan[162], Michael

477 Barnes[163], Michael Christie[164], Michael Misialek[165], Michail Ignatiadis[15], Michiel de

478 Maaker[130], Mieke Van Bockstal[166], Nadia Harbeck[167], Nadine Tung[168], Nele Laudus[169], Nicolas

479 Sirtaine[170], Nicole Burchardi[171], Nils Ternes[2], Nina Radosevic-Robin[172], Oleg Gluz[173], Oliver

480 Grimm[117], Paolo Nuciforo[174], Paul Jank[175], Pawan Kirtani[176], Peter H. Watson[12], Peter

481 Jelinic[142], Prudence A. Francis[7,8], Prudence A. Russell[177], Robert H. Pierce[178], Robert Hills[179],

Roberto Leon-Ferre[114], Roland de Wind[170], Ruohong Shui[44], Samuel Leung[12], Sami Tabbarah[180], Sandra C. Souza[181], Sandra O'Toole[182], Sandra Swain[183], Sarah Dudgeon[4], Scooter Willis[184], Scott Ely[185], Shahinaz Bedri[186], Sheeba Irshad[187], Shi-Wei Liu[188], Shona Hendry[34], Simonetta Bianchi[189], Sofia Bragança[190],Soonmyung Paik[3], Sua Luz[191], Thomas Gevaert[192], Timothy d'Alfons[55], Tom John[193], Tomohagu Sugie[194], Uday Kurkure[160], Veerle Bossuyt[137], Venkata Manem[98], Vincente Peg Cámaea[195], Weida Tong[196], William T. Tran[180], Yihong Wang[197], Yves Allory[198], Zaheed Husain[199], Zsuzsanna Bago-Horvath[200]

Affiliations

[80] Department of Oral and Maxillofacial Diseases, Helsinki, Finland, [81] Department of Pathology, Matsuyama Shimin Hospital, Matsuyama, Japan, [82] Surgical Oncology, Baylor College of Medicine, Texas, USA, [83] Roche Diagnostics, Belgium, [84] PhenoPath Laboratories, Seattle, USA, [85] Research Pathology, Genentech Inc., South San Francisco, USA, [86] University of Turin / Candiolo Cancer Institute - FPO, IRCCS, Candiolo, Italy , [87] Pulmonary Pathology, New York University Center for Biospecimen Research and Development, New York University, New York, New York, USA, [88] Department of Pathology, Johns Hopkins Hospital, Baltimore, USA, [89] Department of Pathology, Insituto Nazionale dei Tumori, Milan, Italy, [90] Department of Pathology, University of Iowa Hospitals and Clinics, Iowa City, USA, [91] Department of Oncology, IVO Valencia, Spain, [92] Cancer Bioinformatics Lab, Cancer Centre at Guy's Hospital, London, UK; School of Life Sciences and Medicine, King's College London, London, UK, [93] Dana Farber Cancer Institute, Boston, USA, [94] Departments of Pathology and Oncology, The Johns Hopkins Hospital, Baltimore, USA, [95] Department of Pathology, Karolinska Institute, Sweden, [96] Department of Pathology, New York University Langone Medical Centre, New York, USA, [97] Department of Pathology and Laboratory

Medicine, UNC School of Medicine, USA, [98] Bioinformatics and Computational Genomics Laboratory, Princess Margaret Cancer Center, Toronto, Canada, [99] Department of Medicine, University of Melbourne, Parkville, Australia, [100] Trev & Joyce Deeley Research Centre, British Columbia Cancer Agency, Victoria, Canada, [101] Providence Cancer Research Center, Portland, Oregon, USA, [102] Department of Medical Oncology, Istituto Europeo di Oncologia, Milan, Italy, [103] Department of Pathology, AZ Turnhout, Turnhout, Belgium, [104] Department of Pathology, St Vincent's University Hospital and University College Dublin, Dublin, Ireland, [105] Department of Computational and Systems Biology, University of Pittsburgh, USA, [106] Azienda AUSL, Regional Hospital of Aosta, Aosta, Italy, [107] Gustave Roussy, Universite Paris-Saclay, Villejuif, France; Université Paris-Sud, Institut National de la Santé et de la Recherche Médicale, Villejuif, France, [108] Department of Clinical Medicine, Macquarie University, Sydney, Australia, [109] HistoGeneX NV, Antwerp, Belgium and AZ Sint-Maarten Hospital, Mechelen, Belgium, [110] Oncology Clinical Development, Bristol-Myers Squibb, Princeton, USA, [111] Institut für Pathologie , UK Hamburg, Germany, [112] Department of Medicine, Vanderbilt University Medical Centre, Nashville, USA, [113] Department of Cancer Biology, Mayo Clinic, Jacksonville, USA, [114] Department of Oncology, Mayo Clinic, Rochester, USA, [115] Roche, Tucson, USA, [116] Tumor Pathology Department, Maria Sklodowska-Curie Memorial Cancer Center, Gliwice, Poland, [117] Pathology and Tissue Analytics, Pharma Research and Early Development, Roche Innovation Center Munich, Penzberg, Germany, [118] Sunnybrook Health Sciences Centre, Toronto, Canada, [119] Tehran University of Medical Sciences, Iran, [120] Oncology Biomarker Development, Genentech-Roche, [121] Division of Medical Oncology, Department of Medicine, University of Colorado Anschutz Medical Campus, Aurora, USA, [122] Department of Pathology, Hospital de Oncología Maria Curie, Buenos Aires, Argentina, [123] Directorate of Surgical Pathology, SA Pathology, Adelaide,

Australia, [124] Department of Pathology, University Hospital Antwerp, Belgium, [125] Translational Health Sciences, Department of Cellular Pathology, North Bristol NHS Trust, University of Bristol UK, [126] Medical Oncology, Department of Medicine, Cedars-Sinai Medical Center, Los Angeles, USA, [127] Helsinki University Central Hospital, Helsinki, Finland, [128] Department of Clinical Pathology, Akademiska University Hospital, Uppsala, Sweden, [129] International Agency for Research on Cancer (IARC), World Health Organization, Lyon, France, [130] Division of Molecular Pathology, The Netherlands Cancer Institute, Amsterdam, The Netherlands, [131] Department of Pathology, Sanatorio Mater Dei, Buenos Aires, Argentina, [132] Institute of Pathology, Medical University of Graz, Austria, [133] Department of Oncology, National Cancer Centre, Singapore, [134] R&D UNICANCER, Paris, France, [135] Translational Sciences, MedImmune, Gaithersberg, USA, [136] Breast Unit, Champalimaud Clinical Centre, Lisboa, Portugal, [137] Department of Pathology, Massachusetts General Hospital, Boston, MA, USA, [138] Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, USA, [139] Department of Oncology and Pathology, Karolinska Institutet and University Hospital, Solna, Sweden, [140] Department of Medicine, Clinical Division of Oncology, Comprehensive Cancer Centre Vienna, Medical University of Vienna, Vienna, Austria, [141] Leicester Cancer Research Centre, University of Leicester, Leicester, and MRC Toxicology Unit, University of Cambridge, UK, [142] Merck & Co., Inc., Kenilworth, USA, [143] GHI Le Raincy-Montfermeil, Chelles, Île-de-France, France, [144] Department of Pathology, Gustave Roussy, Grand Paris, France, [145] Departments of Medicine and Cancer Biology, Vanderbilt University Medical Centre, Nashville, USA, [146] Vm Scope, Germany, [147] Department of Pathology, Breast Pathology Section, Northwestern University, Chicago, USA, [148] Department of Biometrics, The Netherlands Cancer Institute, Amsterdam, The Netherlands, [149] Cancer Biomarkers Working Group, Faculty of Medicine

554 and Pharmacy, Université Mohamed Premier, Oujda, Morocco, [150] Pathology Department,
555 Stanford University Medical Centre, Stanford, USA, [151] Center for Pharmacogenomics and
556 Fudan-Zhangjiang, Center for Clinical Genomics School of Life Sciences and Shanghai Cancer
557 Center, Fudan University, China, [152] GROW - School for Oncology and Developmental Biology,
558 Maastricht University Medical Centre and Department of Pathology, Maastricht University
559 Medical Centre, Maastricht, The Netherlands, [153] Biorepository and Tissue Technology
560 Shared Resources, University of California San Diego, USA, [154] Department of Pathology,
561 Gustave Roussy, Villejuif, France, [155] Departments of Epigenetics and Molecular
562 Carcinogenesis, The University of Texas MD Anderson Cancer Center, Houston, TX,
563 USA, [156] Hospital das Clínicas, Sao Paulo, Brasil; Department of Pathology, Faculty of
564 Medicine, University of São Paulo, Sao Paulo, Brasil, [157] Department of Medical Biology and
565 Pathology, Gustave Roussy Cancer Campus, Villejuif, France, [158] Department of Pathology,
566 Academic Medical Center, Amsterdam, The Netherlands, [159] Institut Jules Bordet, Universite
567 Libre de Bruxelles, Brussels, Belgium, [160] Roche Tissue Diagnostics, Digital Pathology, Santa
568 Clara, CA, USA, [161] Department of Pathology, Microbiology and Immunology, Vanderbilt
569 University Medical Centre, Nashville, USA, [162] Division of Biostatistics, Dana-Farber Cancer
570 Institute, Boston, USA; Harvard Medical School, Boston, USA, [163] Roche Diagnostics
571 Information Solutions, Belmont, CA, USA, [164] Department of Anatomical Pathology, Royal
572 Melbourne Hospital, Parkville, Australia, [165] Vernon Cancer Center, Newton-Wellesley
573 Hospital, Newton, USA. , [166] Service de pathologique, Cliniques universitaires Saint-Luc,
574 Bruxelles, Belgique, [167] Breast Center, Dept. OB&GYN and CCC (LMU), University of Munich,
575 Munich, Germany, [168] Division of Hematology-Oncology, Beth Israel Deaconess Medical
576 Center, Boston, USA, [169] University of Leuven, Belgium. , [170] Department of Pathology,
577 Institut Jules Bordet, Université Libre de Bruxelles, Brussels, Belgium, [171] German Breast

Group GmbH, Germany, [172] Department of Surgical Pathology and Biopathology, Jean Perrin Comprehensive Cancer Centre, Clermont-Ferrand, France, [173] Johanniter GmbH - Evangelisches Krankenhaus Bethesda Mönchengladbach, West German Study Group, Mönchengladbach, Germany, [174] Molecular Oncology Group, Vall d'Hebron Institute of Oncology, Barcelona, Spain, [175] Department of Pathology, University of Marburg, Marburg, Germany, [176] Department of Histopathology, Manipal Hospitals Dwarka, New Delhi, India, [177] Department of Anatomical Pathology, St Vincent's Hospital Melbourne, Fitzroy, Australia, [178] Cancer Immunotherapy Trials Network, Central Laboratory and Program in Immunology, Fred Hutchinson Cancer Research Center, Seattle, USA, [179] Clinical Trial Service Unit & Epidemiological Studies Unit, University of Oxford, UK, [180] Department of Radiation Oncology, Odette Cancer Centre, Sunnybrook Research Institute, Toronto, Canada, [181] Oncology Merck & Co, New Jersey, USA, [182] The Cancer Research Program, Garvan Institute of Medical Research, Darlinghurst, Australian Clinical Labs, Australia, [183] Georgetown University Medical Center, Washington DC, USA, [184] Department of Molecular and Experimental Medicine, Avera Cancer Institute, Sioux Falls, SD, USA, [185] Translational Medicine, Bristol-Myers Squibb, Princeton, USA, [186] Anatomic Pathology, Boston, Massachusetts, USA, [187] Guy's Hospital, London, UK; King's College London, London, UK, [188] Peking University First Hospital Breast Disease Center, Beijing, China, [189] Dipartimento di Scienze della Salute (DSS), Firenze, Italy, [190] Department of Oncology, Champalimaud Clinical Centre, Lisbon, Portugal, [191] Department of Pathology, Fundación Valle del Lili, Cali, Colombia, [192] Department of Development and Regeneration, Laboratory of Experimental Urology, KU Leuven, Leuven, Belgium, [193] Department of Medical Oncology, Austin Health, Heidelberg, Australia, [194] Department of Surgery, Kansai Medical School, Hirakata, Japan, [195] Pathology Department, H.U. Vall d'Hebron, Barcelona,

Spain, [196] Division of Bioinformatics and Biostatistics, U.S. Food and Drug Administration, USA, [197] Department of Pathology and Laboratory Medicine, Rhode Island Hospital and Lifespan Medical Center, Providence, USA, [198] Université Paris-Est, Créteil, France, [199] Praava Health, Dhaka, Bangladesh, [200] Department of Pathology, Medical University of Vienna, Vienna, Austria

