INTERNATIONAL TELECOMMUNICATION UNION

TELECOMMUNICATION STANDARDIZATION SECTOR

STUDY PERIOD 2022-2024

FGAI4H-P-013-A01

ITU-T Focus Group on AI for Health

Original: English

WG(s):	Plen	Helsinki, 20-22 September 2022

DOCUMENT

Source:	TG-Histo Topic Driver
Title:	Att.1 – TDD update (TG-Histo) [same as Meeting I]
Purpose:	Discussion

Contact: Frederick Klauschen
Charite Berlin, Germany
Email: frederick.klauschen@charite.de

Abstract: This document contains the topic area description for the FG-AI4H histopathology topic area.

This version of the TDD is the same as seen in Meeting I (FGAI4H-I-013-A01), reproduced for easier reference as a Meeting N document.

Table of Contents

1 Introduction

1.1 Document Structure

This document is an update of the TDD of the topic group "Histopathology"

1.2 Topic Description

Overview

Tumor Infiltrating Lymphocytes (TILs) are emerging as a very promising biomarker in solid tumors such as breast cancer, lung cancer and melanoma. TILs have been shown to be a reliable and reproducible marker of tumor immunogenicity in breast cancer. It is clear that higher levels of TILs are associated with improved prognosis in certain subtypes of breast cancer while their presence indicates a decreased survival in other subtypes. TILs also indicate a higher probability of achieving therapy response in the neoadjuvant setting. Analysis of TILs in residual disease specimens after neoadjuvant therapy has also been shown to have prognostic value. The evaluation of TILs as a biomarker in breast cancer is expected to be extended from the research domain to the clinical setting in the near future. While TILs are normally assessed by manual estimation, efforts are ongoing for the assessment of TILs by image analysis methods. These methods, and among them particularly AI-based methods, are still experimental and not sufficiently documented and standardized for introduction into clinical trial and daily practice.

We therefore propose to establish a data set for the benchmarking of machine learning based tumor cell detection and TIL quantification algorithms.

Impact

The assessment of TILs by digital image analysis will be useful for accurate and reproducible diagnostics in the future, because this approach can be used to determine the number of TILs per stromal tissue area as an exact measurement contrary to the approximate semi-quantitative evaluation suggested at this moment. In the first International Guidelines on TIL-assessment in breast cancer (Salgado et al., Annals of Oncology 2014), an inter-laboratory quality comparison study was proposed to assess the reproducibility and clinical validity of TIL evaluation. Because conventional image analysis approaches, although capable of identifying lymphocytes relatively easily (Wienert et al., 2012, Scientific Reports), have difficulties in robustly detecting tumor cells due to their broad morphological variability, machine learning approaches have been and are currently being developed that allow for a combined detection of both lymphocytes and cancer cells required for accurate TIL scoring (reviewed in Klauschen et al., 2018, Seminars in Cancer Biology).

1.3 Ethical Considerations

Discussion on ethical considerations is ongoing. We have so far discussed the usage of image data from routine diagnostics for the focus/topic group's purpose. We consider this ethically acceptable, because patients have consented the use of their data for research purposes, and all digitized image data are fully anonymous.

1.4 Existing AI Solutions

Various AI-based histopathology solutions are available and developments are under way. However, none of the TiL scoring approaches is already broadly used in diagnostics and no diagnostics-grade benchmarking approach is available.

1.5 Existing work on benchmarking

Currently, no high-quality annotated data sets on TILs in breast cancer are publicly available.

We intend to provide a comprehensive histological image data set that allows for the evaluation of image analysis methods for tumor cell and lymphocyte detection and quantitative scoring in breast cancer (Fig. 1A,B). These Hematoxylin&Eosin (H&E) image data will be provided in an undisclosed fashion within a compute infrastructure that will be used for the actual benchmarking process.

We will provide a second (disjoint and smaller) data set for public download for participants to assess general features of the data used for benchmarking such as quality, staining and morphological spectrum and to compare these features to local data sets used for training their algorithms.

It is important for clinical-grade validation that the data we provide for public download are not sufficient to fully train the developed algorithm de-novo, but that the classifier is benchmarked with a data set independent of that used for training.

Existing so-called "Challenges" usually make the data publicly available so the participants annotate and train the classifier themselves. Apart from the fact that cheating cannot be ruled out using this approach, training and test data are not sufficiently separate and generalizeability cannot be properly evaluated using this benchmarking design.

