EBioMedicine 70 (2021) 103520

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.elsevier.com/locate/ebiom

# What do we still need to learn on digitally assessed biomarkers?

Balazs Acs[a,b,*], Roberto Salgado[c,d], Johan Hartman[a,b,e]

[a] *Department of Oncology and Pathology, Karolinska Institutet, Stockholm, Sweden*
[b] *Karolinska University Hospital, Stockholm, Sweden*
[c] *Department of Pathology, GZA-ZNA Hospitals, Antwerp, Belgium*
[d] *Division of Research, Peter Mac Callum Cancer Centre, Melbourne, Australia*
[e] *Medtech Labs, Bioclinicum, Stockholm, Sweden*

ARTICLE INFO

*Article History:*
Received 27 July 2021
Accepted 27 July 2021
Available online xxx

Is there a ground truth for digital pathology in TILs or any other computationally assessed biomarker? Is it the concordance between the pathologist and the computationally assessed biomarker, or is it patient outcome or perhaps a combination of both? If it is a combination of both, how can it be used in daily practice?

In their recent manuscript Sun et al. have assessed the concordance between manual TIL-scores in different breast cancer cohorts and computationally assessed TIL-counts [1]. It is not surprising that the concordance was not excellent, as manually and computationally assessed TILs are measured using different methods, and measure different variables. It should not be a surprise either that comparing different computationally tools with each other, all developed and validated in different ways, will probably also not lead to optimal concordances, as all measure different but intrinsically related variables of the same biological process, in this case immunity, exemplified by the TILs. However, in the current study both manual as well as computationally assessed TILs predicted outcome. Is the ground truth then the outcome of the patients? Is it that simple? Furthermore, the authors have demonstrated that the combination of computationally and manual measurement predicts outcome better than either variable alone. How should pathologists consider this finding? Should pathologists first score the TILs manually, and when in doubt use a computational tool? If so, what is the cost-benefit of this approach for the common pathology laboratory?

Many issues are still unsolved, but the authors need to be commended for having performed a thorough and critical evaluation of computationally assessed TILs and the potential pitfalls associated herewith. The same variables that induce variability between pathologists for manual TIL-assessment will also induce variability when computationally assessed TILs are used, exemplified by the heterogeneity and the inclusion or exclusion of TILs in specific locations within the cancer. For example, DCIS and normal lobules can contain many immune cells, whilst the invasive cancer cell component next to it, can have no immune cells. To what extent is this critical for a computational tool? A cut-off on TILs is still elusive, and the difficulties for finding a cut-off for biomarkers in general is fraught with uncertainty [2]. Clinicians take binary decisions -treat or not to treat- suggesting that the biomarker-information should also be binary, using cut-offs. However, as the authors have demonstrated, the TILs as a continuous variable predicted outcome, confirming previous findings [3]. Might computationally assessed biomarkers not better be integrated in nomograms containing additional prognostic variables such as lymph node status, tumor size, age, and TILs as a continuous variable, obviating the need to determine a cut-off? This would eliminate the need to determine a cut-off, also in different ethnicities.

The authors correctly point out that further validation of computational tools for TIL-assessment is important [1]. The use of clinical trials in order to validate their findings is crucial. The International Immuno-Oncology Biomarker Working Group (www.tilsinbreastcancer.org) will organize a public Grand Challenge together with the Computational Pathology Group (Diagnostics Image Analysis Group) of the University of Nijmegen, on computationally assessed TILs using on the one hand the slides used by the Working Group for their RING-studies and phase 3 clinical trials [4]. This will provide the community the opportunity to perform a thorough analytical and clinical validity of their tool, partnering with the Working Group. This exercise will inform the community to what extent much detail is needed in clinical decision making. For example, is it important to know exactly how much TILs there are per mm$^2$, or is an assessment by pathologists sufficient for clinical decision making? How much discrepancy between the pathologists´ assessment and the computational assessment on the one hand, and between different computational tools is clinically acceptable? How much deviation between two measurements, either by pathologists or using computational tools, will affect clinical decision making? The Working Group has labeled this as a the "Clinically Allowable Error Margin" that is still unknown on TILs, and it is also unknown for most morphological biomarkers we use for decades in our daily practices. In addition, will additional TIL-related variables, such as for example the distance between TILs and cancer cells affect outcome? Using a publicly available prognostic model, the Working Group will be able to evaluate this. Answers to all the above is probably related to the

DOI of original article: http://dx.doi.org/10.1016/j.ebiom.2021.103492.

