Therapy Response Evaluation of Malignant Lymphoma in a Multicenter Study: Comparison of Manual and Semiautomatic Measurements in CT

Beurteilung des Therapieansprechens beim Malignen Lymphom: Multizentrischer Vergleich von manuellen und semiautomatischen Messungen im CT

Authors

J. Weßling1, C. Schülke1, R. Koch2, N. Kohlhase1, L. Wassenaar1, R. Mesters3, A. J. Höink1, M. D’Anastasi4, W. Heindel1, B. Buerke1

Affiliations

Affiliation addresses are listed at the end of the article.

Key words

- staging
- abdomen
- mediastinum
- thorax
- computed tomography
- lymphoma

Zusammenfassung

Ziel: Multizentrischer Vergleich von manuellen ein-/bi-dimensionalen Messungen und semi-automatischen ein-/bi-dimensionalen und volumetrischen Messungen zur Beurteilung des Therapieansprechens beim Malignen Lymphom in CT-Verlaufskontrollen.


Abstract

Purpose: Comparison of manual one-/bi-dimensional measurements versus semi-automatically derived one-/bi-dimensional and volumetric measurements for therapy response evaluation of malignant lymphoma during CT follow-up examinations in a multicenter setting.

Materials and Methods: MSCT data sets of patients with malignant lymphoma were evaluated before (baseline) and after two cycles of chemotherapy (follow-up) at radiological centers of five university hospitals. The long axis diameter (LAD), the short axis diameter (SAD) and the bi-dimensional WHO of 307 target lymph nodes were measured manually and semi-automatically using dedicated software. Lymph node volumetry was performed semi-automatically only. The therapeutic response was evaluated according to lymphoma-adapted RECIST.

Results: Based on a single lymph node, semi-automatically derived multidimensional parameters allowed for significantly more accurate therapy response classification than the manual or the semi-automatic unidimensional parameters. Incorrect classifications were reduced by up to 9.6%. Compared to the manual approach, the influence of the study center on correct therapy classification is significantly less relevant when using semi-automatic measurements.

Conclusion: Semi-automatic volumetry and bi-dimensional WHO significantly reduce the number of incorrectly classified lymphoma patients by approximately 9.6% in the multicenter setting in comparison to linear parameters. Semi-automatic quantitative software tools may help to significantly reduce wrong classifications that are associated with the manual assessment approach.

Key Points:

- Semi-automatic volumetry and bi-dimensional WHO significantly reduce the number of incorrectly classified lymphoma patients
and sollten daher insbesondere in klinischen Studien zukünftig aber auch in die klinische Routine implementiert werden.

Kernaussagen:
- Semi-automatisches Volumen und bi-dimensionaler WHO-Messungen reduzieren die Anzahl von Fehlklassifikationen beim Therapieansprechen signifikant (p < 0,05)
- Die manuelle Auswertung von Lymphknoten auf Basis unidimensionaler Parameter ist der semi-automatischen in einem Multicenter-Setting unterlegen
- Semi-automatische quantitative Softwaredtools sollten zur Auswertung in klinischen Studien obligat eingesetzt und zukünftig auch in der klinischen Routine implementiert werden.

Introduction

Revised RECIST 1.1 (Response Evaluation Criteria in Solid Tumors) as well as standardized Non-Hodgkin-Lymphoma (NHL) response criteria have underlined the importance of multidetector computed tomography (MDCT) as the primary lymph node imaging modality in clinical radiology practice [1 – 3]. Although firmly established in the setting of clinical trials, the need for systematic quantitative imaging in the daily routine is questioned to some extent by many radiologists [4]. However, with the worldwide introduction of comprehensive cancer centers, it has become much more apparent that many oncology patients are routinely included in clinical trials by tumor board decisions based – among other criteria – on quantitative imaging data. It is remarkable that 94% of oncologists at 55 U.S. cancer institutions expect oncology patients to undergo quantitative measurements regardless of enrollment in clinical trials [4, 5].

Changes in tumor size are routinely assessed using manually acquired metrics such as long axis diameter (LAD) and short axis diameter (SAD) [1, 6 – 11]. Manual acquisition of these uni-dimensional parameters bears inherent sources of error as demonstrated by the high interobserver and intraobserver variability [12, 13] potentially leading to misinterpretations in tumor response assessment [14].

Previous studies have already demonstrated the technical feasibility of methods for semi-automated lymph node measurement in oncology, specifically addressing measurement precision and the necessity of correction [15 – 19]. Robust, user-friendly semi-automatic tools in particular have shown greater reproducibility compared to their manual counterparts in the assessment of lymph nodes in various oncologic diseases [15, 16, 20].

In view of the increasing mobility of oncological patients between different medical centers, the influence of the reader (different readers in different institutions) becomes more apparent. Thereby, the prerequisites for a reproducible quantitative tumor burden assessment should lie between two extreme poles: a) variance of assessment allows only for a single-center assessment by one and the same radiologist over the whole course of the oncologic disease and b) the method of quantitative measurements is independent of the individual radiologist and institution. To the best of our knowledge, multi-center studies that comparatively define such prerequisites for a) a manual approach and b) a semi-automatic tumor burden assessment are lacking.

