Domain and User-Centered Machine Learning for Medical Image Analysis

by

Katharina Viktoria Hoebel
B.S., Kiel University (2016)
M.D., Heidelberg University (2017)

Submitted to the Harvard-MIT Program in Health Sciences and Technology
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Medical Engineering and Medical Physics
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June 2023

Authored by .....
Katharina Viktoria Hoebel
Harvard-MIT Program in Health Sciences and Technology
February 15, 2023

Certified by .....
Jayashree Kalpathy-Cramer, PhD
Professor of Ophthalmology, Chief of Division of Artificial Medical Intelligence in Ophthalmology, University of Colorado School of Medicine
Visiting Professor in Radiology, Harvard Medical School
Thesis Supervisor

Accepted by .....
Collin M. Stultz, MD, PhD
Director, Harvard-MIT Program in Health Sciences and Technology
Nina T. and Robert H. Rubin Professor in Medical Engineering and Science
Professor or Electrical Engineering and Computer Science

Domain and User-Centered Machine Learning for Medical Image Analysis

by

Katharina Viktoria Hoebel

Submitted to the Harvard-MIT Program in Health Sciences and Technology
on February 15, 2023, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Medical Engineering and Medical Physics

Abstract

The utilization of diagnostic imaging in the United States and worldwide is steadily growing. Due to a shortage of trained staff, the result is an increased and unsustainable workload for radiologists. Consequently, there is a high clinical need for the automation of cognitively challenging tasks, such as analyzing and interpreting medical images, to lighten the burden on radiologists and avoid a further increase in healthcare expenditure. Machine learning (ML), including deep learning (DL) offer a potential solution as these algorithms can learn to automatically recognize subtle patterns from large amounts of data and augment clinical decision-making.

Despite the high enthusiasm for ML algorithms, concerns regarding their readiness for clinical deployment are impeding their clinical translation. In this thesis, we address three fundamental challenges to the translation of ML algorithms into clinical care settings.

First, algorithms must perform robustly in routine clinical care settings. We demonstrate how appropriate image preprocessing improves the stability of hand-crafted radiomic features extracted from brain MRIs. Second, the selected network design must be appropriate for a specific task. Here, we illustrate the advantages of shifting from a strictly discrete (ordinal) model of disease severity distribution to a continuously valued one. We introduce a generalized framework that can recover information lost by discretizing continuous variables into discrete training labels. Furthermore, disagreements in the labels generated by different annotators can be caused by individually varying decision thresholds. Therefore, we present the first design and demonstration of two methods that enable the joint learning of annotators' ordinal classification and their individual biases for a latent, continuously valued target variable like disease severity. Lastly, the performance of ML algorithms needs to be evaluated in a clinically meaningful manner. We address the disconnect between the subjective quality perception of clinical experts and the metrics that are typically used to evaluate performance. Furthermore, we identify criteria that experts use to evaluate the quality of automatically generated segmentations and describe their thought processes as they correct them.

Based on the learnings from our work, we conclude with concrete recommendations for developing robust and trustworthy ML tools for medical imaging.

Thesis Supervisor: Jayashree Kalpathy-Cramer, PhD
Title: Professor of Ophthalmology, Chief of Division of Artificial Medical Intelligence in Ophthalmology, University of Colorado School of Medicine
Visiting Professor in Radiology, Harvard Medical School

Acknowledgments

Doing a PhD is like climbing Mount Everest; it is a long journey that is extremely rewarding – but it is also exhausting, probably even painful at times, and you should not embark on it unprepared and alone. I could not have wished for a better tour guide than Jayashree Kalpathy-Cramer. Jayashree showed me that doing serious research does not mean you cannot have fun along the way. Her compassionate group leadership fostered an environment where every group member could thrive. Jayashree guided me but also gave me the freedom to explore and the encouragement to find my own path.

I am immensely thankful for the guidance from my thesis committee: Polina Golland, Bruce Fischl, and Clifton (Dave) Fuller. Thank you for gently pushing me and shaping my research in new directions.

Being part of the QTIM lab for five years was an extremely rewarding experience. QTIM is a group of dedicated and caring individuals whom I have had the honor to learn from and work with. The lively debates about research, technology, and life have shaped my research interests in new directions. Also, I doubt that any other lab makes better homemade pasta than we do. In Elizabeth Gerstner, I found a mentor and collaborator who is always willing to share her experience as a clinician-scientist and has contributed critically to my research. Chris Bridge taught me that sometimes patience and persistence are among the most valuable assets of a scientist. He has provided me with numerous critical ideas and feedback that have helped me over many hurdles in the last few years.

Throughout my time at QTIM, I have worked with many supportive and collaborative group members: Yi-Fen Yen, Sunakshi Paul, Dania Daye, Samarth Nandekar, Kevin Lou, Ikbeom Jang, Praveer Singh, Matthew Li, Mehak Aggarwal, Sharut Gupta. I am particularly thankful to James Brown, Andrew Beers, and Ken Chang for teaching me that deep learning is not magic but a craft, to Albert Kim and Ben Bearce for countless coffee breaks and watercooler chats, and to Andréanne Lemay and Eshika Saxena for letting me accompany them a part of their research path. I am glad to have been on this excursion with the best team of fellow QTIM grad students I could have wished for: Mishka Gidwani, Jay Patel, Ken Chang, and Syed Rakin Ahmed.

Outside of QTIM, the Martinos Center, NTP, and HST communities were steadfast sources of energy and support. I am particularly indebted to my mentors, Randy Gollup, Bruce Rosen, and Robert Barry. You spent countless hours listening to me as I changed my plans over and over, celebrated small and big successes with me, and supported me up to the mountaintop. I could not have wished for a better program for my PhD than the Health Sciences and Technology program. HST's investment in interdisciplinary research to better healthcare made me feel like I found a home for my research interests. I am thankful to the staff and faculty who make the HST community so special and welcoming: Julie Greenberg, Laurie Ward, Traci Anderson, Joe Stein, and Patty Cunningham. During my PhD, I also had the privilege to be the teaching assistant for my favorite HST class with Rick Mitchell and Bobby Padera - I learned so much from both of you and took away some precious lessons. Lastly,

I want to thank Brett Bouma. When I was a medical student, only starting to toy with the idea of pursuing a PhD, Brett opened the door to his lab and, with it, to the Boston research community and introduced me to the fantastic world of HST.

Outside of HST and the Martinos Center, I have had the honor to collaborate with many excellent and dedicated individuals, including Peter Campbell, Susan Ostmo, Michael Chiang, and Deniz Erdogmus from the i-ROP consortium. During my PhD journey, I also had the pleasure of spending time in Switzerland with the MedGIFT group at HES-SO Valais-Wallis and working with Henning Müller, Adrien Deperusinge, Mara Graziani, Vincent Andrearczyk, Amjad Khan, and Anjani Dhrangadhariya.

I have been fortunate to have met and spent time with many kind individuals who are very special to me. My extended HST cohort: Ashwin Kumar, Taylor Cannon, Brennan Jackson, Emily Alsentzer, Andrew Goldberg, Grissel Cervantes Jaramillo, Hyun Geun Song, Ben Leaker, Nathan Miller, Lily Wang, and Olivia Waring - you made Cambridge feel like home from the start.

I cannot imagine making it up to the mountain top without my truly amazing friends, many of whom I met through organizing the German American Conference: Jonas Lehman, Chris Kranzinger, Philipp Simons, Anshi Wittmann, Affi Maragh, Rahel Dette, Marius Vollberg, Paloma Ocola, Max Biggs, and Rajib Mondal.

Finally, I want to thank the people in my basecamp: My partner Tim Menke – my PhD journey would have been much less fun without you and your support. I look forward to more mountain tops to explore with you. My family, my parents, Birgit and Wolfgang, and my sister Conny – thank you for accompanying me on ups and downs, for your patience, everlasting support, and the never-ending supply of cat pictures – I would not have made it without you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Artificial intelligence – the industrial revolution 4.0

Artificial intelligence (AI) will tremendously impact our everyday life [2], and it has already done so in ways that we may not be fully aware of. It allows us to unlock our phones using face ID, curates news and social media feeds and secures our bank accounts by automatically detecting and blocking fraudulent transactions. AI is expected to transform most, if not all, industrial sectors, and its introduction has been likened to the invention of electricity [3]. The power of modern AI is its ability to create value from data. By this analogy, according to an Economist article published in 2017 [4], in the AI age, data is the new oil, and the few internet companies that are in control of most of the world's data flow wield tremendous power.

Artificial intelligence is "*the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings*," as the Encyclopedia Britannica defines it [5]. Since its inception as the research field of "thinking machines" at a conference at Dartmouth College in 1956, AI has come a long way.

Artificial Intelligence, Machine Learning, and Deep Learning Nowadays, the terms artificial intelligence, machine learning (ML), and deep learning (DL) are often

used interchangeably. However, ML is a branch of AI, and DL is a subfield of ML. ML algorithms make predictions about an input by learning from large amounts of data. The algorithms uncover connections within the available data that often remain hidden from experts and can improve their performance with access to even more data. While in classical ML, the features an algorithms uses to make its predictions are defined by the human user, in DL, it can identify the features within the data without human support using neural networks with hidden layers. As a result, deep learning algorithms depend less on human expertise but are usually more data-hungry than classical ML algorithms.

1.2 Machine learning in medicine

Healthcare systems worldwide are increasingly digitized through the introduction of electronic health records (EHR) and the picture archiving and communication system (PACS) for digital imaging data. The digitalization of healthcare data has created the prerequisite for developing and deploying powerful and data-hungry ML algorithms for the healthcare sector. While critical voices say that medicine has been slow to embrace the potential of ML, the question is not if but how ML will shape the future of healthcare [6]. Indeed, healthcare systems worldwide are in dire need of algorithmic support. In the US and other developed countries, healthcare professionals face a steadily growing amount of data that requires analysis and integration into a final diagnosis and treatment plan [7, 8]. However, the current shortage of physicians is expected to increase in the upcoming decades as the demand grows faster than the supply [9]. In addition, provider burnout and decreased professional satisfaction, intensified by the COVID-19 pandemic, plague the existing workforce [10]. Therefore, the introduction of ML to healthcare comes at a critical time as there is a high desire to automate routine jobs [11] and cut healthcare expenditure [12, 13]. In middle to low-resource countries, AI algorithms can, given the proper computational infrastructure, be used to import medical expertise from abroad to support the limited staff on the ground [14, 15]. However, the World Health Organization projects that healthcare

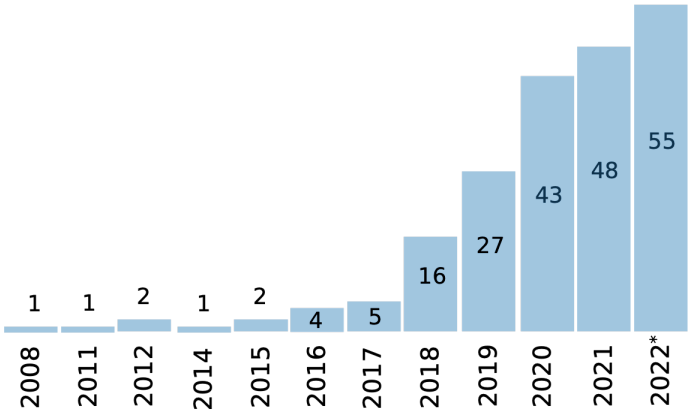

Figure 1-1: **Number of AI algorithms cleared by the FDA.** Sum of AI algorithms that have achieved clearance by the Food and Drug Administration each year starting from 2008. This figure has been adapted from the American College of Radiology (ACR) AI Central. *: as of November 15, 2022 [1]

systems will only benefit from the introduction of ML algorithms if ethics and human rights are core values during the development, deployment, and continued use [16].

Promising applications of ML exist in almost every medical discipline, such as the automatic detection of arrhythmia in cardiology [17], the discovery of signs of depression in psychiatry [18], and the identification of Parkinson's disease from breathing patterns in neurology [19]. Medical imaging, a key diagnostic tool in medicine, is particularly amenable to applying ML methods. Radiology was a trendsetter within medicine, transitioning from analog image acquisition on film to digital in the early 1990s [20]. Nowadays, the PACS system to manage image distribution and diagnostic interpretation is ubiquitous [21]. In addition, radiologists already have decades of experience in working with algorithmic assistance; the first computer-assisted algorithms for the early detection of breast cancer were introduced in the late 1990s [22]. Not surprisingly, radiology has been one of the early adopters of AI in medicine and has seen a steep increase in AI applications that achieved FDA clearance (see Figure 1-1 [23, 1].

Within the medical imaging workflow, as depicted in Figure 1-2, ML algorithms can support the radiology workforce at almost every step. ML algorithms ensure optimal utilization of existing resources [24], design optimized image acquisition protocols

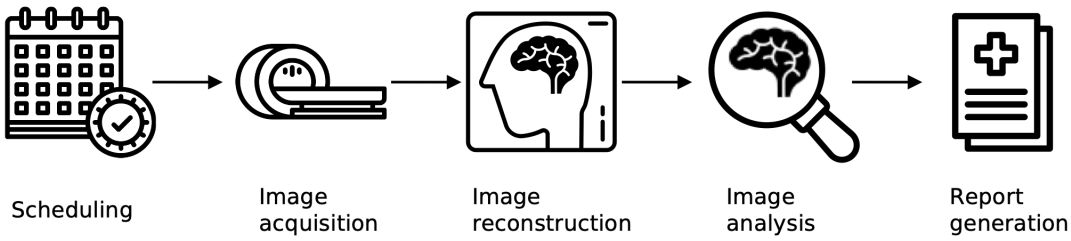

Figure 1-2: **Typical radiology workflow.** Illustration of the steps within a typical workflow in diagnostic radiology at which machine learning algorithms can support the clinical staff.

[25], improve image reconstruction [26], and automatically generate radiology reports [27]. We will focus on using ML for image analysis and interpretation in the following. AI-powered algorithms are already widely deployed, e.g., in recommendation systems, but the clinical deployment of ML image analysis algorithms can still be considered to be in its infancy. While the enthusiasm for the potential of ML for image analysis is unwavering, research has highlighted concerns regarding the readiness for clinical deployment [28, 29]. ML's safety is a significant concern in clinical settings, as even minor errors can seriously harm patients.

The data in medicine is generated in a non-stationary environment, characterized by changing practices, e.g., what is considered normal versus abnormal, and shifts in the patient populations [30]. Therefore, algorithms deployed in the clinic must be robust to these changes over time. However, ML algorithms are sensitive to shifts in the data distribution. As a consequence, their performance decreases under shifting dataset conditions [31] and can fail to provide accurate predictions without prior notice. Furthermore, current ML models are far from providing generalizability across clinical populations [32]. Additionally, they suffer from low test-retest repeatability (as described in Section 3.7.2) and are oftentimes poorly calibrated [33]. Lastly, the metrics that research articles use to report the performance of their newly developed ML do not necessarily represent what matters most to clinicians and their patients - their clinical efficacy [34]. This term has also been coined the "AI chasm" [35].

1.3 Thesis approach

This thesis will address three selected fundamental challenges to the translation of ML algorithms into clinical care settings.

Robustness of ML in routine clinical settings To earn physicians' trust and prove its reliability in clinical settings, it is of utmost importance that the predictions of an ML algorithm are robust. Model performance should not be affected by small imperceptible changes unrelated to the underlying disease. One way to measure this is through test-retest repeatability [36]. While human performance decreases due to physiological factors like hunger, exhaustion, and distraction [37], algorithms are often perceived as stable and objective. However, even supposedly objective machine learning algorithms can be confused, albeit by other factors than humans, and classical ML and DL models face issues with their test-retest repeatability [38, 39, 40]. Minor changes in the images imperceptible to the human eye likely cause these problems. A lack of repeatability hints towards unstable model performance and can lead to substantial clinical errors. Therefore, it is of high importance to limit test-retest variability. We hypothesize that the repeatability of hand-crafted features used for classical ML approaches can be improved by using appropriate image preprocessing steps.

Selecting appropriate network designs While disease severity is a continuous spectrum [41], discrete disease severity classes simplify diagnostics and treatment decisions in clinical practice. However, in addition to the severity class, the position of a case on the continuous spectrum contains valuable clinical information that is not captured by discrete ordinal classification. Following a trend well known in the statistics community and aptly coined "dichotomania" [42], researchers mostly use neural network architectures that are designed to classify nominal categories ignoring the inherent ordinal relationship between the ground truth classes and underlying continuously valued variable [43, 44, 45]. Choosing an inappropriate network design leads to a loss of information and inconsistent behavior for boundary cases [46].

Furthermore, the development of reliable classification models is complicated by noise in the ordinal labels which arises from loosely defined class boundaries and the inability to measure the variable of interest directly. For continuously valued variables, experts tend to disagree mostly about cases close to a decision boundary [47]. Some experts consistently over-call, meaning they have a lower threshold of what can be considered "disease," while others tend to under-call. We hypothesize that the predictive performance can be further improved for continuously valued variables by choosing network designs that respect the underlying distribution and systematic noise in the training labels.

Clinically meaningful evaluation of model performance The performance of segmentation models is routinely assessed through quantitative metrics. The most common metrics are the Dice score or Hausdorff distance between automatic segmentations and their manual ground truth. However, there is a mismatch between human perception of the quality of a segmentation and how popular metrics quantify it [48, 49, 50].

For tumor segmentations on medical imaging, clinical experts may perceive a slight but clinically meaningful aberration as a graver error than an error consisting of a larger volume that would not have any clinical consequences. However, a deeper understanding of the reasons for this mismatch is required for developing segmentation quality evaluation processes that reflect the clinical usefulness of a segmentation. We hypothesize that human experts' perception of brain tumor segmentation quality is context-dependent and aim to identify factors that influence experts' segmentation quality perception.

1.4 Thesis organization

This thesis is structured as follows:

Chapter 2 introduces the background and prior works that are essential to understanding the work presented in this thesis. We provide an overview of two

main methods for supervised machine learning in medical image analysis: radiomics and deep learning using convolutional neural networks. We also review the data annotation process, image preprocessing, and the metrics used to evaluate segmentation algorithms.

In **Chapter 3**, we demonstrate how appropriate image preprocessing can improve the test-retest repeatability of hand-crafted, so-called radiomic features, extracted from brain MRIs of patients newly diagnosed with glioblastoma. Utilizing a unique test-retest dataset of multimodal brain MRI, we assess the repeatability of radiomic features extracted using a popular open-source software package. We assess the influence of common preprocessing strategies on the stability of the features under test-retest conditions and infer guidelines to increase the repeatability of said features.

In **Chapter 4**, we illustrate the advantages of shifting from a strictly discrete (ordinal) view of disease severity to a continuously valued one.

Section 4.4 introduces a generalized framework to predict continuously valued variables using only discrete ordinal labels for model development. We train deep learning models utilizing widely available discrete ordinal labels and convert the models' outputs to continuous scores. The quality of the continuously valued predictions is evaluated using labels on a finer scale than the training ground truth. We demonstrate that it is possible to develop models that can recover the information lost by discretizing the continuous target variable on three datasets: disease severity prediction for retinopathy of prematurity and knee osteoarthritis and breast density estimation from mammograms.

In **Section 4.5**, we show the first conception and demonstration of methods that enable the joint learning of annotators' ordinal classification and their individual biases for a latent, continuously valued target variable. These methods use individual annotators' ordinal labels for training and learn the individual biases with as little as one label per training sample. We demonstrate that the classification performance of ordinal classes can be improved by learning the individual biases of each annotator.

In **Chapter 5**, we address the "chasm" between the evaluation metrics and clinical utility for the segmentation of postoperative brain tumors.

Section 5.3 presents a study of the expert-centered evaluation of deep learning algorithms for brain tumor segmentation. We surveyed the current literature on brain tumor segmentation algorithms for the reported quality evaluation. In addition, we asked expert readers to rate the quality of postoperative brain tumor segmentations subjectively and show that the quality perception of experts does not correlate well with the popular segmentation quality metrics from the literature. Furthermore, the agreement between the quality ratings of individual experts is low.

In **Section 5.4**, we showcase a multiple-method study exploring how clinical experts perceive the quality of postoperative brain tumors. We performed a study consisting of a questionnaire and interview on the perception of the segmentation quality of postoperative brain tumors with ten experts from neuro-oncology and neuroradiology. In the absence of a clear tumor boundary, we illustrate how the clinical context influences experts' quality perception of brain tumor segmentation.

1.5 Omitted Works

I also contributed to the following publications, conference proceedings, and preprints throughout my PhD studies. Some short excerpts from them have been included in this dissertation. In addition, observations and insights from these works have been crucial to developing the described studies.

- Mishka Gidwani, Ken Chang, Jay Biren Patel, **Katharina Hoebel**, Syed Rakin Ahmed, Praveer Singh, Clifton David Fuller, Jayashree Kalpathy-Cramer. "Inconsistent Partitioning and Unproductive Feature Associations Yield Idealized Radiomic Models", Radiology (2022): 220715.

- Andréanne Lemay, **Katharina Hoebel**, Christopher Bridge, Brian Befano, Silvia De Sanjosé, Didem Egemen, Ana Cecilia Rodriguez, Mark Schiffman, John Peter Campbell, and Jayashree Kalpathy-Cramer. "Improving the repeatability of deep learning models with Monte Carlo dropout." npj Digital Medicine 5, no. 1 (2022): 1-11.

- Charles Lu, Andréanne Lemay, Ken Chang, **Katharina Hoebel**, and Jayashree Kalpathy-Cramer. "Fair conformal predictors for applications in medical imaging." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 11, pp. 12008-12016. 2022.

- **Katharina Hoebel**, Christopher Bridge, Andréanne Lemay, Ken Chang, Jay Patel, Bruce Rosen, and Jayashree Kalpathy-Cramer. "Do I know this? segmentation uncertainty under domain shift." In Medical Imaging 2022: Image Processing, vol. 12032, pp. 261-276. SPIE, 2022.

- Raghav Mehta, Angelos Filos, Ujjwal Baid, Chiharu Sako, Richard McKinley, Michael Rebsamen, Katrin Dätwyler, Raphael Meier, Piotr Radojewski, Gowtham Krishnan Murugesan, Sahil Nalawade, Chandan Ganesh, Ben Wagner, Fang F Yu, Baowei Fei, Ananth J Madhuranthakam, Joseph A Maldjian, Laura Daza, Catalina Gómez, Pablo Arbeláez, Chengliang Dai, Shuo Wang, Hadrien Reynaud, Yuanhan Mo, Elsa Angelini, Yike Guo, Wenjia Bai, Subhashis Banerjee, Linmin Pei, Murat AK, Sarahi Rosas-González, Ilyess Zemmoura, Clovis Tauber, Minh Hoang Vu, Tufve Nyholm, Tommy Löfstedt, Laura Mora Ballestar, Veronica Vilaplana, Hugh McHugh, Gonzalo Maso Talou, Alan Wang, Jay Patel, Ken Chang, **Katharina Hoebel** et al. "QU-BraTS: MICCAI BraTS 2020 Challenge on QuantifyingUncertainty in Brain Tumor Segmentation-Analysis of Ranking Scores and Benchmarking Results." Journal of Machine Learning for Biomedical Imaging 1 (2022).

- Nishanth Arun, Nathan Gaw, Praveer Singh, Ken Chang, Mehak Aggarwal, Bryan Chen, **Katharina Hoebel**, Sharut Gupta, Jay Patel, Mishka Gidwani, Julius Adebayo, Matthew D Li, and Jayashree Kalpathy-Cramer. "Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging." Radiology: Artificial Intelligence 3, no. 6 (2021).

- Charles Lu, Andreanne Lemay, **Katharina Hoebel**, and Jayashree Kalpathy-Cramer. "Evaluating subgroup disparity using epistemic uncertainty in mam-

mography." arXiv preprint arXiv:2107.02716 (2021).

- Sharut Gupta, Praveer Singh, Ken Chang, Liangqiong Qu, Mehak Aggarwal, Nishanth Arun, Ashwin Vaswani, Shruti Raghavan, Vibha Agarwal, Mishka Gidwani, **Katharina Hoebel**, Jay Patel, Charles Lu, Christopher Bridge, Daniel Rubin, and Jayashree Kalpathy-Cramer. "Addressing catastrophic forgetting for medical domain expansion." arXiv preprint arXiv:2103.13511 (2021).

- Andrew Beers, James Brown, Ken Chang, **Katharina Hoebel**, Jay Patel, K Ina Ly, Sara Tolaney, Priscilla Brastianos, Bruce Rosen, Elizabeth Gerstner, and Jayashree Kalpathy-Cramer. "DeepNeuro: an open-source deep learning toolbox for neuroimaging." Neuroinformatics 19, no. 1 (2021): 127-140.

- Ken Chang, Andrew Beers, Laura Brink, Jay Patel, Praveer Singh, Nishanth Arun, **Katharina Hoebel**, Nathan Gaw, Meesam Shah, Etta Pisano, Mike Tilkin, Laura Coombs, Keith Dreyer, Bibb Allen, Sheela Agarwal, and Jayashree Kalpathy-Cramer. "Multi-institutional assessment and crowdsourcing evaluation of deep learning for automated classification of breast density." Journal of the American College of Radiology 17, no. 12 (2020): 1653-1662.

- Holger Roth, Ken Chang, Praveer Singh, Nir Neumark, Wenqi Li, Vikash Gupta, Sharut Gupta, Liangqiong Qu, Alvin Ihsani, Bernardo C Bizzo, Yuhong Wen, Varun Buch, Meesam Shah, Felipe Kitamura, Matheus Mendonça, Vitor Lavor, Ahmed Harouni, Colin Compas, Jesse Tetreault, Prerna Dogra, Yan Cheng, Selnur Erdal, Richard White, Behrooz Hashemian, Thomas Schultz, Miao Zhang, Adam McCarthy, B Min Yun, Elshaimaa Sharaf, **Katharina Hoebel**, Jay Patel, Bryan Chen, Sean Ko, Evan Leibovitz, Etta Pisano, Laura Coombs, Daguang Xu, Keith Dreyer, Ittai Dayan, Ram C Naidu, Mona Flores, Daniel Rubin, and Jayashree Kalpathy-Cramer. "Federated learning for breast density classification: A real-world implementation." In Domain adaptation and representation transfer, and distributed and collaborative learning, pp. 181-191. Springer, Cham, 2020.

- Jay Patel, Ken Chang, **Katharina Hoebel**, Mishka Gidwani, Nishanth Arun,

Sharut Gupta, Mehak Aggarwal, Praveer Singh, Bruce R Rosen, Elizabeth R Gerstner, Jayashree Kalpathy-Cramer. "Segmentation, survival prediction, and uncertainty estimation of gliomas from multimodal 3D MRI using selective kernel networks." In International MICCAI Brainlesion Workshop, pp. 228-240. Springer, Cham, 2020.

- Matthew Li, Ken Chang, Ben Bearce, Connie Chang, Ambrose Huang, J Peter Campbell, James Brown, Praveer Singh, **Katharina Hoebel**, Deniz Erdoğmuş, Stratis Ioannidis, William Palmer, Michael Chiang, Jayashree Kalpathy-Cramer. "Siamese neural networks for continuous disease severity evaluation and change detection in medical imaging." NPJ digital medicine 3, no. 1 (2020): 1-9.

- **Katharina Hoebel**, Vincent Andrearczyk, Andrew Beers, Jay Patel, Ken Chang, Adrien Depeursinge, Henning Müller, and Jayashree Kalpathy-Cramer. "An exploration of uncertainty information for segmentation quality assessment." In Medical Imaging 2020: Image Processing, vol. 11313, pp. 381-390. SPIE, 2020.

Chapter 2

Background

The last decade has seen an unprecedented improvement in the capabilities of automatic image analysis algorithms. Figure 2-1 shows the impressive increase in the number of publications using a combination of "artificial intelligence", "deep learning", or "radiomics" with "medical imaging" over the last ten years. Around 2016, the number of AI and DL publications started to grow almost exponentially. This trend is largely driven by the use of Graphics Processing Units (GPU) [51] instead of central processing units (CPU) to speed up the computational processes required for deep learning and concurrent advances in GPU computing power. Compared to CPUs, GPUs are more optimal at performing parallel computations and have a large memory bandwidth which suits the requirements for deep learning.

In this chapter, we review the foundational background information for the remaining chapters of this thesis. We provide an overview of two main methods for supervised machine learning in medical image analysis: radiomics and deep learning using convolutional neural networks. We also review the data annotation process, image preprocessing, and the metrics used to evaluate segmentation algorithms. These concepts will be leveraged throughout this thesis to provide inspiration and solutions for the three fundamental challenges to the translation of ML algorithms into clinical care settings: the robustness of ML in clinical settings (Chapter 3), selecting appropriate network designs (Chapter 4), and the meaningful evaluation of model performance (Chapter 5).

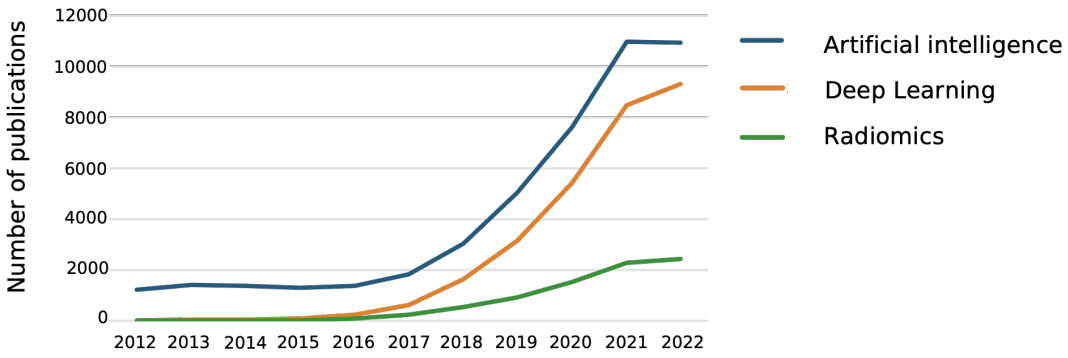

Figure 2-1: **Number of artificial intelligence, deep learning, and radiomics papers published and indexed on PubMed.** Publication trends on PubMed – We searched for publications containing the keywords "artificial intelligence" (blue), "deep learning" (orange), or "radiomics" (green) in combination with the keyword "medical imaging" from January 2012 through October 2022.

2.1 Workflow in a machine learning project

In this section, we outline the typical workflow of a machine learning project for a medical imaging application. The workflow is illustrated in Figure 2-2. It typically starts with a clinical or technical challenge that leads to formulating the research question. Ideally, clinical experts and data scientists or machine learning engineers work together in this step to ensure that the questions asked are clinically valid and technically solvable. Once the research question has been determined, an appropriate dataset needs to be identified. Researchers can choose from publicly available datasets [52, 53, 54] or curate their own datasets from clinical data. Multi-institutional datasets are preferred because algorithms developed using data from a single institution often fail to generalize to cohorts from other institutions [55]. Public datasets have the advantage of allowing performance benchmarking of new methods. They often provide high-quality labels, which significantly reduces the amount of work required for dataset curation. Clinical datasets, on the other hand, often require annotation from scratch as well as careful curation. We outline some challenges and approaches for data annotation in Section 2.4.