2 AI4H Topic group

Currently, the focus of the topic group Histopathology is on breast cancer cell and tumor-infiltrating lymphocyte detection. However, since multiple potential applications for AI-based computational pathology exist, we will add further subtopics after having successfully implemented the benchmarking pipeline for this use case.

We are also planning to extend the scope to molecular diagnostics where AI based approaches are becoming increasingly important (Jurmeister et al., 2019).

3 Method

In the benchmarking process, the participants are expected to submit AI-based solutions that will analyze the histopathological images and

- automatically detect tumor cells and lymphocytes, and/or
- quantify the lymphocyte and tumor cell density (number of cells per square millimeter in the tumor area or in the border area of the tumor), and/or
- predict the semi-quantitative score as diagnosed by pathologists after visual inspection and comparison with reference images (Salgado et al., Annals of Oncology 2014).

The submissions should be evaluated by comparing the AI-based predictions with the cell-wise manual annotations and scores given by pathologists. Different benchmarking metrics are conceivable including statistical measures such as the detection performance (accuracy, F1 score, area under the curve of the receiver operating characteristic etc.) and the quantification error (e.g., the root mean square error). Explanations in visual form that allow humans to interpret why the AI-algorithm eventually came to a conclusion or made a prediction are additional measures to be considered in the benchmarking procedure (see Fig. 1 C for an example).

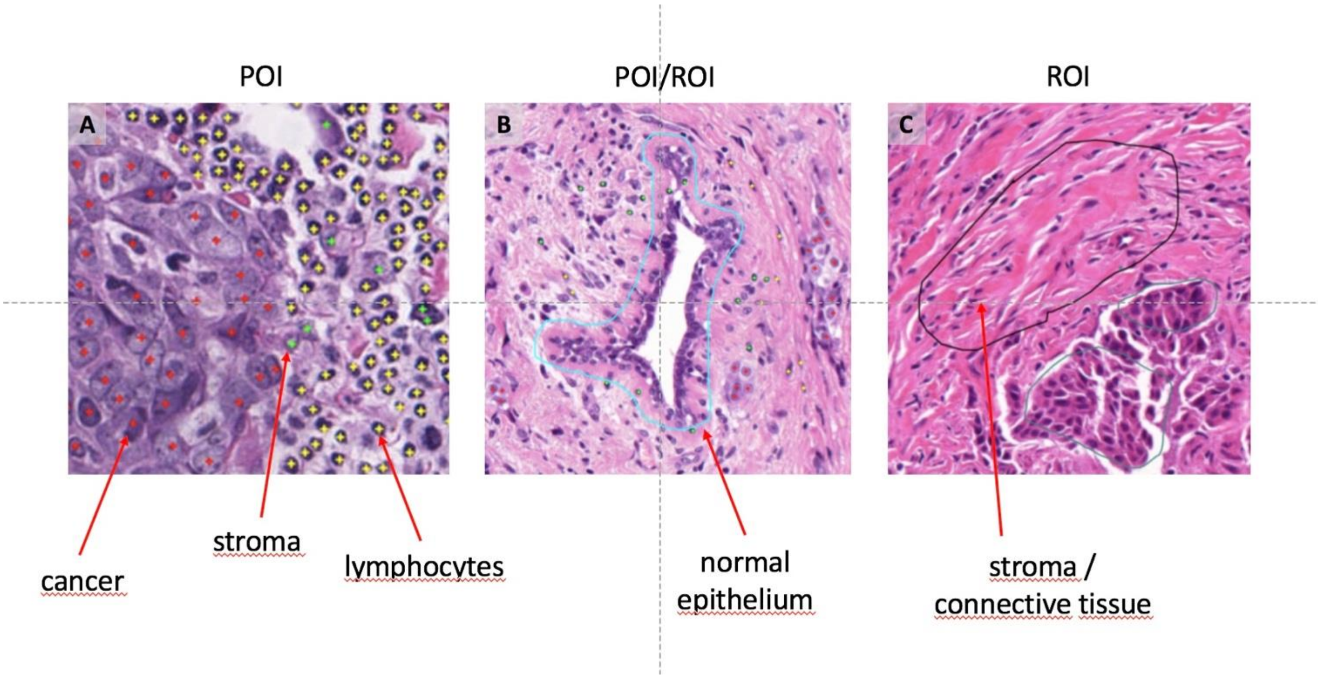

Figure 1: Example of POI and ROI annotations.