Acknowledgements / Funding

RS is supported by a grant from the Breast Cancer Research Foundation (BCRF, grant No. 17-194). SG is supported by Susan G Komen Foundation (CCR18547966) and a Young investigator Grant from Breast Cancer Alliance. TON receives funding support from the Canadian Cancer Society. AM acknowledges research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under award numbers 1U24CA199374-01, R01CA202752-01A1, R01CA208236-01A1, R01 CA216579-01A1, R01 CA220581-01A1, 1U01 CA239055-01, National Center for Research Resources under award number 1 C06 RR12463-01, VA Merit Review Award IBX004121A from the United States Department of Veterans Affairs Biomedical Laboratory Research and Development Service, the DOD Prostate Cancer Idea Development Award (W81XWH-15-1-0558), the DOD Lung Cancer Investigator-Initiated Translational Research Award (W81XWH-18-1-0440), the DOD Peer Reviewed Cancer Research Program (W81XWH-16-1-0329), the Ohio Third Frontier Technology Validation Fund and the Wallace H. Coulter Foundation Program in the Department of Biomedical Engineering and the Clinical and Translational Science Award Program (CTSA) at Case Western Reserve University. JS received funding from NCI grants UG3CA225021 and U24CA215109. CS is a Royal Society Napier Research Professor; this work was supported by the Francis Crick Institute that receives its core

626 funding from Cancer Research UK (FC001169, FC001202), the UK Medical Research Council
627 (FC001169, FC001202), and the Wellcome Trust (FC001169, FC001202); CS is also funded by
628 Cancer Research UK (TRACERx and CRUK Cancer Immunotherapy Catalyst Network), the
629 CRUK Lung Cancer Centre of Excellence, Stand Up 2 Cancer (SU2C), the Rosetrees Trust,
630 Butterfield and Stoneygate Trusts, NovoNordisk Foundation (ID16584), the Prostate Cancer
631 Foundation, the Breast Cancer Research Foundation (BCRF); the research leading to these
632 results has received funding from the European Research Council (ERC) under the European
633 Union's Seventh Framework Programme (FP7/2007-2013) Consolidator Grant (FP7-THESEUS-
634 617844), European Commission ITN (FP7-PloidyNet 607722), ERC Advanced Grant (PROTEUS)
635 has received funding from the European Research Council under the European Union's
636 Horizon 2020 research and innovation programme (grant agreement No. 835297),
637 Chromavision – this project has received funding from the European's Union Horizon 2020
638 research and innovation programme (grant agreement No. 665233); support was also
639 provided to CS by the National Institute for Health Research, the University College London
640 Hospitals Biomedical Research Centre, and the Cancer Research UK University College
641 London Experimental Cancer Medicine Centre. RK and KP-G acknowledge research leading to
642 or reported in this publication was supported by NCI U10CA180868, -180822, UG1-189867,
643 and U24-196067 the Breast Cancer Research Foundation and Genentech.
644

645 **Author Contributions**

646 All authors made a substantial contribution to the conception or design of the work
647 and/or the acquisition, analysis or interpretation of the data. All authors participated in
648 either drafting the work or revising it critically for important intellectual content. All authors

have approved the final completed version of this manuscript and assume accountability for all aspects of the work.

Competing Interests

AJL is a consultant for BMS, Merck, AZ/Medimmune, and Genentech. RS reports research funding from Roche, Puma, Merck; advisory board and consultancy for BMS; travel funding from Roche, Merck, and Astra Zeneca. SG reports Lab research funding from Lilly, Clinical research funding from Eli Lilly and Novartis and is a Paid advisor to Eli Lilly, Novartis, and G1 Therapeutics. JvdL is member of the scientific advisory boards of Philips, the Netherlands and ContextVision, Sweden and receives research funding from Philips, the Netherlands and Sectra, Sweden. SA reports Research funding to institution from Merck, Genentech, BMS, Novartis, Celgene and Amgen and is an uncompensated consultant /steering committee member for Merck, Genentech and BMS. TON has consulted for Nanostring and received compensation and has intellectual property rights / ownership interests from Bioclassifier LLC [not related to the subject material under consideration]. SL receives research funding to institution from Novartis, Bristol Meyers Squibb, Merck, Roche-Genentech, Puma Biotechnology, Pfizer and Eli Lilly, has acted as consultant (not compensated) to Seattle Genetics, Pfizer, Novartis, BMS, Merck, AstraZeneca and Roche-Genentech and acted as consultant (paid to her institution) to Aduro Biotech. SRL has received travel and educational funding from Roche/Ventana. AM is an equity holder in Elucid Bioimaging and in Inspirata Inc., a scientific advisory consultant for Inspirata Inc, has served as a scientific advisory board member for Inspirata Inc, Astrazeneca, Bristol Meyers-Squibb and Merck, has sponsored research agreements with Philips and Inspirata Inc, is involved in a NIH U24 grant with PathCore Inc, and 3 different R01 grants with Inspirata Inc.

and his technology has been licensed to Elucid Bioimaging and Inspirata Inc. GC is on the advisory boards of Roche, BMS, Pfizer, Seattle Genetics and Ellipsis, and reports personal fees from Roche, BMS, Pfizer, Seattle Genetics, and Ellipsis, outside of the submitted work. JH is the director and owner of Vivactiv Ltd. JH is the director and owner of Slide Score B.V. FPL reports funding from Astrazeneca, BMS, Roche, MSD, Pfizer, Novartis, Sanofi, Eli Lilly. JB reports consultancies from Insight Genetics, BioNTech AG, Biotheranostics, Pfizer, RNA Diagnostics and OncoXchange, research funding from Thermo Fisher Scientific, Genoptix, Agendia, NanoString Technologies, Stratifyer GmbH and Biotheranostics, applied for patents, including Jan 2017: Methods and Devices for Predicting Anthracycline Treatment Efficacy, US utility – 15/325,472; EPO – 15822898.1; Canada – not yet assigned; Jan 2017: Systems, Devices and Methods for Constructing and Using a Biomarker, US utility– 15/328,108; EPO – 15824751.0; Canada – not yet assigned; Oct 2016: Histone gene module predicts anthracycline benefit, PCT/CA2016/000247; Dec 2016: 95-Gene Signature of Residual Risk Following Endocrine Treatment, PCT/CA2016/000304; Dec 2016: Immune Gene Signature Predicts Anthracycline Benefit, PCT/CA2016/000305. MAS reports consulting work for Achilles Therapeutics. CS reports receipt of grants/research support from Pfizer, AstraZeneca, BMS and Ventana; receipt of honoraria, consultancy, or SAB Member fees from Pfizer, Novartis, GlaxoSmithKline, MSD, BMS, Celgene, AstraZeneca, Illumina, Sarah Canon Research Institute, Genentech, Roche-Ventana, GRAIL, Medicxi; Advisor for Dynamo Therapeutics; Stock shareholder in Apogen Biotechnologies, Epic Bioscience, GRAIL; Co-Founder & stock options in Achilles Therapeutics. AHB is the co-founder and CEO of PathAI. JK is an employee of PathAI. DD is on the advisory board for Oncology Analytics, Inc, and a consultant for Novartis. DLR is on the advisory board of Amgen, Astra Zeneca, Cell Signaling Technology, Cepheid, Daiichi Sankyo, GSK, Konica/Minolta, Merck, NanoString, Perkin Elmer,

Ventana, Ultivue; receives research support from Astra Zeneca, Cepheid, Navigate BioPharma, NextCure, Lilly, Ultivue; instrument support from Ventana, Akoya/Perkin Elmer, NanoString; paid consultant for Biocept; received travel honoraria from BMS, founder and equity holder for PixelGear and received royalty from Rarecyte. AT reports benefits from ICR's Inventors Scheme associated with patents for one of PARP inhibitors in BRCA1/2 associated cancers, as well as honoraria from Pfizer, Vertex, Prime Oncology, Artios, honoraria and stock in InBioMotion, honoraria and financial support for research from AstraZeneca, Medivation, Myriad Genetics and Merck Serono.

This work includes contributions from, and was reviewed by, individuals at the FDA. This work has been approved for publication by the agency, but it does not necessarily reflect official agency policy. Certain commercial materials and equipment are identified in order to adequately specify experimental procedures. In no case does such identification imply recommendation or endorsement by the FDA, nor does it imply that the items identified are necessarily the best available for the purpose.

This work includes contributions from, and was reviewed by, individuals who received funding from the National Institutes of Health, the U.S. Department of Veterans Affairs and the Department of Defense. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, the U.S. Department of Veterans Affairs, the Department of Defense, or the United States Government.

References

1. Savas P, *et al*. Clinical relevance of host immunity in breast cancer: from TILs to the clinic. Nat Rev Clin Oncol. Apr;13(4):228–41. (2016)

2. Hammerl D, *et al*. Breast cancer genomics and immuno-oncological markers to guide immune therapies. Semin Cancer Biol. Oct 1;52:178–88. (2018)
3. Hudecek J, *et al*. The path to a better biomarker: application of a risk management framework for the implementation of PD-L1 and TILs as immuno-oncology biomarkers into breast cancer clinical trials and daily practice. A report from the International Immuno-Oncology Biomarker Working Group. Npj Breast Cancer. co-submitted article.
4. Adams S, *et al*. Prognostic value of tumor-infiltrating lymphocytes in triple-negative breast cancers from two phase III randomized adjuvant breast cancer trials: ECOG 2197 and ECOG 1199. J Clin Oncol. Sep 20;32(27):2959–66. (2014)
5. Loi S, *et al*. Tumor infiltrating lymphocytes are prognostic in triple negative breast cancer and predictive for trastuzumab benefit in early breast cancer: results from the FinHER trial. Ann Oncol. Aug;25(8):1544–50. (2014)
6. Loi S, *et al*. Prognostic and predictive value of tumor-infiltrating lymphocytes in a phase III randomized adjuvant breast cancer trial in node-positive breast cancer comparing the addition of docetaxel to doxorubicin with doxorubicin-based chemotherapy: BIG 02-98. J Clin Oncol. Mar 1;31(7):860–7. (2013)
7. Denkert C, *et al*. Tumor-associated lymphocytes as an independent predictor of response to neoadjuvant chemotherapy in breast cancer. J Clin Oncol. Jan 1;28(1):105–13. (2010)
8. Salgado R, *et al*. The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014. Ann Oncol. Feb;26(2):259–71. (2015)
9. Loi S, *et al*. Tumor-Infiltrating Lymphocytes and Prognosis: A Pooled Individual Patient Analysis of Early-Stage Triple-Negative Breast Cancers. J Clin Oncol. Mar 1;37(7):559–69. (2019)
10. Dieci MV, *et al*. Prognostic and predictive value of tumor-infiltrating lymphocytes in two phase III randomized adjuvant breast cancer trials. Ann Oncol. Aug;26(8):1698–704. (2015)

11. Denkert C, *et al*. Tumour-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: a pooled analysis of 3771 patients treated with neoadjuvant therapy. Lancet Oncol. Jan 1;19(1):40–50. (2018)
12. Denkert C, *et al*. Tumor-infiltrating lymphocytes and response to neoadjuvant chemotherapy with or without carboplatin in human epidermal growth factor receptor 2-positive and triple-negative primary breast cancers. J Clin Oncol. Mar 20;33(9):983–91. (2015)
13. Issa-Nummer Y, *et al*. Prospective validation of immunological infiltrate for prediction of response to neoadjuvant chemotherapy in HER2-negative breast cancer--a substudy of the neoadjuvant GeparQuinto trial. PloS One. Dec 2;8(12):e79775. (2013)
14. West NR, *et al*. Tumor-infiltrating lymphocytes predict response to anthracycline-based chemotherapy in estrogen receptor-negative breast cancer. Breast Cancer Res. 13(6):R126. (2011)
15. Burstein HJ, *et al*. Estimating the Benefits of Therapy for Early Stage Breast Cancer The St Gallen International Consensus Guidelines for the Primary Therapy of Early Breast Cancer 2019. Ann Oncol. Oct 1;30(10):1541-1557. (2019)
16. Amgad M, *et al*. Key issues in computational assessment of Tumor Infiltrating Lymphocytes in solid tumors. A report from the International Immuno-Oncology Biomarker Working group. Npj Breast Cancer. co-submitted article.
17. Huang J, *et al*. Changes of Tumor Infiltrating Lymphocytes after Core Needle Biopsy and the Prognostic Implications in Early Stage Breast Cancer: A Retrospective Study. Cancer Res Treat. Oct;51(4):1336-1346. (2019)
18. Cha YJ, *et al*. Comparison of tumor-infiltrating lymphocytes of breast cancer in core needle biopsies and resected specimens: a retrospective analysis. Breast Cancer Res Treat. Sep;171(2):295–302. (2018)

