Neuro-Oncology

24(10), 1726–1735, 2022 | https://doi.org/10.1093/neuonc/noac043 | Advance Access date 14 February 2022

Prospective validation of a new imaging scorecard to assess leptomeningeal metastasis: A joint EORTC BTG and RANO effort

Emilie Le Rhun, Patrick Devos, Sebastian Winklhofer, Hafida Lmalem, Dieta Brandsma, Priya Kumthekar, Antonella Castellano, Annette Compter, Frederic Dhermain, Enrico Franceschi, Peter Forsyth, Julia Furtner, Norbert Galldiks, Jaime Gállego Pérez-Larraya, Jens Gempt, Elke Hattingen, Johann Martin Hempel, Slavka Lukacova, Giuseppe Minniti, Barbara O'Brien, Tjeerd J. Postma, Patrick Roth, Roberta Rudà, Niklas Schaefer, Nils O. Schmidt, Tom J. Snijders, Steffi Thust, Martin van den Bent, Anouk van der Hoorn, Guillaume Vogin, Marion Smits, Joerg C. Tonn, Kurt A. Jaeckle, Matthias Preusser, Michael Glantz, Patrick Y. Wen, Martin Bendszus, and Michael Weller

Department of Neurosurgery, Clinical Neuroscience Center University Hospital and University of Zurich, Zurich, Switzerland (E.L.R.); Department of Neurology, Clinical Neuroscience Center University Hospital and University of Zurich, Zurich, Switzerland (E.L.R., P.R., M.W); University of Lille, France (E.L.R., P.D.); Neuro-Oncology, Neurosurgery Department, CHU Lille, France (E.L.R.); Neurology, Medical Oncology Department, Oscar Lambret Center, Lille, France (E.L.R.); CHU Lille, Lille, France (P.D.); Department of Neuroradiology, Clinical Neuroscience Center, University Hospital Zurich, University of Zurich, Zurich, Switzerland (S.W.); EORTC Headquarters, Brussels, Belgium (H.L.); Department of Neuro-Oncology, Netherlands Cancer Institute–Antoni van Leeuwenhoek, Amsterdam, the Netherlands (D.B., A.Co.); Malnati Brain Tumor Institute of The Robert H Lurie Comprehensive Cancer Center of Northwestern University, Chicago, Illinois, USA (P.K.); Neuroradiology Department, Vita-Salute San Raffaele University and IRCCS Ospedale San Raffaele, Milan, Italy (A.Ca.); Radiation Oncology Department, Gustave Roussy University Hospital, Villejuif, France (F.D.); Nervous System Medical Oncology Department, IRCCS Istituto delle Scienze Neurologiche di Bologna, Bologna, Italy (E.F.); Department of NeuroOncology, Moffitt Cancer Center and University of South Florida, Tampa, USA (P.F.); Biomedical Imaging and Image-Guided Therapy, Medical University of Vienna, Vienna, Austria (J.F.); Department of Neurology, Faculty of Medicine and University Hospital Cologne, University of Cologne; Institute of Neuroscience and Medicine (INM-3), Research Center Juelich, Juelich; Center of Integrated Oncology (CIO) Aachen, Bonn, Cologne and Duesseldorf, University of Cologne, Cologne, Germany (N.G.); Health Research Institute of Navarra (IdiSNA), Pamplona, Navarra, Spain (J.G.P-L.); Program in Solid Tumors, Foundation for the Applied Medical Research, Pamplona, Navarra, Spain (J.G.P-L.); Department of Neurology, Clínica Universidad de Navarra, Pamplona, Navarra, Spain (J.G.P-L.); Technical University of Munich, School of Medicine, Klinikum rechts der Isar, Department of Neurosurgery (J.G.); Institute of Neuroradiology, University Hospital Frankfurt/Main, Frankfurt, Germany (E.H.); Department of Neuroradiology, Eberhard Karls University Tübingen, Tübingen, Germany (J.M.H.); Department of Clinical Medicine, Aarhus University, Aarhus, Denmark (S.L.); Department of Oncology, Aarhus University Hospital, Aarhus, Denmark (S.L.); Radiation Oncology Unit, Department of Medicine, Surgery and Neurosciences, University of Siena; IRCCS Neuromed, Pozzilli (IS), Italy (G.M.); Department of Neuro-Oncology, Division of Cancer Medicine, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA (B.O.B.); Department of Neurology, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands (T.J.P.); Department of Neuro-Oncology, City of Health and Science and University of Turin, Turin, Italy (R.R.); Division of Clinical Neuro-oncology, Department of Neurology, University Hospital Bonn, Germany (N.S.); Department of Neurosurgery, University Medical Center, Regensburg, Germany (N.O.S.); Department of Neurology, UMC Utrecht Brain Center, University Medical Center Utrecht, Utrecht, the Netherlands (T.J.S.); Lysholm Department of Neuroradiology, National Hospital for Neurology and Neurosurgery, University College London Hospitals NHS Foundation Trust, London, UK (S.T.); Department of Brain Rehabilitation and Repair, UCL Institute of Neurology, London, UK (S.T.); Brain Tumor Center at Erasmus MC Cancer Institute, University Medical Center Rotterdam, Rotterdam, the Netherlands (M.v.d.B, M.S.); Medical Imaging Center, Department of Radiology, University Medical Center Groningen, Groningen, the Netherlands (A.v.d.H.); IMoPA Ingénierie Moléculaire et Physiopathologie Articulaire UMR7365 CNRS-UL, Vandoeuvre les Nancy, France (G.V.); Centre François Baclesse, Esch-sur-Alzette, Luxemborg (G.V.); Department of Radiology and Nuclear Medicine, Erasmus MC–University Medical Center Rotterdam, Rotterdam, the Netherlands (M.S.); Department of Neurosurgery, Ludwig-Maximilians-University

School of Medicine, Munich, Germany and German Cancer Consortium (DKTK), Partner Site Munich, Munich, Germany (J.T.); Mayo Clinic Florida, Jacksonville, Florida, USA (K.J.); Department of Medicine I, Division of Oncology, Medical University of Vienna, Vienna, Austria (M.P.); Department of Neurosurgery, Penn State College of Medicine, Hershey, Pennsylvania, USA (M.G.); Center for Neuro-Oncology, Dana-Farber/Brigham and Women's Cancer Center, Boston, Massachusetts, USA (P.Y.W.); Department of Neuroradiology, University Hospital Heidelberg, Heidelberg, Germany (M.B.)

