



# Imaging Biomarkers of Glioblastoma Treatment Response: A Systematic Review and Meta-Analysis of Recent Machine Learning Studies

Thomas C. Booth<sup>1,2\*</sup>, Mariusz Grzeda<sup>1</sup>, Alysha Chelliah<sup>1</sup>, Andrei Roman<sup>3,4</sup>, Ayisha Al Busaidi<sup>2</sup>, Carmen Dragos<sup>5</sup>, Haris Shuaib<sup>6,7</sup>, Aysha Luis<sup>2</sup>, Ayesha Mirchandani<sup>8</sup>, Burcu Alparlan<sup>2,9</sup>, Nina Mansoor<sup>2</sup>, Jose Lavrador<sup>10</sup>, Francesco Vergani<sup>10</sup>, Keyoumars Ashkan<sup>10</sup>, Marc Modat<sup>1</sup> and Sebastien Ourselin<sup>1</sup>

## OPEN ACCESS

### Edited by:

Tone Frost Bathen,  
Norwegian University of Science and  
Technology, Norway

### Reviewed by:

Endre Grøvik,  
Møre and Romsdal Hospital Trust,  
Norway  
Weiwei Zong,  
Henry Ford Health System,  
United States

### \*Correspondence:

Thomas C. Booth  
tombooth@doctors.org.uk

### Specialty section:

This article was submitted to  
Cancer Imaging and  
Image-directed Interventions,  
a section of the journal  
Frontiers in Oncology

**Received:** 21 October 2021

**Accepted:** 03 January 2022

**Published:** 31 January 2022

### Citation:

Booth TC, Grzeda M, Chelliah A, Roman A, Al Busaidi A, Dragos C, Shuaib H, Luis A, Mirchandani A, Alparlan B, Mansoor N, Lavrador J, Vergani F, Ashkan K, Modat M and Ourselin S (2022) Imaging Biomarkers of Glioblastoma Treatment Response: A Systematic Review and Meta-Analysis of Recent Machine Learning Studies. *Front. Oncol.* 12:799662. doi: 10.3389/fonc.2022.799662

<sup>1</sup> School of Biomedical Engineering & Imaging Sciences, King's College London, St. Thomas' Hospital, London, United Kingdom,

<sup>2</sup> Department of Neuroradiology, King's College Hospital National Health Service Foundation Trust, London, United Kingdom,

<sup>3</sup> Department of Radiology, Guy's & St. Thomas' National Health Service Foundation Trust, London, United Kingdom, <sup>4</sup> Department

of Radiology, The Oncology Institute "Prof. Dr. Ion Chiricuța" Cluj-Napoca, Cluj-Napoca, Romania, <sup>5</sup> Department of Radiology,

Buckinghamshire Healthcare National Health Service Trust, Amersham, United Kingdom, <sup>6</sup> Department of Medical Physics, Guy's &

St. Thomas' National Health Service Foundation Trust, London, United Kingdom, <sup>7</sup> Institute of Psychiatry, Psychology &

Neuroscience, King's College London, London, United Kingdom, <sup>8</sup> Department of Radiology, Cambridge University Hospitals

National Health Service Foundation Trust, Cambridge, United Kingdom, <sup>9</sup> Department of Radiology, Kocaeli University, Izmit, Turkey,

<sup>10</sup> Department of Neurosurgery, King's College Hospital National Health Service Foundation Trust, London, United Kingdom

**Objective:** Monitoring biomarkers using machine learning (ML) may determine glioblastoma treatment response. We systematically reviewed quality and performance accuracy of recently published studies.

**Methods:** Following Preferred Reporting Items for Systematic Reviews and Meta-Analysis: Diagnostic Test Accuracy, we extracted articles from MEDLINE, EMBASE and Cochrane Register between 09/2018–01/2021. Included study participants were adults with glioblastoma having undergone standard treatment (maximal resection, radiotherapy with concomitant and adjuvant temozolomide), and follow-up imaging to determine treatment response status (specifically, distinguishing progression/recurrence from progression/recurrence mimics, the target condition). Using Quality Assessment of Diagnostic Accuracy Studies Two/Checklist for Artificial Intelligence in Medical Imaging, we assessed bias risk and applicability concerns. We determined test set performance accuracy (sensitivity, specificity, precision, F1-score, balanced accuracy). We used a bivariate random-effect model to determine pooled sensitivity, specificity, area-under the receiver operator characteristic curve (ROC-AUC). Pooled measures of balanced accuracy, positive/negative likelihood ratios (PLR/NLR) and diagnostic odds ratio (DOR) were calculated. PROSPERO registered (CRD42021261965).

**Results:** Eighteen studies were included (1335/384 patients for training/testing respectively). Small patient numbers, high bias risk, applicability concerns (particularly confounding in reference standard and patient selection) and low level of evidence, allow limited conclusions from studies. Ten studies (10/18, 56%) included in meta-analysis gave 0.769 (0.649–0.858) sensitivity [pooled (95% CI)]; 0.648 (0.749–0.532) specificity; 0.706

(0.623-0.779) balanced accuracy; 2.220 (1.560-3.140) PLR; 0.366 (0.213-0.572) NLR; 6.670 (2.800-13.500) DOR; 0.765 ROC-AUC.

**Conclusion:** ML models using MRI features to distinguish between progression and mimics appear to demonstrate good diagnostic performance. However, study quality and design require improvement.

**Keywords:** glioblastoma, machine learning, monitoring biomarkers, meta-analysis, artificial intelligence, treatment response, deep learning, glioma

## 1 INTRODUCTION

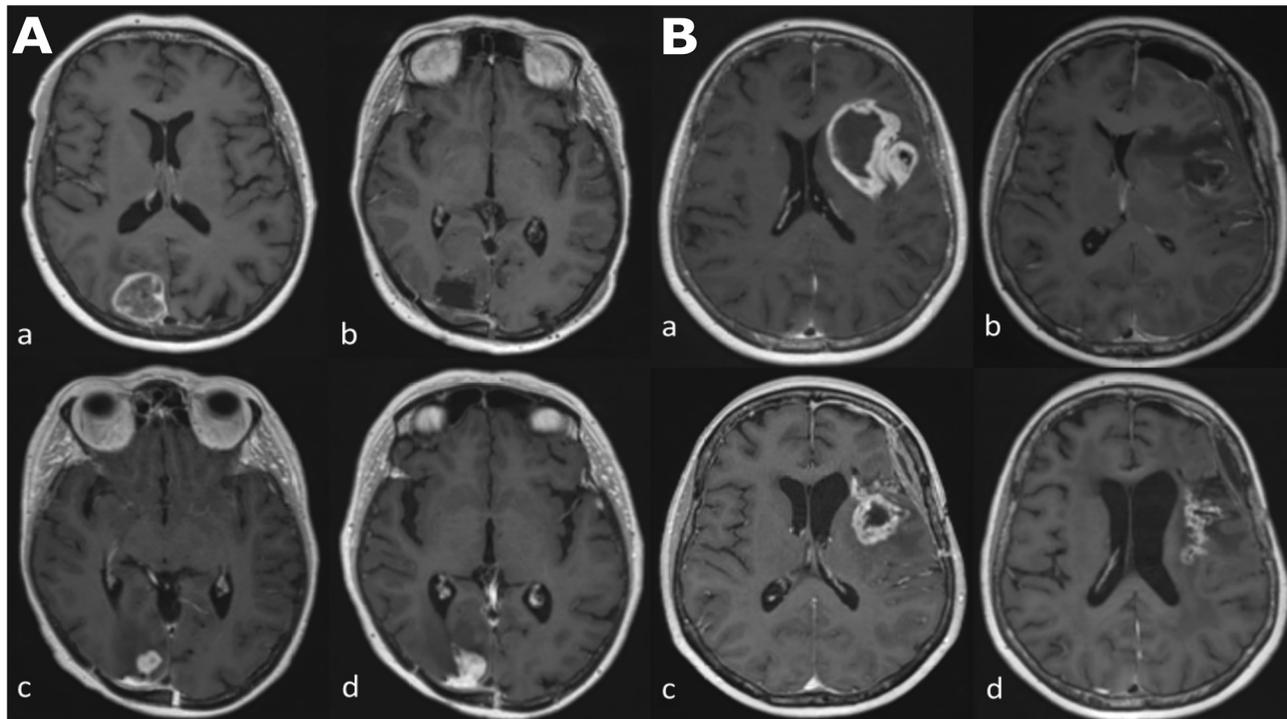
Glioblastoma is the most common primary malignant brain tumor with a median 14.6 month overall survival (1). This is in spite of a standard care regimen comprising maximal debulking surgery, followed by radiotherapy with concomitant temozolomide, followed by adjuvant temozolomide. Monitoring biomarkers (2) identify longitudinal change in the growth of tumor or give evidence of response to treatment, with magnetic resonance imaging (MRI) proving particularly useful in this regard. This is due both to the non-invasive nature of MRI, and its ability to capture the entire tumor volume and adjacent tissues, leading to its recommended incorporation into treatment response evaluation guidelines in trials (3, 4). Yet challenges occur when false-positive progressive disease (pseudoprogression) is encountered, which may take place during the 6 month period following the completion of radiotherapy and is manifest as an increase in contrast enhancement on  $T_1$ -weighted MRI images, which reflects the non-specific disruption of the blood-brain barrier (Figure 1) (5, 6).

Non-specific increased contrast enhancement occurs in approximately 50% of patients undergoing the standard care regimen. There is an approximately equal chance that the tumor may represent pseudoprogression or true progression because pseudoprogression occurs in approximately 10-30% of all patients (7, 8). For more than a decade, researchers have attempted to distinguish pseudoprogression from true progression at the time of increased contrast enhancement because of the substantial potential clinical impact. If there is true progression the treating clinical team typically will initiate a prompt modification in treatment strategy with termination of ineffectual treatment or initiation of second-line surgery or therapies (9). If there is pseudoprogression the treating clinical team typically will continue with the standard care regimen. However, the decision making can only be made retrospectively with current treatment response evaluation guidelines (4). A monitoring biomarker (2) that reliably distinguishes pseudoprogression from true progression at the time of increased contrast enhancement would fully inform the difficult decision contemporaneously.

Under the standard care regimen, pseudoprogression occurs as an early-delayed treatment effect as opposed to radiation necrosis which is a late-delayed radiation effect (10). Radiation necrosis also manifests as non-specific increased contrast enhancement, however, pseudoprogression appears within 6 months of radiotherapy completion whereas radiation necrosis

occurs beyond 6 months. Radiation necrosis occurs with an incidence an order of magnitude less than that of pseudoprogression (11). Another difference between the two entities is that much evidence suggests that pseudoprogression is significantly correlated with O<sup>6</sup>-methylguanine DNA methyltransferase (MGMT) promoter methylation. As with pseudoprogression, there is a need to distinguish radiation necrosis from true progression at the time of increased contrast enhancement because, again, there is substantial potential clinical impact. In particular, if there is true progression the treating clinical team typically would initiate second-line surgery or therapies. However, the decision making can only be made retrospectively with current treatment response evaluation guidelines (3). Therefore, a monitoring biomarker (2) that reliably distinguishes radiation necrosis from true progression at the time of increased contrast enhancement would fully inform the treating clinical team's decision contemporaneously.