19. Luen SJ, *et al*. Prognostic implications of residual disease tumor-infiltrating lymphocytes and residual cancer burden in triple-negative breast cancer patients after neoadjuvant chemotherapy. Ann Oncol. Feb 1;30(2):236–42. (2019)
20. Luen SL, Salgado R, Loi S. Residual disease and immune infiltration as a new surrogate endpoint for TNBC post neoadjuvant chemotherapy. Oncotarget. Jul 23;10(45):4612–4. (2019)
21. Dieci MV, *et al*. Update on tumor-infiltrating lymphocytes (TILs) in breast cancer, including recommendations to assess TILs in residual disease after neoadjuvant therapy and in carcinoma in situ: A report of the International Immuno-Oncology Biomarker Working Group on Breast Cancer. Semin Cancer Biol. Oct;52(Pt 2):16–25. (2018)
22. Denkert C, *et al*. Standardized evaluation of tumor-infiltrating lymphocytes in breast cancer: results of the ring studies of the international immuno-oncology biomarker working group. Mod Pathol. Oct;29(10):1155–64. (2016)
23. Kim RS, *et al*. Stromal Tumor-infiltrating Lymphocytes in NRG Oncology/NSABP B-31 Adjuvant Trial for Early-stage HER2-positive Breast Cancer. J Natl Cancer Inst. Mar 19; pii: djz032. (2019) [Epub ahead of print]
24. Boyce BF. An update on the validation of whole slide imaging systems following FDA approval of a system for a routine pathology diagnostic service in the United States. Biotech Histochem. 92(6):381–9. (2017)
25. Loi S. Host antitumor immunity plays a role in the survival of patients with newly diagnosed triple-negative breast cancer. J Clin Oncol. Sep 20;32(27):2935–7. (2014)
26. Gavrielides MA, Conway C, O'Flaherty N, Gallas BD, Hewitt SM. Observer performance in the use of digital and optical microscopy for the interpretation of tissue-based biomarkers. Anal Cell Pathol (Amst). 2014:157308. (2014)
27. Gavrielides MA, Gallas BD, Lenz P, Badano A, Hewitt SM. Observer variability in the interpretation of HER2/neu immunohistochemical expression with unaided and computer-aided digital microscopy. Arch Pathol Lab Med. Feb;135(2):233–42. (2011)

28. Simon RM, Paik S, Hayes DF. Use of archived specimens in evaluation of prognostic and predictive biomarkers. J Natl Cancer Inst. Nov 4;101(21):1446–52. (2009)
29. Hayes DF, *et al*. Tumor marker utility grading system: a framework to evaluate clinical utility of tumor markers. J Natl Cancer Inst. Oct 16;88(20):1456–66. (1996)
30. Park JH, *et al*. Prognostic value of tumor-infiltrating lymphocytes in patients with early-stage triple-negative breast cancers (TNBC) who did not receive adjuvant chemotherapy. Ann Oncol. Sep 30. pii: mdz395. (2019) [Epub ahead of print]
31. Esteva FJ, Hubbard-Lucey VM, Tang J, Pusztai L. Immunotherapy and targeted therapy combinations in metastatic breast cancer. Lancet Oncol. Mar;20(3):e175–86. (2019)
32. Gong J, Chehrazi-Raffle A, Reddi S, Salgia R. Development of PD-1 and PD-L1 inhibitors as a form of cancer immunotherapy: a comprehensive review of registration trials and future considerations. J Immunother Cancer. Jan 23;6(1):8. (2018)
33. Balar AV, Weber JS. PD-1 and PD-L1 antibodies in cancer: current status and future directions. Cancer Immunol Immunother. May;66(5):551–64. (2017)
34. Hirsch FR *et al*. PD-L1 Immunohistochemistry Assays for Lung Cancer: Results from Phase 1 of the Blueprint PD-L1 IHC Assay Comparison Project. J Thorac Oncol. Feb;12(2):208–22. (2017)
35. Tsao MS, *et al*. PD-L1 Immunohistochemistry Comparability Study in Real-Life Clinical Samples: Results of Blueprint Phase 2 Project. J Thorac Oncol. Sep;13(9):1302–11. (2018)
36. Rimm DL, *et al*. A Prospective, Multi-Institutional Assessment of Four Assays for PD-L1 Expression in NSCLC by Immunohistochemistry. JAMA Oncol. Aug 1;3(8):1051–8. (2017)
37. Loi S, *et al*. Abstract GS2-06: Phase Ib/II study evaluating safety and efficacy of pembrolizumab and trastuzumab in patients with trastuzumab-resistant HER2-positive metastatic breast cancer: Results from the PANACEA (IBCSG 45-13/BIG 4-13/KEYNOTE-014) study. Cancer Res. 78(4 Suppl):Abstract nr GS2-06. (2018)

38. Loi S, *et al*. LBA13Relationship between tumor infiltrating lymphocyte (TIL) levels and response to pembrolizumab (pembro) in metastatic triple-negative breast cancer (mTNBC): Results from KEYNOTE-086. Ann Oncol. Sep 1;28. (2017)
39. Corredor G, *et al*. Spatial Architecture and Arrangement of Tumor-Infiltrating Lymphocytes for Predicting Likelihood of Recurrence in Early-Stage Non-Small Cell Lung Cancer. Clin Cancer Res. Mar 1;25(5):1526–34. (2019)
40. Bera K, Velcheti V, Madabhushi A. Novel Quantitative Imaging for Predicting Response to Therapy: Techniques and Clinical Applications. Am Soc Clin Oncol. May 23;38:1008–18. (2018)
41. Saltz J, *et al*. Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images. Cell Rep. Apr 3;23(1):181-193.e7. (2018)
42. Hida AI, *et al*. Diffuse distribution of tumor-infiltrating lymphocytes is a marker for better prognosis and chemotherapeutic effect in triple-negative breast cancer. Breast Cancer Res Treat. Nov;178(2):283–94. (2019)
43. Klauschen F, *et al*. Scoring of tumor-infiltrating lymphocytes: From visual estimation to machine learning. Semin Cancer Biol.;52(Pt 2):151–7. (2018)
44. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. Jun;15(2):155–63. (2016)
45. Fleiss JL, Shrout PE. Approximate interval estimation for a certain intraclass correlation coefficient. Psychometrika 43: 259. (1978)
46. Kos, Zuzana; Roblin, Elvire; Kim, Rim; Michiels, Stefan; Gallas, Brandon D.; Chen, Weijie; et al.: Metadata supporting data files in the published article: Pitfalls in Assessing Stromal Tumor Infiltrating Lymphocytes (sTILs) in Breast Cancer. *figshare* https://doi.org/10.6084/m9.figshare.11907768 (2020)

Figure Legends

Figure 1. Study flow diagram. Raw data and original scanned images from 3 previously performed ring studies were evaluated (shaded box).

Figure 2. Reference images representing percent sTIL scores. Available at www.tilsinbreastcancer.org.

Figure 3. Standard deviation as a function of mean across all sTILs scores for each slide in 3 ring studies assessing concordance amongst pathologists. (a) Ring study 1, 32 pathologists evaluated 60 scanned core biopsy specimens. (b) Ring study 2, 28 pathologists evaluated 60 scanned core biopsy specimens. (c) Ring study 3, 6 pathologists evaluated 100 scanned whole section specimens. 10% of cases in each study showing the greatest variability in sTIL scores are shown as red squares. Black triangles identify additional cases identified for slide assessment.

Figure 4. Examples of heterogeneity in sTIL distribution as a cause of variation in sTIL assessment in breast cancer, including (a) increased sTILs at the leading edge (blue arrow) compared to the central tumor (yellow arrow); (b) marked heterogeneity in sTIL density within the tumor; and (c) variably spaced apart clusters of cancer cells with a dense tight lymphocytic infiltrate separated by collagenous stroma with sparse infiltrate.

Figure 5. Examples of technical factors as a cause of variation in sTIL assessment in breast cancer, including (a) a poor quality slide as can be seen secondary to prolonged ischemic time, poor fixation or issues during processing; (b) crush artifact ; and (c) out-of-focus scan.

Figure 6. Examples of scoring the wrong area as a cause of variation in sTIL assessment in breast cancer, including (a) difficulty defining the tumor boundary (dashed line) and including fibrous scars (yellow arrow) or lymphoid aggregates (blue arrow) beyond the invasive front; (b) including lymphocytes surrounding ductal carcinoma in situ (DCIS) which may be difficult to distinguish from invasive carcinoma; (c) including lymphocytes associated with an encapsulated papillary carcinoma component of a tumor; and (d) including lymphocytes surrounding benign glands. Shown is invasive carcinoma (yellow arrows) surrounding a benign lobule with associated lymphocytes; adjacent benign lobules (blue arrows) show dense lymphoid aggregates identify the lymphocytic infiltrate to be related to the entrapped lobule rather than the carcinoma.

Figure 7. Examples of scoring the wrong cells as a cause of variation in sTIL assessment in breast cancer, including (a) counting intratumoral TILs (iTILS) ; (b) counting neutrophils; (c) counting histiocytes; (d) misinterpreting apoptotic cells as lymphocytes; and (e) artifactual falling apart of cells mimicking TILs.

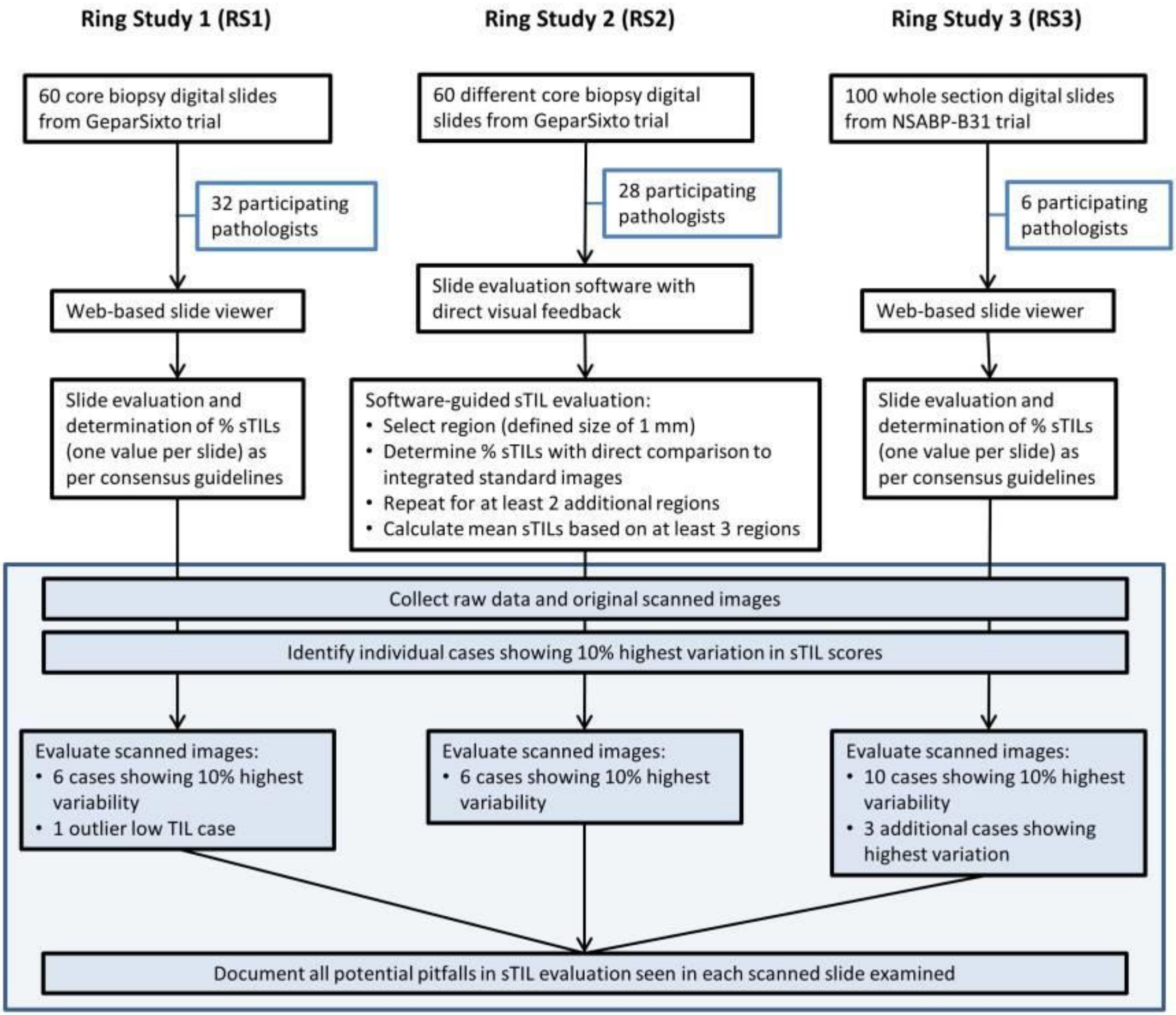
Figure 8. Examples of limited stroma within tumors as a cause of variation in sTIL assessment in breast cancer, including (a) tumor with small volume of intratumoral stroma present for evaluation ; (b) large areas of necrosis which decrease scorable stromal component; and (c) mucinous tumors.

