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. Weßling[1], C. Schülke[1], R. Koch[2], N. Kohlhase[1], L. Wassenaar[1], R. Mesters[3], A. J. Höink[1], M. D´Anastasi[4], M. Karpitschka[4], M. Fabel[5], A. M. Wulff[5], D. Pinto dos Santos[6], A. Kiessling[7], A. Graser[4], L. Bornemann[8], V. Dicken[8], W. Heindel[1], B. Buerke[1]

Affiliations Affiliation addresses are listed at the end of the article.

Key words

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

received 25.6.2013
accepted 10.12.2013

Bibliography
DOI http://dx.doi.org/10.1055/s-0033-1356424
Published online: 4.2.2014
Fortschr Röntgenstr 2014; 186: 768–779 © Georg Thieme Verlag KG Stuttgart · New York · ISSN 1438-9029

Correspondence
Priv.-Doz. Dr. Boris Buerke
Department of Clinical Radiology, University of Muenster
Albert-Schweitzer-Campus 1, Building A1
48149 Muenster
Germany
Tel.: ++ 49/2 51/8 34 58 91
Fax: ++ 49/2 51/8 34 51 27
buerkeb@uni-muenster.de

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 Lymphopm in CT-Verlaufskontrollen.

Material und Methoden: MSCT-Datensätze von Patienten mit Malignem Lymphom wurden vor (baseline) und nach zwei Zyklen Chemotherapie (follow-up) in fünf Universitätsradiologien ausgewertet. Der Langachsen- (LAD), der Kurzachsendurchmesser (SAD) und die bi-dimensionale WHO-Fläche von 307 Target-Lymphknoten wurden manuell und semi-automatisch unter Verwendung einer dedizierten Software bestimmt. Die Lymphknotenvolumetrie wurde lediglich semi-automatisch bestimmt. Das Therapieansprechen wurde anhand Lymphom-adaptierter RECIST-Kriterien beurteilt.

Ergebnisse: Auf der Basis des einzelnen Lymphknotens wiesen semi-automatisch bestimmte mehrdimensionale Parameter einen höheren Anteil in der korrekten Beurteilung des Therapieansprechens als die die manuell oder semi-automatisch bestimmten eindimensionalen Parameter auf. Fehlklassifikationen wurden um bis 9,6 % reduziert. Zusätzlich war im Vergleich zu den manuellen Messungen der Einfluss auf die Beurteilung des Therapieansprechens in den einzelnen Studienzentren bei Anwendung semi-automatischer Messungen geringer.

Schlussfolgerung: Die semi-automatische Volumetrie und die bi-dimensionale WHO-Messung reduzieren die Anzahl von Fehlklassifikationen in der Beurteilung des Therapieansprechens bei Patienten mit Malignem Lymphom signifikant um 9,6 % in einem Multicenterumfeld im Vergleich zu linearen Parametern. Semi-automatische Software-Tools können dazu beitragen, Fehlklassifikationen manueller Messungen zu reduzieren

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 und 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 uni-dimensionaler Parameter ist der semi-automatischen in einem Multicenter-Setting unterlegen
- ▶ Semi-automatische quantitative Softwaretools sollten zur Auswertung in klinischen Studien obligat eingesetzt und zukünftig auch in der klinischen Routine implementiert werden.
- ▶ Manual lymph node evaluation with uni-dimensional parameters is inferior to semi-automatic analysis in a multicenter setting
- ▶ Semi-automatic quantitative software tools should be introduced in clinical study evaluation.

Citation Format:

- ▶ Weßling J, Schülke C, Koch R et al. Therapy Response Evaluation of Malignant Lymphoma in a Multicenter Study: Comparison of Manual and Semiautomatic Measurements in CT. Fortschr Röntgenstr 2014; 186: 768–779

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, Siemens 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 (parailiacal, 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 (mm^2) 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 3 D 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 2 D contours on ill-segmented portions in any of the three 2 D planes, followed by conversion into a 3 D correction using an extrapolation process (**Fig. 2**). The following parameters were automatically displayed: LAD (mm), SAD (mm), volume (ml), and bi-dimensional WHO (mm^2).

