

1 **Performance of Handcrafted Radiomics versus Deep Learning for Prognosticating Head and**
2 **Neck Squamous Cell Carcinoma – A Systematic Review with Critical Appraisal of Quantitative**
3 **Imaging Studies**

4 **Varsha Gouthamchand^{1*}, Louise AF Fonseca³, Frank JP Hoebbers¹, Rianne Fijten¹, Andre Dekker^{1,2}, Leonard**
5 **Wee¹ and Hannah Mary Thomas T⁴**

6 ¹ Department of Radiation Oncology (Maastrro), GROW Research Institute for Oncology and Reproduction, Maastricht
7 University Medical Centre+, Maastricht, The Netherlands

8 ² Brightlands Institute for Smart Society, Faculty of Science and Engineering, Maastricht University, Heerlen, The
9 Netherlands

10 ³ Academic Center for General Practice, Department of Public Health and Primary Care, KU Leuven, Leuven, Belgium.

11 ⁴ Quantitative Imaging Research and Artificial Intelligence Lab, Department of Radiation Oncology, Unit II, Christian
12 Medical College, Vellore, Tamil Nadu, India

13 *** Correspondence:**
14 Corresponding author: Varsha Gouthamchand
15 varsha.gouthamchand@maastro.nl

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31 Abstract

32 Head and neck squamous cell carcinoma (HNSCC) presents a complex clinical challenge due to its heterogeneous nature
33 and diverse treatment responses. This systematic review critically appraises the performance of handcrafted radiomics
34 (HC) and deep learning (DL) models in prognosticating outcomes in HNSCC patients treated with (chemo)-radiotherapy.
35 A comprehensive literature search was conducted up to May 2023, identifying 23 eligible studies that met the inclusion
36 criteria of methodological rigor and long-term outcome reporting. The review highlights the methodological variability
37 and performance metrics of HC and DL models in predicting overall survival (OS), loco-regional recurrence (LRR) and
38 distant metastasis (DM). While DL models demonstrated slightly superior performance metrics compared to HC models,
39 the highest methodological quality was observed predominantly in studies using HC radiomics. The findings underscore
40 the necessity for methodological improvements, including pre-registration of protocols and assessment of clinical utility,
41 to enhance the reliability and applicability of radiomic-based prognostic models in clinical practice.

42 **Keywords:** Radiomics, Deep Learning, Head and neck cancer, Prognosis, Systematic Review, Quality Checklist

43 Introduction

44 Head and neck squamous cell carcinomas (HNSCC) comprise a highly heterogeneous subset of neoplastic diseases
45 originating in the mucosal lining of the oral cavity, pharynx, and larynx [1]. According to GLOBOCAN 2022, HNSCC
46 (including cancers of the lip, oral cavity, larynx, oropharynx, hypopharynx, and salivary glands) accounted for an
47 estimated 826,040 new cases and 445,896 deaths, representing about 4.4% of all cancer cases and 4.6% of all cancer
48 deaths globally [2]. The growing global health burden may be due to alcohol and nicotine consumption patterns
49 correlated with urban/rural migration, economic factors influencing changes in dietary patterns) and a wider exposure to
50 oncogenic viruses such as the human papilloma virus (HPV) in the case of oropharyngeal carcinoma.

51 Locally advanced HNSCC are generally treated with a combination of radiotherapy (RT), chemotherapy and/or surgery.
52 Prognoses for 5-year survival range from almost 90% in HPV-positive oropharyngeal carcinoma (OPC) down to 25% in
53 advanced hypopharyngeal carcinoma (HPC). Long-term side-effects of treatment also vary considerably between persons
54 and may include physical appearance changes (mainly due to surgery), xerostomia, dysphagia, odynophagia, fibrosis,
55 fatigue, and ototoxicity (mainly due to cisplatin chemotherapy). Survivorship within certain subtypes of HNSCC has been
56 improving gradually over time, leading to greater attention towards functional preservation after treatment, psychosocial
57 resilience, and health-related quality of life. Newer treatments such as immunotherapy and proton beam therapy are not
58 readily available in all countries, therefore great diligence is required to identify patients that benefit from expensive
59 novel treatments, or else to reduce disutility of care among patients that do not benefit from aggressive treatment.

60 Genetic diversity and complex pathophysiology imply significant intra- and inter-tumoral heterogeneity in HNSCC (add
61 REF here). Routine oncological imaging with computed tomography (CT), positron emission tomography (PET) and
62 magnetic resonance imaging (MRI) are broadly limited to qualitative (visual) interpretation of the images and/or highly
63 simplified metrics (e.g. measuring the maximum tumor diameter on a single axial slice or using the maximum tracer
64 uptake intensity). The added value of clinical imaging in cancer management is unquestionable, but it remains unclear if
65 such non-invasive investigation sufficiently captures the complicated phenotype of HNSCC to guide risk-based
66 stratification.

67 Radiomics has emerged as a prominent tool for scientific investigation and prognostic modelling of cancer outcomes.
68 Radiomics uses large numbers of quantitative features per subject extracted by a computer algorithm from annotated
69 regions of interest (ROIs) in CT, MRI and PET images [3-5]. More recently, deep learning neural networks (DLNNs) [6,
70 7] have delivered many significant advances in the field of computerized image analysis, hence DLNN-based oncology
71 outcome modelling is now a rapidly growing research topic. The former requires pre-defined mathematical functions to
72 be evaluated on a region of interest in the image, the so-called “hand-crafted (HC) features” approach. In contrast,
73 DLNNs abstract image information as “deep-learning (DL) features” through a consecutive sequence of local convolution
74 and max-pooling steps. Thus, the latter is considered an exclusively data-driven or knowledge-agnostic approach that
75 does not require pre-definition of features.

76 From 2020 onwards, there have been many reviews about HC radiomics and DL for HNSCC prognostication, indicating
77 growing interest and rapid innovations in this field of study [8-16]. Giraud et al. [17] conducted a wide-ranging overview
78 of machine learning for radiotherapy applications (including radiomics) in head and neck cancers but did not provide a

79 systematized synthesis of evidence nor detailed critical appraisals of methodological quality. Spadarella et al. [18]
80 supplied a systematic review of radiomics for nasopharyngeal carcinoma for MRI only. Sanduleanu et al. [19] proposed a
81 radiomics quality scoring system however calibration of this scoring scale remains uncertain. Other authors pointed out
82 possible problems of reducing something as highly nuanced and complex as study quality into a single value [20, 21].
83 Guha et al. [22] systematically reviewed the radiomics literature up till February 2018 for effectiveness of treatment but
84 did not explicitly search for DL imaging studies. Despite these valuable contributions, a critically appraised synthesis
85 comparing both radiomics and deep learning for HNSCC prognostication, covering a broad range of imaging modalities
86 and addressing methodological rigor, remains lacking.

87 The central question addressed in this systematic review with critical appraisal of methodological quality is to estimate
88 the *discriminative performance envelopes of prognostic HNSCC models employing handcrafted radiomics (HC)*
89 *and/or deep learning (DL)*. The performance data shall be gleaned from high-quality primary research articles containing
90 long-term treatment outcomes in locally advanced primary OPC, HPC and laryngeal carcinoma (LC) that are widely
91 treated by (chemo-) radiotherapy, either alone or post-operatively. The primary result expected is a body of evidence for
92 relative efficacy of HC versus DL features for the prognostication tasks in HNSCC

93 In this review we do not cover: (i) studies that are principally about nasopharyngeal carcinoma because it is an
94 epidemiologically distinct disease, and (ii) local cancers in the oral cavity that are managed with surgery alone. We
95 placed emphasis on the methodological reliability of each study and summarized the reviewed models.

96 **Methods**

97 A protocol for this systematic review has not been prospectively registered on a database before performing it.

98 **Eligibility criteria**

99 Eligible studies include only human subjects diagnosed with primary HNSCC that have been treated by (chemo)-
100 radiotherapy either alone or in combination with surgery and post-operative RT. This clinical setting was selected
101 because of nominally standardized and quality-assured protocols (particularly of radiotherapy planning CT that are
102 needed for radiation dosimetry calculations) along with expertly outlined Gross Tumor Volume (GTV) as the region of
103 interest (ROI) by the practicing clinicians.

104 Studies must report at least one clinical outcome (such as all-causes mortality, cancer-related mortality, progression,
105 regression, local and/or regional failure, or distant metastasis). Articles eligible for review contained: (i) HC radiomics
106 and/or DL features derived from pre-treatment clinical imaging, and (ii) TRIPOD 3B (development and validation using
107 separate data) or higher (Type 4: validation only) type of investigation of clinical outcomes modelling [23].

108 **Exclusion criteria**

109 Specific exclusion criteria were: (i) nasopharyngeal carcinoma, (ii) exclusively phantom, *in vitro* or *in silico* studies, and
110 (iii) clinical oncology imaging modality other than CT, PET or MRI.

111 Studies concerning exclusively short-term response immediately following treatment (such as RECIST criteria, tumor
112 expansion/shrinkage) or studies pertaining exclusively to radiomics/DLNN-based diagnostic characterization (such as
113 epidermal growth factor or HPV expression), but without long-term clinical outcomes, were excluded from the review.

114 Excluded studies also lacked peer-reviewed full text from the publishing journal, or if published before 1st January 2011,
115 or if full text was not available in the English language.

116 **Information sources**

117 The primary search for eligible studies up to the end of May 2023 was conducted within the PubMed electronic database
118 after it had been merged with EMBASE. The secondary search was conducted by scanning for eligible studies within the
119 bibliography of reviews and systematic reviews. “Grey” literature sources were not consulted for this review. By-hand
120 searching of individual journal catalogues was not performed. Non-peer reviewed article collections (e.g. arXiv and
121 medArXiv) were omitted from this search.