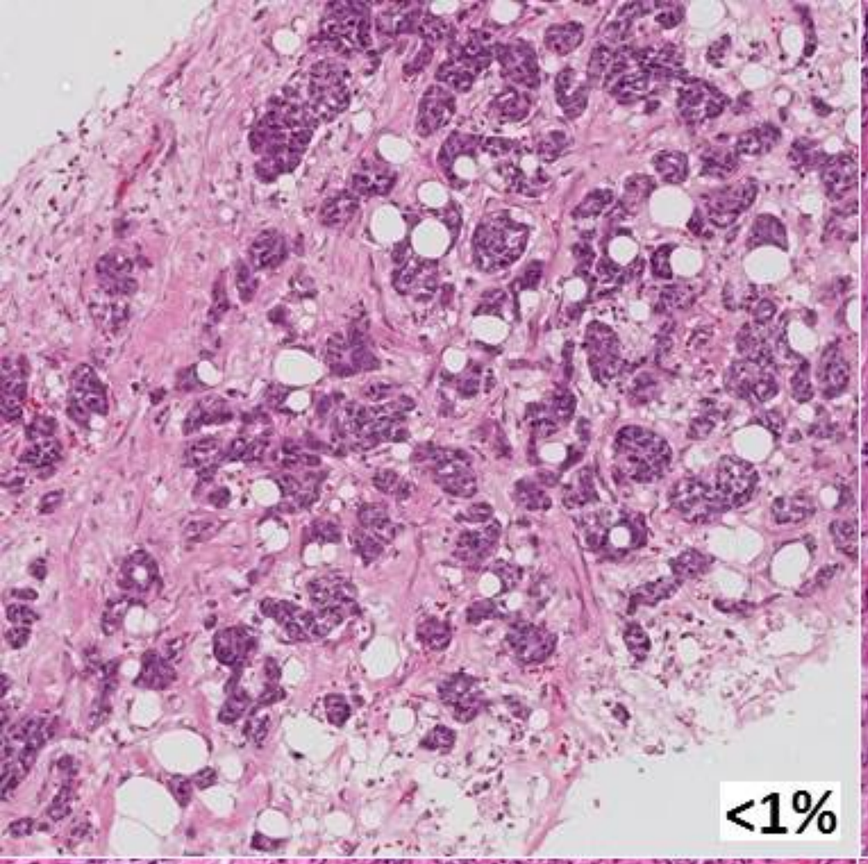
Figure 9. Variation in estimated outcome based on stromal TIL assessment for a 60-year-old patient with a histological grade 3 tumor, 2-5 cm in size and receiving anthracycline + taxane

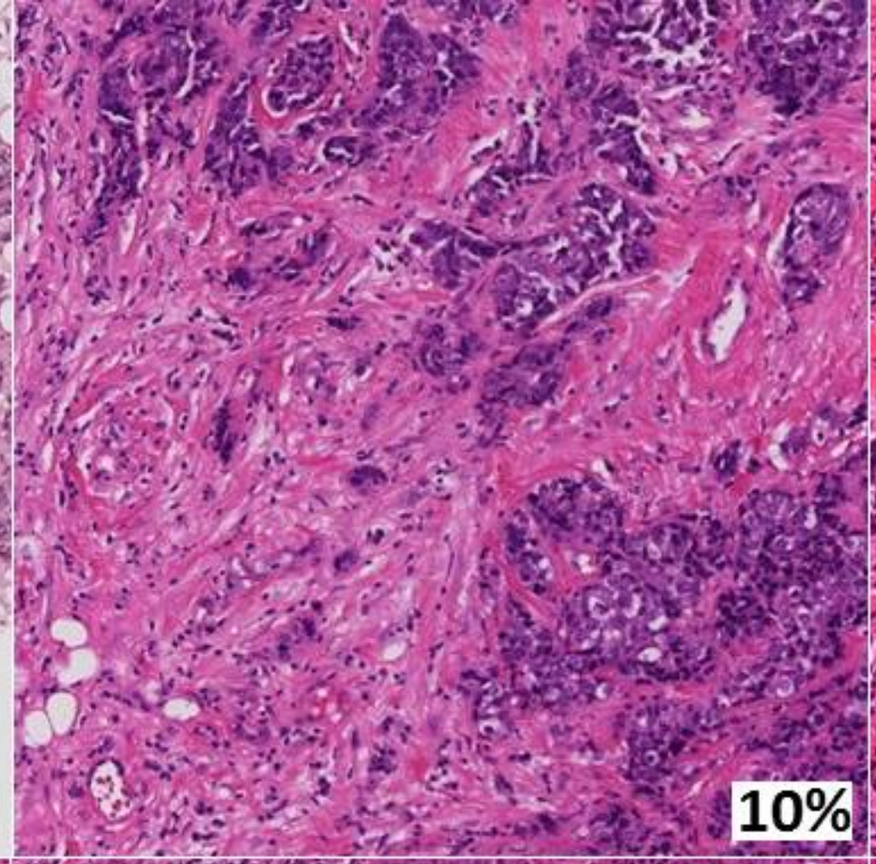
based chemotherapy. Presuming a true value for sTILs of 30%, changes in estimated 5-year iDFS for 5, 10 and 20% deviations (increase and decrease) in sTIL assessments are represented with 95% confidence bands. (All calculations were performed using the online triple-negative breast cancer (TNBC)-prognosis tool[9] available at www.tilsinbreastcancer.org).

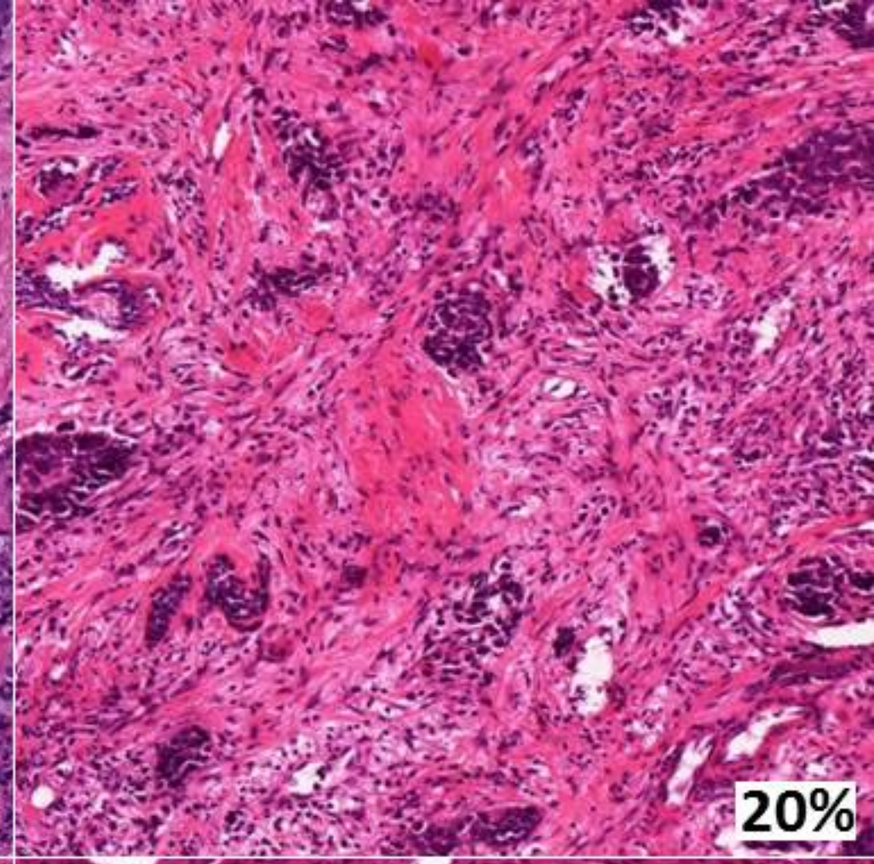
Box 1. ***Key Points***

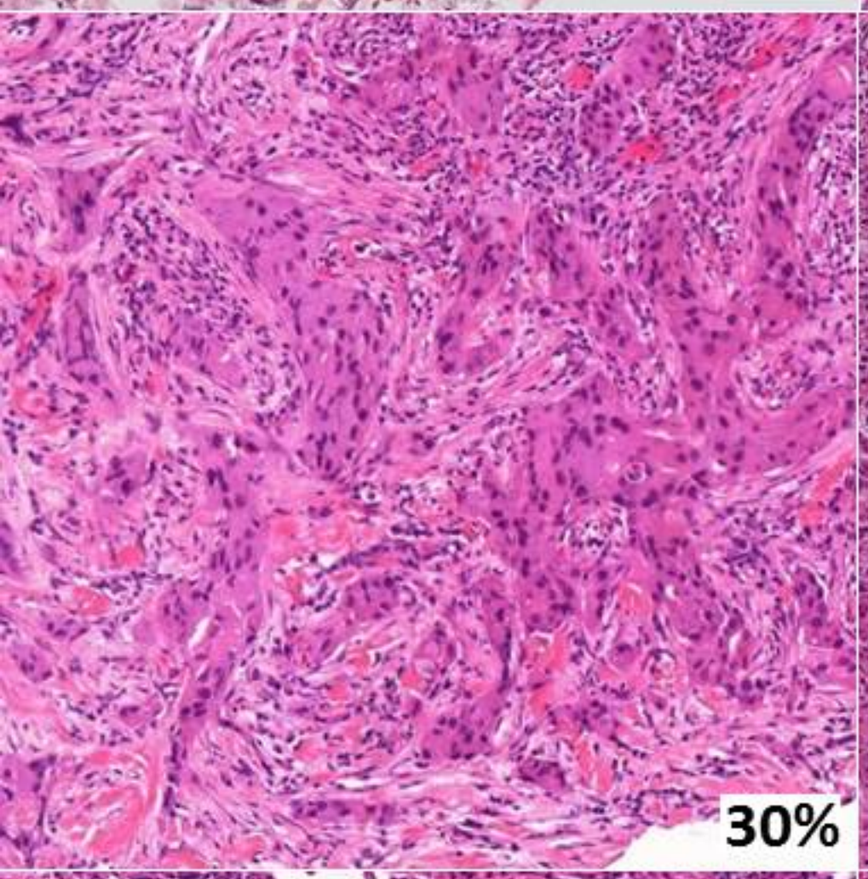
- Stromal TILs are mononuclear cells (predominantly lymphocytes) present *within* the boundary of a tumor that are located within the stroma *between* carcinoma cells without directly contacting the carcinoma cell nests
- Heterogeneity in sTIL distribution is the main contributing factor to variability in assessment
- Two key factors improve consistency of sTIL results:
  - Scoring multiple areas in heterogeneous tumors and averaging results
  - Use of reference images
- Poor sample processing or fixation can increase histological artefacts and compromise assessment of sTILs
- Careful adherence to the definition and morphology of sTILs is required to avoid scoring stromal areas outside of the tumor boundary and mistaken classification of artefacts, mitotic bodies etc as sTILs

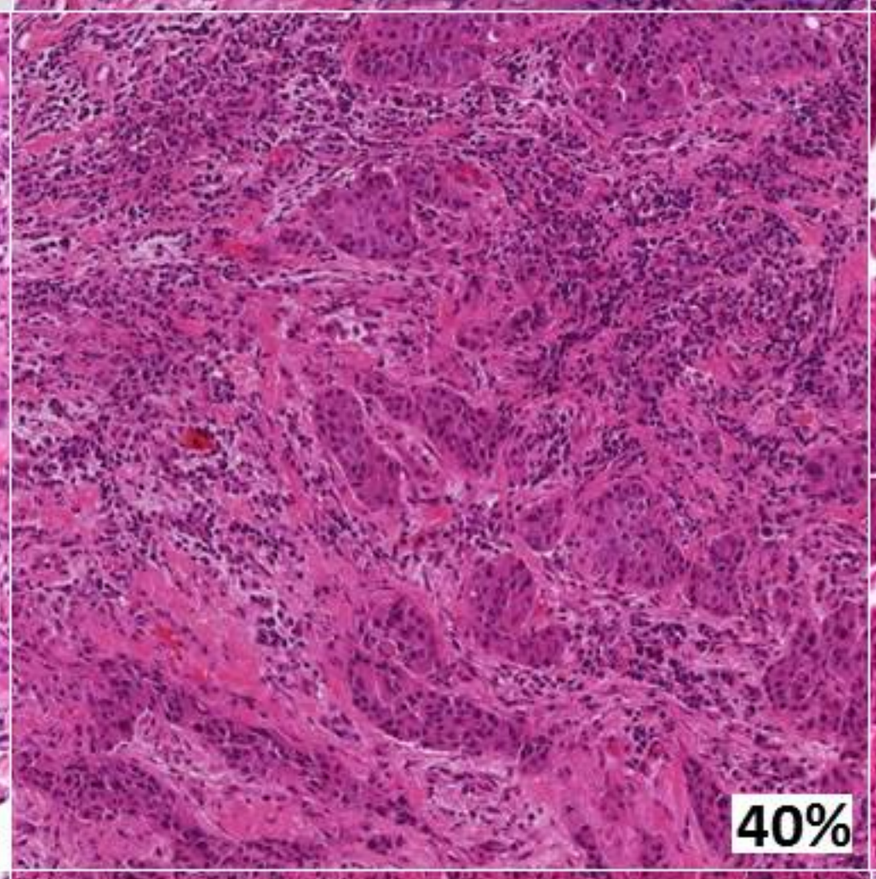

Ring Study 1 (RS1)
Ring Study 2 (RS2)
Ring Study 3 (RS3)
60 core biopsy digital slides from GeparSixto trial
32 participating pathologists
60 different core biopsy digital slides from GeparSixto trial
28 participating pathologists
100 whole section digital slides from NSABP-B31 trial
6 participating pathologists
Web-based slide viewer
Slide evaluation software with direct visual feedback
Web-based slide viewer
Slide evaluation and determination of % sTILs (one value per slide) as per consensus guidelines
Software-guided sTIL evaluation:
• Select region (defined size of 1 mm)
• Determine % sTILs with direct comparison to integrated standard images
• Repeat for at least 2 additional regions
• Calculate mean sTILs based on at least 3 regions
Slide evaluation and determination of % sTILs (one value per slide) as per consensus guidelines
Collect raw data and original scanned images
Identify individual cases showing 10% highest variation in sTIL scores
Evaluate scanned images:
• 6 cases showing 10% highest variability
• 1 outlier low TIL case
Evaluate scanned images:
• 6 cases showing 10% highest variability
Evaluate scanned images:
• 10 cases showing 10% highest variability
• 3 additional cases showing highest variation
Document all potential pitfalls in sTIL evaluation seen in each scanned slide examined

