Title Page

Title: Artificial Intelligence Uncertainty Quantification in Radiotherapy Applications - A Scoping Review

Kareem A. Wahid[a,b]*, Zaphanlene Y. Kaffey[b]*, David P. Farris[c], Laia Humbert-Vidan[b], Amy C. Moreno[b], Mathis Rasmussen[d], Jintao Ren[d], Mohamed A. Naser[b], Tucker J. Netherton[e], Stine Korreman[d], Guha Balakrishnan[f], Clifton D. Fuller[b], David Fuentes[a]**, Michael J. Dohopolski[g]**

[a]Department of Imaging Physics, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA.
[b]Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA.
[c]Research Medical Library, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA.
[d]Department of Oncology, Aarhus University Hospital, Denmark.
[e]Department of Radiation Physics, University of Texas MD Anderson Cancer Center, Houston, TX, USA.
[f]Rice University, Houston, TX, USA.
[g]Department of Radiation Oncology, The University of Texas Southwestern Medical Center, Dallas, Texas, USA.

* co-first authors
** co-corresponding authors

Corresponding authors contact info: David Fuentes: dtfuentes@mdanderson.org; Michael Dohopolski: michael.dohopolski@utsouthwestern.edu.

Funding Statement: KAW was supported by an Image Guided Cancer Therapy (IGCT) T32 Training Program Fellowship from T32CA261856. ZYK's time was supported by a doctoral fellowship from the Cancer Prevention Research Institute of Texas grant #RP210042. MAN receives funding from NIH National Institute of Dental and Craniofacial Research (NIDCR) Grant (R03DE033550). CDF received/receives unrelated funding and salary support from: NIH National Institute of Dental and Craniofacial Research (NIDCR) Academic Industrial Partnership Grant (R01DE028290) and the Administrative Supplement to Support Collaborations to Improve AIML-Readiness of NIH-Supported Data (R01DE028290-04S2); NIDCR Establishing Outcome Measures for Clinical Studies of Oral and Craniofacial Diseases and Conditions award (R01DE025248); NSF/NIH Interagency Smart and Connected Health (SCH) Program (R01CA257814); NIH National Institute of Biomedical Imaging and Bioengineering (NIBIB) Research Education Programs for Residents and Clinical Fellows Grant (R25EB025787); NIH NIDCR Exploratory/Developmental Research Grant Program (R21DE031082); NIH/NCI Cancer Center Support Grant (CCSG) Pilot Research Program Award from the UT MD Anderson CCSG Radiation Oncology and Cancer Imaging Program (P30CA016672); Patient-Centered Outcomes Research Institute (PCS-1609-36195) sub-award from Princess Margaret Hospital; National Science Foundation (NSF) Division of Civil, Mechanical, and Manufacturing Innovation (CMMI) grant (NSF 1933369). CDF receives grant and infrastructure support from MD Anderson Cancer Center via: the Charles and Daneen Stiefel Center for Head and Neck Cancer Oropharyngeal Cancer Research Program; the Program in Image-guided Cancer Therapy; and the NIH/NCI Cancer Center Support Grant

(CCSG) Radiation Oncology and Cancer Imaging Program (P30CA016672). ACM received/receives funding and salary support from: NIDCR (K01DE030524, R21DE031082), the NIH National Cancer Institute (K12CA088084), and the University of Texas MD Anderson Cancer Center Charles and Daneen Stiefel Center for Head and Neck Cancer Oropharyngeal Cancer Research Program. DF was supported by R01CA195524 and NSF-2111147. Disclaimer: The content is solely the responsibility of the authors and does not necessarily represent the official views of the funders.

Conflicts of Interest: KAW serves as an Editorial Board Member for Physics and Imaging in Radiation Oncology. CDF has received travel, speaker honoraria and/or registration fee waiver unrelated to this project from: The American Association for Physicists in Medicine; the University of Alabama-Birmingham; The American Society for Clinical Oncology; The Royal Australian and New Zealand College of Radiologists; The American Society for Radiation Oncology; The Radiological Society of North America; and The European Society for Radiation Oncology.

Acknowledgments: TJN would like to acknowledge the support of the NIH Loan Repayment Award Program.

Declaration of generative AI and AI-assisted technologies in the writing process: During the preparation of this work, the authors used ChatGPT (GPT-4 architecture) to improve the grammatical accuracy and semantic structure of portions of the text. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Main Text

Abstract

Background/purpose: The use of artificial intelligence (AI) in radiotherapy (RT) is expanding rapidly. However, there exists a notable lack of clinician trust in AI models, underscoring the need for effective uncertainty quantification (UQ) methods. The purpose of this study was to scope existing literature related to UQ in RT, identify areas of improvement, and determine future directions.

Methods: We followed the PRISMA-ScR scoping review reporting guidelines. We utilized the population (human cancer patients), concept (utilization of AI UQ), context (radiotherapy applications) framework to structure our search and screening process. We conducted a systematic search spanning seven databases, supplemented by manual curation, up to January 2024. Our search yielded a total of 8980 articles for initial review. Manuscript screening and data extraction was performed in Covidence. Data extraction categories included general study characteristics, RT characteristics, AI characteristics, and UQ characteristics.

Results: We identified 56 articles published from 2015-2024. 10 domains of RT applications were represented; most studies evaluated auto-contouring (50%), followed by image-synthesis (13%), and multiple applications simultaneously (11%). 12 disease sites were represented, with head and neck cancer being the most common disease site independent of application space (32%). Imaging data was used in 91% of studies, while only 13% incorporated RT dose information. Most studies focused on failure detection as the main application of UQ (60%), with Monte Carlo dropout being the most commonly implemented UQ method (32%) followed by ensembling (16%). 55% of studies did not share code or datasets.

Conclusion: Our review revealed a lack of diversity in UQ for RT applications beyond auto-contouring. Moreover, there was a clear need to study additional UQ methods, such as conformal prediction. Our results may incentivize the development of guidelines for reporting and implementation of UQ in RT.

Introduction

Artificial intelligence (AI) in healthcare has become increasingly important due to its potential to enhance diagnosis, treatment, and prognostic prediction [1]. A significant obstacle to the clinical implementation of AI that is receiving growing attention is a relative absence of model uncertainty quantification (UQ) [2]. The ability of an AI model to characterize and communicate its uncertainty, in other words, learning when to say “I don’t know” [3], could enhance clinician trust and facilitate the integration of AI into clinical practice [4–6].

Radiotherapy (RT) is a fundamental pillar of cancer treatment used in approximately 50% of all malignancies [7]. Due to the highly quantitative and structured nature of the RT clinical workflow, AI-based methodologies — namely, machine learning (ML) and deep learning (DL) — have been increasingly investigated to automate and improve a variety of tasks [8]. Advances in DL algorithms trained on increasingly larger, diverse datasets have allowed for impressive performance in a variety of RT-related applications such as image synthesis [9], registration [10], contouring [11], dose prediction [12], and outcome prediction [13–15]. However, despite the impressive performance of these models in research studies, to date there are relatively few standard AI-based tools that are routinely used in RT workflows. This hesitation could be partially attributed to insufficient clinician trust [16,17]. Enhanced UQ could bridge this trust gap, fostering greater confidence in AI applications within the RT field.

Conventionally there are two types of uncertainty: aleatoric and epistemic [18]. Aleatoric uncertainty arises from the noise inherent in the data. An example is the inherent variation in contour “ground truth” among radiation oncologists, each can be “right” but likely slightly different [19]. Epistemic uncertainty stems from incomplete information. For instance, a head and neck tumor contouring model may have limited exposure to certain rare malignancies (e.g., salivary gland cancer) and may generate poor contours with high epistemic uncertainty as these cases were underrepresented in model development. Models trained outside of medicine are often trained on datasets with >1 million samples [20]. Medical datasets, especially in RT, are considerably smaller [21], often ranging from hundreds to thousands of patient samples. Thus, epistemic uncertainty estimation would be particularly important for RT model development. Together, aleatoric and epistemic uncertainty account for the total predictive uncertainty [22]. Illustrative figures related to aleatoric and epistemic uncertainty concepts are shown in **Appendix A Figure A1**.

Within the UQ literature, there exists several methods for providing estimates of uncertainty. Contemporary methods for estimating uncertainty in ML often adopt a Bayesian perspective, treating model predictions as probability distributions rather than single point values. For instance, when predicting if a patient will develop xerostomia after radiation therapy, the model might output an 80% probability instead of simply stating "yes" or "no". These probabilistic measures could enable safer model deployment in various clinical applications [22]. For example, UQ could be used in auto-segmentation for failure detection, flagging cases with a low probability of an accurate segmentation (i.e., high uncertainty) for additional clinical review. UQ

methods such as Monte Carlo dropout and ensembles, which are suggested to be grounded in Bayesian principles [23,24], have surged in popularity in recent years [25]. However, emerging techniques, such as conformal prediction [26], are increasingly drawing on more traditional statistical methodologies.

Finally, worthy of note is that UQ has historically been closely linked to calibration, which measures the agreement between predicted probabilities and observed frequencies. Large-scale ML models — particularly DL models with numerous parameters — often show poor calibration, with output probabilities being higher than observed probabilities, subsequently leading to overconfident predictions [27]. UQ methods can help quantify and mitigate poor calibration; for example Monte Carlo dropout and ensembles often inherently improve confidence calibration [28,29]. For readers interested in more technical reviews on UQ concepts generally and specific to RT, we refer to comprehensive narrative works by Hullermeier & Waegeman et al. [18] and van den Berg & Meliadò [30], respectively.

While previous systematic and scoping reviews have covered the topics of UQ in healthcare generally [25,31] and in relation to medical imaging [32–34], these studies lacked any explicit focus on RT-related applications. Therefore, we conducted this scoping review to synthesize current trends for UQ in RT and provide an outlook for the future of this important research area for clinicians and researchers. An overview of our study is illustrated in **Appendix A Figure A2**.

Materials and Methods

This scoping review was conducted in line with the reporting guidelines of Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) [35]. The pre-registration for this scoping review was performed using the Open Science Foundation Generalized Systematic Review Registration template and can be found online (https://doi.org/10.17605/OSF.IO/E3DQG). We utilized Covidence [36] — a standardized web-based literature review collaboration software platform — to perform all initial study screening and data extraction.

Eligibility Criteria

This scoping review was conducted to summarize the state of literature that implemented AI UQ for RT. We utilized the population, concept, context (PCC) framework to develop a focus question as recommended by the Joanna Briggs Institute Scoping Review Methodology Group [37]. Population was defined as human patients undergoing RT for cancer treatment, concept was defined as utilization of AI and UQ, and context was defined as RT applications (e.g., image acquisition and synthesis, tumor and organ at risk contouring, dose prediction, outcome

prediction, etc.). Additional details on the PCC eligibility criteria and its integration into the search strategy are discussed in **Appendix B**.

Search Strategy

A medical research librarian (D.P.F.) searched MEDLINE (Ovid), Embase (Ovid), PubMed (NLM), Cochrane Library (Wiley), and Web of Science Core Collection (Clarivate) from inception to November 17, 2023, with the search executed on November 20, 2023. A supplementary search of Web of Science Preprint Citation Index (Clarivate) and Google Scholar (Alphabet Inc.) from inception to December 12, 2023 was executed on December 13, 2023 in order to adequately query gray literature such as preprints and conference proceedings. After consultation with the research team, the librarian developed and tailored the search strategy to each database and selected controlled vocabulary (MeSH and Emtree) and natural language terms for the concepts of AI, UQ, and RT. No language, publication date, or other limiters or published search hedges were used. A total of 8974 results were retrieved from the five databases including an original set of 9 key articles supplied by the research team (MEDLINE = 1084; Embase = 1708; PubMed = 1154; Cochrane = 42; Web of Science Core Collection = 4358; Web of Science Preprint Citation Index = 428; Google Scholar = 200). The full search strategy inputs for each database is available in **Appendix C**. Notably, we incorporated 6 additional manuscripts that were not captured in the initial eligibility screening post-hoc via manual citation searching up to January 19, 2024; these manuscripts were principally added because they were formally indexed after the initial search date and were deemed relevant to ensure a more up-to-date review. Search results were uploaded to Covidence; after deduplication, 6017 unique results were identified for eligibility screening. The full PRISMA-ScR flow diagram is shown in **Appendix A Figure A3**.

Study Selection

Initial screening to ensure studies broadly fit within our defined PCC framework was performed by 2 independent reviewers (K.A.W., Z.Y.K.) based on titles and abstracts. Disagreements were mediated by an independent third senior reviewer (M.J.D.). All disagreements were discussed in a group setting with the 3 reviewers; in cases where no consensus was reached the decision of the senior reviewer was implemented. A second full text review of these articles was performed to ensure all inclusion criteria were fully satisfied (additional details in **Appendix B**). Only full English-language preprints, conference proceedings, and standard peer-reviewed publications were included for this study; conference abstracts were excluded. Conference proceedings and preprints were deemed appropriate for inclusion due to their ubiquitous nature in computational fields [38]. Preclinical and animal studies were not included in this review. 56 articles were ultimately selected for final inclusion (**Appendix A Figure A3**). All screening was performed through the Covidence online platform.

Data Extraction

Two reviewers extracted data from the final manuscripts (K.A.W., Z.Y.K.). All extractions were cross-checked by both reviewers and a final third reviewer (M.J.D.) when disagreements occurred. Data were initially extracted using a template generated in Covidence, focusing on four categories: general study characteristics, RT characteristics, AI characteristics, and UQ characteristics. General study characteristics included manuscript type, publication year, geographic location of the study authors, and code/data availability. RT characteristics included intended RT application space (e.g., contouring, dose planning, etc.), specific data types used (e.g., CT, MRI, etc.), and patient cancer type. AI characteristics included algorithmic approach, training/validation/testing sample sizes, and properties of the validation/testing (e.g., separate set, cross-validation, etc.). AI characteristics were adapted from existing related guidelines including TRIPOD [39] and CLAIM [40]. UQ attributes included application category, method type, evaluation metrics, self-described uncertainty type (i.e., aleatoric vs. epistemic), and use of quantitative or qualitative evaluation methods. UQ application categories and definitions were adapted from Kahl et al. [22] and Lambert et al. [34]. Additional specific considerations for each category in the data extraction process are described in detail in **Appendix B**.