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-

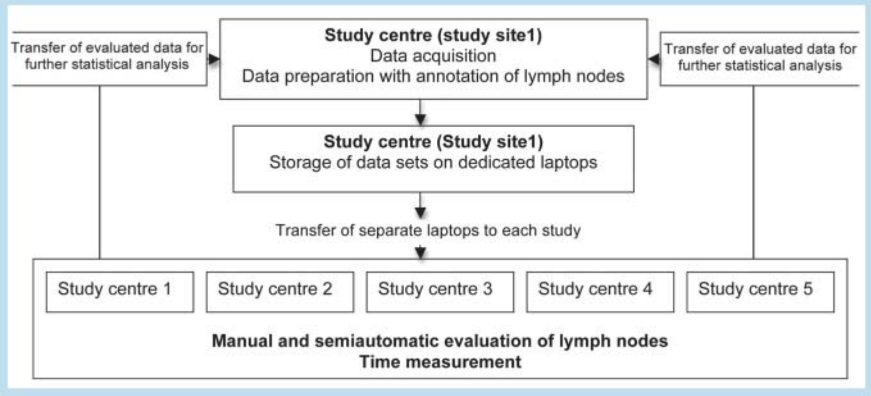
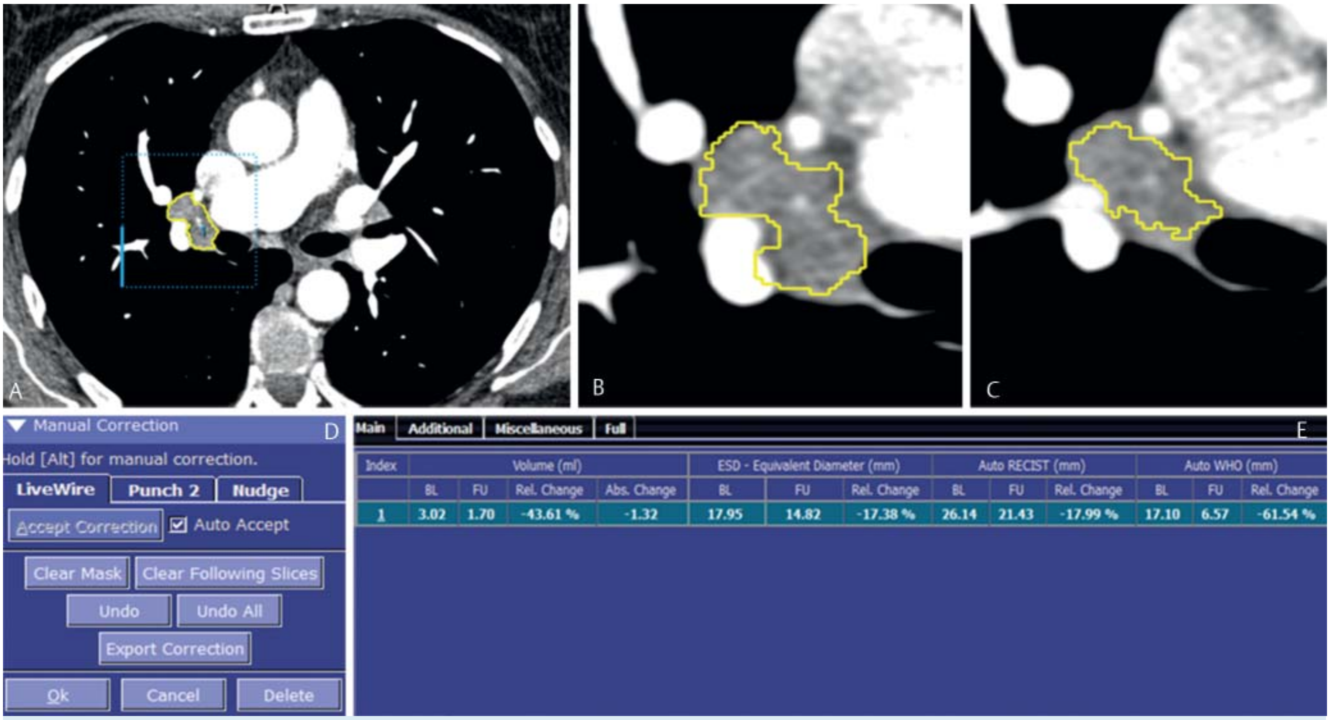

Transfer of evaluated data for further statistical analysis
Study centre (study site1)
Data acquisition
Data preparation with annotation of lymph nodes
Transfer of evaluated data for further statistical analysis
Study centre (Study site1)
Storage of data sets on dedicated laptops
Transfer of separate laptops to each study
Study centre 1
Study centre 2
Study centre 3
Study centre 4
Study centre 5
Manual and semiautomatic evaluation of lymph nodes
Time measurement

Fig. 1 Data transfer and management. The flow chart outlines the organizational structure and data management in this multicenter study with five study sites.

Abb. 1 Datentransfer und Studienmanagement. Das Flowchart gibt einen Überblick über den organisatorischen Aufbau und das Studienmanagement dieser Multicenter-Studie mit fünf Studienzentren.



Index	Volume (ml)				ESD - Equivalent Diameter (mm)			Auto RECIST (mm)			Auto WHO (mm)		
	BL	FU	Rel. Change	Abs. Change	BL	FU	Rel. Change	BL	FU	Rel. Change	BL	FU	Rel. Change
1	3.02	1.70	-43.61 %	-1.32	17.95	14.82	-17.38 %	26.14	21.43	-17.99 %	17.10	6.57	-61.54 %

Fig. 2 Semi-automatic segmentation of a hilar lymph node. The baseline and follow-up scans of each patient are shown on one screen in multiplanar reconstructions **A**. The segmentation result of a malignant lymph node in the baseline and follow-up examination is displayed (yellow lines) a few seconds after starting segmentation **B, C**. If the initial segmentation result is inadequate, corrections can be made using dedicated software tools **D**. Once the segmentation result is confirmed as adequate and corrections have been made, a synopsis of the different segmentation parameters (e. g. volume) with changes in size under therapy is generated automatically **E**.