122 **Search strategy**

123 For PubMed, a sensitive search for diagnostic and prognostic studies was performed using a combination of the broad
124 Haynes [24] and Ingui [25] filters, with an additional modification proposed by Geersing [26]. The search was narrowed
125 using MeSH term for “head and neck cancer”, or text words anywhere in the title and abstract referring to radiomics and
126 deep learning (including common synonyms). Text word searches were first combined with ‘OR’ operators, then
127 integrated to MeSH term and prognostics studies filter using ‘AND’ operators. The plain text of our search string is
128 provided as Supplemental Text Box 1. The search was conducted in two phases, in August 2020 and again on March 31st,
129 2021, and all returned records were merged prior to screening. We also checked the references of all the review articles
130 on ‘head and neck prognostication’ for any additional articles that may have been missed in our electronic database
131 search.

132 **Study selection**

133 We approximated the methodological conventions established by the Cochrane Collaboration for systematic reviews due
134 to the small size of our review team. Two reviewers (VG and LAAF) during the first search phase and two reviewers (VG
135 and HMTT) during the second search phase independently screened PubMed records only by title and abstract to identify
136 potential articles. Disagreements during screening were resolved by unanimous consensus through re-appraisal together
137 with a third reviewer (LW). The full text for candidate articles was obtained through the authors’ institutional library
138 subscriptions. Three reviewers working separately (VG, LW and HMTT) subjected full texts to a detailed reading against
139 inclusion and exclusion criteria, then additional disagreements were resolved by unanimous consensus through
140 reappraisal.

141 **Data extraction**

142 First, general details of the eligible studies were summarized as tables. These included the primary cancer type, imaging
143 modality and image acquisition details, cohort clinical information with sample size, primary outcomes including non-
144 radiomics and non-deep-learning-based comparator factors, and the software base for HC or DL.

145 **Estimating risk of bias in individual studies**

146 There have been several tools proposed to appraise methodological quality of prognostic and diagnostic studies in
147 general, such as QUADAS [27], or radiomics-specific score (RQS) [19].

148 We have based a methodological appraisal on the rationale raised by the RQS but refrained from assigning a single
149 quality number. In its place, we included a brief overview of what, in our view, might have constituted some part of
150 methodological robustness in the study. Each of the three reviewers worked independently on extracting the
151 methodological information and was afterwards cross-checked by another reviewer. The methodological aspects we
152 sought to extract from the studies were the following:

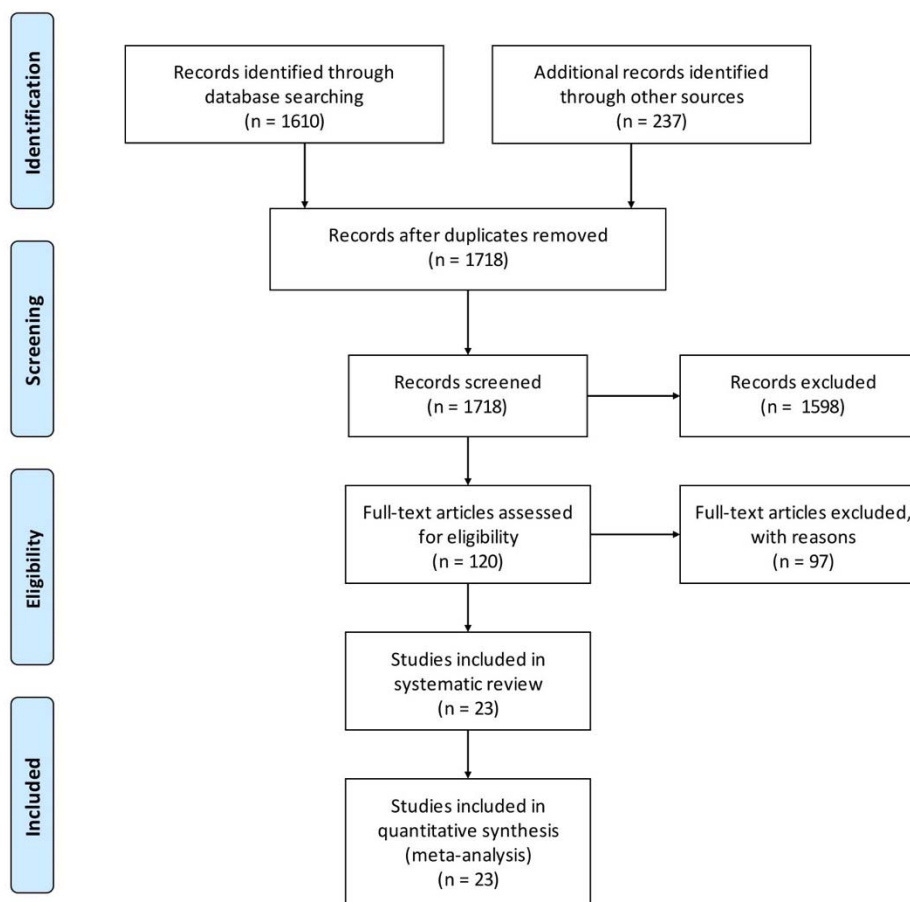
- 153 1. Was the study prospectively registered for the intended analysis methodology in a publicly accessible study
154 database prior to commencement of the analysis?
- 155 2. If imaging data were not publicly available for download, were sufficient details present in the article to identify
156 the scope of validity (e.g., diversity of equipment vendors, whether i.v. contrast was used, image acquisition
157 parameters etc.)?
- 158 3. If digital image pre-processing had been applied (digital filters, isotropic resampling, augmentations such as flips
159 or rotations, and related operations) was enough information provided or standardized steps documented, that
160 would support reproducing the same steps independently?
- 161 4. If some form of model simplification had been performed, such as feature dimensionality reduction or drop-outs,
162 was enough information provided or standardized steps documented, that would support reproducing the same
163 model simplifications independently?
 - 164 • Was some form of model interpretability incorporated into the findings, such as biological correlates of
165 HC features, or attention maps of DL features, that could support a high degree of clinical verification
166 of the output?
- 167 5. If risk stratification groups had been defined (e.g. cut-offs and operating point) was a clinical justification
168 provided rather than solely relying on model fine-tuning for optimal groupings, since the latter might produce
169 overly optimistic results of discrimination?

- 170 6. Was the reference standard of outcomes used in the supervised learning (also known as ground truth) provided by
171 human experts and closely matched with the context of the clinical decision being supported by the proposed
172 model?
173 7. Whether the expected clinical utility of the proposed model had been estimated, through some form of cost-
174 benefit or decision-curve analysis or related measure of decision-making utility?

175 Results

176 Study selection

177 The PRISMA (Preferred Reporting Items for Systematic review and Meta-Analysis) [28, 29] flowchart is provided as
178 Figure 1. 1610 records were identified based on the specified search terms in PubMed and 237 additional records through
179 other sources. After duplicates removal, there were 1718 articles available for screening. Applying the selection criteria
180 led to 120 studies for full-text screening. In the end, 23 articles were deemed eligible.



181
182 **Fig 1.** Meta-analyses (PRISMA) workflow to select articles for review.

183 Characteristics of included studies

184 Table 1 presents an overview of the general characteristics observed in all the studies included. The majority (16/23) of
185 the studies included HC radiomics, six studies included DL, and one study involved both HC and DL features for
186 prediction of outcome following radiation therapy for HNSCC.
187

188 The most widely reported disease subsite was oropharyngeal cancer with almost all studies including this in their dataset
189 (22/23). Next major tumor site represented was larynx (18/23), followed by hypopharynx (13/23) and oral cavity (7/23).
190 Although we did not specifically include nasopharyngeal cancer, some studies reported a mixed subset of patients with
191 NPC (9/23) and paranasal cancer (1/23) in the training or validation cohorts.

192 Most of the studies included radiomics derived from radiotherapy treatment planning images; some of them included
193 FDG PET-CT (8/23), FDG PET (4/23) or CT (14/23). Additionally, only one study reported the use of MRI. Only 5
194 studies reported the use of contrast-enhanced CT images.

195 Most of the patients included in the studies were treated with definitive radiotherapy or chemo-radiation therapy. Cheng
196 et al. included some patients who underwent surgery [30]. One study included patients who had surgery following RT
197 [31]. Zhai et al. specifically excluded patients who underwent elective neck node dissection following RT [32].

198 The sample size of the cohorts reported in this review ranged from 52 to 2552. For HC radiomics studies used training
199 dataset sizes ranging from 124 [33] to 377 [34], (mean 202; SD 101), and their validation dataset sizes varied from 65
200 [35] to 542 [36], (Mean 143; SD 107). In the DL studies, the training datasets sizes ranged from 102 [37] to 2552 patients
201 [38], (Mean 531; SD 826) and independent validation datasets from 52 [37] to 872 patients [38], (Mean 200; SD 133).

202 A wide variety of software tools were used to extract the HC features, with 9 studies reporting the use of custom-built
203 codes using MATLAB [31, 32, 36, 39-43] or Python [44]. The other open-source software reported were Z-Rad [45],
204 IBEX [34], CERR [35], OncoRadiomics [46], Pyradiomics [33] and MIRP [47].

205 Most DL studies [30, 37, 48-50] applied a reasonably consistent CNN architecture, from what could be gleaned in the
206 technical details of the publications. For instance, Le et al. used a 3-layer neural network with self-attention, also known
207 as PreSANet (Pre-Self-Attention Network) [51] while Kazmierski et al. used a deep multitask logistic regression to model
208 the time-to-event [38].