Corresponding Author: Emilie Le Rhun, MD, PhD, Department of Neurosurgery, University Hospital Zurich, Frauenklinikstrasse 10, 8091 Zurich, Switzerland (emilie.lerhun@usz.ch).

Abstract

Background. Validation of the 2016 RANO MRI scorecard for leptomeningeal metastasis failed for multiple reasons. Accordingly, this joint EORTC Brain Tumor Group and RANO effort sought to prospectively validate a revised MRI scorecard for response assessment in leptomeningeal metastasis.

Methods. Coded paired cerebrospinal MRI of 20 patients with leptomeningeal metastases from solid cancers at baseline and follow-up after treatment and instructions for assessment were provided via the EORTC imaging platform. The Kappa coefficient was used to evaluate the interobserver pairwise agreement.

Results. Thirty-five raters participated, including 9 neuroradiologists, 17 neurologists, 4 radiation oncologists, 3 neurosurgeons, and 2 medical oncologists. Among single leptomeningeal metastases-related imaging findings at baseline, the best median concordance was noted for hydrocephalus (Kappa = 0.63), and the worst median concordance for spinal linear enhancing disease (Kappa = 0.46). The median concordance of raters for the overall response assessment was moderate (Kappa = 0.44). Notably, the interobserver agreement for the presence of parenchymal brain metastases at baseline was fair (Kappa = 0.29) and virtually absent for their response to treatment. 394 of 700 ratings (20 patients x 35 raters, 56%) were fully completed. In 308 of 394 fully completed ratings (78%), the overall response assessment perfectly matched the summary interpretation of the single ratings as proposed in the scorecard instructions.

Conclusion. This study confirms the principle utility of the new scorecard, but also indicates the need for training of MRI assessment with a dedicated reviewer panel in clinical trials. Electronic case report forms with "blocking options" may be required to enforce completeness and quality of scoring.

Key Points

- Interobserver agreement with the new scorecard for MRI features of LM is generally acceptable.
- Agreement for overall response assessment in LM is moderate.
- The level of inconsistencies within ratings might be reduced by an improved eCRF and training sessions.

Importance of the Study

This study showed a moderate to substantial agreement for the identification of leptomeningeal metastasis-related MRI features, with the best concordance obtained for hydrocephalus and spinal nodules and the worst concordance noted for linear disease. The agreement for overall response was moderate. Interestingly, the agreement for the identification of brain metastases in the context of leptomeningeal metastasis was low. Electronic case report forms with "blocking solutions" will help to increase the quality of response assessment. Central review among trained raters is probably required for optimal assessment of response to treatment in patients with leptomeningeal metastases in clinical trials.

Leptomeningeal metastasis (LM) is usually a late and life-threatening complication in patients with metastatic cancer. Diagnosis and follow-up of patients with LM is challenging. It is classically based on the triad of clinical evaluation, neuroimaging findings, and CSF cytology. These three levels of assessment are complementary and all have their specific

strengths and weaknesses. The clinical status may be affected by many confounding factors, including concomitant brain metastases, seizures, comedications, and others. Neuroimaging appears to be the most objective and reliable measure to measure disease burden, although our previous effort at validating imaging-based response criteria has not confirmed this view in practice.[1] CSF parameters such as protein or lactate are unspecific, and even qualitative assessment of tumor cells in the CSF has turned out to be challenging. Assessing tumor burden by quantifying CSF tumor cells would be attractive, but would not capture nodular and linear disease detected on MRI.

Gadolinium-enhanced brain and spine MRI scans, in combination with cerebrospinal fluid (CSF) cytology, help to define the subtype of LM.[2] Subclassification according to these criteria helps to estimate prognosis and to guide clinical decision making.[3,4]

No validated or broadly used algorithms for clinical, neuroimaging, or CSF assessments have entered clinical practice. Heterogeneity in patient populations and imaging protocols, as well as lack of validation of assessment tools may have contributed to the disappointing or inconclusive results of early randomized clinical trials performed more than 20 years ago.[5] For neuroimaging, in particular, the complex shape of the space being assessed and the dynamic nature of CSF flow present challenges not encountered when evaluating disease outside the central nervous system. Low resolution or low quality imaging may contribute to challenges of assessment of LM and to disagreement among raters, too.

Given this lack of a validated tool, the assessment of response to treatment in clinical trials for LM from solid tumors represents a challenge, too. The first effort of the Response Assessment in Neuro-oncology (RANO) group LM committee (LANO) addressed the key areas of clinical presentation, MRI presentation, and cerebrospinal fluid findings in LM patients.[6] Yet, a subsequent effort to validate the LANO scorecard proposed to rate the MRI findings failed for multiple reasons: several items were unclear even for an expert reviewer panel, the logic of some tasks appeared to be questionable in retrospect, and rules to apply the scorecard were not sufficiently clear.[7] Accordingly, a new scorecard was developed that aimed at addressing the challenges experienced during the validation of the first scorecard. Here we report on the results of a prospective validation study of the new LM scorecard.

Material and Methods

Study Design

The objective was to prospectively validate the revised scorecard for the neuroimaging assessment of LM from solid tumors. Coded paired cerebrospinal MRI scans without and with gadolinium from 20 adult patients with confirmed or probable diagnosis of LM from a solid tumor according to EANO ESMO criteria[2] were evaluated by 35 raters via an electronic case report form using the EORTC imaging platform. Active members of the RANO Leptomeningeal Metastasis Committee and the CNS Metastasis and Imaging Committees of the EORTC Brain tumor Group were invited to participate. There were 9 neuroradiologists, 17 neurologists, 4 radiation oncologists, 3 neurosurgeons, and 2 medical oncologists. Age was provided by 30 raters, median age was 54 years (interquartile range, IQR: 49.5–58.5). Thirty-one dates of board certification were available, the median number of years since board certification was 12 (IQR: 8.5–25).