Analysis

The final extracted data were analyzed using Python v. 3.10. Descriptive statistics and visual plots were generated using the pandas, seaborn, matplotlib, numpy, geopandas, and squarify Python libraries. We also compared the overlap of extracted publications in our study and publications extracted in previous systematic and scoping reviews in similar topic domains (i.e., UQ in medical applications). To accomplish this, we compiled a comprehensive list of all publications referenced in these studies during the data extraction process along with their respective titles and digital object identifiers (DOIs). Initially, we attempted to automatically match DOIs from studies in our scoping review with those in the existing literature. If no DOI match was found, we proceeded to automatically compare titles using the difflib Python library, setting a sequence match ratio threshold of at least 0.75. All identified matches were then subsequently manually verified.

Data and Code Availability

A CSV file containing the final studies and corresponding extracted data for this scoping review are made publicly available through Figshare (doi: 10.6084/m9.figshare.25535017). All Python code used in the analysis can be found on Github (URL: https://github.com/kwahid/RT_UQ_scoping_review/tree/main).

Results

Table 1 presents an overview of the extracted data from the final 56 manuscripts included in this review.

Table 1. Comprehensive listing of final studies analyzed in this scoping review. Data of interest were split into four main categories: general study characteristics, radiotherapy (RT) characteristics, artificial intelligence (AI) characteristics, and uncertainty quantification (UQ) characteristics. Rows are ordered by ascending publication year and study ID. Additional abbreviations: organ at risk = OAR, machine learning = ML, cross validation = CV, failure detection = FD, active learning = AL, ambiguity modeling = AM, out of distribution detection = OODD, Gaussian Process = GP, Ensemble = ENS, Other Bayesian = OB, Platt Scaling = PS, MC Dropout = MCD, Test-time Augmentation = TTA, Evidential Deep Learning = EDL, Direct softmax output = DSO.

General Study Characteristics						RT Characteristics				AI Characteristics						UQ Characteristics				
Study ID	Paper type	Year	Location	Data avail.	Code avail.	RT applic.	Image data	Additional data	Cancer type	ML type	Training patients	Valid. patients	Valid. type	Testing patients	Testing type	UQ applic.	UQ type	UQ method	UQ metric	UQ experiments
Bukhari 2015 [41]	Standard pub.	2015	Korea	No	No	Motion tracking	NA	Repirator y trace	Lung	Supervised	31	31	CV	31	Separate set [internal]	FD	Unspecified	GP	Variance-based	Quant.
Lee 2015 [42]	Standard pub.	2015	Canada	No	No	Outcome related	NA	Clinical	Lung	Supervised	53	Unspecified	Unspecified	200	Bootstrap	Calib.	Unspecified	ENS; OB	Other	Quant.
Bragman 2018 [43]	Conf. proc.	2018	UK	No	No	Multiple	Multimodal	OAR	Prostate	Mixed	10	Unspecified	Unspecified	5	Cross validation	FD	Both	MCD; OB	Other	Quant. + Qual.
Jungo 2018a [44]	Conf. proc.	2018	SUI	No	No	Contouring	MRI	Target	Brain	Supervised	25	Unspecified	Unspecified	5	Cross validation	FD	Unspecified	MCD	Entropy-based	Quant. + Qual.
Jungo 2018b [45]	Conf. proc.	2018	SUI	No	No	Contouring	MRI	Target	Brain	Supervised	25	Unspecified	Unspecified	5	Cross validation	FD	Unspecified	MCD; OB	Entropy-based	Quant. + Qual.
Ninomiya 2018 [46]	Conf. proc.	2018	Japan	No	No	Contouring	CT	Target+OAR	Prostate	Supervised	43	Unspecified	Unspecified	1	Cross validation	FD	Unspecified	OB	Probability-based	Quant.
Qin 2018 [47]	Standard pub.	2018	China	No	No	Contouring	CT	OAR	Liver	Supervised	90	Unspecified	Unspecified	10	Cross validation	FD	Unspecified	DSO	Entropy-based	Qual.
Sentker 2018 [48]	Conf. proc.	2018	Germany	Yes	Yes	Image registration	CT	Registration_transforms	Multiple	Supervised	59	Unspecified	Unspecified	16	Separate set [multiple external]	FD	Unspecified	MCD	Variance-based	Quant. + Qual.
Karimi 2019 [49]	Standard pub.	2019	Canada	No	No	Contouring	Ultrasound	Target	Prostate	Supervised	540	Unspecified	Unspecified	135	Cross validation	AL; Calib.	Both	MCD; ENS; PS	Other	Quant. + Qual.
Lipkova 2019 [50]	Standard pub.	2019	Germany	Yes	Yes	Tumor growth modeling	Multimodal	Target	Brain	Supervised	8	Unspecified	Unspecified	8	Other	AL; Calib.	Unspecified	OB	Variance-based	Quant.
Chen 2020 [51]	Standard pub.	2020	China	No	No	Contouring	CT	Target	Breast	Supervised	520	80	Separate set	80	Separate set [internal]	FD	Unspecified	DSO	Other	Quant.
Dohopolski 2020 [52]	Standard pub.	2020	USA	No	No	Nodal classification	PET/CT	Target+OAR	Head and neck	Supervised	Unspecified	Unspecified	Separate set	Unspecified	Separate set [internal]	FD	Both	MCD; TTA	Entropy-based	Quant.
Gustafsson 2020 [53]	Standard pub.	2020	Sweden	Yes	Yes	Contouring	MRI	Fiducial	Prostate	Supervised	326	49	CV	39	Separate set [internal]	FD	Unspecified	DSO	Other	Quant.
Hansch 2020 [54]	Standard pub.	2020	Germany	No	No	Contouring	Multimodal	OAR	Brain	Supervised	27	9	Separate set	9	Separate set [internal]	FD	Unspecified	MCD	Entropy-based	Quant. + Qual.

Maspero 2020 [55]	Standard pub.	2020	NL	No	No	Image synthesis	Multimodal	OAR	Brain	Supervised	30	10	CV	20	Separate set [internal]	FD	Unspecified	ENS	Variance-based	Quant. + Qual.
Nomura 2020 [56]	Standard pub.	2020	Japan	Yes	No	Dose prediction	CT	Dose	Head and neck	Supervised	116	39	Separate set	38	Separate set [internal]	FD	Unspecified	Other	Variance-based	Quant. + Qual.
vanHarten 2020 [57]	Conf. proc.	2020	NL	No	No	Image synthesis	Multimodal	NA	Brain	Unsupervised	30	2	Separate set	74	Separate set [multiple external]	FD	Unspecified	ENS	Other	Quant. + Qual.
Balagopal 2021 [58]	Standard pub.	2021	USA	No	Yes	Contouring	CT	Target+OAR	Prostate	Supervised	290	29	CV	50	Separate set [internal]	FD	Unspecified	MCD	Variance-based	Quant. + Qual.
Dasgupta 2021 [59]	Standard pub.	2021	Canada	No	No	Multiple	MRI	Target	Brain	Supervised	42	8	CV	49	Separate set [internal + external]	Calib.	Unspecified	Other; PS	Probability-based	Quant.
Diao 2021 [60]	Standard pub.	2021	China	Yes	Yes	Contouring	PET/CT	Target	Multiple*	Supervised	99	14	Separate set	28	Separate set [internal]	AL	Both	Other	Entropy-based	Quant. + Qual.
Kajikawa 2021 [61]	Standard pub.	2021	Japan	No	No	Image synthesis	CT	OAR	Lung*	Supervised	60	12	CV	11	Separate set [internal]	FD	Epistemic	MCD	Variance-based	Quant. + Qual.
Lei 2021[62]	Standard pub.	2021	China	Yes	Yes	Contouring	CT	OAR	Head and neck	Supervised	177	Unspecified	Unspecified	48	Separate set [internal]	FD	Unspecified	ENS	Entropy-based; Variance-based	Quant. + Qual.
Luo 2021 [63]	Conf. proc.	2021	China	No	Yes	Contouring	MRI	Target	Head and neck	Supervised	180	20	Separate set	58	Separate set [internal]	AL	Unspecified	Other	Other	Quant. + Qual.
Mei 2021[64]	Standard pub.	2021	China	Yes	Yes	Contouring	CT	Target	Head and neck	Supervised	40	Unspecified	Unspecified	10	Separate set [internal]	FD	Unspecified	ENS	Entropy-based; Variance-based	Quant. + Qual.
Nguyen 2021 [65]	Standard pub.	2021	USA	Yes	No	Dose prediction	CT	Dose	Head and neck	Supervised	200	40	Separate set	100	Separate set [internal]	FD	Unspecified	MCD; ENS	Variance-based	Quant. + Qual.
Nomura 2021 [66]	Standard pub.	2021	USA	Yes	No	Image correction	CT	NA	Head and neck*	Supervised	3	1	Separate set	1	Separate set [internal]	FD; Calib.	Both	ENS; Other	Other	Quant.
Remy 2021 [67]	Standard pub.	2021	Canada	No	No	Motion tracking	MRI	Target	Multiple*	Supervised	10	Unspecified	Unspecified	10	Other	FD	Unspecified	OB	Variance-based	Quant.
vanRooij 2021 [68]	Standard pub.	2021	NL	No	No	Contouring	CT	OAR	Head and neck	Supervised	Unspecified	Unspecified	Separate set	Unspecified	Separate set [internal]	FD; Calib.	Unspecified	MCD	Probability-based	Quant. + Qual.
Zhang 2021 [69]	Standard pub.	2021	China	Yes	No	Contouring	CT	Target+OAR	Lung	Supervised	48	10	CV	28	Separate set [internal + external]	AL	Unspecified	MCD	Other	Qual.
Dohopolski 2022 [70]	Preprint	2022	USA	No	Yes	Outcome related	CT	Dose	Head and neck	Supervised	217	Unspecified	CV	54	Separate set [internal]	FD	Both	MCD; TTA; CP; EDL	Entropy-based; Other	Quant.
Li 2022a [71]	Standard pub.	2022	USA	No	No	Contouring	CT	Target	Prostate	Supervised	306	1	Separate set	3	Separate set [internal]	AL; AM	Unspecified	Other	Other	Qual.
Li 2022b [72]	Standard pub.	2022	USA	No	No	Multiple	NA	Clinical	Liver	Supervised	NA	NA	Separate set	182	Separate set [external]	AL	Unspecified	GP	Probability-based	Quant.
Lin 2022 [73]	Standard pub.	2022	China	No	No	Outcome related	CT	Target+Clinical	Esophageal	Supervised	171	57	Separate set	57	Separate set [internal]	AL	Unspecified	Other	Other	Quant.
Liu 2022 [74]	Standard pub.	2022	China	Yes	No	Contouring	CT	Target+OAR	Pancreatic	Supervised	62	Unspecified	Unspecified	21	Cross validation	AL	Unspecified	MCD	Other	Qual.
Lyu 2022 [75]	Preprint	2022	USA	Yes	No	Image synthesis	Multimodal	NA	Multiple	Unsupervised	17	Unspecified	Unspecified	2	Separate set [internal]	FD	Unspecified	Other	Variance-based	Quant. + Qual.

Mody 2022a [76]	Conf. proc.	2022	NL	Yes	Yes	Contouring	CT	OAR	Head and neck	Supervised	33	Unspecified	Unspecified	25	Separate set [internal + external]	FD; Calib.	Unspecified	MCD; Other	Entropy-based	Quant. + Qual.
Mody 2022b [77]	Conf. proc.	2022	NL	Yes	Yes	Contouring	CT	OAR	Head and neck	Supervised	33	5	Separate set	25	Separate set [internal + external]	FD; Calib.	Both	OB	Entropy-based	Quant.
Sun 2022 [78]	Standard pub.	2022	USA	No	No	Outcome related	NA	Dose+Clinical	Lung	Reinforcement	67	Unspecified	Unspecified	67	Other	FD; AL; Calib.	Unspecified	GP	Variance-based	Quant.
Wang 2022 [79]	Standard pub.	2022	USA	No	Yes	Outcome related	PET/CT	Target+Clinical	Head and neck	Supervised	135	45	CV	45	Cross validation	FD	Both	TTA; Other	Entropy-based; Probability-based	Quant.
Yang 2022 [80]	Standard pub.	2022	China	No	No	Dose prediction	NA	Dose	Multiple*	Supervised	Unspecified	Unspecified	Separate set	Unspecified	Separate set [internal + external]	AL; Calib.; FD; OODD	Both	MCD	Entropy-based	Quant.
Zabiholla hy 2022 [81]	Standard pub.	2022	USA	No	No	Contouring	MRI	Target+OAR	Cervical	Supervised	112	Unspecified	Separate set	13	Separate set [internal]	FD	Unspecified	MCD	Variance-based	Qual.
Cubero 2023 [82]	Conf. proc.	2023	Spain	No	No	Contouring	CT	OAR	Head and neck	Supervised	40	Unspecified	CV	8	Separate set [internal]	FD	Unspecified	MCD	Entropy-based	Quant. + Qual.
DeBiase 2023 [83]	Standard pub.	2023	NL	No	No	Contouring	PET/CT	Target	Head and neck	Supervised	113	37	CV	25	Separate set [internal]	FD	Unspecified	ENS	Probability-based	Quant.
Ebadi 2023 [84]	Standard pub.	2023	USA	Yes	Yes	Contouring	CT	Target	Lung	Supervised	438	Unspecified	CV	3	Cross validation	FD; Calib.	Unspecified	MCD; ENS	Entropy-based; Variance-based	Quant. + Qual.
Galapon 2023 [85]	Standard pub.	2023	NL	No	No	Image synthesis	Multimodal	NA	Head and neck	Unsupervised	71	10	Separate set	20	Separate set [internal]	FD	Unspecified	MCD	Variance-based	Quant. + Qual.
Grewal 2023 [86]	Conf. proc.	2023	NL	No	Yes	Contouring	CT	OAR	Cervical	Supervised	1108	Unspecified	CV	95	Separate set [internal]	AL	Epistemic	Other	Entropy-based	Quant.
Huttinga 2023 [87]	Standard pub.	2023	NL	No	No	Motion tracking	MRI	K-space	Cardiac*	Supervised	1	Unspecified	Unspecified	1	Other	FD	Unspecified	GP	Probability-based	Quant.
Luan 2023 [88]	Standard pub.	2023	China	Yes	No	Contouring	CT	OAR	Head and neck	Supervised	70	Unspecified	CV	68	Separate set [multiple external]	AL	Unspecified	DSO	Probability-based	Quant. + Qual.
Min 2023 [89]	Standard pub.	2023	AU	No	No	Contouring	MRI	Target	Prostate	Supervised	393	5	Separate set	49	Separate set [internal]	FD	Unspecified	MCD	Variance-based	Quant. + Qual.
Outeiral 2023 [90]	Standard pub.	2023	NL	No	Yes	Contouring	MRI	Target	Multiple	Supervised	181	Unspecified	Separate set	Unspecified	Separate set [internal]	FD	Unspecified	DSO	Probability-based	Quant. + Qual.
Sahlsten 2023 [91]	Preprint	2023	USA	Yes	No	Contouring	PET/CT	Target	Head and neck	Supervised	224	45	CV	67	Separate set [external]	FD	Both	MCD; ENS	Entropy-based; Variance-based	Quant. + Qual.
Smolders 2023 [92]	Standard pub.	2023	SUI	No	Yes	Image registration	CT	NA	Multiple	Mixed	50	10	Separate set	10	Separate set [multiple external]	FD; Calib.	Unspecified	OB; Other	Variance-based	Quant. + Qual.
Tian 2023 [93]	Standard pub.	2023	Germany	Yes	No	Image synthesis	Multimodal	NA	Pelvic	Supervised	10	4	Separate set	5	Separate set [internal]	FD	Unspecified	MCD	Variance-based	Quant.
DeBiase 2024 [94]	Standard pub.	2024	NL	No	No	Multiple	PET/CT	Dose+Clinical+Target+Probability map	Head and neck	Supervised	168	Unspecified	CV	100	Separate set [internal]	FD; Calib.	Unspecified	ENS	Probability-based	Quant.