209 **Summarized performances of included studies**

210 Table 2 summarizes the endpoints, model building aspects, and performance of the different models. The most studied
211 prognosis endpoint (16/23) was overall survival (OS) followed by local disease failure, recurrence or control and finally
212 distant metastasis. The event to sample ratios for OS ranged from 15% to 81% and distant metastasis (DM) rates varied
213 between 12% and 19%. The rates for loco-regional recurrence (LRR), local recurrence (LR) local failure (LF) or local
214 regional control (LRC), ranged from 7% to 67%.

215 Notably, all sixteen HC radiomics studies employed various feature reduction techniques to streamline their radiomics
216 models. Among these, Spearman's correlation ranking with the other features emerged as one of the popular methods [31,
217 40, 43, 47] followed by LASSO regression models [31, 34]. Some studies employed more than one feature reduction
218 approach before building the prognostic models. For example, [42, 44] utilized 13 feature selection methods for their
219 various machine learning algorithms.

220 The most frequently (9/23) used machine learning model for HC radiomics was multiple Cox regression technique
221 followed by the multiple linear regression. When we analyzed the prognostic performance of these models across the
222 studies based on reported test AUC/C-Indices, the best performing model for OS was reported by Goncalves et al. with an
223 AUC/C-index of 0.91 using HC features [33]. Analyzing the DL studies separately showed that Kazmierski et al. [38]
224 achieved the highest performing OS model with a discriminatory AUC of 0.82. Vallieres et al. [40] reported the highest
225 performing Distant metastasis (DM) prediction model using HC radiomics and clinical parameters with a C-Index of 0.88
226 while Lombardo (2021) [49] achieved an AUC of 0.89 by including DL and clinical features. Interestingly, [48]
227 constructed a CNN model using HC radiomics features, which surpassed all other performances and achieved an AUC/C-
228 index of 0.92.

229 Consequently, the majority of HC studies (7 out of 16) and DL studies (5 out of 6) found that incorporating radiomics,
230 either alone or combined with clinical parameters enhanced the predictive power and offered value compared to the
231 traditional clinical models.

232 **Methodological quality assessment**

233 Given the large number of HC or DL studies for prognostication in head and neck cancer, we restricted the assessment of
 234 methodological quality of studies that adhered to TRIPOD guidelines and described the development or validation of the
 235 model or both. (TRIPOD 3B and 4). Figure 2 gives an overview of the distribution of the methodological quality and
 236 reporting completeness for 23 studies selected for this review. An extended explanation of the reasons for the scores is
 237 made available as part of the supplementary Table S1.

Number	Reference	Prospective registration	Imaging protocol	Image pre-processing	Model simplification and reproducibility	Model interpretability	Risk groupings justified	Reliability of model	Estimation of clinical utility	Number of items rated good
1	Meneghetti:2021 [47]	Red	Green	Green	Green	Green	Green	Green	Red	6
2	Kim:2022 [31]	Red	Yellow	Green	Green	Green	Green	Green	Red	5
3	Fujima:2021 [37]	Red	Green	Green	Green	Yellow	Green	Yellow	Red	4
4	Ger:2019 [34]	Red	Yellow	Green	Green	Green	Green	Yellow	Red	4
5	Lv:2020 [43]	Red	Yellow	Green	Green	Green	Green	Yellow	Red	4
6	Starke:2020 [50]	Red	Yellow	Green	Green	Yellow	Green	Green	Red	4
7	Zhai: 2021 [32]	Red	Yellow	Red	Green	Green	Green	Green	Red	4
8	Aerts:2014 [39]	Red	Yellow	Green	Yellow	Green	Green	Yellow	Red	3
9	Bogowicz:2020 [45]	Red	Yellow	Green	Green	Red	Green	Yellow	Red	3
10	Cheng:2021 [30]	Red	Yellow	Green	Green	Yellow	Green	Yellow	Red	3
11	Kazmierski:2023 [38]	Red	Yellow	Green	Green	Yellow	Green	Yellow	Red	3
12	Keek:2020 [46]	Red	Yellow	Green	Green	Yellow	Green	Yellow	Red	3
13	Le WT:2022 [51]	Red	Yellow	Green	Green	Yellow	Green	Yellow	Red	3
14	Leger:2017 [44]	Red	Yellow	Green	Green	Red	Green	Yellow	Red	3
15	Lombardo:2021 [49]	Red	Yellow	Green	Green	Yellow	Green	Yellow	Red	3
16	Parmar:2015a [41]	Red	Yellow	Green	Green	Yellow	Green	Yellow	Red	3
17	Parmar:2015b [42]	Red	Yellow	Green	Green	Red	Green	Yellow	Red	3
18	Vallières:2017 [40]	Red	Yellow	Red	Green	Green	Green	Yellow	Red	3
19	Diamant:2019 [48]	Red	Yellow	Green	Green	Yellow	Red	Yellow	Red	2
20	Folkert:2017 [35]	Red	Yellow	Yellow	Green	Red	Green	Yellow	Red	2
21	Goncalves:2022 [33]	Red	Yellow	Green	Green	Yellow	Red	Yellow	Red	2
22	Leijenaar:2015 [36]	Red	Yellow	Red	Green	Red	Green	Yellow	Red	2
23	Zhou:2020 [52]	Red	Yellow	Red	Red	Green	Red	Yellow	Red	1

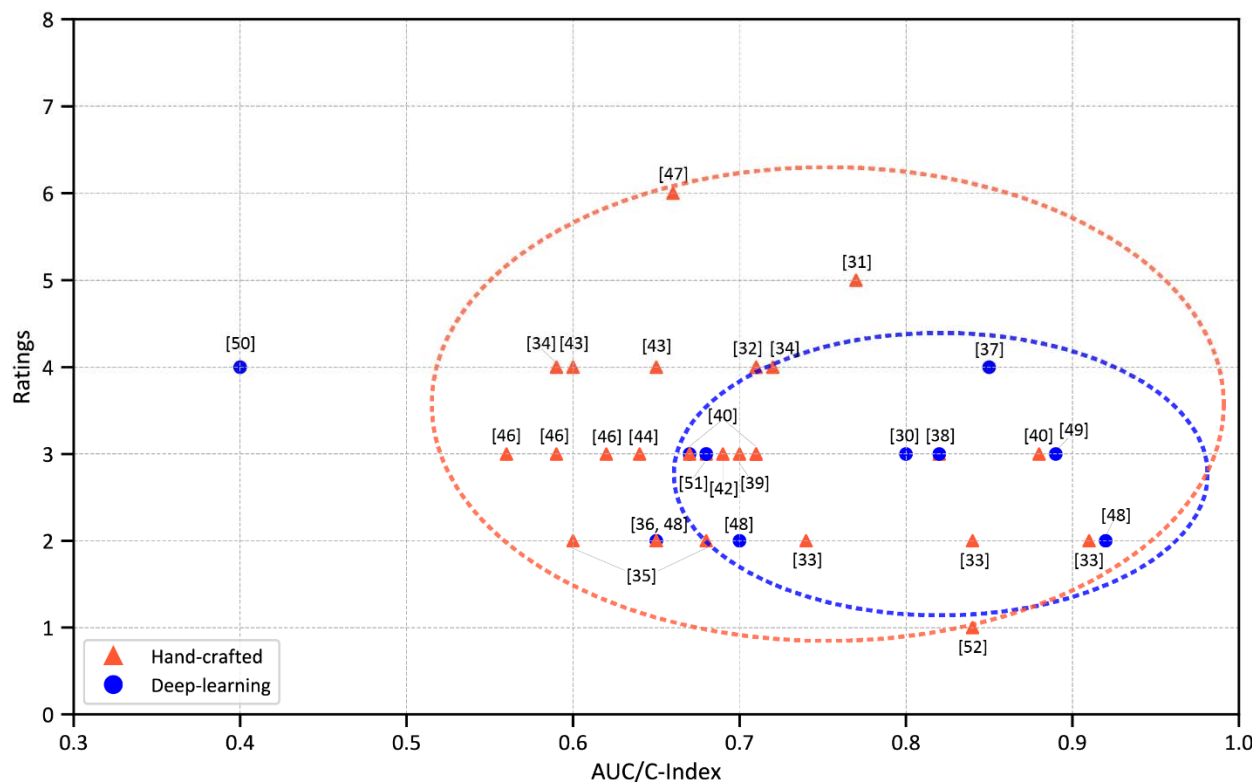
238
 239 **Fig 2.** Methodological quality assessment of included studies. Red, yellow, and green dots represent poor, medium, and good quality,
 240 respectively. The final column shows the total number of good scores (green) out of 8 quality assessments.

241 The methodological assessment involved using an 8-item rating system rating based on the criteria as mentioned earlier
 242 under the Methods section, with red, yellow, and green indicating poor, medium, and good respectively. The highest
 243 score achieved was 6/8 by only one study, falling short due to study not prospectively registered and show the clinical
 244 utility of the models [47]. More than half of the reporting was of potentially suboptimal methodological quality and

245 achieved a score of 3 or lower. None of the studies were prospectively registered prior to the HC or DL analysis, which
246 recapitulates a general limitation seen globally in prediction modelling studies using HC features and DL. Only 2/23
247 studies reported all essential details regarding the imaging acquisition protocol, and 18/23 about any image pre-
248 processing. Regarding assessing the clinical utility of the developed models through methodologies like cost-benefit
249 analysis or decision curve analysis; none of the studies included in this review fulfilled this criterion. Most studies (20/23)
250 included the essential model simplification techniques such as feature selection/multi-dimensionality reduction, hyper-
251 parameters, and dropout rates etc. that allow reproducibility of the models. However, fewer (7/23) studies included
252 measures for interpretability such as comparison with biological correlates. Most studies (20/23) provided appropriate
253 justification for risk grouping/risk cut-offs to delineate risk subgroups; however, three studies [33, 48, 52] did not include
254 any risk stratification.