Definitions of the different items were provided with the scorecard. Nodules should be at least 5 x 5 mm in orthogonal diameters in 2 planes. Rules for calling progression, stable disease, partial or complete response were also defined.[7] MRI scans were obtained retrospectively from Bonn University Hospital, University Hospital Zurich and Lille University Hospital. There was no standardization of scanner or sequences for MRI sets to be included in this study. As a minimum requirement, brain MRI was intended to include unenhanced axial T1, axial FLAIR or axial T2, and 3D postcontrast T1 sequences. Spine MRI was intended to include cervical, thoracic, and lumbar sagittal T1 postcontrast and sagittal T2 sequences. The median interval between first baseline MRI and first follow-up MRI was 3 months (interquartile—IQR: 2.5–4.3 months). All items except the Evan's index[8] for hydrocephalus had to be rated as present, absent, or nonevaluable at baseline and at follow-up. A maximum of three nodules at the brain level and at the spinal level were to be measured dimensionally at baseline and at follow-up. For each item except the Evan's index, the rater had to score the change as improved, no change or worse. The Evan's index was determined using the measures of the two lengths as recommended[8] and the ratio was then derived automatically. A threshold of 0.27 has been reported as corresponding to the normal cut-off value in females aged between 65–69 years.[8] Moreover, in a cohort of 534 participants, 29% of 308 healthy controls had an Evan's index of 0.3 or more,[8] and thus both were explored. Overall response assessment included five categories.[7] The evaluations of all raters were centrally reviewed by ELR and MW. Obvious inconsistencies of the ratings were corrected, e.g., if actual measures of nodules were indicated, but a rater did not indicate that there were indeed nodules. For a defined item, when the value was nonevaluable at baseline or at follow-up, the change score was rated as nonevaluable. Unless specifically indicated otherwise, ratings were analyzed as provided by the raters.

Ethics Statement

The sponsor of the study was the University Hospital Zurich. The Cantonal Ethics Committee of the Canton of Zurich approved the project (2018-00192). Appropriate ethics approvals as required were obtained at participating centers. The study was conducted in accordance with the ethical principles of the Declaration of Helsinki and Good Clinical Practice Guidelines.

Statistical Analyses

The Kappa coefficient is commonly used to evaluate the agreement of 2 raters for a specific item. When more than 2 raters are involved in an analysis, the Krippendorf α coefficient is generally used. Here, there were more raters than

patients, and a relatively high number of missing data. The choice to use pairwise Kappa allowed us to estimate the distribution of the Kappa for each item analyzed and thus to assess its variability according to the couples of raters and to compare the Kappa according to certain subgroups. Analysis of quantitative parameters, such as the Evan`s index, was performed after discretization (higher or lower than a cut-off, 2.7 or 3 for the Evan`s index). Thus, the concordance of the different items was calculated on 20 patients evaluated by 35 raters using the Kappa coefficient.[9] For each item, we estimated the Kappa on all pairwise combinations, corresponding to (35*34)/2 possible values (n = 595). We then summarized the Kappa distribution using the first quartile, median two, and fourth quartile calculated on all 595 observed values. A median Kappa coefficient of 0.81 or more is considered as almost perfect agreement, values between 0.80 and 0.61 are considered as substantial agreement, values between 0.60 and 0.41 as moderate agreement, values between 0.40 and 0.21 as fair agreement, values between 0.20 and 0.01 as in favor of none to slight agreement, and values below 0 show no agreement.

Agreement was analyzed among all raters (n = 35), among neuroradiologists and neurologists pooled (n = 26), among neurologists only (n = 17), and among neuroradiologists only (n = 9). The distributions of Kappa according to neurologists versus neuroradiologists were compared by the nonparametric Wilcoxon test. A P-value of .05 or below was considered as significant. Since the Kappa measure depends on the prevalence of the finding under consideration, it could not be derived for items that were too rare, e.g. spinal leptomeningeal metastases, reported as present at baseline by 17 (2%) of the ratings or epidural metastases, reported as present at baseline by 33 (5%) of the ratings (Supplementary Tables S1 and S2). For the assessment of overall response, ratings were recoded as binary. To explore the compliance of raters with the instructions, analyses were performed using datasets fully assessed by raters, with measurement of brain and spinal nodules and Evan's index at baseline and follow-up, and with linear enhancement assessment in all compartments evaluated at baseline and follow-up. Statistical analyzes were performed by PD using the SAS V9.4 software (Cary, NC, USA).

Results

Rating and Concordance of LM-Related Items

With 35 raters assessing 20 patients at 2 time points, 1400 ratings should have been available for analysis. Individual and pooled ratings per item at baseline are summarized in Supplementary Tables S1 and S2. The vast majority of LM-related items were considered evaluable by most raters, i.e., the maximum percentage of "nonevaluable" responses for any item did not exceed 3%. Missing data were mainly noted for hydrocephalus (n = 18, 3%). For the other items, missing data were observed in less than 1% (Supplementary Table S2).

The concordance of ratings per item at baseline is reported in Table 1 and Figure 1. Among LM-related items, substantial agreement among raters was obtained for hydrocephalus assessed by the Evan's index (Kappa = 0.75 with a cut-off at 0.30 and 0.69 with a cut-off at 0.27), hydrocephalus assessed visually, and spinal nodules (Kappa = 0.63 each). Moderate agreement was noted for brain meningeal nodules (Kappa = 0.47), spinal linear disease (Kappa = 0.46), and brain linear disease (Kappa = 0.45). In the brain, agreement for nodules was similar for cut-off values of 5 x 5 mm versus 10 x 10 mm whereas the cut-off of 10 x 10 mm yielded better agreement in the spinal cord (Supplementary Table S3) (not formally tested). Concordance was better among neuroradiologists for the rating of brain linear disease and for hydrocephalus based on Evan's index with a cut-off of 0.27, but better among neurologists for the rating of spinal linear disease (Table 1 and Figure 1).

Brain nodules at baseline were noted as present on 320 brain scans (46%) (Supplementary Table S2). No quantitative measurement was reported for 35 of 320 nodules (11%) rated as present. Among the 285 nodules with measurements, 16 nodules (6%) were measured by the raters as inferior to 5 x 5 mm and 269 nodules (84%) were measured as 5 x 5 mm or more, including 151 nodules (53%) measured at 10 x 10 mm or more. At the spinal level, 173 nodules were reported, but these were not measured in 54 instances (31%). Among the 119 nodules with measurements, 26 nodules (22%) were measured inferior to 5 x 5 mm, and 93 nodules (78%) were measured as 5 x 5 mm or more, including 5 nodules measured as 10 x 10 mm or more (Supplementary Table S3).