This multi-center study aims to determine the impact of manual and semi-automatic lymph node measurements, measurement parameters (uni- versus multidimensional) and readers (different centers) on therapy response classification in the follow-up of patients with malignant lymphoma.

Material and Methods

Patients

63 consecutive patients (male/female 40 (64 %)/23; 22 – 83 years, mean age of 56 ± 13 years) with histologically confirmed Hodgkin lymphoma (n = 10, 15.4 %) and non-Hodgkin lymphoma (n = 53, 84.6 %) including follicular lymphoma (15.4 %), mantle-cell lymphoma (6.1 %), marginal zone lymphoma (3.1 %), other indolent B-cell lymphoma (47.7 %) and T-cell lymphoma (12.4 %), were included in this retrospective study. The criteria for inclusion were a) initial diagnosis of lymphoma (88 %) or b) relapse of malignant lymphoma (12 %). Patients already on chemotherapy prior to CT were excluded. All patients underwent a contrast-enhanced MDCT scan prior to therapy for staging and after two cycles of chemotherapy (mean time between baseline and final staging: 106 days; range 15 – 448 days).

Written informed consent for MDCT was obtained from all patients before examination. The study was approved by the local ethics committee and conducted according to the guidelines of the institutional review board.

Data acquisition, preparation and transfer

Data acquisition

All examinations were performed at the main study center (study site 1) in order to minimize potential variations due to different scanner geometries and protocol parameters. The standardized CT examinations of the cervico-thoracic and abdominal region were performed using a 64 multislice CT scanner (Somatom Definition; Siemens Medical Solutions, Forchheim, Germany). The contrast agent (Ultravist 370®, Bayer Schering Pharma AG, Leverkusen, Germany) was applied with a constant injection rate of 3 ml/s. The scan delay was adapted to the anatomic regions (cervico-thoracic 45 s and abdominal 85 s). Images were obtained at 120 kV with a 32 × 0.6 mm² collimation, using a special dose-modulation template for radiation exposure reduction (CARE dose®) [21]. All CT data sets were reconstructed at a slice thickness of 1.5 mm with a reconstruction increment of 0.6 mm, which was revealed in a recent study as the optimal slice thickness for segmentation [15]. The scanning protocol did not differ from the standardized protocol used in the clinical routine.

Data preparation with labeling of target lymph nodes

At study site 1 CT data sets were transferred to a separate workstation (Oncology Prototype Software (Fraunhofer MEVIS, Sie-
mens Healthcare, Germany)) for lymph node selection and preparation including annotation. A radiologist unblinded to the diagnosis (4 years oncologic radiology experience) identified pathological target lymph nodes in the cervical, thoracic (axillary, mediastinal and hilar), abdominal (retroperitoneal, mesenteric), and pelvic (paraliiacal, inguinal) region. According to International Workshop Criteria (IWC) guidelines [2, 22], up to six target lymph nodes with an LAD > 15 mm were numbered digitally at baseline and the corresponding follow-up examination in order to minimize correlation and mapping errors, as may occur when readers have to search manually for target lymph nodes in follow-up images.

Data transfer and management
Baseline and follow-up CT data sets with the digitally labeled and numbered lymph nodes were stored on several identically equipped laptops (time measurement and automatic measurement data transfer program). These laptops were transferred from the main study center (study site 1) to four university radiology departments (study site 2, 3, 4, 5). After completion of analysis at each site, the laptops – complete with data sets and Excel® tables (see below) – were returned to the main study center (study site 1) for statistical analysis (Fig. 1).

Data evaluation
Manual evaluation
The lymph nodes were manually evaluated by two blinded radiologists at each study site (each with a minimum of 4 years oncologic radiology experience). Each radiologist separately and independently evaluated the digitally tagged lymph nodes. The data sets of the baseline and follow-up examinations were presented in a randomized fashion in order to avoid memory bias (with regard to diameter level and orientation). Manual assessment encompassed digital caliper measurements of LAD (mm) and SAD (mm) on axial CT images of the reader’s choice (cine mode). Manual bi-dimensional WHO (mm²) was calculated as the product of manual LAD and SAD.

Semi-automatic evaluation
Semi-automatic lymph node segmentation was performed by the same blinded radiologists at each site, separately and independently in a randomized fashion, using dedicated segmentation software. This software includes an algorithm for semi-automated lymph node evaluation based on an extended version of the lung lesion segmentation approach [13, 23 – 25]. The semi-automated segmentation process was started by drawing a stroke on the tagged lymph node of any particular slice. The volume of interest (VOI) and thresholds (histogram analysis within the VOI) for initial segmentation of the lymph node originating at the center of the stroke were estimated automatically. The initial segmentation results were displayed on the basis of region-growing-based algorithms, whereas ellipsoid approximation, distance map calculation and watershed algorithms separated adjacent structures of similar density such as blood vessels and muscle tissue. A 3D viewer, producing multiplanar reconstructions, delivered visual verification of the segmentation result. Dedicated correction tools could be used to modify any unsatisfactory segmentation results by drawing 2D contours on ill-segmented portions in any of the three 2D planes, followed by conversion into a 3D correction using an extrapolation process (Fig. 2). The following parameters were automatically displayed: LAD (mm), SAD (mm), volume (ml), and bi-dimensional WHO (mm²). Approved manual caliper and semi-automatic measurements at each site were transferred automatically into an Excel® table by dedicated software in order to prevent manual data transfer errors.