255 We found only 4/23 of the studies investigated the overall effectiveness of the models in comparison to the traditional
256 clinical models within the healthcare setting. The broad parameters assessed included were if a) the AI models were
257 compared to clinical models b) the models were trained and tested for the outcomes appropriately c) if the outcome was
258 survival, the survival data used were linked to any cancer registries, and d) adequate documentation of the final model
259 that permits reproducibility for external validation. The two studies that reported 'good' in at least 5 out of the 8
260 assessment items [31, 47] used HC radiomics and were both reporting for local disease as the outcome. Of the five studies
261 that had ratings of 4, 3 used HC radiomics and 2 used DL, all reporting varied outcomes. Most of the other studies
262 (16/23) had ratings of 3 or less, which included 10 studies using HC radiomics, 5 using DL methods and 1 study having
263 both HC and DL features.

264 Figure 3 visualizes the reported discriminatory metrics (AUC/C-index) against the number of methodological items rated
265 'good' in this review. The color-codes refer to the type of features used for modelling the outcomes, namely HC or DL.
266 The top two methodological rated studies had a discriminatory performance between 0.66-0.77. We noticed that most of
267 the studies lie within a wide scatter with respect to the performances ranging from 0.58-0.92 and ratings between 2 and 4.
268 From three HC radiomics studies, we observed an AUC/C Index of 0.83 to 0.91 [33, 40, 52] for four outcomes. Five DL
269 studies had discriminatory performance between 0.80-0.92 [30, 37, 38, 48, 49]. To view the reported metrics against our
270 ratings for each outcome individually, please refer to the supplementary materials.
271



272

273 **Fig 3.** Reported discriminatory metrics (AUC/C-index) of included studies with the number of methodological items rated ‘good’ in
274 this review.

275 **Discussion**

276 In this systematic review, we summarized the basic characteristics and reported results of studies and rigorously
277 evaluated the methodological quality of studies that included either HC radiomics or DL methods to predict disease
278 outcome in patients treated for head and neck cancer. The models focused on the prediction of recurrent disease and/or
279 survival and were constructed using either HC or DL based radiomics features or both. Only studies that qualified as
280 TRIPOD 3b or higher (had independent validation of the results in an external dataset) were included in the review.
281 While a handful of the studies have reported encouraging results and hinted at their suitability for clinical use, a
282 considerable portion of studies still fall short in their methodological rigor. Future studies can enhance the quality based
283 on the quality checklist provided which allows them to think about employing more robust methodologies and ensuring
284 documentation for wider implementation.

285 Figure 3 offers an overview of the studies examined in this review, presenting their reported performance metrics
286 alongside ratings from our methodological assessment independently for models validated using either HC or DL based
287 radiomics. The two studies with the highest reported AUC/C index metrics for HC and DL also happen to have very low
288 methodological robustness [33, 48]. Overall, we noted that the DL models exhibited higher performance scores
289 compared to the HC models with the exception of [50] that recorded the lowest performance of AUC/C Index of 0.4.
290 However, despite the higher discriminatory performance, we observed that the methodological aspects were generally
291 better in studies currently using HC radiomics than those involving DL methods. This could be partly attributed to the
292 significant efforts towards standardization of HC radiomics, particularly through initiatives like the Imaging Biomarker
293 Standardization Initiative (IBSI) [53] that has led to clearer definitions, workflows, and best practices. In contrast, deep
294 learning, while rapidly evolving and is sometimes integrated with the HC radiomics workflow, currently lacks the same
295 level of formalized guidelines emphasizing the need for such initiatives.

296 It would have been optimal if the data collection and statistical analysis protocol for radiomics modeling had been pre-
297 registered. Platforms like ClinicalTrials.gov could serve this purpose, providing transparency in the analysis. Regrettably,
298 none of the studies in this review report such a pre-registration. This may be attributed to the absence of widely available
299 consensus on where such protocols can be registered in advance. We also recommend that, as a radiomics community, we
300 should further promote biomedical modeling registries like the AIME registry [54]. These platforms should facilitate
301 review, provide suggestions for collaboration, and offer feedback on statistical protocols before initiating a radiomics
302 project study. Transparent registration helps ensure reproducibility and credibility in radiomics research by minimizing
303 biases and clearly defining the methodological framework, including training and validation strategies.

304 Similarly, we observed that AI prognostication models often overlook the assessments of their clinical implications and
305 applicability for practical use. The evaluation of clinical utility, carried out through methodologies like cost-benefit
306 analysis or decision curve analysis [55, 56], it is imperative for gaining insights on the practical implications of these
307 models. Regrettably, none of the papers included in this review fulfilled both criteria, highlighting a gap in research
308 transparency and pre-analysis protocol documentation.

309 In our methodological assessment, we evaluated studies based on the clarity of outcome definitions and endpoints. While
310 predicting patient prognosis remains a challenge, to ensure the reliability of the prognostic models, a clear definition of
311 primary endpoint is required that are both valid and reliable. Currently, the endpoints were accepted as defined by the
312 clinicians based on the oncology practices. The endpoints studied include OS, LRR and DM. Most studies lacked clear
313 information for how their clinical endpoints were determined, and whether this was accurately and consistently applied.
314 For instance, when modeling OS, better statistics on the date of death of people that can be prospectively collected is
315 preferred to assess survival interval, as opposed to phone surveys with next-of-kin, but it was not always clear how the
316 important endpoint information was obtained. It could be hospital-based or like in the Netherlands, a national population
317 registry for all births/deaths related information. Overall, there was heterogeneity in the broad definitions of the endpoints
318 and follow-up periods available for survival analysis which could have also contributed to the results varying
319 significantly, with C-Index/AUC values ranging between 0.40 [50] to 0.92 [48]. Notable exceptions to where there was
320 clarity on the clinician-defined endpoints include [31, 32, 47, 50].

321 Many studies showed [30, 32, 33, 38-40, 43, 47, 49, 51] that a combined model involving clinical factors and imaging
322 features outperforms the results of just the clinical model. These findings suggest that multi-dimensional data possesses
323 greater predictive capability compared to a predictive model constructed solely with mono-dimensional data. However,
324 Le et al. reported that the addition of PET to either the CT or a combination of CT and clinical DL model showed a
325 marked decrease in performance in the models' predictive capability for all endpoints [51]. It is interesting to note that the
326 same training dataset was used by Goncalves [33] and Vallieres [40].

327 Clinical research and modelling become highly relevant to the clinician only if the study is accurate and reproducible.
328 The prognostic efficacy of the survival models leveraging radiomics features relies on the utilization of stable and
329 reproducible features, alongside transparent imaging protocols [57, 58]. While studies such as [47] and [37] stand out as
330 exemplary studies with comprehensive and reproducible imaging details, other studies were missing key information such
331 as CT image acquisition and reconstruction parameters.

332 Most studies also lacked sufficient details for model reproducibility. To reproduce the model, feature engineering and
333 model building are equally important steps. Feature selection methods work by reducing the number of input variables by
334 eliminating redundant features and selecting the most relevant ones for the model. This process significantly enhances
335 model performance, improves interpretability of findings, and addresses generalization issues. In our methodological
336 assessment, we evaluated studies based on their model simplicity and reproducibility. Most HC radiomics studies have
337 clearly outlined their feature reduction and model selection parameters to ensure reproducibility. However, Aerts et al.
338 [39] provided explanations for some statistical methods used but lacked clarity on certain aspects. Zhou et al. (2020) [52]
339 developed multifaceted radiomics models for predicting distant metastases, incorporating both DL and machine learning
340 classifiers. While their feature extraction was conventional, reproducibility steps were not explicitly detailed. For the DL
341 studies, we noted that the model parameters were typically disclosed to ensure reproducibility across most included
342 studies.

343 It is crucial to understand the biological correlates of features included in the model to improve its interpretability. In
344 some HC radiomics studies we noticed a lack of emphasis on model interpretability, like comparing the model's
345 performance with established clinical parameters [35, 36, 42, 44, 45]. Goncalves et al. [33] incorporated a combined
346 model of radiomics and clinical parameters, but it should be noted that the clinical parameters were not predictors usually
347 reported in literature for that outcome. While [42, 46] integrate clinical parameters in separate models, a combined model
348 for interpreting the biological implications of the radiomics parameters was absent. In DL, the focus shifts to making the
349 activated regions, from which features influencing the chosen outcome are derived, interpretable for clinicians. Attention
350 maps play a significant role in this context. Except [48], none provided minimal and maximal activation maps. However,
351 it should be emphasized that the activation maps did not correlate the model's covariates with any known clinical
352 biomarkers. Except [37], other studies trained models incorporating both DL and known clinical features, with a focus to
353 enhance the comprehension of the biological correlates of deep-extracted features.

354 During our literature review, we encountered papers submitted for the HECKTOR Challenge at MICCAI 2021 and 2022
355 [59, 60], which focused on automatic head and neck tumor segmentation and outcome prediction in PET/CT images.
356 These papers demonstrated that by integrating radiomics features with machine learning algorithms, valuable insights can
357 be provided into the metabolic and morphological properties of tumors, aiding in the prediction of patient outcomes. The
358 challenge participants were given the same data, and their work centered on applying DL and conventional radiomics to
359 head and neck cancer diagnosis and prognosis, specifically Recurrence-Free Survival (RFS), using FDG-PET/CT images
360 and available clinical data. Despite being highly relevant to our search criteria, these studies were not included as they did
361 not meet the TRIPOD criteria.