Among the 20 scans assessed by 35 raters, hydrocephalus was rated as present on 164 scans (23%) and absent on 517 scans (74%) (Supplementary Table S2). Measurements of hydrocephalus allowing to derive Evan's indexes were available for 588 of the 700 ratings. Indexes below 0.27 were reported on 222 scans (37%), in accordance with the absence of hydrocephalus for 213 scans (96%), but were rated as associated with hydrocephalus according to the rater on 2 scans (<1%) or missing data for 7 scans (3%). Indexes between 0.27 and 0.29 (>0.27 and <0.30) were noted in 173 instances (29%), with hydrocephalus rated as absent in 164 instances (95%), present in 7 instances (4%), and missing data in 2 instances (1%). Indexes were measured at 0.30 or more on 193 scans (33%), among these, hydrocephalus was rated as present in 148 instances (77%), absent in 39 instances (20%), and missing data in 6 instances (3%) (data not shown).

Rating and Concordance of Non-LM-Related Items

For non-LM-related items, the rates of nonevaluable or missing data were equally low as for LM-related items (Supplementary Tables S1 and S2). Among the non-LM-related items, the Kappa index could be calculated for brain metastases only because of too low incidence of spinal parenchymal and epidural metastases (Supplementary Tables S1 and S2). Concordance for brain metastases was fair (Kappa = 0.29), inferior to all LM-related items. It was best among neuroradiologists (Table 1 and Figure 1). Among the 257 measures provided at baseline, brain metastases were measured 10 x 10 mm or more in 93 ratings (36%) and had at least a diameter of 10 mm in one dimension in 164 ratings (64%).

Table 1 Overall Concordance Rate for Single Items of the Scorecard Using Simple Kappa Coefficient: Median (Lower-Upper Quartile) at Baseline

MRI findings	All raters (*N* = 35)	Neuroradiologists and neurologists (*N* = 26)	Neurologists only (*N* = 17)	Neuroradiologists only (*N* = 9)	*P*-value between neurologists and neuroradiologists
Number of Kappa analysis	*N* = 595	*N* = 325	*N* = 136	*N* = 36	
LM-related items					
Brain nodules	0.47 (0.3–0.6)	0.42 (0.25–0.56)	0.42 (0.26–0.6)	0.39 (0.27–0.59)	.4940
Brain linear disease	0.45 (0.31–0.58)	0.5 (0.38–0.64)	0.48 (0.34–0.62)	0.59 (0.45–0.7)	**.0039**
Hydrocephalus	0.63 (0.44–0.77)	0.69 (0.53–0.85)	0.69 (0.54–0.86)	0.66 (0.5–0.82)	.1800
Evan's index, cut-off 0.27	0.69 (0.57–0.79)	0.69 (0.57–0.79)	0.69 (0.47–0.79)	0.77 (0.68–0.86)	**.0080**
Evan's index, cut-off 0.30	0.75 (0.58–0.88)	0.78 (0.67–0.89)	0.76 (0.58–0.89)	0.83 (0.72–0.89)	.0588
Spinal nodules	0.63 (0.43–0.76)	0.63 (0.42–0.76)	0.6 (0.39–0.76)	0.64 (0.55–0.76)	.0793
Spinal linear disease	0.46 (0.29–0.6)	0.46 (0.29–0.58)	0.5 (0.34–0.63)	0.43 (0.23–0.55)	**.0185**
Non-LM-related items					
Brain parenchymal metastases	0.29 (0.12–0.45)	0.3 (0.13–0.45)	0.24 (0.11–0.45)	0.41 (0.31–0.52)	**.0002**

LM, leptomeningeal metastasis; *N*, number
Evan's index 0.27: cut-off inferior to 0.27 versus equal or superior to 0.27; Evan's index 0.30: cut-off inferior to 0.30 versus equal or superior to 0.30

Rating Changes of Single LM-Related Items

Supplementary Table S4 summarizes the assessment of changes of each item between baseline and follow-up MRI and the overall response assessment as proposed in the scorecard.[7] Regarding change of single items, nonevaluable responses were observed mainly for response assessment of spinal linear enhancement ($n = 21$, 3%), and missing data mainly for response assessment of brain linear enhancement ($n = 23$, 3%). For the other items, missing data percentages were below 3%. Inconsistencies of rater responses were noted in 16 instances (2%), including assessment of hydrocephalus in 10 instances (1%) (Supplementary Table S4). The interobserver agreement for change of single items was substantial for hydrocephalus (median Kappa = 0.64). The worst agreements were none to slight agreement for brain linear disease (median Kappa = 0.18) and fair agreement for brain meningeal nodules (median Kappa = 0.32). The concordance was moderate for the other LM-related items. The concordance among neuroradiologists was significantly better for the evaluation of response of spinal linear disease and hydrocephalus (Table 2).

Description of the Overall Response Assessment from Raters

Overall response assessments were provided for 644 scans (92%) (Supplementary Table S4). The overall response assessment was rated "nonevaluable" in 21 instances (3%) only, mainly for case 19 ($n = 7$). One rater (rater 3) scored the overall response as nonevaluable 5 times. Five other raters rated response as nonevaluable twice. In 3 instances, progression of brain or spinal nodules was reported, but the response assessment said "nonevaluable". The response was missing in 35 instances (5%). The maximum number of missing data per patient was 3 (9%) and this was noted in 4 patients. One rater (rater 16) did not provide response assessment for 9 patients, another (rater 5) for 5 patients.

The patterns of progression are reported in Supplementary Table S5. According to the raters, progression was determined by the progression of one LM-related item only in 83 instances (31%), by the progression of 2 LM-related items in 117 instances (44%), of 3 LM-related items in 41 instances (15%), of 4 LM-related items in 18 instances (7%) and by progression of all LM-related items in 4 instances (1%). Overall progression was reported by raters in 5 instances, although no progression was reported at the single item level. Four raters (1 neuroradiologist, 2 neurologists, and 1 medical oncologist) reported a progression based on hydrocephalus increase only. Among them, 3 provided a measure for Evan's index at baseline and follow-up. A 25% increase of the Evan's index was noted in 2 instances. Among the ratings according to the instructions, the pattern of progression was mainly determined by progression of one item (Supplementary Table S5). The median Kappa for the concordance among the 700 scores was 0.44, showing a moderate agreement, and no difference was observed among the different groups of raters (Table 2 and Figure 2).