Time measurement
Time measurements taken with a stopwatch are subject to handling errors and limited assessment with a view to sub-processes. We therefore compiled a dedicated program for automatic time measurements without the need for interaction from the examining radiologist. During manual assessment, the time measurement was started when starting to scroll through the tagged lymph node (cine mode) and stopped on finalizing the LAD and SAD caliper measurements. The time for semi-automatic assessment was captured from the point of time at which a stroke was drawn on the tagged lymph node of any particular slice until correctness was verified. The correction time was recorded from activation of a correction tool until confirmation of correctness. An additional 4 – 6 s transfer time from the scanner to the workstation for manual and semi-automatic approach remained out of consideration.

Response assessment
Response criteria
To ensure the comparability of parameters in the same familiar units, measurements need to be converted and standardized as basically described by James et al. [26]. All volume and bi-dimensional measurements were therefore converted to diameters as recently published by different groups [16, 19, 27]. These effec-
Diameters were measured in mm and defined as volume-equivalent and area-equivalent diameters. The volume-equivalent diameter (DVOL; mm) was calculated by inverting the volume formula: $DVOL = \left( \frac{6 \cdot V}{\pi} \right)^{1/3}$, where $V = \text{volume} \ (\text{mm}^3)$ and $DVOL = \text{diameter}$ [23]. The area-equivalent diameter (DS; mm) was calculated using the following formula: $DS = 2\left( \frac{1}{\pi \cdot LAD \cdot SAD} \right)^{1/2}$. For clarity, the equivalent diameters are referred to below as "volume" and "bi-dimensional WHO".

Assouline et al. applied a modified RECIST concept to response assessment in lymphoma, using uni-dimensional tumor measurements [28]. We adapted this modified RECIST system in light of the uni-dimensionality of our parameters. The following response criteria modified from RECIST 1.1 were used to compare the manually and semi-automatically measured parameters in this study:

- $+20\% = \text{progressive disease}$; $-20\% < \text{to} \leq 20\% = \text{stable disease}$; $-50\% < \text{to} \leq -20\% = \text{good response}$; $-99\% \text{to} \leq -50\% = \text{very good response}$.

For the purposes of measurement and gaining a better impression of the effects of measurement errors over such a wide range, partial responses have been subdivided into "good" and "very good".

As in a study published recently, the standard of reference consisted of a combination of manual and semi-automatic LAD and SAD and the independently determined volume-equivalent and area-equivalent diameter [29].

**Response classification**

Response classification in this study was based on two different assumptions.

a) In order to avoid a selection and averaging bias and to examine the measurement quality and measurement precision of the different evaluation techniques, the "response classification per lymph node" was determined based on changes in the size of each single lymph node, irrespective of the patient concerned. As a restriction, this approach cannot be applied to clinically utilized classification systems.

b) The "response classification per patient" was based on target groups, i.e. in each patient the diameters of up to six target lymph nodes were summarized. The sum of each parameter was recorded at baseline and compared with the sum of the diameters at follow-up. Wrong classifications were assumed for sum diameter changes aberrant to the reference standard. This clinically applied classification system reduces measurement deviations by accepting averaging biases.

**Statistical analysis**

Statistical analyses were performed using SAS software, version 9.3 of the SAS system for Windows. Inferential statistics are intended to be exploratory (hypothesis generating), not confirmatory, and are interpreted accordingly. The comparison-wise type-I error rate is controlled instead of the experiment-wise error rate. The local significance level is set to 0.05. No adjustment
was made for multiple testing, hence an overall significance level was not determined and cannot be calculated.

Standard descriptive statistical analyses were performed for the target parameters of manual and semi-automatic LAD and SAD, semi-automatic bi-dimensional WHO and volume. Results are shown as mean values ± standard deviation.

In order to compare semi-automatic and manual time parameters, the Student's t-test for independent groups was applied to log-transformed time data. Time data are presented as median values [25% quantile, 75% quantile].

According to the relative change of lymph node sizes, a classification of response criteria was derived for each measurement parameter (see response criteria). The reference standard response was defined as the mean relative change across all parameters (i.e., reference = mean(relative change SAD manual, relative change SAD semi-automatic, relative change LAD manual, relative change LAD semi-automatic, relative change volume (as uni-dimensional equivalent diameters)). Each single parameter was compared to this reference standard with respect to the response classification. The classification results were described in terms of relative frequencies or odds ratios (95% confidence limits).

Situation A (response classification per lymph node) entailed the process of fitting generalized linear mixed models. The dependent variable was the binary response classification (right/ wrong). The logit function was chosen as the link function with binomial distribution as the corresponding distribution. The measurement method was treated as a fixed effect. In order to account for multiple ratings of one lymph node, measurement correlations from each individual rater, and dependencies between lymph nodes in one patient, the parameters lymph node, reader and patient parameters were modeled as random effects with a compound symmetry covariance structure. To compare semi-automatic with manual measurement classification, an additional, fixed effect.