362 We acknowledge several limitations of the current systematic review that future research could address. Firstly, this
363 review was not prospectively registered prior to commencement. Second, we were unable to perform a quantitative meta-
364 analysis owing to the significant heterogeneity in the outcomes analyzed, the methodological and mathematical process
365 involved for HC and DL-based modelling. Instead, we provided a visual synthesis of reported model performance in
366 relation to methodological robustness (Figure 3). Third, despite our rigorous efforts to evaluate methodological
367 procedures using objective criteria, independent raters, and consensus, we believe some degree of subjectivity and
368 potential debatable assessments may remain. Additional detailed notes on methodology are provided in the
369 supplementary table S1 to improve transparency. Lastly, we introduced some inclusion bias by only considering full-text
370 articles in English. This decision was made pragmatically, as all authors of this review are proficient in English, ensuring
371 that the selected material is accessible and understandable to readers who may wish to inspect the individual papers
372 themselves. Additionally, since January 2024 TRIPOD has released an update called TRIPOD+AI. There may be some

373 things in the new TRIPOD that we did not align or incorporate in this present review. However, this may be useful in
374 future work in this question.

375 Conclusion

376 This systematic review provides a critical evaluation of the current state of handcrafted radiomics and deep learning
377 models for prognostication in head and neck squamous cell carcinoma. Despite promising advancements, significant
378 methodological heterogeneity and gaps in reporting standards were identified. The review emphasizes the need for
379 standardized methodologies, including pre-registration of study protocols and detailed reporting of imaging and model
380 development procedures, to improve the reproducibility and clinical utility of these models. Future research should also
381 focus on integrating clinical factors with radiomics features to enhance predictive accuracy and on conducting
382 comprehensive assessments of the clinical implications and cost-effectiveness of these models. Such efforts will be
383 crucial in advancing personalized treatment strategies and improving outcomes for HNSCC patients.

384

385 References

- 386 1. Johnson, D.E., et al., *Head and neck squamous cell carcinoma*. Nature Reviews. Disease Primers, 2020. **6**(1): p.
387 92.
- 388 2. Bray, F., et al., *Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for*
389 *36 cancers in 185 countries*. CA: A Cancer Journal for Clinicians, 2024/05/01. **74**(3).
- 390 3. Lambin, P., et al., *Radiomics: Extracting more information from medical images using advanced feature*
391 *analysis*. European journal of cancer (Oxford, England : 1990), 2012. **48**(4): p. 441-446.
- 392 4. Kumar, V., et al., *QIN "Radiomics: The Process and the Challenges"*. Magnetic resonance imaging, 2012.
393 **30**(9): p. 1234-1248.
- 394 5. Gillies, R.J., P.E. Kinahan, and H. Hricak, *Radiomics: Images Are More than Pictures, They Are Data*.
395 Radiology, 2016. **278**(2): p. 563-577.
- 396 6. Schmidhuber, J., *Deep learning in neural networks: An overview*. Neural Networks, 2015. **61**: p. 85-117.
- 397 7. LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning*. Nature, 2015. **521**(7553): p. 436-444.
- 398 8. Chinnery, T., et al., *Utilizing Artificial Intelligence for Head and Neck Cancer Outcomes Prediction From*
399 *Imaging*. Canadian Association of Radiologists Journal, 2021. **72**(1): p. 73-85.
- 400 9. Tanadini-Lang, S., et al., *Radiomic biomarkers for head and neck squamous cell carcinoma*. Strahlentherapie
401 und Onkologie, 2020. **196**(10): p. 868-878.
- 402 10. Bibault, J.-E., et al., *Radiomics: A primer for the radiation oncologist*. Cancer/Radiothérapie, 2020. **24**(5): p.
403 403-410.
- 404 11. Scheckenbach, K., L. Colter, and M. Wagenmann, *Radiomics in Head and Neck Cancer: Extracting Valuable*
405 *Information from Data beyond Recognition*. ORL, 2017. **79**(1-2): p. 65-71.
- 406 12. Caudell, J.J., et al., *The future of personalised radiotherapy for head and neck cancer*. The Lancet. Oncology,
407 2017. **18**(5): p. e266-e273.
- 408 13. Giannitto, C., et al., *Radiomics-based machine learning for the diagnosis of lymph node metastases in patients*
409 *with head and neck cancer: Systematic review*. Head & Neck, 2023/02/01. **45**(2).
- 410 14. Rasheed Omobolaji Alabi, M.E., Ilmo Leivo, Alhadi Almangush, Antti A. Mäkitie, *Artificial Intelligence-Driven*
411 *Radiomics in Head and Neck Cancer: Current Status and Future Prospects*. International Journal of Medical
412 Informatics, 2024/08/01. **188**.
- 413 15. Tortora, M., et al., *Radiomics Applications in Head and Neck Tumor Imaging: A Narrative Review*. Cancers,
414 2023/02. **15**(4).
- 415 16. Li, S., et al., *Application of PET/CT-based deep learning radiomics in head and neck cancer prognosis: a*
416 *systematic review*. Radiology Science, 2022.

- 417 17. Giraud, P., et al., *Radiomics and Machine Learning for Radiotherapy in Head and Neck Cancers*. Frontiers in
418 Oncology, 2019. **9**.
- 419 18. Spadarella, G., et al., *MRI based radiomics in nasopharyngeal cancer: Systematic review and perspectives using
420 radiomic quality score (RQS) assessment*. European Journal of Radiology, 2021. **140**: p. 109744.
- 421 19. Sanduleanu, S., et al., *Tracking tumor biology with radiomics: A systematic review utilizing a radiomics quality
422 score*. Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology,
423 2018. **127**(3): p. 349-360.
- 424 20. Whiting, P., R. Harbord, and J. Kleijnen, *No role for quality scores in systematic reviews of diagnostic accuracy
425 studies*. BMC Medical Research Methodology, 2005. **5**(1): p. 19.
- 426 21. Fornaçon-Wood, I., et al., *Radiomics as a personalized medicine tool in lung cancer: Separating the hope from
427 the hype*. Lung Cancer (Amsterdam, Netherlands), 2020. **146**: p. 197-208.
- 428 22. Guha, A., et al., *Radiomic analysis for response assessment in advanced head and neck cancers, a distant dream
429 or an inevitable reality? A systematic review of the current level of evidence*. The British Journal of Radiology,
430 2020. **93**(1106): p. 20190496.
- 431 23. Moons, K.G.M., et al., *Transparent Reporting of a multivariable prediction model for Individual Prognosis or
432 Diagnosis (TRIPOD): explanation and elaboration*. Annals of Internal Medicine, 2015. **162**(1): p. W1-73.
- 433 24. Haynes, R.B., et al., *Optimal search strategies for retrieving scientifically strong studies of treatment from
434 Medline: analytical survey*. BMJ, 2005. **330**(7501): p. 1179.
- 435 25. Ingui, B.J. and M.A.M. Rogers, *Searching for Clinical Prediction Rules in Medline*. Journal of the American
436 Medical Informatics Association, 2001. **8**(4): p. 391-397.
- 437 26. Geersing, G.-J., et al., *Search Filters for Finding Prognostic and Diagnostic Prediction Studies in Medline to
438 Enhance Systematic Reviews*. PLOS ONE, 2012. **7**(2): p. e32844.
- 439 27. Whiting, P., et al., *The development of QUADAS: a tool for the quality assessment of studies of diagnostic
440 accuracy included in systematic reviews*. BMC Medical Research Methodology, 2003. **3**(1): p. 25.
- 441 28. Moher, D., et al., *Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA
442 Statement*. PLOS Medicine, 2009. **6**(7): p. e1000097.
- 443 29. Page, M.J., et al., *The PRISMA 2020 statement: an updated guideline for reporting systematic reviews*. BMJ,
444 2021: p. n71.
- 445 30. Cheng, N.-M., et al., *Deep Learning for Fully Automated Prediction of Overall Survival in Patients with
446 Oropharyngeal Cancer Using FDG-PET Imaging*. Clinical Cancer Research, 2021. **27**(14): p. 3948-3959.
- 447 31. Kim, M., et al., *Development and Validation of a Model Using Radiomics Features from an Apparent Diffusion
448 Coefficient Map to Diagnose Local Tumor Recurrence in Patients Treated for Head and Neck Squamous Cell
449 Carcinoma*. Korean Journal of Radiology, 2022. **23**(11): p. 1078-1088.
- 450 32. Zhai, T.-T., et al., *External validation of nodal failure prediction models including radiomics in head and neck
451 cancer*. Oral Oncology, 2021. **112**: p. 105083.
- 452 33. Gonçalves, M., et al., *Radiomics in Head and Neck Cancer Outcome Predictions*. Diagnostics, 2022. **12**(11): p.
453 2733.
- 454 34. Ger, R.B., et al., *Radiomics features of the primary tumor fail to improve prediction of overall survival in large
455 cohorts of CT- and PET-imaged head and neck cancer patients*. PLOS ONE, 2019. **14**(9): p. e0222509.
- 456 35. Folkert, M.R., et al., *Predictive modeling of outcomes following definitive chemoradiotherapy for oropharyngeal
457 cancer based on FDG-PET image characteristics*. Physics in Medicine & Biology, 2017. **62**(13): p. 5327.
- 458 36. Leijenaar, R.T.H., et al., *External validation of a prognostic CT-based radiomic signature in oropharyngeal
459 squamous cell carcinoma*. Acta Oncologica, 2015.
- 460 37. Fujima, N., et al., *Prediction of the local treatment outcome in patients with oropharyngeal squamous cell
461 carcinoma using deep learning analysis of pretreatment FDG-PET images*. BMC Cancer, 2021. **21**(1): p. 900.