Abb. 2 Semi-automatische Segmentierung eines hilären Lymphknotens. Baseline and Follow-up Untersuchungen werden in multiplanaren Rekonstruktionen angezeigt **A**. Das Ergebnis der Segmentierung des malignen Lymphknotens in der Baseline- und der Follow-up-Untersuchung wird innerhalb von wenigen Sekunden nach Beginn der Segmentierung angezeigt (gelbe Linie) **B, C**. Bei inadequatem Segmentierungsergebnis können Korrekturen mit entsprechenden Software-Tools korrigiert werden **D**. Nach Bestätigung des Segmentierungsergebnisses werden die verschiedenen Parameter (z. B. Volumen) in einer Übersicht unter Angabe der Größenänderungen unter Therapie angezeigt.

tive diameters were measured in mm and defined as volume-equivalent and area-equivalent diameters. The volume-equivalent diameter (D_{VOL}; mm) was calculated by inverting the volume formula: $D_{VOL} = (6^*V / \pi)^{1/3}$, where V = volume ($mm^3$) and D_{VOL} = diameter [23]. The area-equivalent diameter (D_S; mm) was calculated using the following formula: $D_S = 2(1/\pi^*LAD^*SAD)$^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 % = progressive disease; -20 % < to ≤ 20 % = stable disease; -50 % < to ≤ -20 % = good response; -99 % to ≤ -50 % = 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 semiautomatic with manual measurement classification, a dummy variable was included as a fixed effect. The same was done to compare one-dimensional and multidimensional measurement classifications. To detect differences between anatomic regions, additional generalized linear mixed models were computed with the anatomic region of the lesion as a fixed effect. The influence of the study centers on classification results was also analyzed separately in terms of manual, semiautomatic, uni-dimensional and multidimensional parameters with the study center as an additional, fixed effect.

In situation B (response classification per patient), McNemar's test for clustered data was applied to compare the frequencies of agreement with the reference standard between the measurement methods. The resulting p-values were adjusted for repeated ratings of each lesion by all readers [30]. Further subgroup analyses were performed for each anatomic region and each study center.

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 smaller anatomic volume, fewer lymph nodes were tagged in the cervical (n = 53) region.

Table 1 Manual and semi-automatic lymph node analysis.

Tab. 1 Manuelle und semi-automatische Messungen von Lymphknoten.

	baseline	**follow-up**
parameter	**mean ± SD (all sites)**	**mean ± SD (all sites)**
manual		
LAD [mm]	24.2 ± 9.9	17.5 ± 9.1
SAD [mm]	15.9 ± 7.3	10.9 ± 6.3
bi-dimensional WHO [mm]	22.0 ± 9.2	15.4 ± 8.2
semi-automatic		
LAD [mm]	24.4 ± 10.3	17.5 ± 9.4
SAD [mm]	15.9 ± 7.7	10.7 ± 6.7
bi-dimensional WHO [mm]	22.1 ± 9.8	15.3 ± 8.7
Volume [mm]	20.5 ± 8.9	14.9 ± 8.0

Mean ± standard deviation (SD) of manual and semi-automatic long axis diameter (LAD, mm) and short axis diameter (SAD, mm) for baseline and follow-up examinations across all study sites. Bi-dimensional WHO and volume are given as unidimensional equivalent diameters in mm. Lymph nodes in the cervical, thoracic (axillary, mediastinal and hilar), abdominal (retroperitoneal, mesenteric), and pelvic (parailiac, inguinal) region were analyzed.

Mittelwerte ± Standardabweichung (SD) von manuellem und semi-automatischem Langachsendurchmesser (LAD, mm) und Kurzachsendurchmesser (SAD, mm) in den Baseline- und Follow-up-Untersuchungen über alle Studienzentren. Bi-dimensionaler WHO und Volume werden als äquivalente eindimensionale Durchmesser in mm angegeben. Analysiert wurden cervikale, thorakale (axillär, mediastinale und hilär) und abdominale Lymphknoten (retroperitoneal, mesenterial, parailiakal und inginal).

Time evaluation

Fig. 3 summarizes the time expenditure. Manual lymph node measurements (LAD and SAD) took a median time of 12.1 s [9.5 s/17.9s] on average across all sites and readers. Without any further need for correction (41.9 % of all cases), semi-automated segmentation (12.2s; [9.3 s/16.0s]) was equivalent to the manual approach at all study sites. The time required for semi-automatic segmentation increased consecutively from 12.2 s [9.5 s/15.3s] to 38.2 s [26.7 s/ 57.8s] when correction tools were used (56.6 % of all cases). With regard to all lymph node segmentations (with and without the use of correction tools), the median time of 23.0 s [12.9 s/ 42.4s] was significantly higher than with the manual approach (12.2 s, p < 0.001).

Therapy response classification

Over all sites

Table 2 details the numbers of correctly and incorrectly classified therapy response (A. per lymph node and B. per patient classification) across all centers. In addition, **Table 3, 4** outline the corresponding levels of significance for each of the different manual and semi-automatic parameters.

a) *"Response classification per lymph node":* Across all sites, the precision of the correct therapy response classification was comparable between the manual uni-dimensional parameters (LAD 79.5 %/SAD 77.1 %) and the semi-automatic SAD (SAD 77.5 %). Statistical significance was observed for semi-automatic LAD (83.1 %) compared to manual LAD/SAD (p = 0.0007/p < 0.0001). Semi-automatically derived multidimensional parameters allowed for a significantly more accurate response to therapy classification than either the manual or the semi-automatic uni-dimensional parameters (e.g.