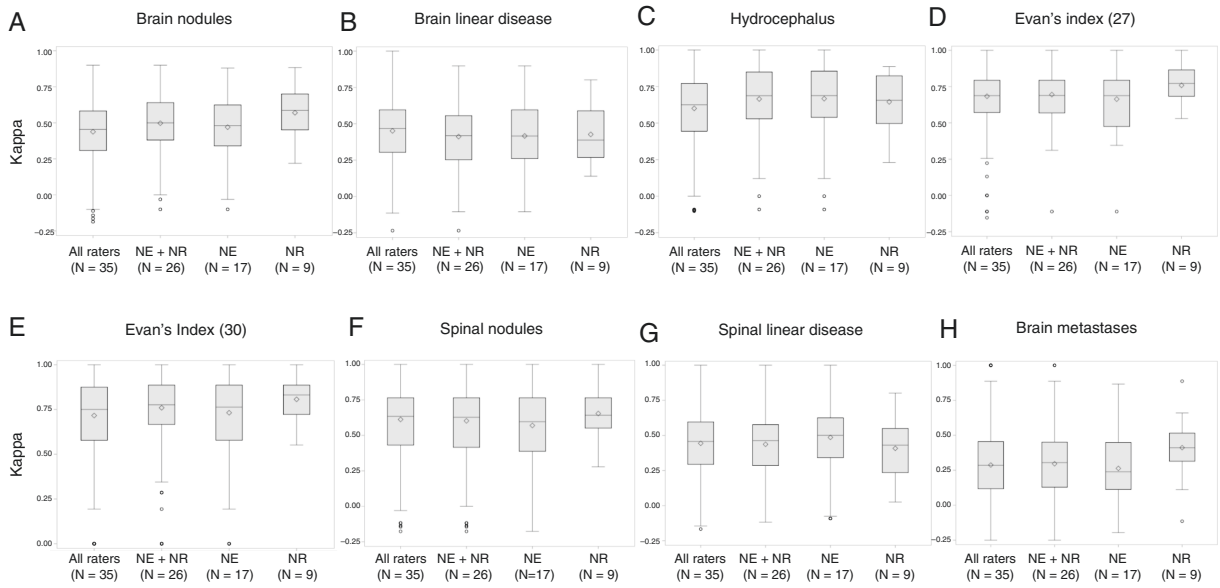

Fig. 1 Distribution of the median Kappa for baseline single items among neurologists and neuroradiologists together, neurologists and neuradiologists. a: brain meningeal nodules, b: brain linear meningeal enhancement, c: hydrocephalus, d: Evan's index cut-off 27, e: Evan's index cut-off 30, f: spinal meningeal nodules, g: spinal linear meningeal enhancement, h: parenchymal brain metastases.

Rating Changes of Single Items and of Overall Response of Non-LM-Related Items

For non-LM-related items, the percentage of nonevaluable items was 3 or less and the percentage of missing data 4 or less (Supplementary Table S4). Only the agreement between raters for brain metastases could be analyzed for non-LM items. The median K-value showed no agreement (median Kappa = 0) (Table 2).

Compliance with the Instructions

Of 700 ratings, corresponding to 20 patients assessed by 35 raters, 394 ratings (56%) were complete, with measurement of brain and spinal nodules and Evan's index at baseline and follow-up, and with all linear enhancements evaluated at baseline and follow-up. The differential frequency of responses for the single items was similar between the 2 groups of ratings (Supplementary Table S2). The overall response assessment in the 394 complete ratings was concordant for 308 ratings (78%), including concordance for CR in 3 instances (1%), PR in 19 instances (5%), SD in 155 instances (39%), and PD in 131 instances (33%) (Table 3 top, and Supplementary Table S4). Discordances included PD according to the raters versus SD according to the instructions (n = 17, 4%), PR according to the raters but CR according to the instructions (n = 15, 4%) and PR according to the rater versus SD according to the instructions (n = 14, 4%). Divergences in the interpretation of response were mainly related to the assessment of linear meningeal enhancement.

To evaluate the potential impact of these discordances on the discontinuation or continuation of a therapeutic intervention, we pooled CR, PR, and SD and opposed these responses to PD (Table 3, bottom, Supplementary Table S4). A PD of the raters changed into CR, PR, or SD according to the instructions was observed in 23 instances (6%), whereas CR, PR or SD according to the raters turned into PD according to the instructions in 12 instances (3%). The reasons were *de novo* linear enhancement in 5, PD of brain nodules in 3, and PD of spinal nodules in 4 instances. The overall response was declared as nonevaluable or was missing although all items were rated and could have led to an overall scoring in 15 instances (3.5%).

Discussion

The main scope of the standardized LM scorecard studied here is to assess the overall imaging response of patients with LM from solid tumors in clinical trials in a reproducible fashion and thus to improve the quality of the evaluation of new drugs or therapeutic approaches and to improve the validity of therapeutic decision making. The new scorecard was developed based on experience with a similar validation effort as reported here for the first assessment tool proposed by the RANO group many years ago.[6,7] We anticipate that the new scorecard may also be useful in routine practice outside the setting of a clinical trial.

The best interobserver agreement for the assessment of single items met criteria for "substantial"[9] and was observed for hydrocephalus and spinal meningeal nodules (Table 1). The concordance for the evaluation of nodules was better at the spinal level than at the brain level where LM nodules may be more difficult to distinguish from parenchymal metastases. Overall the agreement for measurable parameters like hydrocephalus or nodules was better than for nonmeasurable items, e.g., linear disease.

Table 2 Overall Concordance Rates of the Change of the Different Single Items and Overall Response Using Simple Kappa Coefficient: Median (Lower-Upper Quartile) Between Baseline and Follow-Up MRI

MRI findings	All raters (*N* = 35)	Neuroradiologists and neurologists (*N* = 26)	Neurologists only (*N* = 17)	Neuroradiologists only (*N* = 9)	*P*-value between neurologists and neuroradiologists
Number of Kappa analysis	*N* = 595	*N* = 325	*N* = 153	*N* = 136	
LM-related items					
Brain nodules	0.32 (0.14–0.5)	0.29 (0.12–0.48)	0.29 (0.14–0.48)	0.31 (–0.07 to 0.57)	.3702
Brain linear disease	0.18 (–0.08 to 0.37)	0.18 (–0.07 to 0.35)	0.15 (–0.07 to 0.35)	0.22 (–0.07 to 0.35)	.3050
Hydrocephalus	0.64 (0.44–0.77)	0.64 (0.45–0.77)	0.64 (0.45–0.77)	0.69 (0.54–0.83)	**.0259**
Spinal nodules	0.58 (0.38–0.74)	0.58 (0.34–0.73)	0.61 (0.36–0.77)	0.46 (0.31–0.65)	.0511
Spinal linear disease	0.44 (0.22–0.63)	0.52 (0.31–0.65)	0.45 (0.29–0.6)	0.66 (0.45–0.84)	**.0004**
Overall response	0.44 (0.29–0.58)	0.46 (0.31–0.6)	0.46 (0.34–0.61)	0.4 (0.29–0.6)	.2174
Non-LM-related items					
Brain parenchymal metastases	0 (–0.09 to 0.49)	0 (–0.09 to 0.45)	0.21 (–0.09 to 0.48)	–0.08 (–0.08 to 0.38)	.1093