- 462 38. Kazmierski, M., et al., *Multi-institutional Prognostic Modeling in Head and Neck Cancer: Evaluating Impact*
463 *and Generalizability of Deep Learning and Radiomics*. Cancer Research Communications, 2023. **3**(6): p. 1140-
464 1151.
- 465 39. Aerts, H.J.W.L., et al., *Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics*
466 *approach*. Nature Communications, 2014. **5**(1): p. 4006.
- 467 40. Vallières, M., et al., *Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer*.
468 Scientific Reports, 2017. **7**(1): p. 10117.
- 469 41. Parmar, C., et al., *Radiomic feature clusters and Prognostic Signatures specific for Lung and Head & Neck*
470 *cancer*. Scientific Reports, 2015. **5**(1): p. 11044.
- 471 42. Parmar, C., et al., *Radiomic Machine-Learning Classifiers for Prognostic Biomarkers of Head and Neck Cancer*.
472 Frontiers in Oncology, 2015. **5**.
- 473 43. Lv, W., et al., *Multi-Level Multi-Modality Fusion Radiomics: Application to PET and CT Imaging for*
474 *Prognostication of Head and Neck Cancer*. IEEE Journal of Biomedical and Health Informatics, 2020. **24**(8): p.
475 2268-2277.
- 476 44. Leger, S., et al., *A comparative study of machine learning methods for time-to-event survival data for radiomics*
477 *risk modelling*. Scientific Reports, 2017. **7**(1): p. 13206.
- 478 45. Bogowicz, M., et al., *Privacy-preserving distributed learning of radiomics to predict overall survival and HPV*
479 *status in head and neck cancer*. Scientific Reports, 2020. **10**(1): p. 4542.
- 480 46. Keek, S., et al., *Computed tomography-derived radiomic signature of head and neck squamous cell carcinoma*
481 *(peri)tumoral tissue for the prediction of locoregional recurrence and distant metastasis after concurrent*
482 *chemo-radiotherapy*. PLOS ONE, 2020. **15**(5): p. e0232639.
- 483 47. Rabasco Meneghetti, A., et al., *Definition and validation of a radiomics signature for loco-regional tumour*
484 *control in patients with locally advanced head and neck squamous cell carcinoma*. Clinical and Translational
485 Radiation Oncology, 2021. **26**: p. 62-70.
- 486 48. Diamant, A., et al., *Deep learning in head & neck cancer outcome prediction*. Scientific Reports, 2019. **9**(1): p.
487 2764.
- 488 49. Lombardo, E., et al., *Distant metastasis time to event analysis with CNNs in independent head and neck cancer*
489 *cohorts*. Scientific Reports, 2021. **11**(1): p. 6418.
- 490 50. Starke, S., et al., *2D and 3D convolutional neural networks for outcome modelling of locally advanced head and*
491 *neck squamous cell carcinoma*. Scientific Reports, 2020. **10**(1): p. 15625.
- 492 51. Le, W.T., et al., *Cross-institutional outcome prediction for head and neck cancer patients using self-attention*
493 *neural networks*. Scientific Reports, 2022. **12**(1): p. 3183.
- 494 52. Zhou, Z., et al., *Multifaceted radiomics for distant metastasis prediction in head & neck cancer*. Physics in
495 Medicine & Biology, 2020. **65**(15): p. 155009.
- 496 53. Zwanenburg, A., et al., *The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics*
497 *for High-Throughput Image-based Phenotyping*. Radiology, May 2020. **295**(2).
- 498 54. Matschinske, J., et al., *The AIMe registry for artificial intelligence in biomedical research*. Nature Methods,
499 2021. **18**(10): p. 1128-1131.
- 500 55. Vickers, A.J. and E.B. Elkin, *Decision Curve Analysis: A Novel Method for Evaluating Prediction Models*.
501 Medical Decision Making, 2006. **26**(6): p. 565-574.
- 502 56. Guerra, A., et al., *Clinical application of machine learning models in patients with prostate cancer before*
503 *prostatectomy*. Cancer Imaging, 2024. **24**(1): p. 24.
- 504 57. Jha, A.K., et al., *Repeatability and reproducibility study of radiomic features on a phantom and human cohort*.
505 Scientific Reports, 2021. **11**(1): p. 2055.

- 506 58. Dirnagl, U., et al., *Reproducibility, relevance and reliability as barriers to efficient and credible biomedical*
507 *technology translation*. *Advanced Drug Delivery Reviews*, 2022. **182**: p. 114118.
- 508 59. Andrearczyk, V., et al., *Overview of the HECKTOR Challenge at MICCAI 2021: Automatic Head and N.*
509 *Lecture Notes in Computer Science*, 2022.
- 510 60. Andrearczyk, V., et al., *Overview of the HECKTOR Challenge at MICCAI 2022: Automatic Head and Neck*
511 *Tumor Segmentation and Outcome Prediction in PET/CT*. *Head and neck tumor segmentation and outcome*
512 *prediction : third challenge, HECKTOR 2022, held in conjunction with MICCAI 2022, Singapore, September*
513 *22, 2022, Proceedings. Head and Neck Tumor Segmentation Challenge (3rd : 2022 : Singapor, 2023. 13626*.
- 514 61. *USZ Medical Physics*. [cited 2024 August 24]; Available from: <https://medical-physics-usz.github.io/>.
- 515 62. *Py-Radiomics*. [cited 2024 August 24]; Available from: <https://www.radiomics.io/pyradiomics.html>.
- 516 63. *Radiogenomics*. [cited 2024 August 24]; Available from: <https://github.com/jieunp/radiogenomics>.
- 517 64. *Medical Image Radiomics Processor*. [cited 2024 August 24]; Available from: <https://github.com/oncoray/mirp>.
- 518 65. *DeepPET-OPSCC-Example*. [cited 2024 August 24]; Available from: [https://github.com/deep-med/DeepPET-](https://github.com/deep-med/DeepPET-OPSCC-Example)
519 [OPSCC-Example](https://github.com/deep-med/DeepPET-OPSCC-Example).
- 520 66. *UHN RADCURE Prognostic Modelling Challenge 2020*. [cited 2024 August 24]; Available from:
521 <https://github.com/bhklab/uhn-radcure-challenge>.
- 522 67. Lombardo, E. *dl_based_prognosis*. 2020; Available from: [https://gitlab.physik.uni-muenchen.de/LDAP_ag-](https://gitlab.physik.uni-muenchen.de/LDAP_ag-E2ERadiomics/dl_based_prognosis)
523 [E2ERadiomics/dl_based_prognosis](https://gitlab.physik.uni-muenchen.de/LDAP_ag-E2ERadiomics/dl_based_prognosis).
- 524 68. *cnn-hnsc*. [cited 2024 August 25]; Available from: <https://github.com/oncoray/cnn-hnsc>.

525

526

527

528

529

530

531

532

533

534

535

536

537

538

539

540

541

542

543

544

545 **Statements & Declarations**

546 **Funding**

547 Authors acknowledge financial support from the NWO funded personal health Train for RAdiation oncology in India and
548 the Netherlands - TRAIN project (629-002-212), the Stichting Hanarth Fonds and the Proton Therapy Research
549 Infrastructure – ProTRAIT project (RUG 2017-8254) which is supported by the Dutch Cancer Society. Additionally,
550 HMTT acknowledges the DBT/Wellcome Trust India Alliance Early Career Fellowship [Grant number:
551 IA/E/18/1/504306] for the support.

552 **Competing Interest**

553 Andre Dekker is a founder, shareholder and employee of Medical Data Works B.V. All other authors do not have any
554 competing interest.

555 **Author Contributions**

556 Varsha Gouthamchand (VG), Louise AF Fonseca (LAAF), Leonard Wee (LW), and Hannah Mary Thomas T (HMTT)
557 contributed to the literature search, study design, eligibility assessment, and manuscript editing. Frank JP Hoebbers (FH),
558 as the senior clinician, reviewed the draft to ensure clinical relevance. Rianne Fijten (RF), Andre Dekker (AD), Leonard
559 Wee (LW), and Hannah Mary Thomas T (HMTT) supervised the review process. All authors have read and approved the
560 final version of the manuscript for publication.

561 **Ethics approval**

562 This review was performed in accordance with the relevant ethical guidelines and regulations of the institution, and did
563 not require any informed consent.