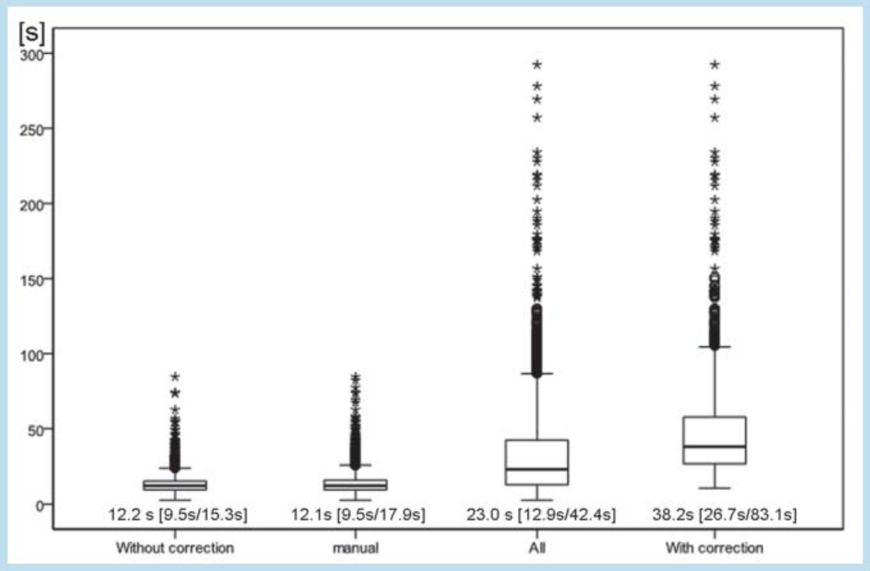


Fig. 3 Time expenditure for manual and semi-automatic lymph node evaluation with and without correction. Boxplot of time expenditure for lymph node evaluation. The median is indicated by the thicker black line within the box. The horizontal edges of the box display the upper and lower quartile. Median time expenditure for semi-automatic 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-automatic segmentation, whereas manually acquired results have to be documented manually, thereby increasing the total operation time in the clinical routine.

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.

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

		manual			semi-automatic			
	response classification	**LAD**	**SAD**	**bi-dimensional WHO**	**LAD**	**SAD**	**bi-dimensional WHO**	**volume**
A	correct	1942 *(79.5 %)*	1884 *(77.1 %)*	2123 *(86.9 %)*	2030 *(83.1 %)*	1893 *(77.5 %)*	2185 *(89.4 %)*	2126 *(87.0 %)*
	incorrect	502 *(20.5 %)*	560 *(22.9 %)*	321*(13.1 %)*	414 *(16.9 %)*	551 *(22.5 %)*	259 *(10.6 %)*	318 *(13.0 %)*
B	correct	431 *(84.7 %)*	419 *(82.3 %)*	467 *(91.8 %)*	454 *(89.2 %)*	408 *(80.2 %)*	478 *(93.9 %)*	470 *(92.3 %)*
	incorrect	78 *(15.3 %)*	90 *(17.7 %)*	42 *(8.2 %)*	55 *(10.8 %)*	101 *(19.8 %)*	31 *(6.1 %)*	39 *(7.6 %)*
	false better	28 *(5.5 %)*	57 *(11.2 %)*	24 *(4.7 %)*	21 *(4.1 %)*	798 *(15.5 %)*	24 *(4.7 %)*	9 *(1.8 %)*
	false worse	50 *(9.8 %)*	33 *(6.5 %)*	18 *(3.5 %)*	34 *(6.7 %)*	22 *(4.3 %)*	7 *(1.4 %)*	30 *(5.9 %)*

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.

Korrektheit 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.

semi-automatic bi-dimensional WHO/semi-automatic volumetry 89.4 %/87.0 % compared for example with manual LAD 79.5 %, p < 0.00 001/p < 0.00 001). Furthermore, manual bi-dimensional WHO (86.9 %) was found to be significantly inferior (p = 0.0045) to its semi-automatic counterpart (89.4 %) and semi-automatic volumetry (87.0 %, p = 0.0068). ▶ **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.

b) *"Response classification per patient"* revealed that significantly (each p-value p < 0.012) more patients were correctly classified by using multidimensional semi-automatic measure-

Table 3 Response classification per lymph node based on manual and semi-automatic measurements (A).

Tab. 3 Beurteilung des Therapieansprechens pro Lymphknoten bei manuellen und semi-automatischen Messungen (A).

parameter	LAD	LAD	SAD	SAD	bi-dimensional WHO	bi-dimensional WHO	volume
	manual	semi-automatic	manual	semi-automatic	manual	semi- automatic	
LAD manual	X	*3.6*	*2.4*	*2.0*	*7.4*	*9.9*	*7.5*
LAD semi-automatic	0.0007	X	*6.0*	*5.6*	*3.8*	*6.3*	*3.9*
SAD manual	0.0326	< 0.0001	X	*0.4*	*9.8*	*12.3*	*9.9*
SAD semi-automatic	0.0702	< 0.0001	0.7435	X	*9.4*	*11.9*	*9.5*
bi-dimensional WHO manual	< 0.0001	0.0001	< 0.0001	< 0.0001	X	*2.5*	*0.1*
bi-dimensional WHO semi-automatic	< 0.0001	< 0.0001	< 0.0001	X< 0.0001	0.0045	X	*2.4*
volume	< 0.0001	< 0.0001	< 0.0001	< 0.0001	0.0068	0.8946	X

Results of the generalized linear mixed model: dependent variable = binary response classification, fixed effect = measurement method, random effects = patient, rater. Repeated observations = lymph node (all compound symmetry covariance structure). Volumetry and semi-automatic bi-dimensional WHO revealed significantly fewer misclassifications compared to all unidimensional parameters (i. e. LAD, SAD), whether derived manually or semi-automatically. This was consistent across all sites. Additionally, the differences (modulus) between the relative numbers (%) of correct therapy evaluation are displayed in italics.