Li 2024 [95]	Standard pub.	2024	SUI	No	No	Multiple	Multimodal	Dose	Brain	Unsupervised	64	8	CV	10	Separate set [internal]	FD	Unspecified	OB	Other	Quant. + Qual.
Rusanov 2024 [96]	Standard pub.	2024	AU	No	No	Image synthesis	CT	NA	Prostate	Unsupervised	40	5	Separate set	5	Separate set [internal]	FD; Calib.	Both	MCD; TTA	Variance-based	Quant. + Qual.

* Dataset specific notes. Nomura 2021 – phantom replicas derived from human head and neck patient data. Remy 2021 – volunteer data but with specific application to radiotherapy applications. Diao 2021 – utilized two disease sites (soft tissue sarcoma and lymphoma) but combined data into one dataset. Kajikawa 2021 – patients had primary diseases that were not cancer but application of study was specific to radiotherapy. Yang 2022 – primary study involved glioma, lung, and liver cancer patients, out-of-distribution experiments involved breast, cervical, esophageal, tongue, and lung cancer patients. Huttinga 2023 – in vivo studies using volunteers and a ventricular tachycardia patient who received radioablation.

General Study Characteristics

Twelve countries of origin were represented, with the majority of studies emanating from the United States (23%), China (20%), or Netherlands (20%) (**Fig 1A**, **Fig 1B**). Most studies were standard peer-reviewed publications (75%) (**Fig 1A**). The range of publication dates included in this study was 2015-2024, with most studies taking place in 2021, 2022, or 2023 (**Fig 1C**). The majority of studies did not publicly release data or code (55%), with only 32% releasing data, 29% releasing code, and 16% releasing both data and code; relative code and data availability increased in 2021, 2022, and 2023 (**Fig 1C**).

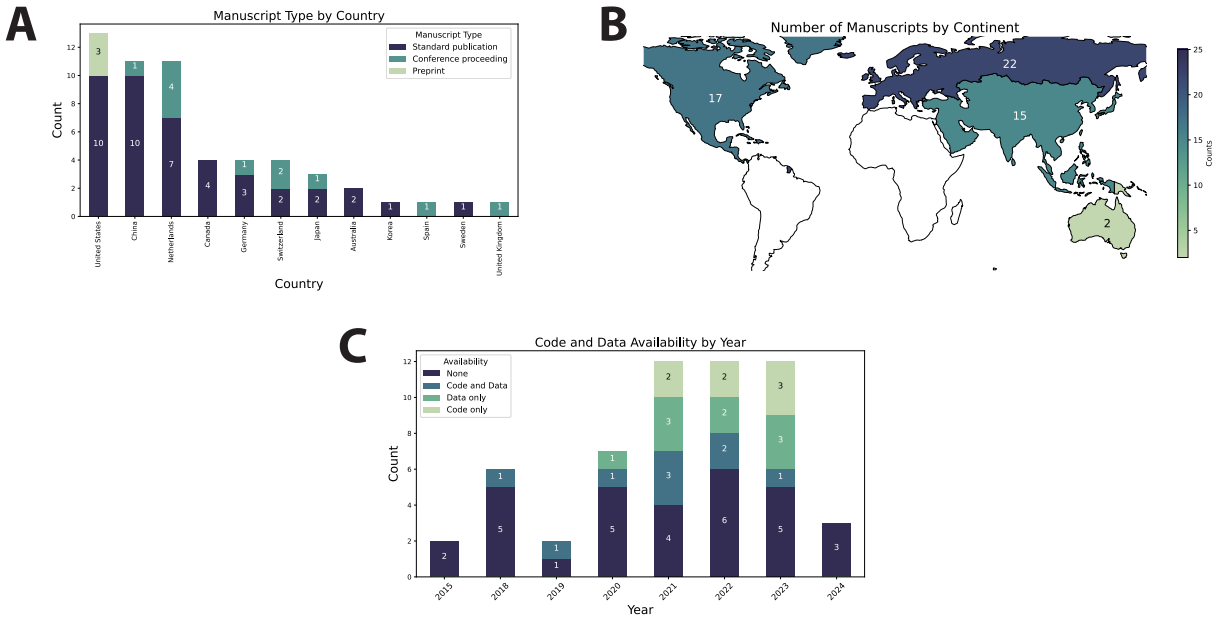

Figure 1. General study characteristics. (**A**) Stacked barplot showing total number of publications per country by publication type. (**B**) Heatmap of the number of studies by continent where green indicates a low number of publications and blue indicates a high number of publications; continents where no studies were extracted from are represented in white. (**C**) Stacked barplot showing code and data availability over time. Each item in the barplots corresponds to one study.

Radiotherapy Characteristics

Multiple disease sites were included: head and neck, prostate, brain, lung, cervical, liver, esophageal, pancreatic, cardiac, breast, pelvic. Most studies were applied to head and neck cancer patient populations (32%) (**Fig 2A**). Ten RT application domains were involved: contouring, image synthesis, outcome-related, motion tracking, dose planning, image registration, nodal classification, tumor growth modeling, image correction. Most applications

were focused on contouring (50%) (**Fig 2A**). Most used medical imaging data in some capacity — 45% of studies utilized CT data (**Fig 2B**); only 9% did not utilize medical imaging. The majority of studies also utilized target structures (29%), OARs (21%), or both (11%) as input data in their algorithms (**Fig 2B**); only 13% of included studies used RT dose in their algorithms.

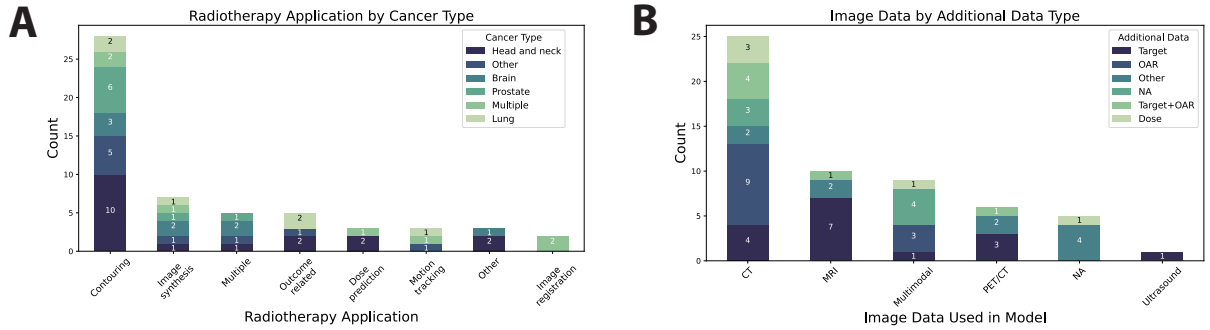

Figure 2. Radiotherapy characteristics. (**A**) Stacked barplot showing cancer disease site per each radiotherapy application domain. “Other” category for cancer type included cervical, liver, esophageal, pancreatic, cardiac, breast, pelvic. “Other” category for radiotherapy application included nodal classification, tumor growth modeling, and image correction. (**B**) Stacked barplot showing additional data per each imaging modality represented. “Other” category for additional data included registration transforms, respiratory trace, K-space, fiducial, clinical data, target+clinical data, dose+clinical data, and dose+clinical data+target+probability map. Each item in the barplots correspond to one study.

AI Characteristics

The vast majority of the studies (88%) used labeled data for model training, i.e., supervised learning. Median (interquartile range) patient sample sizes were 63 (145.25), 10 (31.5), and 25 (46) for training, validation, and test datasets (**Fig 3A**). Most studies used a separate dataset for model validation (40%) compared to cross-validation approaches (30%), while 30% did not mention their validation methods. Most studies used a separate test set composed of internal, i.e., single source data (55%); only 7% of studies used multiple external validation datasets for testing (**Fig 3B**).

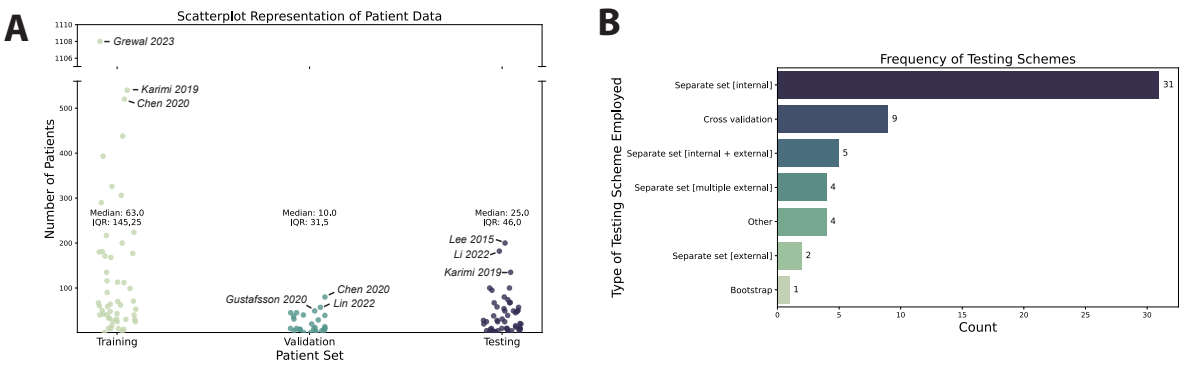

Figure 3. Artificial intelligence characteristics. (**A**) Scatter plot showing number of training, validation, and testing patients used in studies. Only studies that explicitly reported patient-level sample sizes are included. The three studies with the highest sample sizes in each category are annotated. (**B**) Bar plot showing types of testing strategies used in studies. Each item in the barplot corresponds to one study.

Uncertainty Quantification Characteristics

Most studies investigated failure detection applications (60% of reported applications) followed by calibration (19%) and active learning (18%), with only a few studies investigating ambiguity modeling or out-of-distribution detection (**Fig 4A**). The majority of studies used MC Dropout (32% of reported methods), followed by ensembles (16%) and other methods (16%), with a smaller number of studies using other Bayesian methods, direct softmax outputs, test-time augmentation, gaussian processes, Platt scaling, conformal predictions, and evidential deep learning (**Fig 4B**). In terms of calculated uncertainty metrics, most studies reported using variance-based methods (34% of reported metrics) and entropy-based metrics (27%), followed by other self-defined metrics (23%), with the smallest number reporting probability based metrics (16%) (**Fig 4C**). Most studies did not explicitly report if they investigated aleatoric or epistemic uncertainty (77% of studies) and used a combination of quantitative and qualitative experiments for investigating uncertainty (52%).

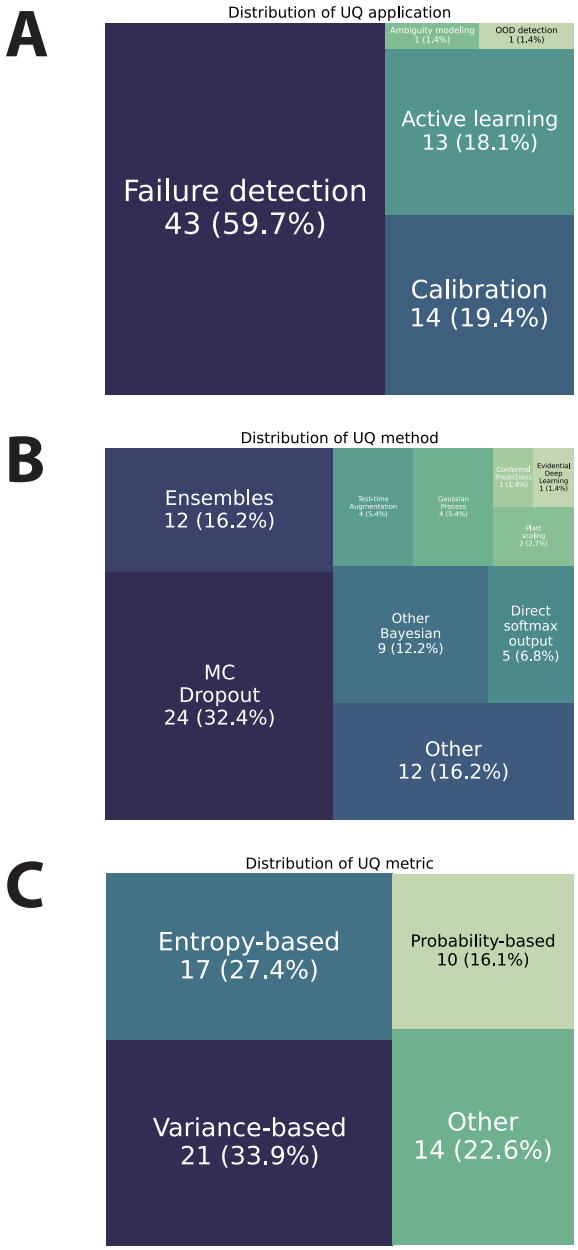

Figure 4. Uncertainty quantification characteristics. (**A**) Tree map of uncertainty quantification applications represented in the studies. (**B**) Tree map of uncertainty quantification methods represented in the studies. (**C**) Tree map of uncertainty quantification metrics represented in the studies. Each item in the tree maps correspond to a reported item (could be multiple per study).

Study Overlap with Previous Reviews

Five systematic/scoping review papers related to UQ in medicine were selected for the overlap comparison. Only six studies investigated in these review papers overlapped with our 56 extracted manuscripts (**Table 2**).