LM, leptomeningeal metastasis; *N*, number

The best agreement for response assessment of single items, that is, change between two MRI series, was also noted for hydrocephalus and spinal meningeal nodules (Table 2).The apparent differences between concordance of neurologists versus neuroradiologists may be a chance finding since concordance for linear disease was better for neuroradiologists in the brain, but better for neurologists in the spinal cord.

Hydrocephalus had not been included in the previous scorecard proposal. A good interobserver concordance using the Evan's index has been reported among healthy elderly males and females or Alzheimer's patients,[8] but no data are available for LM. The prognostic value and a cut-off still need to be determined in the LM population, but the strong agreement among raters warrants further exploration of the Evan's index to be included in the assessment of response to treatment in LM.

The interobserver agreement for overall response assessment was moderate (Table 2). No specific patterns of concordance by the specialty of raters emerged, except for the response assessment of spinal linear disease and hydrocephalus with a better concordance among neuroradiologists than among neurologists.

Divergent outcomes between rater's assessment of response and calculation according to the instructions were mainly related to the calling of partial response based on reduction of linear meningeal disease only which does not conform to the instructions.[7] Similarly, CR, which requires resolution of all contrast-enhancing LM-related measurable lesions, without an increase in ventricular size, was sometimes rated as PR because linear enhancement persisted. Conversely, PR requires regression of all nodules by 50% or more, without an increase in ventricular size according to the instructions, but can thus be called even with an apparent increase of pre-existing linear nonmeasurable disease. The contribution of linear disease to the overall response deserves to be better defined in a subsequent effort.

Interestingly, the interobserver agreement for the detection of brain metastases at baseline was lower than the agreement for the LM-related items, with only fair agreement obtained, although it was better (moderate) among neuroradiologists. No agreement was found for the response assessment of brain metastases either. This observation suggests that tools are needed to improve the assessment of brain metastases, too, notably in the context of LM, where leptomeningeal nodules might be scored as parenchymal lesions and vice versa, and probably also in the context of small parenchymal lesions as present in most of our patients. Such small lesions are often evaluated in clinical trials on systemic pharmacotherapy in patients with brain metastases.[10–13] Also, brain metastases may have merely been scored less thoroughly because this was not the focus of the study.

Obvious inconsistencies of the ratings were observed and had to be corrected, but this affected not more than 2% of the ratings. For example, linear disease rated as present at baseline and absent at follow-up, but with a change score of progression, or measurements of nodules greater of 5 x 5 mm, but no documentation of the presence of nodules. We also noted that the instructions were not always followed by some raters. Although the definition of the nodule was explicitly defined on the scorecard, 6% of scans rated as "modules present" had measurements less than 5 x 5 mm.The assessment of overall response was different between scores as provided by the raters and the score recalculated according to the instructions in 22% of the ratings (Table 3), with 9% of assessments potentially resulting in differing therapeutic consequences (CR, PR, SD versus PD). Such inconsistencies and missing data as well as disregard for the instructions could be avoided by

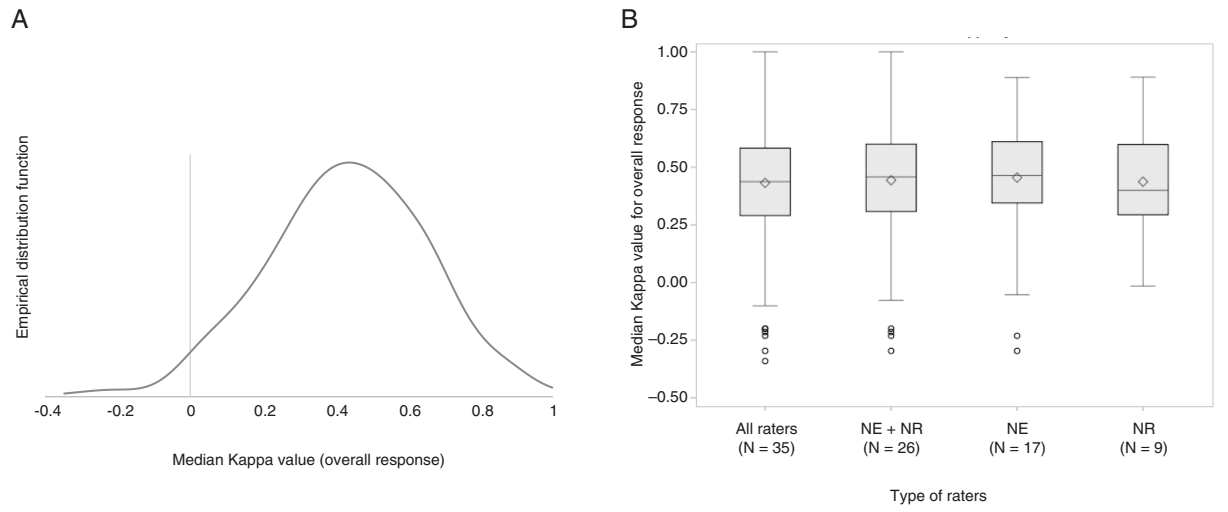

Fig. 2 Box plots showing the distribution of the Kappa for the assessment of overall response. a. Distribution of the Kappa among all raters, showing a normal distribution of the values. b. Distribution of the median Kappa among neurologists and neuroradiologists together, neurologists and neuradiologists.

a computerized system that precludes such errors and automatically documents items according to the instructions (a so-called "blocking solution"). For example, the item "presence" of nodules could appear automatically once a measurement corresponding to the definition of a nodule is provided. Automatic calculation of size variation could also guide the response assessment for each item and for the overall response.