564 **Consent to publish**

565 Not applicable

566

567 **Tables**

568 **Table 1 Summary of general study characteristics.**

Reference	Type	Imaging modality	Cohort description and sample sizes	Type(s) of image features	Primary treatment	Study design	Potential comparators	Software access (or code base)
Aerts:2014 [39]	OPC, LC	Treatment planning FDG-PET-	Training datasets comprised of only NSCLC cases, validation in separate HNSCC datasets.	HC	The majority were RT only, the	R	TNM staging, HPV status and tumor	Radiomics; in-house code (Matlab base)

		CT Treatment planning CT	Validation 1: OPC (64%), LC (36%) n=136, all stages, HPV negative 72%, from 2004. Validation 2: OPC (100%) n=95, all stages, 2000-2006, HPV negative 81%.		remaining CRT.		volume	Code access n.r.
Bogowicz:2020 [45]	OPC, HPC, LC, Oral Cavity	Treatment planning CECT	Six institutional cohorts, all stages, years n.r Cohorts were used for internal-external validation according to Steyerberg method. Cohort sizes ranged from n=100 up to n=441, HPV negative rates ranged from 11% to 100%.	HC	Definitive RT or CRT	R	None; only compared radiomics models	Z-Rad (Python base) Code access - [61]
Folkert:2017 [35]	OPC	Treatment planning FDG-PET	Training: One academic center in USA (n=174) Only stage III-IV, 2002-2009, HPV status n.r. Validation: Independent academic center in USA (n=65) Only stage III-IV, 2003-2009, HPV status n.r.	HC	Definitive CRT	R	Compared with multivariable clinical models.	Computational Environment for Radiotherapy Research (CERR), (Matlab base) Code access n.r.
Ger:2019 [34]	OPC, LC	Treatment planning CECT F18-FDG-PET	Model development and validation of CT and PET features separately. CT - Training: (n=377) All stages, 2004-2013, HPV negative 41%. PET - Training: (n=345), all stages, 2004-2013, HPV negative 40%. Validation: HNSCC datasets from Aerts:2014.	HC	Definitive RT or CRT	R	CT radiomics: Tumor volume and HPV status. PET radiomics: HPV status	IBEX (Matlab base)
Goncalves:2022 [33]	OPC, HPC, NPC, LC	Treatment planning CT	Training 1: Two centers in Montreal (n=124), all stages, 2006-2014, HPV negative 28%. Validation 1: Two other centers in Montreal (n=70), stage II-IV, 2008-2014, HPV negative 1%.	HC	Definitive RT or CRT	R	Compared with multivariable clinical model.	PyRadiomics (Python base) Code - [62]
Keek:2020 [46]	OPC, HPC, LC	CECT	Training: (n=301) from 4 Dutch academic hospitals, all stages, years n.r, HPV negative 69%. Validation (n=143) from 4 other European academic hospitals, all stages, years n.r , HPV negative 45%.	HC	Definitive CRT	R	Compared with multivariable clinical features.	RadiomiX toolbox (OncoRadiomics)
Kim:2022 [31]	OPC, HPC, LC, Oral Cavity, NPC, Paranasal sinus	MRI T2WI, CE-T1WI	Training: Single Dutch academic center (n=161) all stages, 2014-2019, HPV n.r. Internal Validation: Same as training center (n=54) all stages, 2014-2019, HPV n.r. External Validation: University hospital in Korea (n=70), all stages, 2014-2019, HPV n.r.	HC	Definitive RT or CRT. Surgery followed by RT	R	Compared with multivariable clinical model.	Radiomics; in-house code (Matlab base) Code access - [63]
Leger:2017 [44]	Oral cavity, OPC, HPC, LC	Non-CECT only	Training: Combined from two centers (n=213), 1999-2011, all stages, HPV negative 77%. Validation: Combined from two centers (n=80), all stages, 2005-2012, HPV negative 49%.	HC	All definitive CRT	R	None; only compared radiomics models	Radiomics; in-house code (Python base) Code access n.r.

Leijenaar:2015 [36]	OPC	Treatment planning CT	Only external validation of previously published model (ie Aerts:2014). OPC (100%) n=542, all stages, 2005 – 2010, HPV negative 24%,	HC	About equal proportion of definitive RT and CRT.	R	TNM staging, HPV status and tumor volume	As in Aerts:2014
Lv:2020 [43]	OPC, HPC, NPC, LC	Treatment planning FDG-PET-CT	(n=296) Subset of data from Vallières:2017, from the 4 cancer centers in Montreal, Canada.	HC	Definitive RT or CRT	R	Compared with multivariable clinical model.	Radiomics Analysis (SERA) package (Matlab base)
Meneghetti:2021 [47]	OPC, HPC, LC, Oral Cavity	Treatment planning CT	Training: Two German centers combined (n=233) stages II-IV, 2005-2013, HPV negative 61%. Validation: One of the German centers in training combined with two other independent centers (n=85) stage III-IV, 2005-2013, HPV negative 61%.	HC	Definitive CRT	R	Compared with tumor volume	MIRP by Oncoray Code access - [64]
Parmar:2015a [41]	OPC, LC	Treatment planning FDG-PET-CT Treatment planning CT	HNSCC datasets from Aerts:2014	HC	The majority were RT only, the remaining CRT.	R	None; used radiomics to predict stage and HPV status	As in Aerts:2014
Parmar:2015b [42]	OPC, LC	Treatment planning FDG-PET-CT Treatment planning CT	HNSCC datasets from Aerts:2014	HC	Majority RT only, Remaining CRT.	R	None; compared different radiomics models	As in Aerts:2014
Vallières:2017 [40]	OPC, HPC, NPC, LC	Treatment planning FDG-PET-CT	Training: Two centers HGJ and CHUS in Montreal (n=194), all stages, 2006-2014, HPV negative 20%. Validation: Two other independent centers HMR and CHUM in Montreal (n=106), stage II-IV, 2008-2014, HPV negative 3%.	HC	Definitive RT or CRT	R	Compared with multivariable clinical models.	Radiomics; in-house code (Matlab base) Code access n.r.
Zhai: 2021 [32]	Oral cavity, OPC, NPC, HPC, LC	CECT	Only external validation of previously published model (Zhai:2020). Training: One academic center (n=165), all stages, 2007-2016, HPV negative 29%. Validation 1: Same center as training (n=112), all stages, 2007-2016, HPV negative 27%. Validation 2: Independent academic center (n=113), all stages, 2007-2016, HPV negative 24%.	HC	Definitive RT or CRT; Excluding elective neck dissection immediately following RT	R	Compared with multivariable clinical models.	Radiomics; in-house code (Matlab base) Code access n.r.
Zhou:2020 [52]	OPC, HPC, NPC, LC	Pre-treatment FDG-PET/CT	Subset (n=188) of the cohorts used previously by Vallières:2017	HC	Definitive RT or CRT	R	None	SSAE, IMIA and IMIA-II Code access n.r.
Cheng:2021 [30]	OPC	FDG-PET	Training: Single institution (n=268), 2006-2017, HPV negative 79%. Validation 1: (n=353) combined	DL	Definitive RT or	R	Compared with multivariable clinical	3D UNet model and 3D

			datasets from Vallières:2017, Ger:2019 and Aerts:2014, all stages, 2003-2014, overall HPV negative 30%. Validation 2: Chinese academic centre (n=31), all stages, 2011-2013, HPV negative 3%.		CRT Included some surgery patients		model.	ConvCox Code access - [65]
Diamant:2019 [48]	OPC, HPC, NPC, LC	Treatment planning FDG-PET-CT	Same training and validation datasets as Vallieres:2017	DL	Definitive RT or CRT	R	Compared DL to Vallieres radiomics model(s).	CNN in Keras with Tensorflow Code access n.r.
Fujima:2021 [37]	OPC	Treatment planning FDG-PET	Training: Single institution (n=102), all stages, 2007-2017, HPV negative 25%. Validation: Another independent institution (n=52), all stages, 2007-2017, HPV negative 15%.	DL	Definitive RT or CRT	R	Compared with multivariable clinical models.	Several 2D CNNs: AlexNet, GoogLeNet Inception v3 and ResNet-101 Code access n.r.
Kazmierski:2023 [38]	Multiple sites incl. OPC, LC, NPC, Oral Cavity	Treatment planning CECT	Training: Single institution dataset (n=2552) from Toronto, Canada; all stages, years n.r., HPV negative 17%. Validation 1: First validation dataset as was used by Aerts:2014. Validation 2: Same institution as Ger:2019 (n=444), all stages, years n.r., HPV negative 9%. Validation 3: Private Polish dataset (n=298), all stages, years n.r., HPV negative 6%.	HC and DL	Definitive RT or CRT	R	Compared with multivariable clinical model and tumour volume.	PyRadiomics and DeepMTLR Code access-[66]
Le WT:2022 [51]	OPC, HPC, NPC, LC	Treatment planning FDG-PET-CT	Same training and validation datasets as Vallieres:2017. External Validation: New dataset from one of the Montreal hospitals (n=371), all stages, 2011-2019, HPV negative 24%.	DL	Definitive RT or CRT	R	Compared with multivariable clinical model.	PreSASNet Code access n.r.
Lombardo:2021 [49]	Various subtypes of HNSCC	CT	Training: Dataset used by Diamant:2019. Validation 1: The validation set #1 used by Aerts:2014. Validation 2: Subset of Canadian data previously used by Kazmierski:2023. Validation 3: Single center set from Italy (n=110), all stages, 2017-2019, HPV n.r.	DL	Definitive RT or CRT	R	Compared with multivariable clinical model.	2D and 3D CNNs [67]
Starke:2020 [50]	Oral cavity, OPC, HPC, LC	Treatment planning CT	Same training and validation datasets as used by Leger:2017.	DL	Definitive CRT	R	Compared with multivariable clinical model.	2D and 3D CNNs Code access - [68]

569
570
571
572

Abbreviations: CECT – Contrast Enhanced Computed Tomography; CNN – Convolutional Neural Network; CRT – Chemoradiotherapy; CT - Computed Tomography; DM – Distant Metastasis; DL – Deep Learning; FDG-PET - Fluorodeoxyglucose-Positron Emission Tomography; HC – Hand crafted; HNSCC - Head-and-Neck Squamous Cell Carcinoma; HPC – Hypopharyngeal Carcinoma; HPV – Human Papillomavirus; IMRT - Intensity-modulated radiation therapy (which also includes volume-modulated arc therapy and helical tomotherapy); LC - Laryngeal Carcinoma; LRC – Loco-

573
574
575

Regional Tumor Control; LRR – Locoregional Recurrence; LF – Local Failure; NPC – Nasopharyngeal Carcinoma; n.r - not reported; NSCLC - Non-Small Cell Lung Cancer; OPC – Oropharyngeal Carcinoma; OS – Overall Survival; R – Retrospective; RT – Radiotherapy;