Table 2. Study overlap with previous systematic and scoping reviews. Papers contained in previous systematic/scoping reviews related to uncertainty quantification in medicine were compared with papers extracted for our scoping review. Only final papers that were used for data extraction were compared.

Review Paper	Overlapping Citations/Total Citations (%)	Specific Overlapping Manuscript(s)
Zou et al. (2023) [32]	3/56 (5%)	Jungo et al. (2018) [44], Jungo et al. (2018) [45], Bragman et al. (2018) [43]
Kurz et al. (2022) [33]	0/22 (0%)	None
Lambert et al. (2024) [34]	4/217 (2%)	Jungo et al. (2018) [44], Jungo et al. (2018) [45], Balagopal et al. (2021) [58], Mei et al. (2021) [64]
Loftus et al. (2022) [31]	0/30 (0%)	None
Seoni et al. (2023) [25]	2/144 (1%)	Jungo et al. (2018) [45], Lipkova et al. (2019) [50]

Discussion

The field of RT is increasingly incorporating AI into its various workflows. Although AI UQ is well-established in computer science, its adaptation to medicine and RT is still in its early stages. Incorporating UQ in AI models used in RT workflows has the potential to increase clinician confidence, helping bridge the translational gap between single institutional model development to multi-institutional clinical implementation. Our scoping review is a pioneering effort to systematically examine the application of UQ concepts within RT.

We identified several trends in UQ research in RT, likely driven by technical innovations within the broader AI research community. Mirroring practices from computer science [38], a considerable number of manuscripts were conference proceedings rather than traditional publications. Notably, we found a predominant contribution of studies from the European Union (EU). Given stringent EU data protection laws — such as The General Data Protection Regulation (GDPR) which poses challenges for secondary data use in research [97] — this raises considerations for how practitioners of UQ in RT should value data sharing. There exists

a known tension between open science principles and protecting patient privacy. Although code and data availability have become standard in AI-related research, medical applications lag in this regard [98–100]. Our analysis revealed a gradual increase in code and data availability over time, reflecting a slowly evolving open science ethos in the RT community [21]. Notably, the National Institutes of Health (NIH) new Data Management and Sharing policy, effective January 2023, mandates broader data sharing for NIH-supported research [101], aligning with these open science principles. In light of these findings, we advocate for the publication of code and anonymized data in AI UQ research in RT wherever feasible to enhance reproducibility. When code or data sharing is not possible, future work should encourage privacy-preserving methodologies, like federated learning [102], as viable alternatives to conventional data sharing practices.

The extracted manuscripts covered various RT application spaces and disease sites. Auto-contouring was the most common application, aligning with its prevalence in AI-based RT [11,103,104]. Many studies focused on head and neck cancer, likely due to the complexity of this disease site, which requires precise delineation of numerous organs at risk (OARs) and challenging tumor-related target structures [105]. Most auto-contouring studies investigated OARs on CT imaging. Target structures were often generated with imaging modalities beyond CT, such as MRI for enhanced soft tissue contrast or PET for metabolic activity incorporation, matching physician practice patterns [106]. The variability observed in OARs and target structures can be characterized as aleatoric uncertainty driven by physician judgment [104,107,108]. Interestingly, Karimi et al. [49] showed that reducing aleatoric uncertainty may not be as critical as ensuring large training set sizes, at least in their specific prostate cancer target use-case. This finding suggests that, for institutional model training and fine-tuning, focusing on expanding dataset size could be more impactful than minimizing underlying contour variability (i.e., addressing factors associated with epistemic uncertainty). Of note, it has recently been suggested that RT auto-contouring performance is saturating [109], driving the need for research into additional research spaces such as UQ. Future research may benefit from exploring UQ techniques specifically tailored to address aleatoric uncertainty in auto-contouring models, considering the differences in variability between OARs and target structures.

A distinct facet in RT workflows compared to other oncologic research areas is the presence of multidimensional and complex dosimetric treatment data. DL-based dose prediction is emerging as a promising alternative to traditional knowledge-based planning approaches, offering the potential for improved accuracy, reliability, and efficiency in patient-specific plan optimization [110]. The uncertainty in DL-based dose prediction models could be critical, as it could determine when model-generated plans should be directly accepted, or if manual interventions from physicians and physicists are required to improve plan quality [80]. Surprisingly, in our review there were relatively few manuscripts directly investigating model UQ in dose prediction applications [56,65,72,80,95]. This scarcity is mirrored in outcome prediction research, where only a few studies explored dose-related toxicities [42,70,72,78], as opposed to broader oncologic outcomes like survival. Naturally, a major challenge in outcome-related research stems from the limited availability of training samples. Compared with studies that leverage granular inputs (e.g., multiple image slices representing one patient), dose-related toxicity

outcomes often can only be represented at the patient level, which may explain the relative scarcity of literature.

Consistent with similar medical domains reliant on imaging [111], the majority of studies in our review employed supervised learning techniques, which involve training models on labeled data to make predictions based on new, unseen data. A minority of studies explored unsupervised learning approaches [43,57,75,85,95,96] where models learn patterns and relationships from data without explicit labels. In our review, these unsupervised methods were particularly useful for image synthesis tasks. Only one study utilized reinforcement learning [78], a technique where an agent learns to make decisions based on rewards and punishments. Regardless of the ML technique employed, training dataset sizes were generally small, with the three largest patient cohorts corresponding to auto-contouring studies (520-1108 patients) [49,51,86]. ML models often struggle with small sample sizes, especially when considering complex, multidimensional data like medical images, where models must learn intricate generalizable spatial relationships. As previously noted, this challenge is intensified when prediction outputs are restricted to broad patient-level information, such as toxicity or prognosis, with each patient representing a single data point. Notably, tasks that utilize more granular training information, like auto-contouring or image synthesis, can effectively utilize the numerous data points within each image, allowing these models to achieve reasonable performance despite the limited number of patients [109,112,113]. However, given these relatively small patient sample sizes, it is likely that intrinsic epistemic uncertainty would be high. Subsequently, carefully designed UQ may help identify patients for whom the model's predictions are more reliable. Finally, despite the importance of using diverse and heterogeneous data for uncertainty experiments, particularly for determining how well models handle new and unknown data scenarios [22], only a handful of studies attempted to utilize multiple external test datasets [48,57,88,92]. Interestingly, this was in stark contrast to a previous scoping review on AI UQ in a broader medical context which identified a predominance of external dataset testing [31].

Ideally, UQ methods should be validated across a broad spectrum of downstream uncertainty tasks [22]. However, in our review only the study by Yang et al. [80] explored a comprehensive approach. Focusing on RT dose delivery, they evaluated several UQ applications such as active learning, calibration, failure detection, and out-of-distribution detection. Most other studies focused on singular UQ applications, with failure detection being the most common. In these studies, UQ is used as a quality assurance tool, such as flagging contours below a pre-defined quality threshold. Interestingly, model calibration, which attempts to ensure that predicted probabilities align with observed outcomes, appears somewhat underexplored in the reviewed studies, despite its historical importance in uncertainty discussions [27]. This oversight might stem from an inherent assumption that some UQ model outputs are already calibrated [28], which may not always hold true [29]. A minority of studies also used uncertainty in active learning frameworks, where the model selects the most informative data points for labeling based on their uncertainty, to improve model training. Ambiguity modeling and out-of-distribution detection were vastly underrepresented, with only one study each (Li et al. [71], and Yang et al. [80], respectively) investigating these areas.

In line with previous review literature [31,33], Monte Carlo dropout was the most frequently used UQ method in our scoping review. Monte Carlo dropout has gained widespread acceptance for its simplicity and effectiveness as a scalable approach to approximate Bayesian inference. Notably, Monte Carlo dropout, and other common methods such as ensembles often yield comparable results [114], so the superiority of specific methods is unclear and likely context-specific. In terms of uncertainty metrics, the majority of the reviewed studies favored variance-based or entropy-based metrics, aligning with their established prevalence in the literature [31,33]. There appears to be a noticeable gap in the adoption of newer, innovative approaches. For instance, only one study in our review, conducted by Dohopolski et al. on head and neck cancer outcome prediction [70], explored newer methods like evidential deep learning and statistically rigorous approaches like conformal prediction. Moreover, while qualitative analyses through heatmap visualizations were common in our extracted studies, it has been argued that conventional methods (e.g., Monte Carlo dropout, ensembles) fall short in providing spatially-correlated pixel-wise estimates [115], which may ultimately limit their clinical utility and incentivize the development of alternative approaches.

Historically, the AI UQ research community has placed significant emphasis on distinguishing between aleatoric and epistemic uncertainty. However, recent literature suggests that the ability to differentiate aleatoric from epistemic uncertainty using popular contemporary UQ techniques may not be as clear-cut as previously thought [116]. Our review revealed that the majority of studies surveyed did not explicitly identify whether their models captured aleatoric or epistemic uncertainty. This observation suggests that, at least within the RT community, the distinction between these types of uncertainty may not be deemed critical enough to warrant specific mention. Moreover, the practical significance of distinguishing between epistemic and aleatoric uncertainty may vary depending on the study's objectives; for instance, if the primary goal is to quickly flag errors for broader human oversight (e.g., failure detection), a detailed separation of these uncertainty types might not be crucial.

Although a principal motivation behind UQ in medical AI is often believed to be the enhancement of clinician trust [5], none of the studies we reviewed explicitly investigated the influence of UQ estimates on end-user trust or decision making. This is particularly interesting given that most studies dealt with failure detection applications which would necessitate a secondary review by a clinician. Literature within diagnostic imaging applications has demonstrated that the presentation of differential outputs from an AI algorithm can impact user performance, confidence, and reliance to varying degrees [117,118]. Examining how UQ influences clinician trust and decision-making in RT through targeted human-machine interaction experiments could further elucidate the real-world impact of these tools, suggesting a vital direction for future research.

To further emphasize the novelty and need of our study, we investigated the intersection between the publications we reported on and those already cited across existing related systematic and scoping review papers. Importantly, our examination revealed just six instances of citation overlap across five distinct review papers, highlighting the originality of our research and a significant gap within the current academic discourse.

Our study, while striving for a structured and comprehensive overview of existing literature, has some limitations. Firstly, the landscape of available studies was largely limited to those indexed in the queried databases, although we supplemented our search with hand-selected literature to ensure broader coverage. Secondly, the emergent field of AI UQ presents challenges in applying traditional study quality assessment guidelines, as tailored guidelines are not yet available. However, we have taken inspiration from existing reporting guidelines, such as TRIPOD [39] and CLAIM [40], to extract relevant modeling information for our review. Furthermore, we have incorporated aspects of the newly proposed ValUES framework which aims to provide a systematic approach to validating uncertainty estimation in semantic segmentation [22], adapting its principles to enrich our review process.

Finally, an important facet of UQ that we have not considered, but should be the focus of future work is related to model bias. Notably, UQ is often part of broader discussions related to AI explainability, which is often more directly related to identifying and addressing bias. Although explainability and bias in medical AI has garnered significant attention [119], investigation of these topics in RT remains limited [17]. Moreover, while a recent small-scale study indicated that geographic biases in RT auto-contouring models are minimal [120], the necessary broader investigations across various applications have yet to be conducted. The potential for perpetuating biases and inequalities escalates when AI models function as “black boxes” with obscured decision-making processes [121]. UQ, potentially in combination with other explainability methods, could ultimately allow for improved bias detection and mitigation [6].

Conclusions

The escalating use of UQ for RT applications signifies a key shift towards potentially more clinically impactful AI tools. Our scoping review uncovered a broad spectrum of RT applications and disease sites that have benefited from UQ. However, we observed a concentration of efforts in specific areas, such as auto-contouring, while crucial domains like dose and outcome prediction were underrepresented. Moreover, although established techniques like Monte Carlo dropout and ensembles were frequently used for failure detection applications, the exploration of alternative methods, such as conformal predictions, was limited. Notably, the majority of studies lacked code and dataset sharing suggesting a need for improved transparency and reproducibility in AI UQ research for RT. Additionally, the absence of standardized guidelines for implementing and reporting AI UQ in RT highlights a crucial area for future research. Addressing these gaps by broadening UQ applications, fostering model transparency, and developing comprehensive guidelines could significantly advance UQ in RT research.

References

[1] Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022;28:31–8.
[2] Wells L, Bednarz T. Explainable AI and Reinforcement Learning-A Systematic Review of Current Approaches and Trends. Front Artif Intell 2021;4:550030.
[3] Shashikumar SP, Wardi G, Malhotra A, Nemati S. Artificial intelligence sepsis prediction algorithm learns to say “I don’t know.” NPJ Digit Med 2021;4:134.
[4] Begoli E, Bhattacharya T, Kusnezov D. The need for uncertainty quantification in machine-assisted medical decision making. Nature Machine Intelligence 2019;1:20–3.
[5] Abdar M, Khosravi A, Islam SMS, Rajendra Acharya U, Vasilakos AV. The need for quantification of uncertainty in artificial intelligence for clinical data analysis: increasing the level of trust in the decision-making process. IEEE Systems, Man, and Cybernetics Magazine 2022;8:28–40.
[6] Faghani S, Moassefi M, Rouzrokh P, Khosravi B, Baffour FI, Ringler MD, et al. Quantifying Uncertainty in Deep Learning of Radiologic Images. Radiology 2023;308:e222217.
[7] Chaput G, Regnier L. Radiotherapy: Clinical pearls for primary care. Can Fam Physician 2021;67:753–7.
[8] Lastrucci A, Wandael Y, Ricci R, Maccioni G, Giansanti D. The Integration of Deep Learning in Radiotherapy: Exploring Challenges, Opportunities, and Future Directions through an Umbrella Review. Diagnostics 2024;14:939.
[9] Spadea MF, Maspero M, Zaffino P, Seco J. Deep learning based synthetic-CT generation in radiotherapy and PET: A review. Med Phys 2021;48:6537–66.
[10] Teuwen J, Gouw ZAR, Sonke J-J. Artificial Intelligence for Image Registration in Radiation Oncology. Semin Radiat Oncol 2022;32:330–42.
[11] Isaksson LJ, Summers P, Mastroleo F, Marvaso G, Corrao G, Vincini MG, et al. Automatic Segmentation with Deep Learning in Radiotherapy. Cancers 2023;15. https://doi.org/10.3390/cancers15174389.
[12] Wang M, Zhang Q, Lam S, Cai J, Yang R. A Review on Application of Deep Learning Algorithms in External Beam Radiotherapy Automated Treatment Planning. Front Oncol 2020;10:580919.
[13] Huang S, Yang J, Fong S, Zhao Q. Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. Cancer Lett 2020;471:61–71.
[14] Appelt AL, Elhaminia B, Gooya A, Gilbert A, Nix M. Deep Learning for Radiotherapy Outcome Prediction Using Dose Data - A Review. Clin Oncol 2022;34:e87–96.
[15] Tan D, Mohd Nasir NF, Abdul Manan H, Yahya N. Prediction of toxicity outcomes following radiotherapy using deep learning-based models: A systematic review. Cancer Radiother 2023;27:398–406.
[16] Hallows R, Glazier L, Katz MS, Aznar M, Williams M. Safe and Ethical Artificial Intelligence in Radiotherapy - Lessons Learned From the Aviation Industry. Clin Oncol 2022;34:99–101.
[17] Heising L. Accelerating Implementation of Artificial Intelligence in Radiotherapy through Explainability. Joint 1st World Conference on eXplainable Artificial Intelligence: Late-Breaking Work, Demos and Doctoral Consortium, xAI-2023: LB-D-DC, vol. 3554, Rheinisch-Westfaelische Technische Hochschule Aachen * Lehrstuhl Informatik V; 2023, p. 217–24.
[18] Hüllermeier E, Waegeman W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Mach Learn 2021;110:457–506.
[19] Lin D, Lapen K, Sherer MV, Kantor J, Zhang Z, Boyce LM, et al. A Systematic Review of Contouring Guidelines in Radiation Oncology: Analysis of Frequency, Methodology, and Delivery of Consensus Recommendations. Int J Radiat Oncol Biol Phys 2020;107:827–35.

