

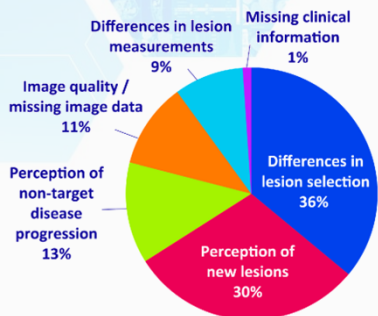
Inter-observer Agreement of The Response To Therapy Assessment in Advanced Lung Cancer within a Normative Measurement Environment

H. Beaumont¹, E. Oubel¹, A. Iannesi², D. Wormanns³

¹MEDIAN Technologies, Valbonne, France; ²Centre Antoine Lacassagne, Nice, France; ³ELK Berlin Chest Hospital, Berlin, Germany

Problem Statement

The variability in assessing the response to therapy is a global issue affecting not only the reliability in clinical trials but also longitudinal assessments in routine oncology practice. Specifically, imaging assessments present the same limitations which sources and consequences have extensively been investigated in the literature.



Analysis of blinded independent central review (BICR) data from 40 oncology clinical trials in 12 indications including 12,299 subjects (Borradaile et al., 2010)

Objectives

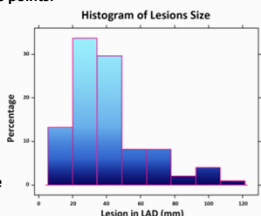
By assessing the inter-reader variability in different contexts, using the volume of lesions as a biomarker, we intended to evaluate the impact of a normative measurement environment - standardizing and automating the workflow and the measurements - on the reproducibility of volume-based response assessment. From this analysis, we aim at defining guidelines enabling more reliable assessments and a normative measurement environment that addresses some of the sources of variability.

Method and Data

The data, from a retrospective study, includes 10 patients with Non-Small Cell Lung Cancer (NSCLC) with an average of 7 Time Points (TP) per patient. In 70% of the patients, image acquisition protocols were different among the time points. 50% of the lesions were reported uneasy to segment, having spiculated shapes or being close to complex anatomical structures.

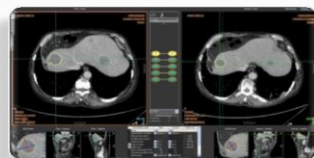
At each TP, five readers, imaging scientists and radiologists, segmented each lesion to extract its volume. The same reader reviewed all the TPs of each patient twice, through a manual segmentation then, six months later, through a semi-automated segmentation. The amount of time required to measure the volume was recorded. The response to treatment was assessed by applying +/-30% thresholds on the sum of lesion volumes.

Random mixing of measurements from different readers simulated a change of reader during patient follow-up.

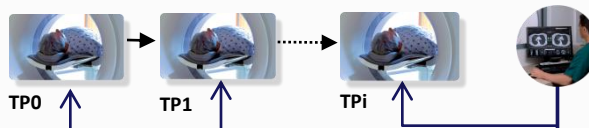


Normative Measurement Environment Automation with Lesion Management Solutions - LMS

- Lesion Segmentation
- Scan Registration
- Lesion Pairing
- Biomarker Extraction
- Detection of Changes
- Reporting



Quality Control through Clinical Trial Imaging Services - CTIS procedures



Standardized Workflow

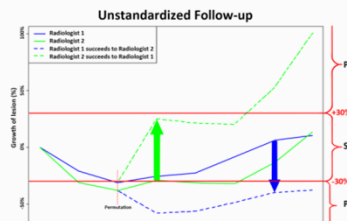
The involvement of several radiologists in the assessment of the TPs of the same patient often leads to misclassification even if both radiologists would have agreed on the global outcome when assessing all TPs independently from each other.

The inter-reader agreement under a normalized and automated workflow involving a single reader per patient, and several readers per patient were evaluated using kappa statistics. The involvement of several readers in assessing each individual patient degrades the agreement.

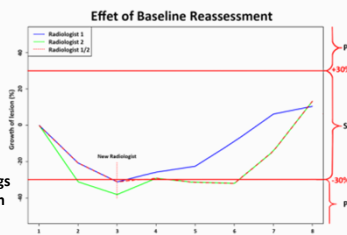
The effect induced by the change of reader during the assessment can be mitigated by:

- Baseline re-assessment
- Checking coherence of the nadir
- Using side by side follow-up display

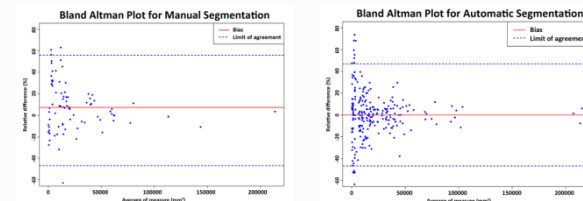
The re-assessment of the baseline brings continuity between the analysis of both readers.



One reader per patient	Several readers per patient
0.69 [0.59; 0.79]	0.50 [0.39, 0.61]



Automation & Reader's bias



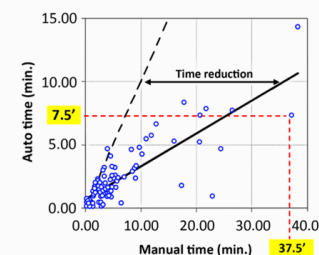
Similar Limit of Agreement between manual and semi-automated segmentations

Manual	Semi-Automatic
0.51 [0.32; 0.7]	0.69 [0.6; 0.79]

Improved inter-reader agreement according to Kappa coefficient with semi automated segmentation

For a given reader, mixing semi-automated and manual segmentations results in a poor intra-reader agreement (Kappa = 0.31 [0.10; 0.52])

Automation & Time Saving



Time required to perform the segmentation was reduced by a factor of 4 with no change in variability according to the Kruskal-Wallis test ($p > 0.05$)

For lesions larger than 12mm in diameter, automation allowed for a faster segmentation in 95% of cases.

Automation enabled volume measurement for large lesions that would otherwise be too time consuming with manual segmentation

Conclusions

- Standardized workflow is key in reducing variability. Involving multiple readers or mixing measurement paradigms in assessing a patient introduces strong discrepancies. A single reader using a single segmentation paradigm for all TPs of a given patient is the way to go. Baseline reassessment could be used as a contingency if several readers are involved.
- Automation, such as semi-automated segmentation, improves inter-reader agreement in reducing readers' individual biases.
- Automation brings strong time saving with no additional variability, making volume measurement for large lesions feasible

Acknowledgments

The authors thank MERCK SHARP & DOME Corp. and Dr. David Mozley for the data they provided and Simon Souchet (statistics) for his contribution at MEDIAN technologies.