Developing monitoring biomarkers to determine treatment response has been the subject of many studies, with many incorporating machine learning (ML). A review of such neuro-oncology studies up to September 2018 showed that the evidence is relatively low level, given that it has usually been obtained in single centers retrospectively and often without hold-out test sets (11, 12). The review findings suggested that those studies taking advantage of enhanced computational processing power to build neuro-oncology monitoring biomarker models, for example deep learning techniques using convolutional neural networks (CNNs), have yet to show benefit compared to ML techniques using explicit feature engineering and less computationally expensive classifiers, for example using support vector machines or even multivariate logistic regressions. Furthermore, studies show that using ML to make neuro-oncology monitoring biomarker models does not appear to be superior to applying traditional statistical methods when analytical validation and diagnostic performance is considered (the fundamental difference between ML and statistics is that statistics determines population inferences from a sample, whereas ML extracts generalizable predictive patterns). Nonetheless, the rapidly evolving discipline of applying radiomic studies to neuro-oncology imaging reflects a recent exponential increase in published studies applying ML to neuroimaging (13), and specifically to neuro-oncology imaging (14). It also mirrors the notable observation that in 2018, arXiv (a repository where computer science papers are self-archived before publication in a peer-reviewed journal) surpassed 100 new ML pre-prints per day (15). Given these developments, there is a need to appraise the



**FIGURE 1** | Longitudinal series of MRI images in two patients (**A, B**) with glioblastoma, IDH-wildtype. All images are axial  $T_1$ -weighted after contrast administration. Images (**Aa–Ad**) demonstrate tumor progression. (**Aa**) Pre-operative MRI of a glioblastoma in the occipital lobe. (**Ab**) Post-operative MRI five days after resection; there is no contrast enhancement therefore no identifiable residual tumor. (**Ac**) The patient underwent a standard care regimen of radiotherapy and temozolomide. A new enhancing lesion at the inferior margin of the post-operative cavity was identified on MRI at three months after radiotherapy completion. (**Ad**) The enhancing lesion continued to increase in size three months later and was confirmed to represent tumor recurrence after repeat surgery. Images (**Ba–Bd**) demonstrate pseudoprogression. (**Ba**) Pre-operative MRI of a glioblastoma in the insula lobe. (**Bb**) Post-operative MRI at 24 hours after surgery; post-operative blood products are present but there is no contrast enhancement therefore no identifiable residual tumor. (**Bc**) The patient underwent a standard care regimen of radiotherapy and temozolomide. A new rim-enhancing lesion was present on MRI at five months after radiotherapy completion. (**Bd**) Follow-up MRI at monthly intervals showed a gradual reduction in the size of the rim-enhancing lesion without any change in the standard care regimen of radiotherapy and temozolomide or corticosteroid use. The image shown here is the MRI four months later.

evidence of ML applied to monitoring biomarkers determining treatment response since September 2018.

The aim of the study is to systematically review and perform a meta-analysis of diagnostic accuracy of ML-based treatment response monitoring biomarkers for glioblastoma patients using recently published peer-reviewed studies. The study builds on previous work to incorporate the rapidly growing body of knowledge in this field (11, 16), providing promising avenues for further research.

## 2 MATERIALS AND METHODS

This systematic review and meta-analysis are registered with PROSPERO (CRD42021261965). The review was organized in line with the Preferred Reporting Items for Systematic Reviews and Meta-Analysis: Diagnostic Test Accuracy (PRISMA-DTA) (17) incorporating Cochrane review methodology relating to “developing criteria for including studies” (18), “searching for studies” (19), and “assessing methodological quality” (20).

Pseudoresponse (bevacizumab-related response mimic), an important concern in the United States where it is licensed, was not the focus of the systematic review and meta-analysis.

### 2.1 Search Strategy and Selection Criteria

Recommendations were followed to perform a sensitive search (with low precision), including the incorporation of subject headings with exploded terms, and without any language restrictions (19). Search terms were applied to MEDLINE, EMBASE and the Cochrane Register to capture original research articles published from September 2018 to January 2021 (**Supplementary Table S1**). Pre-prints and non-peer reviewed material were excluded.

#### 2.1.1 Inclusion Criteria

Study participants included were adult glioblastoma patients treated with a standard care regimen (maximal debulking surgery, followed by radiotherapy with concomitant temozolomide, followed by adjuvant temozolomide) who underwent follow-up imaging to determine treatment response

status (explicitly, differentiating true progression/recurrence from mimics of progression/recurrence (defined below), and designated as the target condition of the systematic review).

### 2.1.2 Exclusion Criteria

Studies were excluded if they focused on pediatrics, pseudoresponse, or had no ML algorithm employed in the extraction or selection of features, or in classification/regression.

### 2.1.3 Index Test and Reference Standard

The ML model determined the treatment response outcome, and was designated as the index test of the systematic review. Either clinicoradiological follow up or histopathology at re-operation or a combination of both, were designated as the reference standard of the systematic review. The bibliography of each included article was checked manually for other relevant studies.

A neuroradiologist, T.C.B., and a data scientist, A.C., with 16 and 2 years, respectively, of experience in neuroimaging applied to neuro-oncology, independently performed the literature search and selection.

## 2.2 Data Extracted and Risk of Bias Assessment

For every study, risk of bias as well as concerns regarding applicability, were assessed by applying QUADAS 2 methodology (21) alongside proformas incorporating items from the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) (22). Data was extracted from published studies to determine: whether the datasets analyzed contained any tumors other than glioblastomas, especially anaplastic astrocytomas and anaplastic oligodendrogliomas; the index test ML algorithm and any cross validation processes; training and hold-out test set information; what reference standard(s) were employed; non-imaging features and MRI sequence(s) included in the analysis.

The appropriateness of reference standard follow-up imaging protocols was reviewed. The handling of confounding factors such as second-line medication therapy, temozolomide cessation, and steroid use were assessed. It was also determined whether the treatment response (target condition) used in the published study was appropriate. Under the standard care regimen, contrast-enhancing lesions enlarging due to pseudoprogression typically occur within 0-6 months after radiotherapy, whereas contrast-enhancing lesions enlarging due to radiation necrosis typically occur beyond this 6 month window, according to the evidence. When “post-treatment related effects” (PTRE) is employed as a term for treatment response outcome, the phenomena of pseudoprogression and radiation necrosis are both included (23, 24). These three terms therefore capture detail regarding the time period when the mimics of progression/recurrence occur. Deviations in the use of the three terms defined here were noted. Data on the length of follow-up imaging after contrast-enhancing lesions enlarged were additionally extracted and evaluated. Clinicoradiological strategies considered optimal in designating outcomes as PTRE or true progression/recurrence included the following: assigning an MRI scan as baseline after radiotherapy (25); excluding outcomes based on  $T_2$ -w lesion enlargement (25); permitting a period of 6-month follow up from the first time

when contrast-enhancing lesions enlarged; during this 6-month period having two subsequent follow-up scans as opposed to a single short interval “confirmatory” follow-up scan. Two follow-up scans mitigate against some scenarios where the contrast-enhancing lesions due to PTRE continue to enlarge over a short interval, and this continued enlargement is seen at a short interval scan confounding assessment by falsely “confirming” true progression (26, 27). This might be termed an “upslope effect”.

A neuroradiologist (US attending, UK consultant), T.C.B., and a data scientist, A.C., with 16 and 2 years, respectively, of experience in neuroimaging applied to neuro-oncology, independently performed the data extraction and quality assessment. Discrepancies between the two reviewers were considered at a research meeting chaired by a third neuroradiologist (US attending, UK consultant), A.A-B. (8 years experience of neuroimaging applied to neuro-oncology), until a consensus was reached.

## 2.3 Data Synthesis and Statistical Analysis

### 2.3.1 Performance Accuracy for Individual Studies

Based on the published study data, 2 x 2 contingency tables were made for hold-out test sets from which the principal diagnostic accuracy measures of sensitivity (recall) and specificity were calculated. The area under the receiver operating characteristic curve (ROC-AUC) values and confidence intervals were extracted in studies where these were published. Additional secondary outcome measures of balanced accuracy, precision (positive predictive value) and F1-score were also determined from the contingency tables. In those studies where there was a discrepancy in the principal diagnostic accuracy measures and the accessible published study raw data, this was highlighted. If both internal and external hold-out test sets were published in a study, the principal diagnostic accuracy measures for the external test set alone were calculated. In studies without hold-out test sets, “no test set” was recorded (22) and the training set principal diagnostic accuracy measures from the training set were summarized. The unit of evaluation was per-patient. All test set data included glioblastoma.

### 2.3.2 Meta-Analysis

The principal diagnostic accuracy measures of sensitivity (recall) and specificity were subject to meta-analysis. We determined two pooled primary measures of accuracy: the true positive rate (sensitivity/recall) and the false negative rate (1-specificity). A bivariate random-effect model (28), which allows for two important circumstances (29–31) (**Supplementary Statistical Information**), was chosen to determine the two pooled primary measures of accuracy. Briefly, the circumstances are first, that the values of the selected principal diagnostic accuracy measures are usually highly related to one another through the cut-off value. With an increase of sensitivity, specificity is likely to decrease and, as a consequence, these two measures are usually negatively correlated. Second, a relatively high level of heterogeneity is commonly observed among the results of diagnostic studies. This is verified in various ways ranging from visual assessment through chi-square based tests to random-intercept models decomposing total variance of results into between- and within- study levels. The bivariate random-effect model not only allows for the simultaneous analysis

of diagnostic measures but also addresses their heterogeneity (28). Bivariate joint modelling of the primary measures of accuracy assumes that the logits of these quantities follow a bivariate normal distribution and allows for a non-zero correlation. Based on this assumption, a linear random-effect model is applied to the data and estimates of mean true positive rate (sensitivity) and false positive rate (1-specificity), along with their variances and correlation between them, can be obtained. The pooled estimates of true positive rate and false positive rate are initially estimated on the logit scale (**Supplementary Statistical Information**). To be interpretable they require transformation back to the original probability scale (ranging within 0-1 limits).

The parameters of this model also allowed us to plot the summary ROC (SROC) curve and determine the summary ROC-AUC. Using a resampling approach (32), the model estimates were also used to derive the pooled measures of balanced accuracy as well as the positive and negative likelihood ratios and the diagnostic odds ratio.

The meta-analysis was conducted by a statistician, M.G., with 15 years of relevant experience. All the statistical analyses were performed in R (v 3.6.1). The R package *mada* (v 0.5.10) (33) was used for the bivariate model. Since some of the 2 x 2 contingency table input cell values (true positive, false positive, false negative, true negative) derived from the individual studies contained zeros, a continuity correction (0.5) was applied.

## 2.4 Prognostic Biomarkers Predicting Future Treatment Response

Most studies of prognostic imaging biomarkers in glioblastoma predict the outcome measure of overall survival using baseline images. Nonetheless, we found a small group of studies using ML models that predicted the outcome measure of future treatment response using baseline images. The studies were examined using identical methodology to that applied to monitoring biomarkers.

## 3 RESULTS

### 3.1 Characteristics and Bias Assessment of Studies Included

In all, 2362 citations fulfilled the search criteria of which the full text of 57 potentially eligible articles were reviewed (**Figure 2**). Twenty-one studies from September 2018 to January 2021 (including the publication of “online first” articles prior to September 2018) were included, 19 of which were retrospective. The total number of patients in the training sets were 1335 and in the test set 384. The characteristics of the 18 monitoring biomarker studies are presented in **Table 1** and the characteristics of the 3 studies that applied the ML models to serve as prognostic biomarkers to predict future treatment response using baseline images (or genomic alterations) are presented in **Table 2**.

#### 3.1.1 Treatment Response Target Conditions

The treatment response target conditions varied between studies (**Table 1**). Around a quarter of studies (5/18, 28%) designated only 0-12 weeks after radiotherapy as the time period when pseudoprogression appears – as opposed to the entire 6-month

time period when pseudoprogression might occur. A third of studies (6/18, 33%) assigned PTRE as the target condition. No study assigned radiation necrosis alone as the target condition. Five studies in the systematic review (5/18, 28%) included grade 3 gliomas. Only two of these five studies employed test sets; the test set in one study did not contain any grade 3 gliomas and the number in the test set in the other study was unclear although the number was small (14% grade 3 in combined training and test datasets). Therefore, as a minimum, all but one test set in the systematic review and meta-analysis contained only glioblastoma, the previous equivalent of glioma grade 4 according to c-IMPACT classification (“glioblastomas, IDH-wildtype” or “astrocytoma, IDH-mutant, grade 4”) (55).