[20] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet Large Scale Visual Recognition Challenge. Int J Comput Vis 2015;115:211–52.
[21] Wahid KA, Glerean E, Sahlsten J, Jaskari J, Kaski K, Naser MA, et al. Artificial Intelligence for Radiation Oncology Applications Using Public Datasets. Semin Radiat Oncol 2022;32:400–14.
[22] Kahl K-C, Lüth CT, Zenk M, Maier-Hein K, Jaeger PF. ValUES: A Framework for Systematic Validation of Uncertainty Estimation in Semantic Segmentation. arXiv [csCV] 2024.
[23] Gal Y, Ghahramani Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In: Balcan MF, Weinberger KQ, editors. Proceedings of The 33rd International Conference on Machine Learning, vol. 48, New York, New York, USA: PMLR; 20--22 Jun 2016, p. 1050–9.
[24] Wilson AG. Deep ensembles as approximate Bayesian inference. Deep Ensembles as Approximate Bayesian Inference 2019. https://cims.nyu.edu/~andrewgw/deepensembles/ (accessed April 8, 2024).
[25] Seoni S, Jahmunah V, Salvi M, Barua PD, Molinari F, Acharya UR. Application of uncertainty quantification to artificial intelligence in healthcare: A review of last decade (2013-2023). Comput Biol Med 2023;165:107441.
[26] Vazquez J, Facelli JC. Conformal Prediction in Clinical Medical Sciences. Int J Healthc Inf Syst Inform 2022;6:241–52.
[27] Guo C, Pleiss G, Sun Y, Weinberger KQ. On Calibration of Modern Neural Networks. In: Precup D, Teh YW, editors. Proceedings of the 34th International Conference on Machine Learning, vol. 70, PMLR; 06--11 Aug 2017, p. 1321–30.
[28] Mehrtash A, Wells WM, Tempany CM, Abolmaesumi P, Kapur T. Confidence Calibration and Predictive Uncertainty Estimation for Deep Medical Image Segmentation. IEEE Trans Med Imaging 2020;39:3868–78.
[29] Ghoshal B, Tucker A. On Calibrated Model Uncertainty in Deep Learning. arXiv [csLG] 2022.
[30] van den Berg CAT, Meliadò EF. Uncertainty Assessment for Deep Learning Radiotherapy Applications. Semin Radiat Oncol 2022;32:304–18.
[31] Loftus TJ, Shickel B, Ruppert MM, Balch JA, Ozrazgat-Baslanti T, Tighe PJ, et al. Uncertainty-aware deep learning in healthcare: A scoping review. PLOS Digit Health 2022;1. https://doi.org/10.1371/journal.pdig.0000085.
[32] Zou K, Chen Z, Yuan X, Shen X, Wang M, Fu H. A Review of Uncertainty Estimation and its Application in Medical Imaging. arXiv [eessIV] 2023.
[33] Kurz A, Hauser K, Mehrtens HA, Krieghoff-Henning E, Hekler A, Kather JN, et al. Uncertainty Estimation in Medical Image Classification: Systematic Review. JMIR Med Inform 2022;10:e36427.
[34] Lambert B, Forbes F, Doyle S, Dehaene H, Dojat M. Trustworthy clinical AI solutions: A unified review of uncertainty quantification in Deep Learning models for medical image analysis. Artif Intell Med 2024;150. https://doi.org/10.1016/j.artmed.2024.102830.
[35] Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann Intern Med 2018;169:467–73.
[36] Babineau J. Product Review: Covidence (Systematic Review Software). J Can Health Libr Assoc 2014;35:68.
[37] Pollock D, Peters MDJ, Khalil H, McInerney P, Alexander L, Tricco AC, et al. Recommendations for the extraction, analysis, and presentation of results in scoping reviews. JBI Evid Synth 2023;21:520–32.
[38] Freyne J, Coyle L, Smyth B, Cunningham P. Relative status of journal and conference publications in computer science. Commun ACM 2010;53:124–32.

[39] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC Med 2015;13:1.
[40] Mongan J, Moy L, Kahn CE Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. Radiol Artif Intell 2020;2:e200029.
[41] Bukhari W, Hong S-M. Real-time prediction and gating of respiratory motion using an extended Kalman filter and Gaussian process regression. Phys Med Biol 2015;60:233–52.
[42] Lee S, Ybarra N, Jeyaseelan K, Faria S, Kopek N, Brisebois P, et al. Bayesian network ensemble as a multivariate strategy to predict radiation pneumonitis risk. Med Phys 2015;42:2421–30.
[43] Bragman FJS, Tanno R, Eaton-Rosen Z, Li W, Hawkes DJ, Ourselin S, et al. Uncertainty in Multitask Learning: Joint Representations for Probabilistic MR-only Radiotherapy Planning. Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, Springer International Publishing; 2018, p. 3–11.
[44] Jungo A, Meier R, Ermis E, Herrmann E, Reyes M. Uncertainty-driven Sanity Check: Application to Postoperative Brain Tumor Cavity Segmentation. arXiv [csCV] 2018.
[45] Jungo A, Meier R, Ermis E, Blatti-Moreno M, Herrmann E, Wiest R, et al. On the Effect of Inter-observer Variability for a Reliable Estimation of Uncertainty of Medical Image Segmentation. Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, Springer International Publishing; 2018, p. 682–90.
[46] Ninomiya K, Arimura H, Sasahara M, Hirose T, Ohga S, Umezu Y, et al. Bayesian delineation framework of clinical target volumes for prostate cancer radiotherapy using an anatomical-features-based machine learning technique. Medical Imaging 2018: Image-Guided Procedures, Robotic Interventions, and Modeling, vol. 10576, SPIE; 2018, p. 472–7.
[47] Qin W, Wu J, Han F, Yuan Y, Zhao W, Ibragimov B, et al. Superpixel-based and boundary-sensitive convolutional neural network for automated liver segmentation. Phys Med Biol 2018;63:095017.
[48] Sentker T, Madesta F, Werner R. GDL-FIRE4D: Deep Learning-Based Fast 4D CT Image Registration. Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, Springer International Publishing; 2018, p. 765–73.
[49] Karimi D, Zeng Q, Mathur P, Avinash A, Mahdavi S, Spadinger I, et al. Accurate and robust deep learning-based segmentation of the prostate clinical target volume in ultrasound images. Med Image Anal 2019;57:186–96.
[50] Lipkova J, Angelikopoulos P, Wu S, Alberts E, Wiestler B, Diehl C, et al. Personalized Radiotherapy Design for Glioblastoma: Integrating Mathematical Tumor Models, Multimodal Scans, and Bayesian Inference. IEEE Trans Med Imaging 2019;38:1875–84.
[51] Chen X, Men K, Chen B, Tang Y, Zhang T, Wang S, et al. CNN-Based Quality Assurance for Automatic Segmentation of Breast Cancer in Radiotherapy. Front Oncol 2020;10:524.
[52] Dohopolski M, Chen L, Sher D, Wang J. Predicting lymph node metastasis in patients with oropharyngeal cancer by using a convolutional neural network with associated epistemic and aleatoric uncertainty. Phys Med Biol 2020;65:225002.
[53] Gustafsson CJ, Swärd J, Adalbjörnsson SI, Jakobsson A, Olsson LE. Development and evaluation of a deep learning based artificial intelligence for automatic identification of gold fiducial markers in an MRI-only prostate radiotherapy workflow. Phys Med Biol 2020;65:225011.
[54] Hänsch A, Hendrik Moltz J, Geisler B, Engel C, Klein J, Genghi A, et al. Hippocampus segmentation in CT using deep learning: impact of MR versus CT-based training contours. J Med Imaging (Bellingham) 2020;7:064001.
[55] Maspero M, Bentvelzen LG, Savenije MHF, Guerreiro F, Seravalli E, Janssens GO, et al. Deep learning-based synthetic CT generation for paediatric brain MR-only photon and

proton radiotherapy. Radiother Oncol 2020;153:197–204.
[56] Nomura Y, Wang J, Shirato H, Shimizu S, Xing L. Fast spot-scanning proton dose calculation method with uncertainty quantification using a three-dimensional convolutional neural network. Phys Med Biol 2020;65:215007.
[57] van Harten LD, Wolterink JM, Verhoeff JJC, Išgum I. Automatic online quality control of synthetic CTs. Medical Imaging 2020: Image Processing, vol. 11313, SPIE; 2020, p. 399–405.
[58] Balagopal A, Nguyen D, Morgan H, Weng Y, Dohopolski M, Lin M-H, et al. A deep learning-based framework for segmenting invisible clinical target volumes with estimated uncertainties for post-operative prostate cancer radiotherapy. Med Image Anal 2021;72:102101.
[59] Dasgupta A, Geraghty B, Maralani PJ, Malik N, Sandhu M, Detsky J, et al. Quantitative mapping of individual voxels in the peritumoral region of IDH-wildtype glioblastoma to distinguish between tumor infiltration and edema. J Neurooncol 2021;153:251–61.
[60] Diao Z, Jiang H, Han X-H, Yao Y-D, Shi T. EFNet: evidence fusion network for tumor segmentation from PET-CT volumes. Phys Med Biol 2021;66. https://doi.org/10.1088/1361-6560/ac299a.
[61] Kajikawa T, Kadoya N, Maehara Y, Miura H, Katsuta Y, Nagasawa S, et al. A deep learning method for translating 3DCT to SPECT ventilation imaging: First comparison with 81m Kr-gas SPECT ventilation imaging. Med Phys 2022;49:4353–64.
[62] Lei W, Mei H, Sun Z, Ye S, Gu R, Wang H, et al. Automatic segmentation of organs-at-risk from head-and-neck CT using separable convolutional neural network with hard-region-weighted loss. Neurocomputing 2021;442:184–99.
[63] Luo X, Liao W, Chen J, Song T, Chen Y, Zhang S, et al. Efficient Semi-supervised Gross Target Volume of Nasopharyngeal Carcinoma Segmentation via Uncertainty Rectified Pyramid Consistency. Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Springer International Publishing; 2021, p. 318–29.
[64] Mei H, Lei W, Gu R, Ye S, Sun Z, Zhang S, et al. Automatic segmentation of gross target volume of nasopharynx cancer using ensemble of multiscale deep neural networks with spatial attention. Neurocomputing 2021;438:211–22.
[65] Nguyen D, Sadeghnejad Barkousaraie A, Bohara G, Balagopal A, McBeth R, Lin M-H, et al. A comparison of Monte Carlo dropout and bootstrap aggregation on the performance and uncertainty estimation in radiation therapy dose prediction with deep learning neural networks. Phys Med Biol 2021;66:054002.
[66] Nomura Y, Tanaka S, Wang J, Shirato H, Shimizu S, Xing L. Calibrated uncertainty estimation for interpretable proton computed tomography image correction using Bayesian deep learning. Phys Med Biol 2021;66:065029.
[67] Remy C, Ahumada D, Labine A, Côté J-C, Lachaine M, Bouchard H. Potential of a probabilistic framework for target prediction from surrogate respiratory motion during lung radiotherapy. Phys Med Biol 2021;66. https://doi.org/10.1088/1361-6560/abf1b8.
[68] van Rooij W, Verbakel WF, Slotman BJ, Dahele M. Using Spatial Probability Maps to Highlight Potential Inaccuracies in Deep Learning-Based Contours: Facilitating Online Adaptive Radiation Therapy. Adv Radiat Oncol 2021;6:100658.
[69] Zhang G, Yang Z, Huo B, Chai S, Jiang S. Automatic segmentation of organs at risk and tumors in CT images of lung cancer from partially labelled datasets with a semi-supervised conditional nnU-Net. Comput Methods Programs Biomed 2021;211:106419.
[70] Dohopolski M, Wang K, Wang B, Bai T, Nguyen D, Sher D, et al. Uncertainty estimations methods for a deep learning model to aid in clinical decision-making -- a clinician's perspective. arXiv [csLG] 2022.
[71] Li X, Bagher-Ebadian H, Gardner S, Kim J, Elshaikh M, Movsas B, et al. An uncertainty-aware deep learning architecture with outlier mitigation for prostate gland segmentation in