The appropriate network design is then chosen based on the research question and the amount and quality of the available data. In Sections 2.2 and 2.3, we outline the

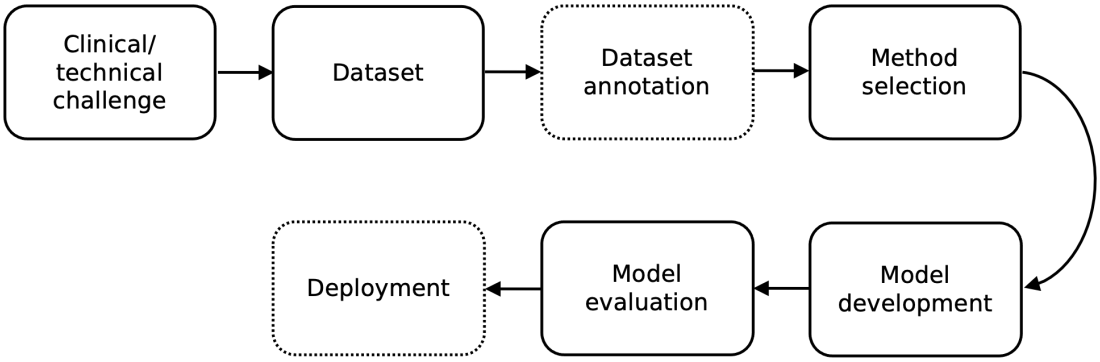

Figure 2-2: **Simplified workflow of a machine learning image analysis project.** First, a research question is established in an interdisciplinary collaboration between machine learning engineers and clinicians. Subsequently, an adequate dataset is identified and annotated if needed. This step is followed by the definition of a network architecture, which is then trained on the data. Lastly, the performance is evaluated and the algorithm can be proposed for clinical deployment if that is the intended end point for the project.

main characteristics of two popular methods for analyzing medical images: radiomics and deep learning using convolutional neural networks. In Chapter 4, we highlight the importance of respecting the statistical distribution of the target variable in selecting the right network design.

After the model development phase, the trained algorithm is evaluated using the appropriate data. The evaluation step ensures that the machine learning model makes accurate predictions on unseen data and is safe to be deployed. In Section 5.3, we analyze how brain tumor segmentation algorithms are evaluated in the literature and illustrate the relevance of incorporating clinical experts in the process because standard quantitative metrics (see Section 2.6) do not capture the clinical usefulness of segmentations. Furthermore, in Section 5.4, we explore what clinical experts value as they evaluate the quality of brain tumor segmentation. Lastly, if that is the intended goal for a project, the algorithm will be deployed and continuously evaluated in the clinic.

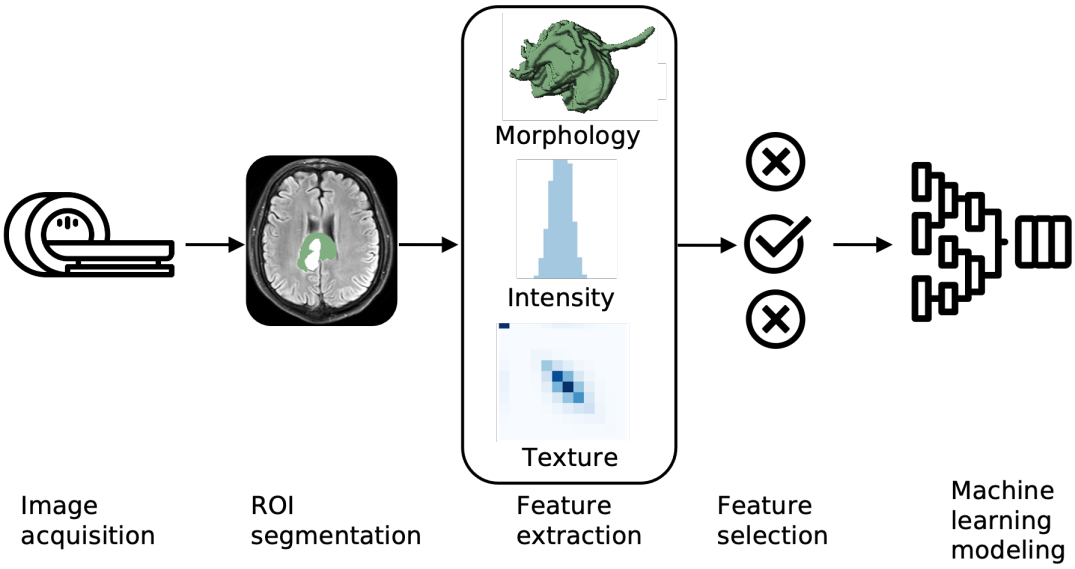

Figure 2-3: **Typical radiomics image analysis workflow.** For the development of a radiomics image analysis pipeline, first, a region of interest (ROI) is outlined on an image. Subsequently, pre-defined radiomic like morphology, intensity, and texture features are computed based on the shape of the ROI and the voxel intensities within the ROI. The most promising features are then selected and used as input for a machine learning algorithm.

2.2 Radiomics

Radiomics is a quantitative medical imaging analysis method that utilizes hand-crafted features extracted from a region of interest and connects them to a patient's disease status [56]. The predictions based on radiomic features are intended to represent an objective, quantitative, and reproducible decision support for physicians [57]. While some have extended the term radiomics to include the use of deep learning for medical image analysis [58], here, we explicitly refer to the extraction of predefined, hand-crafted features. These features are subsequently used to predict a variable of interest, e.g., tumor phenotype [59], response to a treatment [60], or survival [61].

2.2.1 Radiomics workflow

A typical radiomics analysis pipeline, as outlined in Figure 2-3, consists of 4 steps. The imaging modalities most commonly analyzed using radiomics pipelines are CT, MRI, and PET. The first step in the pipeline is the segmentation of the region of

interest (ROI). This step can be performed either manually or (semi-)automatically [62], followed by impage preprocessing. Next, a radiomics software is used to extract multiple predefined features. Different groups of features as defined by the Image Biomarker Standardization Initiative are [63]:

- **Morphology**: These are features like the volume and diameter of the ROI or its volume-to-surface ratio. Morphology features characterize the size and shape of the ROI and are independent of the underlying image.

- **Intensity statistics**: Intensity features describe the characteristics of the intensity all voxels within the ROI, e.g., the mean intensity, kurtosis, or entropy.

- **Intensity-histogram**: These features are computed based on the discretized intensity-histogram of the ROI and includes texture features that characterize the heterogeneity of the voxel intensities. Texture features are derived from matrices that capture the variability and spatial relationship of the intensity values. One such matrix is the gray-level-co-occurrence matrix (GLCM) which summarizes the intensity differences between neighboring voxels. Examples of GLCM features are the correlation or inverse difference moment.

More features can be obtained by first filtering the image using, e.g., Wavelet filters [64] and then extracting features.

The number of features extracted from one image often exceeds 100. Therefore, to avoid overfitting [65], the original number of features is then reduced to limit the number of features that are highly correlated and select the most promising ones for the desired prediction task [66, 67]. Lastly, the selected features are used as input to train a classical machine learning model such as a random forest [68] or support vector machine to predict the desired outcome [69].

2.2.2 Hope and reality

Since its introduction in 2012, radiomics has gained increasing traction in medical image analysis. While it was initially designed for the use in oncology [70], radiomics

image analysis has been applied to various tasks, e.g., the detection of infection with the human papilloma virus [71], cardiac inflammation [72], and the identificiation of the etiology of liver cirrhosis [73]. Although radiomic features were originally intended to capture radiologists' qualitative impressions of images, most common radiomics approaches, albeit quantitatively interpretable, have been criticized for bearing no biological meaning [74].

The hope that radiomics would revolutionize image analysis in clinical practice has been disappointed. Over the years, a growing number of publications have highlighted problems with the reproducibility and generalizability of radiomics image analysis pipelines [38, 75, 76]. Choices at every step of a radiomics analysis pipeline can influence its robustness:

Image acquisition. The scanner type, software, and image preprocessing used to acquire, reconstruct, and prepare an image for analysis have a significant influence on the numerical values of the radiomic features – and therefore on the robustness of a radiomic model [77, 78]. The use of standardized image acquisition, reconstruction, and preprocessing protocols improves the reproducibility of radiomic models [79].

ROI segmentation. Small differences in the definition of an ROI can lead to the extraction of vastly different radomic features [80]. To alleviate this source of variability, researchers recommend using automatically generated segmentations, e.g., outputs of deep learning segmentation algorithms [81]. Alternatively, the features can be extracted from only a core region of the ROI, which limits the influence of small variations in the definition of the segmentation [82].

Feature extraction. Although each radiomic feature is mathematically defined, different software packages may implement the computation differently, resulting in different values for the same features between packages [83].

Feature reduction. By incorrectly performing the feature reduction step using the full dataset, including the validation and test data splits, information from the test data can influence the model development and lead to false-high performance estimates and poor generalization [84].

In summary, numerous factors, including design choices, like the software package

used for feature extraction, and clinical infrastructure, like the available scanner types, influence the robustness of radiomic image analysis models, rendering them difficult to reproduce. The search for solutions to make radiomics algorithms a reliable tool in the clinic is an ongoing research question. In this thesis, we will focus on the role of image preprocessing, particularly normalization and brain extraction, for the test-retest robustness of radiomic features.

2.3 Medical image analysis with deep learning

While in applications of classical machine learning, e.g., radiomics as described in Section 2.2, features are pre-defined, deep learning algorithms learn to identify the features themselves. In an end-to-end fashion, purely based on the training data and without human input, those features are subsequently used, e.g., to classify an image. Several studies have found that DL algorithms outperform radiomics models [85, 86, 87].

2.3.1 Deep Learning image analysis workflow

The method of choice for analyzing images using deep learning are convolutional neural networks (CNN) [88, 89]. These networks learn the weights of kernels to identify patterns that are related to the desired learning task at different resolutions going from high to successively lower resolutions. In the process, CNNs use pooling operations to compress the dimensions of the feature map. Through their ability to identify complex patterns in images, CNNs can be trained to fulfill tasks that previously had to be performed by a small pool of highly specialized experts. If trained appropriately, these algorithms can perform as well as specialists.

DL models based on CNNs have achieved and, in some cases, even exceeded human performance in disease detection and automatic severity classification for numerous diseases such as diabetic retinopathy [90, 91], retinopathy of prematurity (ROP) [45], osteoarthritis [92, 93], and lung diseases [94]. Convolutional neural networks can be used for the classification [90], regression [95], detection [96], segmentation [97], and

registration [98] of medical images. For classification and regression tasks, the features extracted at the lowest resolution are fed into fully connected layer(s) to produce the final prediction as illustrated in Figure 2-4A. To generate a segmentation, so-called encoder-decoder architectures [99] with a downsampling (encoding) and successive up-sampling (decoding) arm are the most commonly used architectures. Particularly the introduction of the U-Net [100] (see Figure 2-4B) and its variations have led to significant improvements in automatic segmentation performance.

Several design choices need to be made development process of a deep learning algorithm. In the following, we describe the most important design choices:

Data preprocessing. Due to a high heterogeneity in image acquisition parameters, data preprocessing is a crucial element of every deep learning pipeline. Preprocessing improves the consistency within a dataset and allows researchers to enhance image features that are considered critical for the subsequent image analysis task. We highlight some essential preprocessing techniques in Section 2.5.

Data augmentation. The size of most medical imaging datasets ranges between around one hundred and a few thousand images and can be considered too small for the requirements of DL. For comparison, the most famous natural image classification datasets ImageNet [101] and CIFAR-100 [102] consist of 14,197,122 images across 20,000 classes and 60,000 images across 100 categories, respectively. Therefore, the risk of overfitting on the training dataset is substantial for the small medical imaging datasets [103].

Data augmentation is used to expand the training dataset's size and quality, reducing the risk of overfitting and improving test performance [104, 105]. Data transformations like vertical and horizontal flips, rotations, and the applications of filters augment the dataset during algorithm development. To lower the memory requirements during training, large 3D volumes, such as MRIs or CT scans, can be broken up into smaller patches, allowing oversampling from specific regions to decrease the class imbalance between fore- and background for segmentation tasks [106]. While more data augmentation can generally result in better performance, the techniques should be chosen to reflect the distribution of the target dataset.

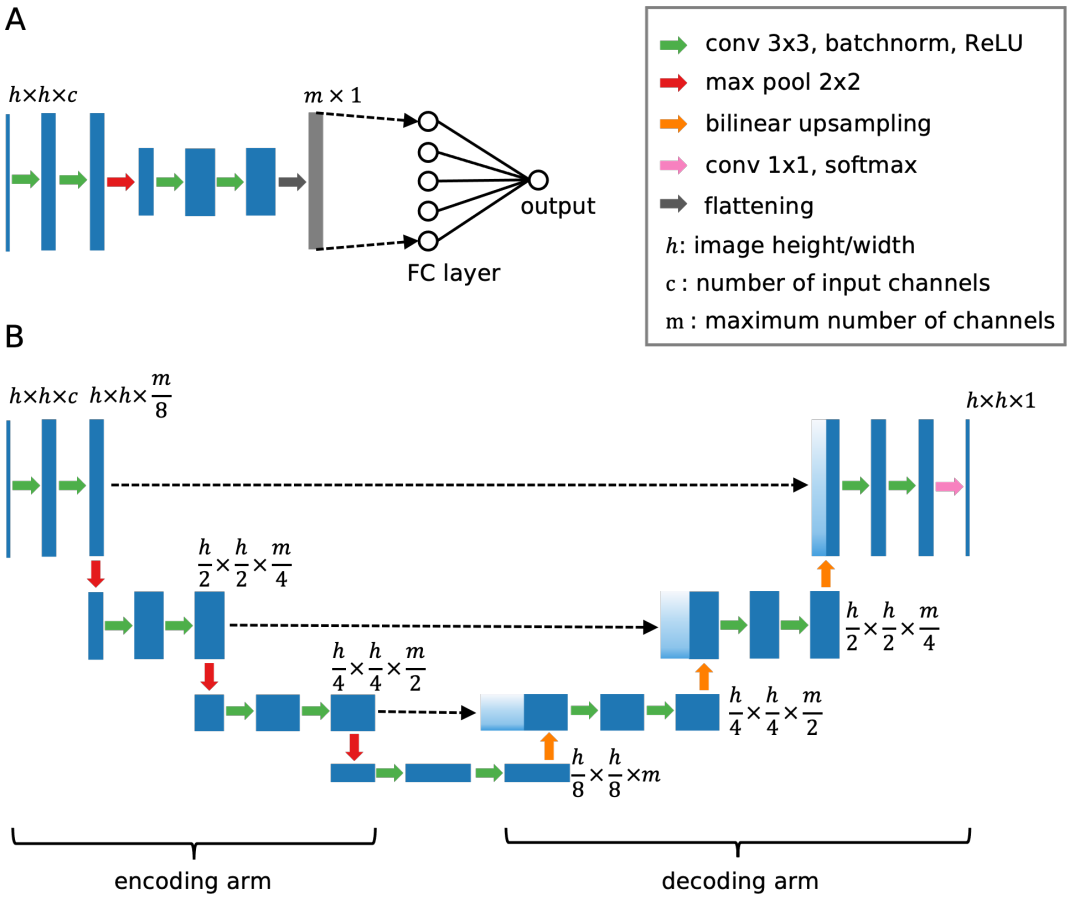

Figure 2-4: **Convolutional neural networks for classification and segmentation.** A: Classification networks consist of several layers with convolutional (green arrows) followed by pooling (downsampling; red arrows) operations. After the final convolution, the features are flattened into a vector (gray arrow) which is fed into one or more fully connected layers to produce the final prediction. B: U-Net-architecture consisting of an encoding and decoding arm with skip connections between both arms. The encoding arm resembles a classification network. In the decoding arm, convolutional operations are preceded by bilinear upsampling (orange arrows) and concatenation with the output of the corresponding level form the the encoding arm.

Network architecture. The neural network architecture that is the most appropriate for a given project depends primarily on the task a deep learning model is expected to solve, i.e., classification versus segmentation. However, other factors like the size of the training dataset and the computational resources available for development and deployment should also be considered [107]. We highlight the importance of respecting the underlying distribution of the target variable when choosing a certain network design in Chapter 4.

Loss function. The loss function measures how well algorithm's predictions and the associated ground truth agree and guides the network to its optimal solution during training. Like the model architecture, the loss function must be chosen to fit the task at hand. Besides choosing the proper loss function for classification or segmentation, some loss functions have been developed for specific training needs, e.g., noisy training labels [108] or a high class imbalance [109].

Evaluation metrics and model selection. Evaluation metrics monitor the model performance on unseen data that is not used for model optimization. They are used to decide when the training process needs to be stopped to avoid overfitting, which is usually when the performance on the validation data starts to plateau while the loss on the training data is still decreasing. The evaluation metric can be the same as the loss metric and should be chosen to reflect the neural network's intended use case, i.e., an algorithm for disease screening should be evaluated primarily based on its sensitivity and only secondarily on its specificity.

Model evaluation on test data. Lastly, the model needs to be evaluated on the test data that reflects the distribution of the target population and has not been used to either train or select the model (validation). This step is required to get an impression of the real-world performance that an algorithm can be expected to show. In Chapter 5, we address the "chasm" between commonly used evaluation metrics and clinical utility for the segmentation of post-operative brain tumors.

While some of the stability problems that radiomics is confronted by do not apply to deep learning, the field has been grappling with its own set of challenges. DL models have been critized for their "black box" nature, using features that lack any inherent

interpretability [110]. Additionally, they are sensitive to shifts in the distribution of the data they are applied to [31]. Lastly, even though algorithms are expected to be objective, they have been shown to reflect and even enhance existing biases in our society and medical system [111]

2.4 Data annotation

To develop machine learning algorithms that meet or exceed human-level performance, large amounts of data are essential; this includes high-quality annotations that function as ground truth. The ground truth annotations, which represent the optimal prediction performance, are used to optimize the neural network during training and later evaluate its performance. The annotation process requires expert annotators with domain expertise. In this section, we outline the challenges of annotating medical images and annotation strategies that are commonly deployed. We focus on the generation of labels for classification on an image level and segmentation on a voxel level. Since segmentation can be seen as a voxelwise classification problem, there are considerable parallels between the two annotation processes.

2.4.1 Challenges in label annotation

Data annotation is a known bottleneck in developing supervised ML algorithms for medical image analysis. Medical datasets are very costly to annotate, as significant expertise is required to interpret and label training data accurately. The annotation process requires a lot of time, particularly for segmentations.

Besides the high cost of acquiring annotations, substantial inter- and even intra-annotator variability make it challenging to establish low noise ground truth standards [112]. Even experts can come to different conclusions when presented with the same clinical information [113], as the interpretation of medical imaging is often inherently ambiguous, and experts interpret the available information differently. For manual segmentations, low-intensity gradients between the target and the surrounding tissue make it challenging to tell different structures apart and cause high variability between

manual outlines. Small-sized lesions, imaging artifacts, and comorbidities complicate the labeling process further [114].

2.4.2 Data annotation strategies

The easiest way to establish a ground truth for a medical imaging dataset is using single reads of each image [115, 116]. These labels can be the clinical diagnoses, parsed from radiology reports either manually or by using natural language processing [117]. However, given the considerable frequency of diagnostic errors [118, 119], using single reads can lead to low-quality labels [120]. Therefore, integrating the impressions from multiple annotators into one label is preferable [121, 122].

Several approaches have been established to combine labels for the same sample that are provided by several annotators. Segmentation labels can be easily aggregated by taking the union, the intersection, or the majority vote of all available labels. By modeling the reliability of single annotators, the Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm generates fused ground truth segmentations that are superior to the simple aggregation approaches listed above [123]. STAPLE has established itself as the standard for generating ground truth annotations for public datasets [124, 125].

Single image-level classification labels are frequently combined using majority vote [126]. However, critical findings are oftentimes easily overlooked and may be recognized by only a minority of annotators. These findings can be under-reported if labels are aggregated using majority vote [127]. A solution to this problem is implementing a label adjudication process in which disagreements between individual annotators are resolved, e.g., through case discussions [117, 128]. Lastly, incorporating additional clinical information further enhances the label quality. Examples are the use of the discharge diagnosis, results from histologic assessments, or additional imaging studies in the process of annotation or adjudication.

Numerous tools have been developed to streamline the collection of annotations. An excellent overview and evaluation of these tools are provided by Aljabari et al. [129]. Furthermore, commercial options to outsource the annotation process are also

available. Other approaches are the use of generative networks to generate realistic medical images and ground truth [130]; weakly supervised training, using image-level labels to obtain voxel-level predictions [131, 132]; or self-supervision [133].

2.5 Preprocessing of MRI brain imaging

This section discusses the steps typically performed during the preprocessing of medical images for ML analysis pipelines. We emphasize the methods used in the preprocessing of brain MRIs as it applies to the research presented in Chapters 3 and 5. The goals of image preprocessing are manifold: it harmonizes the images or full datasets, removes artifacts, results in a simplification of the learning process, and increases the performance of the resulting algorithms. A typical set of preprocessing steps for brain MRI data includes bias field correction, brain extraction, normalization, and image registration. We use the term voxel instead of pixel, as we focus on 3D volumes, but all methods can also be performed on a 2D pixel level.

2.5.1 Bias field correction

The voxel intensities within a tissue can vary across an MRI volume due to corruption by a low-frequency bias field signal caused by inhomogeneities in the magnetic field [134, 135]. While human perception is usually not affected by this distortion, many image analysis algorithms, e.g., for image registration and segmentation, explicitly rely on voxel intensities. Therefore, their performance can be impaired by the presence of the bias field distortions [106].

Over the years, numerous methods have been proposed to remove this undesired signal content. These methods can be broadly categorized as segmentation-based using expectation-maximization [136, 137], filtering [138], and intensity distribution-based methods. Out of the latter, the N4 bias correction is the most notable [139], as it has evolved to become the de-facto standard for bias correction. It is available in several open-source software packages; here, we use the implementation distributed through the Nipype (Neuroimaging in Python: Pipelines and Interfaces) Python package [140].

2.5.2 Brain extraction

Brain extraction, often also referred to as skull stripping, is an image-processing step in which the brain is separated from non-brain tissue, like the scalp, skull, and dura mater. The goal is to eliminate parts of the image that are not required and can cause unwanted variability in the analysis process. An additional benefit of brain extraction is the de-identification of brain scans, as patients can be identified based on renderings of the facial characteristics present in 3D CT and MRI brain images [141, 142]. This step is crucial if a dataset is shared across hospitals or made publicly available.

Researchers can choose from numerous conventional and deep learning methods for brain extraction. Among the conventional methods are methods based on region growing [143], deformable models that are fit to the brain surface [144], or hybrid generative-discriminative models [145]. However, these methods are sensitive to artifacts, anatomical variability, and pathologies, such as brain tumors [146]. More recently, deep learning-based methods have been proposed as a potentially more robust alternative to the conventional models [147, 148, 149]. However, these methods require access to GPU computing. More details on conventional and DL-based brain extraction methods are provided in a review by Rehman et al. [150].

In Chapter 3, we demonstrate how erroneous brain extraction influences a downstream radiomics analysis of brain MRI scans.

2.5.3 Intensity normalization and standardization

Intensity normalization is a routine preprocessing step for ML image analysis pipelines, not just in medical imaging. It facilitates the harmonization of the imaging data by transforming all features onto a similar scale. Normalizing the input data for an ML algorithm can considerably improve its performance, particularly for DL algorithms. Intensity normalization can be performed on a dataset, patient, image, or slice level.

The two image normalization methods we deploy in this work are zero-mean unit-variance or z-score normalization, one of the, if not the most commonly deployed image normalization method, and histogram matching. For z-score normalization, the

following operation is performed on every voxel within an image:

$$z = \frac{x - \mu}{\sigma}.$$

Here, researchers can either use the mean μ and standard deviation σ over all voxels in the full dataset or the voxels of the present image volume. If dataset statistics are used, they must be computed only based on the training dataset to avoid information leaking from the test data into the development of a model.

During histogram matching, the intensity histogram of an image is modified to match the intensity histogram of a designated target image. By matching the histograms of all images within a dataset to the same target image, harmonization of the intensity distributions of all images within a dataset is achieved. In Chapter 3, we use the histogram-matching method proposed by Nyú and Udupa [151].

2.5.4 Image registration

Image registration is the process of aligning a target image with the coordinate system of a source image. In medicine, image registration is often performed to either allow direct comparisons between patients (inter-individual) or align images of the same patient (intra-individual) either between different imaging modalities acquired at the same time-point (inter-modal) or at different time points. The registration of images to a shared space simplifies the manual analysis of images, e.g., to collect information about the localization of lesions, and improves the training of DL models [152]. Within the work considered in this thesis, we perform intra-individual registration of multi-modal MRI scans either taken at the same time or at different time points. Like the other pre-processing methods outlined in this section, conventional [153] and deep learning-based methods [98] have been devised to perform the registration of medical images efficiently. We use the BRAINSfit module implemented in the 3D Slicer image analysis platform [154, 155].

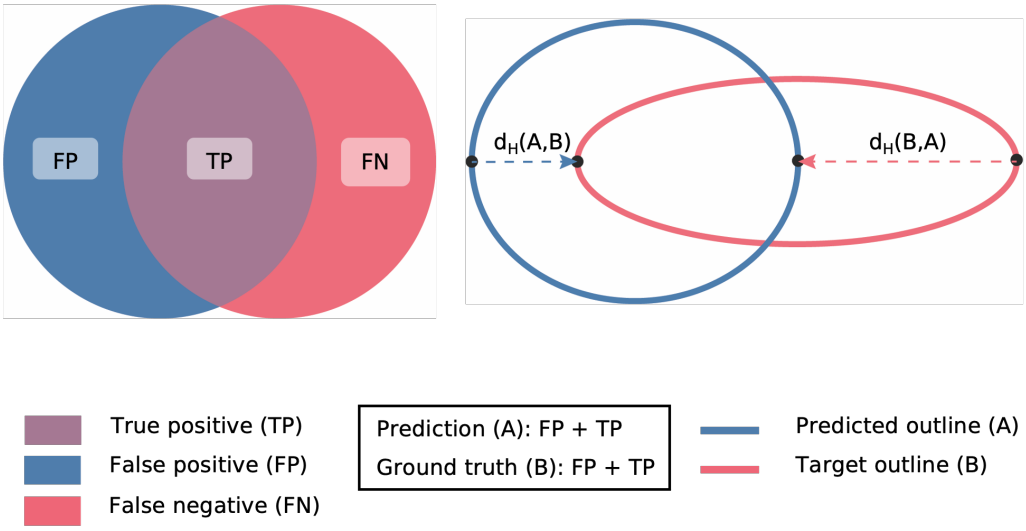

Figure 2-5: **Overlap and distance-based segmentation metrics.** Illustration of the principles behind overlap (left) and distance-based metrics (right) for the quantitative evaluation of segmentation quality. The prediction (A) is shown in blue, the ground truth (B) in red, and the overlap between the prediction and ground truth is in purple.

2.6 Segmentation quality metrics

Choosing the right metrics to assess the performance of a segmentation model is of utmost importance for the success of an ML project [156]. Here, we introduce popular metrics typically used to evaluate the quality of segmentations quantitatively. While this thesis also contains work on classification in Chapter 4, the metrics used in this chapter are standard, and providing details on these metrics is beyond the scope of this section. Reinke et al. offer an excellent and comprehensive overview and discussion of the characteristics and pitfalls of metrics [157].

We organize the metrics into the following four groups: overlap-based metrics, like the Dice score; distance-based metrics, like the Hausdorff distance; volume-based metrics, like the relative volume error; and voxel-level confusion matrix-based metrics, like the sensitivity.

2.6.1 Overlap-based metrics

The most crucial overlap-based and generally most widely used segmentation quality metric is the Dice-Sorensen coefficient or Dice score [158, 159]. As illustrated in Figure 2-5, it is computed using the following equation:

$$D_{score} = \frac{2 * |A \cap B|}{|A| + |B|} = \frac{2 * TP}{2 * TP + FP + FN}$$

The Dice score is equivalent to the F1-score or F-measure and closely related to the Jaccard index or Intersection over Union (IoU). In fact, the Dice score can be interpreted as an approximation of the IoU and vice versa [160]. Furthermore, it can be interpreted as a special case of the kappa index, a chance-corrected measurement of agreement [161]. We will only discuss the Dice score here, but all observations and comments in the following paragraphs also apply to the IoU.

Advantages of the Dice score The Dice score is an excellent metric for segmentation quality as it provides an intuitive measure of the agreement between the ground truth and a prediction in their size and location. Furthermore, a differentiable version of the Dice score, the so-called soft Dice [162], is among the most popular loss functions in medical imaging for the training deep learning segmentation algorithms. Both the Dice score and the soft Dice are stable towards class imbalances, which is a considerable advantage during training and for the evaluation of segmentations, particularly for small targets against large background volumes, as is often the case in medical image segmentation.

Issues with the Dice score However, the value of the Dice score is known to be influenced by the size of the segmentation target. For large segmentation volumes, a few false positive or false negative voxels do not substantially influence the value of the Dice score. On the other hand, the Dice score is quite sensitive to a few missed pixels for small segmentation volumes. It can also be said that the Dice score is overly confident for larger segmentation volumes. As for numerous medical segmentation

targets, methods to determine the exact outlines on macroscopic imaging have yet to be discovered; an ideal segmentation quality metric should be insensitive to minor differences, for large and small segmentation targets alike. Another consequence is the unequal treatment of over and under-segmentation, particularly for smaller segmentation targets. A single-voxel thick rim of false positives pixels (over-segmentation) results in a higher Dice score than a layer of missed pixels (under-segmentation) of the same thickness. Even though it could be argued that the error is the same [163, 164]. Lastly, the Dice score has been criticized for not accurately reflecting the clinical utility of a segmentation [165], mainly due to its inability to differentiate between systematic and random errors [166].

2.6.2 Distance-based metrics

Distance-based metrics focus on the distance between the outlines of a prediction and the ground truth. The most common distance-based metric is the Hausdorff distance [167] (see Figure 2-5B). It represents the maximum of all the shortest distances between two outlines and is computed based on the following equation:

$$H(A, B) = \inf\left(d_H(A, B), d_H(B, A)\right),$$

where

$$d_H(A, B) = \sup_{a \in A}\left(\inf_{b \in B}\left(\|a - b\|\right)\right).$$

The Hausdorff distance is highly sensitive to small outliers, which deep learning segmentation models are prone to produce. Therefore, the 95th percentile of the Hausdorff distance is often used as a more robust metric unaffected by small outliers. Another popular distance-based metric is the Average Symmetric Surface Distance (ASSD) [168], representing the average of all minimum pointwise distances between two surfaces.

The so-called surface Dice [169] is an outline-focused alternative to the conventional Dice score. It also quantifies an overlap, not of the total segmentation volume but only

of the surfaces of the ground truth and prediction. However, the user needs to choose the value of a tolerance term, which defines the acceptable deviation between the predicted and ground truth boundary. Since it is unclear which values are clinically acceptable for this deviation term, the results can be challenging to interpret.

Because of the importance of correctly identifying the border between tumorous and healthy tissue in radiation therapy planning, these metrics are particularly suited to evaluate the quality of segmentations for applications in radiation oncology. Distance-based metrics are also insensitive to holes within a predicted segmentation (false negative areas surrounded by true positive voxels) [163]. As overlap and distance metrics are often not highly correlated, metrics from both groups can, if chosen wisely, be complementary in evaluating segmentation performance [170].

2.6.3 Voxel-level confusion matrix-based metrics

Voxel-level confusion matrix-based metrics are primarily used for the evaluation of classification performance but are also popular for segmentation, as it can be interpreted as a voxel-level classification. Accuracy is among this group's most commonly reported metrics, primarily due to its simplicity in the interpretation. However, accuracy is sensitive to class imbalances; therefore, its value does not accurately reflect the segmentation performance for segmentation problems with a high-class imbalance between fore- and background [171]. Along similar lines, the specificity for segmentation models with small foreground-to-background ratios will always be relatively high and can be misleading if not complemented by other metrics. Lastly, depending on the application, missing parts of the segmentation target may represent a graver error than over-segmentation. Therefore, more than one segmentation metric is needed to capture the importance of different mistakes accurately. Furthermore, an estimate of the class imbalance should be reported to allow readers to interpret the presented results accurately [172].

2.6.4 Volume-based metrics

Metrics in this group measure the similarity between the target volume and the predicted volume. This group's most used metrics are the relative and absolute volume error and volume similarity. While for several applications, e.g., treatment response monitoring, only the accuracy of the segmented volume matters, these metrics do not consider the agreement in the location between the ground truth and the prediction. A segmentation approximately the same size as the target but in the wrong place would be considered a high-quality segmentation based on volume-based metrics. However, this segmentation would perform poorly when assessed by other metrics, such as overlap-based ones, illustrating the importance of utlizing several metrics from different metric groups.