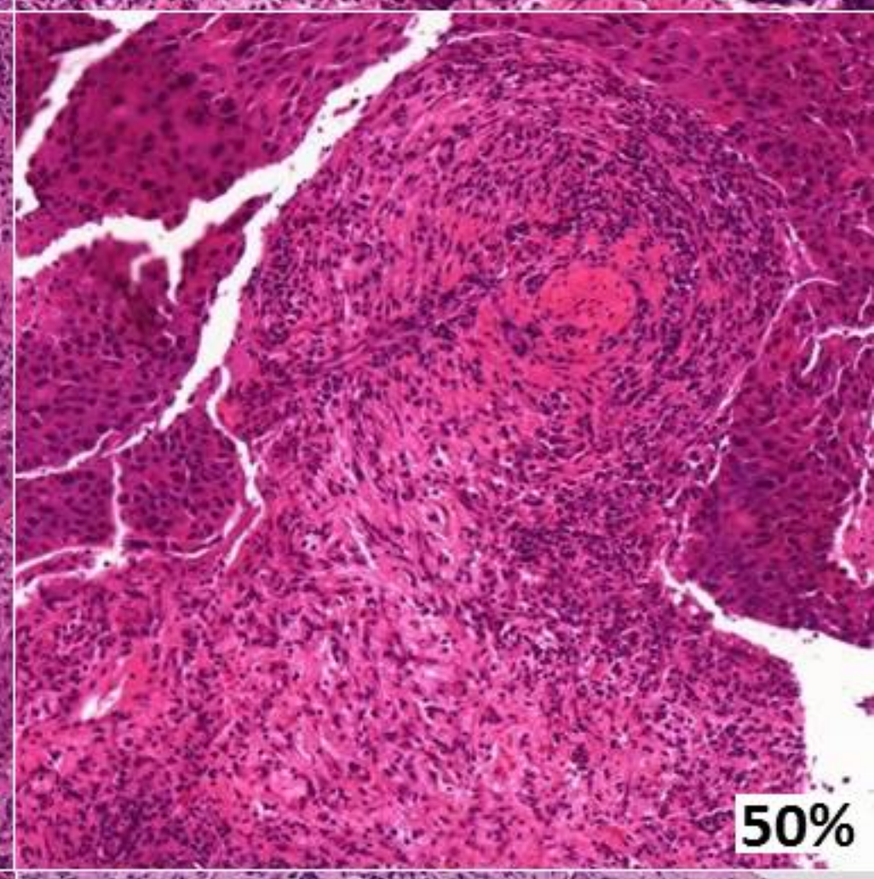

<1%

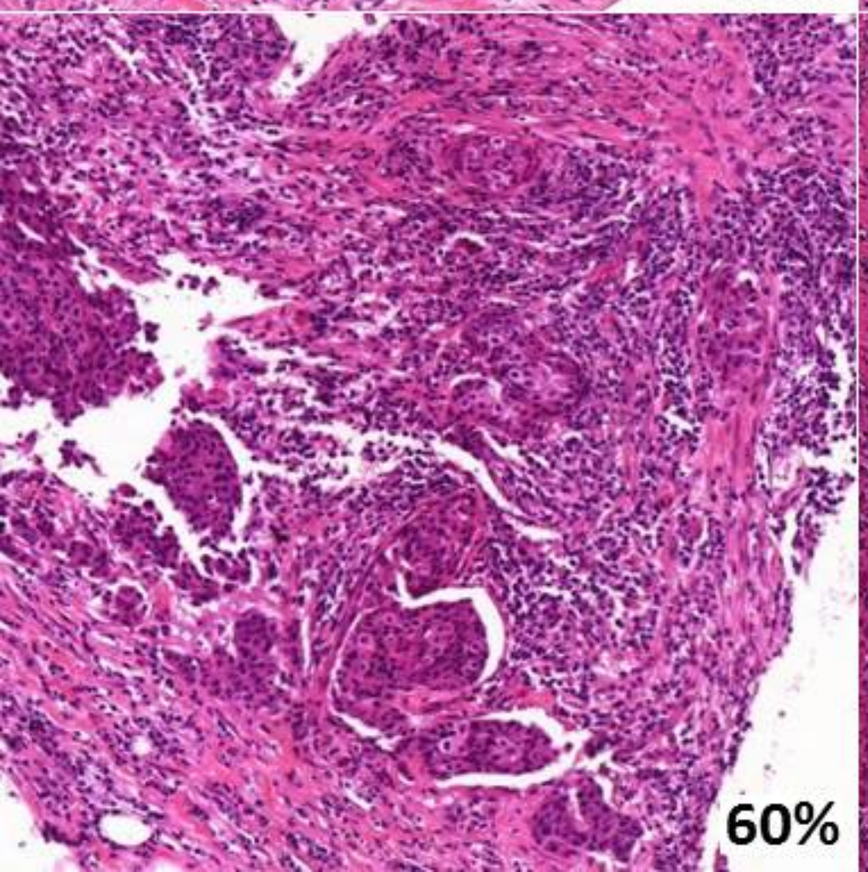

10%

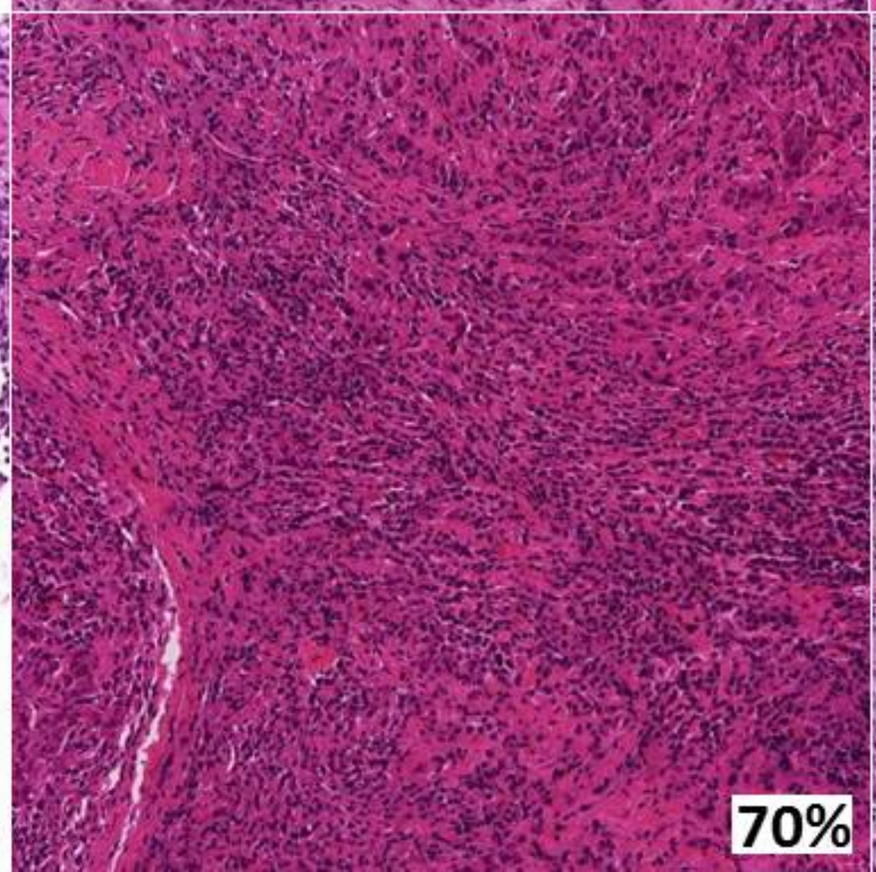

20%

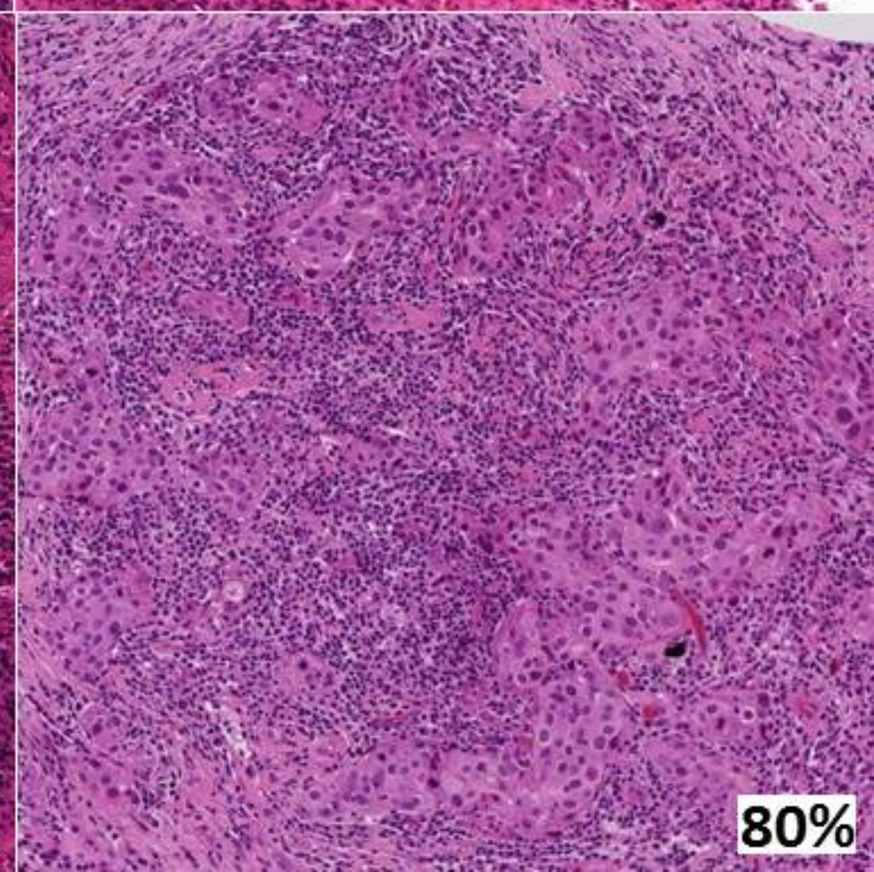

30%


40%

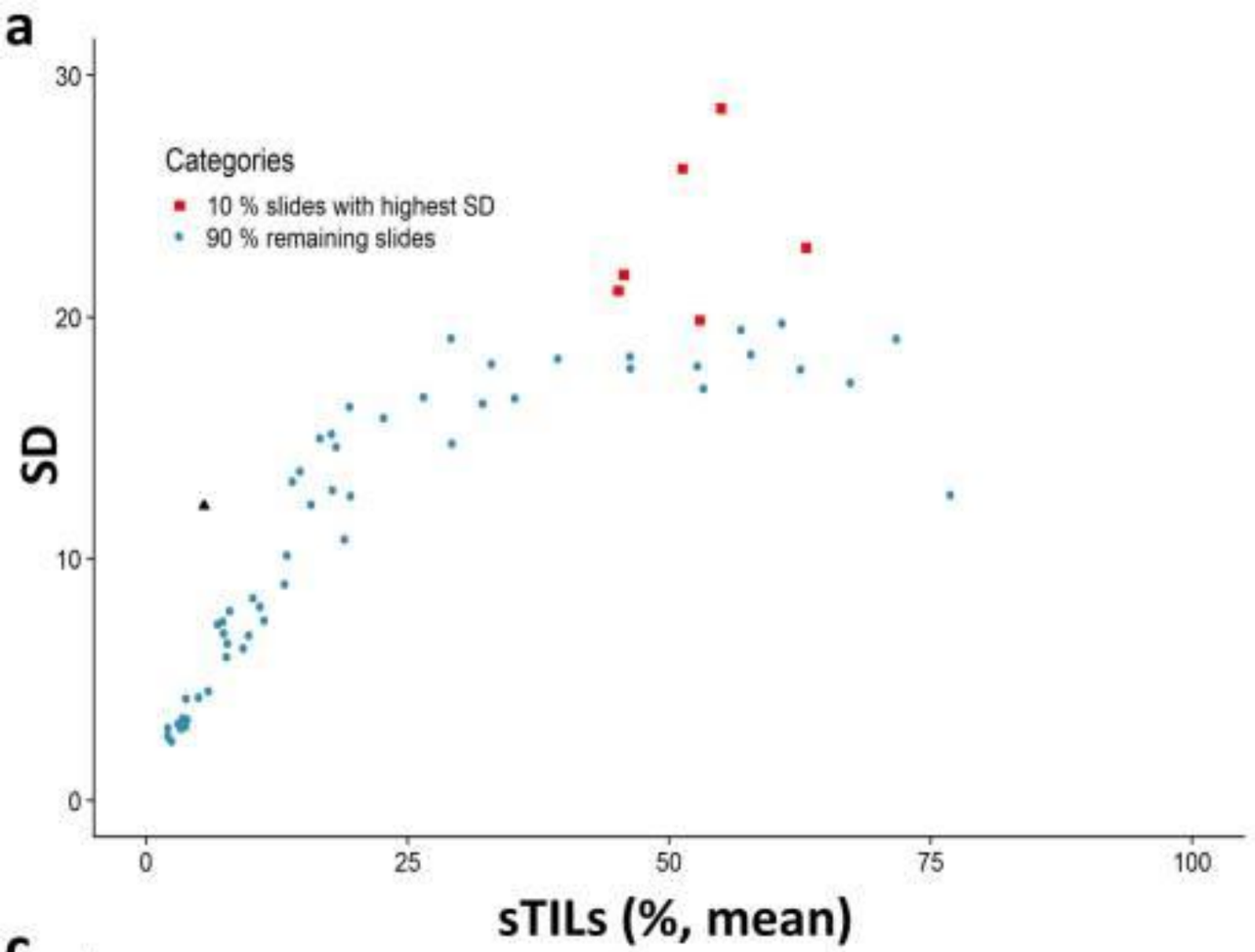