#### 3.1.2 Reference Standards: Clinicoradiological Follow-Up and Histopathology Obtained at Re-Operation

The majority of studies (13/18, 67%) employed a combination of clinicoradiological follow up and histopathology at re-operation, to distinguish true progression from a mimic. A few individual studies employed one reference standard for one decision (true progression) and another reference standard for the alternative decision (mimic); this and other idiosyncratic rules led to a high risk of bias in terms of the reference standard used, as well as how patients were selected, in several studies.

#### 3.1.3 Selected Features

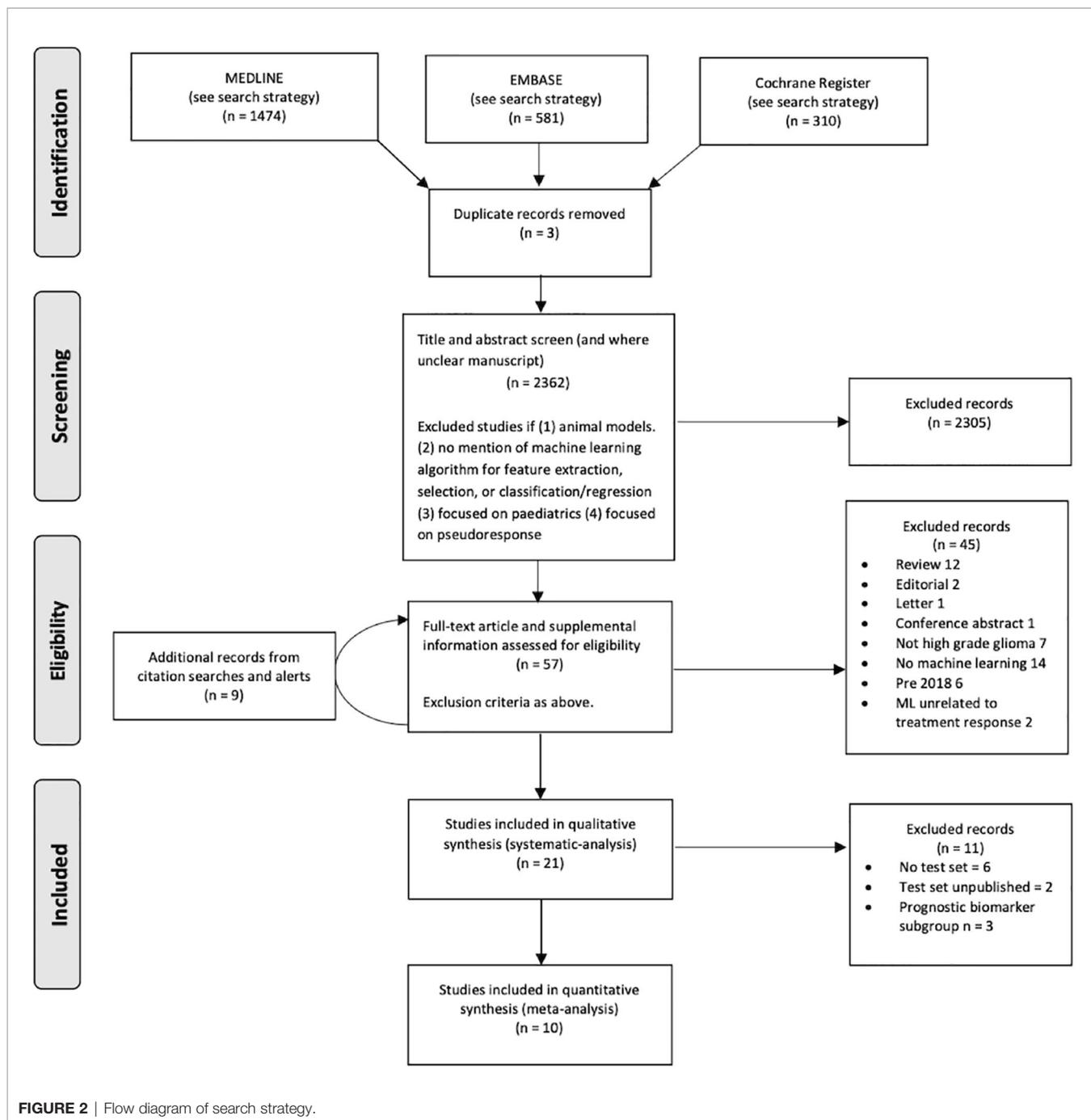
Most studies only analyzed imaging features alone (15/18, 83%) whereas the remainder incorporated additional non-imaging features. A third of studies (6/18, 33%) used deep learning methodology to derive features (specifically, convolutional neural networks).

#### 3.1.4 Test Sets

A third of studies did not have hold-out test sets (6/18, 33%) and instead the performance accuracy was determined using training data through cross-validation (**Table 1**). Therefore, there was a high risk of bias for the index test used in these six studies. A third of studies had external hold-out test sets (6/18, 33%). The ranges of mean diagnostic accuracy measures in these six studies were: recall (sensitivity) = 0.61-1.00; specificity = 0.47-0.90; precision (positive predictive value) = 0.58-0.88; balanced accuracy = 0.54-0.83; F1 score = 0.59-0.94; ROC-AUC = 0.65-0.85.

#### 3.1.5 Bias Assessment and Concerns Regarding Applicability Summary

The risk of bias evaluation for each study was summarized (**Supplementary Figure S1**). All or most studies were assigned to the highest class for risk of bias in terms of the reference standard (18/18, 100%) and patient selection (15/18, 83%) QUADAS 2 categories respectively. A third or nearly a half of studies were either in the highest class for risk of bias or the risk was unclear in terms of flow and timing (6/18, 33%) and the index test (8/28, 44%) QUADAS 2 categories respectively. The results from the “concerns regarding applicability” evaluation largely mirrored the results of the risk of bias evaluation.



### 3.1.6 Prognostic Biomarkers Predicting Future Treatment Response (Subgroup)

There were two studies which were prospective, both of which had a small sample size ( $n = 10$ ); the third study in this subgroup was retrospective. One study applied genomic alterations alone as features to predict future MRI treatment response. All studies (3/3, 100%) were in the highest class for risk of bias in terms of the reference standard, patient selection

and index test QUADAS 2 categories (**Supplementary Figure S2**). In terms of “concerns regarding applicability” evaluation, the results mirrored the risk of bias evaluation exactly. Diagnostic accuracy measures could not be calculated because of study design. Design constraints included units of assessment in one study being per-lesion whilst another was per-voxel. One study also incorporated a prognostic metric of 1-year progression free survival for the predicted treatment

**TABLE 1** | Studies using machine learning in the development of glioblastoma monitoring biomarkers.

Author	Target condition	Reference standard	Dataset(s)	Available demographic information	Methodology	Features selected	Test set performance
<sup>a</sup> Kim J.Y. et al. (34)	Early true progression or Early pseudoprogression	Mixture of histopathology and imaging follow up	Training = 61 Testing = 34 T <sub>1</sub> C, FLAIR, DWI, DSC	Training = age mean ± SD (range) 58 ± 11 (34–83) male 38 (62%) Testing = age mean ± SD 62 ± 12 male 25 (74%) Data from Korea	Retrospective 2 centers: 1 train & 1 external test set. LASSO feature selection with 10-fold CV Linear generalized model	First-order, volume/shape (texture), wavelet. Second-order (texture), wavelet. ADC & CBV parameters included.	Recall 0.71 Specificity 0.90 Precision 0.83 BA 0.81 F1 0.77 AUC 0.85 (CI 0.71 – 0.99)
Kim J.Y. et al. (35)	Early true progression or Early pseudoprogression	Mixture of histopathology and imaging follow up	Training = 59 Testing = 24 T <sub>1</sub> C, FLAIR, DTI, DSC	Training = age mean ± SD 61 ± 11 male 37 (63%) Testing = age mean ± SD 59 ± 12 male 9 (38%) Data from Korea	Retrospective 1 center LASSO feature selection with 10-fold CV Linear generalized model	First-order, Second-order (texture), wavelet. FA & CBV parameters included.	Recall 0.80 Specificity 0.63 Precision 0.36 BA 0.72 F1 0.50 AUC 0.67 (0.40 – 0.94)
Bacchi S. et al. (36)	True progression or PTRE (HGG)	Histopathology for progression and imaging follow up for pseudoprogression	Training = 44 Testing = 11 T <sub>1</sub> C, FLAIR, DWI	Combined = age mean ± SD 56 ± 10 male 26 (47%) Data from Australia	Retrospective 1 center 3D CNN & 5-fold CV	CNN. FLAIR & DWI parameters	Recall 1.00 Specificity 0.60 Precision 0.75 BA 0.80 F1 0.86 AUC 0.80
Elshafeey N. et al. (37)	True progression or <sup>b</sup> PTRE	Histopathology	Training = 98 Testing = 7 DSC, DCE	Training = age mean ± SD 50 ± 13 male 14 (58%) No testing demographic information Data from USA	Retrospective 3 centers mRMR feature selection. 1 test. 1) decision tree algorithm C5.0 2) SVM including LOO and 10-fold CV	K <sub>trans</sub> & CBV parameters	Insufficient published data to determine diagnostic performance (CV training results available recall 0.91; specificity 0.88)
Verma G. et al. (38)	True progression or Pseudoprogression	Mixture of histopathology and imaging follow up	Training = 27 3D-EPSI	Training = age mean ± SD 64 ± 10 male 14 (52%) Data from USA	Retrospective 1 center Multivariate logistic regression LOOCV	Cho/NAA & Cho/Cr	No test set (CV training results available recall 0.94; specificity 0.87)
Ismail M. et al. (39)	True progression or Pseudoprogression	Mixture of histopathology and imaging follow up	Training = 59 Testing = 46 T <sub>1</sub> C, T <sub>2</sub> /FLAIR	Training = age mean(range) 61 (26–74) male 39 (66%) Testing = age mean (range) 56 (25–76) male 30 (65%) Data from USA	Retrospective 2 centers: 1 train & 1 external test set. SVM & 4-fold CV	Global & curvature shape	Recall 1.00 Specificity 0.67 Precision 0.88 BA 0.83 F1 0.94
<sup>a</sup> Bani-Sadr A. et al. (40)	True progression or Pseudoprogression	Mixture of histopathology and imaging follow up	Training = 52 Testing = 24 T <sub>1</sub> C, FLAIR MGMT promoter status	Combined = age mean ± SD 58 ± 11 male 45 (59%) Data from France	Retrospective 1 center Random Forest.	Second-order features +/- MGMT promoter status	Recall 0.94 (0.71 - 1.00) Specificity 0.38 (0.09 - 0.76) Precision 0.36 BA 0.66 F1 0.84 AUC 0.77 & non-MRI: Recall 0.80 (0.56 - 0.94) Specificity 0.75 (0.19 - 0.99)

(Continued)