Chapter 3

Pitfalls in the repeatability of radiomics

One essential requirement for the successful deployment of ML algorithms in the clinic is that they are stable in their predictions, meaning given the same patient at the same time point, they will produce the same prediction. In this chapter, we assess the influence of preprocessing on the repeatability and redundancy of radiomics features extracted using a popular open-source radiomics software package in a scan-rescan glioblastoma MRI study. We found that shape features, which are independent of voxel intensities, show higher repeatability than voxel intensity dependent features, like intensity or texture features. Furthermore, the repeatability of radomics features could be increased through the careful selection of preprocessing steps, such as intensity normal normalization and feature extraction settings, like intensity quantization.

This chapter has been adapted from the following manuscript: Katharina Hoebel, Jay Patel, Andrew Beers, Ken Chang, Praveer Singh, James Brown, Marco Pinho, Tracy T Batchelor, Elizabeth Gerstner, Bruce Rosen, Jayashree Kalpathy-Cramer. "Radiomics repeatability pitfalls in a scan-rescan MRI study of glioblastoma." Radiology: Artificial Intelligence 3, no. 1 (2020): e190199.

with additional excepts from the following two works:

- Mishka Gidwani, Ken Chang, Jay Biren Patel, Katharina Hoebel, Syed Rakin

Ahmed, Praveer Singh, Clifton David Fuller, Jayashree Kalpathy-Cramer. "Inconsistent Partitioning and Unproductive Feature Associations Yield Idealized Radiomic Models", Radiology (2022): 220715.

- Andréanne Lemay, Katharina Hoebel, Christopher Bridge, Brian Befano, Silvia De Sanjosé, Didem Egemen, Ana Cecilia Rodriguez, Mark Schiffman, John Peter Campbell, and Jayashree Kalpathy-Cramer. "Improving the repeatability of deep learning models with Monte Carlo dropout." npj Digital Medicine 5, no. 1 (2022): 1-11.

- Katharina Hoebel, Christopher Bridge, Andréanne Lemay, Ken Chang, Jay Patel, Bruce Rosen, and Jayashree Kalpathy-Cramer. "Do I know this? segmentation uncertainty under domain shift." In Medical Imaging 2022: Image Processing, vol. 12032, pp. 261-276. SPIE, 2022.

- Katharina Hoebel, Vincent Andrearczyk, Andrew Beers, Jay Patel, Ken Chang, Adrien Depeursinge, Henning Müller, and Jayashree Kalpathy-Cramer. "An exploration of uncertainty information for segmentation quality assessment." In Medical Imaging 2020: Image Processing, vol. 11313, pp. 381-390. SPIE, 2020.

3.1 Introduction

In recent years, radiology has experienced a shift toward more quantitative analysis of imaging to aid medical decision-making. Among the most prominent techniques leading this shift is radiomics, which is defined by the extraction of quantitative features from a region of interest (ROI) on medical images [56]. As described in Section 2.2, these quantitative descriptors can then be used to build predictive models for clinical variables, such as molecular markers, treatment response, and prognosis [173]. This approach has the advantage that, in comparison to biopsies, features can reflect the full diversity of the ROI and factors such as tumor heterogeneity can be captured more easily and in a noninvasive manner [174, 175]. However, if these models are used for patient stratification with potential treatment decisions based on

their predicted outcomes, the features used must fulfill two criteria: repeatability and reproducibility.

Repeatability refers to the *"variability of the quantitative image biomarker when repeated measurements are acquired on the same experimental unit under identical or nearly identical conditions"* to determine the measurement error [36]. Reproducibility refers to *"variability in the quantitative image biomarker measurements associated with using the imaging instrument in real-world clinical settings,"* such as different settings of a software package attempting to identify and separate measurement errors from the reproducibility conditions [36]. Both repeatability and reproducibility of radiomic features have been described to be sensitive to various factors, such as image acquisition, resolution, reconstruction, preprocessing, and the software package used to extract them [176].

Most published studies have described repeatability on CT [173, 176, 177, 178]. In contrast to CT, absolute voxel intensities on MRI do not have tissue-specific values, and changing signal intensities can leave tissue contrast unaltered [179]. Therefore, intensity normalization might be needed to correct for these changes in intensity to make features comparable between and within patients, especially when scanned under slightly different conditions [180].

Relatively few studies have examined the repeatability and reproducibility of radiomic features extracted from contrast-enhanced MRI [181, 182, 183]. One potential reason is the challenge with test-retest studies that require the use of contrast agents. Of the few studies that have examined this topic, very little has been reported on the underlying causes for the lack of robustness of features. In a recent study on the repeatability of radiomic features for small prostate tumors, Schwier et al. showed that different features extracted with different MRI sequences might require different settings to increase their repeatability [181]. However, this study only evaluated the effect of preprocessing and feature extraction configurations on the intraclass correlation coefficient (ICC) for features extracted from a small ROI and for a relatively small dataset of 15 patients.

In this chapter, we examine the repeatability and feature redundancy of radiomic

features in a unique scan-rescan dataset of patients with newly diagnosed glioblastoma, to understand some of the reasons for the observed lack of repeatability.

3.2 Methods

3.2.1 Study population

We present a secondary analysis of prospectively collected data from two clinical trials (ClinicalTrials.gov ID NCT00662506 and NCT00756106) at Massachusetts General Hospital and Dana-Farber Cancer Institute [184]. All patients underwent the same imaging protocol, and both studies were approved by the institutional review board. The full dataset (scan, re-scan, and all follow-up visits) contained a total of 713 post-operative MRI visits. A total of 54 adult patients (mean age, 57 years [age range, 22–77 years]; 33 men, 21 women) were included in the initial evaluation. All patients had undergone tumor biopsy or partial tumor resection with a remaining contrast-enhancing tumor of at least 1 cm in diameter at the time of enrollment. Patients received either chemoradiation with cediranib (NCT00662506) or standard chemoradiation (NCT00756106).

Patients underwent two pretreatment scans 2–6 days apart (mean, 3.7 days apart). Figure 3-1 illustrates the imaging timeline with respect to the start of treatment. Patients for whom one or both of the baseline scans were unavailable were excluded from this study, resulting in a cohort size of 48 (mean age, 56 years [age range, 22–77 years]; 27 men, 21 women). The patients received no treatment between scan and rescan. None of the tumors had clinically significant changes between scans, as measured by the change in contrast-enhancing tumor volume or fluid-attenuated inversion recovery (FLAIR) hyperintensity.

3.2.2 MRI acquisition

All MRI images from both clinical studies were acquired at the same research institution, using an identical imaging protocol, and were obtained with the same model of 3.0-T

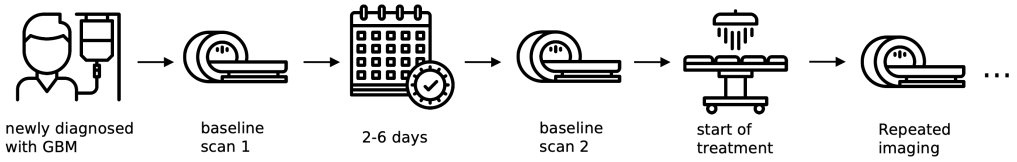

Figure 3-1: **Clinical study imaging protocol.** The study protocol included two baseline scans 2-6 days apart before the start of treatment and repeated follow-up imaging at defined intervals until the patients left the study cohort.

MRI system (TimTrio; Siemens Medical Solutions, Malvern, Pa) and 32-channel head coils. A total of 40 of 48 patients underwent both scan and rescan using identical MRI scanners. To improve scan-to-scan reproducibility, AutoAlign (Siemens) was used to ensure automatic alignment of the slice positions in a standard reproducible way for each scan.

Axial T2-weighted FLAIR images were acquired with a repetition time of 10 000 msec, an echo time of 70 msec; T1-weighted images were acquired with a repetition time of 600 msec, an echo time of 12 msec before and after the injection of a bolus of 0.1 mmol per kilogram of body weight of Magnevist (Bayer Healthcare, Warrendale, Pa). All imaging volumes have 5-mm slice thickness, 1-mm intersection gap, 23 sections, 0.43-mm in-plane resolution, matrix size of 256 x 216, and a field of view of 185mm x 220mm. The pre-treatment scans used for this analysis are publicly available from the TCIA [185].

3.2.3 Segmentation, annotation, and preprocessing

Manual segmentation Segmentations of enhancing lesions based on T1-weighted pre- and postcontrast sequences and areas of T2 abnormality on T2-weighted FLAIR sequences were performed by an expert neuro-oncologist with 12 years of experience and a neuroradiologist with 11 years of experience, respectively. The segmentations for each patient were performed by a single expert. The annotators were blinded to patient identity, order of scans, and patient treatment status. To exclude postoperative blood products that appear hyperintense on T2-weighted FLAIR, the experts additionally evaluated the imaging presentation on T1-weighted pre- and post-contrast MRIs. All

manual segmentations were performed using the 3D Slicer software [155].

Image preprocessing After segmentation, each patient's T1-weighted postcontrast sequences were registered to corresponding T2-weighted FLAIR sequences using the BRAINSfit module in 3D Slicer [154, 155]. The N4 bias-correction algorithm was applied to all images using the Python package Nipype [140]. Whole-brain extraction was performed on T1-weighted postcontrast images using the ROBEX (RObust Brain EXtraction) algorithm [145], and the resulting brain mask was applied to T2-weighted FLAIR images.

Normalization of input images was performed as part of the feature extraction (built-in z-score normalization) or by using a histogram-matching technique as a separate step before feature extraction. The built-in normalization normalizes each input volume such that the mean of the voxel intensity distribution is centered at zero with unit variance (z-score normalization). Histogram matching of the non-ROI region is a common normalization technique in radiomics [186]. In our study, we implemented histogram matching using the method described by Nyúl and Udupa [151], in which a piecewise linear transformation is applied such that the histogram of a source image is matched to that of a chosen reference image. A randomly chosen patient was used as a reference to which the histograms of all other patients were matched. More detailed information on the performed preprocessing steps can be found in Section 2.5.

In addition to the aforementioned manual masks, we derived union masks of both visits by registering the rescan to the scan and taking the union of both masks separately for the enhancing tumor ROI on T1-weighted postcontrast images and total tumor ROI on T2-weighted FLAIR images. These masks were then registered back to the nonregistered images for feature extraction.

3.2.4 Radiomics software

Radiomics features were extracted using the open-source Python package PyRadiomics [177]. Features for the scan and rescan were extracted separately from both T1-weighted postcontrast and T2-weighted FLAIR images. Whenever indicated, the

package default image normalization was applied to brain-extracted images as part of the feature extraction process (z-score normalization). All features defined as default by PyRadiomics were extracted from three-dimensional tumor volumes.

We limited our analysis of texture features to features derived from gray-level co-occurrence matrices (GLCMs) and excluded the following features from further analysis: compactness1, compactness2, and spherical disproportion are perfectly correlated with sphericity; and homogeneity1 and homogeneity2 are directly correlated with inverse difference moment. For each experimental setting and sequence (T1-weighted postcontrast and T2-weighted FLAIR) we extracted 13 shape, 17 intensity, and 23 texture features.

3.2.5 Statistical analysis

For each feature extracted from both T1-weighted postcontrast and T2-weighted FLAIR sequences, we calculated the intra-class correlation coefficient (ICC) [187] between the feature value extracted from the scan and rescan over the sample of 48 patients. We used a two-way model of the ICC (unit, single; type, consistency; 95% CI) as implemented in the R statistical software (version 3.5.2) "IRR" package (version 0.84). Features were then grouped into shape (or morphology describing size and shape), intensity, and texture features, as proposed by Kalpathy-Cramer et al. [83] and following the image biomarker standardization initiative classes for feature groups for further analysis [188].

To determine the association between features, we calculated the pairwise Spearman correlation coefficient between features for all patients (one scan) and took the absolute value to reflect the strength of the correlation. For comparison of the ROI intensity distributions between scan and rescan, we chose the maximum range of voxel values of both images and divided it into 100 bins. These bins were then used to derive the intensity histograms for both visits. We used these histograms to calculate the Jensen-Shannon divergence (JSD) between visits. On the basis of the Kullback-Leibler divergence, the JSD has the advantage of being both symmetric and an unbiased measure of the similarity between two probability distributions [189].

Statistical significance between feature groups was assessed using a Kruskal-Wallis test followed by post hoc pairwise Dunn multiple-comparisons tests with Bonferroni correction to determine the relationship between the individual means [190, 191]. Analysis of statistical differences between normalization approaches was performed with the paired Wilcoxon test with respect to the chosen baseline (no normalization) and Bonferroni correction for multiple comparisons. The significance threshold for adjusted p-values was 0.05. Statistical analysis was performed using R statistical software (version 3.5.2).

3.3 Results

3.3.1 Repeatability of feature extraction from unnormalized MRI

First, we examined the repeatability of shape, intensity, and texture features using the PyRadiomics default settings (no normalization, intensity quantization with constant bin width set to 10). Figure 3-2 shows the distribution of the ICC scores for each feature group for both sequences. The ICC is computed based on the full study population. For both sequences, purely segmentation-dependent features in the shape group are highly repeatable between the scan and rescan, with a median ICC of 0.98 (range, 0.88–0.99) for T2-weighted FLAIR images and 0.96 (range, 0.78–0.98) for T1-weighted postcontrast images. Features in the intensity and texture feature groups, which depend on voxel intensity values, show low ICCs and high variability in the ICCs within the groups for both sequences, with median ICC values for T2-weighted FLAIR and T1-weighted postcontrast images of 0.60 (range, 0.38–0.84) and 0.71 (range, 0.36–0.83), respectively, for intensity, and 0.68 (range, 0.10–0.94) and 0.78 (range, 0.48–0.86), respectively, for texture features.

We observed differences in the ICC distribution for T2-weighted FLAIR and T1-weighted postcontrast images, respectively, between shape and intensity (p-values < 0.001 and < 0.001, adjusted for three comparisons) and shape and texture features

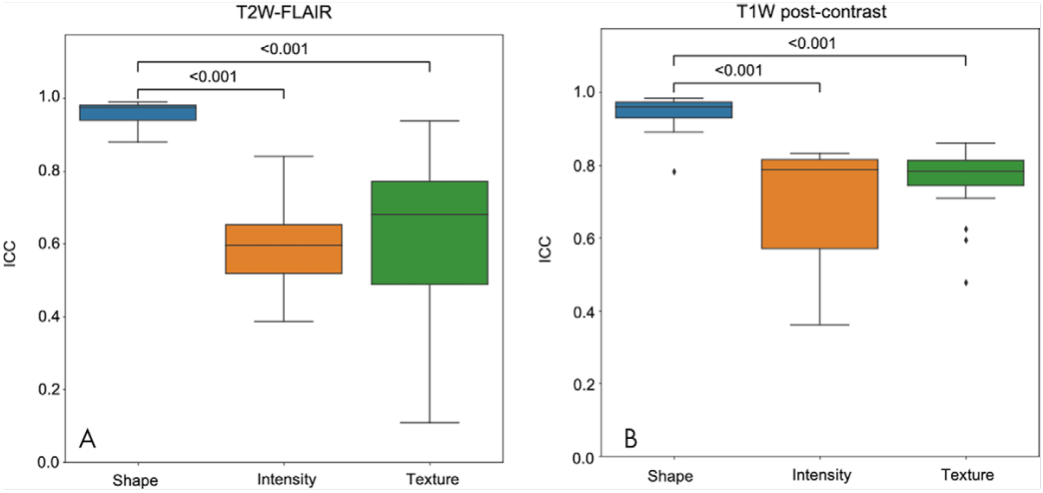

Figure 3-2: **Distribution of intraclass correlation coefficient (ICC) values per feature group under default feature extraction settings**. Each boxplot represents the distribution of one radiomics feature group (shape, intensity, texture) between scan and rescan for the cohort of 48 patients. A: T2-weighted fluid-attenuated inversion recovery (T2W-FLAIR); B: T1-weighted (T1W) postcontrast. Features were extracted from non-normalized images using the PyRadiomics default settings (no normalization, constant bin width for intensity quantization)

(p-values < 0.001 and < 0.001, adjusted for three comparisons) on the pairwise Dunn test. Accordingly, we assessed how the repeatability of features in the intensity and texture groups that are calculated based on voxel intensities can be improved.

3.3.2 Effect of normalization on the intensity distribution and intensity quantization on within-scan feature correlation

Influence of normalization on the ROI intensity distribution Intensity features describe the distribution of voxel intensity values in the segmented region. GLCM features are computed based on the GLCM, which represents the relationships of the voxel intensities of neighboring voxels in the ROI. Before we studied the repeatability of intensity and GLCM features, we first assessed the effect of normalization on the ROI intensity histogram of both the scan and the rescan and voxel intensity quantization on the correlation between GLCM features.

The voxel values for MRI are not normalized, and there are no tissue-specific intensity ranges, so features based on voxel intensities showed great variability in ICC

between scan and rescan. Therefore, we first studied the effect of normalization on the intensity distribution of the segmented tumor region (ROI intensity histogram) by comparing the voxel intensity histograms between scan and rescan before turning to the repeatability of intensity features. We used (a) the built-in normalization (z-score normalization over all voxels in the input volume) and (b) histogram matching to a reference case. The overlap between histograms was measured by the JSD between the ROI intensity histogram of the scan and rescan and the effect of normalization as change in JSD before and after normalization for all 48 patients.

For our study population, both normalization techniques (z-score and histogram normalization of brain-extracted images) significantly improved the similarity between the histograms, as measured by JSD on T2-weighted FLAIR and T1-weighted post-contrast images (paired Wilcoxon test against the not-normalized baseline without comparisons between the normalized groups, adjusted p-values for two comparisons, z-score and histogram matching, respectively, of < 0.001 and < 0.001 on T2-weighted FLAIR images and 0.002 and 0.03 on T1-weighted postcontrast images). Panels A and C in Figure 3-3 show the ROI intensity histograms for both scans before normalization (column 1) and the change in the overlap between the histograms owing to normalization (columns 2 and 3). Figure 3-3A illustrates the effect of normalization techniques for a representative case.

In some cases, normalization caused an increase in JSD between scan and rescan instead of the expected decrease. Failure cases were defined as cases for which normalization resulted in an increase in JSD for both T1-weighted postcontrast and T2- weighted FLAIR sequences (this analysis was constrained to z-score normalization). Six of 48 patients' scans were identified as failure cases. Visual assessment of these cases revealed that for all of them, the brain extraction step was not performed properly. We identified two patterns: (a) either too aggressive or (b) total or partial failed brain extraction (leaving either the full skull or parts of the skull behind). The latter mode of failure is illustrated in Figure 3-3C, with representative axial slices of T2-weighted FLAIR and T1-weighted postcontrast scan and rescan in panel D, illustrating the corresponding brain extraction failure pattern.

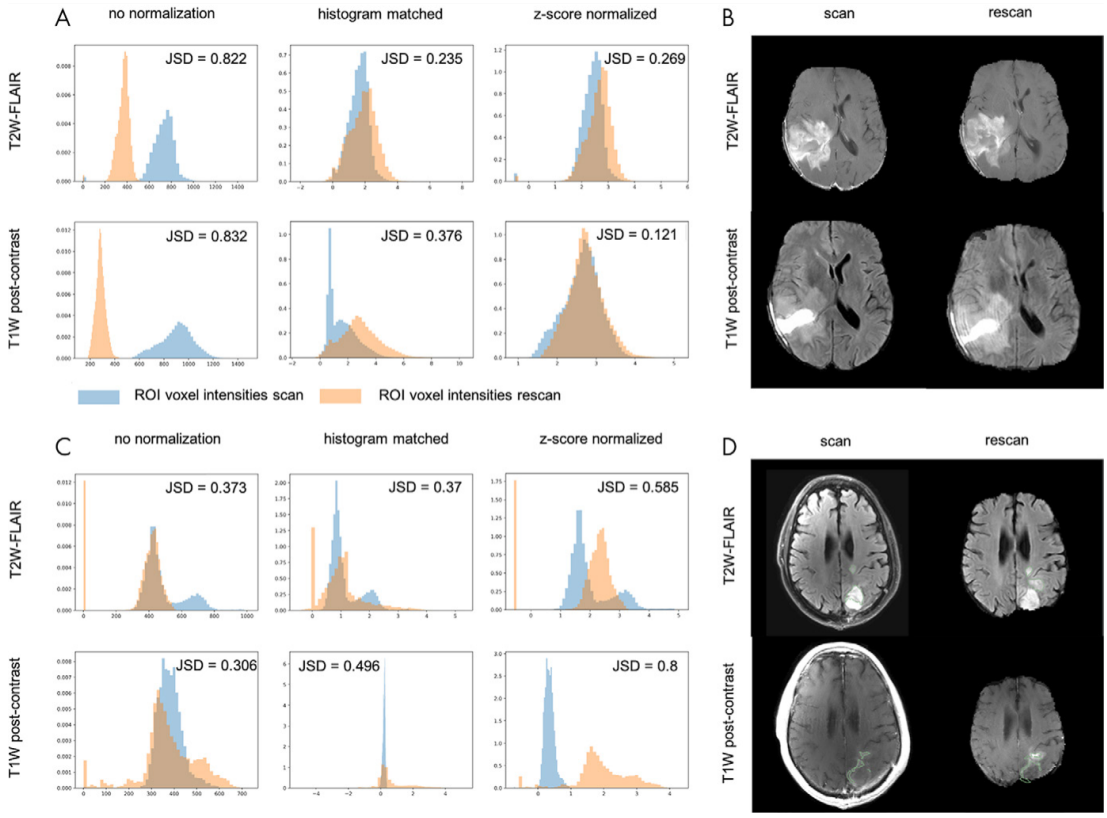

Figure 3-3: **Effect of normalization on the region of interest (ROI) intensity histograms.** A, C: Intensity histograms of the ROI segmentations from the scan (blue) and rescan (orange) of representative cases on both T2-weighted fluid-attenuated inversion recovery (T2W-FLAIR) and T1-weighted (T1W) postcontrast sequences of a representative case (A) and a failure case (C). The first column shows ROI intensity histograms without preprocessing; the second column, after brain extraction and normalization via histogram matching; and the third column, after brain extraction and z-score normalization. The overlap between the histograms is quantified by the Jensen-Shannon divergence (JSD). B, D: Axial sections from the T2-weighted FLAIR and T1-weighted postcontrast scan and rescan after brain extraction of the corresponding cases, A, C, respectively.

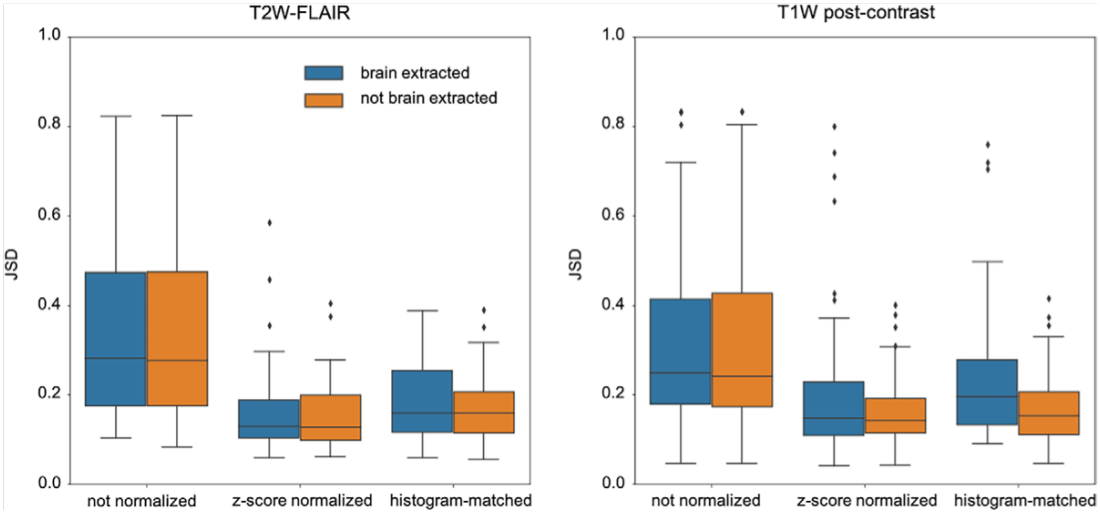

Figure 3-4: **Jensen-Shannon divergence (JSD) distributions with and without brain extraction.** Distribution of the JSD between the region of interest intensity histograms of the scan and rescan for the entire cohort using T2-weighted fluid-attenuated inversion recovery (T2W-FLAIR) (left) and T1- weighted (T1W) postcontrast (right) for not-normalized, z-score normalized, and histogram-matched images, each with (blue) and without (orange) brain extraction performed before normalization. For each normalization approach (no normalization, z-score normalization, histogram-matched), the absence of brain extraction before normalization did not have a significant effect on the JSD.

We therefore additionally examined the JSD distributions of images that were normalized without previous brain extraction. As shown in Figure 3-4, the JSD values of the scan and rescan ROI intensity histograms of images normalized without previous brain extraction were not significantly different from brainextracted and normalized images.

Influence of voxel intensity quantization on feature correlation In a manner similar to how intensity features describe the distribution of intensity values in the ROI, GLCM features describe the GLCM. For the computation of the GLCM, intensity values first need to be quantized into discrete intensity ranges. This quantization step can be performed using either a defined bin width (absolute binning) or a preset number of bins (relative binning) adapted to the range of intensity values in the ROI.

Assuming the user has normalized the intensities, using the default intensity quantization settings as implemented in PyRadiomics (constant bin width set to

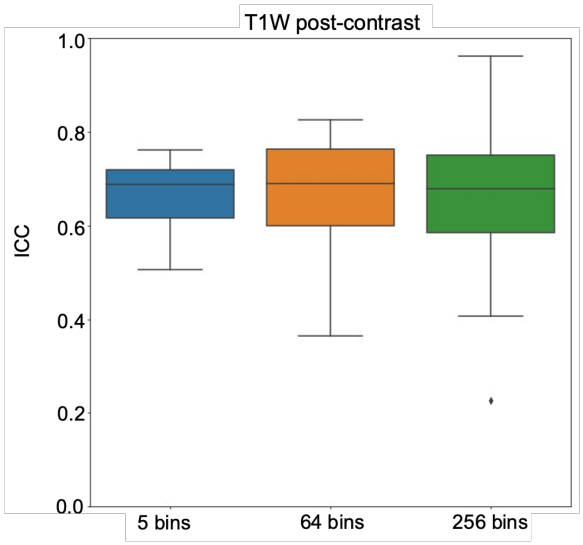

Figure 3-5: **Distribution of intraclass correlation coefficient (ICC) values for GLCM texture features.** Each boxplot represents the distribution of GLCM features repeatability between scan and rescan for the cohort of 48 patients using a different with 5 (blue), 64 (orange), and 254 (green) bins for intensity quantization. Features were extracted from T1-weighted postcontrast images.

10) results in nonsensical binning for the computation of the GLCM (i.e., all voxel intensities are placed into only two bins, as this choice of bin width is too coarse for the existing range of voxel values following normalization). Texture features calculated based on this GLCM do not capture the true variability in image intensity that is present within the images. This setting results in extremely high correlations between texture features (mean Spearman correlation coefficient for all features, 0.95) on T1-weighted postcontrast images.

By explicitly specifying the number of bins (relative binning) rather than a fixed bin width, the aforementioned effect can be avoided. With increasing numbers of bins and quantization levels, the overall correlation between texture features decreases (mean Spearman correlation coefficient for all GLCM features, 0.52 [five bins], 0.47 [64 bins], and 0.43 [256 bins]; T1-weighted postcontrast imaging), without an adverse effect on the repeatability of these features (Figure 3-5). The same effect can be observed on T2-weighted FLAIR images. On the basis of these results, for data reported in the following sections, we did not use the constant bin width setting; rather, we explicitly set the number of intensity value bins for intensity quantization to 256.

3.3.3 Influence of normalization on the repeatability of intensity and texture features

As described previously, the application of z-score normalization and histogram matching improved the overlap between the ROI intensity histograms of the scan and rescan. Furthermore, the previous results highlight the importance of an appropriate binning strategy. Building on these results, we examined the influence of the normalization on the repeatability of intensity and texture features between scan and rescan, using relative binning with 256 bins for intensity quantization, for features extracted from not normalized, z-score normalized, and histogram-matched scans. For comparison, we also included the ICC data computed on features extracted using z-score normalization in combination with the default absolute intensity quantization setting (constant bin width, 10).

The effect of the choice of the normalization technique on the ICC between both scans for intensity and texture features is presented in Figure 3-6(top row, intensity; bottom row, texture features). The ICCs of single features are shown in Figure 3-7 for intensity and Figure 3-8 for texture features, illustrating the variability in the effect of normalization on the repeatability of single features.

Intensity features While both normalization techniques lead to an improved overlap between the ROI intensity histograms of scan and rescan for both T2-weighted FLAIR and T1-weighted postcontrast sequences, the effect of normalization on the repeatability of intensity features varies between the sequences (Figure 3-6A, B). On T2-weighted FLAIR images (Figure 3-6A), both z-score normalization and histogram matching improved the repeatability of intensity features with respect to the not-normalized baseline (relative binning with 256 bins; paired Wilcoxon test against the not-normalized baseline without comparisons between the normalized groups; adjusted p-values for three comparisons, 0.003 [z-score normalization] and 0.002 [histogram matching]). On T1-weighted postcontrast images, however, neither z-score normalization nor histogram matching resulted in a significant effect on the ICC of intensity features between scan and rescan (Figure 3-6B).

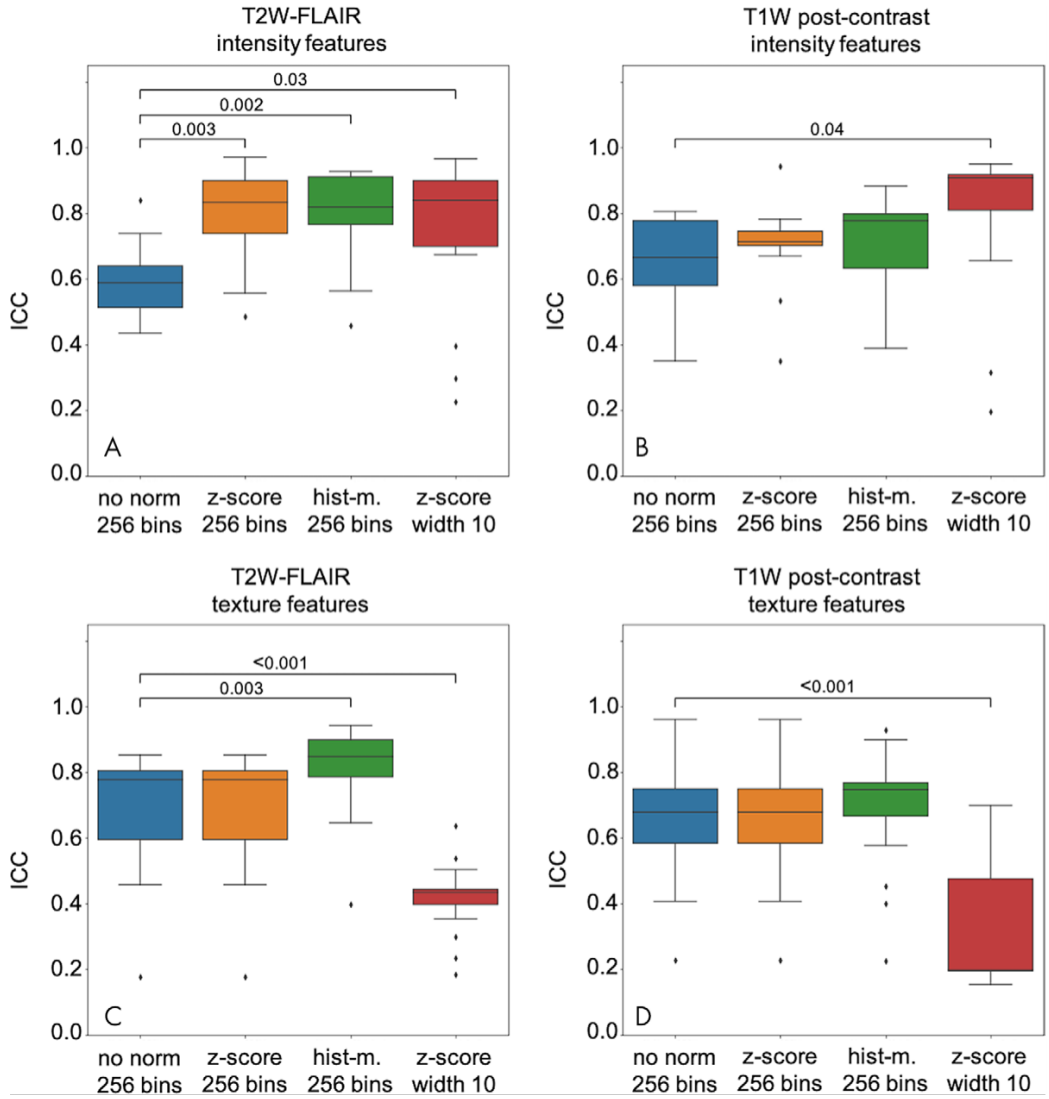

Figure 3-6: **Distribution of intensity and texture intraclass correlation coefficient (ICC) values under different conditions.** ICC for intensity (A, B) and texture features (C, D) extracted from T2-weighted fluid-attenuated inversion recovery (T2W FLAIR) (left) and T1-weighted (T1W) postcontrast (right) using either z-score normalization (z-score) or histogram matching (hist-m.) compared with features extracted from not-normalized (no norm) images. Significant differences in the feature group mean ICC between feature extraction strategies (paired Wilcoxon test) are indicated with brackets.