Results

Lymph node characteristics

Table 1 provides a summary of the manual/semi-automatic measurement results. In total, 614 lymph nodes (307 baseline, 307 follow-up) were measured manually and semi-automatically in 63 patients (4.8 ± 3.3 lymph nodes/patient) at each site. The lymph nodes were evenly distributed in the thoracic (n = 129) and abdominal/pelvic (n = 125) region. Due to the relatively small-
Anzahl von Misklassifikationen um 9,6 % bei Verwendung von semi-automatischen bi-dimensionalem WHO und dem Volumen im Vergleich zu manuellem LAD und SAD reduziert.

Korrektheit der Beurteilung des Therapieansprechens anhand Annahme A)

Tab. 2 Beurteilung des Therapieansprechens in allen Studienzentren auf der Basis der in dieser Studie gemachten Annahmen (pro Lymphknoten/pro Patient).

<table>
<thead>
<tr>
<th>response classification</th>
<th>manual LAD</th>
<th>manual SAD</th>
<th>semi-automatic bi-dimensional WHO LAD</th>
<th>semi-automatic bi-dimensional WHO SAD</th>
<th>volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>A correct</td>
<td>1942 (79.5 %)</td>
<td>1884 (77.1 %)</td>
<td>2123 (86.9 %)</td>
<td>2030 (83.1 %)</td>
<td>2185 (89.4 %)</td>
</tr>
<tr>
<td>B incorrect</td>
<td>502 (20.5 %)</td>
<td>560 (22.9 %)</td>
<td>321 (13.1 %)</td>
<td>414 (16.9 %)</td>
<td>551 (22.5 %)</td>
</tr>
<tr>
<td>B correct</td>
<td>431 (84.7 %)</td>
<td>419 (82.3 %)</td>
<td>467 (91.8 %)</td>
<td>454 (89.2 %)</td>
<td>408 (80.2 %)</td>
</tr>
<tr>
<td>incorrect</td>
<td>78 (15.3 %)</td>
<td>90 (17.7 %)</td>
<td>42 (8.2 %)</td>
<td>55 (10.8 %)</td>
<td>101 (18.8 %)</td>
</tr>
<tr>
<td>false better</td>
<td>28 (5.5 %)</td>
<td>57 (11.2 %)</td>
<td>24 (4.7 %)</td>
<td>21 (4.1 %)</td>
<td>798 (15.5 %)</td>
</tr>
<tr>
<td>false worse</td>
<td>50 (9.8 %)</td>
<td>33 (6.5 %)</td>
<td>18 (3.5 %)</td>
<td>34 (6.7 %)</td>
<td>22 (4.3 %)</td>
</tr>
</tbody>
</table>

Correctness of therapy response classification according to A) "Response classification per lymph node" and B) "Response classification per patient". (A), Response classification was summarized and calculated across all study sites (n = 614 lymph nodes or n = 126 patients). Assumption B revealed a mean reduction in wrongly classified patients of 9.6 % for semi-automatic bi-dimensional WHO and volume compared to manual LAD and SAD.

Correktheit der Beurteilung des Therapieansprechens anhand Annahme A) „Beurteilung des Therapieansprechens pro Lymphknoten“ und B) „Beurteilung des Therapieansprechens pro Patient“. (A), Die Beurteilung des Therapieansprechens wurde über alle Studienzentren (n = 614 Lymphknoten oder n = 126 Patienten) bestimmt. Unter Annahme B wurde die Anzahl von Misklassifikationen um 9,6 % bei Verwendung von semi-automatischen bi-dimensionalem WHO und dem Volumen im Vergleich zu manuellem LAD und SAD reduziert.

Abb. 3 Zeitaufwand für die manuelle und semi-automatische Lymphknotensegmentierung mit und ohne Korrektur. Boxplot für den Zeitaufwand für die Lymphknotenauswertung. Der Median wird durch die dickere Linie innerhalb des Kastens angezeigt. Die horizontalen Kanten des Kastens verdeutlichen das obere und untere Quartil. Der mittlere Zeitaufwand für die semi-automatische Segmentierung ohne Korrektur (12.2 s [9.5 s/15.3 s]) entspricht dem bei manuellen Messungen (12.1 s [9.5 s/17.9 s]). Die Anwendung von Korrektur-Tools war in 56.6 % der Fälle erforderlich, was zu einem Anstieg des Zeitaufwands auf 38.2 s [26.7 s/83.1 s] führte. Der mittlere Zeitaufwand (23.0 s [12.9 s/42.4 s]) für semi-automatische Messungen (mit und ohne Korrektur) war signifikant höher im Vergleich zu den manuellen Messungen (p < 0.001). Bei der semi-automatischen Segmentierung ist im Gegensatz zu den manuellen Messungen die Dokumentation der Messergebnisse bereits eingeschlossen, während die manuell akquirierten Messergebnisse zusätzlich erfaßt werden müssen, was in einen höheren Zeitaufwand in der klinischen Routine führt.

Table 2 Therapy response classification across all sites based on the two different assumptions (per lymph node/per patient) in this study.