[*] Corresponding author.
*E-mail address:* balazs.acs@ki.se (B. Acs).

https://doi.org/10.1016/j.ebiom.2021.103520

biological importance of the biomarker. If immunity is very active, and consequently if many immune cells are present, probably a limited degree of discordance may conceptually not matter that much. This needs however to be proven. For other morphological biomarkers, a more stringent concordance might be needed. Nevertheless, it remains crucial that pathologists score the biomarker as reliably as they are able to, with adequate training and with support of reference materials.

Finally, how can the authors' findings inform subsequent developments in computationally assessed TILs? First, the lessons learned with TILs can be a good paradigm for the development of other computational tools that assess biomarkers on haematoxylin and eosin (HE) slides. There is currently a lack of publicly available annotated HE-images that can inform and help the scientific community to validate their computational tools. It is advised to scientific journals and computational pathology research groups to make their annotations publicly available. Second, there is a lack of reference materials that can be used to compare different computational tools with each other. The Working Group is partnering with the FDA and other organizations to develop these materials that can inform and help the scientific community [5]. Third, building on the current Ki67-narrative, the Ki67 Working Group has recently issued a recommendation that Gene Expression Profiles can be used in that Ki67-category where pathologists are probably not so concordant, namely in the category between 5% and 30% Ki67-expression [6]. Can a similar reasoning conceptually be applied to the use of computationally assessed TILs? Fourth, to what extent should computational tools be validated according to the subtype is unknown. In practice, quite probably pathologists will not use different computational tools on TILs for luminal disease, HER2+ and TNBC, as the “TILs all look the same” in all these subtypes. Can there be a generic computational tool on TILs irrespective of the subtype? Furthermore, can this tool even ben generic for the sample type, namely for core-biopsies vs surgical HE-slides? Fifth, considering the diminished immune cell counts in metastatic sites, with different TIL-levels and with a different microenvironment according to the metastatic site, can a computational tool that was validated on primary samples be used for metastatic sites [7]? If not, should it be validated according to each metastatic site? Can this ever be practicable? All the above illustrates that there are still many issues that need to be solved. Training of pathologists remains crucial, for manual assessment, and trained pathologists for TIL-assessment are crucial for optimal and thorough validation of computational tools, exemplifying that both are needed before optimal implementation of computational tools on TILs or any other morphological biomarker can be considered in the workflow of all pathologists. This is an active field in progress. Computational tools will probably not be deemed by necessarily more precision, but rather by providing the practicing pathologists alternatives to help determine the biomarker where this is more needed in their practices. A transparent cooperation and communication between all stakeholders, patients, pathologists, clinicians, industry, and the regulatory instances is needed to make this a success, for the sake of our patients.

## Contributors

All authors have contributed equally.

## Declaration of Competing Interest

JH has received speaker's honoraria or advisory board remunerations from Roche, Novartis, Merck, Eli Lilly, Pfizer and MSD. JH is co-founder and shareholder of Stratipath AB. JH has received institutional research grants from Cepheid and Novartis. Other authors have no conflicts of interest to disclose.

## References

[1] Sun P, He J, Chao X, et al. A computational tumor-infiltrating lymphocyte assessment method comparable with visual reporting guidelines for triple-negative breast cancer. EBioMedicine 2021;70:103492. doi: 10.1016/j.ebiom.2021.103492.

[2] Fundytus A, Booth CM, Tannock IF. How low can you go? PD-L1 expression as a biomarker in trials of cancer immunotherapy. Ann Oncol 2021;32(7):833–6.

[3] Loi S, Drubay D, Adams S, et al. Tumor-infiltrating lymphocytes and prognosis: a pooled individual patient analysis of early-stage triple-negative breast cancers. J Clin Oncol 2019;37(7):559–69.

[4] Denkert C, Wienert S, Poterie A, 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 2016;29(10):1155–64.

[5] Dudgeon SN, Wen S, Hanna MG, et al. A pathologist-annotated dataset for validating artificial intelligence: a project description and pilot study. *arXiv preprint arXiv:201006995*.

[6] Nielsen TO, Leung SCY, Rimm DL, et al. Assessment of Ki67 in breast cancer: updated recommendations from the international Ki67 in breast cancer working group. J Natl Cancer Inst 2021;113(7):808–19.

[7] Loi S, Adams S, Schmid P, et al. Relationship between tumor infiltrating lymphocyte (TIL) levels and response to pembrolizumab (pembro) in metastatic triple-negative breast cancer (mTNBC): results from KEYNOTE-086. Annal Oncol 2017;28:v608.