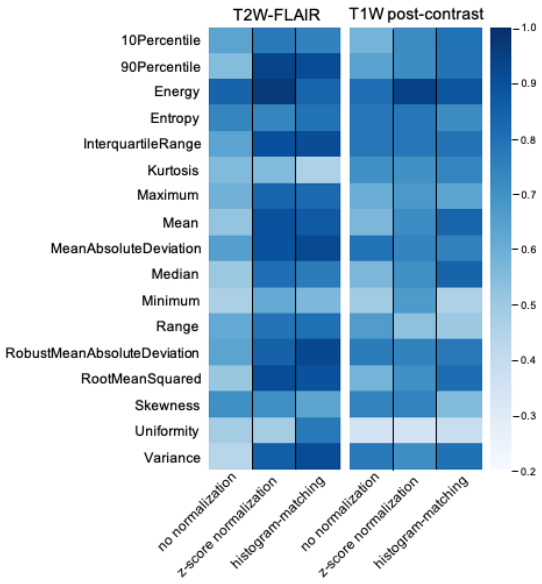

Figure 3-7: **Intensity feature stability for different normalization strategies.** ICC of single intensity features (rows) extracted under different normalization conditions using relative intensity quantization with 256 quantization levels (columns) from T2W-FLAIR (left) and T1W post-contrast (right). Particularly for features extracted from T2W-FLAIR normalization improved the ICC.

Texture features As in the case of intensity features, normalization techniques have a different effect on both sequences. For T2-weighted FLAIR images, z-score normalization did not change the ICC distribution of texture features compared with no normalization (relative intensity quantization, 256 bins), whereas histogram matching improved the repeatability (paired Wilcoxon test against the not-normalized baseline without comparisons between the normalized groups; adjusted p-value = 0.003 for three comparisons) (Figure 3-6C). For T1-weighted postcontrast images, neither of the normalization techniques improved the repeatability of GLCM features (Figure 3-6D). The ICC distribution of texture features extracted from z-score normalized scans using the default bin width setting is presented in the fourth column in Figure 3-6C and D. The coarse intensity quantization, effectively reducing the total number of bins to two for the majority of images, decreases the repeatability of GLCM features significantly on both sequences (paired Wilcoxon test, adjusted p-value = 0.001 for both T2-weighted FLAIR and T1-weighted postcontrast).

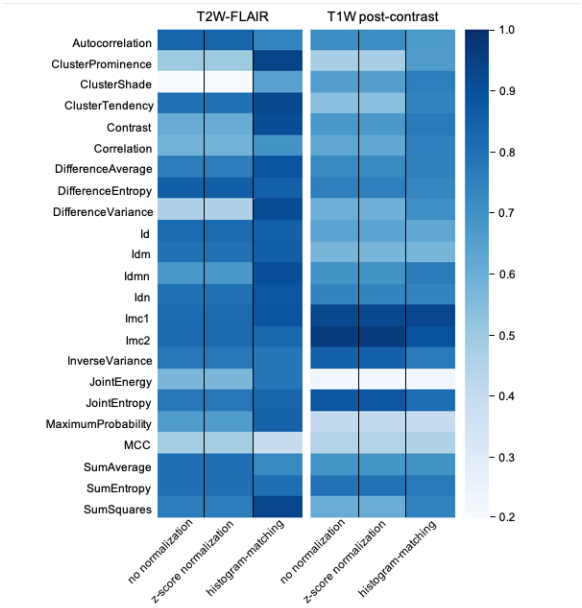

Figure 3-8: **Texture feature stability for different normalization strategies.** ICC of single intensity features (rows) extracted under different normalization conditions using relative intensity quantization with 256 quantization levels (columns) from T2W-FLAIR (left) and T1W post-contrast (right). Using relative intensity quantization, z-score normalization does not have an effect on the ICC of texture features as the constant offset and scaling does not affect the relative intensity difference between neighboring voxels.

3.3.4 Independence of feature extraction repeatability of the ROI

To exclude all segmentation-dependent factors that might influence the repeatability of radiomics intensity and texture features, we extracted features using the union of the ROI of both scan and rescan. However, this approach did not produce higher ICC values for intensity and texture features (both from T2-weighted FLAIR and T1-weighted postcontrast images) than using manual masks separately defined for scan and rescan (one-sided analysis of variance). The ICC distributions are illustrated in Figure 3-9. This finding suggests that the low repeatability of intensity and texture features in our study is driven by differences in voxel intensities within the ROI between scan and rescan as opposed to intrarater variability in segmentations.

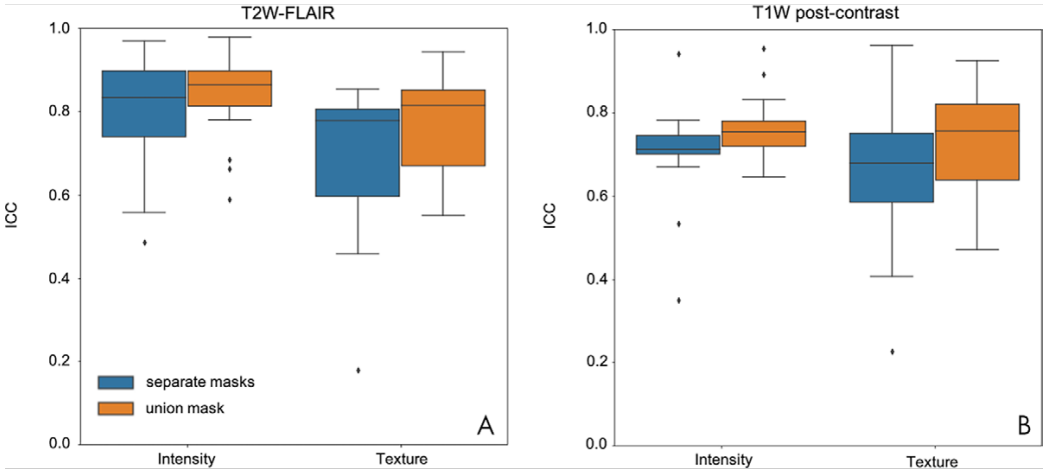

Figure 3-9: **Distribution of intensity and texture intraclass correlation coefficient (ICC) values depending on the region of interest (ROI) definition.** ICC for intensity and texture features extracted from, A, T2-weighted fluid-attenuated inversion recovery (T2W FLAIR) and, B, T1-weighted (T1W) postcontrast using manual ROI masks separately outlined for scan and rescan (blue) or the union of both masks to extract features from the scan as well as rescan (orange). There is no statistically significant difference (paired Wilcoxon test) in the ICC distributions between the ROI definitions.

3.4 Discussion

In this study, we analyzed the influence of normalization (including voxel intensity quantization) on the repeatability of radiomic feature extraction from brain MRI (T2-weighted FLAIR and T1-weighted postcontrast sequences) using the open-source software package PyRadiomics for feature extraction. The high repeatability of shape features, which are computed exclusively based on the provided manual segmentations (Figure 3-2) indicates that the segmentations are very consistent between scan and rescan. While it has been reported that features are susceptible to variations in manual segmentations [192], we excluded this as a major driver for the low repeatability of intensity and texture features in this study based on the high consistency of the segmentations. Furthermore, using the same mask to extract radiomic features from the scan and rescan to eliminate segmentation effects did not result in an improvement in the ICC between visits. This serves as additional support to the idea that image acquisition and patient-related factors have a greater influence on the lack of radiomic feature repeatability than intra-rater variability in segmentation between scan and

rescan. As such, we investigated whether the repeatability of intensity and texture features can be improved by the application of the appropriate normalization technique in combination with an adaptation of the intensity quantization strategy.

The intensity values' dependency on MRI on scanner properties, image acquisition, and image processing requires standardization of the image intensity to enable a comparison of features across patients [180]. Accordingly, there are many approaches to normalize medical imaging, particularly for the normalization of brain MRI [193]. We chose two of the most widely used normalization techniques in radiomics pipelines: z-score normalization (built into PyRadiomics) and histogram matching to a reference case [194]. The optimal intensity normalization technique is expected to result in a good overlap between the intensity histograms of the scan and rescan. Both normalization techniques significantly improved the overlap between the ROI intensity histograms of scan and rescan. However, we could identify cases in which, because of either too aggressive or insufficient brain extraction, normalization efforts had an adverse effect (Figure 3-3C and D). This is consistent with a previous study, which showed that the most commonly used brain extraction algorithms can fail in the presence of disease [195], thereby introducing another factor that can harm standardization and intensity normalization efforts.

For cohort-level analysis, we recommend manual auditing of the results of automatic brain extraction to ensure that brain extraction did not fail. Manual correction of sporadic failures might not be needed for cohort-level analyses. However, analyzing individual scans (e.g., for treatment stratification) might require manual checks of every scan to ensure appropriate brain extraction. If necessary, manual correction of the automatic brain extraction must be performed to ensure that the analysis is not impaired by flawed brain extraction and its downstream effects. We could not detect a significant difference in the JSD of normalized cases with and without brain extraction (Figure 3-4), but we did not examine downstream effects on feature repeatability.

Both normalization techniques show better repeatability for T2-weighted FLAIR images than for T1-weighted postcontrast images. One reason for this is that the repeatability of T1-weighted postcontrast scans can be complicated by variations in

contrast application and the timing of image acquisition after injection, notwithstanding the controlled research conditions under which the scans used in this study were acquired. The normalization approaches used in our study do not account for these differences, as they are based on the intensity distribution of the full input volume, including the contrast-enhancing region. These findings are consistent with those in He et al., which showed that variability introduced by contrast enhancement could negatively affect the diagnostic performance of radiomics models on CT [196].

Additionally, the parameters for feature extraction, especially the choice of voxel intensity quantization, can have marked effects not just on the repeatability of radiomic feature extraction but also on the correlation between features [182, 183]. Features that are calculated based on binned or quantized values (e.g., GLCM features) are sensitive to the choice of this setting. This is reflected in the poor ICC for texture features using a bin width of 10 on z-score normalized images (Figure 3-6C and D). Given the lack of standardized intensity ranges in MRI, relative binning is a more reasonable choice, as it results in improved repeatability.

Importantly, the effects of intensity quantization require additional examination of the correlation between features. Increasing the number of histogram bins after brain extraction and normalization results in a decrease of the correlation between GLCM features within one scan while having no adverse effect on feature repeatability. Highly redundant features may harm downstream predictive pipelines.

3.5 Limitations

There were some limitations to our study. First, we limited the examination of texture features to GLCM features because of the popularity of these descriptors with respect to other texture features. Future studies will need to thoroughly examine other classes of texture features (e.g., Laws energy, Gabor) [197, 198]. Second, features were extracted from two-dimensional axial sequences, and differences in slice placement can have an additional influence on the repeatability of radiomic feature extraction. Moreover, most researchers use radiomics features for some task (e.g., survival analysis,

disease diagnosis) to be solved via some machine learning model (e.g., random forest, support vector machine classifier). In this study, we only tested for the repeatability of features. We did not test whether trained machine learning models using these radiomic descriptors are repeatable. Last, our findings and, therefore, our recommendations, may only be valid for radiomic features extracted from newly diagnosed and untreated glioblastoma, as this was the use case in our study.

3.6 Conclusions

In summary, our findings that the optimal setting for feature extraction may vary from feature group to group (and maybe even within the separate groups) are consistent with results presented by Schwier et al. on the repeatability of radiomic feature extraction from MRI on a dataset of small prostate tumors [181]. The extraction of repeatable intensity and GLCM radiomic features from MRI requires robust standardized preprocessing and careful selection of feature extraction settings. Based on our results, we recommend using a normalization strategy (especially for unenhanced sequences) and using relative binning strategies to account for varying intensity ranges within images. Furthermore, we recommend checking the within-scan correlations between features during feature selection and using a higher number of bins to avoid feature redundancy.

3.7 Perspectives on the stability and robustness of machine learning models

In this chapter, we presented a study of the repeatability of radiomics features extracted from brain MRI and demonstrated that appropriate preprocessing could improve feature repeatability. During my PhD research, I led and contributed to additional projects on the stability and trustworthiness of machine learning algorithms for medical image analysis. These projects point towards common problems that ML models encounter on their way to clinical deployment: artificial performance inflation

through data leaks and lacking stability in predictions.

3.7.1 Flawed practices lead to performance inflation of radiomics models

While many radiomics studies report impressive performance, radiomics has lately come under scrutiny. As we have outlined in Section 2.2, besides the feature extraction, other steps in a radiomics pipeline are also prone to introducing errors. For an in-depth analysis of the quality of radiomics models published in the literature, we performed a selective review of 50 publications reporting on the development of radiomics models. The analysis of the methodology focused on model development following feature extraction. We identified a median of six flaws in the methodology of each article. These flaws could be attributed to information leaks and unproductive feature associations.

Information leaks To allow objective evaluation of the performance of a radiomics algorithm, test data must be kept strictly separated from the development dataset during normalization, feature selection, model training, and hyperparameter tuning [199, 200]. If that is not the case, we show that the performance can be artificially inflated by utilizing information from the test dataset during model development.

Unproductive feature associations Radiomics features are high-dimensional. The number of features extracted from a single case is oftentimes higher than the number of patients in the entire dataset. This can lead to an overestimation of causal relationships between radiomics features and other variables due to spurious correlations.

We identified multiple serious methodological flaws in a selected subset of the radiomics literature. Furthermore, we experimentally demonstrated how information leaks due to inconsistent data partitioning and the reliance on unproductive feature associations lead to an over-optimistic estimation of a model's performance. The

methodological mistakes outlined above were found in research articles that had all undergone peer review. Peer review is known to be a flawed process as it is characterized by a high level of randomness [201], often fails to detect substantial errors [202], and its effectiveness in improving the quality of publications is questionable [203]. Therefore, research articles on radiomics models should be read with care. Lastly, neither information leaks due to inconsistent partitioning nor spurious correlations [204, 205] are problems exclusive to radiomics – deep learning suffers from them as well.

3.7.2 Improving the repeatability of deep learning classification predictions

Low repeatability affects not just radiomics features but also the predictions of DL classification models. During model development and evaluation, much attention is paid to the classification performance of models. However, their repeatability, a crucial indicator of a model's robustness and requirement to win the trust of healthcare professionals and patients alike, is rarely assessed. The lack of attention to model robustness leads to the development of models with low test-retest repeatability [39], which are unsafe for clinical deployment. Given the importance of developing reliable deep-learning algorithms for medicine, we developed strategies to improve their repeatability.

We studied the test-retest repeatability of binary, multi-class, and ordinal classification and regression with and without Monte Carlo (MC) dropout. Monte Carlo dropout is a straightforward approach to prevent models from making over-confident predictions [206]. We will introduce MC dropout in more detail in Section 4.2.3. For this study, we utilized four medical image ordinal classification tasks from public and private datasets: knee osteoarthritis, cervical cancer screening, breast density estimation, and retinopathy of prematurity (see Section 4.2.1 for more details on the datasets).

Through extensive experimental validation, we showed that different modeling

approaches led to drastically different test-retest repeatability. Because the classification labels for all problems we assessed were ordinal, most variability occurred for cases close to the decision boundary between two classes. We found that using regression led to the highest repeatability for models trained without MC dropout, followed by ordinal classification. Conventional nominal classification models showed the lowest repeatability performance. Additionally, using Monte Carlo dropout led to significantly higher repeatability without decreasing and, in some cases, even improving classification performance.

In summary, we demonstrated how the test-retest repeatability of DL algorithms could be improved using Monte Carlo dropout. This represents an easy-to-implement method to develop robust models that deserve the trust of healthcare professionals and patients.

3.7.3 Automatic quality assessments using predictive uncertainty

Segmentation uncertainty allows the detection of suboptimal segmentations The performance of DL algorithms is usually evaluated on a full dataset consisting of many cases. However, in clinical practice, the validity of every single prediction is essential. While DL algorithms have been touted as high-performing automatic diagnostic tools that are more cost-efficient and faster than humans, they can silently fail for various and oftentimes unknown reasons such as inconsistencies in the data format, noise, and domain shift [207]. In short, DL algorithms cannot be trusted blindly [208]; the quality of every single prediction, classification and segmentation alike must be checked by human experts. Consequently, obtaining an automatic assessment of a prediction's reliability would be highly desirable.

While a classification prediction is either right or wrong, the quality of a segmentation can be assessed on a continuous quality spectrum depending on the agreement between a ground truth label and the algorithm's output, e.g., through the Dice score. Recently, Bayesian approximation methods, like deep ensembles and Monte Carlo

dropout, have been proposed as tools to measure a model's uncertainty associated with a prediction. However, the influence of design choices, like the cost function used for model development, on the quality of these uncertainty metrics is unknown. Therefore, we examined the potential of three different methodological approaches [209, 210, 211, 212] that are trained using different cost functions. We obtain measures of segmentation uncertainty for the 3D segmentation of lung nodules on low-dose CT scans [213].

We found that areas of high uncertainty were localized at the margins of segmentations. We visualized the spatial distribution of segmentation uncertainty by assigning an uncertainty value to each voxel in a CT volume based on the algorithm's predictive uncertainty. Upon visual assessment of these spatial uncertainty distributions, areas of high uncertainty were localized in the periphery of the predicted segmentation. They agreed with false-labeled voxels, i.e., areas of false positives and false negatives were visually aligned with areas of high uncertainty.

Furthermore, uncertainty metrics associated with the predictions of all three methods showed high correlations with the Dice score of a segmentation. The strong relationship between ground truth-independent uncertainty metrics and segmentation performance indicates that predictive uncertainty metrics may have the potential to be used as an indicator of a segmentation's quality.

Lastly, we found that uncertainty metrics can be used to provide a reliable estimate of a segmentation's quality. These quality estimates can potentially be used to flag low-quality predictions and speed up the evaluation of large study cohorts by flagging only cases of expected low quality for human review. They could furthermore assist in the generation of smart worklists [214] for clinicians checking and correcting the output of automatic segmentation pipelines, by prioritizing challenging samples (with an expected low quality) early during a shift when the raters can focus better [215].

Monitoring segmentation uncertainty under domain-shift While uncertainty metrics have substantial potential to increase trust in "black box" neural networks by allowing them to detect low-quality output automatically, these estimates should

remain stable under naturally occurring domain shifts. In an additional study, we evaluated the relationship between epistemic uncertainty and segmentation quality under domain shift within two clinical contexts: optic disc segmentation in retinal photographs and brain tumor segmentation from multi-modal brain MRI. Specifically, we assessed the behavior of uncertainty metrics derived from the same three methods described in the previous section: conventional neural networks, deep ensembles, and Monte Carlo dropout algorithms trained using either soft Dice or weighted categorical cross-entropy. We introduced domain shifts by excluding a group with a known characteristic (glaucoma for optic disc segmentation and low-grade glioma for brain tumor segmentation) from model development and using the excluded data as additional, domain-shifted test data.

As expected, the performance of all models dropped slightly on the domain-shifted test data compared to the in-domain test set. While there was no change in the Pearson correlation coefficients, we did observe a change in the slope of the linear relationship between the uncertainty metrics and the Dice scores of the segmentations. To describe and quantify the observed changes under domain shift, we introduced a new metric, the *empirical strength distribution*. We found that shifts in the empirical strength distributions between training, in-domain, and domain-shifted test datasets caused a decrease in the performance of uncertainty-based automatic estimation of segmentation quality. Therefore, quality assessment tools based on the strong relationship between epistemic uncertainty and segmentation quality can be stable under small domain shifts if the empirical strength distribution is not affected. Developers should thoroughly evaluate the strength relationships for all available data and, if possible, under domain shift to ensure the validity of these uncertainty estimates on unseen data.

3.7.4 Recommendations for the development of robust ML algorithms

Despite the tremendous optimism that ML will change medicine, surprisingly, few models are used in clinical decision support. In a review of ML models proposed for

the diagnosis or prognosis of COVID-19 from chest X-ray or CT images, Roberts et al. report that out of 62 studies, none of the models are of potential clinical use [216]. The reasons ranged from missing robustness or sensitivity analysis of the models to insufficient reporting on limitations, potential biases, or generalizability issues. The successful translation will require authors and reviewers alike to adhere to rigorous standards and best practices, like the ones outlined in the CLAIM checklist [217].

AI research – with medical AI being no exception – is sometimes described as a race to develop new methodologies that outperform the previous state-of-the-art based on specific metrics. However, optimizing metrics has become a central aspect of AI research, leading to manipulation or gaming of results, sometimes even inadvertently, e.g., by not adhering to best practices and a disproportionate focus on short-term goals [218]. As noted by Charles Goodhart and what is today known as Goodhart's law: *"When a measure becomes a target, it ceases to be a good measure"* [219]. Thomas and Uminsky proposed a framework promoting the balanced use of metrics for evaluating AI algorithms [218]. Their recommendations include using multiple metrics, including qualitative accounts, to get a more detailed picture of model performance and the involvement of stakeholders in the evaluation process. As we outline in Chapter 5, this potentially requires the development of new metrics that allow rigorous and truthful performance evaluation.

Chapter 4

Network designs for latent, continuously valued variables

Many variables of interest in clinical medicine, like disease severity, are recorded using discrete ordinal categories such as normal/mild/moderate/severe. These labels are used to train and evaluate disease severity prediction models. However, ordinal categories represent a simplification of an underlying continuous severity spectrum. Using continuous scores instead of ordinal categories is more sensitive to detecting small changes in disease severity over time. Additionally, for these continuously valued variables that are measured through ordinal labels, characteristic noise patterns around the boundaries between classes have been described. Some experts tend to consistently over-call, meaning that they have a lower threshold of what they consider to be "abnormal." In contrast, others tend to under-call, causing a pattern of inter-annotator variability.

In this chapter, we present two applications based on the concept of a latent, continuously valued variable of interest. We first present a generalized framework that accurately predicts continuously valued variables using only discrete ordinal labels during model development. Our novel framework was validated on three clinical prediction tasks and bridged the gap between discrete ordinal labels and the underlying continuously valued variables. Second, we show that previously described approaches for identifying differences between annotators' labeling behavior are inappropriate for

a latent continuously valued variable. Additionally, we present two new techniques to jointly learn classification and the individual annotators' bias in the binarization threshold.

The work presented in this chapter was performed in collaboration with Andréanne Lemay, John Peter Campbell, Susan Ostmo, Michael Chiang, Christopher Bridge, Matthew Li, Praveer Singh, Aaron Coyner, and Jayashree Kalpathy-Cramer. A manuscript is currently in preparation.

4.1 Introduction

4.1.1 Role of continuously valued variables in medical image analysis

Many clinical variables, like disease severity, are communicated and recorded as discrete ordinal classes. However, in reality, they are distributed on a continuous spectrum [41]. The discretization of continuously valued variables facilitates documentation and communication and standardizes treatment decisions at the cost of losing information. As illustrated in Figure 4-1, two patients that fall into the same severity category will receive the same label. Consequently, if their data is used to develop a prediction model, it will be treated exactly the same during training, regardless of their exact position along the continuous spectrum.

Most of the tremendous successes of deep learning based algorithms for disease severity prediction, have been built on the simplifying assumption that disease severity prediction can be formulated as a simple classification task. Researchers mostly use DL architectures that are intended for the classification of nominal categories and ignore the inherent ordinal nature of the available training labels [43, 44]. More so, disease severity prediction tasks are often simplified even more by treating them as binary problems, e.g., disease detection [90], the identification of severe disease [220], or cases for referral [221, 222].

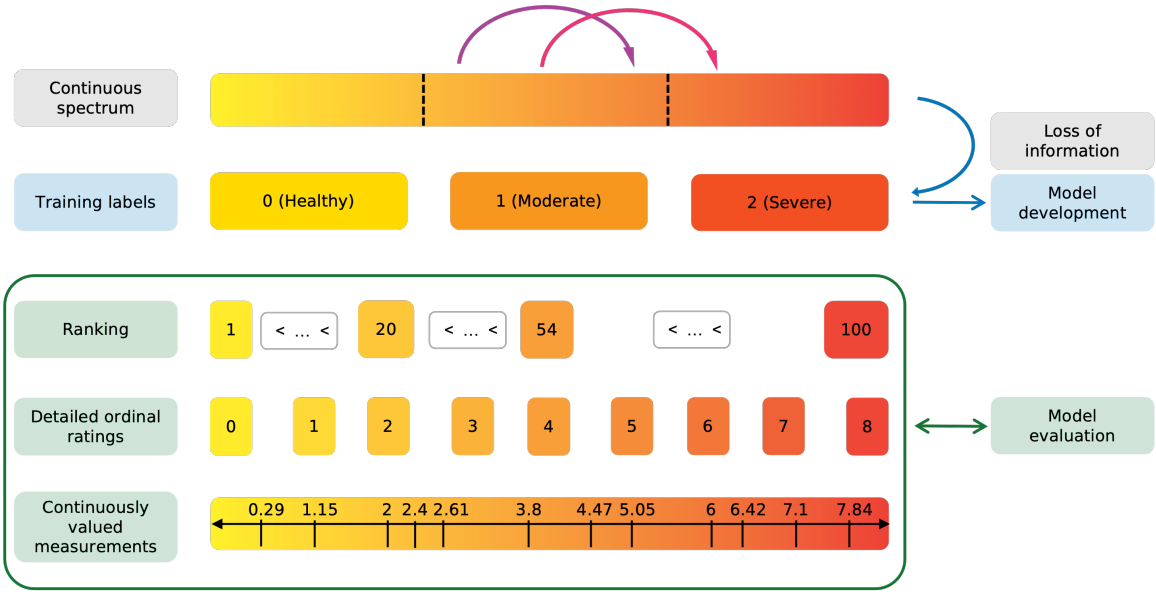

Figure 4-1: **Relationship between the underlying continuous variable of interest and the training and evaluation labels.** The conversion of a latent continuously distributed variable into discrete ordinal variables for model development represents a loss of information. The purple and magenta arrows represent temporal changes in disease severity. The available annotation types for the evaluation of the continuous predictions are presented in the green box: Rankings, ordinal ratings on a more detailed scale than the training labels, and continuously valued measurements.

Advantages of continuous scores In addition to the class, the position of a case on the continuous spectrum contains valuable clinical information that is not captured by current approaches [223]. Therefore, the use of continuous scores to describe clinical variables that are distributed on a continuous spectrum provides several advantages over discrete ordinal variables. First, continuous metrics allow for the detection and quantification of changes within a class, i.e., an increase in disease severity that does not constitute a transition between class n and $n+1$ [224, 225]. In Figure 4-1 the purple and magenta arrows represent similar large increases in disease severity. While the magenta transition would get detected by the traditional classification approach, the purple transition would not. The only difference between the two transitions is that the magenta one crosses the class boundary from moderate to severe, while the purple arrow represents a within class change. The detection of within-class changes allows to detect disease deterioration earlier and act upon it if required.

Second, the higher degree of information presented in continuous versus ordinal

scores can be useful for efficient patient stratification, particularly the identification of cases close to a decision boundary. Third, expert perception of class boundaries can be subject to changes over time [226]. Therefore, models ignoring the continuous nature of disease severity could become less valuable over time as, e.g., the perception of what constitutes mild versus moderate disease severity shifts. Lastly, an algorithm that predicts a continuous score is more likely to fulfill notions of individual fairness as similar individuals that are close to the label decision boundaries will be more likely to receive similar scores compared to using a simple classification algorithm [227].

4.1.2 Predicting granular disease severity information

Previous attempts to predict continuous disease severity scores have made use of either nominal classification networks or Siamese networks. Redd et al. proposed to aggregate the softmax outputs from a nominal 3-class convolutional network into one continuously valued vascular severity score for disease severity classification in ROP [228]. This score strongly correlates with experts' ranking of overall disease severity and groups with different disease severity show significantly different severity score distributions [228, 229]. Furthermore, changes in this score over time accurately reflect disease progression [225]. However, the training objective of multi-class classification models is to separate the latent space representation of classes as much as possible. This could therefore lead to unstable predictions and confusion at the class boundaries.

Using Siamese networks, Li et al. showed that the continuously valued difference relative to a reference pool of images correlates with expert's rankings of disease severity and reflects temporal changes in severity in knee osteoarthritis and ROP [224]. Similarly, in a study on an x-ray based severity score for COVID-19, a score generated by a Siamese network highly correlated with radiologist-determined severity scores [230]. However, the performance of Siamese networks for the predictions of continuous scores has not been compared to other methods and their calibration has not been studied yet. The limited availability of datasets and labels with more granular information than the ordinal labels typically used in clinical practice hampers efforts to develop and validate DL models that predict continuous instead of discrete

values.

4.1.3 Noisy ordinal annotations

The ground truth of many medical imaging datasets can be considered noisy, as described in Section 2.4. Due to a high ambiguity in numerous clinical image classification tasks, experts often do not agree with each other, causing inter-annotator variability and inconsistencies in the ground truth of datasets. Given deep learning algorithms' tendency to memorize noisy training data, many methods have been developed to limit the influence of label noise on the training process and prevent the neural network to overfit the noise in the training data. Approaches include the development of loss functions that are robust to noise [108, 231], reducing the influence of noisy data points on the training process [232], or correcting potentially wrong labels [233]. However, these methods rely on the assumption that the noise is random, independent of the person who generated a label, and impeding the training process.

Annotator-specific noise patterns In some situations, label noise can contain informative patterns that reveal information about the data or the people who labeled it. In crowdsourcing data annotations, for example, some workers are more reliable than others [234]. And in nominal classification tasks, some classes tend to be more similar than others and are therefore more frequently confused by annotators [235].

For continuously valued variables, experts tend to disagree mostly about cases close to a decision boundary [47]. A case that can be considered normal but already shows some of the characteristics of the disease may be classified as normal by one expert. Still, another expert may call it moderate disease severity. Interestingly, this noise component is not independent of the annotators. Some experts tend to consistently over-call, meaning that they have a lower threshold of what they consider to represent "abnormal," while others tend to under-call. One such scenario for a binary classification problem with two annotators who have different binarization thresholds is illustrated in Figure 4-2A. In our model, the reader-dependent noise can therefore be characterized by their individual decision thresholds.