radiotherapy treatment planning. Med Phys 2023;50:311–22.
[72] Li P, Taylor JMG, Boonstra PS, Lawrence TS, Schipper MJ. Utility based approach in individualized optimal dose selection using machine learning methods. Stat Med 2022;41:2957–77.
[73] Lin Z, Cai W, Hou W, Chen Y, Gao B, Mao R, et al. CT-Guided Survival Prediction of Esophageal Cancer. IEEE J Biomed Health Inform 2022;26:2660–9.
[74] Liu S, Liang S, Huang X, Yuan X, Zhong T, Zhang Y. Graph-enhanced U-Net for semi-supervised segmentation of pancreas from abdomen CT scan. Phys Med Biol 2022;67. https://doi.org/10.1088/1361-6560/ac80e4.
[75] Lyu Q, Wang G. Conversion Between CT and MRI Images Using Diffusion and Score-Matching Models. arXiv [eessIV] 2022.
[76] Mody PP, Chaves-de-Plaza N, Hildebrandt K, van Egmond R, de Ridder H, Staring M. Comparing Bayesian models for organ contouring in head and neck radiotherapy. Medical Imaging 2022: Image Processing, vol. 12032, SPIE; 2022, p. 100–9.
[77] Mody P, Chaves-de-Plaza NF, Hildebrandt K, Staring M. Improving Error Detection in Deep Learning Based Radiotherapy Autocontouring Using Bayesian Uncertainty. Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, Springer Nature Switzerland; 2022, p. 70–9.
[78] Sun W, Niraula D, El Naqa I, Ten Haken RK, Dinov ID, Cuneo K, et al. Precision radiotherapy via information integration of expert human knowledge and AI recommendation to optimize clinical decision making. Comput Methods Programs Biomed 2022;221:106927.
[79] Wang K, Dohopolski M, Zhang Q, Sher D, Wang J. Towards reliable head and neck cancers locoregional recurrence prediction using delta-radiomics and learning with rejection option. Med Phys 2023;50:2212–23.
[80] Yang X, Li S, Shao Q, Cao Y, Yang Z, Zhao Y-Q. Uncertainty-guided man-machine integrated patient-specific quality assurance. Radiother Oncol 2022;173:1–9.
[81] Zabihollahy F, Viswanathan AN, Schmidt EJ, Lee J. Fully automated segmentation of clinical target volume in cervical cancer from magnetic resonance imaging with convolutional neural network. J Appl Clin Med Phys 2022;23:e13725.
[82] Cubero L, Serrano J, Castelli J, De Crevoisier R, Acosta O, Pascau J. Exploring Uncertainty for Clinical Acceptability in Head and Neck Deep Learning-Based OAR Segmentation. 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), IEEE; 2023, p. 1–4.
[83] De Biase A, Sijtsema NM, van Dijk L, Langendijk JA, van Ooijen PMA. Deep learning aided oropharyngeal cancer segmentation with adaptive thresholding for predicted tumor probability in FDG PET and CT images. Phys Med Biol 2023. https://doi.org/10.1088/1361-6560/acb9cf.
[84] Ebadi N, Li R, Das A, Roy A, Nikos P, Najafirad P. CBCT-guided adaptive radiotherapy using self-supervised sequential domain adaptation with uncertainty estimation. Med Image Anal 2023;86:102800.
[85] Galapon AV Jr, Thummerer A, Langendijk JA, Wagenaar D, Both S. Feasibility of Monte Carlo dropout-based uncertainty maps to evaluate deep learning-based synthetic CTs for adaptive proton therapy. Med Phys 2023. https://doi.org/10.1002/mp.16838.
[86] Grewal M, van Weersel D, Westerveld H, Bosman PAN, Alderliesten T. Clinically Acceptable Segmentation of Organs at Risk in Cervical Cancer Radiation Treatment from Clinically Available Annotations. arXiv [eessIV] 2023.
[87] Huttinga NRF, Akdag O, Fast MF, Verhoeff JJC, Mohamed Hoesein FAA, van den Berg CAT, et al. Real-time myocardial landmark tracking for MRI-guided cardiac radio-ablation using Gaussian Processes. Phys Med Biol 2023;68. https://doi.org/10.1088/1361-6560/ace023.

[88] Luan S, Wu K, Wu Y, Zhu B, Wei W, Xue X. Accurate and robust auto-segmentation of head and neck organ-at-risks based on a novel CNN fine-tuning workflow. J Appl Clin Med Phys 2024;25:e14248.
[89] Min H, Dowling J, Jameson MG, Cloak K, Faustino J, Sidhom M, et al. Clinical target volume delineation quality assurance for MRI-guided prostate radiotherapy using deep learning with uncertainty estimation. Radiother Oncol 2023;186:109794.
[90] Rodríguez Outeiral R, Ferreira Silvério N, González PJ, Schaake EE, Janssen T, van der Heide UA, et al. A network score-based metric to optimize the quality assurance of automatic radiotherapy target segmentations. Phys Imaging Radiat Oncol 2023;28:100500.
[91] Sahlsten J, Jaskari J, Wahid KA, Ahmed S, Glerean E, He R, et al. Application of simultaneous uncertainty quantification for image segmentation with probabilistic deep learning: Performance benchmarking of oropharyngeal cancer target delineation as a use-case. medRxiv 2023. https://doi.org/10.1101/2023.02.20.23286188.
[92] Smolders A, Lomax A, Weber DC, Albertini F. Deep learning based uncertainty prediction of deformable image registration for contour propagation and dose accumulation in online adaptive radiotherapy. Phys Med Biol 2023;68. https://doi.org/10.1088/1361-6560/ad0282.
[93] Tian L, Lühr A. Proton range uncertainty caused by synthetic computed tomography generated with deep learning from pelvic magnetic resonance imaging. Acta Oncol 2023;62:1461–9.
[94] De Biase A, Ma B, Guo J, van Dijk LV, Langendijk JA, Both S, et al. Deep learning-based outcome prediction using PET/CT and automatically predicted probability maps of primary tumor in patients with oropharyngeal cancer. Comput Methods Programs Biomed 2024;244:107939.
[95] Li X, Bellotti R, Meier G, Bachtiary B, Weber D, Lomax A, et al. Uncertainty-aware MR-based CT synthesis for robust proton therapy planning of brain tumour. Radiother Oncol 2024;191:110056.
[96] Rusanov B, Hassan GM, Reynolds M, Sabet M, Rowshanfarzad P, Bucknell N, et al. Transformer CycleGAN with uncertainty estimation for CBCT based synthetic CT in adaptive radiotherapy. Phys Med Biol 2024;69. https://doi.org/10.1088/1361-6560/ad1cfc.
[97] Peloquin D, DiMaio M, Bierer B, Barnes M. Disruptive and avoidable: GDPR challenges to secondary research uses of data. Eur J Hum Genet 2020;28:697–705.
[98] McDermott MBA, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M. Reproducibility in machine learning for health research: Still a ways to go. Sci Transl Med 2021;13. https://doi.org/10.1126/scitranslmed.abb1655.
[99] Venkatesh K, Santomartino SM, Sulam J, Yi PH. Code and Data Sharing Practices in the Radiology Artificial Intelligence Literature: A Meta-Research Study. Radiol Artif Intell 2022;4:e220081.
[100] Moassefi M, Rouzrokh P, Conte GM, Vahdati S, Fu T, Tahmasebi A, et al. Reproducibility of Deep Learning Algorithms Developed for Medical Imaging Analysis: A Systematic Review. J Digit Imaging 2023;36:2306–12.
[101] 2023 NIH data management and sharing policy n.d. https://oir.nih.gov/sourcebook/intramural-program-oversight/intramural-data-sharing/2023-nih-data-management-sharing-policy (accessed May 2, 2024).
[102] Sheller MJ, Edwards B, Reina GA, Martin J, Pati S, Kotrotsou A, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci Rep 2020;10:12598.
[103] Ger RB, Netherton TJ, Rhee DJ, Court LE, Yang J, Cardenas CE. Auto-contouring for Image-Guidance and Treatment Planning. In: El Naqa I, Murphy MJ, editors. Machine and Deep Learning in Oncology, Medical Physics and Radiology, Cham: Springer International Publishing; 2022, p. 231–93.
[104] Rong Y, Chen Q, Fu Y, Yang X, Al-Hallaq HA, Wu QJ, et al. NRG Oncology Assessment

of Artificial Intelligence Deep Learning-Based Auto-segmentation for Radiation Therapy: Current Developments, Clinical Considerations, and Future Directions. Int J Radiat Oncol Biol Phys 2023. https://doi.org/10.1016/j.ijrobp.2023.10.033.

[105] Riegel AC. Applications of Artificial Intelligence in Head and Neck Radiation Therapy n.d.

[106] Jensen K, Al-Farra G, Dejanovic D, Eriksen JG, Loft A, Hansen CR, et al. Imaging for Target Delineation in Head and Neck Cancer Radiotherapy. Semin Nucl Med 2021;51:59–67.

[107] Lin D, Wahid KA, Nelms BE, He R, Naser MA, Duke S, et al. E pluribus unum: prospective acceptability benchmarking from the Contouring Collaborative for Consensus in Radiation Oncology crowdsourced initiative for multiobserver segmentation. J Med Imaging (Bellingham) 2023;10:S11903.

[108] Baroudi H, Brock KK, Cao W, Chen X, Chung C, Court LE, et al. Automated Contouring and Planning in Radiation Therapy: What Is "Clinically Acceptable"? Diagnostics (Basel) 2023;13. https://doi.org/10.3390/diagnostics13040667.

[109] Wahid KA, Cardenas CE, Marquez B, Netherton TJ, Kann BH, Court LE, et al. Evolving Horizons in Radiotherapy Auto-Contouring: Distilling Insights, Embracing Data-Centric Frameworks, and Moving Beyond Geometric Quantification. Advances in Radiation Oncology 2024:101521.

[110] Kui X, Liu F, Yang M, Wang H, Liu C, Huang D, et al. A review of dose prediction methods for tumor radiation therapy. Meta-Radiology 2024;2:100057.

[111] Kelly BS, Judge C, Bollard SM, Clifford SM, Healy GM, Aziz A, et al. Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE). Eur Radiol 2022;32:7998–8007.

[112] Fang Y, Wang J, Ou X, Ying H, Hu C, Zhang Z, et al. The impact of training sample size on deep learning-based organ auto-segmentation for head-and-neck patients. Phys Med Biol 2021;66. https://doi.org/10.1088/1361-6560/ac2206.

[113] Tappeiner E, Pröll S, Fritscher K, Welk M, Schubert R. Training of head and neck segmentation networks with shape prior on small datasets. Int J Comput Assist Radiol Surg 2020;15:1417–25.

[114] Czolbe S, Arnavaz K, Krause O, Feragen A. Is Segmentation Uncertainty Useful? Information Processing in Medical Imaging, Springer International Publishing; 2021, p. 715–26.

[115] Mehta R, Filos A, Baid U, Sako C, McKinley R, Rebsamen M, et al. QU-BraTS: MICCAI BraTS 2020 Challenge on Quantifying Uncertainty in Brain Tumor Segmentation - Analysis of Ranking Scores and Benchmarking Results. J Mach Learn Biomed Imaging 2022;2022. https://doi.org/10.1007/978-3-030-46640-4_21.

[116] Wimmer L, Sale Y, Hofman P, Bischl B, Hüllermeier E. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In: Evans RJ, Shpitser I, editors. Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence, vol. 216, PMLR; 31 Jul--04 Aug 2023, p. 2282–92.

[117] Cecil J, Lermer E, Hudecek MFC, Sauer J, Gaube S. Explainability does not mitigate the negative impact of incorrect AI advice in a personnel selection task. Sci Rep 2024;14:9736.

[118] Tang JSN, Lai JKC, Bui J, Wang W, Simkin P, Gai D, et al. Impact of Different Artificial Intelligence User Interfaces on Lung Nodule and Mass Detection on Chest Radiographs. Radiol Artif Intell 2023;5:e220079.

[119] Albahri AS, Duhaim AM, Fadhel MA, Alnoor A, Baqer NS, Alzubaidi L, et al. A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. Inf Fusion 2023;96:156–91.

[120] McQuinlan Y, Brouwer CL, Lin Z, Gan Y, Sung Kim J, van Elmpt W, et al. An

investigation into the risk of population bias in deep learning autocontouring. Radiother Oncol 2023;186:109747.

[121] Mensah GB. Artificial intelligence and ethics: A comprehensive review of bias mitigation, transparency, and accountability in AI systems 2023. https://doi.org/10.13140/RG.2.2.23381.19685/1.

Appendices

Appendix A: Additional Figures

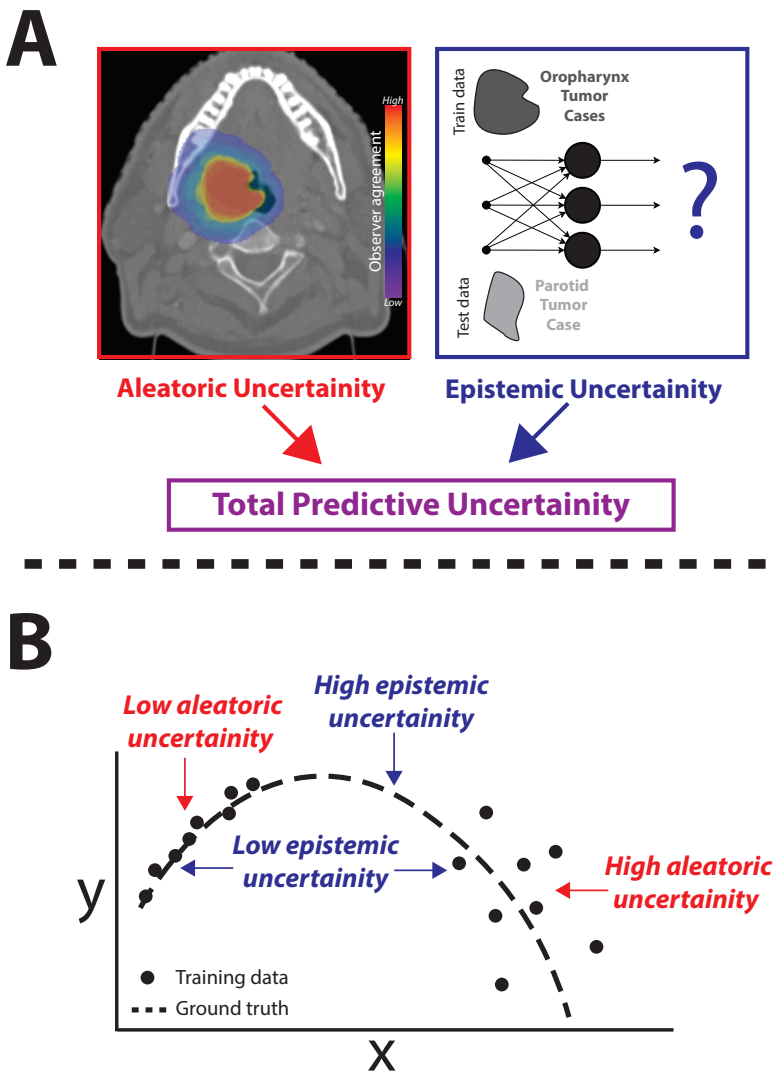

Figure A1. Illustrative examples of aleatoric and epistemic uncertainty concepts. (**A**) Left: A computed tomography image of an oropharyngeal cancer patient, overlaid with a probability map of interobserver agreement, illustrates aleatoric uncertainty in segmentation. Example data derived from expert contours from the Contouring Collaborative in Radiation Oncology (doi: 10.1038/s41597-023-02062-w). Right: A hypothetical tumor contouring model trained using oropharyngeal cancer cases would yield high epistemic uncertainty when presented with a parotid tumor case as a byproduct of insufficient training data. The combination of aleatoric and epistemic uncertainties contributes to the total predictive uncertainty. (**B**) A scatterplot of hypothetical variables x and y demonstrates high aleatoric uncertainty in regions with noisy data points and high epistemic uncertainty in regions with sparse data points.