Ergebnisse des Generalized Linear Mixed Models. Bei Volumen und semi-automatischem bi-dimensionalem WHO zeigten sich signifikant weniger Fehlklassifikationen im Vergleich zu allen unidimensionalen Parametern (z. B. LAD, SAD) sowohl manuell als auch semi-automatisch. Diese Ergebnisse waren konsistent in allen Studienzentren. Zusätzlich sind Unterschiede als Betrag zwischen den relativen Anteilen (%) der korrekten Beurteilungen des Therapieansprechens unter Berücksichtigung der verschiedenen Parameter kursiv angegeben.

Table 4 Response classification per patient (B).

Tab. 4 Beurteilung des Therapieansprechens pro Patient (B).

parameter	LAD	LAD	SAD	SAD	bi-dimensional WHO	bi-dimensional WHO	volume
	manual	semi-automatic	manual	semi-automatic	semi-automatic	manual	
LAD manual	X	*4.5*	*2.4*	*4.5*	*9.2*	*7.1*	*7.6*
LAD semi-automatic	0.041	X	*6.9*	*9.0*	*4.7*	*2.6*	*3.1*
SAD manual	0.477	0.018	X	*4.5*	*11.6*	*9.5*	*10.0*
SAD semi-automatic	0.278	0.018	0.496	X	*13.7*	*11.6*	*12.1*
bi-dimensional WHO semi-automatic	0.003	0.018	< 0.001	< 0.0001	X	*2.1*	*1.6*
bi-dimensional WHO manual	0.001	0.152	< 0.001	0.003	0.300	X	*0.5*
volume	0.012	0.170	0.003	0.002	0.453	0.811	X

Results of the McNemar's test for clustered data adjusted for multiple ratings. Volumetry and semi-automatic bi-dimensional WHO revealed significantly fewer misclassifications compared to most unidimensional parameters (i. e., manual LAD, SAD) across all sites. No significant differences were found between semi-automatic bi-dimensional WHO and volume. Additionally, the differences (modulus) between the relative numbers (%) of correct therapy evaluation are displayed in italics.

Ergebnisse des McNemar´s Tests. Bei Volumen und semi-automatischem bi-dimensionalem WHO zeigten sich signifikant weniger Fehlklassifikationen im Vergleich zu allen unidimensionalen Parametern (z. B. LAD, SAD) sowohl manuell als auch semi-automatisch. Diese Ergebnisse waren konsistent in allen Studienzentren. Statistisch signifikante Unterschiede zeigten sich zwischen semi-automatischen bi-dimensionalem WHO und Volumen nicht. Zusätzlich sind Unterschiede als Betrag zwischen den relativen Anteilen (%) der korrekten Beurteilungen des Therapieansprechens unter Berücksichtigung der verschiedenen Parameter kursiv angegeben.

ments (e. g. bi-dimensional WHO/semi-automatic volumetry 93.9 %/92.3 %) compared to uni-dimensional measurements, e. g. manual LAD 84.7 % and SAD 82.3 %, respectively. The mean reduction in wrongly classified patients was calculated to be 9.6 %. This was consistent across all study centers (study site 1: 9.5 %; study site 2: 13.5 %; study site 3: 11.7 %; study site 4: 7.9 %; study site 5: 8.6 %; interval across all study sites: 7.9 – 13.5 %). Based on the number of patients included in this study, therapy response was wrongly classified in seven patients. On a patient level manual bi-dimensional WHO (91.8 %) did not differ significantly (p = 0.3003) from semi-automatic bi-dimensional WHO (93.9 %) and semi-automatic volumetry (92.3 %, p = 08 111).

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-automatic 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 uni-dimensional 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

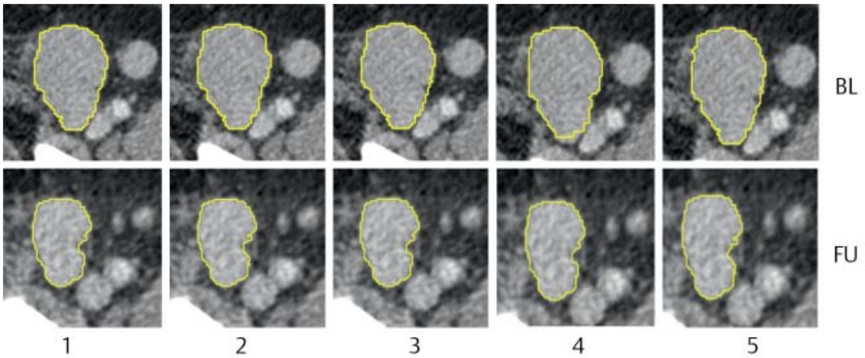


Fig. 4 Size regression of lymph node under therapy and segmentation. Segmentation results on baseline (BL) and follow-up (FU) scans of a sample inguinal lymph node under therapy at the five study sites (1 – 5). The yellow contour lines indicate adequate segmentation results that largely correspond between the five study sites.

Abb. 4 Größenabnahme eines segmentierten Lymphknotens unter Therapie. Segmentierungsergebnisse in der Baseline (BL)- und der Follow-up (FU)-Untersuchung unter Therapie in den fünf Studienzentren (1 – 5). Die gelben Konturlinien verdeutlichen die Segmentierungsergebnisse, die zwischen den Studienzentren eine große Übereinstimmung aufweisen.