3.1 AI Input Data Structure

- Annotations should be flexibly reusable with different patch sizes extractable from annotation coordinates (saved in xml-format)
- Annotation procedure (single cell “point” vs. area “region” annotation)
    - positive annotations
        - <u>point annotations</u> (POI): cell nuclei are marked, relevant for heterogeneous tissues (e. g. individual lymphocytes between cancer cells)
        - <u>region annotations</u> (ROI): regions containing at least 95% cells of respective class
    - negative annotations
        - <u>region annotations</u> (ROI): regions negative of a certain class, i. e. region may contain any cells, but none of the respective

3.2 AI Output Data Structure

- Classifier provides binary classification for each image patch.

3.3 Test Data Labels

- cancer tissue
    - multiple subtypes
    - focus on NST (no-special-type) and invasive-lobular breast cancer
- normal tissue
    - normal breast gland and duct epithelium
    - connective tissue (fibers, cells)
    - fatty tissue
    - bone tissue

  - blood and lymphatic vessels
  - nerves
- immune system
  - lymphocytes
  - granulocytes
  - monocytes/macrophages
  - plasma cells
- necrotic tissue
- artifacts
- background

3.4 Scores & Metrics

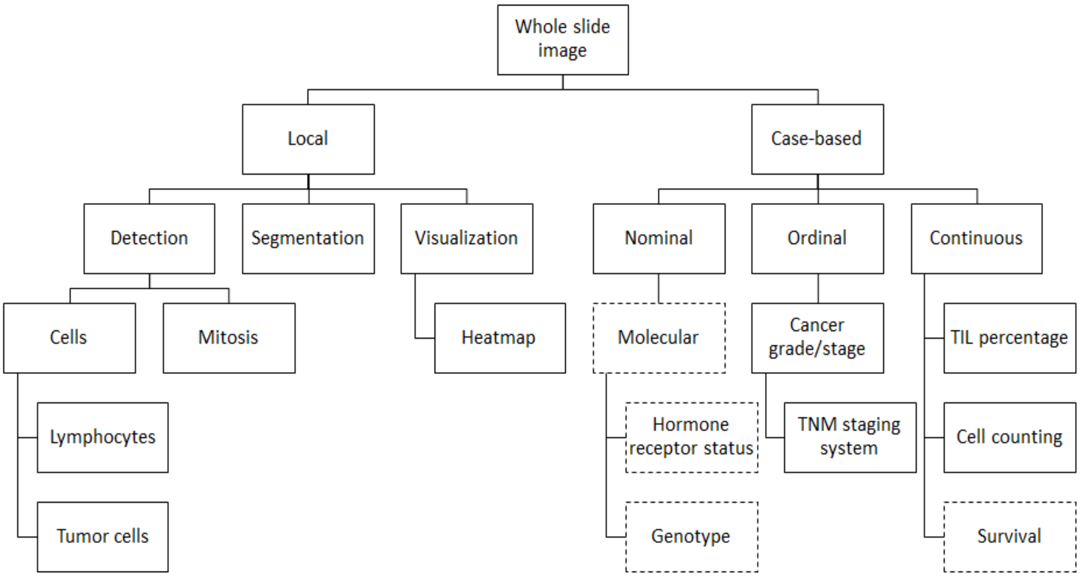

Figure 2:: Different benchmarking measures.

3.5 Undisclosed Test Data Set Collection

Our current data set consists of 90 2000x2000 histopathological breast cancer images (standard histological slides; stained with Hematoxylin&Eosin) at 400x resolution. The data set contains 258k patches. Out of this data set 80 images will compose the undisclosed data set and a disjoint set of 10 images will be made publicly available. The images were digitized with a whole-slide-scanner (3DHistech, Budapest).

4 Benchmarking Methodology and Architecture

The benchmarking process consists of 3 independent stakeholders.

- Data annotation and provision
- AI development and algorithm submission
- Data collection, server hosting, running of algorithm, report generation

4.1 Multi-site data annotation

One of the major open questions in providing benchmarking data for histopathology is the required scope and redundancy of annotations. How many pathologists need to agree on a how large and diverse data set? Using a data set that will be made publicly available with a recent publication (Hägele et al., 2020) and a web-based annotation tool, we will provide data to an international group of pathologists to perform annotations on the same data sets. The associated data base facilitates keeping track of annotation redundancies/differences and will be the basis for an updated AI4H benchmarking data set for breast and lung cancer.