TABLE 1 | Continued

Author	Target condition	Reference standard	Dataset(s)	Available demographic information	Methodology	Features selected	Test set performance
Gao X.Y. et al. (41)	True progression or PTRE (HGG)	Mixture of histopathology and imaging follow up	Training = 34 Testing = 15 (per lesion) $T_1$ C, FLAIR	Combined = age mean $\pm$ SD 51 $\pm$ 11 male 14 (36%) (per patient) Data from China	Retrospective 2 centers SVM & 5-fold CV	$T_1$ C, FLAIR subtraction map parameters	Precision 0.86 BA 0.74 F1 0.83 AUC 0.85 Recall 1.00 Specificity 0.90 Precision 0.83 BA 0.95 F1 0.91 AUC 0.94 (0.78 – 1.00)
Jang B-S. et al. (42)	True progression or Pseudoprogression	Mixture of histopathology and imaging follow up	Training = 59 Testing = 19 $T_1$ C & clinical features & IDH/MGMT promoter status	Training = age median (range) 56 (22–77) male 41 (70%) Testing = age mean $\pm$ SD 53 (28–75) male 10 (53%) Data from Korea	Retrospective 2 centers 1 train & 1 external test set. CNN LSTM & 10-fold CV (compared to Random Forest)	CNN $T_1$ C parameters +/- Age; Gender; MGMT status; IDH mutation; radiotherapy dose and fractions; follow-up interval	Recall 0.64 Specificity 0.50 Precision 0.64 BA 0.57 F1 0.63 AUC 0.69 & non-MRI: Recall 0.72 Specificity 0.75 Precision 0.80 BA 0.74 F1 0.76 AUC 0.83
Li M. et al. (43)	True progression or $^b$ PTRE	Imaging follow up	Training = 84 DTI	No demographic information Data from USA	Retrospective. 1 center DC-AL GAN CNN with SVM including 5 and 10 and 20-fold CV (compared to DCGAN, VGG, ResNet, and DenseNet)	CNN. DTI	No test set (CV training results only available: Recall 0.98 Specificity 0.88 AUC 0.95)
Akbari H. et al. (44)	True progression or Pseudoprogression	Histopathology	Training = 40 Testing = 23 Testing = 20 $T_1$ C, $T_2$ /FLAIR, DTI, DSC, DCE	Combined internal = age mean (range) 57 (33–82) male 38 (60%) No external demographic information Data from USA	Retrospective 2 centers. 1 train & test. 1 external test set. imagenet_vgg_f CNN SVM & LOOCV	First-order, second-order (texture). CBV, PH, TR, $T_1$ C, $T_2$ /FLAIR parameters included.	Recall 0.70 Specificity 0.80 Precision 0.78 BA 0.75 F1 0.74 AUC 0.80
Li X. et al. (45)	Early True progression or early pseudoprogression (HGG)	Mixture of histopathology and imaging follow up	Training = 362 $T_1$ C, $T_2$ , multi-voxel & single-voxel 1H-MRS, ASL	Training = age mean (range) 50 (19–70) male 218 (60%) Data from China	Retrospective Gabor dictionary and sparse representation classifier (SRC)	Sparse representations	No test set (CV training results only available: Recall 0.97 Specificity 0.83)
Manning P et al. (46)	True progression or pseudoprogression	Mixture of histopathology and imaging follow up	Training = 32 DSC, ASL	Training = age mean $\pm$ SD 56 $\pm$ 13 male 22 (69%) Data from USA	Retrospective 1 center Linear discriminant analysis & LOOCV	CBF and CBV parameters included.	No test set (CV training results only available: Recall 0.92 Specificity 0.86 AUC 0.95)
Park J.E. et al., 2020 (47)	Early True progression or early pseudoprogression	Mixture of histopathology and imaging follow up	Training = 53 Testing = 33 $T_1$ C	Training = age mean $\pm$ SD 56 $\pm$ 11 male 31 (59%) Testing = age mean $\pm$ SD	Retrospective 2 centers. 1 train & test. 1 external test set.	First-order, volume/shape, Second-order (texture), wavelet	Recall 0.61 Specificity 0.47 Precision 0.58 BA 0.54

(Continued)

TABLE 1 | Continued

Author	Target condition	Reference standard	Dataset(s)	Available demographic information	Methodology	Features selected	Test set performance
				62 ± 12 male 25 (76%) Data from Korea	Random Forest feature selection with 10-fold CV (Automated segmentation)	parameters included.	F1 0.59 AUC 0.65 (0.46 – 0.84)
Lee J. et al. (48)	True progression or <sup>b</sup> PTRE (HGG)	Histopathology	Training = 43 <i>T</i> <sub>1</sub> , <i>T</i> <sub>1</sub> C, <i>T</i> <sub>2</sub> , FLAIR, (subtractions: <i>T</i> <sub>1</sub> C - <i>T</i> <sub>1</sub> , <i>T</i> <sub>2</sub> . FLAIR) ADC parameters.	Training = age mean ± SD (range) 52 ± 13 (16–74) male 24 (56%) Data from USA	Retrospective 1 center CNN-LSTM. 3-fold CV	CNN-LSTM parameters.	No test set (CV training results only available: AUC 0.81 (0.72 – 0.88))
Kebr S. et al. (49)	True progression or <sup>b</sup> PTRE	Imaging follow up	Training = 30 Testing = 14 O-(2[ <sup>18</sup> F]-fluoroethyl)- L-tyrosine (FET)	Combined = age mean ± SD (range) 57 ± 11 (34-79) male 34 (77%) Data from Germany	Retrospective 1 center Linear discriminant analysis. 3-fold CV	TBR <sub>mean</sub> TBR <sub>max</sub> TTP <sub>min</sub> parameters.	Recall 1.00 Specificity 0.80 Precision 0.90 BA 0.92 F1 0.95 AUC 0.93 (0.78 – 1.00)
Cluceru J. et al. (50)	Early True progression or early pseudoprogression (HGG)	Histopathology	Training = 139 DSC, MRSI, DWI, DTI	Training = age median (range) 52 (21–84) Male 83 (60%) Data from USA Ethnicity: White 112 (80%) American Indian 1 (1%) Asian 6 (4%) Pacific Islander 2 (1%) Other 18 (13%)	Retrospective 1 center Multivariate logistic regression. 5-fold CV	Cho, Cho/Cr, Cho/NAA & CBV parameters.	No test set (CV training results only available: Recall 0.65 (0.33 – 0.96); Specificity 0.62 (0.21 – 1.00) AUC 0.69 (0.51 – 0.87))
Jang B.S. et al. (51)	True progression or <sup>b</sup> PTRE	Mixture of histopathology and imaging follow up (including PET)	(i) (trained model = 78) testing = 104 (ii) all training = 182 <i>T</i> <sub>1</sub> C & clinical, molecular, timings, radiotherapy data	Testing = age median (range) 55 (25-76) male 59 (67%) Data from Korea	Retrospective (i) 6 centers 1 external test set. CNN LSTM (ii) 7 centers 1 training set CNN LSTM & 10-fold CV	CNN <i>T</i> <sub>1</sub> C parameters and Age; Gender; MGMT status; IDH mutation; radiotherapy dose and fractions; follow-up interval	(i) Insufficient published data to determine diagnostic performance (ii) No test set (CV training results available AUPRC 0.87)

<sup>a</sup>Within publication some data appears mathematically discrepant.

<sup>b</sup>Within publication discrepant or unclear information (e.g. interval after radiotherapy).

Unless otherwise stated, glioblastoma alone was analyzed.

PTRE, post-treatment related effects; HGG, high-grade glioma.

MRI sequences: *T*<sub>1</sub> C, postcontrast *T*<sub>1</sub>-weighted; *T*<sub>2</sub>, *T*<sub>2</sub>-weighted; FLAIR, fluid-attenuated inversion recovery; DSC, dynamic susceptibility-weighted; DCE, dynamic contrast-enhanced; DWI, diffusion-weighted imaging; DTI, diffusor tensor imaging; ASL, arterial spin labelling; MRI parameters: ADC, apparent diffusion coefficient; FA, fractional anisotropy; TR, trace (DTI); CBV, cerebral blood volume; PH, peak height; *K*<sub>trans</sub>, volume transfer constant.

Magnetic resonance spectroscopy: 1H-MRS, 1H-magnetic resonance spectroscopy; 3D-EPSI, 3D echo planar spectroscopic imaging.

1H-MRS parameters: Cr, creatine; Cho, choline; NAA, N-acetyl aspartate.

Nuclear medicine: TBR, tumor-to-brain ratio; TTP, time-to-peak.

Molecular markers: MGMT, O6-methylguanine-DNA methyltransferase; IDH, isocitrate dehydrogenase.

Machine learning methodology: CV, cross validation; LOOCV, leave-one-out cross validation; SVM, support vector machine; CNN, convolutional neural network; LASSO, least absolute shrinkage and selection operator; LSTM, long short-term memory; mRMR, minimum redundancy and maximum relevance; VGG, Visual Geometry Group (algorithm); DCGAN, deep convolutional generative adversarial network; DC-AL GAN, DCGAN with AlexNet.

Statistical measures: CI, confidence intervals; BA, balanced accuracy; AUC, area under the receiver operator characteristic curve; AUPRC, area under the precision-recall curve.

**TABLE 2** | Studies applying machine learning models to baseline MRI images (or genomic signatures) to operate as glioblastoma prognostic biomarkers to predict future treatment response.

Author	Target condition	Reference standard	Dataset(s)	Available demographic information	Methodology	Features selected	Test set performance
Wang S. et al. (52)	True progression or pseudoprogression (immunotherapy for EGFRvIII mutation) Baseline prediction	Histopathology	model testing set = 10 DTI, DSC and 3D-EPSI	Testing = age mean (range) 55 (45-77) ± 8 male 4 (40%) Data from USA	Prospective. 1 center. Multivariate logistic regression.	CL, CBV, FA parameters	Insufficient data to determine per patient diagnostic performance (per lesion results only available: Recall = 0.86 Specificity = 0.60)
Yang K. et al. (53)	True progression or not (stable disease, partial & complete response & pseudoprogression) Baseline prediction	Imaging follow up	Training = 49 Genomic alterations	Training = age median (range) 57 (22-82) male 30 (61%) Data from Korea	Retrospective. 1 center. Analysis including Gene Set Enrichment Analysis (GSEA).	Genomic alterations including CDKN2A and EGFR mutations	No test set (Insufficient data to determine per patient diagnostic performance. From training dataset: 1-year PFS for responder 45%; non-responder 0%)
Lundemann M. et al. (54)	Early recurrence or not (voxel-wise) Baseline prediction	Mixture of histopathology and imaging follow up	Training = 10 18F-FET PET/CT; 18F-FDG PET/MRI; T <sub>1</sub> C; T <sub>2</sub> /FLAIR; DTI; DCE	Training = age mean (range) 54 (40-71) male 7 (78%) Data from Denmark	Prospective. 1 center. Multivariate logistic regression LOOCV.	FET; FDG; MD, FA; F, Vb, Ve, Ki, and MTT parameters.	No test set (Insufficient data to determine per patient diagnostic performance. From training dataset: Voxel-wise recurrence probability AUC 0.77)

EGFR, epidermal growth factor receptor; EGFRvIII, epidermal growth factor receptor variant III; CDKN2A, cyclin-dependent kinase Inhibitor 2A.

MRI sequences: T<sub>1</sub> C, post-contrast T<sub>1</sub>-weighted; T<sub>2</sub>, T<sub>2</sub>-weighted; FLAIR, fluid-attenuated inversion recovery; DSC, dynamic susceptibility-weighted; DCE, dynamic contrast-enhanced; DTI, diffusor tensor imaging.

Other imaging techniques: 3D-EPSI, 3D echo planar spectroscopic imaging; PET/CT, positron emission tomography and computed tomography; PET/MRI, positron emission tomography and magnetic resonance imaging; 18F-FDG, [18F]-fluorodeoxyglucose; 18F-FET, [18F]-fluoroethyl-L-tyrosine.

MRI parameters: FA, fractional anisotropy; MD, mean diffusivity; CL, linear anisotropy; CBV, cerebral blood volume; MTT, mean transit time; F, blood flow; Ve, extra-vascular extra-cellular blood volume; Vb, vascular blood volume; Ki, vascular permeability.

Statistical and machine learning methodology: LOOCV, leave-one-out cross validation; AUC, area under the receiver operator characteristic curve; PFS, progression free survival.

response target condition. Overall, the studies are best considered as proof of concept. Overall, there was insufficient data to perform a subgroup meta-analysis.