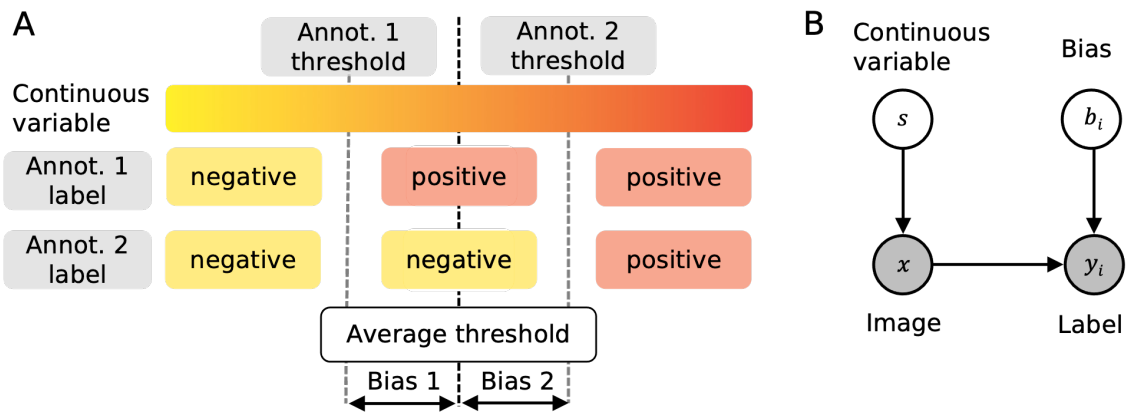

Figure 4-2: **Continuously valued variables and annotator bias.** A: Binary labels for a continuously valued variable generated by two annotators with different binarization thresholds. The difference between the individual binarization thresholds causes disagreements between the labels from both annotators. B: Model for label generation. White nodes represent variables that are not directly measurable. The latent continuously valued variable of interest s is represented in an image x. A binary label y_i is then provided by annotator i. This label, in turn, is influenced by the annotator's individual bias b_i.

4.1.4 Approaches to learning annotator-specific characteristics

Learning annotators' individual characteristics, like their reliability or decision thresholds, not only reveals important information about every annotator but also enables mitigation of the noise in the training data. In medicine, acquiring such a gold standard is almost impossible for most variables. One example of an approach based on expectation-maximization is the STAPLE algorithm, a popular method to generate a ground truth segmentation from noisy segmentations provided by different annotators [123]. The STAPLE algorithm learns each annotator's sensitivity and specificity and uses this information to generate a joint segmentation that can be considered the ground truth given the noisy input. One substantial downside of such expectation maximization approaches is that they require several labels for each sample. However, obtaining annotations for medical imaging is highly time and cost-intensive; therefore, acquiring multiple labels for one case is not feasible in most cases.

Traditionally, reader characteristics have been learned using multiple annotations per sample, e.g, using expectation-maximization techniques [236] or comparing them

to a gold standard ground truth. In medicine, acquiring such a gold standard is almost impossible for most variables. One example of an approach based on expectation-maximization is the STAPLE algorithm, a popular method to generate a ground truth segmentation from noisy segmentations provided by different annotators [123]. The STAPLE algorithm learns each annotator's sensitivity and specificity and uses this information to generate a joint segmentation that can be considered the ground truth given the noisy input. One substantial downside of such expectation maximization approaches is that they require several labels for each sample. However, obtaining annotations for medical imaging is highly time and cost-intensive; therefore, acquiring multiple labels for one case is not feasible in most cases.

Only a few articles in the literature address the problem of reader-dependent noise distributions in deep learning. In their work presenting the DoctorNet, Guan al. assume that the annotations provided by some physicians are more reliable than others [237]. The model learns to predict how each member of a group of experts would annotate an image. Subsequently, DoctorNet learns an individual weighing factor for each expert, reflecting their reliability. The final prediction is a weighted average of the model's prediction for each expert in the training group. Following this training procedure, DoctorNet outperformed a conventional deep learning model trained on adjudicated ground truth labels.

If we had access to annotators' patterns of confusion between nominal label classes, label probabilities could be corrected for the training of a deep learning model. Khetan et al. propose Model Bootstrapped Expectation Maximization (MBEM), an extension of expectation-maximization approaches that works with only a few annotations per sample [238]. The authors propose to train a deep learning model on aggregated noisy labels and use the predictions from this model to estimate confusion matrices for each annotator. In an iterative process, the confusion matrices are then used to correct the labels and retrain the prediction model using the corrected labels.

An all-in-one solution for the joint learning of classification and reader confusion matrices has been proposed by Tanno et al. [239]. The solution consists of a convolutional neural network extended by learnable confusion matrices. The joint learning

of the prediction model and the confusion matrices improved the model's robustness to label noise. However, the currently available methods are developed for nominal classification, assuming no ordinal relationship between the classes. In particular, they are unable to capture a continuous distribution underlying the ordinal labels.

4.1.5 Chapter outline

In this chapter, we describe how the shift from a purely discrete ordinal view of medical variables to a detailed one of a continuous spectrum can be achieved through careful network design.

A generalized framework to predict continuous scores from discrete ordinal labels First, we aim to identify model development strategies that lead to the prediction of accurate continuous scores. Most importantly, while we utilize widely available discrete ordinal labels for training, the models' performance to predict accurate continuous scores is evaluated using labels on a finer scale than the training ground truth (illustrated in the green box in Figure 4-1) on three datasets: disease severity prediction for ROP and knee osteoarthritis and breast density estimation from mammograms. Following this process, we demonstrate that it is possible to develop models that are capable of recovering the information lost through the discretization of the continuous target variable.

Learning the bias of individual annotators from single ordinal labels In the second study presented in this chapter, we assume that a substantial amount of inter-annotator disagreement that can be observed for ordinal annotation problems is caused by annotator-specific biases in their inter-class thresholds. We illustrate how these individual biases that affect the generation of reliable ordinal labels for continuous variables can be characterized with the help of deep learning approaches. To this end, we introduce MBEM$_{cts}$, an extension of MBEM to continuously valued variables, and BiasNet, to learn a robust classification and biased binarization thresholds of individual annotators from "noisy" single labels.

4.2 Methods to learn continuously valued variables from ordinal labels

4.2.1 Datasets

All images were de-identified prior to data access. The dataset of knee x-rays is publicly available and our institution's Institutional Review Board waived the requirement to obtain informed consent for the breast density and ROP datasets. Dataset splits were performed on a patient level. The size of all dasets is listed in Table 4.1 and class distributions for each dataset are listed in Appendix A.1.

Retinopathy of prematurity ROP is an eye disorder mainly developed by prematurely born babies and is one of the leading causes of preventable childhood blindness [240]. It is characterized by a continuous spectrum of abnormal growth of retinal blood vessels which is typically categorized into three discrete severity classes: normal, pre-plus, or plus [41, 241]. We used the same dataset, labels, and preprocessing as described by Brown et al. [45]. In addition to the standard diagnostic labels, the test set was labeled by five raters on a scale from 1 to 9 and five experts ranked an additional 100 ROP photographs based on severity [45, 229]. To verify the out-of-distribution performance, 7893 retinal images acquired in Nepal (7565 normal, 217 pre-plus, 111 plus) were classified into the same 1 to 9 scale as the in-distribution test set by a single expert rater.

Knee osteoarthritis The global prevalence of knee osteoarthritis is 22.9% for individuals over 40, causing chronic pain and functional disability [242]. Knee osteoarthritis can be diagnosed with radiographic images and disease severity is typically evaluated using the Kellgren-Lawrence (KL) scale consisting of the following severity categories: none, doubtful, mild, moderate, and severe [243]. We used the Multicenter Osteoarthritis Study (MOST) dataset. 100 images from the test were ranked by their severity by three experts [224]. All images were center cropped to 224x224 pixels and intensity scaled between 0 and 1 as preprocessing.

Breast density Female breast cancer is estimated to be the most commonly diagnosed cancer worldwide and is the leading cause of cancer death in women [244]. Breast density is typically categorized as fatty, scattered, heterogeneous, or dense, depending on the amount of fibroglandular tissue present [245]. Women with high breast density are at a higher risk of developing breast cancer and require additional MRI screening [246, 247]. We used a subset of the Digital Mammographic Imaging Screening Trial (DMIST) dataset [248]. Furthermore, for 1892 mammographs from the test dataset, an automatic assessment of the volumetric breast density was obtained using the commercially available Volpara Density software which has demonstrated a good agreement with expert ratings of breast density (see Figure 4-10A) [249, 250]. Preprocessed mammograms were of size 224x224 pixels.

Table 4.1: Summary of dataset size and training/validation/test splits of the three datasets used for this study: disease severity prediction in retinopathy of prematurity (ROP) and knee osteoarthritis (OA) and breast density prediction

Dataset	**Size**	**Training**	**Validation**	**Test**
ROP	5611	4322	722	467 (9-point scale) 100 (ranked)
Knee OA	14273	12268	1905	100 (ranked)
Breast Density	83034	70293	10849	1892 (Volpara Density)

4.2.2 Network designs

We developed algorithms using four different network designs: nominal multi-class and ordinal classification, regression, and Siamese. The model output was converted to a continuous score value for each model to represent the underlying severity spectrum of the medical task.

Nominal classification All classification models for this study were trained with categorical cross-entropy loss. The continuous severity score was computed as the sum of the softmax outputs weighted by their class (Equation 4.1) [228], leading to scores from 0 to k-1.

$$Cl_{score} = \sum_{i=1}^{k} p_i \times i - 1 \tag{4.1}$$

with k being the number of classes and p_i the softmax probability of class i.

Ordinal classification In ordinal classification, the multi-class classification task is broken up into $k - 1$ binary classification tasks, leading to one output unit less than the number of classes [251]. During training, the ordinal loss function penalizes larger misclassification errors more than smaller errors (e.g., predicting class 2 when the ground truth label is 0 is penalized more than if the model predicts class 1). We used the CORAL loss as described by Cao et al. for model optimization [252].

A continuous score was generated by summing over the output probabilities (Equation 4.2), resulting in values ranging from 0 to k-1.

$$O_{score} = \sum_{i=1}^{k} p_i \tag{4.2}$$

Regression Similar to ordinal models, regression models require the ordinality of the target variable. However, unlike ordinal models, the output of regression models is a continuous value rather than a discrete class. The regression models were trained using the mean squared error loss function with the class number as the target value. The raw model output yielded a continuous value; hence, no conversion was required to receive a continuous score.

Siamese Siamese models compare pairs of images to evaluate their similarity and were shown to be effective in generating continuously valued predictions [253, 224]. They are composed of two branches of identical sub-networks with shared weights where each of the two images is processed in one of the branches. The lower the Euclidean distance between the outputs of each branch, the higher the similarity between the inputs. During training, the model is encouraged to yield small Euclidean distances for images of the same class and large distances for images from different classes using a contrastive loss [254] paired with a cross-entropy loss. Following a

procedure described by Li et al., at test time, the target images were compared to ten anchor images associated with class 0 [224]. Here, the continuous score was the median of the Euclidean distances between the target and the ten anchor images.

4.2.3 Monte Carlo dropout

Dropping out single neurons in fully connected layers or full activation maps in convolutional layers during the training of a neural network mitigates overfitting and helps to regularize the learning process [255, 256]. Utilizing dropout not just during training but also at test time, yields N slightly different Monte Carlo (MC) predictions [206]. The MC predictions can subsequently be averaged, to obtain the final output prediction. All models referred to as *MC models* were trained with spatial dropout after each residual block of the ResNet models, and $N = 50$ MC iterations at test time. The dropout rates are 0.2, 0.2, and 0.1 for ROP, knee osteoarthritis, and breast density, respectively, and were selected empirically and based on current literature.

4.2.4 Model training

Model parameters were selected based on initial data exploration and empirical results. A detailed description of the training parameters can be found below.

ROP ROP models had a ResNet18 architecture and were trained with a batch size of 24, a learning rate of 1.0×10^{-4} for 25 epochs, and the best model was selected using the highest accuracy on the validation set. Balanced class sampling mitigated the class imbalance during training. Data augmentation consisted of random rotations of up to $\pm$ 15 degrees with a probability of 0.5, random flips with a probability of 0.5, and random zooms of 0.9 to 1.1 with a probability of 0.5.

Knee osteoarthritis A ResNet50 architecture was selected for the knee osteoarthritis model and was trained with a batch size of 16 and an initial learning rate of 5.0×10^{-6} for 75 epochs. The final model was chosen based on the best loss value on the validation set. The data sampler used balanced weights during training to help

with data imbalance. Images were randomly rotated of ± 15 degrees with a probability of 0.5 and randomly flipped with a probability of 0.5 as data augmentation.

Breast density Breast density models were trained with a ResNet50 architecture for 75 epochs using batches of 8 images with a learning rate of 5.0×10^{-5}. The best model was selected using the best loss score on the validation set. The same data augmentation as the knee osteoarthritis model was applied for breast density.

4.2.5 Evaluation

Metrics

Ranked datasets The model performance was evaluated based on the ranked test data using the following three metrics. First, we computed Spearman's rank coefficient between the rank and the predicted score. A monotonic increase between both metrics was expected; hence, a Spearman coefficient of 1 corresponds to a perfect correlation. Second, we computed the agreement between the ground truth rank and the rank based on the continuous score using mean squared error (MSE) to quantify the correspondence between the predictions and ground truth. Here, the ranks were normalized to the maximum rank. Finally, the classification performance was assessed using the area under the receiver-operator-curve for clinically relevant binary classification. We defined the clinically relevant classification and normal/pre-plus versus plus for ROP, none/doubtful versus mild/moderate/severe for knee osteoarthritis, and fatty/scattered versus heterogeneous/dense for breast density.

ROP A subset of the ROP test set had expert ratings from 1 to 9 based on the quantitative scale previously published by Taylor et al. [229]. The correspondence between the expert rating and the predicted continuous scores was measured using the MSE.

Statistical analysis

Metrics were bootstrapped (500 iterations) and 95% confidence intervals were evaluated for statistical analysis. Bootstrapped metrics yielding two-sided t-test with a p-value ≤ 0.05 were considered statistically different.

4.3 Methods to learn the bias of individual annotators

Our methods to learn the characteristics of individual annotators are based on the assumption illustrated in Figure 4-2A: the only difference between the annotators is a bias term. We also assume that the annotators are reliable and consistently biased. We present two methods to learn individual readers' biases. The first method is an extension to the MBEM method introduced by Khetan et al. for continuous values, which we refer to as MBEM$_{cts}$. Second, we present the BiasNet, in which the conventional logistic sigmoid operator after the last layer is extended by a learnable individual bias parameter for every annotator who labeled samples in the dataset.

4.3.1 Assumptions and Notation

For simplicity, we are describing and testing our methods using binary labels. Both the MBEM$_{cts}$ and the BiasNet can easily be extended to ordinal multi-class settings. We assume that the underlying ground truth is a continuously valued variable s, for example, disease severity. Furthermore, s is considered a latent variable that cannot be used for training. It is measured through the binary label (or discrete ordinal label in a multi-class setting) y, based on its appearance on an image x. However, the binarization threshold b_i is annotator-dependent; therefore, the binarized label y_i is corrupted and characterized by the annotator. A graphical representation of this model is presented in Figure 4-2B.

For simplicity, we will refer to the annotator-dependent label noise as "bias," while we use the term "noise" for random noise that is independent of the annotator. As

depicted in Figure 4-2A, we use the term threshold for the individual binarization thresholds. In contrast, the term "bias" denotes the difference between the average threshold of all annotators and an individual threshold.

In a conventional deep learning setting, the prediction of a model is described as

$$f(x|\theta) = f_{sig}(\hat{s}) = \hat{y}.$$

Here, θ denotes the learned model weights and f_{sig} the logistic sigmoid function. With the addition of the annotator-dependent bias, we obtain

$$f(x|\theta, r_i) = f_{sig}(\hat{s} + \hat{b}_i) = \hat{y}_i,$$

where r_i denotes annotator i.

4.3.2 MBEM$_{cts}$: Model bootstrapped expectation maximization for continuous variables

Following the process outlined by Khetan et al. and illustrated in Figure 4-3A, we trained a deep learning algorithm using the majority vote of the available labels for each sample as the ground truth. Therefore, the training process is agnostic to the annotators' identities. After training the initial model, we used it to obtain the continuously valued predictions $\hat{s}$ for the training samples. These predictions were then used to identify an optimal decision threshold for the labels provided by each annotator. We define the empirical optimal decision threshold $\hat{t}_i$ for annotator i, as the binarization threshold that results in an equal sensitivity and specificity based on the labels provided by annotator i. The training labels were then updated based on the continuously valued predictions and the annotator-dependent binarization thresholds. In an iterative process, the majority vote of the updated training labels was used to re-train the deep learning model and re-update the labels. This process was repeated until the number of updated training labels was lower than a pre-specified threshold.

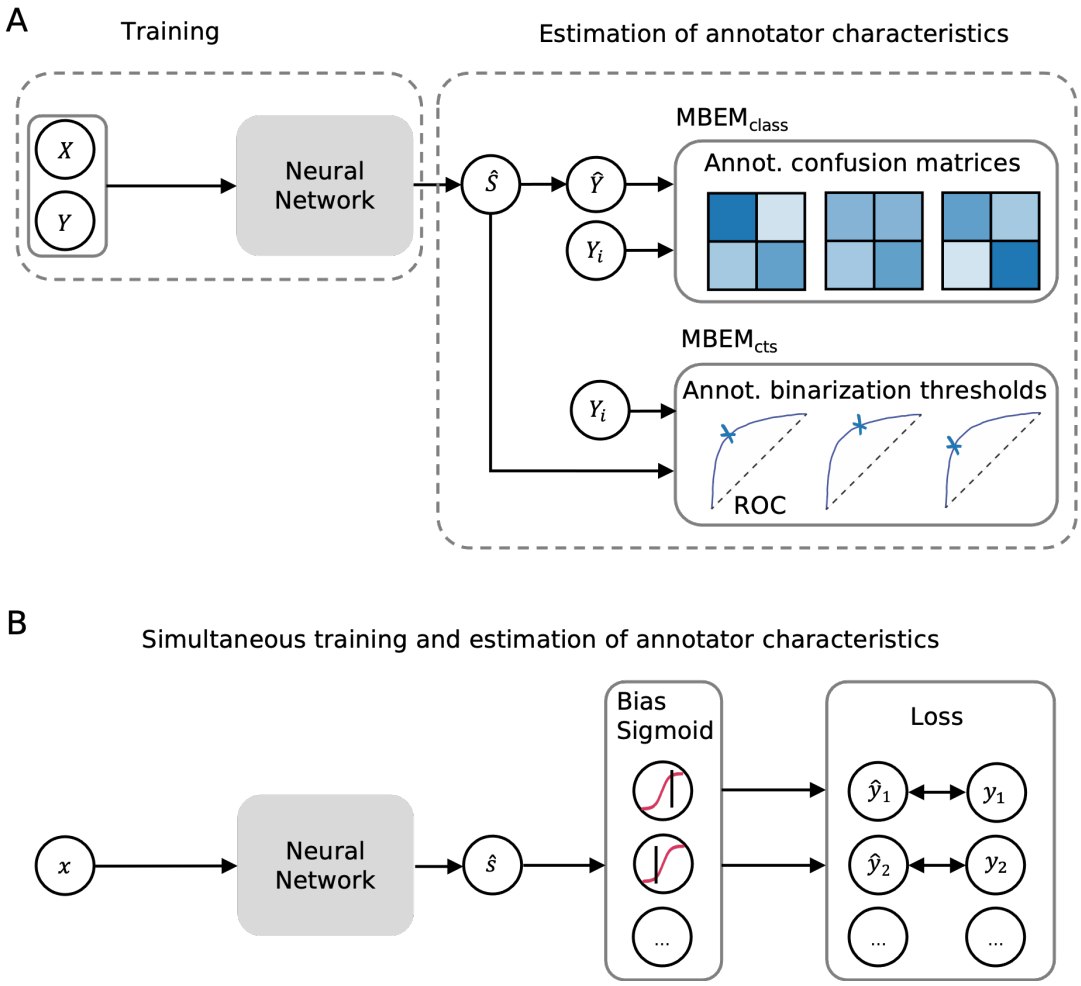

Figure 4-3: **Training schematic of MBEM$_{class}$, MBEM$_{cts}$, and BiasNet.** A: MBEM$_{class}$ and MBEM$_{cts}$: A neural network is trained on a set of training images X and the associated anonymous labels Y. To generate the annotator-specific confusion matrices, the binarized predictions $\hat{Y}$ are compared to $\hat{Y}_i$, the training labels provided by annotator i. The annotator-specific binarization thresholds are identified by determining the optimal binarization threshold based on the continuously valued, pre-sigmoid predictions $\hat{S}$ and $\hat{Y}_i$. B: A BiasNet model consists of a neural network with a specialized sigmoid layer with a trainable bias parameter for each annotator. During training, a separate loss is computed for the labels from each annotator and then used to update the neural network and annotator-specific bias parameter.

4.3.3 BiasNet - Biased sigmoid layer

To convert any neural classification network into a BiasNet, the biased sigmoid layer can be used instead of the unbiased logistic sigmoid function. The logistic sigmoid

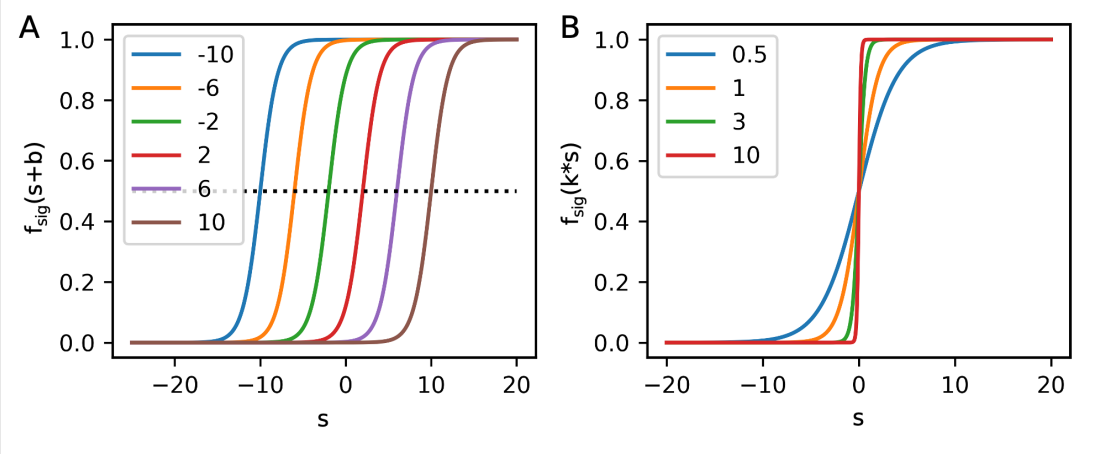

Figure 4-4: **Influence of bias and width parameters on the logistic sigmoid function** A: Different values of the bias parameter b influence the position of the logistic sigmoid function along the x-axis. The dotted black line indicates the conventional binarization threshold. B: Different values of the width parameter, k, influence the slope or width of the logistic sigmoid function.

function is defined as

$$f_{sig}(s) = \frac{1}{1 + e^{-k(s+b)}},$$

where b denotes the bias term and k is a scaling factor determining the sigmoid's width. To receive a conventional logistic sigmoid shape, one chooses $k = 1$ and $b = 0$. Figure 4-4 illustrates how different values of b (plot A) and k (plot B) influence the position and shape of the logistic sigmoid function.

As depicted in Figure 4-3B, the model was trained using all available labels for each sample. No label aggregation would be needed if several labels were available for each sample. During training, a separate loss term is computed based on the labels for each annotator. Therefore, we obtained an annotator-dependent loss L_i to update the model's weights and the bias parameter for annotator i. The model was trained until both the model itself and the trainable bias parameters converge.

4.3.4 Two component noise model

For our experiments, we used a two-component noise model for human annotators: The first component was random noise. We generated random Gaussian noise by sampling

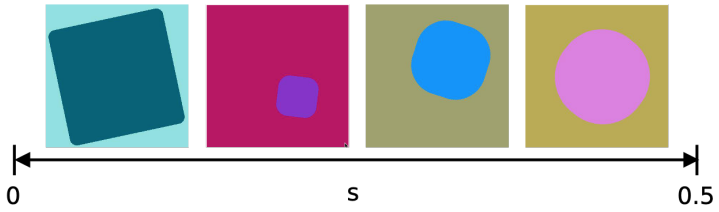

Figure 4-5: **Rounded squares dataset.** The synthetic dataset consisted of squares with rounded corners. The roundness of the corners was used as continuously valued ground truth which ranged from 0 (perfect square) to 0.5 (perfect circle).

from a Gaussian distribution with mean 0 and standard deviation σ_n. The noise was then added to the continuously valued ground truth to obtain a noisy version of the ground truth. We then generated the second, annotator-specific noise component by binarizing the noisy continuous ground truth using annotator-dependent binarization thresholds. To model the biased decision thresholds of individual annotators, we sampled the thresholds from a Gaussian distribution with a mean of 0.25 and standard deviation $\sigma_r = 0.05$. We assumed that the random noise is the same for every reader.

4.3.5 Generation of the synthetic dataset

To validate our method, we required a dataset with access to the underlying continuous ground truth without any noise. Therefore, we generated a synthetic dataset of squares with rounded corners. The geometric forms ranged from a perfect square to a circle. We defined the continuously valued ground truth as the ratio between the rounded corner's radius and the square's side length. The ground truth ranged from 0 to 0.5. We randomly sampled the side length of the square, its rotation angle, and the fore- and background colors (three channels) from uniform distributions. A detailed description of the variables used to generate the dataset can be found in Appendix B.1. The dataset consisted of 10000 images. 9500 were used for training and 500 for validation. Some examples are depicted in Figure 4-5. No further image augmentation was performed for training.

4.3.6 Regularization terms

The values of the learned bias terms depend on the range of the pre-sigmoid continuous score $\hat{s}$. Since the range of $\hat{s}$ is unbounded, learning annotators' biases is an ill-defined problem. As the optimization process during the training of a deep learning model incentivizes the model to separate the positive from the negative class as much as possible, the range of $\hat{s}$ may not converge during training. Thus, we added the following two regularization terms to the loss to limit the range of $\hat{s}$:

$$L_{range} = \frac{1}{N} \sum_{i=1}^{N} \begin{cases} (|\hat{s}| - m)^2, & |\hat{s}| > m \\ 0, & \text{otherwise} \end{cases}.$$

Here, $m \geq 0$ denotes the limit term. As a result, all values of $\hat{s}$ that are larger than m were penalized. We achieved an additional speedup of the bias learning process by varying the width of the logistic sigmoid via the scaling factor k. The combination of $k = 5$ and $m = 1$ was chosen empirically.

Additionally, we experimentally found that a small direct regularization of the bias parameters sped up the convergence of the bias parameters:

$$L_{bias} = \frac{1}{R} \sum_{i=1}^{R} \hat{b}_i^2,$$

where R denotes the number of annotators. By constraining the range for the continuous score $\hat{s}$, the values of b_i were lower, and the model converged faster.

Lastly, we added a regularization term to minimize the sum of the bias parameters

$$L_{sum} = \frac{1}{R} \sum_{i=1}^{R} \hat{b}_i.$$

Ideally, the sum of the bias parameters is very close to zero, meaning that the unbiased model predictions are centered around the average of all annotators' thresholds.

We used binary cross entropy as loss function. With the penalty terms, the loss

therefore became:

$$L = L_{BCC} + \lambda_{range} L_{range} + \lambda_{bias} L_{bias} + \lambda_{sum} L_{sum}$$

4.3.7 Training process

We used a ResNet18 [257] as the classification model backbone for all three methods. The individual thresholds for all annotators were randomly initialized as described in Section 4.3.4. We generated labels for each annotator using different levels of random noise σ_n. The standard deviation of the random noise σ_n ranged from 0.025 to 0.1 in steps of 0.005. In the following, we specify the strength of the random noise by the ratio between σ_n and σ_r, where the latter is the standard deviation used to generate the biased binarization thresholds. For our experiments, we assumed that only one biased label is available for each sample during model development, However, we generated labels for each annotator for the validation data to be used during the evaluation phase.

The first step in the training of the two MBEM models was the same as the labels are not yet modified. The models were trained on the only available label for each sample. No label aggregation was required. The BiasNet with three output neurons, one for each annotator, was trained following the procedure outlined in Section 4.3.3. The hyperparameters for the regularization terms described in Section 4.3.6 are listed in Table 4.2.

Table 4.2: **Hyperparameters for the regularization terms**.

model	λ_{range}	λ_{bias}	λ_{sum}	k	m
MBEM	0.1	0	0	5	1
Bias Layer	0.1	0.05	0.1	5	1

To ensure that the parameters in the ResNet18 and the bias parameters of the BiasNet converged around the same epoch, we chose different learning rates for the main network ($lr_{net} = 1.0 \times 10^{-6}$) and the bias parameters ($lr_{bias} = 1.0 \times 10^{-4}$). Each model was trained for 400 epochs, with 2000 samples per epoch. We used the Adam optimizer [258] with a weight decay parameter of 0.01. No further data augmentation

was performed.

4.4 A generalized framework to predict continuous scores from discrete ordinal labels

4.4.1 Results

Predicted scores compared with severity rankings

Agreement between predicted score and severity rankings We first assessed how well the predicted continuous scores reflected a ranking of the images in each dataset. Retinal photographs and knee radiographs were ranked by domain experts with increasing disease severity. The mammograms were ranked with increasing density based on the quantitative continuously valued Volpara breast density score.

The relationship between the ground truth rankings and the predicted continuous scores are presented in Figure 4-6. For all datasets, the multi-class models without MC dropout displayed horizontal plateaus around the class boundaries where the predicted score was more or less constant with increasing rank. Similar patterns could be observed for the Siamese and MC Siamese models, especially for normal ROP and knee osteoarthritis cases.

Agreement between predicted and ground truth rankings A linear correlation between the predicted continuous score and the consensus rank cannot be assumed as the predicted score will increment variably depending on the severity increase from a patient of rank n and $n+1$. Therefore, we used the Spearman correlation coefficient and MSE to quantify the agreement between the ground truth ranking and the ranking based on the predictions (see Table 4.3 and Appendix A.2).

All MC dropout models were associated with a statistically significant higher Spearman correlation coefficient and lower MSE compared to their non-MC counterparts (p-value $< 2.2 \times 10^{-4}$, see Appendix A.3 for pair-wise statistical comparisons between the models). The higher Spearman correlation coefficients and lower MSE indicate

Table 4.3: **Model performance overview (mean ± 95% CI)**. Bold values indicate a statistical difference (p-value < 0.05) was observed. Spearman's rank correlation coefficient and the AUC are measured on the predicted continuous score while the MSE is measured between the normalized ground truth rank and the predicted rank generated from continuous scores. AUC was measured between normal and pre-plus vs. plus for ROP, none and doubtful vs. mild, moderate, severe for knee osteoarthritis, and fatty and scattered vs. heterogeneous and dense for breast density.