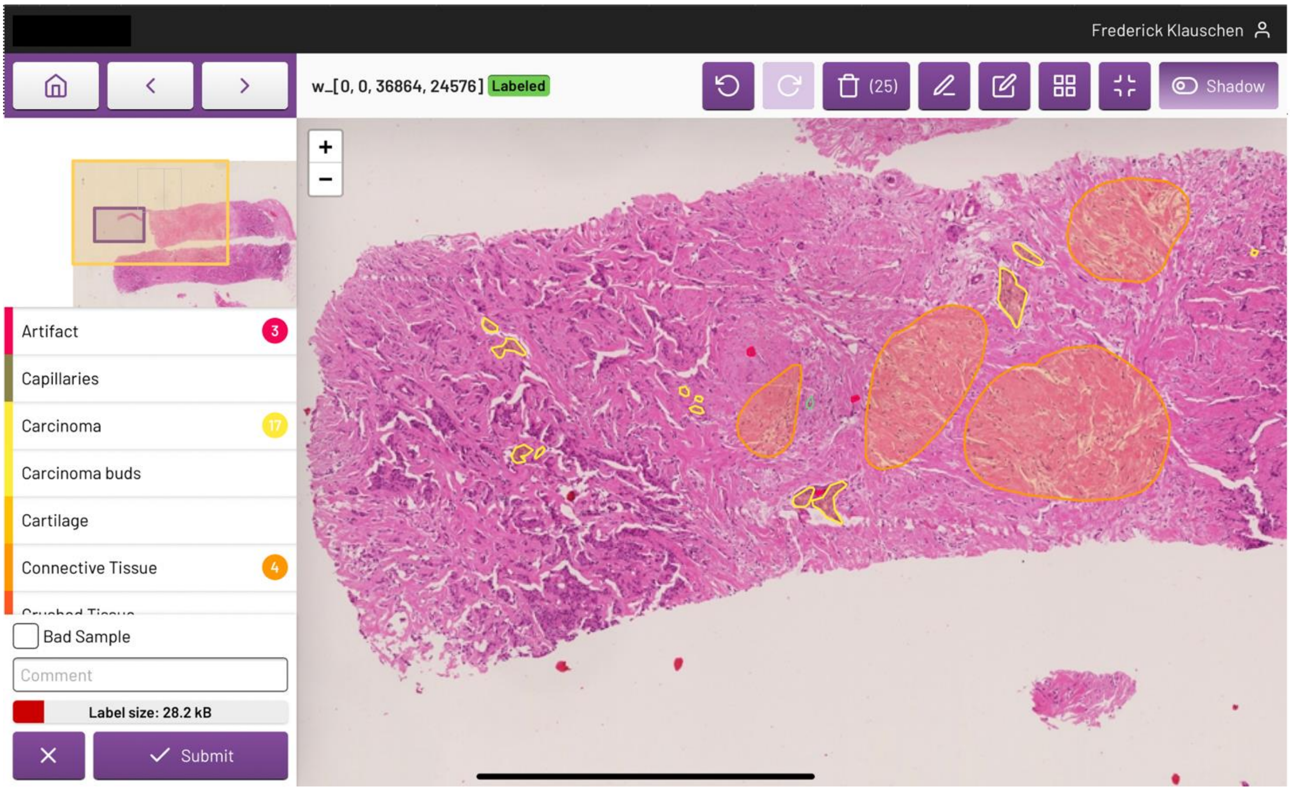
[redacted]

Figure 3: Web-based annotation tool for multi-site benchmarking data generation with redundant annotation.