50%


60%

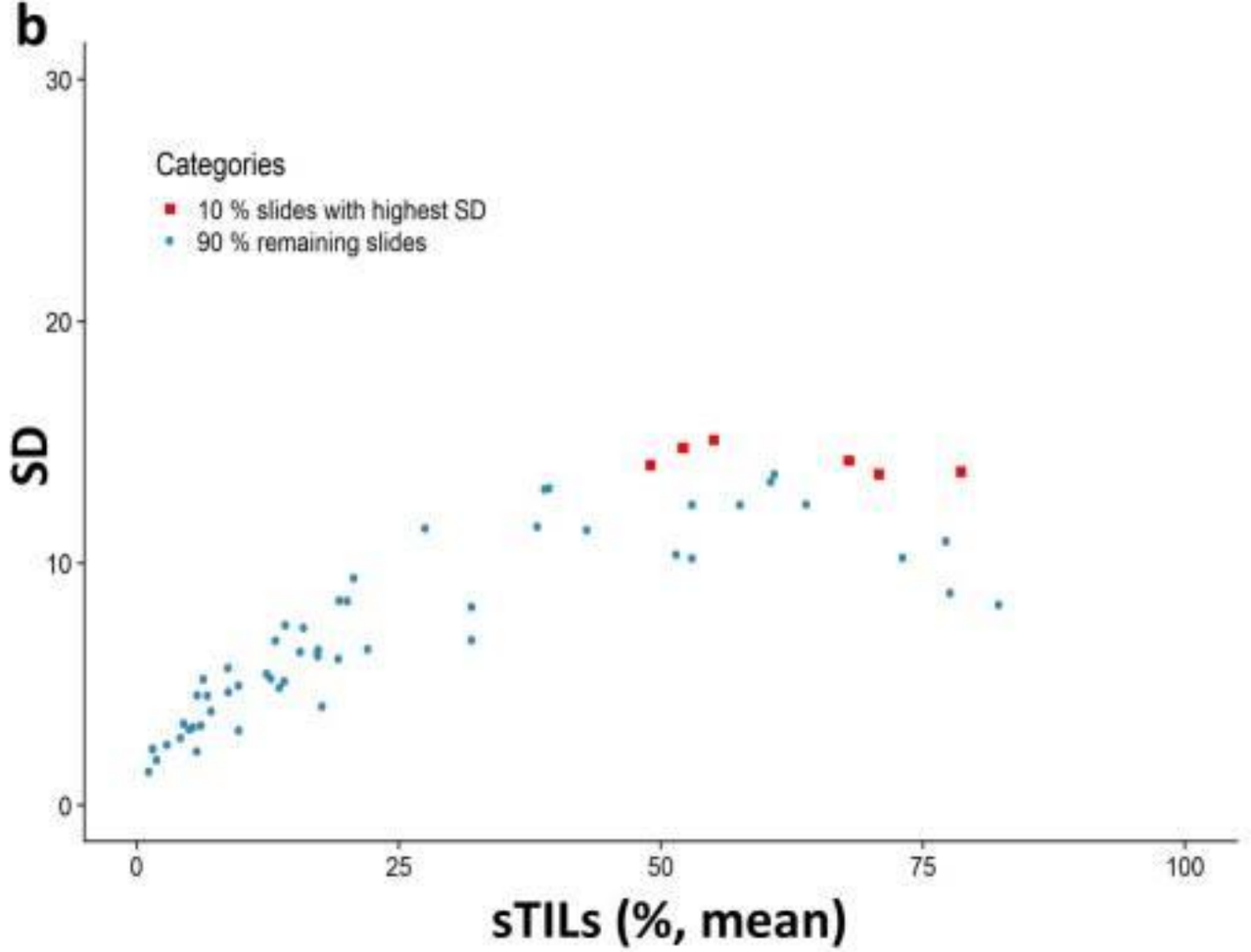

70%


80%

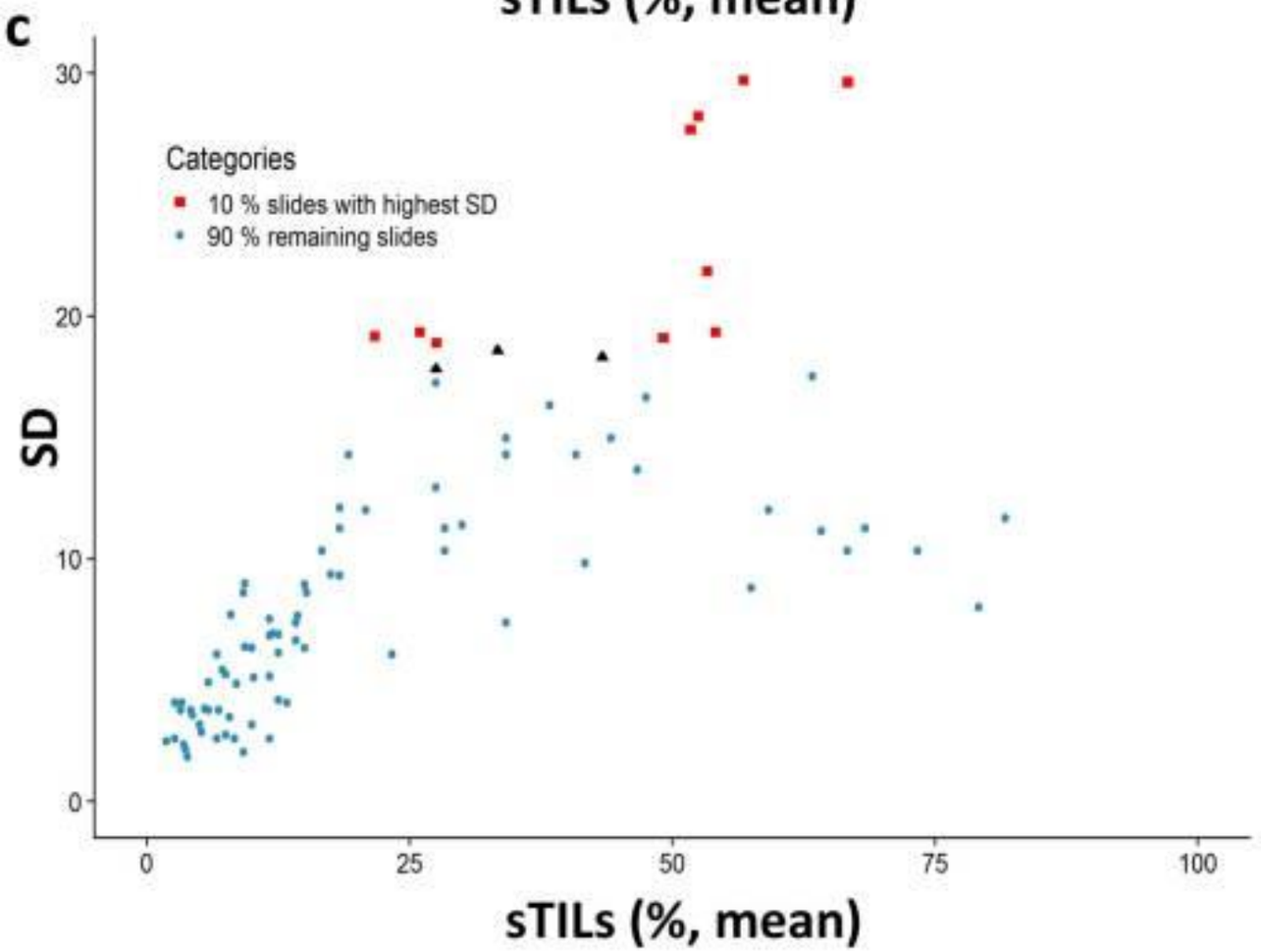
a

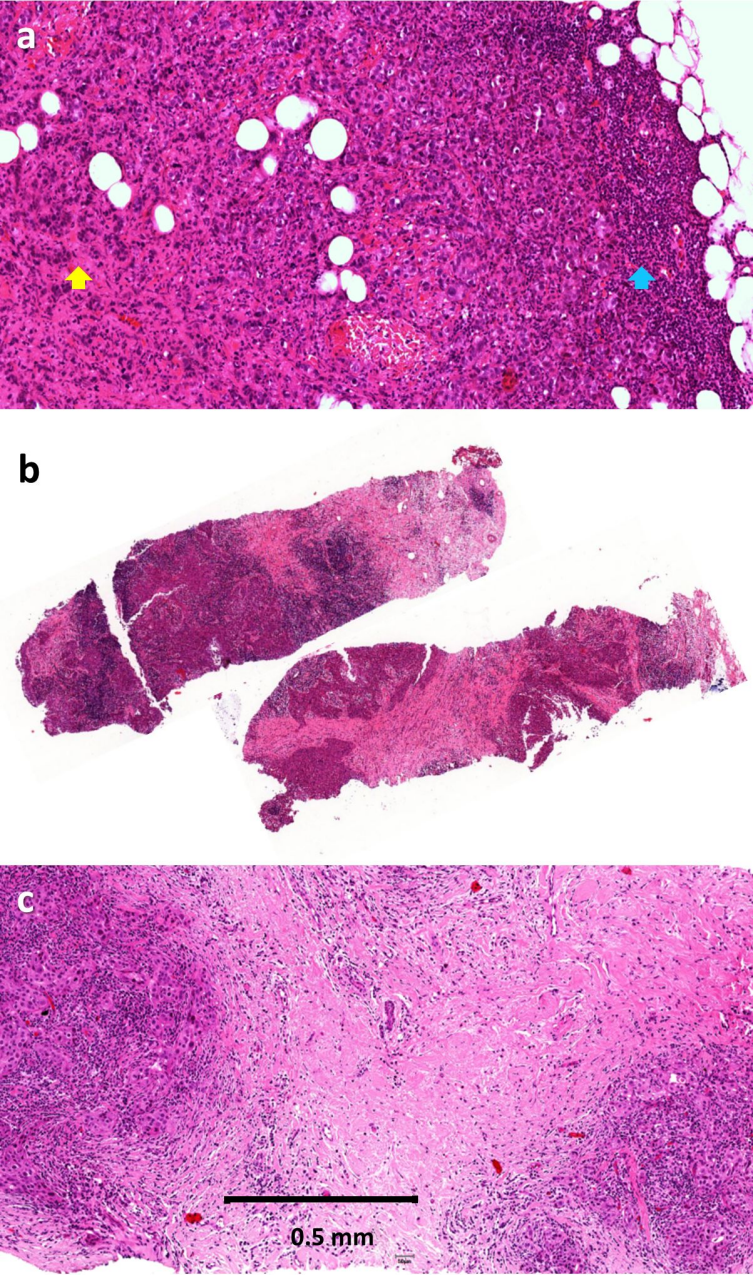

Categories
10 % slides with highest SD
90 % remaining slides
SD
sTILs (%, mean)