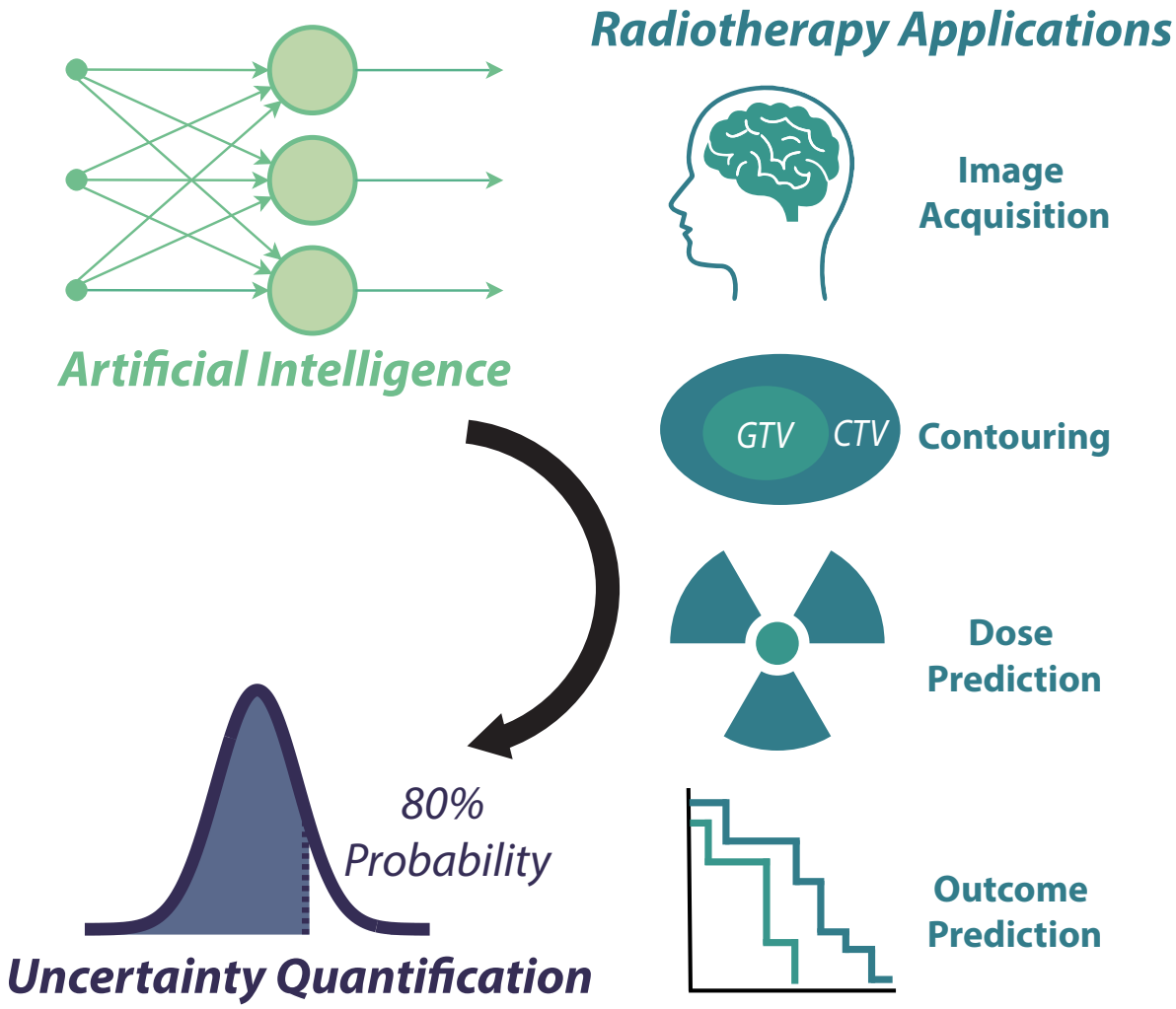

Figure A2. Study overview. This scoping review aims to comprehensively evaluate the literature on artificial intelligence models designed to quantify model uncertainty, specifically within the context of radiotherapy applications such as image acquisition, contouring, dose prediction, and outcome prediction, among others.

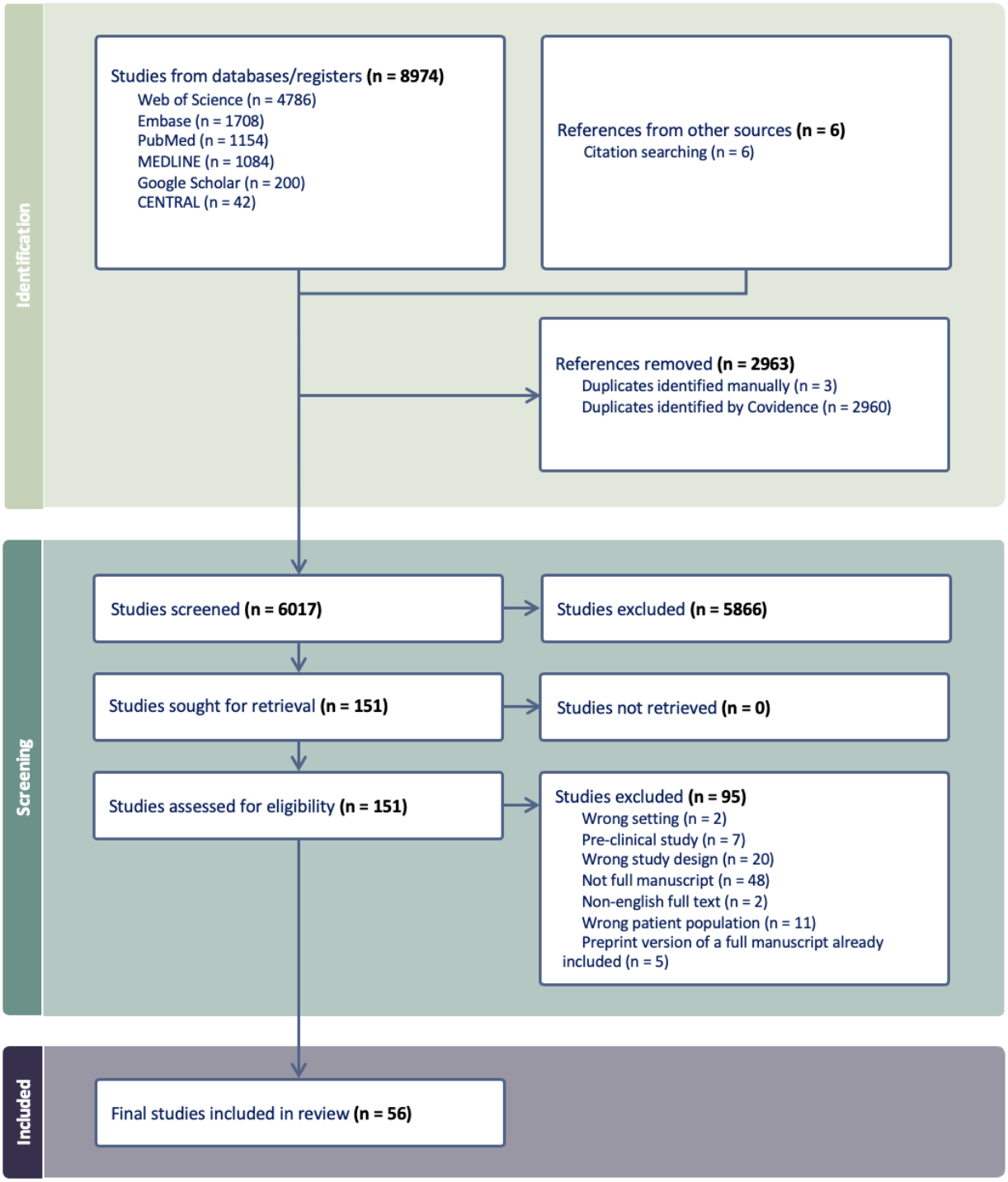

Figure A3. Preferred Reporting Items for Systematic Reviews and Meta-Analyses diagram illustrating systematic screening of identified studies. Ultimately, 56 studies out of the initially identified 8980 were included for the final analysis.

Appendix B: Additional information on manuscript screening and data extraction

Additional manuscript screening details

Two rounds of screening were performed by all reviewers which took into account all the inclusion criteria listed below. The initial screening (first round) was intended as a quick filtering process based on titles and abstracts to reduce the number of manuscripts into a workable size for eventual full-text screening (second round). Two reviewers (K.A.W. and Z.Y.K.) performed the screening process through Covidence which allows for rapid categorization of articles into inclusion/exclusion piles. Any disagreements were automatically flagged for additional review. Flagged cases underwent further scrutiny via virtual video meetings, where the two initial reviewers were joined by an independent, senior third reviewer (M.J.D.) to meticulously evaluate the contentious manuscripts. This collaborative review culminated in a final vote to decide whether the article merited inclusion. This process was repeated both for initial screening and full-text screening.

Population, concept, context (PCC) criteria for study inclusion:

1. *Population - Human patients undergoing radiotherapy for cancer treatment.* The study should explicitly mention that human patients that were actively undergoing radiotherapy, had plans to undergo radiotherapy, or had already completed radiotherapy were the subjects from which data were derived from. Studies using only preclinical samples (e.g., cell line, animal model, phantom studies, synthetic data) were excluded even if results could eventually be extrapolated to human patients. In rare circumstances where human and preclinical data was combined (e.g., mixed human data and phantom data), these studies were included. Moreover, in rare circumstances data derived from non-cancer patients were included only if directly applicable for radiotherapy specific indications (e.g., radioablation treatment of arrhythmia).
2. *Concept - Utilization of artificial intelligence and uncertainty quantification.* The study should explicitly mention that the underlying modeling technique is related to artificial intelligence or machine learning (e.g., deep learning related or more traditional methods), and must provide a method to quantify the uncertainty or confidence of the underlying model. Ideally, studies should explicitly list training and testing sample sizes, but this was not a strict requirement, particularly for older studies where this stratification was not yet standard. Studies only investigating underlying uncertainties of a radiotherapy related process (e.g., proton range uncertainty, segmentation interobserver variability, etc.) without any indication of a method to quantify predictive model uncertainty were not included.
3. *Context - Radiotherapy applications.* The study should explicitly mention that radiotherapy is the target application domain of the study or belong to a predefined list of radiotherapy applications recognized by the authors (image synthesis, image

registration, contouring, dose prediction, outcome prediction). Studies in other related but distinct medical application domains (e.g., diagnostic radiology, interventional radiology, surgical oncology, medical oncology) were excluded unless the study investigated multiple applications within the same paper (e.g., diagnostic radiology applications AND radiotherapy-related applications).

Additional criteria for study inclusion:

1. Full text must be accessible to screeners. Conference abstracts must be linked to a full text (e.g., conference proceeding) or were excluded from the search.
2. Full text must be available in written English, so that it could be appropriately evaluated by all screeners.

Additional data extraction details

Two human extractors worked in parallel to manually extract data from the final manuscripts. Specific extraction items are detailed below. These items were initially presented as a Covidence template that was used in the data extraction process and then refined if needed to fit into categorical values. All extractions were cross-checked by both reviewers (K.A.W., Z.Y.K.) and a final third reviewer (M.J.D.) when disagreements were found. Extracted data was transformed into machine readable format after initial collection based on agreement between reviewers using a version-controlled online Google Sheets document.

General Study Characteristics

1. Manuscript type - If the manuscript is a standard publication (i.e., published in a peer-reviewed journal), a conference proceeding (could be peer-reviewed or not), or a preprint. Articles extracted from preprint servers (e.g., arXiv) would be considered conference proceedings if explicitly indicated in the uploaded document (e.g., *this paper has been accepted to X conference*) and/or a corresponding entry was found on the conference website.
2. Publication year - Year of manuscript upload (in case of preprint) or year of publication as reported by publisher (in case of standard publication or conference proceeding).
3. Geographic location of the study authors - Which country the authors were from as determined from author affiliation information. If not all authors were from the same country, the following hierarchy was used. 1. Country where the majority of authors were from, 2. In the unlikely event of a tie, the country of the corresponding author was reported.
4. Code/data availability - If code and/or data were made publicly available. Relevant datasets DOIs and GitHub URLs were collected and reported where applicable.

RT Characteristics

1. Radiotherapy application space - What specific end use the manuscript is developing an artificial intelligence model for? Initially was collected with a free text option but was condensed into a categorical variable with the following possible values: dose planning, image correction, image registration, image synthesis, motion tracking, nodal

classification, outcome related, contouring.

2. Specific data types used - What input data is being used for the artificial intelligence models? Initially was collected with a free text option but was condensed into two categorical variables with the following possible values:
    a. Image data: CT, MRI, Multimodal, PET/CT, ultrasound, NA (i.e., none).
    b. Additional data: Clinical, dose, dose+clinical, dose+clinical+target+probablity map, fiducial, K-space, organ at risk, registration transforms, respiratory trace, target, target+clinical, target + organ at risk, NA (i.e., none)
3. Cancer type of patients in the study - What were the underlying diagnoses of the patients used in the study? Initially was collected with a free text option but was condensed into a categorical variable with the following possible values: brain, breast, cardiac, cervical, esophageal, head and neck, liver, lung, multiple, pancreatic, pelvic, prostate.

AI Characteristics

1. Algorithmic approach - What type of underlying algorithm was used in the study? Initially was collected with a free text option but was condensed into a separate free-text variable and a categorical variable with the following possible values:
    a. Machine learning type: Which overarching domain of machine learning the algorithm is categorized as: supervised, unsupervised, reinforcement, or mixed.
2. Training/validation/testing sample sizes - Specific numbers of training, validation, and testing datapoints used in the study. Data is extracted at most granular level (e.g., some algorithms use axial slices or images as input) and at the patient level. Could be NA if this information was not reported in the manuscript. We chose to focus on patient level data for reported data in our review since it was more clinically relevant and easier to compare between studies.
3. Characteristics of the validation/testing sets - How authors utilized validation and testing sets. Best practices often require separate hold-out sets but this may not always be feasible given dataset constraints. Initially was collected with a free text option but was condensed into two categorical variables with the following possible values:
    a. Validation type: Cross-validation, not specified, separate set.
    b. Testing type: Bootstrap, cross-validation, separate set [external], separate set [internal + external], separate set [internal], separate set [multiple external], other.