### 3.2 Results of Meta-Analysis

Eleven studies appeared eligible for inclusion in a meta-analysis of monitoring biomarker studies as there was information regarding internal or external hold-out test set data. However, one test was ineligible ( $n < 10$ ; 3 cells in the 2 x 2 contingency table  $n = 0$ ). Ten (10/18, 56%) remaining studies were subject to further analyses. Forest plots of sensitivity and specificity (**Figure 3**) graphically showed a high level of heterogeneity. Also, chi-square tests were applied separately to both primary measures. The p values resulting from these tests were 0.017 and 0.110 for sensitivities and specificities, respectively thus indicating the significant heterogeneity. This supported the choice of the bivariate random-effect model. The pooled true positive rate (sensitivity) = 0.769 (0.649 - 0.858) and the pooled false positive rate (1-specificity) = 0.352 (0.251 - 0.468).

A scatter plot of false positive rates (1-specificity) and true positive rates (sensitivity) shown in **Figure 4** demonstrates individual ROC point estimates and a summary ROC (SROC) curve giving summary ROC-AUC = 0.765.

The derived pooled measures of balanced accuracy = 0.706 (0.623-0.779); positive likelihood ratio = 2.220 (1.560-3.140); negative likelihood ratio = 0.366 (0.213- 0.572); diagnostic odds ratio = 6.670 (2.800-13.500).

## 4 DISCUSSION

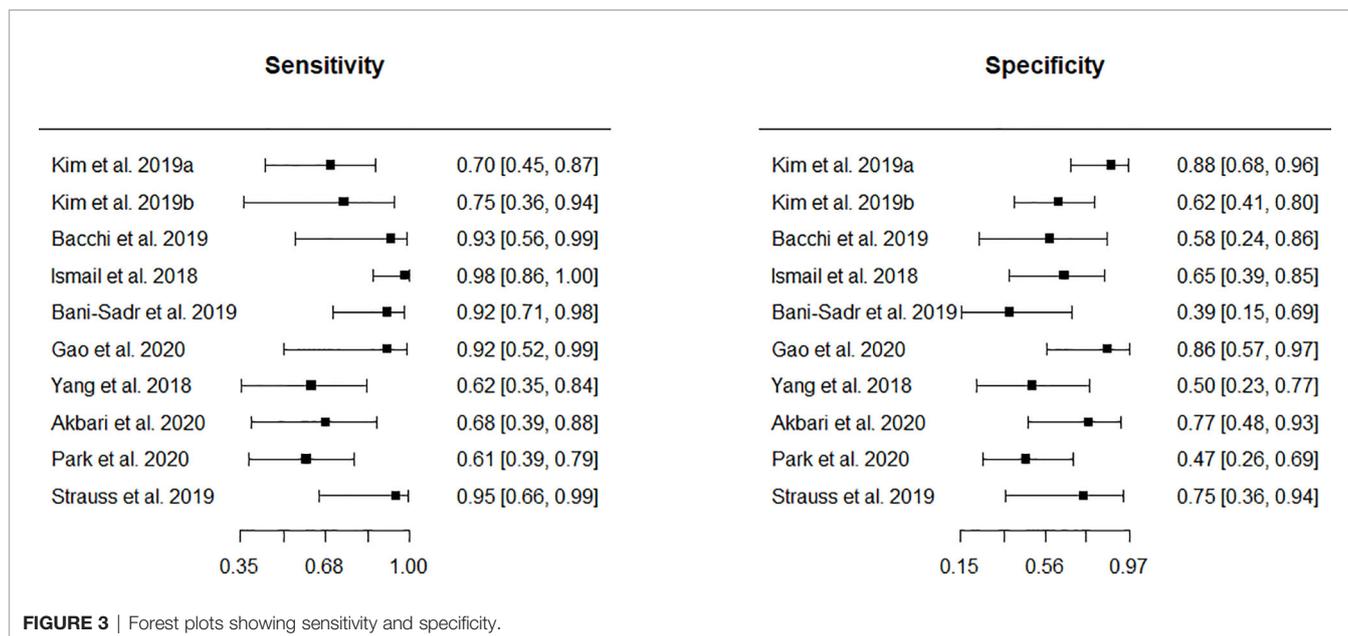
### 4.1 Summary of Findings

To date, available evidence is relatively low level (12) for determining the diagnostic accuracy of ML-based glioblastoma treatment response monitoring biomarkers in adults. The available evidence is subject to a number of limitations because recent studies are at a high risk of bias and there are concerns about its applicability, especially when determining the status of response to treatment using the reference standards of follow-up imaging or pathology at re-operation. There are similar and associated concerns regarding the selection of study patients. A third of the studies did not include any type of hold-out test set. Most of the studies employed classic ML approaches based on radiomic features. A third of studies employed deep learning methodologies.

### 4.2 Limitations

#### 4.2.1 Studies Assessed

Limitations encompassed three main areas. First, the reference standards used in all studies resulted in a high risk of bias and concerns about applicability. With the exception of the prognostic biomarker subgroup of studies, all the studies were retrospective, which increased the risk of confounding. Confounding factors, in relation to imaging follow-up and pathology at re-operation reference standards, were second-line drug therapy and cessation of temozolomide, all of which were rarely considered. Likewise, the



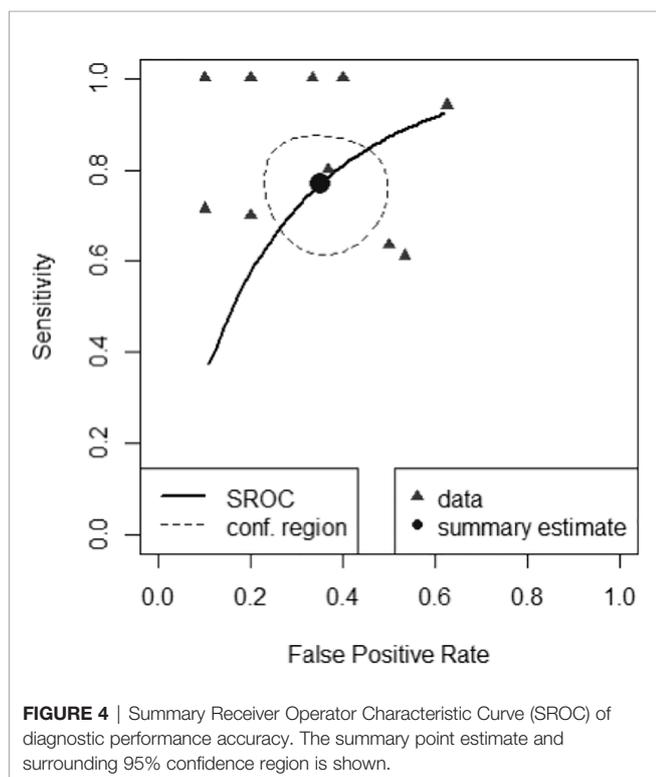
**FIGURE 3** | Forest plots showing sensitivity and specificity.

use of corticosteroids was rarely considered despite being a confounding factor in relation to the imaging follow-up reference standard. If unaccounted for, an increase in corticosteroid dose may cause false negative treatment response. Some authors provided a statement within their methodology that they followed RANO guidelines (4) which if followed meticulously would surmount some of these clinoradiological limitations, such as the use of

corticosteroids which is integrated with the imaging assessment. One limitation in using the RANO guidelines, however, is that in some scenarios the contrast-enhancing lesions due to PTRE continue to enlarge over a short interval, confounding assessment by falsely confirming true progression if continued enlargement is seen at a second short interval scan; RANO guidelines do not account for this upslope effect (26, 27).

Second, patient selection was problematic and is associated with confounding. For example, patients receiving second-line drug therapy should have been excluded as response assessment may be altered. It is also noteworthy that astrocytoma, IDH-mutant, grade 4 are biologically and prognostically distinct from glioblastomas, IDH-wildtype (55). Variable proportions in individual studies introduces between-study heterogeneity and therefore this is a source of potential confounding when comparing or pooling data. Nonetheless, it is acknowledged that for grade 4 tumors, IDH-mutants have a prevalence an order of magnitude less than IDH-wildtype, likely limiting the impact of such confounding.

Third, hold-out test sets should be used for diagnostic accuracy assessment in ML studies (22) as it is a simple demonstration as to whether the trained model overfits data; nonetheless more than a third of studies did not use either an internal or external hold-out test set. Nonetheless, six studies did use external hold-out tests which might be considered optimal practice for determining generalizability.



**FIGURE 4** | Summary Receiver Operator Characteristic Curve (SROC) of diagnostic performance accuracy. The summary point estimate and surrounding 95% confidence region is shown.

#### 4.2.2 Review Process

Imaging reference standards, especially RANO trial guidelines (4) and later iterations (25), are rarely applied correctly and are themselves confounded (56). Because tumors have a variety of shapes, may have an outline that is difficult to delineate, and may be located only within the cavity rim, it can be challenging to perform seemingly simple size measurements (11). For example, large, cyst-

like glioblastomas may be “non-measurable” unless a solid nodular component of the rim fulfils the “measurable” criteria.

As well as the scenario described above highlighting the upslope effect of PTRE (26, 27), another limitation of RANO is a failure to acknowledge that pseudoprogression appears over a 6-month period rather than a 3-month period (although it is accepted that even a 6 month cut-off is arbitrary) (26). Follow-up imaging of adequate duration is therefore required in study design. This leads to a further limitation of this or other systematic reviews – it is extremely difficult to design studies with enough nuance to be at low risk of bias in regards to the reference standard.

Another limitation of this systematic review is that pathology at re-operation, where used as a reference standard, is typically not an entirely reliable reference standard for two reasons (57). First, there is the potential for biopsy sampling bias because the entire enhancing tissue may represent an admixture of PTRE and tumor (58). Second, there is a lack of pathological standardization causing a variety of inter-observer diagnostic interpretations given the background of extensive post-therapy related changes (59). Nonetheless, in the absence of more reliable available reference standards at re-operation, it was pragmatically included as an acceptable reference standard. Additionally, according to many authors, it is closer to being a more accurate reference standard compared to follow-up imaging.

Publication bias may also have affected the range of diagnostic accuracy of the monitoring biomarkers included in this systematic review and meta-analysis. Related to this, the exclusion of pre-prints and non-peer reviewed material may exacerbate publication bias. In particular, given that some in the data science community may not submit their work in peer-reviewed journals as peer review is relatively slow compared to the speed at which data science develops, it is plausible that publication bias relates to the make-up of the researcher team. For example, more clinically-orientated teams may be more inclined to publish in a peer reviewed journal compared to more data science-orientated teams.

### 4.3 Explanation of the Results in the Context of Other Published Evidence

After treatment, “monitoring biomarkers” are measured serially to detect change in the extent of tumor infiltration or to provide evidence of response to treatment (2). In nearly all glioblastomas the integrity of the blood brain barrier is disrupted and MRI is used to take advantage of this. Following intravenous administration of gadolinium-based contrast agents, the hydrophilic contrast molecules diffuse from the vessel lumen and accumulate in the extravascular extracellular space, manifesting on  $T_1$ -weighted sequences as contrast-enhancing hyperintense regions (60). Subsequently, MRI has been incorporated into recommendations for determining response to treatment in trials (4). In these recommendations, treatment response assessment is based on simple linear metrics of contrast-enhancing tumor, specifically, the product of maximal perpendicular cross-sectional dimensions in “measurable” lesions defined as  $> 10$  mm in all perpendicular dimensions. The recommendations are based on expert opinion

informed by observational studies and derived from the biologically plausible assumption that an increase in the size of a tumor identifies disease progression, potentially resulting in a lead time improvement for therapeutic intervention before the tumor becomes clinically apparent (61). The rationale is that there may be advantages in altering management early on before the onset of irreversible disability or the tumor extent precludes intervention. Justification for enhancement as a proxy for tumor has been inferred from data showing that the size of the enhancing region and extent of resection of the enhancing region are “prognostic biomarkers” (2) at both initial presentation and confirmed recurrence (62–64).