The contribution of educational sessions on a web-based platform to improve the reproducibility of scoring LM has been explored.[14] Eight MRI series were scored by 4 radiologists (2 neuroradiology fellows and 2 radiology residents) without specific training and 3 neuroradiologists who had received instructions and task-specific pictorial examples and description of the tools prior to the scoring of patients. A better interobserver agreement was observed among reviewers who had received the training. These results also favor the use of central imaging review in clinical trials. However, planning real time imaging central review by only one trained expert or a group of experts may be difficult in a trial. Regular training sessions, based on speed of enrollment, could improve the homogenization of the local assessment at the site level between centers. Standardized MRI protocols as recently developed for glioblastoma[15] or brain metastases[16] would certainly also improve the reliability and reproducibility of neuroradiological assessment of LM. A preliminary consensus proposal for LM that could be used in clinical trials has been formulated.[2]

Limitations of this study include—potentially—the inclusion of too many nonneuroradiologists and the failure to conduct a virtual training session for the raters (who are all authors of the manuscript). Additional potential limitations include the large (44%) number of incomplete responses and the relatively high frequency (22%) of responses that did not conform to instructions for assessment instructions. This suggests that some assessors provided responses that preferentially favored their own idiosyncratic conceptions of "response" and "failure" over the formalized algorithms of the assessment tool. Furthermore, the sample may not have been fully representative of the full spectrum of LM.

Given the complexity of response assessment in LM by neuroimaging, one might raise the question of whether alternative approaches of disease monitoring should be prioritized over imaging. However, as noted above, the clinical status of LM patients may be determined by factors other than the LM disease burden and the usefulness of CSF markers still needs to be confirmed, furthermore, CSF assessment does not capture nonfloating tumor cell burden, e.g., linear and nodular disease. Finally, neuroimaging data may be more suitable for post hoc central review than films of patients or digitalized CSF cytology studies.

In conclusion, we believe that the present scorecard presents a major advance in developing tools for the assessment of LM[7] and can be used in practice, provided that the challenges experienced in this validation effort are taken into consideration. A modified ready-to-use version is included as Supplementary Table S6. In clinical practice, a limited set of items of the grid chosen individually by the clinical practitioner could also be useful.

Although clear instructions were provided, the agreement among raters was only moderate with a median Kappa of 0.44 for overall response assessment. Electronic case report forms with "blocking solutions" are probably required to reinforce completeness and quality of scoring. This study confirms the necessity of central review as well as the need for training of MRI assessment even for local raters, ideally board-certified neuroradiologists, in clinical trials.

Future prospective studies may have to explore whether artificial intelligence might be a complementary

Table 3 Concordance in Overall Response Between Rater's Scores and Scores Calculated Based on Their Single Ratings and According to the Instructions Provided in the Grid

Instructions \ Raters	Complete response	Partial response	Stable disease	Progression	Nonevaluable	Empty	Total
Complete response	3 (1%)	15 (4%)	0	1 (<0.5%)	0	0	**19 (5%)**
Partial response	0	19 (5%)	4 (1%)	5 (1%)	0	1 (<0.5%)	**29 (7%)**
Stable disease	3 (1%)	14 (4%)	155 (39%)	17 (4%)	0	5 (1%)	**194 (49%)**
Progression	0	2 (0.5%)	10 (2%)	131 (33%)	2 (0.5%)	7 (2%)	**152 (39%)**
Total	**6 (1%)**	**50 (13%)**	**169 (43%)**	**154 (39%)**	**2 (0.5%)**	**13 (3%)**	394
Instructions \ Raters	**Complete response, partial response or stable disease**			**Progression**	**Nonevaluable**	**Empty**	**Total**
Complete response, partial response or stable disease	213 (54%)			23 (6%)	0	6 (1%)	**242 (61%)**
Progression	12 (3%)			131 (33%)	2 (0.5%)	7 (2%)	**152 (38%)**
Total	**225 (57%)**			**154 (39%)**	**2 (0.5%)**	**13 (3%)**	394

or alternative approach to response assessment in LM as recently proposed for recurrent glioblastoma.[17] Finally, neuroradiological assessment is only one part of the overall assessment of patients with LM and similar efforts to establish reliable tools for clinical assessment and for CSF evaluation would also facilitate the conduct of successful clinical trials.

Supplementary material

Supplemental material is available at *Neuro-Oncology* online.

Keywords

brain | feasibility | MRI | response | validation

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors, except support for the set-up of the imaging platform by the European Organisation for Research and Treatment of Cancer (EORTC) Brain Tumor Group.

Conflict of interest statement. ELR has received honoraria for lectures or advisory board from Adastra, Abbvie, Bayer, Daiichi Sankyo, Leo Pharma, Tocagen, Seattle Genetics. PK has received advisory board participation or consulting fees from the following for-profit companies: Janssen, Gerson Lehrman Group (GLG), Angiochem, Biocept, Bliss Bio, SDP Oncology, Celularity, Novocure, Affinia, and Orbus Therapeutics. EF received honoraria for advisory board participation from the following for-profit companies: Karyopharm. J GP-L has received honoraria for lectures, consultation or advisory board participation from EISAI, Janssen, Pharmaventures, and research support from DNAtrix. JG has received honoraria for lectures, consultation from BrainLab, Zeiss, Medtronic. The following for-profit companies have supported clinical trials and contracted research conducted by JG with payments made to his institution: Zeiss, BrainLab, Pfizer, Medtronic. SL has received honoraria for lecture and advisory board participation from Bayer. PR has received research grants from MSD and Novocure and honoraria for advisory board participation or lectures from Bristol-Myers Squibb, Covagen, Debiopharm, Medac, MSD, Novartis, Novocure, QED, Roche and Virometix. RR has received honoraria for lectures or consultation from UCB, Novocure. NS has received honoraria for advisory board participation from Bayer. MVDB has received honoraria for consultation from Abbvie, Agios, Carthera, Celgene, Bayer, Nerviano, Karyopharm, Boehringer. MS has received honoraria from GE Healthcare (paid to institution). JT has received honoraria for lectures from CarThera. The following for-profit companies have supported clinical trials and contracted research conducted by JT with payments made to his institution: novocure, Munich surgical imaging. MP has received honoraria for lectures, consultation or advisory board participation from the following for-profit companies: Bayer, Bristol-Myers Squibb, Novartis, Gerson Lehrman Group (GLG), CMC Contrast, GlaxoSmithKline, Mundipharma, Roche, BMJ Journals, MedMedia, AstraZeneca, AbbVie, Lilly, Medahead, Daiichi Sankyo, Sanofi, Merck Sharp & Dome, Tocagen. The following for-profit companies have supported clinical trials and contracted research conducted by MP with payments made to his institution: Böhringer-Ingelheim, Bristol-Myers Squibb, Roche, Daiichi Sankyo, Merck Sharp & Dome, Novocure, GlaxoSmithKline, AbbVie. PYW has received research support from Agios, AstraZeneca/Medimmune, Beigene, Celgene, Eli Lily, Genentech/Roche, Kazia, MediciNova, Merck, Novartis, Nuvation Bio, Oncoceutics, Vascular Biogenics, VBI Vaccines and honoraria for advisory board participation or consultation from Agios, AstraZeneca, Bayer, Black Diamond, Boston Pharmaceuticals, Chimerix, CNS Pharmaceuticals, Imvax, Merck, Morphosys, Mundipharma, Novartis, Novocure, Nuvation Bio, Prelude Therapeutics, Vascular Biogenics, VBI Vaccines, Voyager, Celularity, Sapience. MB reports personal fees from Teva, BBraun, Vascular Dynamics, Bayer, Grifols, Merck, Neuroscios, Boehringer Ingelheim, Novartis, research grants from Hopp Foundation, DFG, BMBF, European Union, Stryker, Siemens, all outside the submitted work. MW has received research grants from Abbvie, Adastra, Dracen, Merck,