Uncertainty Quantification Characteristics

1. Uncertainty application category - What is the general use-case of the uncertainty methodology applied? Categories were adapted from existing literature (Kahl et al., [doi: 10.48550/arXiv.2401.08501], Lambert et al. [doi: 10.48550/arXiv.2210.03736]). Studies could investigate multiple applications simultaneously. The following specific categories were utilized:
    a. Active learning: Utilization of uncertainty estimates for improving the model training process.
    b. Ambiguity modeling: Comparison of model uncertainty estimates to a ground truth measure of uncertainty. For example, in segmentation, this could refer to

computing the normalized cross-correlation or the generalized energy distance between the pixel-wise model measures and pixel-wise ground truth probability measures.

   c. Calibration: Measurement of agreement between model estimated probabilities and true underlying data distribution probabilities. Popular methods of measuring calibration would include the Expected Calibration Error and the Brier score.
   d. Failure detection: Utilize numerical model uncertainty to determine which cases should be flagged for further inspection. For example, in a segmentation framework the uncertainty estimate could be correlated to a geometric value (e.g., DSC) and subsequently binarized to classify samples below and above an expected correlated geometric value. Related to Misclassification Detection Protocol and Rejection Protocol in Lambert et al. (doi: 10.48550/arXiv.2210.03736).
   e. Out-of-distribution detection: Conceptually similar to failure detection in that an uncertainty measure is used to flag cases. Typically requires a priori identification of in-distribution and out-of-distribution properties for samples (e.g., normal and abnormal images). Typically requires multiple external (out-of-distribution) datasets to implement.

2. Type of uncertainty quantification method used - Specific approach to calculate model uncertainty. Initially was collected with a free text option but was condensed into a single categorical variable. Studies could investigate multiple applications simultaneously. The following specific categories were utilized: Monte Carlo Dropout, Ensembles, Direct Softmax Output, Gaussian Process, Test-time Augmentation, Conformal Predictions, Evidential Deep Learning, Other Bayesian (an explicitly defined Bayesian approach that did not fall into a previous category), Other (bespoke approach developed in a specific paper that did not fall into a previous category).
3. Metrics used for UQ experiments - Any numerical indicators used in the computation of model uncertainty. Initially was collected with a free text option but was condensed into a single categorical variable. Studies could utilize multiple metrics simultaneously. The following specific categories were utilized: Entropy-based, Variance-based, Other (bespoke approach developed in a specific paper).
4. Self-described uncertainty type studied - Whether the study explicitly mentioned they were investigated epistemic and/or aleatoric uncertainty. Possible values of epistemic, aleatoric, both, or unspecified. Only explicit mentions of these terms (or related terms homoscedastic uncertainty and heteroscedastic uncertainty) in the manuscript were considered, otherwise this variable was labeled as unspecified (i.e., no inference about methods was performed on our part).
5. Utilization of quantitative and/or qualitative methods - Whether a uncertainty was presented in a quantitative and/or qualitative manner. Examples of qualitative experiments would include visualizing heatmap pixel-wise representations of model uncertainty in a segmentation problem.

Appendix C: Additional information on database search criteria

Ovid MEDLINE (R) ALL 1946 to November 17, 2023

#	Searches	Results
1	exp Artificial Intelligence/	183179
2	((artificial or machine or deep) adj (learning or intelligence)).ti,ab.	150066
3	("neural net*" or "support vector" or "decision tree" or "random forest" or "gradient boost*" or bagging or ensemble or radiom*).ti,ab.	198847
4	or/1-3 [AI]	370737
5	Uncertainty/	18287
6	(uncertain* or aleatoric or epistemic or "monte carlo*" or dropout or Bayes* or "conformal prediction" or "variational inference" or "temperature scaling" or platt or entropy).ti,ab.	400538
7	or/5-6 [Uncertainty]	404310
8	4 and 7 [AI + Uncertainty]	24799
9	exp Radiotherapy/	208974
10	exp Radiotherapy Planning, Computer-Assisted/	25613
11	exp Radiation Oncology/	5869
12	(radiotherap* or "radio-therap*" or irradiat* or radiat* or chemoradi* or radiochemo* or "chemo-radi*" or "radio-chemo*" or "intensity modulated" or IMRT or EBRT or photon* or proton* or radiosurgery or "radio-surgery" or brachytherapy or "brachy-therapy").ti,ab.	1086055
13	or/9-12 [Radiotherapy]	1129582
14	8 and 13 [AI + Uncertainty + Radiotherapy]	1084

Ovid Embase Classic+Embase 1947 to 2023 November 17

#	Searches	Results
1	exp Artificial Intelligence/	89799
2	((artificial or machine or deep) adj (learning or intelligence)).ti,ab.	176472
3	("neural net*" or "support vector" or "decision tree" or "random forest" or "gradient boost*" or bagging or ensemble or radiom*).ti,ab.	234393
4	or/1-3 [AI]	385870
5	Uncertainty/	50932

6	(uncertain* or aleatoric or epistemic or "monte carlo*" or dropout or Bayes* or "conformal prediction" or "variational inference" or "temperature scaling" or platt or entropy).ti,ab.	483767
7	or/5-6 [Uncertainty]	489765
8	4 and 7 [AI + Uncertainty]	27219
9	exp Radiotherapy/	736137
10	Radiation Oncology/	7613
11	(radiotherap* or "radio-therap*" or irradiat* or radiat* or chemoradi* or radiochemo* or "chemo-radi*" or "radio-chemo*" or "intensity modulated" or IMRT or EBRT or photon* or proton* or radiosurgery or "radio-surgery" or brachytherapy or "brachy-therapy").ti,ab.	1462005
12	or/9-11 [Radiotherapy]	1652692
13	8 and 12 [AI + Uncertainty + Radiotherapy]	1708

PubMed (NLM)

(("artificial intelligence"[MeSH Terms] OR "artificial learning"[Title/Abstract] OR "artificial intelligence"[Title/Abstract] OR "machine learning"[Title/Abstract] OR "machine intelligence"[Title/Abstract] OR "deep learning"[Title/Abstract] OR "deep intelligence"[Title/Abstract] OR "neural net*"[Title/Abstract] OR "support vector"[Title/Abstract] OR "decision tree"[Title/Abstract] OR "random forest"[Title/Abstract] OR "gradient boost*"[Title/Abstract] OR "bagging"[Title/Abstract] OR "ensemble"[Title/Abstract] OR "radiom*"[Title/Abstract])

AND

("uncertainty"[MeSH Terms] OR "uncertain*"[Title/Abstract] OR "aleatoric"[Title/Abstract] OR "epistemic"[Title/Abstract] OR "monte carlo*"[Title/Abstract] OR "dropout"[Title/Abstract] OR "bayes*"[Title/Abstract] OR "conformal prediction"[Title/Abstract] OR "variational inference"[Title/Abstract] OR "temperature scaling"[Title/Abstract] OR "platt"[Title/Abstract] OR "entropy"[Title/Abstract])

AND

("radiotherapy"[MeSH Terms] OR "radiotherapy planning, computer assisted"[MeSH Terms] OR "radiation oncology"[MeSH Terms] OR "radiotherap*"[Title/Abstract] OR "radio therap*"[Title/Abstract] OR "irradiat*"[Title/Abstract] OR "radiat*"[Title/Abstract] OR "chemoradi*"[Title/Abstract] OR "radiochemo*"[Title/Abstract] OR "chemo radi*"[Title/Abstract] OR "radio chemo*"[Title/Abstract] OR "intensity modulated"[Title/Abstract] OR "IMRT"[Title/Abstract] OR "EBRT"[Title/Abstract] OR

"photon*"[Title/Abstract] OR "proton*"[Title/Abstract] OR "radiosurgery"[Title/Abstract] OR "radio-surgery"[Title/Abstract] OR "brachytherapy"[Title/Abstract] OR "brachy-therapy"[Title/Abstract])) Results = 1154

Cochrane Library (Wiley)

ID	Search	Results
#1	MeSH descriptor: [Artificial Intelligence] explode all trees	2958
#2	("artificial learning" or "artificial intelligence" or "machine learning" or "machine intelligence" or "deep learning" or "deep intelligence" or "support vector" or "decision tree" or "random forest" or bagging or ensemble or radiom*):ti,ab or ((neural NEXT net*) or (gradient NEXT boost*)):ti,ab	7392
#3	{or #1-#2}	8923
#4	MeSH descriptor: [Uncertainty] explode all trees	421
#5	(uncertain* or aleatoric or epistemic or dropout or Bayes* or "conformal prediction" or "variational inference" or "temperature scaling" or platt or entropy):ti,ab or (monte NEXT carlo*):ti,ab	28181
#6	{or #4-#5}	28249
#7	#3 and #6	609
#8	MeSH descriptor: [Radiotherapy] explode all trees	10381
#9	MeSH descriptor: [Radiotherapy Planning, Computer-Assisted] explode all trees	467
#10	MeSH descriptor: [Radiation Oncology] explode all trees	82
#11	(radiotherap* or irradiat* or radiat* or chemoradi* or radiochemo* or "intensity modulated" or IMRT or EBRT or photon* or proton* or radiosurgery or "radio-surgery" or brachytherapy or "brachy-therapy"):ti,ab or (((radio NEXT therap*) or (radio NEXT chemo*) or (chemo NEXT radi*))):ti,ab	63161
#12	{or #8-#11}	64059
#13	#7 and #12	42

Web of Science Core Collection (Clarivate)

Entitlements: WOS.IC: 1993 to 2023; WOS.CCR: 1985 to 2023; WOS.SCI: 1900 to 2023; WOS.AHCI: 1975 to 2023; WOS.BHCI: 2005 to 2023; WOS.BSCI: 2005 to 2023; WOS.ESCI: 2005 to 2023; WOS.ISTP: 1990 to 2023; WOS.SSCI: 1900 to 2023; WOS.ISSHP: 1990 to 2023

ID	Search	Results

#1	TI=((artificial or machine or deep) NEAR/1 (learning or intelligence)) OR AB=((artificial or machine or deep) NEAR/1 (learning or intelligence))	560769
#2	TI=("neural net*" OR "neural net*" or "support vector" or "decision tree" or "random forest" or "gradient boost*" or bagging or ensemble or radiom*) OR AB=("neural net*" or "support vector" or "decision tree" or "random forest" or "gradient boost*" or bagging or ensemble or radiom*)	988580
#3	#2 OR #1	1344163
#4	TI=(uncertain* or aleatoric or epistemic or "monte carlo*" or dropout or Bayes* or "conformal prediction" or "variational inference" or "temperature scaling" or platt or entropy) OR AB=(uncertain* or aleatoric or epistemic or "monte carlo*" or dropout or Bayes* or "conformal prediction" or "variational inference" or "temperature scaling" or platt or entropy)	1533339
#5	TI=(radiotherap* or "radio-therap*" or irradiat* or radiat* or chemoradi* or radiochemo* or "chemo-radi*" or "radio-chemo*" or "intensity modulated" or IMRT or EBRT or photon* or proton* or radiosurgery or "radio-surgery" or brachytherapy or "brachy-therapy") OR AB=(radiotherap* or "radio-therap*" or irradiat* or radiat* or chemoradi* or radiochemo* or "chemo-radi*" or "radio-chemo*" or "intensity modulated" or IMRT or EBRT or photon* or proton* or radiosurgery or "radio-surgery" or brachytherapy or "brachy-therapy")	2578649
#6	#3 AND #4 AND #5	4358

Web of Science Preprint Citation Index (Clarivate)

ID	Search	Results
#1	TI=((artificial or machine or deep) NEAR/1 (learning or intelligence)) OR AB=((artificial or machine or deep) NEAR/1 (learning or intelligence))	79524
#2	TI=("neural net*" OR "neural net*" or "support vector" or "decision tree" or "random forest" or "gradient boost*" or bagging or ensemble or radiom*) OR AB=("neural net*" or "support vector" or "decision tree" or "random forest" or "gradient boost*" or bagging or ensemble or radiom*)	99437
#3	#2 OR #1	153713

#4	TI=(uncertain* or aleatoric or epistemic or "monte carlo*" or dropout or Bayes* or "conformal prediction" or "variational inference" or "temperature scaling" or platt or entropy) OR AB=(uncertain* or aleatoric or epistemic or "monte carlo*" or dropout or Bayes* or "conformal prediction" or "variational inference" or "temperature scaling" or platt or entropy)	147982
#5	TI=(radiotherap* or "radio-therap*" or irradiat* or radiat* or chemoradi* or radiochemo* or "chemo-radi*" or "radio-chemo*" or "intensity modulated" or IMRT or EBRT or photon* or proton* or radiosurgery or "radio-surgery" or brachytherapy or "brachy-therapy") OR AB=(radiotherap* or "radio-therap*" or irradiat* or radiat* or chemoradi* or radiochemo* or "chemo-radi*" or "radio-chemo*" or "intensity modulated" or IMRT or EBRT or photon* or proton* or radiosurgery or "radio-surgery" or brachytherapy or "brachy-therapy")	142649
#6	#3 AND #4 AND #5	428

Google Scholar (first 200 results)

(artificial learning OR machine learning OR deep learning OR artificial intelligence OR machine intelligence OR deep intelligence OR neural network OR neural networks OR neural networking OR support vector OR support vectors OR decision tree OR decision trees OR random forest OR gradient boost OR gradient boosts OR bagging OR ensemble OR radiomic OR radiometric OR radiomorhometric) AND (uncertainty OR aleatoric OR epistemic OR "monte carlo*" OR dropout OR Bayes OR "conformal prediction" OR "variational inference" OR "temperature scaling" OR platt OR entropy) AND (radiotherapy OR irradiation OR radiation OR chemoradiation OR radiochemotherapy OR "intensity modulated" OR IMRT OR EBRT OR photon* OR proton* OR radiosurgery OR brachytherapy)

Key Articles

To ensure a comprehensive inclusion of relevant articles, the following PubMed IDs were used as “key articles” in shaping our initial search queries: “33179605" or "33503599" or "33778184" or "34111573" or "36112996" or "36484346" or "36865296" or "37414257" or "37820691".