(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 parameters [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

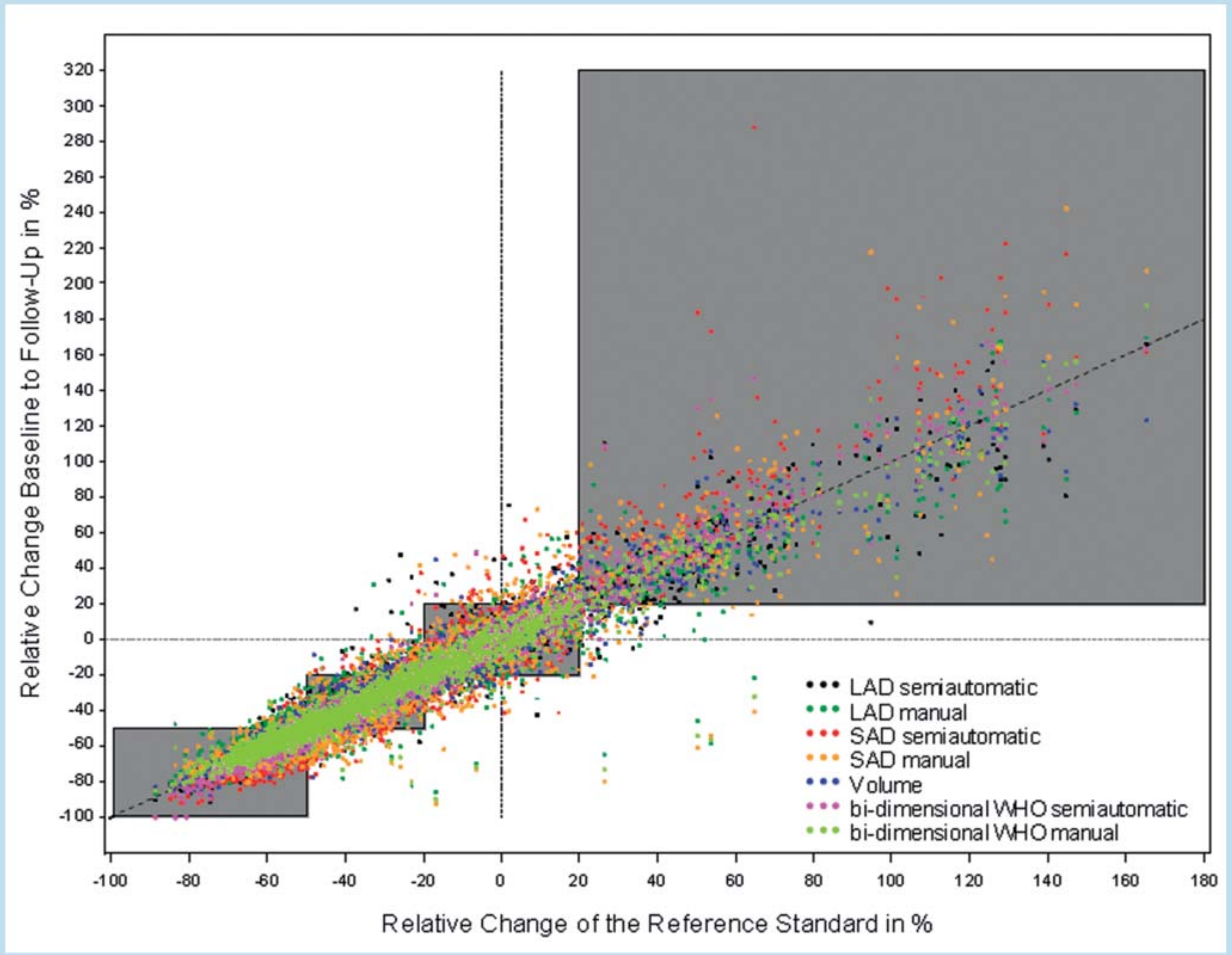


Fig. 5 Scatter plot synopsis of therapy response classification per lymph node (A) across all five study sites based on manual and semi-automatic lymph node evaluation. Response classification per lymph node adapted to RECIST 1.1 [28] superimposed on the scatter plot. The diagonal black line indicates no difference from the reference standard. The closer the measurements come to the reference line, the more accurate they are with regard to therapy response classification. Misclassifications are indicated by measurements located outside the response squares. Across all study sites, semi-automatic volume (blue dots) and bi-dimensional WHO (pink dots) are consistently located closer to the reference line and corresponding points lie superposed, evidently with fewer classification outliers compared to the metric measurements.

Abb. 5 Streudiagramm der Beurteilung des Therapieansprechens pro Lymphknoten (A) über alle fünf Studienzentren bei manueller und semi-automatischer Lymphknotenauswertung. Streudiagramm der Beurteilung des Therapieansprechens pro Lymphknoten anhand adaptierter RECIST 1.1-Kriterien [28]. Die diagonale schwarze Linie zeigt eine Übereinstimmung mit dem Referenzstandard. Je näher ein Punkt der Refenzlinie liegt, desto höher ist der Anteil korrekter Beurteilungen des Therapieansprechens. Fehlklassifikationen werden durch Punkte außerhalb der Kästen verdeutlicht, die die Klassifikation des Therapieansprechens angeben. Über alle Studienzentren zeigten das semi-automatische Volumen (blaue Punkte) und der bi-dimensionale WHO (pinke Punkte) eine geringe Abweichung von der Referenz, was mit einer geringeren Anzahl an Fehlklassifikationen im Vergleich zu den metrischen Messungen verbunden ist. Daher liegen die Punkte dicht bei der Referenzlinie übereinander.