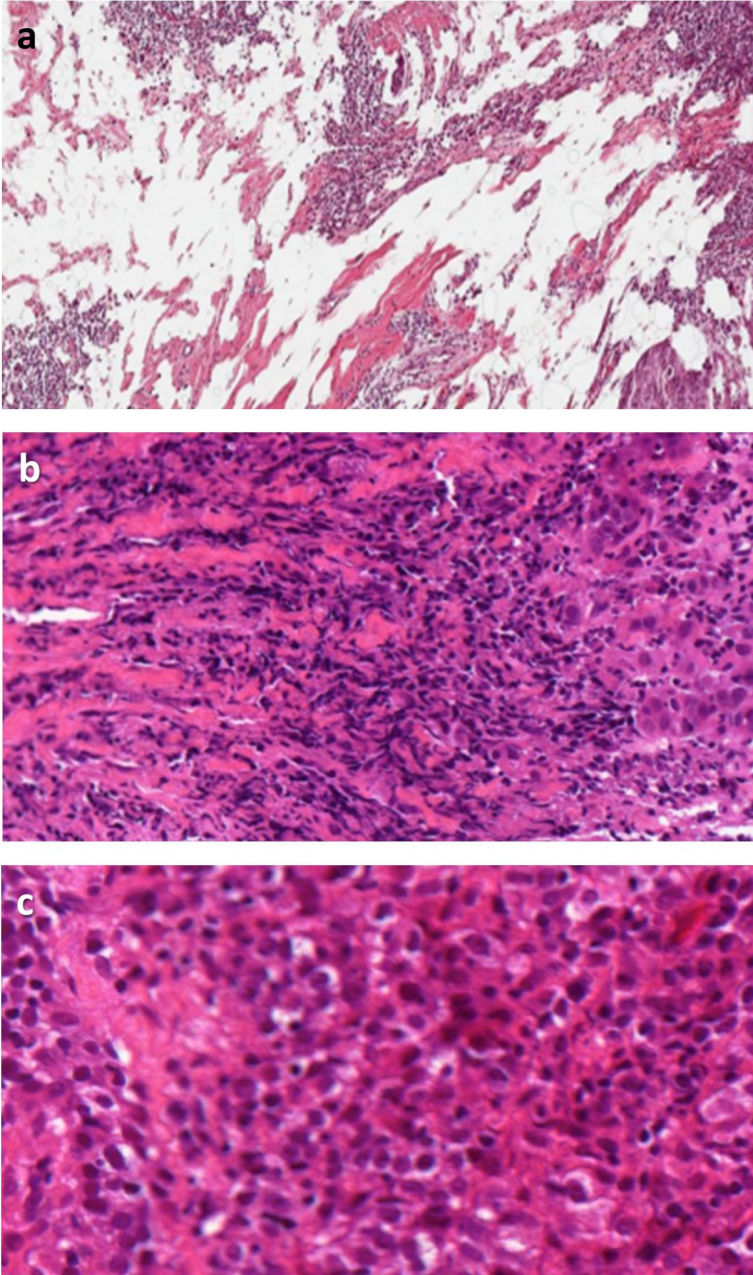
b

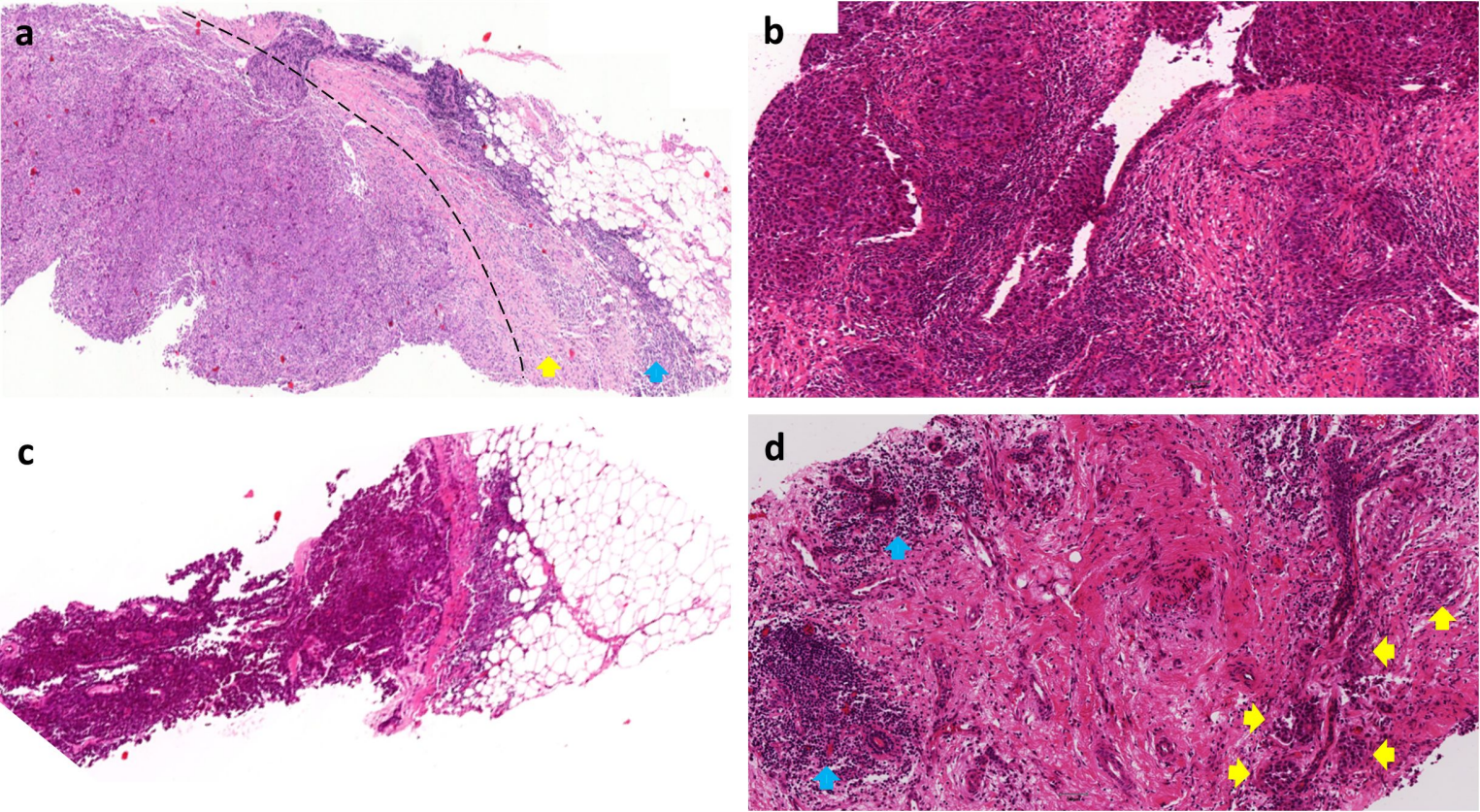

Categories
10 % slides with highest SD
90 % remaining slides
SD
sTILs (%, mean)

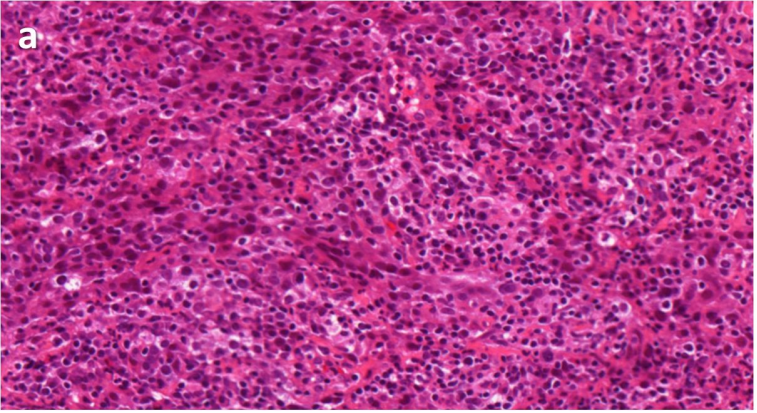
c

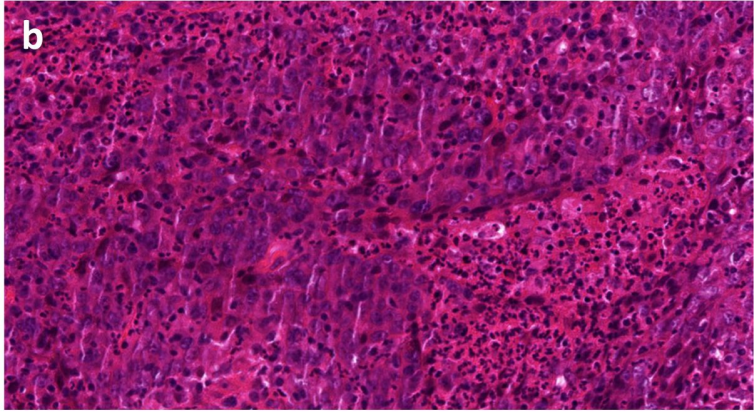

Categories
10 % slides with highest SD
90 % remaining slides
SD
sTILs (%, mean)