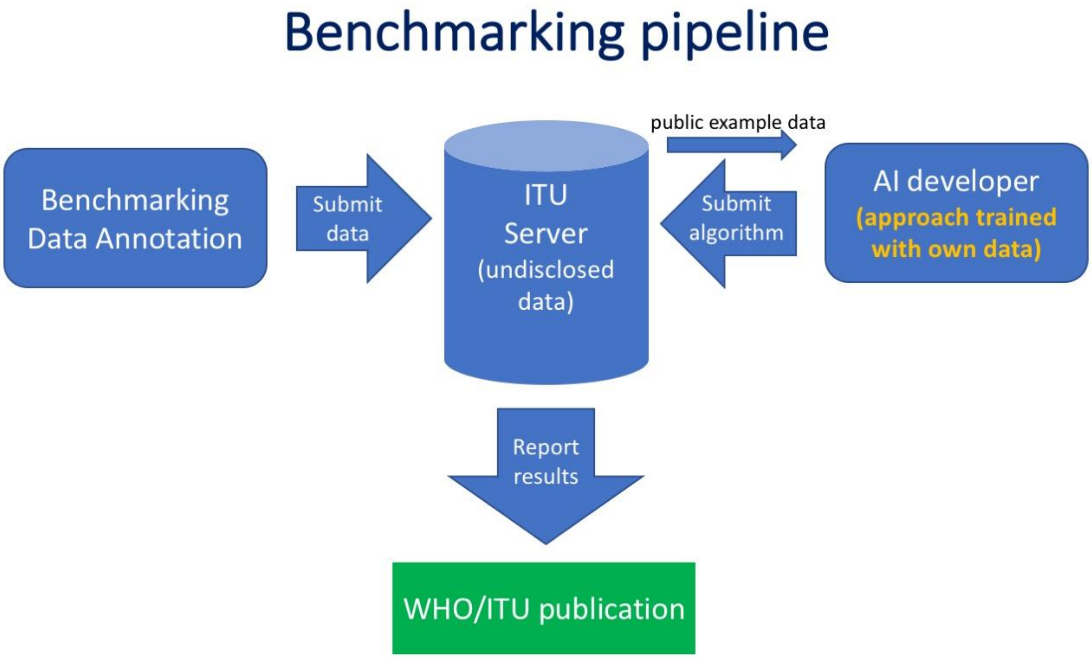

Figure 4: Benchmarking process. The annotated data is submitted to the ITU. A small subset of the data is made publicly available (max. 10% of the data set). The remaining data is stored at the ITU and not disclosed. The AI developer submits the AI approach, which was trained on a different data set owned by the AI developer. The ITU staff runs the submitted algorithm on the undisclosed data and computes statistical measures which are then reported to the AI developer and published by the ITU.

5 Reporting Methodology

- For the current proof-of-concept study, statistical measures were reported by e-mail. Structured reports and reporting guidelines need to be discussed and implemented.

5.1 Results

The first benchmarking run was performed on May 28th 2019. An algorithm for breast cancer cell detection was submitted by Prof. Dr. Alexander Binder, Singapore University of Technology. This proof-of-concept run yielded a true positive rate of tp=0.91 and a true negative rate of tn=0.88.

Heatmaps and further statistical measures were not available for this initial proof-of-concept.

5.2 Discussion

The discussion in ongoing on how annotations are compared with algorithm output. While we currently use patch-wise classification in the proof-of-principle validation run, algorithms may output predicted cell coordinates or probability maps. Here, various alternatives exist for comparing ground truth and AI output and biases need to be avoided.

We will also need to provide the exact physical resolution since classical magnification numbers (400x) do not correspond to exact same slide scanner resolutions.

Discussions are also still ongoing on the number of pathologists who need to agree on annotations to consider them sufficient to be a standard used in benchmarking.

Benchmarking of the value of explanatory heatmaps will have to be defined.

Benchmarking should be further refined by offering average results but also provide information on a case basis, including the identification of of outliers (i. e. cases which are partial/complete fails) and compute tail accuracies (number of cases for which the AI approach can achieve at least x% (e. g. x=95, x=99) accuracies.

5.3 Declaration of Conflict of Interest

- The authors of this document declare no conflict of interest.