Model	**MSE ↓**	**Spearman ↑**	**clinically relevant AUC ↑**
ROP			
Multi-class	0.027 ± 0.009	0.84 ± 0.07	0.98 ± 0.02
MC multi-class	0.010 ± 0.003	0.94 ± 0.02	0.99 ± 0.01
Ordinal	0.015 ± 0.005	0.91 ± 0.04	0.99 ± 0.01
MC ordinal	0.009 ± 0.003	0.94 ± 0.02	0.99 ± 0.02
Regression	0.026 ± 0.010	0.85 ± 0.07	0.98 ± 0.03
MC regression	0.017 ± 0.005	0.89 ± 0.05	0.98 ± 0.02
Siamese	0.020 ± 0.007	0.88 ± 0.05	0.99 ± 0.01
MC Siamese	0.013 ± 0.004	0.92 ± 0.03	0.98 ± 0.02
Knee osteoarthritis			
Multi-class	0.023 ± 0.008	0.86 ± 0.06	0.97 ± 0.02
MC multi-class	0.019 ± 0.006	0.89 ± 0.05	0.99 ± 0.01
Ordinal	0.024 ± 0.007	0.85 ± 0.07	0.98 ± 0.02
MC ordinal	0.022 ± 0.007	0.86 ± 0.06	0.99 ± 0.02
Regression	0.023 ± 0.009	0.86 ± 0.06	0.99 ± 0.01
MC regression	0.019 ± 0.006	0.88 ± 0.05	0.98 ± 0.02
Siamese	0.022 ± 0.007	0.87 ± 0.05	0.97 ± 0.03
MC Siamese	0.020 ± 0.005	0.88 ± 0.04	0.97 ± 0.03
Breast density			
Multi-class	0.018 ± 0.001	0.89 ± 0.01	0.93 ± 0.01
MC multi-class	0.016 ± 0.001	0.90 ± 0.01	0.94 ± 0.01
Ordinal	0.016 ± 0.001	0.90 ± 0.01	0.93 ± 0.01
MC ordinal	0.015 ± 0.001	0.91 ± 0.01	0.94 ± 0.01
Regression	0.015 ± 0.001	0.91 ± 0.01	0.94 ± 0.01
MC regression	0.011 ± 0.001	0.93 ± 0.01	0.94 ± 0.01
Siamese	0.013 ± 0.001	0.92 ± 0.01	0.91 ± 0.01
MC Siamese	0.012 ± 0.001	0.93 ± 0.01	0.92 ± 0.01

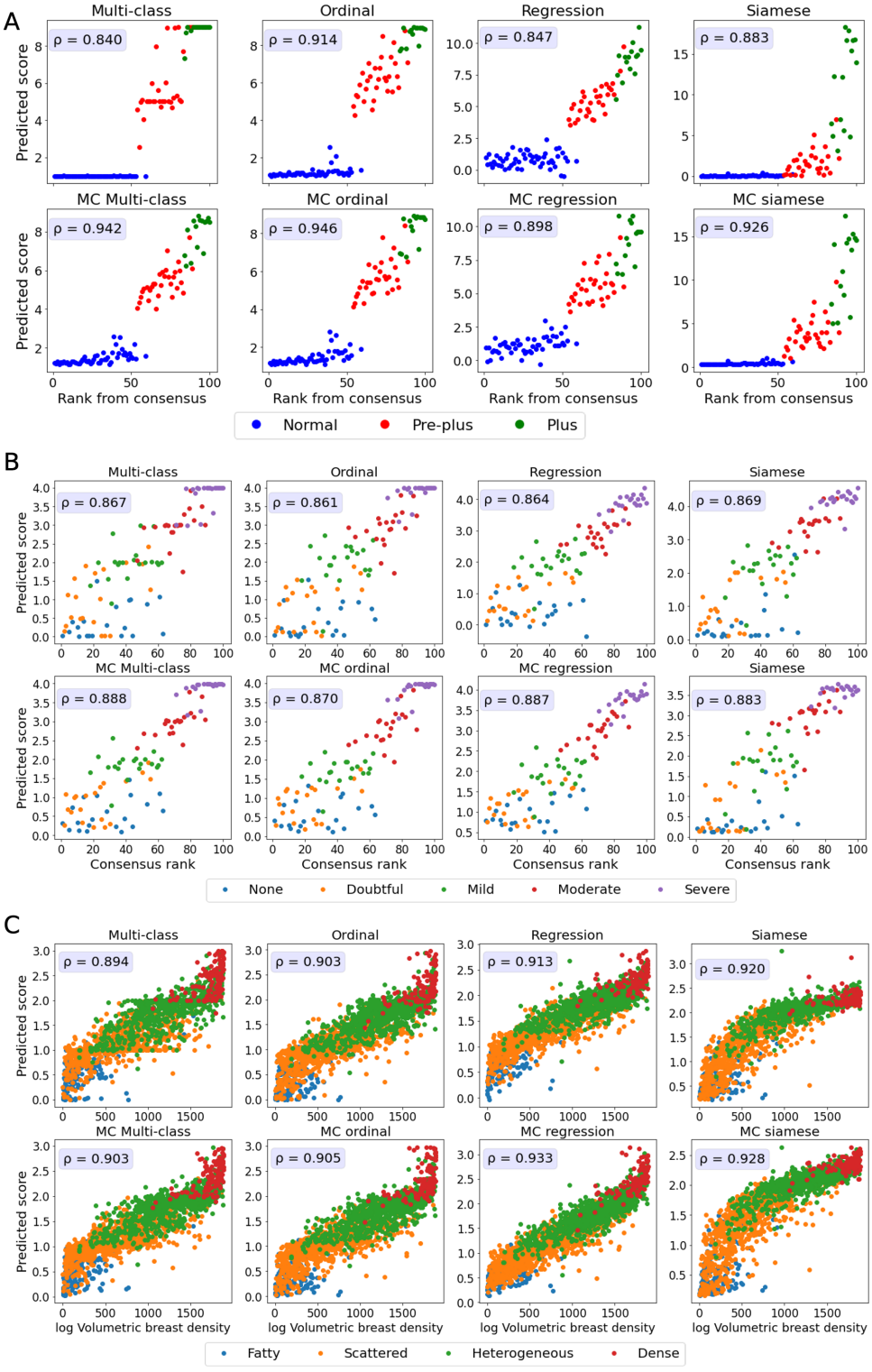

Figure 4-6: **Correspondence between predicted scores and severity ranking.** A: Retinopathy of prematurity; B: Knee osteoarthritis; C: Breast density. For each model, the Spearman correlation coefficient (ρ) is displayed in the upper left corner.

that the addition of MC dropout during training and inference improves the ability of DL models to correctly rank the images based on the continuous predictions. The models with the best correspondence between actual and predicted rank were MC multi-class and MC ordinal models for ROP, MC multi-class and MC regression for knee osteoarthritis, and MC regression and MC Siamese networks for breast density.

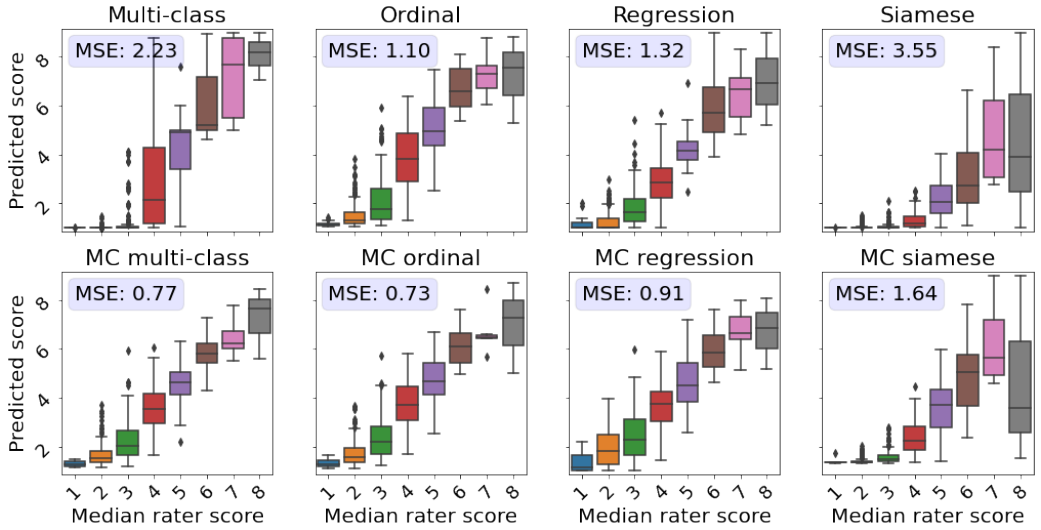

Figure 4-7: **Correspondence between predicted and consensus ROP severity on an in-distribution test set.** The consensus ROP score was obtained by calculating the median of five ratings from different experts. The predicted scores from multi-class, ordinal, and regression models that were trained to predict values from 0 to 2 were scaled and shifted to match the 1 to 9 range ($\text{score}_{rescaled} = \text{score}_{model} \times 2 + 1$). Siamese networks predict values from 0 to infinity and are not fully bounded. The Siamese scores were hence only shifted by 1 ($\text{score}_{rescaled} = \text{score}_{Siamese} + 1$). In accordance with the severity scale used, Siamese rescaled scores were also clipped to values between 1 and 9. All MSE measurements reported in this figure are statistically different (p-value $= 1.2 \times 10^{-41}$).

Classification performance All MC dropout models showed a slightly higher or comparable classification performance, as assessed by AUC, to their non-MC equivalent (see Table 4.3). The only exceptions were the Siamese models for ROP and knee osteoarthritis and the regression knee osteoarthritis model. In these three cases, though statistically significant, the AUC of the MC models was only 0.01 (or less) lower the one of their non-MC equivalents. The model associated with the best classification did not necessarily correspond to the best continuous severity scores.

The following models have the overall best performance for each dataset: MC multi-class (knee osteoarthritis), MC ordinal (ROP), and MC regression (breast density).

Comparison of predicted ROP scores with disease severity ratings

Next, we evaluated the correspondence between the predicted scores and more detailed severity ratings generated by domain experts. An subset of the test dataset was rated by five experts on a scale from 1 to 9 instead of the standard scale from 1 to 3 [229]. This dataset allowed us to evaluate the quality of the continuous model outputs on a more granular scale than the 3-class labels the models were trained on. Perfect continuous predictions would result in increasing disease severity scores with increasing ground severity ratings.

All MC models showed a higher correspondence between the true severity ratings and predicted scores, as reflected by a lower MSE in comparison with their conventional counterparts (see Figure 4-7). The models predicting the experts' ratings the most accurately are the MC multi-class and MC ordinal models. While Siamese networks showed decent correspondence between the predicted score and the ranked severity, a direct comparison with the severity ratings reveals that the predictions from these models are not well calibrated.

The multi-class model without MC showed the second worst performance in this analysis. Images rated from 1 to 3 by experts mainly obtained scores near 0, which does not highlight the severity differences as perceived by human experts. Furthermore, for retinal photographs associated with a score of 4, the model predicted values on the entire spectrum, i.e., from 1 to 9, which is undesirable.

Detection of temporal changes in disease severity Another important characteristic of a reliable severity score is its ability to reflect slight changes in disease severity over time. The disease evolution was quantified as the difference in the ground truth severity ratings or predicted severity scores between photographs of the same patient taken at different time points. We then compared the difference in experts'

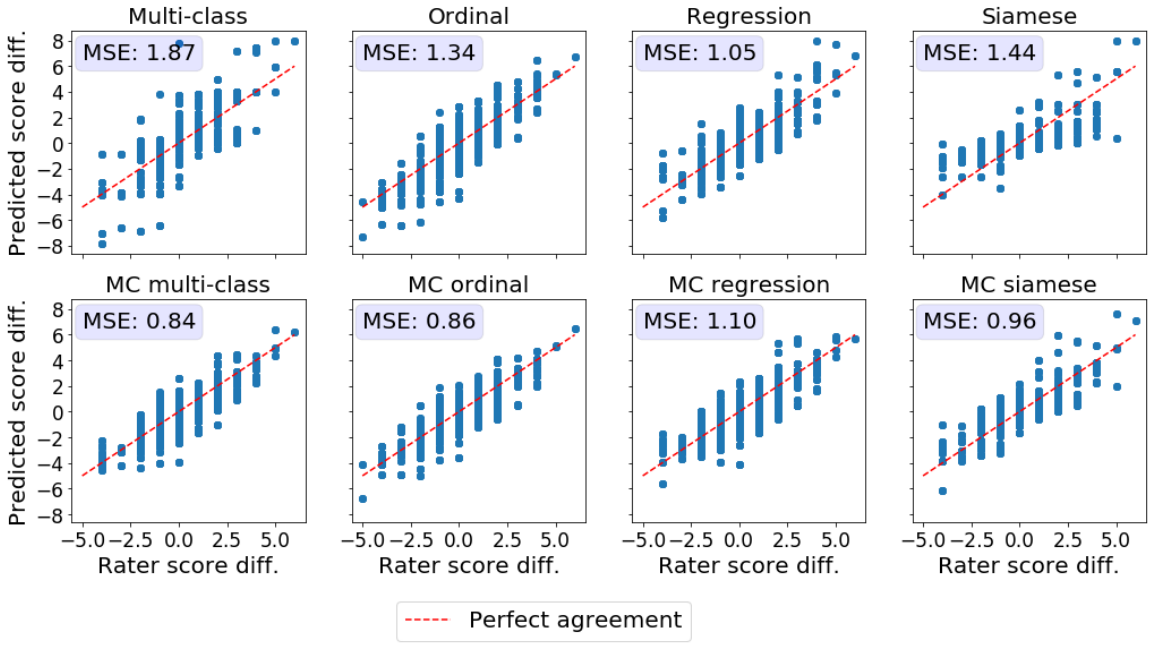

Figure 4-8: **Correspondence between the perceived rater score difference on longitudinal images from the same ROP patient and the predicted score difference.** The red dashed line is the identity line and indicates the expected region were the data points should fall. All MSE measurements reported in this figure are statistically different(p-value $= 4.3 \times 10^{-28}$).

ratings to the difference in the predicted scores using MSE (see Figure 4-8). Ideally, the difference in the expert's scores should be equal to the difference in the models' predictions. MC dropout improved the correspondence between the disease evolution as perceived by experts and predicted by the DL models for multi-class, ordinal, and Siamese models. Prediction differences from MC multi-class and MC ordinal models matched the severity shifts in the experts' ratings most closely. The conventional multi-class model showed multiple outliers and was associated with the highest MSE.

Continuous scores on an out-of-distribution dataset The ROP models were further tested on a dataset from a different population and acquired at different centers from the training dataset. Figure 4-9 illustrates the correspondence between the predicted and rater scores for this out-of-distribution dataset. Similar to the in-distribution test set (see Figure 4-7), all the MC models had a better MSE compared with the non-MC corresponding models and MC multi-class and MC ordinal were

the best performing models. The multi-class and regression models for most severity scores predicted a wide range of values, often from 1 to 9 which could lead to medical errors. The miscalibration of Siamese models is especially noticeable in Figure 4-9 as visually, the predicted and rater score do not match for high severity values. This out-of-distribution dataset contained only a few plus and pre-plus images, i.e., only 328 plus and pre-plus cases compared to 7565 normal cases. Driven by a large number of outliers, particularly within images with the lower disease severity ratings (normal cases), the MSE was particularly high for the conventional multi-class, ordinal classification, and regression models. The low number of images with higher disease severity scores also explains why the MSE is not extremely high for the Siamese networks, even though they are visually miscalibrated.

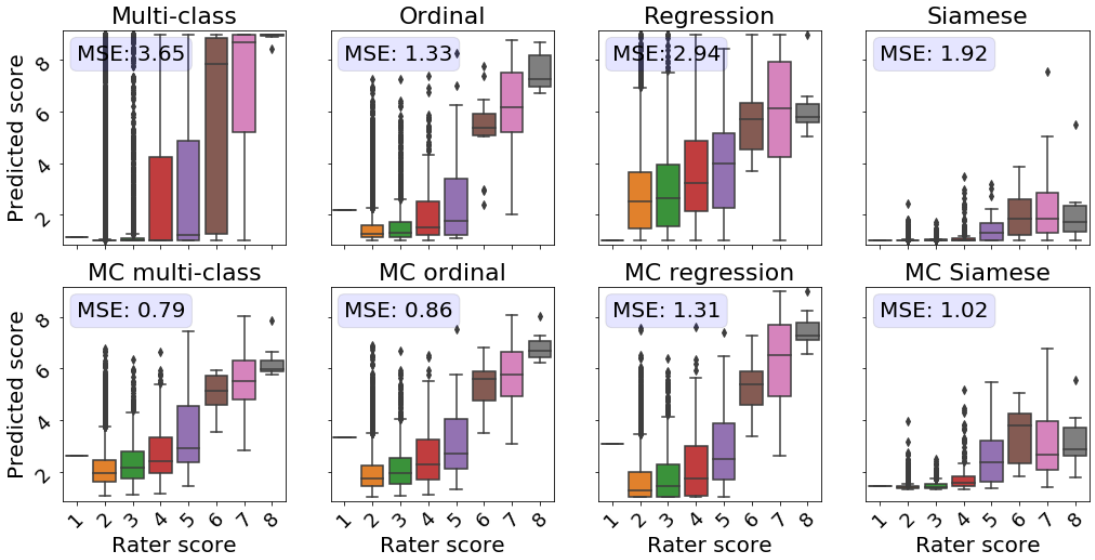

Figure 4-9: **Correspondence between predicted and consensus ROP severity on an out-of-distribution test set.** The rater score is obtained from a single rater. The predicted scores from multi-class, ordinal, and regression models that were trained to predict values from 0 to 2 were scaled and shifted to match the 1 to 9 range ($\text{score}_{rescaled} = \text{score}_{model} \times 2 + 1$). Siamese networks predict values from 0 to infinity and is not fully bounded. The Siamese scores were hence only shifted by 1 ($\text{score}_{rescaled} = \text{score}_{Siamese} + 1$). All MSE measurements reported in this figure are statistically different (p-values < 0.05).

Comparing predicted breast density scores with continuously valued ground truth breast density measurements

Lastly, we evaluated the ability of the breast density prediction models trained to accurately reflect the continuously valued Volpara Density measurements. The subset of mammograms with the Volpara Density measurements provided us with the unique opportunity to evaluate algorithms trained using ordinal labels on a continuously valued ground truth. Therefore, unlike with the ranked score analysis presented in Section 4.4.1, here we directly compared the Volpara Density scores with the continuously valued model predictions. Ideal continuously valued predictions would correlate linearly with the Volpara Density measurements. We first assessed the relationship between the Volpara Density scores and the discrete ground truth labels generated by domain experts used for training. As illustrated in the boxplot in Figure 4-10A, and by the Spearman correlation coefficient of 0.73, there was a high agreement between the ground truth labels and the Volpara Density scores. The MC multiclass model's predictions, both class and continuous score, were a close proxy to the volumetric breast density measurements, as seen in Figure 4-10B and C with a Spearman correlation coefficient of 0.803 (classification) and Pearson correlation coefficient of 0.91 (continuous scores). The high correlation between the continuous breast density predictions and Volpara Density measurements indicates that our model can generate an accurate continuous prediction while being trained on only a finite number of classes.

4.4.2 Discussion

The underlying continuous nature of many prediction targets for DL image analysis tasks, such as breast density and disease severity, has to be taken into account in the process of model design. Here, we studied the capability of DL models to intrinsically learn a continuous score while being trained using discrete ordinal labels. Our results show that training a conventional multi-class classification model without MC dropout did not lead to predictions that reflect the underlying continuous nature of the target

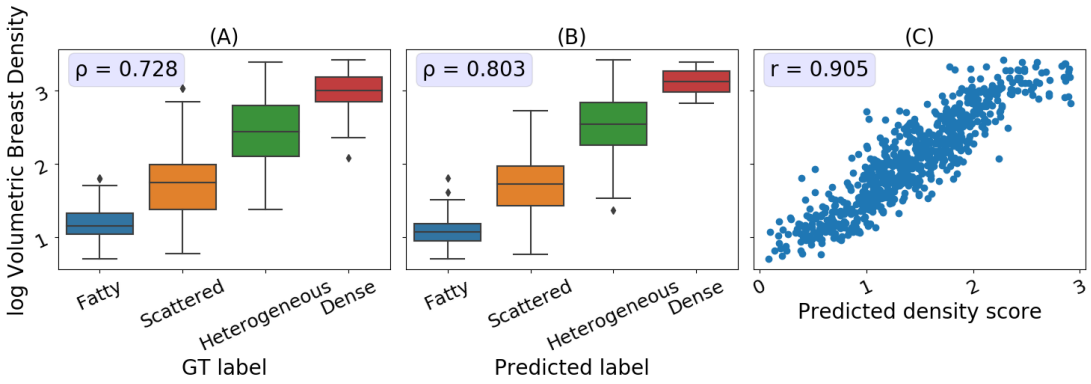

Figure 4-10: **Correspondence between volumetric breast density measurement and predicted or true breast density.** The predicted values are from the breast density MC multi-class model. A: The relationship between the expert ratings and the quantitative Volpara measurement; B: Correspondence between the predicted labels obtained by taking the class with the highest softmax score and volumetric breast density; C: Relationship between the continuous predicted score and the volumetric breast density. ρ is the Spearman correlation coefficient for each metric pair.

variable. Approaches that model the relationship between the ordinal labels, such as ordinal classification, regression, and Siamese networks, provide continuous predictions that closely capture the continuity of the target variable even without the use of MC dropout. Finally, using MC dropout during training and inference increased the ability of the DL models to predict meaningful continuous scores. MC dropout multi-class classification ranked among the best performing models in this study.

Multi-class classification models Ignoring the ordinal relationship between the training label classes caused conventional multi-class prediction models to return predictions that are clustered around the values of the training labels. This behavior is reflected in the plateaus visible in Figure 4-6, and the medians in Figures 4-7, and 4-9, a lower Spearman correlation coefficient and higher MSE.

Due to the definition of the training objective, multi-class classification models are optimized to precisely predict a specific class and discouraged from predicting scores at the class boundaries. This behavior is desirable for nominal classification, where the classes should be separated as clearly as possible with minimal overlap in the feature latent space to avoid ambiguous predictions. However, the approach is

not appropriate for problems with a target variable with an underlying continuous nature and explains the limited performance of the multi-class classification models to predict meaningful continuous scores.

Siamese networks Siamese networks showed decent correspondence between the ranked severity and the predicted score (Figure 4-6). However, a direct comparison between the predicted score and the severity determined by domain experts (Figure 4-7), revealed that the predictions were not well calibrated. The predictions did not accurately reflect disease severity on a more granular scale than the labels used for model training. Siamese networks are not trained to predict a specific value, unlike the other models, but rather to detect whether two images stem from the same or different classes [259]. Therefore, they can pick up subtle differences in disease severity [224]. Here, we obtained predictions comparing the input image of interest to a pool of anchor images that are typical representations of the class corresponding to the lowest label score. While the predicted difference between the anchor images and the target images resulted in accurate ordinal predictions (Figure 4-6), it was not well calibrated to the underlying continuous variable, particularly at the extremes.

MC dropout improves prediction of continuous variables Through the use of MC dropout, all four model types evaluated showed an improvement in the quality of the continuous scores as reflected in significantly higher Spearman correlation coefficients and lower MSE (see Table 4.3). MC multi-class classification networks were consistently among the highest performing models for all tasks and datasets evaluated, making them the top-performing models in our study.

MC dropout presents a simple way to obtain meaningful continuous predictions from models trained using ordinal labels. In our experiments, we did not have to sacrifice classification performance for improved quality of the continuous predictions and in some cases, even significantly improved predictive performance (see Table 4.3). However, MC dropout comes at a higher computational cost as inference requires multiple passes of the same input image to obtain the final prediction. If the additional

computational burden is a concern, ordinal classification or regression are alternatives to conventional multi-class classification models that are easy to train and provide decent continuous predictions without the use of MC dropout.

4.4.3 Limitations

There are some limitations to this study. First, due to the latent nature of the variable of interest, for most of our analysis, we had to rely on proxy variables such as rankings and more granular expert disease severity ratings. Second, MC dropout predictions were based on 50 samples, an empirically chosen value based on common practices and our own experience. Lastly, we treated the available ordinal labels as ground truth. For all three image analysis tasks analyzed here, high inter-rater variability, particularly around the decision boundaries between severity classes, have been reported [41, 47, 260, 261]. In Section 4.5, we explore the influence of noisy and biased ordinal ratings for the task of learning and predicting a continuous variable and how annotator-specific biases can be learned with only one label per sample available.

4.4.4 Conclusions

In this section, we presented a generalizable framework to predict meaningful continuous scores while only using discrete ordinal labels for model development. Our findings are particularly relevant for disease severity prediction tasks as the available labels are usually coarse and ordinal, but continuous disease severity predictions could provide crucial information that allows for earlier detection of deterioration and more personalized treatment planning.

Code and data availability

The code used to train the models can be found at `https://github.com/andreanne-lemay/gray_zone_assessment`. Access to the MOST dataset for knee osteoarthritis can be requested through the NIA Aging Research Biobank `https://agingresearchbiobank.nia.nih.gov/`. The breast density, and ROP datasets are not publicly accessible due

to patient privacy restrictions.

4.5 Learning the bias of individual annotators from single ordinal labels

4.5.1 Results

We compared three methods: conventional MBEM (MBEM_{class}) as described by Khetan et al., our proposed extension to MBEM for continuously valued latent variables MBEM_{cts} and a BiasNet model with our proposed trainable biased sigmoid layer.

We randomly initialized biased thresholds for experimental validation by sampling from a Gaussian distribution with a mean of 0.25 and a standard deviation σ_r of 0.05. The thresholds and biases (difference between the individual and average threshold) for three annotators were set to the following values:

- Annotator 0: 0.281 (threshold) and 0.058 (bias)
- Annotator 1: 0.211 (threshold) and -0.012 (bias)
- Annotator 2: 0.177 (threshold) and -0.046 (bias)

Based on the annotator thresholds, the average threshold representing the optimal binarization threshold for a model is 0.223.

To obtain the initial training labels for MBEM, we used one label for each sample without any further label aggregation. We varied the standard deviation of the random noise component σ_n between $0.5 * \sigma_r$ and $2 * \sigma_r$. The relationship between the random additive noise, the bias levels of the simulated annotators, and the resulting frequency of misclassifications by binarizing the noisy ground truth are shown in Figure 4-11.

At each noise level, we trained five models using different random seeds. We evaluated and compared the three methods according to the following four evaluation categories:

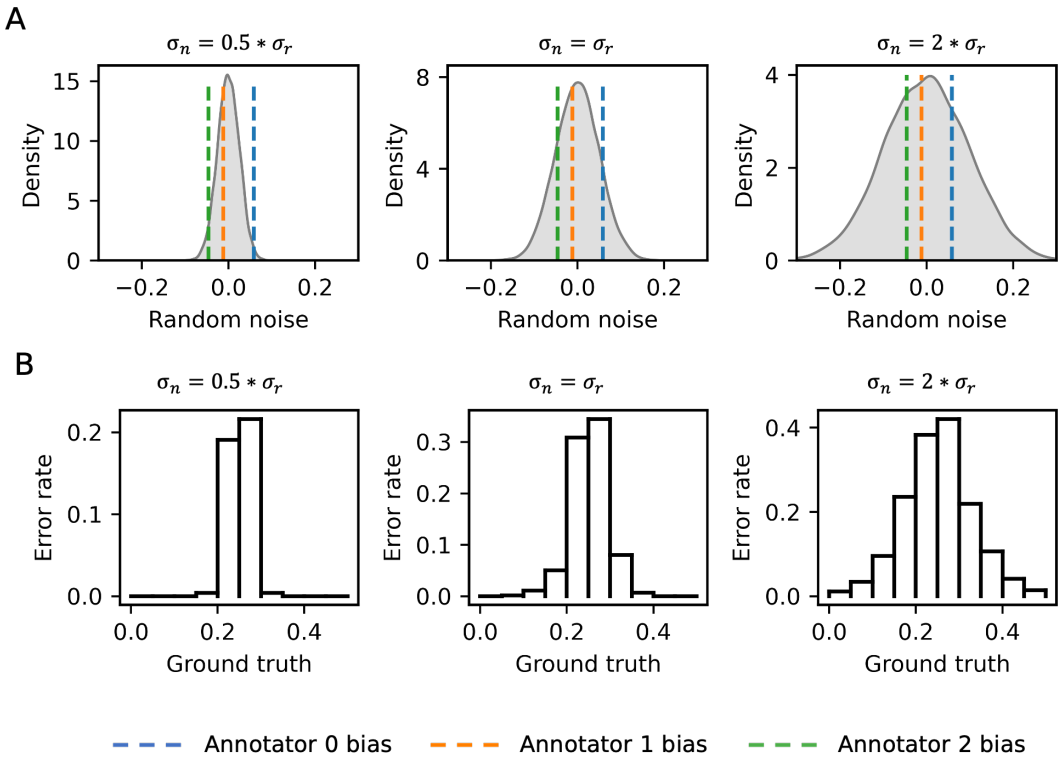

Figure 4-11: **Influence of random noise on label error rate.** A: Distribution of the random noise component added to the continuous ground truth before binarization. To illustrate the relationship between the distribution of the random noise and annotators' biases, the biases are indicated by dashed lines. B: Effect of increasing random noise on the labels after binarization, depending on the uncorrupted ground truth value. With increasing noise, samples with a continuous ground truth value further away from the binarization threshold (here: 0.25) are affected by errors in the binary labels.

1. The label quality improvement through the label update step for the MBEM methods

2. The improvement in the classification performance that can be achieved by taking the annotator-dependent characteristics into account

3. The quality of the inferred individual thresholds of all annotators

4. The quality of the continuously valued model predictions

MBEM – Label quality improvement

To achieve higher performance, the MBEM framework contains a label update step that explicitly uses the inferred annotator characteristics. During this step, the training labels are corrected based on the learned annotator characteristics—confusion matrices in the case of MBEM$_{class}$ and individual thresholds in the case of MBEM$_{cts}$. Subsequently, the updated labels are used to train a new model and the annotator characteristics are re-calculated. The process can be repeated until the difference between the updated labels and the training labels from the previous round is below a pre-specified threshold or no performance improvement can be obtained by training a model with the most recent updated labels. Since we have access to the non-noisy ground truth in our experimental setting, we can compute the quality improvement of the label update step.

Following the process of MBEM$_{class}$, no label was updated based on the learned confusion matrices in this scenario. To result in a label update in a binary classification setting, one of the diagonal entries in the annotator's confusion matrices must be lower than 0.5. Even though the bias of annotator 0 can be considered substantial, the overall error rate was low. In annotator 0's confusion matrix, the diagonal entries were larger than the off-diagonal entries. Because no training labels were updated, the MBEM$_{class}$ process was considered complete after the first round.

MBEM$_{cts}$, our proposed extension of MBEM for continuously valued variables, results in a small improvement of label quality after the first round of training as depicted in Figure 4-12. We measure the quality of the labels by computing Cohen's kappa between the noise-corrupted training labels and the uncorrupted ground truth. To obtain a sense of the improvement obtained by the label update step, we compare the label quality of the initial training labels with the quality of the training labels after the label update step. We found a small improvement in label quality. This effect shrinks with increasing noise. However, using the updated training labels for training did not improve the models' classification performance. Therefore, the process was aborted after the first round for both MBEM$_{class}$ and MBEM$_{cts}$

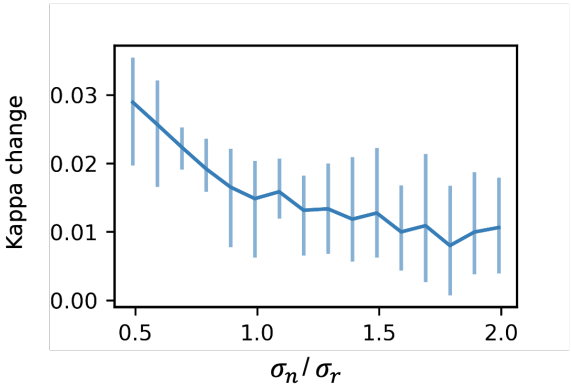

Figure 4-12: **Quality improvement through MBEM$_{cts}$ label update.** Difference in Cohen's kappa of the training labels before and after the label update step for varying levels of random label noise. Cohen's kappa is computed based on the uncorrupted ground truth. The solid line indicates the mean change, and the error bars indicate the highest and lowest values of five independently trained models.

4.5.2 Classification performance

Next, we tested whether the agreement between the model predictions and each annotator's labels could be improved by using the information that each model learns about the individual annotators. Classification performance is measured using Cohen's kappa between the predictions and the binarized uncorrupted ground truth labels. We compare the kappa scores of the unbiased and biased predictions to quantify the performance improvement that can be achieved by using individual characteristics. We expect that the classification predictions generated using individual binarization thresholds have a higher agreement with each annotator's biased labels. This step is performed separately for the labels from each annotator to capture differences between the annotators.

MBEM$_{class}$ Due to the limited information that can be inferred about each annotator from MBEM$_{class}$, we were not able to use biased thresholds to generate individual predictions. Therefore, we used the annotator confusion matrices to correct predictions, following the inverse procedure used to correct training labels. However, this step did not result in any updates in the predictions, for the same reason the MBEM$_{class}$ label update step led to no changes in the training labels (see Section 4.5.1).