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. This finding was consistently observed at all study sites. 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)

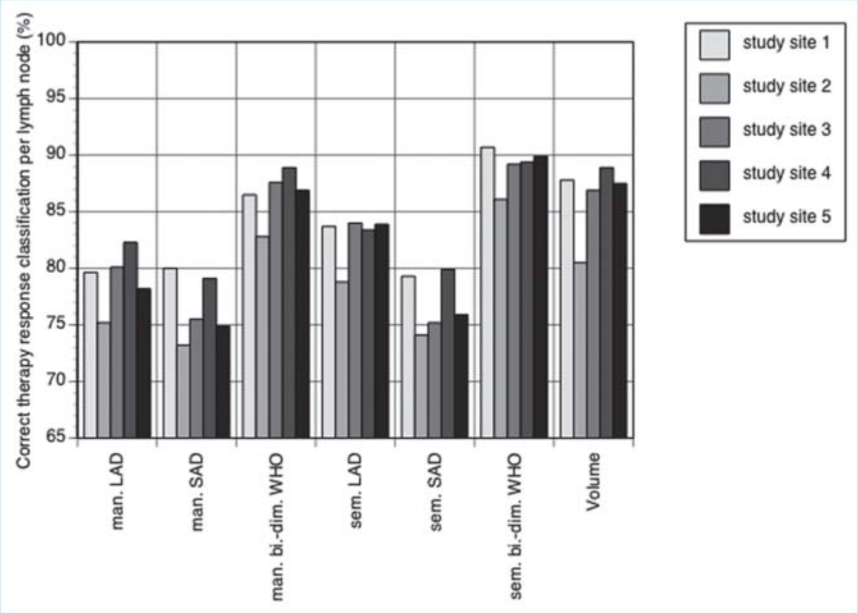


Fig. 6 Therapy response evaluation per lymph node with regard to study site. Relative rates of correct therapy response evaluation per lymph node with regard to study sites. Across all the study sites, the correctness of the therapy response evaluation was lower for manual and semi-automatic LAD and SAD, while the rate for bi-dimensional WHO and volume was higher.

Abb. 6 Beurteilung des Therapieansprechens pro Lymphknoten unter Berücksichtigung des Studienzentrums. Relative Raten korrekter Beurteilung des Therapieansprechens pro Lymphknoten unter Berücksichtigung des Studienzentrums. Über alle Studienzentren war der Anteil korrekter Beurteilungen des Therapieansprechens geringer für den manuellen und semi-automatischen LAD und SAD während diese für die den bi-dimensionalen WHO und das Volumen höher war.

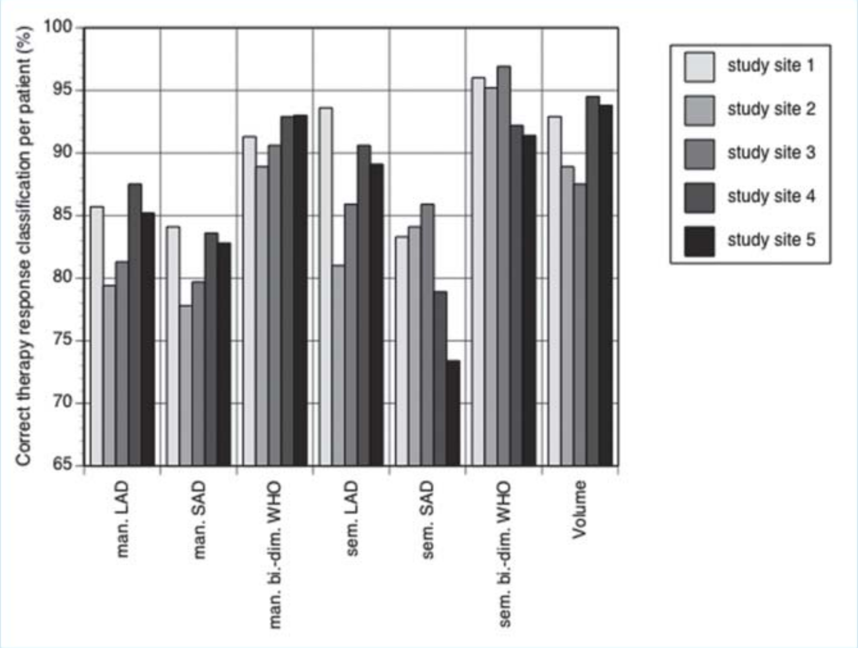


Fig. 7 Therapy response evaluation per patient with regard to study site. Relative rates of correct therapy response evaluation per patient with regard to study sites. Across all the study sites, the correctness of therapy response evaluation was consistently lower for manual and semi-automatic LA-D and SAD, while the rate for bi-dimensional WHO and volume was higher.

Abb. 7 Beurteilung des Therapieansprechens pro Patienten unter Berücksichtigung des Studienzentrums. Relative Raten korrekter Beurteilung des Therapieansprechens pro Patienten unter Berücksichtigung des Studienzentrums. Über alle Studienzentren war der Anteil korrekter Beurteilungen des Therapieansprechens geringer für den manuellen und semi-automatischen LAD und SAD, während diese für den bi-dimensionalen WHO und das Volumen höher waren.