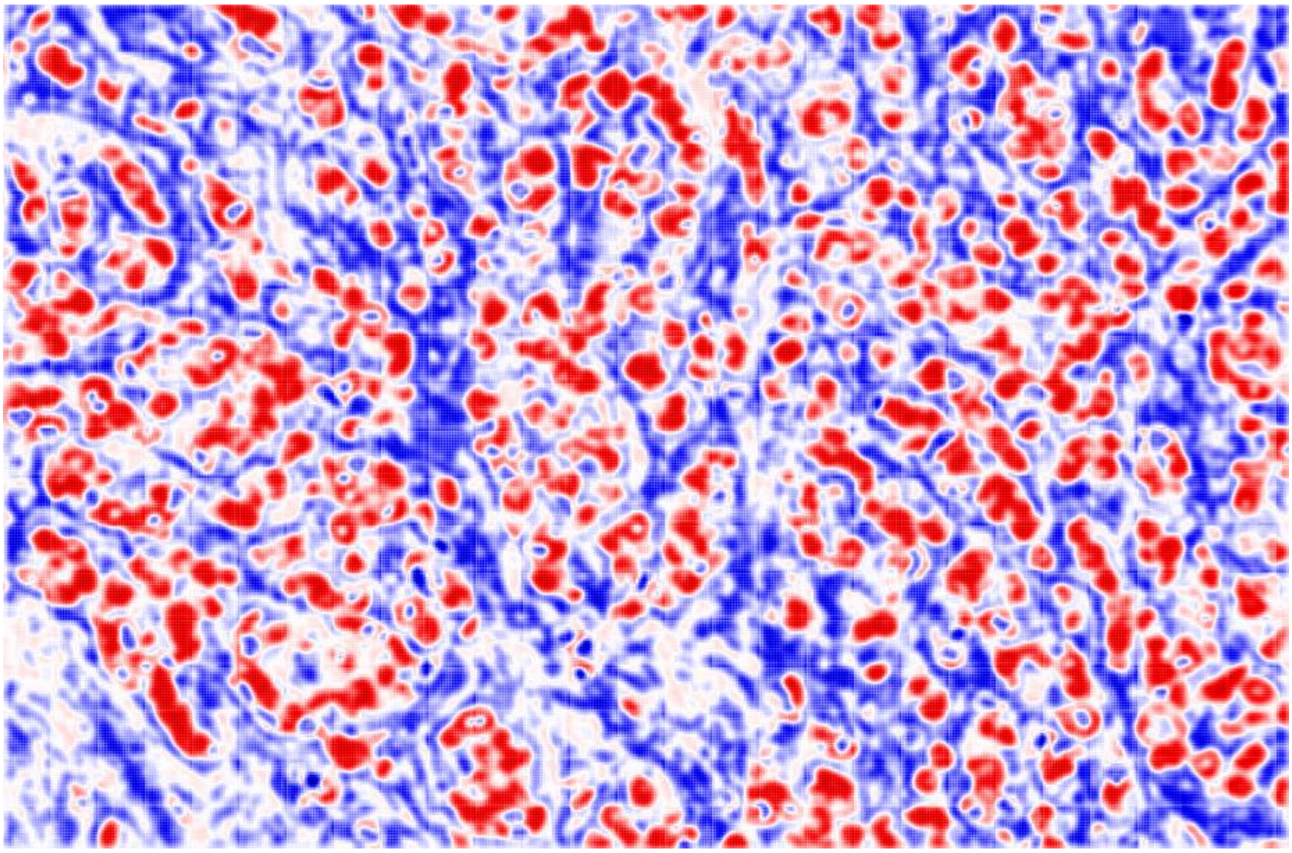

Figure 5: Example heatmap for cancer cell detection (red) vs. normal tissue (blue).

References

Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. Hägele M, Seegerer P, Lapuschkin S, Bockmayr M, Samek W, Klauschen F*, Müller KR*, Binder A*. ***Scientific Reports.*** 2020 Apr 14;10(1):6423.

Machine learning analysis of DNA methylation profiles distinguishes primary lung squamous cell carcinomas from head and neck metastases. Jurmeister P, Bockmayr M, Seegerer P, Bockmayr T, Treue D, Montavon G, Vollbrecht C, Arnold A, Teichmann D, Bressem K, Schüller U, von Laffert M, Müller KR, Capper D, Klauschen F. ***Science Transl Med***. 2019 Sep 11;11(509).

Salgado, R., Denkert, C., Demaria, S., Sirtaine, N., Klauschen, F., Pruneri, G., ... & Perez, E. A. (2014). The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014. ***Annals of oncology***, *26*(2), 259-271. https://doi.org/10.1093/annonc/mdu450

Wienert, S., Heim, D., Saeger, K., Stenzinger, A., Beil, M., Hufnagl, P., ... & Klauschen, F. (2012). Detection and segmentation of cell nuclei in virtual microscopy images: a minimum-model approach. ***Scientific reports***, *2*, 503. https://doi.org/10.1038/srep00503

Klauschen, F., Müller, K. R., Binder, A., Bockmayr, M., Hägele, M., Seegerer, P., ... & Michiels, S. (2018, July). Scoring of tumor-infiltrating lymphocytes: from visual estimation to machine learning. In ***Seminars in cancer biology***. Academic Press. https://doi.org/10.1016/j.semcancer.2018.07.001

____