The trial assessment recommendations, incorporated in a less stringent form during routine clinical assessment (65), allow for an early change in treatment strategy (9). However, there are important challenges using conventional structural MRI protocols.

First, treatment response assessment typically is made in a retrospective manner as confirmatory imaging is required to demonstrate a sustained increase or a sustained decrease in enhancing volume. This leads to a delay in diagnosis.

Second, contrast enhancement is biologically non-specific, which can result in false negative, false positive, and indeterminate outcomes, especially in regards to the post-treatment related pseudophenomena observed in glioblastoma patients (61). Pseudoprogression is an early post-treatment related effect characteristically appearing within 6 months of glioblastoma patients completing radiotherapy and concomitant temozolomide, whereas pseudoresponse (not examined in this systematic review) appears after patients have been treated with anti-angiogenic agents such as bevacizumab. False-negative treatment response and false-positive progression appear as a decrease or an increase in the volume of MRI contrast enhancement, respectively. Delayed post-treatment related effects caused by radiation necrosis similarly appear as an increase in volume of MRI contrast enhancement, again potentially causing false-positive progression. A different scenario where contrast enhancement is biologically non-specific includes post-operative peritumoral parenchymal enhancement after operative “tissue handling”; or after operative infarction.

Conventional structural MRI protocols are therefore limited and contemporaneous, accurate and reliable monitoring biomarkers are required for glioblastoma treatment response assessment. Three potential solutions are highlighted here:

First, an emerging alternative approach is to harness the potential value of circulating biomarkers (including circulating tumor cells, exosomes, and microRNAs) to monitor disease progression in glioma patients (66). However, as with any potential monitoring blood or cerebral spinal fluid biomarker, potential use requires further evaluation and validation in large scale prospective studies before implementation into standard clinical practice can be envisaged.

Second, another promising approach is to use advanced imaging techniques (67). The last three decades have seen considerable technical developments in MRI (for example, those related to perfusion, permeability and diffusion), 1H-MR spectroscopic imaging, and positron emission tomography (for example using radiolabelled amino acids). A meta-analysis of 28 perfusion and

permeability imaging studies showed that the pooled sensitivities and specificities of each study's best performing parameter were 90% and 88% (95% confidence interval (CI), 0.85 - 0.94; 0.83 - 0.92) and 89% and 85% (95% CI, 0.78 - 0.96; 0.77 - 0.91) for dynamic susceptibility-weighted (DSC) and dynamic contrast-enhanced (DCE) MRI, respectively (68). Clinical translation is far from ubiquitous (65) reflecting that further investigation and consensus standardization is required before implementing any particular widespread quantitative strategy (68). Indeed, advanced imaging is not yet recommended for determining treatment response in trials (4), and there is a lack of evidence that using advanced MRI techniques leads to a reduction in morbidity or mortality (61). However, compared to ML where accuracy-driven performance metrics have resulted in increasingly opaque models, particularly when using structural images, the underlying biological processes relating to advanced imaging appear to be well understood whilst also demonstrating high performance accuracy.

A third approach is to use ML, whether applied to conventional structural MRI, advanced imaging techniques or a combination of both imaging and non-imaging features. Indeed, an advantage of machine learning applied to MRI is that wide data can be handled relatively easily (11) which might allow the wide spectrum of signatures from multiparametric advanced MRI to be captured together to improve performance accuracy. However, a disadvantage when compared to a single modality approach is that combinations of outputs from individual modalities that are without frameworks for technical and clinical use, might compound inter-center variability and reduce generalizability considerably. The advantages and disadvantages of using ML-based monitoring biomarkers for glioblastoma treatment response assessment have been described recently (summarised in **Table 3**) (61). However, a number of factors demonstrate that only limited conclusions on performance can be drawn from recent studies in our systematic review. These include the high risk of bias and

concerns about applicability in study designs, the small number of patients analysed in ML studies, and the low level of evidence of the monitoring biomarker studies given their retrospective nature.

Nonetheless, overall there appears to be good diagnostic performance of ML models using MRI features to differentiate between progressive disease and mimics. For now, if ML models are to be used they may be best confined to the centers where the data was obtained from, badged as research tools and undergo further improvement.

Concordant with a previous review of studies published up to Sept 2018 (11), the diagnostic performance of ML using implicit features did not appear to be superior to ML using explicit features. However, the small number of studies precluded meaningful quantitative comparison.

#### 4.4 Implications for Clinical Practice and Future Research

The results demonstrate that glioblastoma treatment response monitoring biomarkers using ML are promising but are still at the early development stage and are not yet ready to be integrated into clinical practice. All studies would benefit from the improvements in methodology described above. Methodological profiles or standards might be developed through consortiums such as the European Cooperation in Science and Technology (COST) Glioma MR Imaging 2.0 (GliMR) (67) initiative or the ReSPOND Consortium (76). Determining an accurate reference standard for treatment response is challenging and performing prospective studies capturing contemporaneous detailed information on steroids and second line treatments is likely to mitigate the effects of confounding. Additionally, multiple image-localized biopsies at recurrence may lessen sampling bias due to PTRE and tumor admixture.

In future studies, it would be beneficial to perform analytical validation using external hold-out tests as epitomized by several studies in the current review. Using larger datasets which include a

**TABLE 3** | Advantages and disadvantages of using ML-based monitoring biomarkers for glioblastoma treatment response assessment (61).

Advantages	Disadvantages
<p>Using ML requires less formal statistical training given the huge developments in software (69), and the programming expertise for researchers has now been transformatively reduced, enabled by standardized implementations of open source software (70, 71). Wide data can be handled relatively easily (11) and ML can be applied to conventional structural MRI, advanced imaging techniques or a combination of both imaging and non-imaging features. ML models have the ability to determine implicitly any complex nonlinear relationship between independent and dependent variables (69), and have the ability to determine all possible interactions between predictor variables (73).</p>	<p>The clinical context may not be represented with a decreased ability to perform holistic evaluations of patients, with loss of valuable and irreducible aspects of the human experience such as psychological, relational, social, and organizational issues (72). Linking the empirical data to a categorical analysis can neglect an intrinsic ambiguity in the observed phenomena (72), which might adversely affect the intended performance (69). Overreliance on the capabilities of automation can lead to the related phenomenon of radiologist deskilling (74). Algorithms may be unreliable due to several technical constraints: domain adaptation is currently limited, and more solutions are required to help algorithms extrapolate well to new centers. Ultimately models may require calibration or retraining. Robustness to unintended data, such as artifacts, is also a technical constraint that needs to be overcome. Finally, the presence of more than one pathology (e.g., stroke or abscess associated with a tumor following treatment) can also confound algorithms as these cases are scarce and often unlabeled. Accuracy-driven performance metrics have led to a trend towards increasingly opaque models (73), although recent developments in interpretability and explainability may help to mitigate this to some extent (75).</p>

wider range of tumors and mimics as well as parameters from different sequences, manufacturers and coils, and thereby reduce overfitting, would also improve future studies. Multidisciplinary efforts and multicenter collaborations are therefore necessary (61). However, datasets will always be relatively small in neuro-oncological imaging even if distributed machine learning approaches such as federated learning, where the model comes to the data rather than the data comes to the model, overcome data sharing regulatory bottlenecks (61). Therefore, strategies to improve ML performance using small datasets, some of which are at the research stage, should be exploited further. Strategies include data augmentation (generate more varied image examples, within a single classification task) and the related process of meta-augmentation (generate more varied tasks, for a single example) (77) as well as transfer learning and the overlapping process of one- or few-shot learning (78). Transfer learning aims to learn representations from one domain (does not need to consist of brain tumors) and transfer the learned features to a closely related target domain (glioblastoma). Few-shot learning allows classifiers to be built from very small labelled training sets. Another research direction could be reducing the demand for image labelling. This field is known as self-supervised learning (79). Finally, an entirely different approach to counter the challenges of small datasets is to use synthetic data, for example using generative adversarial networks (80).

Predictions can also be made more informative through the modelling of prediction uncertainty including the generation of algorithms that would “know when they don’t know” what to predict (11).

Further downstream challenges for clinical adoption will be the completion of clinical validation (2) as well as the deployment of the clinical decision support (CDS) software to clinical settings. Clinical validation consists of evaluating the CDS software containing the locked machine learning model in a clinical trial thereby producing high level evidence (12). The CDS software deployment brings both technical and non-technical challenges. In terms of technical challenges, the CDS software must be easily integrated into the radiologist’s workflow (electronic health record system and picture archiving and communication system) and preferably deliver a fully automated process that analyzes images in real time and provides a quantitative and probabilistic report. Currently there has been little translation of CDS software into radiological departments however there are open source deployment solutions (71, 81).

Non-technical challenges relate to patient data safety and privacy issues; ethical, legal and financial barriers to developing and distributing tools that may impact a patient’s treatment course; medical device regulation; usability evaluation; clinical acceptance and medical education around the implementation of

CDS software (14, 82). Medical education includes articulating the CDS software limitations to ensure there is judicious patient and imaging selection reflecting the cohort used for validation of the model (11).

## 5 CONCLUSION

A range of ML-based solutions primed as glioblastoma treatment response monitoring biomarkers may soon be ready for clinical adoption. To ensure clinical adoption, it would be beneficial during the development and validation of ML models that studies include large, well-annotated datasets where there has been meticulous consideration of the potential for confounding.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding author.

## AUTHOR CONTRIBUTIONS

TB: experimental design and implementation, analysis and interpretation of the data, performance accuracy statistical analysis, writing of draft manuscript, approval of the final version. AC, AR, AAB, CD, HS, AL, AM, BA, NM, JL, FV, KA, SO, MM: the implementation, analysis, and interpretation of the data and approval of the final version. MG: implementation, analysis and interpretation of the data, meta-analysis statistical analysis, writing of draft manuscript, approval of the final version. All authors contributed to the article and approved the submitted version.

