Deep learning for Circulating Tumor Cell (CTC) identification with the CELLSEARCH system: towards optimal standardization

Luca Biasiolli[1]
Pietro Ansaloni[1], Nicolò Gentili[1], Ramona Miserendino[1], Chiara Rossi[2], Giulio Signorini[1] and Gianni Medoro[1]
[1] Menarini Silicon Biosystems S.p.A.
[2] Alten Italia S.p.A.

Background & objectives

CTC enumeration in blood samples with CELLSEARCH® is a prognostic biomarker in metastatic breast, prostate and colorectal cancer. CTC identification is performed by reviewers through visual assessment, which is time-consuming and potentially affected by subjective interpretations. Artificial Intelligence (AI) can automate CTC identification, producing faster and more reproducible results. Recently we presented an automated algorithm (research use only) for CTC identification in CELLSEARCH® images (Ansaloni, EACR2022).

This study aims to measure AI performance compared to human reviewers and study inter- and intra-rater differences between repeated reviews with and without AI support.

Methods

The AI algorithm was trained on fluorescent images of 7255 CTCs and 32876 non-CTC events from 90 breast, 122 prostate and 54 colorectal cancer samples. 10 reviewers, qualified for CTC image analysis, performed blind labeling of a separate dataset (26 breast and 28 prostate cancer samples). The ground truth (GT) was generated by 3 experienced reviewers to assess CTC classification performance of the other 7 reviewers and AI. 3 reviewers (3R) repeated blind labeling with CTC selection suggested by AI to test if it could improve human performance and variability compared to their first review.

Results

GT identified 1182 CTCs out of 5352 events (test dataset). Image classification performance of AI (accuracy=95.7%; F1=90.5%) was close to the top of the reviewers' range (accuracy=92.7-96.1%; F1=82.4-91.7%). The average F1 of 3R improved from 86.3% to 87.5% and accuracy from 94.2% to 94.6%.

Counting CTCs per sample and applying the clinically validated cutoff (CTC $\geq$ 5) for favorable/unfavorable prognosis, AI scored perfect agreement with GT (accuracy=100%; F1=100%) compared to reviewers (accuracy=94.4-100%; F1=91.4-100%). The average

accuracy of 3R increased from 95.0% to 97.5% and F1 from 92.7% to 96.2%. Their inter-rater agreement of 3R on image classification improved (Fleiss Kappa=0.76vs0.79) and their variability on CTC enumeration decreased (mean CV=20.7%vs22.7% on 25 samples with GT≠0CTC).

Conclusion

AI performed better than most reviewers and helped increase their accuracy and reduce subjectivity in CTC identification. These results show that introducing AI could improve standardization while reducing review time.