576 **Table 2 Summary of model discriminative performances.**

Reference	Primary outcome	Event to sample size ratio	Model simplification/reduction	Type of model	Discriminative performance in validation dataset(s)	Added value of radiomics/DL
Aerts:2014 [39]	OS	n.r (231)	Stability ranks for feature selection	Multivariable Cox proportional hazards regression	Radiomics: C-index 0.69 Clinical C-index 0.68 - 0.69 Combined C-index 0.69 - 0.70	Comparable to tumor volume and TNM stage.
Bogowicz:2020 [45]	2yr OS	5 datasets with LOOCV - 68% (1064)	Hierarchical clustering and univariate logistic regression	Multivariable Logistic regression	Centralized AUC 0.69 - 0.82 Distributed AUC 0.73 - 0.80	n.r.
Folkert:2017 [35]	OS, LF, DM	Train: 27% OS 7% LF, 19% DM (174) Validate: 48% ACM, 15% LF, 17% DM (65)	Forward feature selection	Multivariable Logistic regression (Tested stratification in Kaplan-Meier but did not report any time-to-event discrimination metric)	OS: AUC 0.60 LF: AUC 0.68 DM: AUC 0.65	n.r.
Ger:2019 [34]	OS	CT Train 26% (377), Validate 21% (349) PET Train 22% (345), Validate 15% (341)	Clinical variables– Forward feature selection using Akaike information criteria (AIC) > 2 Radiomics - LASSO regression	Multivariable Cox proportional hazards regression dichotomized at 3 years	OS: CT AUC 0.72 OS: PET AUC 0.59	Tumor volume alone was superior to radiomics and clinical model PET covariates not associated with OS
Goncalves:2022 [33]	LRR, DM, OS	Training: LRR – 14%, DM – 18%, OS – 18% (125) Validation: LRR – 20%, DM – 19%, OS – 27% (70)	Feature importance (XGBoost)	Multiple machine learning algorithms Multilayer perceptron, extreme gradient boosting, logistic regression, random forest, and decision trees	Best performing model XGBoost. Combined model LRR: AUC 0.74 DM: AUC 0.84 OS: AUC 0.91 Radiomics only LRR: AUC 0.58 DM: AUC 0.84 OS: AUC 0.82	Combined model outperforms radiomics model
Keek:2020 [46]	OS, LRR, DM	Training: n.r. (301) Validation: n.r. (143)	Relative feature importance (random survival forest (RSF))	Multivariable Cox proportional hazards regression and random survival forest	Clinical model OS: C-index 0.77 (RSF) LRR: C-index 0.79 (RSF) DM: C-index 0.84 (RSF) Radiomics model OS: C-index 0.62 (Cox) LRR: C-index 0.59 (RSF)	Clinical models outperform radiomics models

					DM: C-index 0.56 (Cox)	
Kim:2022 [31]	LR	Training: 57% (161) Validation: 67% (54) External validation: 49% (70)	Spearman's correlation + LASSO logistic model	Multivariable Logistic regression	Radiomics AUC 0.77 (CI 0.40–0.88) Clinical AUC 0.53 (CI 0.39–0.67)	Radiomics model outperforms clinical model
Leger:2017 [44]	LRC, OS	Training: LRC- 40%, OS-56% (213) Validation: LRC-32%, OS-65% (80)	13 feature selection methods	Multiple machine learning algorithms	LRC: C-index 0.71 (Random Forest) OS: C-index 0.64 (Boosting Tree)	n.r.
Leijenaar:2015 [36]	OS	n.r. (542)	Not applicable	Multivariable Cox proportional hazards regression	OS: C-index 0.65	n.r.
Lv:2020 [43]	RFS, MFS, OS	Training: RFS - 14%, MFS - 14%, OS - 17% (190) Validation: RFS - 8%, MFS - 4%, OS - 7% (106)	Univariate cox analysis + 3-fold cross validation + Spearman's correlation	Multivariable Cox proportional hazards regression	Combined clinical and radiomics RFS C-index: 0.54 - 0.60 MFS C-index: 0.61 - 0.71 OS C-index: 0.60 - 0.65 Clinical FS C-index: 0.58 MFS C-index: 0.61 OS C-index: 0.62	Combined model outperformed clinical model
Meneghetti:2021 [47]	LRC	Training: n.r. (233) Validation: n.r.(85)	3 feature-selection algorithms: Spearman's correlation, minimal redundancy maximum relevance and regularized Cox regression	Multivariable Cox proportional hazards regression	Combined C-index 0.66 (0.55-0.75) Clinical C-index 0.56 [0.49-0.62]	Combined model outperformed clinical model
Parmar:2015a [41]	OS	Training n.r. (136) Validation n.r. (95)	Unsupervised clustering methods	Multivariable Cox proportional hazards regression	Radiomics C-index 0.68 Clinical C-index 0.63	Radiomics model outperforms clinical model
Parmar:2015b [42]	3yr OS	Training 37% (101) Validation 34% (95)	13 feature selection methods	Multiple machine learning algorithms	AUC 0.61 - 0.69	n.r.
Vallières:2017 [40]	OS, LR, DM	Training: LR - 15%, DM - 13%, OS - 16% (194) Validation: LR - 15%, DM - 13%, OS - 23% (106)	Stepwise forward feature selection + Spearman rank correlation + maximal information coefficient	Clinical: Random forest classifier Radiomics: Cox regression model	Clinical OS: AUC 0.78 C-Index 0.76 LR: AUC 0.72 C-Index 0.69 DM: AUC 0.55 C-Index 0.60 Combined OS: AUC 0.74 C-index 0.71 LR: AUC 0.69 C-index	Combined models superior to clinical models for OS and DM

					0.67; DM: AUC 0.86 C-index 0.88;	
Zhai: 2021 [32]	NF	5.3% (113)	N.A	Multivariable Cox proportional hazards regression	Radiomics C-Index 0.71 Clinical C-Index 0.57 Combined C-Index 0.71	External validation confirms superiority of combined model.
Zhou:2020 [52]	DM	16% (188)	Deep learning with stacked sparse autoencoder	Single objective (SO), multi-objective (MO) and multi-faceted models capable of combining many base classifiers (M-radiomics)	SO AUC 0.81 MO AUC 0.76 M-radiomics 0.84	n.r
Cheng:2021 [30]	OS	Training: 50% (268) Validation: 77% (384)	N.A	3D CNN-based cox proportional hazards (ConvCox)	Clinical AUC 0.75 (CI 0.65- 0.84) Combined AUC 0.80 (CI 0.73-0.87)	Combined model outperforms clinical model
Diamant:2019 [48]	DM, LRF, OS	Same as Vallieres:2017	N.A	CNN discriminator	CNN DM: AUC 0.86 - 0.88 LRF: AUC 0.50 - 0.65 OS: AUC 0.65 - 0.70 Handcrafted Radiomics and CNN DM: AUC 0.92 LRF: AUC n.r OS: AUC n.r	Handcrafted radiomics and CNN model outperforms CNN model for DM
Fujima:2021 [37]	LF, PFS	Training: n.r (n=102) Validation: n.r (n=52)	N.A	Deep learning based multivariable Cox proportional hazards regression	Clinical LF: AUC 0.59 - 0.74 DL LF: AUC 0.61 - 0.85	Deep learning models outperforms clinical models
Kazmierski:2023 [38]	OS	Training: 59% (1802) Test: 81% (750) Validation: 67% (872)	Maximum relevance-minimum redundancy method (MRMR)	Multiple machine learning and Deep learning models using clinical (EMR, Volume) and/or imaging features	DL + Clinical AUC 0.72 - 0.82 (Best model EMR + Volume) Radiomics + Clinical AUC 0.72 - 0.82 (Best model EMR + Volume)	Similar performances for machine and deep learning models; Deep Clinical models outperform deep imaging models
Le WT:2022 [51]	DM, LR, OS	Same as Vallieres:2017	N.A	Multiple Deep learning Models compared with proposed Pseudo-volumetric convolutional neural network with deep preprocessor module and self-attention (PreSANet)	DM: AUC 0.67 [CI 0.61–0.73] LR: AUC 0.68 [CI 0.65–0.72] OS: AUC 0.68 [CI 0.65–0.71]	Proposed Deep learning model outperformed other reported models for LR and OS
Lombardo:2021 [49]	DM	Training: 13% (294) Test: 12% (744)	N.A	Comparing 2D and 3D CNN	2D CNN + Clinical AUC 0.66-0.89 3D CNN + Clinical AUC 0.66-0.87	2D CNN + clinical outperformed 3D deep learning models

Starke:2020 [50]	LRC	Same as Leger:2017	N.A	Comparing 2D and 3D CNN	Clinical C-Index 0.39 2D CNN + Volume C-Index 0.40 3D CNN C-Index 0.31	2D CNN + volume outperformed clinical and other deep learning models
---------------------	-----	-----------------------	-----	----------------------------	---	---

577
578
579

Abbreviations: CNN - convolutional neural network; DL - deep learning; DM - Distant Metastasis; LF - Local Failure; LR - Local Recurrence; LRC - Loco-regional tumor control; LRR - Locoregional Recurrence; MFS- Metastatic Free survival; NF - Nodal Failure; OS - Overall Survival; PFS - Progression-Free Survival; RFS - Recurrence-free Survival; RT - Radiotherapy; T1 WI-T1 weighted image; T2 WI-T2 weighted image;

Identification

Records identified through
database searching
(n = 1610)

Additional records identified
through other sources
(n = 237)

Screening

Records after duplicates removed
(n = 1718)

Records screened
(n = 1718)

Records excluded
(n = 1598)

Eligibility

Full-text articles assessed
for eligibility
(n = 120)

Full-text articles excluded,
with reasons
(n = 97)

Included

Studies included in
systematic review
(n = 23)

Studies included in
quantitative synthesis
(meta-analysis)
(n = 23)