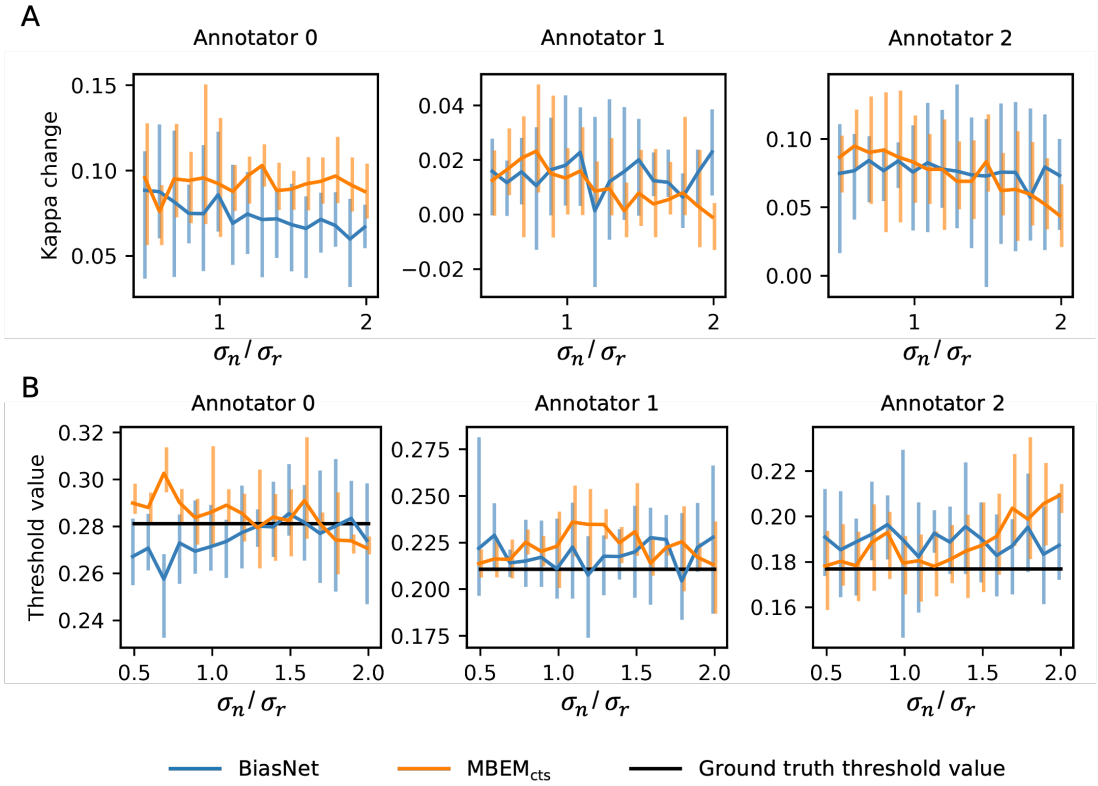

Figure 4-13: **Classification improvement and learned biases for MBEM$_{class}$ and BiasNet.** A: Change in Cohen's kappa by using individual binarization thresholds for each annotator instead of a generic one; B: Learned individual binarization threshold for each annotator. The solid lines indicate the average performance, and the error bars are the maximum and minimum values from five independent experiments at each noise level. Results from the BiasNet experiments are presented in blue, and results for the MBEM$_{cts}$ are presented in orange. The ground truth of each annotator's binarization threshold is depicted in black.

MBEM$_{cts}$ and BiasNet As shown in Figure 4-13, the classification performance could be improved for the MBEM$_{cts}$ and BiasNet. Due to the different sizes of the biases, the improvement is different for each annotator. The largest improvement in classification performance could be achieved for the annotator with the largest bias (annotator 0) and the smallest, if any, for the annotator with the smallest bias (annotator 1).

Learning of the annotator-specific characteristics

MBEM$_{class}$ While both MBEM$_{cts}$ and the BiasNet allowed learning each annotator's individual bias or binarization threshold, we were unable to infer these parameters from the confusion matrices generated by MBEM$_{class}$. Therefore, we used each annotator's estimated sensitivity and specificity as proxies for their bias. Based on the sensitivity and specificity, we can determine whether an annotator is an over- or under-caller and order the annotators according to their binarization thresholds. We describe our approach to obtain the order of annotator thresholds in Algorithm 1.

Algorithm 1: Determine order of annotator bias from confusion matrices

```
Data: Annotator confusion matrices
Result: Annotators sorted from lowest to highest biases
for annotator r = 0 to R do
    compute annotator sensitivity and specificity based on their estimated
     confusion matrix;
    if sensitivity > specificity then
        label annotator r as over-caller;
    else if sensitivity < specificity then
        label annotator r as under-caller;
    else
        label annotator r as average caller;
Sort over-caller from lowest to highest specificity (highest to lowest negative
 bias);
Sort under-caller from highest to lowest sensitivity (lowest to highest positive
 bias);
The order of all annotators is then given by: sorted over-callers, average-caller,
 sorted under-callers
```

Each annotator was first categorized as an over- or under-caller, or an average caller.

An over-caller is an annotator with a sensitivity higher than their specificity. Their bias would be negatively valued since their binarization threshold is lower than the average threshold. Every annotator with a specificity higher than their sensitivity was categorized as an under-caller. Their bias is positively valued, and their binarization threshold is higher than the average threshold. If an annotator had the same sensitivity as specificity (symmetric error), they were categorized as an average caller.

Even after just one round of MBEM$_{class}$, we could classify each of the modeled annotators as over- or under-caller with 100% accuracy (five models per noise setting) and the rankings of the individual thresholds reflected the ground truth ranking perfectly with a Spearman correlation coefficient of 1.0 for all noise settings.

MBEM$_{cts}$ and BiasNet To measure how well a method can learn the annotators' individual biases, we converted the learned bias (BiasNet) or threshold (MBEM$_{cts}$) into the same space as the ground truth threshold. The learned thresholds for both methods are illustrated in the plots in the bottom row of Figure 4-13. Both the MBEM and the BiasNet learned the biases of the modeled annotators' as expected. Surprisingly, the ability of the methods to learn each annotator's threshold was not affected by the amount of random noise that corrupted the training labels. Only the estimation for the threshold of annotator 2 showed an increasing error with more extensive noise corruption.

Continuously valued predictions

The primary underlying assumption for this work was that the variable of interest is continuously distributed. Therefore, we evaluated how well the unbiased pre-sigmoid output $\hat{s}$ of the different methods reflect the continuous value of the target variable s. As we were using synthetically generated data, we had access to s, which is rarely the case for natural data measured using discrete ordinal labels.

The agreement between $\hat{s}$ and the uncorrupted ground truth was measured using the mean squared error (MSE) and Pearson correlation coefficient. A higher agreement between the estimated and the ground truth continuous variable would result in a

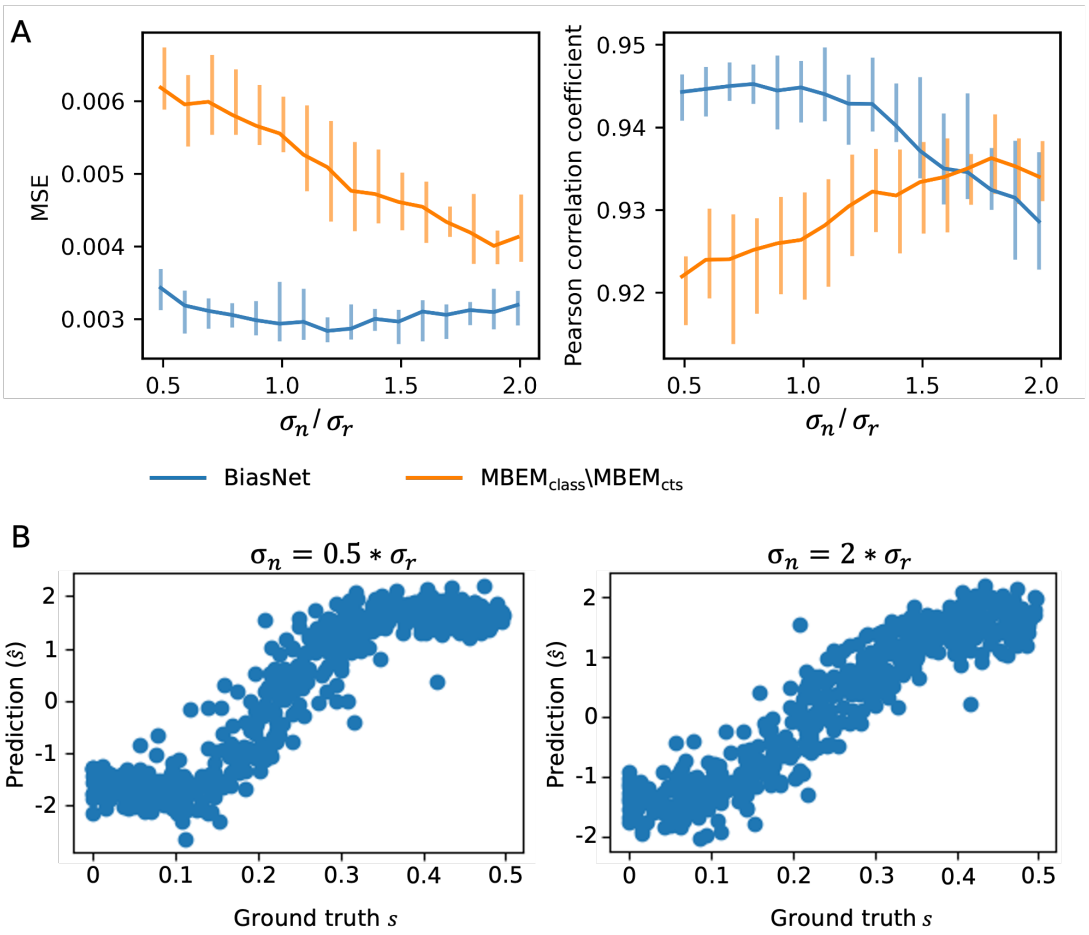

Figure 4-14: **Agreement between continuously valued predictions and ground truths.** A: Mean squared error (MSE) and Pearson correlation coefficient between the pre-sigmoid predictions $\hat{s}$ and uncorrupted ground truth s. The solid lines indicate the average performance, and the error bars are the maximum and minimum values from five independent experiments at each noise level. Results from the BiasNet experiments are presented in blue, and results for the MBEM$_{class/cts}$ are shown in orange. B: Relationship between $\hat{s}$ and s for selected MBEM$_{class/cts}$ models at the minimum and maximum noise level.

lower MSE and higher Pearson correlation coefficient. Since $\hat{s}$ and s have different ranges, we re-scaled $\hat{s}$ to the range of the ground truth $[0, 0.5]$. Furthermore, as we only performed one round of training for MBEM$_{cts}$ and MBEM$_{class}$, the classification models were equivalent. At this stage, the only difference between MBEM$_{cts}$ and MBEM$_{class}$ was the generation of "biased" predictions.

The MSE and Pearson correlation coefficients for both the BiasNet and MBEM$_{class}$/ MBEM$_{cts}$ are illustrated in Figure 4-14A. At low noise settings, the continuous, pre-

sigmoid outputs from the BiasNet appeared to model s better than the MBEM$_{cts}$ and MBEM$_{class}$, which was reflected in a low MSE and high Pearson correlation coefficient. However, while the MSE for the BiasNet models remained stable for higher levels of random noise in the binary training labels, the Pearson correlation coefficient increased, indicating a lower linear correlation between $\hat{s}$ and the continuous ground truth. At the same time, the MSE and Pearson correlation coefficient from the MBEM$_{cts}$ and MBEM$_{class}$ models were improving with increasing label noise. The plots in Figure 4-14B depict the relationship between $\hat{s}$ and the continuous ground truth s for representative models at the minimum and maximum noise levels. From these plots, it can be appreciated that with increasing noise, the s-shaped agreement between $\hat{s}$ and s is turning into a more linear one.

4.5.3 Discussion

In this section, we described how sub-class level differences between annotators could be detected and quantified. Additionally, the learned information about individual annotators could be used to improve the classification performance of DL prediction models. Our work applies to all ordinal classification settings in which the underlying variable of interest, albeit not directly measurable, is continuous. In these cases, the training labels can be affected by individually varying binarization thresholds.

We presented the first description of two methods to infer individual differences between annotators for ordinal variables that work with as little as one label per sample. Similarly to the framework outlined in Section 4.4, these methods only require access to discrete ordinal labels but enable researchers to accurately infer features on a more granular scale.

MBEM$_{class}$ is not appropriate for continuously valued variables

While we could learn the order of the annotators' binarization thresholds and accurately classify them as over- or under-callers using MBEM$_{class}$, our experiments clearly showed the limitations of the method. MBEM$_{class}$ was developed for nominal classification.

Therefore, it does not support the assumption of an ordinal relationship between the labels. Our experimental results, using a synthetic dataset with a latent continuous variable and ordinal labels, highlighted the importance of identifying samples close to the decision boundary. These findings illustrates that the assumptions of nominal classifications are insufficient for the present problem.

Furthermore, MBEM$_{class}$ was developed using a hammer-spammer model to describe the quality of each annotator's labels. In this model, each annotator is either a hammer or a spammer. A hammer generates almost perfect labels with a very low random error rate. A spammer, on the other hand, is an unreliable annotator who randomly assigns labels to each sample. While it has been shown that MBEM$_{class}$ can learn systematic errors beyond simply categorizing each annotator as a hammer or spammer, i.e., cases in which an annotator systematically confuses two specific classes, and correct them [239], our results indicate that the assumption of nominal classification is not appropriate for ordinal classification with systematic errors that occur at the class boundaries.

To infer more detailed information about the annotator's binarization thresholds using MBEM$_{class}$, additional assumptions about the distribution of the underlying continuous variable would need to be made, e.g., assume it is uniformly distributed. However, since this information is not routinely available for ordinal data, we decided against this step. Therefore, the originally proposed MBEM algorithm is inappropriate for settings with an underlying continuous variable that is measured in ordered (binary) categories. Consequently, we developed an extension of MBEM$_{class}$ for continuously valued variables, MBEM$_{cts}$.

New methods are required to learn annotator's binarization thresholds

MBEM$_{cts}$ Our proposed extension of MBEM for continuously valued variables, MBEM$_{cts}$, utilizes the continuously valued, pre-sigmoid output $\hat{s}$ of a binary classification model. In our experiments, this output had a strong linear relationship with the continuously valued ground truth, which is reflected in Pearson correlation coefficients above 0.9. Ignoring this strong relationship and using the binarized model predictions

lead to a loss of information in MBEM$_{class}$. In MBEM$_{cts}$, however, the strong linear relationship is utilized to identify accurate individual binarization thresholds for each annotator. While the agreement between $\hat{s}$ and the continuous ground truth improved with increasing random noise in the training data (see Figure 4-14), the quality of the learned thresholds and, therefore, the improvement that can be achieved by using these thresholds did not (see Figure 4-13). As shown in Figure 4-14B, the improvement for high noise-settings, is mostly due to an improvement in the linearity of the relationship between s and $\hat{s}$ at the extremes of the prediction range. Since we did not observe changes around the value range of the binarization thresholds, the quality of the learned thresholds was unaffected.

The improvement in the relationship could be a consequence of two effects: First, we limited the range of the outputs by penalizing outputs with a high value, using the regularization term introduced in Section 4.3.6. Second, with increasing random noise, the range of values of the continuous variable affected by noise in the binarized labels was increasing, as illustrated in Figure 4-11. The increasing noise in the binary training labels may have served as an additional regularization of the model training that prevented the model from becoming over-confident in its predictions, particularly at the extremes.

BiasNet Both the MBEM extension for continuously valued variables and the BiasNet worked well for learning the individual thresholds. The BiasNet picked up the relationship between the unbiased continuous predictions $\hat{s}$ and the continuously valued ground truth at lower noise levels. For the present application, only one round of training was required for MBEM$_{cts}$ to accurately learn the individual binarization thresholds. However, variables that are more challenging to learn and affected by more complicated noise patterns than those presented here may require several training rounds, rendering the MBEM methods unwieldy in practice. The BiasNet training process, however, consists of one, not several iterative training steps. Additionally, as the BiasNet does not explicitly rely on the prediction performance on the training dataset, it may be less sensitive to overfitting on the training data. However, it requires

optimization of the regularization weights for the bias parameters.

Applications and future directions

Viewing differences between the annotators as meaningful information that can be exploited during model development and deployment has several promising applications, particularly in disease severity prediction. As described in Section 2.4, the labels for most medical imaging datasets are frequently generated by collecting several annotations for each sample. These labels are subsequently aggregated using either an algorithmic tool [262] or following a label adjudication process, potentially involving several data sources [128]. Since the cost of obtaining annotations from highly specialized annotators is high, this elaborate process can be prohibitive to assembling a dataset with high-quality annotations. Furthermore, label aggregation processes have been shown to introduce representational biases [263]. Consequently, modeling single readers by learning their labeling characteristics is an exciting avenue for training deep learning models on single labeled "noisy" data. In this section, we demonstrated that individual annotators' biased binarization thresholds can be learned based on only one label per sample without requiring repeated annotations. Indeed, a costly adjudication process may not be necessary for generating discrete ordinal labels in this scenario, and the noise in the data can be exploited during training.

Human annotators are often not aware of their own biases in generating discrete ordinal labels. Therefore, providing physicians feedback on their labeling practices may be beneficial for their training and the deployment of algorithms in decision support. For example, a case slightly below the decision boundary between normal and abnormal may be labeled as abnormal by a reader considered an over-caller. An automatic classification system providing a second read, in addition to a classification label of normal, could offer individualized feedback to the reader. It could inform them that based on their personal history, they would likely call this case abnormal. This would allow the reader to rethink their decision and lead to informed, harmonized decisions.

Notably, both the MBEM$_{cts}$ and BiasNet methods can be fine-tuned to learn the

individual thresholds of new readers that are not represented in the training data, given enough annotated data, particularly around the decision boundary.

For this work, we used a simplified model in which the underlying continuous ground truth depends on only one variable (the roundness of the square). Still, most medical variables, like disease severity, are influenced by multiple pathophysiological factors. Different physicians may weigh single factors differently and therefore come to different conclusions in interpreting disease severity from imaging. Medical professionals who create annotations for datasets resemble black boxes as they are rarely asked to justify their labeling decisions. Incorporating explainability methods for disease severity prediction [264] may be an exciting avenue to explore in future research. Furthermore, our methods may have implications for federated learning due to variability in annotation practices at different institutions. Consequently, the influence of different data and class distributions for each annotator should also be investigated.

4.5.4 Limitations

There were some limitations to this work. First, the dataset noise was modeled as a simple two-component model consisting of a random noise component identical for all annotators and the annotator-specific biased binarization threshold. However, in practice, some annotators may be more reliable than others, resulting in an annotator-dependent random noise component. Additionally, the random noise component was entirely independent of the underlying data. In the clinical setting, low-quality images suffering from artifacts or the presence of co-morbidities can lead to diagnostic and labeling errors [114], adding a source of heteroscedastic, input-dependent label noise. Consequently, it would be desirable for future work to explore more complex noise models. Lastly, we worked with a binary classification problem for simplicity, but all three methods can be extended to (ordinal) multi-class settings with multiple thresholds between the classes.

4.5.5 Conclusions

In this section, we showed that using deep learning, sub-class level differences between annotators can be detected and quantified. This work is the first conception and demonstration of methods that enable the joint learning of annotators' ordinal classification and their individual biases for a latent, continuously valued target variable. The methods use individual annotators' ordinal labels for training and learn the individual biases with as little as one label per training sample. We also demonstrate that the classification performance of ordinal classes can be improved by learning the individual biases of each annotator.

4.6 Perspectives on working with continuously valued variables

In this chapter, we highlighted the advantages of shifting from a purely discrete ordinal picture of disease severity to a continuous one. In Section 4.4, we demonstrated that the gap between continuously distributed variables and their discrete ordinal labels can be closed using appropriate modeling strategies. Models like ordinal classification and regression generated continuously valued predictions that truthfully reflect the continuous ground truth. A similar effect could be achieved by calibrating nominal classification models using MC dropout. Our results illustrate that information lost through discretizing a continuous variable can be recovered utilizing appropriate models.

Based on the concept of a latent continuous variable, we also introduced two new methodologies to learn sub-class level biases that are manifest as inter-individual differences in binarization thresholds. We hypothesize that biases of individual annotators represent a substantial component of inter-annotator differences in generating labels for continuously distributed variables. Lastly, the repeatability of disease severity predictions can be improved by using models that respect the ordinal distribution and MC-based models (see Section 3.7.2). This effect is particularly strong for cases at

the boundaries between two classes.

While ordinal classification is a common task in medical image analysis, particularly given the importance of disease severity prediction, its importance is limited outside medicine. Consequently, datasets and methods for ordinal classification are limited, in particular well-curated datasets that offer detailed information beyond the severity class labels, e.g., annotator agreement or long-term outcomes. In order to validate the methods introduced in Section 4.5, access to the single labels that were used to generate adjudicated training labels would be of great importance.

Future directions To study the advantages of using continuous over discrete ordinal predictions, it would be desirable to investigate their usefulness for risk modeling and their robustness towards temporal changes in diagnostic categorization trends. Furthermore, we envision that the methods for the joint learning of disease severity prediction and individual annotator's biases will support efforts to develop a general noise model for the annotation process of ordinal variables. Lastly, disease severity is usually not determined by a single but by many factors. It would be valuable to investigate if disease severity can be decomposed into these factors and how they contribute to the final prediction.

Chapter 5

Clinically meaningful evaluation of brain tumor segmentations

Brain tumor segmentations are an integral part of the clinical management of patients with glioblastoma, the deadliest primary brain tumor in adults. The manual delineation of tumors is time-consuming and highly provider dependent. These two problems are expected to be addressed by introducing automated, deep-learning-based segmentation tools. In developing deep learning segmentation algorithms, performance is optimized and evaluated based on various quantitative metrics. However, the quality perception of expert raters shows a low concordance with known and widely-used metrics. And these quantitative metrics do not correlate with the clinical quality of a segmentation. A detailed understanding of experts' assessment of segmentation quality is required to enable a successful collaboration between physicians and segmentation algorithms. This chapter presents two studies contributing to a clinically meaningful evaluation of brain tumor segmentation. We first investigate the relationship between the commonly-used quantitative metrics and experts' ratings of segmentation quality. Additionally, we report findings from a study on the role of contextual information in the perception of the quality of brain tumor segmentation

The work presented in this chapter is adapted from:

- Katharina Hoebel, Christopher P Bridge, Sara Ahmed, Oluwatosin Akintola,

Caroline Chung, Raymond Huang, Jason Johnson, Albert Kim, K Ina Ly, Ken Chang, Jay Patel, Marco Pinho, Tracy T Batchelor, Bruce Rosen, Elizabeth Gerstner, Jayashree Kalpathy-Cramer. "Expert-centered evaluation of deep learning algorithms for brain tumor segmentation" (under review)

- Katharina Hoebel, Christopher P Bridge, Albert Kim, Elizabeth Gerstner, K Ina Ly, Francis Deng, Matthew DeSalvo, Jorg Diettrich, Raymond Huang, Susie Y Huang, Stuart Pomerantz, Saivenkat Vaglvala, Bruce Rosen, and Jayashree Kalpathy-Cramer. "Not without context - A multiple methods study on evaluation and correction of automated brain tumor segmentations by experts" (in preparation)

Parts are published in the conference proceedings of Society of Photo-Optical Instrumentation Engineers (SPIE) - Medical Imaging 2022:
Katharina Hoebel, Christopher P. Bridge, Sara Ahmed, Oluwatosin Akintola, Caroline Chung, Raymond Huang, Jason Johnson et al. "Is this good enough? On expert perception of brain tumor segmentation quality." In Medical Imaging 2022: Image Perception, Observer Performance, and Technology Assessment, vol. 12035, pp. 165-175. SPIE, 2022.

5.1 Introduction

5.1.1 Brain tumor segmentation

Glioblastomas (GBM) are the most common and aggressive primary brain tumor in adults [265]. Their aggressive growth patterns lead to an unclear appearance of the tumor boundaries on imaging, caused by the infiltration of tumor cells into the surrounding healthy tissue. Additionally, the tumor boundaries can be quite irregular, and tumors can be multi-focal or multi-centric [266]. Therefore, manual segmentation – creating an accurate outline of the tumor – is a challenging task for physicians. However, accurate tumor segmentations are an important aspect of the state-of-the-art multi-modal treatment approach for GBM. Volumetric segmentations are required for

radiation treatment planning and are expected to play a growing role in treatment monitoring and surgical planning, particularly the visualization of the resection area and the surrounding anatomy [267, 268].

Manual segmentations of GBMs are considered the gold standard for clinical applications. However, the quality of manual tumor contours, not just for brain tumors, is impaired by a high inter and intra-observer variability [269, 270, 271] and several factors have been identified that influence the variability in brain tumor outlines [272]. Due to the poorly defined boundaries, there are multiple potential solutions to the task of segmenting a given brain tumor, and it can therefore be seen as ill-defined [273]. Inconsistent protocols for the interpretation of the available imaging data between institutions, differences in patient anatomy, and varying tumor topography are the main drivers in the inter-observer variability of brain tumor segmentation [274, 275].

Additionally, differences in tumor outlines can be caused by differences in the conceptual understanding of patterns of tumor spread [276]. Apart from the variability in brain tumor segmentation, the manual segmentation of GBM is a time-consuming task. Even experienced physicians can spend up to one hour per patient, depending on the complexity of the case [277, 278].

5.1.2 Deep learning segmentation of brain tumors

Over the last five years, deep learning-based automated segmentation algorithms for brain tumors have substantially improved and can nowadays achieve a performance comparable to human experts [97]. These algorithms have the potential to speed up the laborious process of obtaining accurate volumetric GBM segmentations and limit the variability introduced by individual observers, resulting in more reliable and standardized treatment response assessment and radiation treatment planning [170, 195, 279, 280].

Nonetheless, even high-performing deep learning algorithms can fail on individual cases for various reasons and suffer from performance degradation over time due to shifts in characteristics of the input data [207, 281]. The potentially unreliable performance

limits physicians' trust in the capabilities of DL algorithms [282]. Therefore, the first generations of clinical deep learning-based segmentation tools are generally designed as support tools and will assist clinicians in making decisions of clinical importance. Instead of spending long hours segmenting highly challenging target structures, clinicians would be tasked to determine whether an initial segmentation provided by an algorithm requires manual correction or is already of sufficient quality to be used, e.g., for treatment planning. If corrections are needed, these would then be performed by domain experts and could be used in an active learning framework to improve the model performance continuously [283].

5.1.3 Perception of segmentation quality

Recent research has shown that human experts have a more contextual response to segmentation quality. The metrics that are conventionally used to optimize, evaluate, and report the performance of brain tumor segmentation models are only lowly to moderately correlated with human perception of the segmentation quality [49, 50]. Additionally, similarly to the substantial inter-rater variability in manual tumor delineations, the perception of what constitutes a good segmentation differs among human experts, which is reflected in high disagreements in peer-review for radiation therapy planning [284]. Therefore, the currently available quantitative metrics may not be sufficient proxies to evaluate the clinical acceptability of segmentations [285]. Consequently, using DL algorithms as an effective support system for the generation of brain tumor delineations will require a detailed understanding of the perception of brain tumor segmentation quality by domain experts.

5.1.4 Chapter outline

In this chapter, we describe how we studied experts' perceptions of the quality of brain tumor segmentations using quantitative and qualitative research approaches.

Expert-centered evaluation of brain tumor segmentation Inspired by recent findings on the disconnect between experts' quality perception and commonly used quantitative metrics, we describe a study on expert-centered evaluation of DL brain tumor segmentation models. We first studied how DL segmentation models are currently being evaluated. To this end, we reviewed studies on DL segmentation models of brain tumors and assessed what quantitative metrics are used and whether clinical experts were involved in assessing the models. Second, we performed an experimental study to assess a) the inter-rater variability in segmentation quality perception and b) the variability in the agreement between quantitative metrics and experts' quality perception for segmentations of post-operative glioblastomas.

We found that the most commonly used quantitative metrics for the evaluation of segmentation performance showed a low agreement with the quality of the segmentations of post-operative brain tumors as perceived by domain experts. Additionally, similar to the low inter-observer agreement in manual tumor segmentations, human experts had high inter-observer variability in their assessments of segmentation quality. We conclude that the quality perception of experts is likely more contextual than popular quantitative metrics. While it is evident that humans do not evaluate segmentations in the same way as quantitative metrics, like the Dice score or Hausdorff distance do, no study has identified the aspects that influence their quality perception. These findings are described and discussed in Section 5.3.

The role of context in experts' perception of brain tumor segmentation quality Consequently, we performed a multiple-methods study focusing on qualitative methods to gain deeper insights into experts' perceptions of the quality of brain tumor segmentations. The goal of this study, as outlined in Figure 5-2, is two-fold: First, we aimed to identify factors influencing experts' quality perception of post-operative brain tumor segmentation through a questionnaire. Second, we performed semi-structured interviews to identify the thought processes of experts while they were correcting imperfect segmentations generated by a DL algorithm.

Through this study, we identified five quality criteria that participants used to judge

the quality of segmentations. These criteria were partially aligned with commonly used quantitative metrics. But in contrast to conventional metrics, the identified criteria are largely conditioned on the context of a segmentation error. Furthermore, contextual information was heavily used to limit uncertainties in correcting imperfect brain tumor segmentations. Finally, experts' decisions were influenced by their personal beliefs about whether ambiguous areas should be included in a segmentation and individual biases toward the performance of segmentation algorithms. Results from this study are described and discussed in Section 5.4.

5.2 Methods

5.2.1 Literature review

We searched PubMed for English-language original research articles published between 08/2017 and 09/2022 reporting on DL segmentation models for macroscopic brain tumors. Only articles that described a segmentation algorithm for human brain tumors on macroscopic imaging and contained a performance evaluation of the segmentations were included in the analysis. We recorded the metrics used in the performance evaluation of the segmentation models. We also categorized the quantitative metrics into six groups: overlap-based metrics, such as Dice score, volume-based metrics such as relative volume error, voxel-level confusion matrix-derived metrics such as sensitivity, distance-based metrics such as Hausdorff distance, threshold-metrics such as area under the receiver-operator-curve, and information-based metrics such as variation of information. We provide a overview over the main characteristics of most metric groups in Section 2.6.

List of all metrics in each metric group Here, we list all evaluation metrics that appeared in more than one article. Synonymous terms and standard abbreviations are listed in parentheses.

- Overlap-based metrics: Dice score (F1-score), Jaccard index (intersection-over-union), weighted F-measure

- Confusion matrix-based metrics: sensitivity (true positive rate, recall), specificity (true negative rate), precision (positive predictive value), accuracy, Matthew's correlation coefficient, Cohen's kappa, false positive rate, false negative rate
- Distance-based metrics: Hausdorff distance, average symmetric surface distance, boundary F1-score
- Volume-based metrics: volume, volume similarity, absolute volume difference, relative volume difference
- Threshold-based metrics: area under the receiver-operator curve (AUROC), area under the precision-recall curve (AUPRC)
- Information-based metrics: variation of information, normalized mutual information

5.2.2 Post-operative brain tumor segmentation model

Dataset The study population, image acquisition, and generation of the manual ground truth segmentation for the dataset used in this study are described in Sections 3.2.1, 3.2.2, and 3.2.3.

Data preprocessing Since the dataset originated from longitudinal clinical studies, it contained imaging from multiple study visits of each patient. We split the dataset into training/validation/test subsets on the patient level, such that all available images of a respective patient were part of only one subset. The training/validation/test datasets consisted of the imaging of 34 (464)/ 9 (128) / 11 (119) patients (study visits). We used T1-weighted pre-and post-contrast and T2-weighted FLAIR sequences as the three input channels for model development. All images were registered to the T2-weighted FLAIR image of the respective study visit. Preprocessing consisted of brain extraction with manual correction if needed, N4 bias correction, and z-score normalization of the brain region of each scan as described in Section 2.5.