Fig. 3 Time expenditure for manual and semi-automatic lymph node evaluation with and without correction. Boxplot of time expenditure for lymph node evaluation. The median is indicated by the thicker black line within the box. The horizontal edges of the box display the upper and lower quartile. Median time expenditure for semi-automated segmentation without correction (12.2 s [9.5 s/15.3 s]) is equivalent to the conventional manual measurement approach (12.1 s [9.5 s/17.9 s]). The use of correction tools was necessary in 56.6 % of all cases and the time expenditure increased to 38.2 s [26.7 s/83.1 s]. Thus, the average time expenditure (23.0 s [12.9 s/42.4 s]) for semi-automatic measurements of all segmentations (with and without correction) was significantly higher in comparison to manual measurements (p < 0.001). However, automatic documentation of the measurement results is included in this time by semi-automated segmentation, whereas manually acquired results have to be documented manually, thereby increasing the total operation time in the clinical routine.

Fig. 4 illustrates the regression in size of an exemplary inguinal lymph node under therapy. Fig. 5 presents the response evaluation per lymph node across all study centers.

semi-automatic bi-dimensional WHO/semi-automatic volume 89.4%/87.0 % compared for example with manual LAD 79.5 %, p < 0.0001 (p < 0.0001). Furthermore, manual bi-dimensional WHO (86.9 %) was found to be significantly inferior (p = 0.0045) to its semi-automated counterpart (89.4 %) and semi-automatic volume (87.0 %, p = 0.0068).

Korrektheit der Beurteilung des Therapieansprechens anhand Annahme B) „Beurteilung des Therapieansprechens pro Patient“. (A), Die Beurteilung des Therapieansprechens wurde über alle Studienzentren (n = 614 Lymphknoten oder n = 126 Patienten) bestimmt. Unter Annahme B wurde die Anzahl von Misklassifikationen um 9,6 % bei Verwendung von semi-automatischen bi-dimensionalem WHO und dem Volumen im Vergleich zu manuellem LAD und SAD reduziert.
Tab. 3  Beurteilung des Therapieansprechens pro Lymphknoten bei manuellen und semi-automatischen Messungen (A).

<table>
<thead>
<tr>
<th>parameter</th>
<th>LAD manual</th>
<th>LAD semi-automatic</th>
<th>SAD manual</th>
<th>SAD semi-automatic</th>
<th>bi-dimensional WHO manual</th>
<th>bi-dimensional WHO semi-automatic</th>
<th>volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAD manual</td>
<td>X</td>
<td>3.6</td>
<td>2.4</td>
<td>2.0</td>
<td>7.4</td>
<td>0.9</td>
<td>7.5</td>
</tr>
<tr>
<td>LAD semi-automatic</td>
<td>0.0007</td>
<td>X</td>
<td>6.0</td>
<td>5.6</td>
<td>3.8</td>
<td>6.3</td>
<td>3.9</td>
</tr>
<tr>
<td>SAD manual</td>
<td>0.0326</td>
<td>&lt;0.0001</td>
<td>X</td>
<td>0.4</td>
<td>9.8</td>
<td>12.3</td>
<td>9.9</td>
</tr>
<tr>
<td>SAD semi-automatic</td>
<td>0.0702</td>
<td>&lt;0.0001</td>
<td>0.7435</td>
<td>X</td>
<td>9.4</td>
<td>11.9</td>
<td>9.5</td>
</tr>
<tr>
<td>bi-dimensional WHO manual</td>
<td>&lt;0.0001</td>
<td>0.0001</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
<td>X</td>
<td>2.5</td>
<td>0.1</td>
</tr>
<tr>
<td>bi-dimensional WHO semi-automatic</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
<td>0.0045</td>
<td>X</td>
<td>2.4</td>
</tr>
<tr>
<td>volume</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
<td>0.0068</td>
<td>0.8946</td>
<td>X</td>
</tr>
</tbody>
</table>


Table 4  Response classification per patient (B).

<table>
<thead>
<tr>
<th>parameter</th>
<th>LAD manual</th>
<th>LAD semi-automatic</th>
<th>SAD manual</th>
<th>SAD semi-automatic</th>
<th>bi-dimensional WHO manual</th>
<th>bi-dimensional WHO semi-automatic</th>
<th>volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAD manual</td>
<td>X</td>
<td>4.5</td>
<td>2.4</td>
<td>4.5</td>
<td>9.2</td>
<td>7.1</td>
<td>7.6</td>
</tr>
<tr>
<td>LAD semi-automatic</td>
<td>0.041</td>
<td>X</td>
<td>6.9</td>
<td>9.0</td>
<td>4.7</td>
<td>2.6</td>
<td>3.1</td>
</tr>
<tr>
<td>SAD manual</td>
<td>0.477</td>
<td>0.018</td>
<td>X</td>
<td>4.5</td>
<td>11.6</td>
<td>9.5</td>
<td>10.0</td>
</tr>
<tr>
<td>SAD semi-automatic</td>
<td>0.278</td>
<td>0.018</td>
<td>0.496</td>
<td>X</td>
<td>13.7</td>
<td>11.6</td>
<td>12.1</td>
</tr>
<tr>
<td>bi-dimensional WHO semi-automatic</td>
<td>0.003</td>
<td>0.018</td>
<td>&lt;0.001</td>
<td>&lt;0.0001</td>
<td>X</td>
<td>2.1</td>
<td>1.6</td>
</tr>
<tr>
<td>bi-dimensional WHO manual</td>
<td>0.001</td>
<td>0.152</td>
<td>&lt;0.001</td>
<td>0.003</td>
<td>0.300</td>
<td>X</td>
<td>0.5</td>
</tr>
<tr>
<td>volume</td>
<td>0.012</td>
<td>0.170</td>
<td>0.003</td>
<td>0.002</td>
<td>0.453</td>
<td>0.811</td>
<td>X</td>
</tr>
</tbody>
</table>