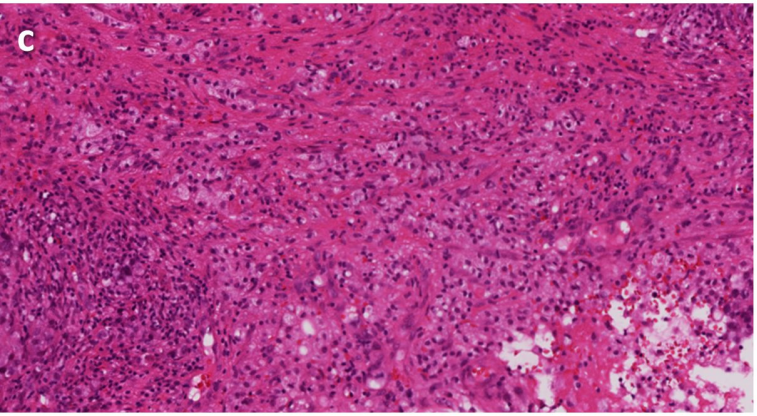

a
b
c
0.5 mm

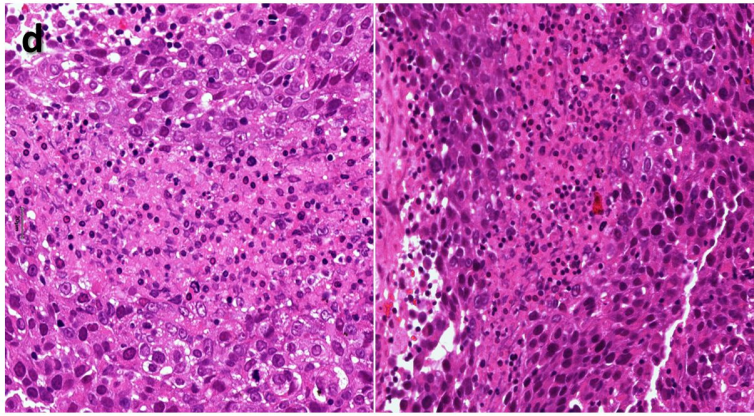

a
b
c

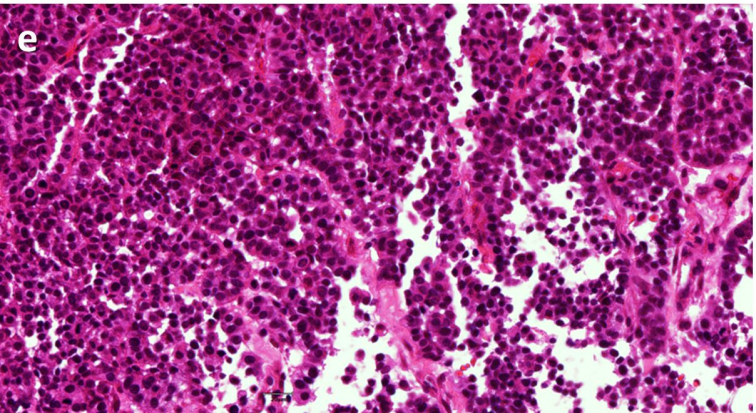

a
b
c
d

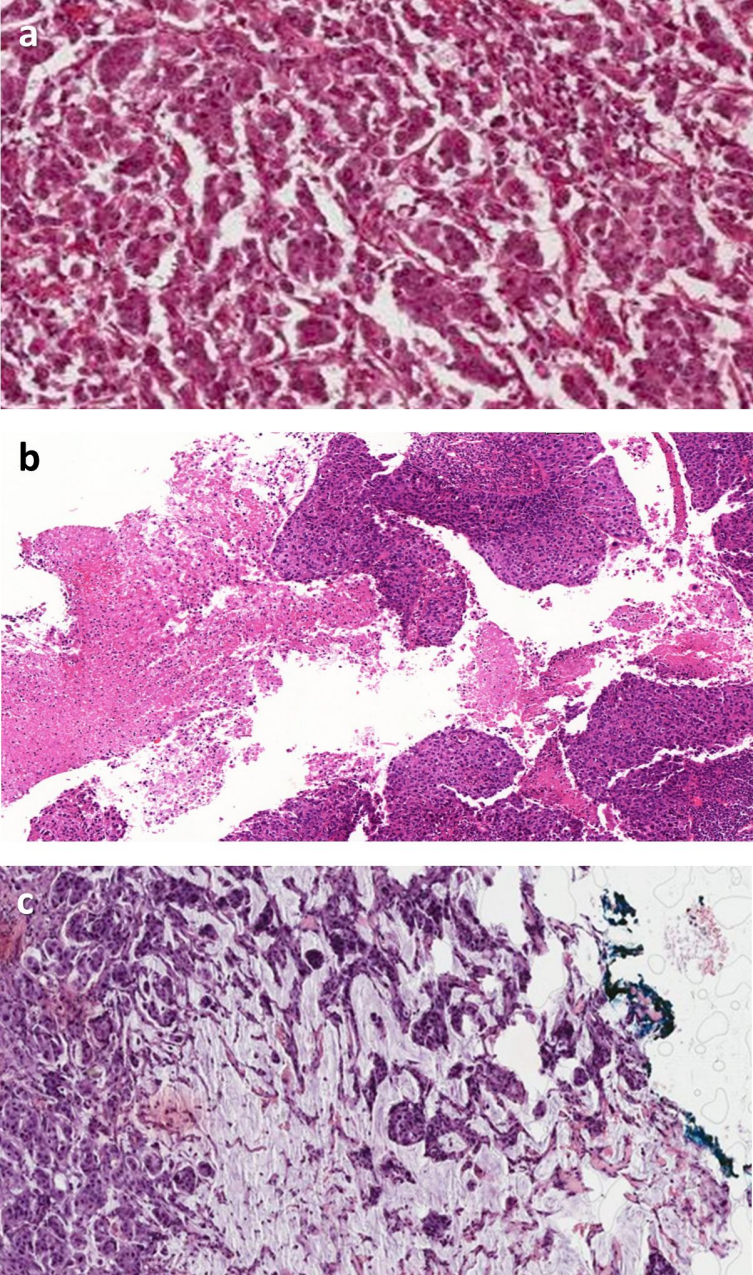

a


b


c


d


e


a
b
c

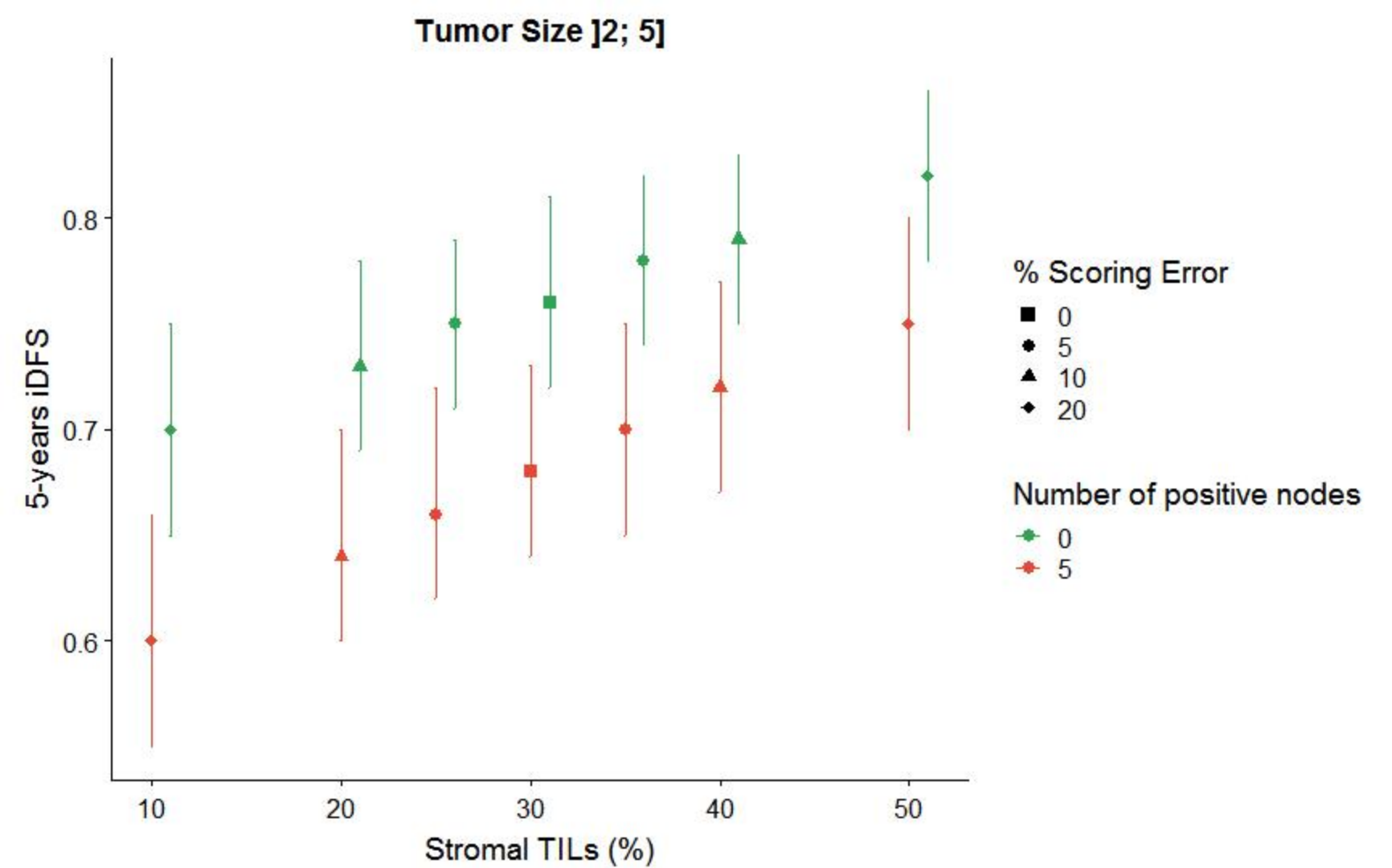

Tumor Size]2; 5]
5-years iDFS
0.8
0.7
0.6
10
20
30
40
50
Stromal TILs (%)
% Scoring Error
0
5
10
20
Number of positive nodes
0
5

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 [1]			
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)

Abbreviations: ICC, intraclass correlation coefficient; TILs, tumor-infiltrating lymphocytes.
[1]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.

Table 2. Pitfalls in sTIL assessment in breast cancer slides identified from cases showing the highest variation in 3 ring studies (RS) – Heterogeneity of lymphocyte distribution

Pitfall	Frequency seen	Recommendation
Heterogeneity	15 /26 (58 %)	
Increased sTILs at the leading edge compared to central tumor (Figure 4a)	RS1: 3/7 (43%) RS2: 1/6 (17%) RS3: 7/13 (54%)	Increased density of lymphocytes at the leading front should be included as long as the lymphocytes are within the boundary of the tumor. Scoring multiple areas and averaging the results can help with heterogeneous tumors.
Marked hterogeneity in sTIL density within the tumor (Figure 4b)	RS1: 2/7 (29%) RS2: 0 RS3: 0	All stroma within the boundary of a single tumor is included in sTIL assessment. Scoring multiple distinct areas encompassing the range of sTIL density and averaging the results can assist in providing a more reproducible overall sTIL score.
Variably spaced apart clusters of cancer cells with a dense tight lymphocytic infiltrate separated by collagenous stroma with sparse infiltrate (Figure 4c)	RS1: 2/7 (29%) RS2: 3/6 (50%) RS3: 0	All stroma within a single tumor is included within the sTIL assessment. In this situation, both the higher density areas closely associated with (but not touching) epithelial clusters and the lower density areas located between epithelial clusters are included. [The exception is a central hyalinized scar, which is excluded from scoring.] Scoring multiple areas and averaging the results can help with heterogeneous tumors.

RS1 Ring Study 1, *RS2* Ring Study 2, *RS3* Ring Study 3

Table 3. Pitfalls in sTIL assessment in breast cancer slides identified from cases showing the highest variation in 3 ring studies (RS) – Technical factors

Pitfall	Frequency seen	Recommendation
Technical factors	13/26 (50%)	
Poor quality slides / Histological artifacts secondary to prolonged ischemic time, poor fixation or issues during processing (Figure 5a)	RS1: 0 RS2: 0 RS3: 11/13 (85%)	Thankfully, in the current era, with greater awareness and monitoring of preanalytical and analytic variables, these sorts of poor quality H&E slides should not be an issue. If presented with such a case, only intact, morphologically assessable areas should be included in sTIL score. If applicable, one can cut and stain an additional section or select a different block for assessment.
Crush artifact (Figure 5b)	RS1: 1/7 (14%) RS2: 0 RS3: 0	More commonly seen in biopsy samples, crush artifact can compromise sTIL assessment. Areas of crushing should be excluded from sTIL evaluation.
Out-of-focus scan (Figure 5c)	RS1: 1/7 (14%) RS2: 1/6 (17%) RS3: 0	As part of a study one may struggle with scoring an out-of-focus scan. In clinical practice, however, particularly as sTILs are poised to impact patient management, there is no good justification to not rescan the slide. If this is not a possibility most computer programs have some capability of image correction.

RS1 Ring Study 1, *RS2* Ring Study 2, *RS3* Ring Study 3

Table 4. Pitfalls in sTIL assessment in breast cancer slides identified from cases showing the highest variation in 3 ring studies (RS) – Score wrong area of cells

Pitfall	Frequency seen	Recommendation
Scoring wrong area or cells	12/26 (46%)	
Defining tumor boundary and scoring outside of tumor (Figure 6a)	RS1: 0 RS2: 2/6 (33%) RS3: 2/13 (15%)	Do not include fibrous scars (image; yellow arrow) or lymphoid aggregates (blue arrow) beyond the invasive front of the tumor.
Including lymphocytes surrounding DCIS (Figure 6b)	RS1: 2/7 (29%) RS2: 1/6 (17%) RS3: 0	Lymphocytes surrounding DCIS are excluded from assessment of sTILs. Myoepithelial stains can be used if there is doubt as to whether a particular focus is invasive or in situ.
Including lymphocytes associated with encapsulated papillary carcinoma (Figure 6c)	RS1: 1/7 (14%) RS2: 0 RS3: 0	Only score sTILs associated with conventional invasive carcinoma. Similar to DCIS, lymphocytes associated with encapsulated papillary carcinoma should not be included in the sTIL assessment of the invasive component.
Including lymphocytes surrounding benign glands (Figure 6d)	RS1: 1/7 (14%) RS2: 0 RS3: 0	Lymphocytes associated with benign lobules or ducts should be excluded from sTIL counts when carcinoma (image; yellow arrows) surrounds benign structures. Similar lymphocytic infiltrates outside of the tumor boundary (blue arrows) can identify these as not tumor-related.
Including intratumoral TILs (iTILS) (Figure 7a)	RS1: 2/7 (29%) RS2: 1/6 (17%) RS3: 0	Certain cases show dense lymphocytic infiltrates within the tumor epithelial nests, sometimes obscuring the boundary between tumor cells and stroma. It is important to be aware that intratumoral TILs are excluded from the assessment, which only includes TILs within the intervening stroma. If necessary, a cytokeratin stain may assist with defining tumor from stroma.
Including neutrophils (Figure 7b)	RS1: 1/7 (14%) RS2: 0 RS3: 0	Only lymphocytes and plasma cells are included in sTIL evaluation. Pathologists should assess slides at a sufficiently high power to be able to differentiate between types of immune cells. Neutrophils, eosinophils, basophils, and histiocytes/ macrophages are all excluded from sTIL assessment.
Including histiocytes (Figure 7c)	RS1: 1/7 (14%) RS2: 0 RS3: 0	Only lymphocytes and plasma cells are included in sTIL evaluation. Pathologists should assess slides at a sufficiently high power to be able to differentiate between types of immune cells. Neutrophils, eosinophils, basophils, and histiocytes/ macrophages are all excluded from sTIL counts.
Misinterpreting apoptotic cells as lymphocytes (Figure 7d)	RS1: 1/7 (14%) RS2: 0 RS3: 0	At low power apoptotic cells can mimic lymphocytes. Pathologists should assess slides at a sufficiently high power to differentiate this mimic.
Artifactual falling apart of cells mimicking TILs Figure 7e)	RS1: 1/7 (14%) RS2: 0 RS3: 0	Artifactual falling apart of tumor cells is more common in biopsy specimens, particularly along the edge. At low power discohesive tumor cells can mimic lymphocytes. Pathologists should assess slides at a sufficiently high power to differentiate this mimic.

RS1 Ring Study 1, *RS2* Ring Study 2, *RS3* Ring Study 3

Table 5. Pitfalls in sTIL assessment in breast cancer slides identified from cases showing the highest variation in 3 ring studies (RS) – Limited tumor stroma

Pitfall	Frequency seen	Recommendation
Limited stroma within tumor for evaluation	8/26 (31%)	
Small volume of intratumoral stroma present for evaluation (Figure 8a)	RS1: 0 RS2: 0 RS3: 6/13 (46%)	Assessing % sTILs is difficult when the denominator is very small. Evaluation should be restricted to areas where there is clear stroma. The leading edge ought to provide at least some tumor stroma for assessment.
Large areas of necrosis (decreases scorable stromal component) (Figure 8b)	RS1: 1/7 (14%) RS2: 0 RS3: 0	Necrosis and associated granulocytes are excluded from sTIL assessment. Some tumors show extensive necrosis with only a thin rim of viable cells at the periphery. Only lymphocytes associated with viable tumor should be included. Even in highly necrotic tumor, there are typically at least some viable areas along the invasive front.
Mucinous tumors (Figure 8c)	RS1: 0 RS2: 0 RS3: 2/13 (15%)	Lymphocytes generally are absent within extracellular mucin. Thin septa and fibrous bands are often present providing a stromal component for assessment. Stroma associated with any 'no special type' component should be included.

RS1 Ring Study 1, *RS2* Ring Study 2, *RS3* Ring Study 3