## FUNDING

This research was supported by the Wellcome/EPSRC Centre for Medical Engineering (WT 203148/Z/16/Z) (TB, MG, AC, MM, SO) which includes open access fees, The Royal College of Radiologists (TB) and King’s College Hospital Research and Innovation (TB).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fonc.2022.799662/full#supplementary-material>

## REFERENCES

1. Stupp R, Mason WP, van den Bent MJ, Weller M, Fisher B, Taphoorn MJB, et al. Radiotherapy Plus Concomitant and Adjuvant Temozolomide for Glioblastoma. *N Engl J Med* (2005) 352:987–96. doi: 10.1056/NEJMoa043330
2. FDA-NIH Biomarker Working Group. *BEST (Biomarkers, EndpointS, and Other Tools) Resource. 1st edn*. Silver Spring, MD: Food and Drug Administration (US), co-published by Bethesda, MD: National Institutes of Health US (2016). Available at: <https://www.ncbi.nlm.nih.gov/books/NBK326791>.
3. MacDonald D, Cascino TL, Schold SC, Cairncross JG. Response Criteria for Phase II Studies of Supratentorial Malignant Glioma. *J Clin Oncol* (2010) 8:1277–80. doi: 10.1200/JCO.1990.8.7.1277
4. Wen PY, Macdonald DR, Reardon DA, Cloughesy TF, Sorensen AG, Galanis E, et al. Updated Response Assessment Criteria for High-Grade Gliomas: Response Assessment in Neuro-Oncology Working Group. *J Clin Oncol* (2010) 28:1963–72. doi: 10.1200/JCO.2009.26.3541

5. Booth TC, Tang Y, Waldman AD, Quigley A-M, Lewis D, Soloviev D, et al. Neuro-Oncology Single-Photon Emission CT: A Current Overview. *Neurographics* (2011) 01:108–20. doi: 10.3174/ng.3110014
6. Chamberlain MC, Glantz MJ, Chalmers L, Van Horn A, Sloan AE. Early Necrosis Following Concurrent Temodar and Radiotherapy in Patients With Glioblastoma. *J Neurooncol* (2007) 82:81–3. doi: 10.1007/s11060-006-9241-y
7. Brandsma D, Stalpers L, Taal W, Sminia P, van den Bent MJ. Clinical Features, Mechanisms, and Management of Pseudoprogression in Malignant Gliomas. *Lancet Oncol* (2008) 9(5):453–61. doi: 10.1016/S1470-2045(08)70125-6
8. Radbruch A, Fladt J, Kickingeder P, Wiestler B, Nowosielska M, Baumer P, et al. Pseudoprogression in Patients With Glioblastoma: Clinical Relevance Despite Low Incidence. *Neuro Oncol* (2015) 17(1):151–9. doi: 10.1093/neuonc/nou129
9. Dhermain FG, Hau P, Lanfermann H, Jacobs AH, van den Bent MJ. Advanced MRI and PET Imaging for Assessment of Treatment Response in Patients With Gliomas. *Lancet Neurol* (2010) 9(9):906–20. doi: 10.1016/S1474-4422(10)70181-2
10. Verma N, Cowperthwaite MC, Burnett MG, Markey MK. Differentiating Tumor Recurrence From Treatment Necrosis: A Review of Neuro-Oncologic Imaging Strategies. *Neuro Oncol* (2013) 15(5):515–34. doi: 10.1093/neuonc/nos307
11. Booth TC, Williams M, Luis A, Cardoso J, Ashkan K, Shuaib H. Machine Learning and Glioma Imaging Biomarkers. *Clin Radiol* (2020) 75:20–32. doi: 10.1016/j.crad.2019.07.001
12. Howick J, Chalmer I, Glasziou P, Greenhalgh T, Heneghan C, Liberati A, et al. *Oxford Centre for Evidence-Based Medicine The Oxford 2011 Levels of Evidence* (2016). Oxford. Available at: <http://www.cebm.net/index.aspx?o/45653> (Accessed 1 August, 2018).
13. Lui YW, Chang PD, Zaharchuk G, Barboriak DP, Flanders AE, Wintermark M, et al. Artificial Intelligence in Neuroradiology: Current Status and Future Directions. *AJNR Am J Neuroradiol* (2020) 41(8):E52–9. doi: 10.3174/ajnr.A6681
14. Jin W, Fatehi M, Abhishek K, Mallya M, Toyota B, Hamarneh G. Artificial Intelligence in Glioma Imaging: Challenges and Advances. *J Neural Eng* (2020) 17(2):021002. doi: 10.1088/1741-2552/ab8131
15. Dean J, Patterson D, Young C. A New Golden Age in Computer Architecture: Empowering the Machine-Learning Revolution. *IEEE Micro* (2018) 38(2):21–9. doi: 10.1109/MM.2018.112130030
16. Booth TC, Akpinar B, Roman A, Shuaib H, Luis A, Chelliah A, et al. Machine Learning and Glioblastoma: Treatment Response Monitoring Biomarkers in 2021. In: SM Kia, H Mohy-ud-Din, A Abdulkadir, C Bass, M Habes, JM Rondina, et al, editors. *Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-Oncology. MLCN 2020, RNO-AI 2020. Lecture Notes in Computer Science*, vol. 12449. Heidelberg: Springer (2020). p. 212–8.
17. McInnes MDF, Moher D, Thoms BD, McGrath TA, Bossuyt PM, the PRISMA-DTA Group. Preferred Reporting Items for a Systematic Review and Meta-Analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement. *JAMA* (2018) 319:388–96. doi: 10.1001/jama.2017.19163
18. Bossuyt PM, Leeflang MM. Developing Criteria for Including Studies. In: PM Bossuyt, editor. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 0.4*. London: The Cochrane Collaboration (2008). p. 1–7.
19. de Vet HCW, Eisinga A, Riphagen II, Aertgeerts B, Pewsner D. Searching for Studies. In: PM Bossuyt, editor. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 0.4*. London: The Cochrane Collaboration (2008). p. 1–60.
20. Reitsma JB, Rutjes AWS, Whiting P, Vlassov VV, Leeflang MMG, Deeks JJ. Assessing Methodological Quality. In: JJ Deeks, PM Bossuyt, C Gatsonis, editors. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0.0*. London: The Cochrane Collaboration (2009). p. 1–28.
21. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. *Ann Intern Med* (2011) 155(8):529–36. doi: 10.7326/0003-4819-155-8-201110180-00009
22. Mongan J, Moy L, Kahn CE. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. *Radiol Artif Intell* (2020) 2(2):e200029. doi: 10.1148/ryai.2020200029
23. Booth TC, Ashkan K, Brazil L, Jäger R, Waldman AD. Re: “Tumour Progression or Pseudoprogression? A Review of Post-Treatment Radiological Appearances of Glioblastoma”. *Clin Radiol* (2016) 71(5):495–6. doi: 10.1016/j.crad.2016.01.014
24. Booth TC, Waldman AD, Jefferies S, Jäger R. Comment on “The Role of Imaging in the Management of Progressive Glioblastoma. A Systematic Review and Evidence-Based Clinical Practice Guideline”. *J Neurooncol* (2015) 121(2):423–4. doi: 10.1007/s11060-014-1649-1
25. Ellingson BM, Wen PY, Cloughesy TF. Modified Criteria for Radiographic Response Assessment in Glioblastoma Clinical Trials. *Neurotherapeutics* (2017) 14:307–20. doi: 10.1007/s13311-016-0507-6
26. Booth TC, Larkin TJ, Yuan Y, Dawson SN, Scoffings D, Canuto HC, et al. Analysis of Heterogeneity in T2-Weighted MR Images can Differentiate Pseudoprogression From Progression in Glioblastoma. *PLoS One* (2017) 12(5):e0176528. doi: 10.1371/journal.pone.0176528
27. Gahramanov S, Muldoon LL, Varallyay CG, Li X, Kraemer DF, Fu R, et al. Pseudoprogression of Glioblastoma After Chemo- and Radiation Therapy: Diagnosis by Using Dynamic Susceptibility-Weighted Contrast-Enhanced Perfusion MR Imaging With Ferumoxytol Versus Gadoteridol and Correlation With Survival. *Radiology* (2013) 266(3):842–52. doi: 10.1148/radiol.12111472
28. Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate Analysis of Sensitivity and Specificity Produces Informative Summary Measures in Diagnostic Reviews. *J Clin Epidemiol* (2005) 58:982–90. doi: 10.1016/j.jclinepi.2005.02.022
29. Zhou X-H, Obuchowski NA, McClish DK. *Statistical Methods in Diagnostic Medicine. 2nd ed.* Chichester: John Wiley & Sons (2011) p. 55.
30. Harbord RM, Whiting P, Metandi: Meta-Analysis of Diagnostic Accuracy Using Hierarchical Logistic Regression. *Stata J* (2009) 9(2):211–29. doi: 10.1177/1536867X0900900203
31. Schwarzer G, Carpenter JR, Rücker G. *Meta-Analysis With R*. Heidelberg: Springer (2015). 252 p.
32. Zwinderman A, Bossuyt P. We Should Not Pool Diagnostic Likelihood Ratios in Systematic Reviews. *Stat Med* (2008) 27:687–97. doi: 10.1002/sim.2992
33. Doebler P. *Mada: Meta-Analysis of Diagnostic Accuracy. R Package Version 0.5.10* (2020). Available at: <http://www.cran.R-project.org/packages/mada> (Accessed Apr 21, 2021).
34. Kim JY, Park JE, Jo Y, Shim WH, Nam SJ, Kim JH, et al. Incorporating Diffusion- and Perfusion-Weighted MRI Into a Radiomics Model Improves Diagnostic Performance for Pseudoprogression in Glioblastoma Patients. *Neuro Oncol* (2019) 21:404–14. doi: 10.1093/neuonc/noy133
35. Kim JY, Yoon MJ, Park JE, Choi EJ, Lee J, Kim HS. Radiomics in Peritumoral non-Enhancing Regions: Fractional Anisotropy and Cerebral Blood Volume Improve Prediction of Local Progression and Overall Survival in Patients With Glioblastoma. *Neuroradiology* (2019) 61:1261–72. doi: 10.1007/s00234-019-02255-4
36. Bacchi S, Zerner T, Dongas J, Asahina AT, Abou-Hamden A, Otto S, et al. Deep Learning in the Detection of High-Grade Glioma Recurrence Using Multiple MRI Sequences: A Pilot Study. *J Clin Neurosci* (2019) 70:11–3. doi: 10.1016/j.jocn.2019.10.003
37. Elshafee N, Kotrotsou A, Hassan A, Elshafei N, Hassan I, Ahmed S, et al. Multicenter Study Demonstrates Radiomic Features Derived From Magnetic Resonance Perfusion Images Identify Pseudoprogression in Glioblastoma. *Nat Commun* (2019) 10:3170. doi: 10.1038/s41467-019-11007-0
38. Verma G, Chawla S, Mohan S, Wang S, Nasrallah M, Sherif S, et al. Three-Dimensional Echo Planar Spectroscopic Imaging for Differentiation of True Progression From Pseudoprogression in Patients With Glioblastoma. *NMR BioMed* (2019) 32:4042. doi: 10.1002/nbm.4042
39. Ismail M, Hill V, Statsevych V, Huang R, Prasanna P, Correa R, et al. Shape Features of the Lesion Habitat to Differentiate Brain Tumor Progression From Pseudoprogression on Routine Multiparametric MRI: A Multisite Study. *AJNR Am J Neuroradiol* (2018) 39:2187–93. doi: 10.3174/ajnr.A5858
40. Bani-Sadr A, Eker OF, Berner LP, Ameli R, Hermier M, Barritault M, et al. Conventional MRI Radiomics in Patients With Suspected Early- or Pseudo-Progression. *Neurooncol Adv* (2019) 1:1–9. doi: 10.1093/oaajnl/vdz019
41. Gao XY, Wang YD, Wu SM, Rui WT, Ma DN, Duan Y, et al. Differentiation of Treatment-Related Effects From Glioma Recurrence Using Machine Learning Classifiers Based Upon Pre-And Post-Contrast T1WI and T2 FLAIR