Per study site  
- Fig. 6, 7 illustrate the percentage of correct therapy response classifications per study site, as well as for each manual and semi-automated parameter. Irrespective of the evaluation assumption (response classification per lymph node or per patient), the fraction of correctly classified therapy responses was found to be consistently and significantly higher for multi-dimensional parameters (e.g. manual or semi-automatic bi-dimensional WHO and volume) as compared to unidimensional parameters obtained either manually or semi-automatically.

Influence of the measurement approach (manual versus semi-automatic) on correct lesion classification

The precision of the therapy response classification was significantly affected by the measurement approach, whether manual or semi-automatic. However, with an odds ratio of 1.18 times
(95 % CI 1.08 – 1.29, p = 0.0003) the probability of correct classification was significantly higher using the semi-automatic instead of the manual approach.

**Center-specific influence on correct patient classification**

As revealed by the generalized linear mixed model analyses, the study center has a significant influence on therapy response classification, irrespective of the chosen approach (manual or semi-automatic). Compared to the manual approach, however, the influence of the study center on correct therapy classification is significantly less relevant when semi-automatic methods are used.

**Discussion**

In oncological decision processes manually obtained tumor metrics in CT imply a degree of precision which has to be viewed critically in terms of measurement variability and reproducibility [12, 16, 18, 25, 29]. Consequently, the Quantitative Imaging Biomarker Alliance of the Radiological Society of North America encourages further multidisciplinary research into the enhancement of the value and practicality of quantitative imaging, especially by reducing variability across devices, patients and time [31]. In the past decade, semi-automatic tumor segmentation and measurement tools have demonstrated their technical feasibility with regard to lung nodule and liver lesion segmentation [13, 20, 32 – 34]. Recent studies specifically addressed the aspects of reproducibility and variability between different readers in the semi-automated segmentation of tumor-affected lymph nodes, and found inter-user differences to be reduced by a factor of approximately 1.4 to 3.0 compared to the manual approach [16, 19, 35].

These feasibility studies are limited by their lack of data on follow-up examinations, whereby inter-user differences could be aggravated, e.g. due to variable lymph node orientation caused by shrinkage. In view of the increasing mobility of oncological patients, the influence of the reader (different readers in different institutions) on the assessment of follow-up examinations and therapy response classification is becoming more apparent. The ideal quantitative measurement tool should naturally provide highly reproducible measurements that are more or less uninfluenced by the attending radiologist and institution, do not demand an excessive amount of time and are robust when it comes to their use in the clinical routine. To the best of our knowledge, it is not yet clear whether semi-automatic software tools and the derived uni- and multidimensional parameters (volumetry) harbor such potential with respect to assessing therapeutic response.

We addressed these key questions in a multicenter setting involving a total of five university sites. We adopted a modified RECIST approach to data analysis according to Assouline et al. in order to detect and reliably classify minor changes close to decision-relevant remission limits [28]. Response classification in this study was based, furthermore, on two different assumptions, namely the presence of a lymph node and a clinical patient-based classification. In line with this and the IWC guidelines, the diameters of up to six target lymphoma lesions were summarized and compared between baseline and follow-up [2, 22]. This differentiation was essential for unmasking the selection and averaging biases of clinically applied classification systems based on target groups with sum diameters.

Only rudimentary investigations into the effects of interobserver variability on tumor response classification in follow-up examinations have been undertaken. Fabel et al. [18] did not find any differences in response classification when using semi-automatic measurements in melanoma patients, but admitted to constraints in the selected classification limits. In another recently published single-center study of semi-automatic lymph node segmentation, semi-automatic volumetry and bi-dimensional WHO permitted classifications that were significantly more accurate than those based on manual one-dimensional diameters [29]. According to this study, one of the main findings is that, on a “per lymph node” as well as on a “per patient” basis, multidimensional parameters – whether obtained manually or semi-automatically – allowed for a significantly more accurate therapy response classification than uni-dimensional parameters (e.g. volume 87.0 % vs. manual SAD 79.5 %, p < 0.001). In this study, these findings were confirmed for all study centers, irrespective of the anatomic region. The inferior performance of the one-dimensional parameters SAD and LAD in our study is therefore an argument against proposals to promote uni-dimensional metrics in follow-up assessments of malignant lymphomas [22].

