Inter-observer agreement of the response to therapy assessment in advanced lung cancer within a normative measurement environment

H. Beaumont[1], E. Oubel[1], A. Iannessi[2], D. Wormanns[3]
[1]Median Technologies, Valbonne, France
[2]Centre Antoine Lacassagne, Nice, France
[3]ELK Berlin Chest Hospital, Berlin, Germany

Purpose: Image-based biomarkers play an increasing role in the assessment of response to therapy. The value of a biomarker comes, in part, from its ability to guarantee reproducibility in a varying context. This study aims at evaluating the impact of workflow normalization and automation on the reproducibility of volume-based response assessment. This impact is measured in terms of inter-reader agreement (IRA).

Method: A retrospective study was performed on 10 patients with Non-Small Cell Lung Cancer (NSCLC) lesions followed over 7 time points (TP) on average with Computed Tomography. Five imaging scientists measured sequentially the volume of each lesion at each TP and the time required to perform segmentations. We relied on a software providing semi-automatic segmentation capabilities and follow-up (FU) display. After 6 months, a second reading session used no automation for segmentation. The response to treatment was assessed according to +/-30% thresholds as recommended by the Quantitative Imaging Biomarker Alliance (QIBA). The IRA was measured by using Kappa coefficient.
From the initial reading, where the same reader reviewed consecutively all TPs from the same patient, additional IRA assessments where performed by random mixing of measurements from different readers. Different types of mixing patterns simulated several deviations from normalization, this corresponding to FUs involving more than one radiologist or method.

Results: The IRA of a normalized and automated workflow yielded a significantly higher kappa = 0.69 [0.59; 0.79] compared to mixed manual segmentations where kappa was 0.24 [0.06; 0.42]. Analyzed separately, both single-reviewer assessment and semi-automated segmentation led to higher reproducibility. The recourse to semi-automated segmentation reduces the average segmentation time by a factor of 4.

Conclusions: Normalization and automation of the measurements improved significantly the IRA. Both normalization and automation contribute to improved reproducibility. Even small deviations from a normalized review may impair the global reliability of FUs. Single-reviewer reading and automation must be considered for a highly reproducible assessment of response to therapy.