Sharp & Dohme (MSD), Merck (EMD) and Novocure, and honoraria for lectures or advisory board participation or consulting from Abbvie, Basilea, Bristol Meyer Squibb (BMS), Celgene, Medac, Merck, Sharp & Dohme (MSD), Merck (EMD), Nerviano Medical Sciences, Orbus, Philogen, Roche and Tocagen. PD, SW, HL, DB, Aca, ACo, FD, PF JF, NG, EH, JMH, GM, BOB, TJP, NOS, TJS, ST, AVDH, GV, KJ, MG have nothing to disclose.

Authorship statement. Experimental design and implementation: ELR, MW. Acquisition, analysis, or interpretation of data: all authors. Statistical analysis: PD. Manuscript preparation: ELR, MW. Manuscript approval: all authors.

References

1. Le Rhun E, Devos P, Boulanger T, et al. The RANO Leptomeningeal Metastasis Group proposal to assess response to treatment: lack of feasibility and clinical utility and a revised proposal. *Neuro Oncol.* 2019;21(5):648–658.
2. Le Rhun E, Weller M, Brandsma D, et al. EANO-ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up of patients with leptomeningeal metastasis from solid tumours. *Ann Oncol.* 2017;28(suppl_4):iv84–iv99.
3. Nevel KS, DiStefano N, Lin X, et al. A retrospective, quantitative assessment of disease burden in patients with leptomeningeal metastases from nonsmall-cell lung cancer. *Neuro-oncology.* 2020;22(5):675–683. doi:10.1093/neuonc/noz208
4. Le Rhun E, Devos P, Weller J, et al. Prognostic validation and clinical implications of the EANO ESMO classification of leptomeningeal metastasis from solid tumors. *Neuro Oncol.* 2021;23(7):1100–1112. doi:10.1093/neuonc/noaa298.
5. Chamberlain M, Soffietti R, Raizer J, et al. Leptomeningeal metastasis: a Response Assessment in Neuro-Oncology critical review of endpoints and response criteria of published randomized clinical trials. *Neuro-oncology.* 2014;16(9):1176–1185.
6. Chamberlain M, Junck L, Brandsma D, et al. Leptomeningeal metastases: a RANO proposal for response criteria. *Neuro-oncology.* 2017;19(4):484–492. doi:10.1093/neuonc/now183
7. Le Rhun E, Devos P, Boulanger T, et al. The RANO leptomeningeal metastasis group proposal to assess response to treatment: lack of feasibility and clinical utility, and a revised proposal. *Neuro-oncology.* 2019;21(5):648–658. doi:10.1093/neuonc/noz024
8. Brix MK, Westman E, Simmons A, et al. The Evans' Index revisited: new cut-off levels for use in radiological assessment of ventricular enlargement in the elderly. *Eur J Radiol.* 2017;95:28–32.
9. Cohen J. A coefficient of agreement for nominal scales. *Educat Psychol Measur.* 1960;20:37–46.
10. Lin NU, Borges V, Anders C, et al. Intracranial efficacy and survival with tucatinib plus trastuzumab and capecitabine for previously treated HER2-Positive breast cancer with brain metastases in the HER2CLIMB trial. *JCO.* 2020;38(23):2610–2619. doi:10.1200/JCO.20.00775
11. Gadgeel SM, Lukas RV, Goldschmidt J, et al. Atezolizumab in patients with advanced non-small cell lung cancer and history of asymptomatic, treated brain metastases: exploratory analyses of the phase III OAK study. *Lung Cancer.* 2019;128:105–112.
12. Tawbi HA, Forsyth PA, Algazi A, et al. Combined nivolumab and ipilimumab in melanoma metastatic to the brain. *N Engl J Med.* 2018;379(8):722–730.
13. Long GV, Atkinson V, Lo S, et al. Combination nivolumab and ipilimumab or nivolumab alone in melanoma brain metastases: a multicentre randomised phase 2 study.5. *Lancet Oncol.* 2018;19:672–681.
14. Deol M, Palotai M, Pinzon AM, et al. Identification and characterization of leptomeningeal metastases using SPINE, a web-based collaborative platform. *J Neuroimaging.* 2021;31(2):324–333.
15. Ellingson BM, Bendszus M, Boxerman J, et al. Consensus recommendations for a standardized brain tumor imaging protocol in clinical trials. *Neuro-oncology.* 2015;17(9):1188–1198.
16. Kaufmann TJ, Smits M, Boxerman J, et al. Consensus recommendations for a standardized brain tumor imaging protocol for clinical trials in brain metastases (BTIP-BM). *Neuro-oncology.* 2020;22(6):757–772. doi:10.1093/neuonc/noaa030
17. Kickingereder P, Bonekamp D, Nowosielski M, et al. Radiogenomics of glioblastoma: machine learning-based classification of molecular characteristics by using multiparametric and multiregional MR imaging features. *Radiology* 2016;281(3):907–918.