On the patient level, semi-automatic “volumetry” and “bi-dimensional WHO” significantly reduced the number of wrongly classified lymphoma patients consistently across all study sites by approximately 9.6 % (7.9 – 13.5 %), thus confirming the results of an
earlier single-center study [29] on semi-automatic lymph node segmentation, which transferred lower measurement variability into a reduction of wrongly classified lymphoma patients of 10%. There is a further implication from our data, namely that manual bi-dimensional WHO was comparable to its semi-automatic counterpart and semi-automatic volumetry on a patient level. Consequently, the relevance of using semi-automatic software tools, especially in clinical trials, has to be questioned. On a per lymph node level, however, the precision of manual WHO (86.9%) was found to be significantly inferior to semiautomatic WHO (89.4%, \( p = 0.0045 \)) and semi-automatic volumetry (87.0%, \( p = 0.0068 \)).

Indeed, a number of patients have only a limited number of target lesions, e.g. two or three, and the radiological approach (manual vs. semi-automatic) becomes a significant factor in correct response classification. Therefore, it seems reasonable to use semi-automatic multidimensional parameters in a clinical or study setting, while we do not see any advantages for semi-automatic volumetry over semi-automatic bi-dimensional WHO. Our results furthermore indicate that the correctness of therapy response classification across all study centers is significantly affected by both measurement approaches (manual or semi-automatic). The odds ratios showed a 1.18 times (95 CI: 1.08 – 1.29) higher probability of correct patient classification (\( p = 0.0003 \)).
using the semi-automatic instead of manual approach. As revealed by the generalized linear mixed model analyses, we found the study center to have a significant influence on the therapy response classification and also found the semi-automatic quantitative measurement tools to be dependent on the radiologist and institution concerned. However, compared to the manual approach, the influence of the study center on therapy classification is significantly less relevant when using the semi-automatic method. Semi-automatic quantitative software tools may therefore help to significantly reduce wrong classifications that arise from manual assessment and the examining institution, thus favoring semi-automatic therapy evaluation, especially in a study environment.

Time expenditure – among other criteria – has a decisive influence on the acceptance and dissemination of segmentation software tools in oncology. In line with a recent study [29], semi-automatic lymph node segmentation in our investigations allowed for true-to-detail lymph node segmentation across all study sites at the first attempt in 43.4% of lymph nodes with a comparable time expenditure as compared to the manual approach (12.1 s manually vs. 12.2 s semi-automatically). The evaluation time for all segmentations at all study sites using the semi-automatic software tool was almost twice that of manual evaluation (12.1 s manually vs. 23.0 s semi-automatically). The increased time expenditure for the semi-automatic approach is highly consistent with the results published by Fabel et al. (mean 37s; range 20 – 70s [35]), but has to be put into perspective: Uni-dimensional and multi-dimensional measurements are automatically displayed without the need for further manual interaction and are automatically transferrable into oncologic reporting systems in
an RIS/PACS environment. In the manual approach, the docu-
mentation and calculation of manual WHO – which was not in-
cluded in the time measurements of this study – have to be per-
formed by the attending radiologist and are likely to increase the
total operation time as well as cause transfer biases through
manual interactions.

This study is limited to the extent that it does not allow compar-
ison with an exact reference standard, as is the case with a phan-
tom study. We used the manually and semi-automatically ob-
tained metric parameter as an internal reference standard,
which is accepted in the literature for the analysis of segmenta-
tion results, e.g. in pulmonary nodules and lymph nodes [13, 16].

The transferability of these results from lymphoma patients to
other malignant diseases seems reasonable but must be sup-
ported by additional studies. Furthermore, correlation analysis
of therapy response evaluation based on semi-automatic lymph
node measurements and therapy outcome was not covered by
this study and should be included in a future analysis. This study
also departs from previous studies by being the first to provide
data on multi-observer/multicenter variability and the influence
on therapy response classification. In summary this multicenter study revealed semi-automatic seg-
mentation to be robust and time efficient, with acceptable time
In summary this multicenter study revealed semi-automatic seg-
mentation results, e.g. in pulmonary nodules and lymph nodes [13, 16].

The transferability of these results from lymphoma patients
to other malignant diseases seems reasonable but must be sup-
ported by additional studies. Furthermore, correlation analysis
departs from previous studies by being the first to provide

therapy outcome was not covered by
this study and should be included in a future analysis. This study
also departs from previous studies by being the first to provide
data on multi-observer/multicenter variability and the influence
on therapy response classification.

In this study multicenter study revealed semi-automatic seg-
mentation to be robust and time efficient, with acceptable time
expenditure compared to conventional manual lymph node as-
essment. With regard to therapy response classification, semi-
automated multidimensional parameters (“volumetry” and “bi-
dimensional WHO”) significantly reduce the number of wrongly
classified lymphoma patients across all study sites by approxi-
ately 9.6 % (interval across all study sites: 7.9 – 13.5 %, p < 0.05) and
permit a significantly more accurate therapy response classifica-
tion than uni-dimensional parameters. Semi-automatic
quantitative software tools may help to significantly reduce
wrong classifications that arise from manual assessment meth-
ods and differences between the examining institutions. In con-
clusion, semi-automatic quantitative software tools should be
implemented in clinical studies and desirably in the clinical rou-
tine.