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 – 70 s [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 documentation and calculation of manual WHO – which was not included in the time measurements of this study – have to be performed 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 comparison with an exact reference standard, as is the case with a phantom study. We used the manually and semi-automatically obtained metric parameter as an internal reference standard, which is accepted in the literature for the analysis of segmentation 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 supported 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 segmentation to be robust and time efficient, with acceptable time expenditure compared to conventional manual lymph node assessment. With regard to therapy response classification, semi-automated multidimensional parameters (“volumetry” and “bidimensional WHO”) significantly reduce the number of wrongly classified lymphoma patients across all study sites by approximately 9.6 % (interval across all study sites: 7.9 – 13.5 %, $p < 0.05$) and permit a significantly more accurate therapy response classification than uni-dimensional parameters. Semi-automatic quantitative software tools may help to significantly reduce wrong classifications that arise from manual assessment methods and differences between the examining institutions. In conclusion, semi-automatic quantitative software tools should be implemented in clinical studies and desirably in the clinical routine.

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, Pfistner B, Juweid ME et al.* Revised response criteria for malignant lymphoma. J Clin Oncol 2007; 25: 579 – 586

3 *Schwartz LH, Bogaerts J, Ford R et al.* Evaluation of lymph nodes with RECIST 1.1. Eur J Cancer 2009; 45: 261 – 267

4 *Jaffe TA, Wickersham NW, Sullivan DC.* Quantitative imaging in oncology patients: Part 1, radiology practice patterns at major U. S. cancer centers. Am J Roentgenol 2010; 195: 101 – 106

5 *Jaffe TA, Wickersham NW, Sullivan DC.* Quantitative imaging in oncology patients: Part 2, oncologists' opinions and expectations at major U. S. cancer centers. Am J Roentgenol 2010; 195: 19 – 30

6 *Erasmus JJ, Gladish GW, Broemeling L et al.* Interobserver and intraobserver variability in measurement of non-small-cell carcinoma lung lesions: implications for assessment of tumor response. J Clin Oncol 2003; 21: 2574 – 2582

7 *Castelijns JA, van den Brekel MW.* Imaging of lymphadenopathy in the neck. Eur Radiol 2002; 12: 727 – 738

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 imaging. J Nucl Med 2004; 45: 1509 – 1518

11 *van den Brekel MW, Castelijns JA, Snow GB.* Imaging of cervical lymphadenopathy. Neuroimaging Clin N Am 1996; 6: 417 – 434

12 *van den Brekel MW, Stel HV, Castelijns JA et al.* Cervical lymph node metastasis: assessment of radiologic criteria. Radiology 1990; 177: 379 – 384

13 *Wormanns D, Kohl G, Klotz E et al.* Volumetric measurements of pulmonary nodules at multi-row detector CT: in vivo reproducibility. Eur Radiol 2004; 14: 86 – 92

14 *Schwartz LH, Colville 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 segmentation 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, Mueter S et al.* Measurement accuracy and reproducibility of semiautomated metric 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 manifestations 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 volumetric analysis of lymph node metastases in patients with malignant melanoma stage III/IV–a feasibility study. Eur Radiol 2008; 18: 1114 – 1122

20 *Puesken M, Juergens KU, Edenfeld A et al.* Semiautomatische Segmentierung von Leberläsionen in der MSCT: Einfluss der Schichtdicke auf die Segmentierungsqualität, Messgenauigkeit und Interobservervariabilität. 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 (CARE Dose). Eur Radiol 2002; 12: 1571 – 1576

22 *Cheson BD.* Staging and evaluation of the patient with lymphoma. Hematol Oncol Clin North Am 2008; 22: 825 – 837

23 *Bondiau PY, Malandain G, Chanalet S et al.* Atlas-based automatic segmentation of MR images: validation study on the brainstem in radiotherapy context. Int J Radiat Oncol Biol Phys 2005; 61: 289 – 298

24 *Kuhnigk JM, Dicken V, Bornemann L et al.* Morphological segmentation and partial volume analysis for volumetry of solid pulmonary lesions in thoracic CT scans. IEEE Trans Med Imaging 2006; 25: 417 – 434

25 *Keil S, Plumhans C, Behrendt FF et al.* Automated measurement of lymph nodes: a phantom study. Eur Radiol 2008; 19: 1079 – 1086

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, Moskowitz 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-Verlaufskontrollen beim malignen Lymphom: Vergleich manueller linearer Messungen mit semi-

automatischen Lymphknotensegmentierungen zur Beurteilung des Therapieansprechens. Fortschr Röntgenstr 2012; 184: 795 – 804

30 *Obuchowski NA*. On the comparison of correlated proportions for clustered data. Stat Med 1998; 17: 1495 – 1507

31 Radiological Society of North America Quantitative Imaging Biomarkers Alliance. In: Radiological Society of North America Web site. www.rsna.org/Research/qiba_intro.cfm November 24, 2009

32 *Bolte H*, *Jahnke T*, *Schafer FK et al.* Interobserver-variability of lung nodule volumetry considering different segmentation algorithms and observer training levels. Eur J Radiol 2007; 64: 285 – 295

33 *Bolte H*, *Riedel C*, *Jahnke T et al.* Reproducibility of computer-aided volumetry of artificial small pulmonary nodules in ex vivo porcine lungs. Invest Radiol 2006; 41: 28 – 35

34 *Heussel CP*, *Meier S*, *Wittelsberger S et al.* Quantitative CT-Verlaufskontrolle von Lebermalignomen nach RECIST und WHO im Vergleich zur Volumetrie. Fortschr Röntgenstr 2007; 179: 958 – 964

35 *Fabel M*, *Bolte H*, *von Tengg-Kobligk H et al.* Semi-automated volumetric analysis of lymph node metastases during follow-up-initial results. Eur Radiol 2011; 21: 683 – 692