- Subtraction Features: A Two-Center Study. *Cancer Manag Res* (2020) 12:3191–201. doi: 10.2147/CMARS.S244262
42. Jang BS, Jeon SH, Kim IH, Kim IA. Prediction of Pseudoprogression Versus Progression Using Machine Learning Algorithm in Glioblastoma. *Sci Rep* (2019) 8:12516. doi: 10.1038/s41598-018-31007-2
  43. Li M, Tang H, Chan MD, Zhou X, Qian X. DC-AL GAN: Pseudoprogression and True Tumor Progression of Glioblastoma Multiform Image Classification Based on DCGAN and AlexNet. *Med Phys* (2020) 47:1139–50. doi: 10.1002/mp.14003
  44. Akbari H, Rathore S, Bakas S, Nasrallah MP, Shukla G, Mamourian E, et al. Histopathology - Validated Machine Learning Radiographic Biomarker for Noninvasive Discrimination Between True Progression and Pseudo-Progression in Glioblastoma. *Cancer* (2020) 126(11):2625–36. doi: 10.1002/ncr.32790
  45. Li X, Xu G, Cao Q, Zou W, Xu Y, Cong P. Identification of Glioma Pseudoprogression Based on Gabor Dictionary and Sparse Representation Model. *NeuroQuantology* (2018) 16:43–51. doi: 10.14704/nq.2018.16.1.1178
  46. Manning P, Daghighi S, Rajaratnam MK, Parthiban S, Bahrami N, Dale AM, et al. Differentiation of Progressive Disease From Pseudoprogression Using 3d PCASL and DSC Perfusion MRI in Patients With Glioblastoma. *J Neurooncol* (2020) 147(3):681–90. doi: 10.1007/s11060-020-03475-y
  47. Park JE, Ham S, Kim HS, Park SY, Yun J, Lee H, et al. Diffusion and Perfusion MRI Radiomics Obtained From Deep Learning Segmentation Provides Reproducible and Comparable Diagnostic Model to Human in Post-Treatment Glioblastoma. *Eur Radiol* (2021) 31(5):3127–37. doi: 10.1007/s00330-020-07414-3
  48. Lee J, Wang N, Turk S, Mohammed S, Lobo R, Kim J, et al. Discriminating Pseudoprogression and True Progression in Diffuse Infiltrating Glioma Using Multi-Parametric MRI Data Through Deep Learning. *Sci Rep* (2020) 10(1):20331. doi: 10.1038/s41598-020-77389-0
  49. Kebir S, Schmidt T, Weber M, Lazaridis L, Galldiks N, Langen KJ, et al. A Preliminary Study on Machine Learning-Based Evaluation of Static and Dynamic FET-PET for the Detection of Pseudoprogression in Patients With IDH-Wildtype Glioblastoma. *Cancers* (2020) 12(11):3080. doi: 10.3390/cancers12113080
  50. Cluceru J, Nelson SJ, Wen Q, Phillips JJ, Shai A, Molinaro AM, et al. Recurrent Tumor and Treatment-Induced Effects Have Different MR Signatures in Contrast Enhancing and non-Enhancing Lesions of High-Grade Gliomas. *Neuro Oncol* (2020) 22(10):1516–26. doi: 10.1093/neuonc/noaa094
  51. Jang BS, Park AJ, Jeon SH, Kim IH, Lim DH, Park SH, et al. Machine Learning Model to Predict Pseudoprogression Versus Progression in Glioblastoma Using MRI: A Multi-Institutional Study (KROG 18-07). *Cancers* (2020) 12(9):2706. doi: 10.3390/cancers12092706
  52. Wang S, O'Rourke DM, Chawla S, Verma G, Nasrallah MP, Morrissette JJD, et al. Multiparametric Magnetic Resonance Imaging in the Assessment of Anti-EGFRvIII Chimeric Antigen Receptor T Cell Therapy in Patients With Recurrent Glioblastoma. *Br J Cancer* (2019) 120:54–6. doi: 10.1038/s41416-018-0342-0
  53. Yang K, Jung SW, Shin H, Lin DH, Lee JI, Kong DS, et al. Cancer Genetic Markers According to Radiotherapeutic Response in Patients With Primary Glioblastoma - Radiogenomic Approach for Precision Medicine. *Radiother Oncol* (2019) 131:66–74. doi: 10.1016/j.radonc.2018.11.025
  54. Lundemann M, Munck af Rosenschöld P, Muhic A, Larsen VB, Poulsen HS, Engelholm SA, et al. Feasibility of Multi-Parametric PET and MRI for Prediction of Tumour Recurrence in Patients With Glioblastoma. *Eur J Nucl Med Mol Imaging* (2019) 46:603–13. doi: 10.1007/s00259-018-4180-3
  55. Louis DN, Perry A, Wesseling P, Brat DJ, Cree IA, Figarella-Branger D, et al. The 2021 WHO Classification of Tumors of the Central Nervous System: A Summary. *Neuro Oncol* (2021) 23(8):1231–51. doi: 10.1093/neuonc/noab106
  56. Buwanabala J, Mirchandani A, Booth TC. (2019). The (Mis)Use of Imaging Criteria in the Assessment of Glioblastoma Treatment Response. In: *Proceedings of the 57th American Society of Neuroradiology*, Boston, USA, 2019 May 18-23. p. 1. Oak Brook.
  57. Holdhoff M, Ye X, Piotrowski AF, Strowd RE, Seopaul S, Lu Y, et al. The Consistency of Neuropathological Diagnoses in Patients Undergoing Surgery for Suspected Recurrence of Glioblastoma. *J Neurooncol* (2019) 141:347–54. doi: 10.1007/s11060-018-03037-3
  58. Müller DMJ, Robe PAJT, Eijgelhaar RS, Witte MG, Visser M, de Munck JC, et al. Comparing Glioblastoma Surgery Decisions Between Teams Using Brain Maps of Tumor Locations, Biopsies, and Resections. *JCO Clin Cancer Inform* (2019) 3:1–12. doi: 10.1200/CCI.18.00089
  59. Haider AS, van den Bent M, Wen PY, Vogelbaum MA, Chang S, Canoll PD, et al. Toward a Standard Pathological and Molecular Characterization of Recurrent Glioma in Adults: A Response Assessment in Neuro-Oncology Effort. *Neuro Oncol* (2020) 22(4):450–6. doi: 10.1093/neuonc/noz233
  60. Cao Y, Sundgren PC, Tsien CI, Chenevert TT, Junck L. Physiologic and Metabolic Magnetic Resonance Imaging in Gliomas. *J Clin Oncol* (2006) 24(8):1228–35. doi: 10.1200/JCO.2005.04.7233
  61. Booth TC, Thompson G, Bulbeck H, Boele F, Buckley C, Cardoso J, et al. A Position Statement on the Utility of Interval Imaging in Standard of Care Brain Tumour Management: Defining the Evidence Gap and Opportunities for Future Research. *Front Oncol* (2021) 11:620070. doi: 10.3389/fonc.2021.620070
  62. Ellingson BM, Harris RJ, Woodworth DC, Leu K, Zaw O, Mason WP, et al. Baseline Pretreatment Contrast Enhancing Tumor Volume Including Central Necrosis Is a Prognostic Factor in Recurrent Glioblastoma: Evidence From Single- and Multicenter Trials. *Neuro Oncol* (2017) 19:89–98. doi: 10.1093/neuonc/now187
  63. Lacroix M, Abi-Said D, Fourney DR, Gokaslan ZL, Shi W, DeMonte F, et al. A Multivariate Analysis of 416 Patients With Glioblastoma Multiforme: Prognosis, Extent of Resection, and Survival. *J Neurosurg* (2001) 95:190–8. doi: 10.3171/jns.2001.95.2.0190
  64. Sanai N, Polley M-Y, McDermott MW, Parsa AT, Berger MS. An Extent of Resection Threshold for Newly Diagnosed Glioblastomas. *J Neurosurg* (2011) 115:3–8. doi: 10.3171/2011.2.JNS10998
  65. Booth TC, Luis A, Brazil L, Thompson G, Daniel RA, Shuaib H, et al. Glioblastoma Post-Operative Imaging in Neuro-Oncology: Current UK Practice (GIN CUP Study). *Eur Radiol* (2021) 31(5):2933–43. doi: 10.1007/s00330-020-07387-3
  66. Müller Bark J, Kulasinghe A, Chua B, BW D, Punyadeera C. Circulating Biomarkers in Patients With Glioblastoma. *Br J Cancer* (2020) 122(3):295–305. doi: 10.1038/s41416-019-0603-6
  67. Clement P, Booth T, Borovechi F, Emblem KE, Figueiredo P, Hirschler L, et al. GliMR: Cross-Border Collaborations to Promote Advanced MRI Biomarkers for Glioma. *J Med Biol Eng* (2021) 41(2):115–25. doi: 10.1007/s40846-020-00582-z
  68. Patel P, Baradaran H, Delgado D, Askin G, Christos P, Tsiouris J, et al. MR Perfusion-Weighted Imaging in the Evaluation of High-Grade Gliomas After Treatment: A Systematic Review and Meta-Analysis. *Neuro-Oncology* (2017) 19:118–27. doi: 10.1093/neuonc/now148
  69. White H. Learning in Artificial Neural Networks: A Statistical Perspective. *Neural Comput* (1989) 1(4):425–64. doi: 10.1162/neco.1989.1.4.425
  70. Ratib O, Rosset A, Heuberger J. Open Source Software and Social Networks: Disruptive Alternatives for Medical Imaging. *Eur J Radiol* (2011) 78(2):259–65. doi: 10.1016/j.ejrad.2010.05.004
  71. Medical Open Network for AI (MONAI). Available at: [https://docs.monai.io/projects/monai-deploy-app-sdk/en/latest/release\\_notes/v0.2.0.html](https://docs.monai.io/projects/monai-deploy-app-sdk/en/latest/release_notes/v0.2.0.html) (Accessed 30 Dec 2021).
  72. Cabitza F, Rasoini R, Genisi GF. Unintended Consequences of Machine Learning in Medicine. *JAMA* (2017) 318(6):517–8. doi: 10.1001/jama.2017.7797
  73. Tu JV. Advantages and Disadvantages of Using Artificial Neural Networks Versus Logistic Regression for Predicting Medical Outcomes. *J Clin Epidemiol* (1996) 49(11):1225–31. doi: 10.1016/s0895-4356(96)00002-9
  74. Hoff T. Deskillling and Adaptation Among Primary Care Physicians Using Two Work Innovations. *Health Care Manage Rev* (2011) 36(4):338–48. doi: 10.1097/HMR.0b013e31821826a1
  75. Reyes M, Meier R, Pereira S, Silva CA, Dahlweid F-M, von Tengg-Kobligh K, et al. On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities. *Radiol Artif Intell* (2020) 2(3):e190043. doi: 10.1148/ryai.2020190043
  76. Davatzikos C, Barnholtz-Sloan JS, Bakas S, Colen R, Mahajan A, Quintero CB, et al. AI-Based Prognostic Imaging Biomarkers for Precision Neurooncology: The ReSPOND Consortium. *Neuro-Oncology* (2020) Pii:noaa045. doi: 10.1093/neuonc/noaa045
  77. Rajendran J, Irpan A, Jang E. *Meta-Learning Requires Meta-Augmentation. Advances in Neural Information Processing Systems* 33 (2020). Available at: <https://papers.nips.cc/paper/2020/hash/3e5190eeb51ebe6c5bb45ee8950c548-Abstract.html>.

78. Paul A, Tang YX, Summers RM. Fast Few-Shot Transfer Learning for Disease Identification From Chest X-Ray Images Using Autoencoder Ensemble. *Med Imaging* (2020). doi: 10.1117/12.2549060
79. Jing L, Tian Y. Self-Supervised Visual Feature Learning With Deep Neural Networks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. (2021) 43:4037–58. doi: 10.1109/TPAMI.2020.2992393
80. Shin HC, Tenenholtz NA, Rogers JK, Schwarz CG, Senjem ML, Gunter JL. Medical Image Synthesis for Data Augmentation and Anonymization Using Generative Adversarial Networks. In: A Gooya, O Goksel, I Oguz, N Burgos, editors. *Simulation and Synthesis in Medical Imaging. SASHIMI 2018. Lecture Notes in Computer Science*, vol. 11037. Heidelberg: Springer (2018). p. 1–11.
81. Davatzikos C, Rathore S, Bakas S, Pati S, Bergman M, Kalarot R, et al. Cancer Imaging Phenomics Toolkit: Quantitative Imaging Analytics for Precision Diagnostics and Predictive Modeling of Clinical Outcome. *J Med Imaging (Bellingham)* (2018) 5:011018. doi: 10.1117/1.JMI.5.1.011018
82. Senders JT, Harary M, Stopa BM, Staples P, Broekman MLD, Smith TR, et al. Information-Based Medicine in Glioma Patients: A Clinical Perspective. *Comput Math Methods Med* (2018), 8572058. doi: 10.1155/2018/8572058

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Booth, Grzeda, Chelliah, Roman, Al Busaidi, Dragos, Shuaib, Luis, Mirchandani, Alparslan, Mansoor, Lavrador, Vergani, Ashkan, Modat and Ourselin. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.