Affiliations
1 Dept. of Clinical Radiology, University of Muenster, Germany
2 Dept. of Medical Informatics and Biomathematics, University of Muenster, Germany
3 Dept. of Oncology, University of Muenster, Germany
4 Dept. of Clinical Radiology, University of Munich, Germany
5 Clinic for Diagnostic Radiology, University Hospital Schleswig-Holstein
Campus Kiel, Germany
6 Department for Diagnostic and Interventional Radiology, University of
Mainz, Germany
7 Dept. of Diagnostic Radiology, University of Marburg, Germany
8 MEVIS, Fraunhofer, Bremen, Germany

References
1 Eisenhauer EA, Therasse P, Bogaerts J et al. New response evaluation
criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J
Cancer 2009; 45: 228 – 247
2 Cheson BD, Pfisterer B, Juweid ME et al. Revised response criteria for ma-
3 Schwartz LH, Bogaerts J, Ford R et al. Evaluation of lymph nodes with
4 Jaffe TA, Wickersham NW, Sullivan DC. Quantitative imaging in oncolo-
gy patients: Part 1, radiology practice patterns at major U.S. cancer
5 Jaffe TA, Wickersham NW, Sullivan DC. Quantitative imaging in oncolo-
6 Erasmus JJ, Gladish GW, Bremel RJ et al. Interobserver and intraob-
server variability in measurement of non-small-cell carcinoma lung
2003; 21: 2574 – 2582
7 Castlejöa, van den Brekel MW. Imaging of lymphadenopathy in the
8 Steinkamp Hj, Cornehl M, Hosten N et al. Cervical lymphadenopathy:
ratio of long- to short-axis diameter as a predictor of malignancy. Br J
Radiol 1995; 68: 266 – 270
9 Steinkamp Hj, Hosten N, Richter C et al. Enlarged cervical lymph nodes
at helical CT. Radiology 1994; 191: 795 – 798
10 Torabi M, Aquino SL, Harisinghani MG. Current concepts in lymph node
11 van den Brekel MW, Castlejöa, Snow GB. Imaging of cervical lymph-
12 van den Brekel MW, Stel HV, Castlejöa et al. Cervical lymph node
metastasis: assessment of radiologic criteria. Radiology 1990; 177:
379 – 384
13 Wermanns D, Kohl G, Klotz E et al. Volumetric measurements of pulmo-
nary nodules at multi-row detector CT: in vivo reproducibility. Eur
Radiol 2004; 14: 86 – 92
14 Schwartz LH, Collville JA, Ginsberg MS et al. Measuring tumor response
and shape change on CT: esophageal cancer as a paradigm. Ann Oncol
2006; 17: 1018 – 1023
15 Buerke B, Puesken M, Beyer F et al. Semiautomatic lymph node segmen-
tation in multislice computed tomography: impact of slice thickness
on segmentation quality, measurement precision, and interobserver
variability. Invest Radiol 2010; 45: 82 – 88
16 Buerke B, Puesken M, Muenter S et al. Measurement accuracy and repro-
ducibility of semiautomated and volumetric lymph node analysis
in MDCT. Am J Roentgenol 2010; 195: 979 – 985
17 Puesken M, Buerke B, Gerss J et al. Prediction of lymph node manifesta-
tions in malignant lymphoma: significant role of volumetric compared
with established metric lymph node analysis in multislice computed
tomography. J Comput Assist Tomogr 2010; 34: 564 – 569
18 Fabel M, Biederer J, Jochens A et al. Semi-automated volumetric analysis
of artificial lymph nodes in a phantom study. Eur J Radiol 2011; 80:
451 – 457
19 Fabel M, von Tengg-Kobligk H, Giesel FL et al. Semi-automated volumet-
ic analysis of lymph node metastases in patients with malignant mel-
1122
20 Puesken M, Juergens KI, Edenfeld A et al. Semiautomatic Segmenta-
tierung von Leberlasionen in der MSCT: Einfluss der Schichtdicke
die Segmentierungsgüte, Messgenauigkeit und Interobservervaria-
tilität. Fortschr Röntgenstr 2009; 181: 67 – 73
21 Greess H, Nomayr A, Wolf H et al. Dose reduction in CT examination of
children by an attenuation-based on-line modulation of tube current
22 Cheson BD. Staging and evaluation of the patient with lymphoma. He-
23 Bondiau PY, Malandain G, Chanulet S et al. Atlas-based automatic seg-
mentation of MR images: validation study on the brainstem in radio-
24 Kahlng JK, Dicken V, Bornemann L et al. Morphological segmentation
and partial volume analysis for volumetry of solid pulmonary lesions
26 James K, Eisenhauer E, Christian M et al. Measuring response in solid
tumors: unidimensional versus bidimensional measurement. J Natl
Cancer Inst 1999; 91: 523 – 528
27 Zhao B, Schwartz LH, Moskowski CS et al. Lung cancer: computerized
quantification of tumor response – initial results. Radiology 2006; 241:
892 – 898
28 Assouline S, Meyer RM, Infante-Rivard C et al. Development of adapted
RECIST criteria to assess response in lymphoma and their comparison
to the International Workshop Criteria. Leuk Lymphoma 2007; 48:
513 – 520
29 Wessling J, Puesken M, Koch R et al. MSCT-Verlaufs kontrolle beim ma-
ligen Lymphom: Vergleich manueller linearer Messungen mit semi-
automatischen Lymphknotensegmentierungen zur Beurteilung des Therapieansprechens. Fortschr Röntgenstr 2012; 184: 795–804