Segmentation model Monte Carlo (MC)-dropout networks approximate Bayesian neural networks by dropping out full activation maps after each convolutional layer at training and inference time [212] (see Section 4.2.3). At inference time, slightly varying segmentations can be sampled from an approximate posterior distribution and used to quantify model uncertainty.

We trained an MC-dropout 3D U-Net (see Section 2.3), on patches of size 64x64x16 to segment areas of T2-weighted FLAIR abnormality using weighted cross-entropy loss. The U-Net segmentation model consisted of four layers with 32/64/128/256 channels per layer and spatial dropout, dropping out full activation maps, with a dropout rate of 0.2 after each activation layer. The model weights were randomly initialized. To augment the training data, we used horizontal flips and patch augmentation with over-sampling from tumor regions, such that 70% of the patches contained tumor regions and 30% did not contain any tumorous areas. We used a weighted categorical cross-entropy loss with weights of 3 on the positive class (brain tumor) and 0.1 on the negative class (healthy brain tissue/background). The model was trained for 300 epochs using the Adam optimization algorithm and an initial learning rate of 1.0×10^{-4} [258]. The model with the highest average Dice score on the validation dataset was chosen for evaluation and further experiments.

The STAPLE algorithm created the final segmentation from $N = 10$ MC-dropout samples [123]. The segmentation model was implemented in DeepNeuro [286]. Quantitative segmentation quality metrics were computed using the python package Pymia [287].

Model uncertainty An uncertainty map $U_{MC}(x)$ was obtained by averaging over the voxelwise entropy of all $N = 10$ segmentation samples:

$$U_{MC}(x) = -\frac{1}{N}\sum_{i=1}^{N} p_i(x) \log\left(p_i(x)\right),$$

where $p_i(x)$ denotes the model's sigmoid output for each pixel x for sample i. To describe image-level model segmentation uncertainty, we computed the mean uncertainty

of the segmented region:

$$U_{mean} = \frac{1}{\sum_x b\left(p_{\text{STAPLE}(x)}\right)} \sum_{p_{\text{STAPLE}(x)} \geq 0.5} U_{MC}\left(x\right),$$

where $b(x)$ is the binarization function [288].

5.2.3 Expert ratings of segmentation quality

Experts We recruited eight experts (fellows and attendings from the neuro-oncology, neuroradiology, and radiation oncology departments at three academic centers in the United States) to provide quality ratings for automatically generated brain tumor segmentations. The eight experts were split into two groups stratified by specialty and experience. The first group (experts 1, 2, 6, and 8) consisted of two neuro-oncologists, one with more than five and now with more than ten years of experience, and two neuro-radiologists, one with more than five and one with more than ten years of experience. The second group of experts (experts 3, 4, 5, and 7) consisted of two neuro-oncologists with more than five years of experience, a radiation oncologist specializing in the treatment of neurological cancers with more than ten years of experience, and a neuro-radiologist with more than five years of experience.

Ratings of segmentation quality We split 60 randomly selected cases from the test set into two datasets with 30 cases each and assigned one group of experts to each set. Due to the slice thickness of 5 mm of the available imaging, the experts reviewed the MRIs as a stack of 2D axial slices. The segmentations were represented by red outlines overlayed on the T2-weighted FLAIR images. The experts were provided with two different modes of reviewing the segmentations: either as a mosaic consisting of all axial slices of the T2-weighted FLAIR volume (see Figure 5-1 for an example) or as a stack of the axial slices that they could scroll through. The quality of each segmentation was graded with a score between 1 and 4. The rating categories were as follows:

- 1: not acceptable

- 2: acceptable with moderate changes
- 3: acceptable with minor changes
- 4: acceptable without changes

The ratings were recorded via a dropdown menu on a centrally hosted spreadsheet.

Each expert was provided the same 15 cases to practice, and six experts rated four cases twice during different sessions to assess intra-rater variability. All experts were blinded toward patient identity and treatment status and viewed the cases in a randomized order.

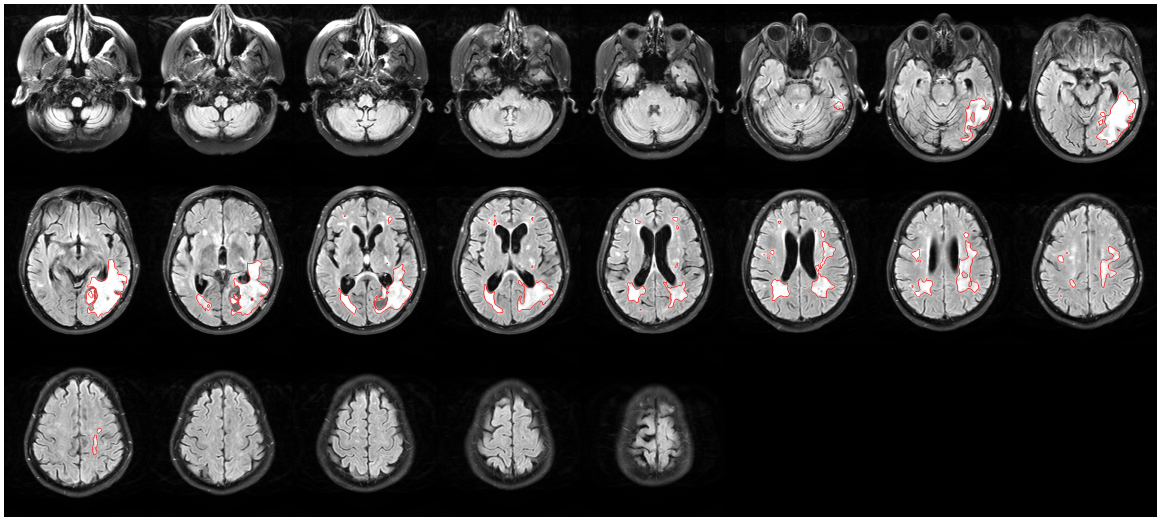

Figure 5-1: **Sample case with segmentation.** Mosaic of 2D axial slices of T2-weighted FLAIR overlayed with the outline of the automatically generated segmentation in red.

5.2.4 Qualitative study design

We conducted a two-part qualitative study with neuroradiologists and neuro-oncologists experienced in working with brain tumor imaging. One part consisted of participants completing an online questionnaire. We identified factors that influenced their perception of segmentation quality based on the responses. Additionally, we performed semi-structured interviews in which participants were also asked to correct imperfect brain tumor segmentations. Figure 5-2 provides an overview over the study process.

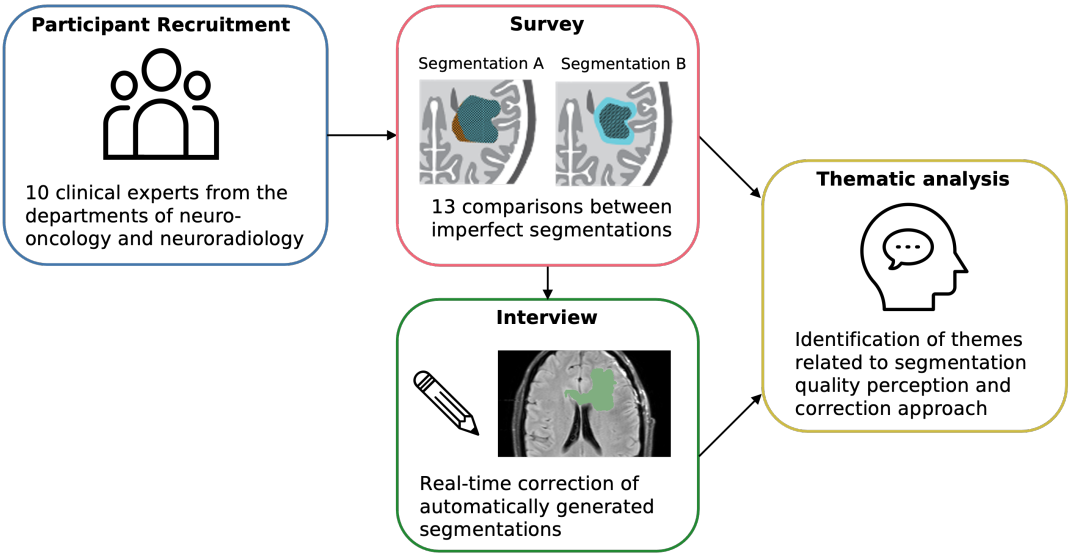

Figure 5-2: **Study overview.** We recruited participants from two major academic teaching hospitals (blue box). Participants first completed an online questionnaire comparing imperfect brain tumor segmentations and provided free text justifications (red box). Subsequently, they participated in a virtual semi-structured interview and corrected proposed segmentations of post-operative glioblastomas (green box). Lastly, we performed a thematic analysis of the questionnaire responses and interview transcripts (yellow box)

The study was approved by the institutional review board (IRB) of Mass General Brigham.

Participants We recruited participants using a convenience sample from the two main sites of our hospital system (Massachusetts General Hospital and Brigham and Women's Hospital, Boston, MA, USA). Volunteers were contacted via email by members of the research team. To be eligible to participate in the study, volunteers were required to be residents, fellows, or attendings in neuro-oncology or neuroradiology. Additionally, they were required to have at least one year of experience working with brain tumor imaging at the time of their participation in the study. Written and oral consent was given for the questionnaire and interview parts of the study.

5.2.5 Collection of qualitative data

Questionnaire

To identify factors influencing physicians' assessment of brain tumor segmentation quality, we showed each participant 13 comparisons between imperfect segmentations. We guided participants' attention to specific differences between the depicted segmentations by providing comparisons. Each comparison focused on one specific aspect we identified as a potential source of disagreement in quality perception through the quantitative study on expert-centered segmentation quality perception described in Section 5.3.1.

Comparison categories in the questionnaire The comparisons in the questionnaire were categorized into four groups:

1. **Over versus under-segmentation:** Importance of over-segmenting (capturing non-tumorous areas around the tumor margins) in comparison to under-segmenting (missing tumorous areas at the tumor margins)

2. **Holes:** Importance of missed areas (holes) within a segmentation

3. **Distance:** Importance of the distance between a false positive area and the primary tumor

4. **False negatives versus false positives:** Importance of missed areas (false negatives) in comparison to false positive areas.

The categories are illustrated in Figure 5-10. The difference between the imperfect segmentations was also quantified using the Dice score and the 95th percentile Hausdorff distance.

We used illustrations of an axial cross-section of a brain with a tumor. Sample illustrations are depicted in Figure 5-3, A1-3. Each illustration showed one tumor lesion. Using illustrations instead of real MRI images, participants focused on the segmentations themselves without any ambiguities caused by blurry tumor margins,

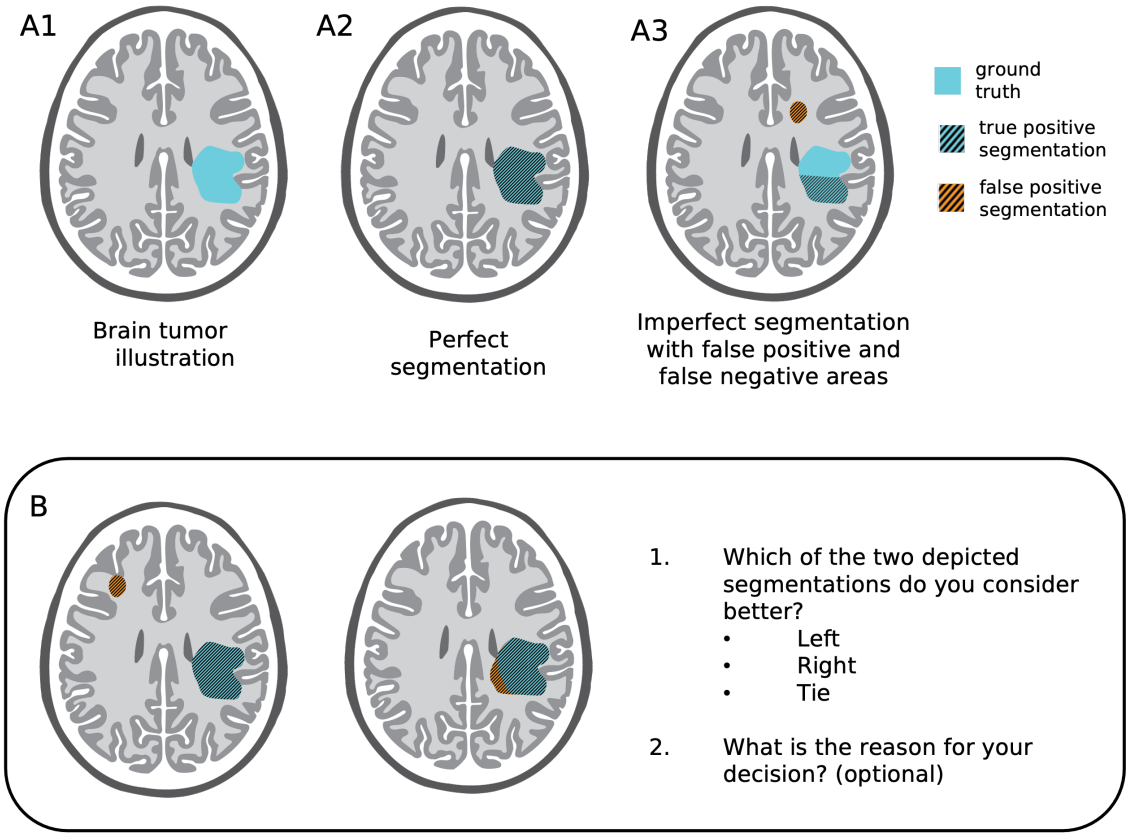

Figure 5-3: **Brain tumor segmentation illustrations for the questionnaire.** A1: Ground truth with tumor in turquoise and non-tumorous brain in gray tones; A2: The segmentation is indicated by a black striped pattern; A3: True positive areas appear as turquoise areas with black stripes, false negative areas are turquoise (without stripes), false positive areas (no tumor present but selected as part of the segmentation) are highlighted in orange (with black stripes). B: Example comparison of two brain tumor illustrations: both segmentations are imperfect with a false positive area. On the left, the false positive area is separated from the primary lesion; on the right, it is connected to the primary lesion.

additional non-tumor abnormalities, and imaging artifacts. Furthermore, this way we were able to control the sources of variability.

For each of the 13 comparisons, we asked participants to indicate whether they preferred one of the two imperfect segmentations or considered them to be of equal quality (tie) as shown in Figure 5-3B. They also had the opportunity to provide a free text answer to justify their decisions. Illustrations were prepared using Adobe Illustrator and segmentation quality metrics were computed using the Python package pymia [287]. We collected and managed the responses to the questionnaire using

REDCap (Research Electronic Data Capture) in compliance with our institution's IRB requirements [289, 290].

Interview

The semi-structured interview aimed to identify experts' thought processes while correcting brain tumor segmentations. Therefore, we asked participants to correct four selected segmentations of post-operative high-grade gliomas. We generated the brain tumor segmentations using the deep learning algorithm described in Section 5.2.2. Each case contained a different characteristic that made the segmentation challenging, e.g., particularly blurry boundaries or a specific error in the automatic segmentation.

The interviews were performed virtually. Participants shared their screens so the interviewer could observe the segmentation correction process. The goal of the segmentation process was to capture areas of T2-weighted FLAIR abnormality corresponding to the total tumor burden. We used the 3D Slicer software to display and edit the automatically generated segmentations [155]. Participants were not expected to have experience working with 3D Slicer. Therefore, the interviewer walked them through the required steps during the interview. To perform the corrections, the participating experts had access to the following MRI sequences: T1W-weighted pre-contrast, T1-weighted post-contrast, T2-weighted, and T2-weighted FLAIR.

The interview guide was developed to follow the process of correcting the four automatically generated segmentations and was informed by our findings from Section 5.3. It focused on the overall perception of the quality of the provided segmentation and contained questions specific to the characteristics of each case, and the participant's approach to correcting an imperfect segmentation of post-operative brain tumors. Additionally, we encouraged participants to think out loud during the correction process. Interviews were scheduled to last less than 90 minutes. We used Microsoft Teams as the platform for virtual meetings in accordance with guidelines specified by the IRB. Meetings were recorded and automatically transcribed by Teams. The interview transcripts were assessed for their accuracy and corrected if needed.

We continued to collect responses to the questionnaire and to perform interviews until we achieved thematic saturation, defined as no additional themes could be identified from the questionnaires and the interviews.

5.2.6 Qualitative data analysis

We performed a thematic analysis of the free-text responses to the questionnaire and interview transcripts using RQDA, an R package for qualitative data analysis [291]. We used a combined deductive and inductive coding approach. The initial set of codes for the questionnaire was based on findings reported in Section 5.3.1. The initial set of codes for the interview was based on preliminary results from the questionnaire. A research team member (KVH) then reviewed the first four complete sets of replies to the questionnaire and interview transcripts and developed two separate codebooks for the questionnaire and the interview. The same researcher coded all questionnaire responses and interview transcripts and the codebook was updated as needed. The final codebook for the questionnaire contained 11 codes related to the role of tumor biology, clinical applications, false-positive areas, and false-negative areas in evaluating segmentation quality. The final codebook for the interview part of the study contained 26 codes related to the brain tumor, the patient, clinical aspects, and the expert performing the edits. All transcripts were reviewed a second time using the final lists of codes. Themes were identified using thematic analysis. Lastly, we performed a respondent validation of the identified themes with three expert participants.

5.2.7 Statistical analysis

Agreement between segmentation quality ratings was assessed using Krippendorff's alpha (type ordinal) [292] for more than two ratings per case and Gwet's AC2 [293] with linear weights for pairwise expert comparisons. We chose Gwet's AC2 instead of Cohen's kappa for pairwise comparisons and the assessment of intra-rater reliability because Gwet's AC is not affected by the frequency distribution of the ratings [294]. Comparison between the distributions of continuously valued features was determined

using the Kruskal-Wallis test [190]. Agreement between participants' preference for segmentations and their quantitative superiority on the questionnaire was determined using a Binomial test. Statistical significance was defined as a p-value ≤ 0.05. Data analysis was performed in python 3.6 and R 4.0.2.

5.3 Expert-centered evaluation of brain tumor segmentation

5.3.1 Results

Literature review

We searched PubMed for full-text articles that covered deep learning segmentation algorithms of brain tumors. The initial search identified 248 articles. We excluded 53 papers after screening titles and abstracts and 15 more after a review of the full-text articles. As a result of this process, we included 180 studies in the final analysis. Figure 5-4 illustrates the literature selection process.

Quantitative evaluation of segmentation model performance Among the 180 articles we reviewed, the three most popular strategies to evaluate segmentation quality were the Dice score as the only evaluation metric (42 articles, 23.3%), a combination of Dice score, Hausdorff distance, sensitivity, and specificity (26 articles, 14.4%), or a combination of Dice score and Hausdorff distance (21 articles, 11.7%). Overall, the Dice score was used in 170 of the 180 articles (94.4%) either exclusively or in combination with other metrics. Sensitivity and Hausdorff distance were used in combination with other metrics in 86 (47.8%) and 69 (38.3%) articles, respectively (Figure 5-5A). We grouped each metric with other metrics that are used to evaluate similar concepts, i.e., both the Dice score and Jaccard index are computed based on the overlap between the ground truth and the predicted segmentation and are therefore categorized as overlap-based metrics. The most widely used metric groups were overlap-, confusion matrix- and distance-based metrics (Figure 5-5B). 28.3% of

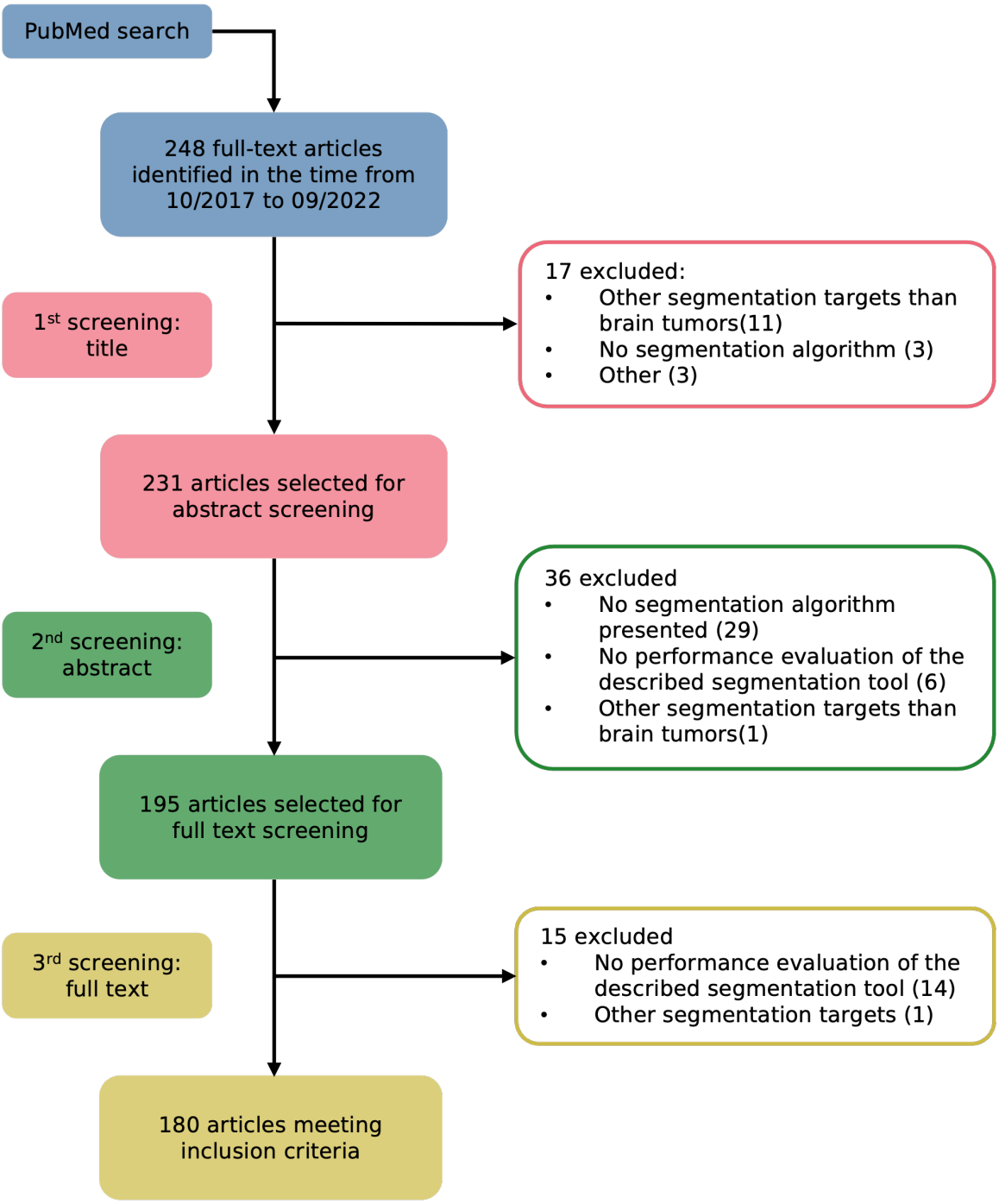

Figure 5-4: **Consort diagram of the literature review process.** The process to identify suitable articles for review consisted of an initial screening of PubMed (blue), a first screening of the titles (red), a second screening of the abstracts (green), and final screening of the full texts (yellow).

articles used metrics from only one, 38.3% from two, and 31.1% from three metric groups (Figure 5-5C). 179 of the 180 studies included at least one metric from the

overlap-, confusion matrix-, or distance-based groups in their analysis. We found that overlap-based metrics are most frequently combined with confusion matrix-based metrics. Distance-based metrics are always associated with overlap-based metrics, as illustrated in Figure 5-5D. 42 studies (23.3%) combined metrics from all three of the most common metric groups.

Segmentation model performance evaluation by clinical experts In addition to the purely quantitative evaluation of segmentation performance described above, five of the 180 studies (2.8%) included an assessment by clinical experts. Table 5.1 contains a list of these studies. The diverse evaluation approaches involved measuring the time it took clinical experts to manually correct an automatically generated segmentation, assessing the consensus between automatic segmentations edited by experts, and rating the quality of the segmentations. The number of clinical experts involved in the evaluation ranged from 2 to 20. The most common clinical specialties were neuroradiology, radiation oncology, and neurosurgery. None of the studies compared the assessment of their clinical experts with quantitative segmentation quality metrics.

User-study on segmentation quality perception of clinical experts

Our literature review on DL segmentation model evaluation showed that most publications rely on a purely quantitative assessment of model performance. While a few articles included qualitative evaluation performed by clinical experts, none linked quantitative and qualitative segmentation quality measurements. A systematic determination of such measurements could inform future algorithm development and clinical validation of such algorithms. Therefore, we conducted a user study on the quality perception of postoperative brain tumor segmentation. The study's goal was to determine the inter-rater variability in segmentation quality perception and its relationship with quantitative measurements.

Our brain tumor segmentation model achieved a mean Dice score of 0.72 on the held-out test dataset. The previously reported Dice score for this dataset, using the same dataset splits, was 0.70 [195], and the average Dice score for the inter-reader

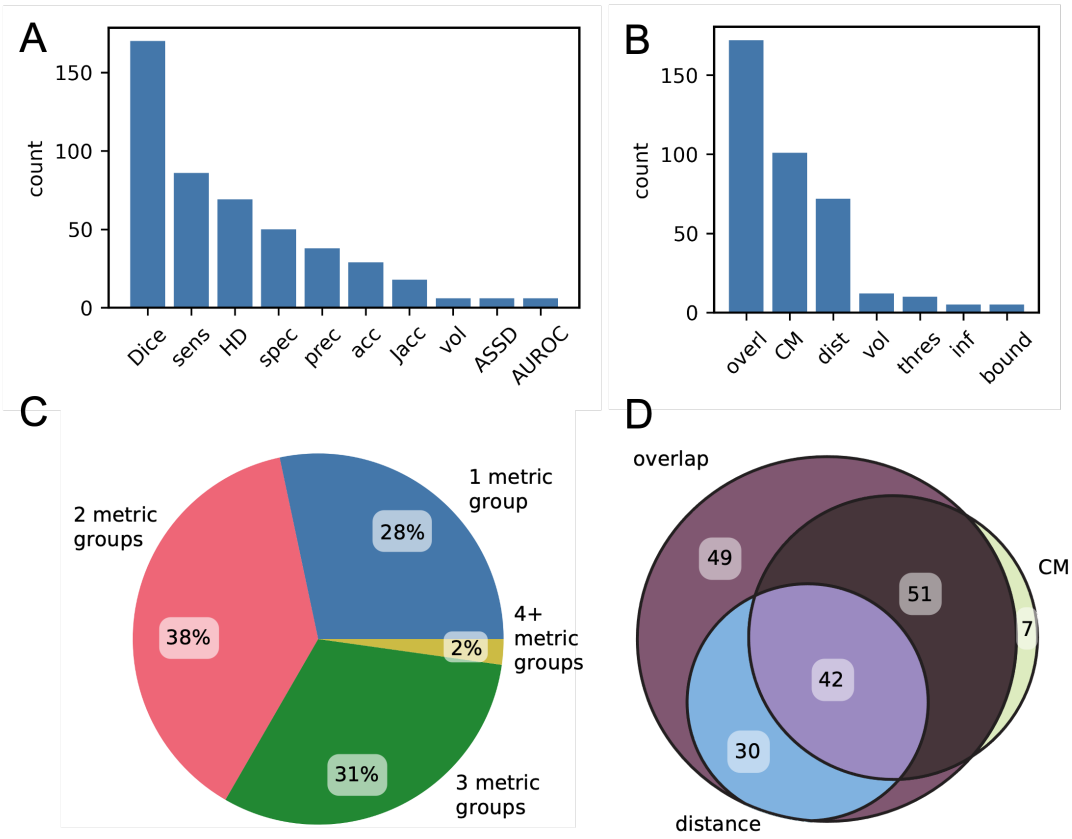

Figure 5-5: **Use of quantitative segmentation quality metrics in the reviewed literature.** A: Count of how often the ten most popular segmentation quality metrics were used to evaluate the performance of the segmentation models; B: Count of how often metrics belonging to one of the seven defined metric groups were used to evaluate the performance of the segmentation models; C: percentage of studies that used metrics from one/two/three/four or more metric groups; D: Venn diagram illustrating the frequency of metric group combinations for segmentation model evaluation between the three most popular groups of segmentation quality metrics. Sens: sensitivity; HD: Hausdorff distance; spec: specificity, prec: precision, acc: accuracy; Jacc: Jaccard index; vol: volume; ASSD: average symmetric surface distance; overl: overlap-based metrics; CM: voxel-level confusion matrix-based metrics; dist: distance-based metrics; vol: volume-based metrics; thres: threshold-based metrics; inf: information-based metrics; bound: boundary-based metrics. We provide an overview over the characteristics of most of the metric groups in Section 2.6.

Table 5.1: **Overview over publications with expert-centered segmentation model evaluation.** *: this study tested an interactive segmentation algorithm, the user interactions were used to improve an initial segmentation; N: number;

Author, year	**Segmentation target**	**Expert evaluation metric**	**Experts (N)**	**Expert background (N)**
Lu et al., 2021[295]	Brain metastases Meningeoma Acoustic neuroma	Time to correct Agreement between corrected segmentations Time to correct	8	Neuroradiology(1) Radation oncology (5) Neurosurgery (2)
Conte et al., 2021 [296]	Glioma (pre-op)	Time to correct	2	Neuroradiology (2)
Di Ieva et al., 2021 [297]	Glioma (pre-op)	Binary quality classification (acceptable/not acceptable)	4	Neuroradiology (1) Radiation oncology (1) Neurosurgery (2)
Mitchell et al., 2020 [298]	Glioma (pre-op)	Preference manual/ automatic Segmentation quality (scale 0-10)	20	Neuroradiology (20)
Wang et al., 2018 [299]	Glioma (pre-op) Fetal organs	User interaction time*	2	Radiology (1) Obstetrics (1)

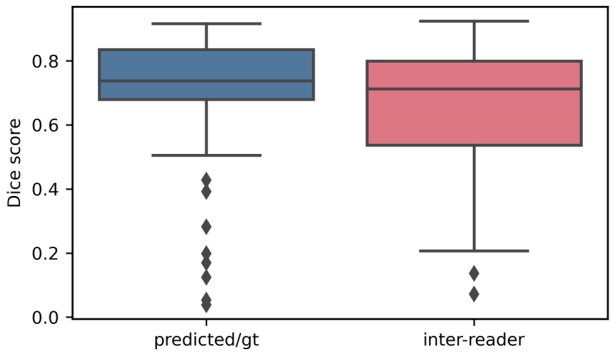

Figure 5-6: **Segmentation model performance compared to inter-rater variability.** Dice score distribution between the predicted segmentation and the manually defined ground truth (gt) segmentations on the held-out test dataset (blue) and between two independent manual segmentations on a subset of the full dataset (red).

agreement was 0.64. The distribution of the Dice score for the model predictions and inter-reader agreement are illustrated in Figure 5-6. The sensitivity, relative volume error and 95th percentile Hausdorff distance were 0.77, 0.33, and 6.5 [voxel], respectively. We obtained 264 segmentation quality ratings, including 24 double reads to determine intra-rater variability from six of the eight participating domain experts, for the 60 cases.

Differences among experts The intra-rater agreement based on double-reads ranged from 0.24 to 1 with a median of 0.88 (Gwet's AC2). The inter-rater agreement among all quality ratings was low, with a Krippendorff's alpha value of 0.34. On pairwise comparison, we found that the agreement between the ratings of individual experts showed a wide variability and ranged between 0.37 and 0.79 (median: 0.59, Gwet's AC2).

Panels A1 and A2 of Figure 5-7 illustrate the pairwise agreement between experts. The low agreement between raters is possibly caused by different internal quality class thresholds. The different cutoffs between rating categories become visible after sorting the cases in each subset according to their mean rating. Panels A1 and A2 of Figure 5-8 illustrate the varying thresholds between the rating categories. The threshold between what constitutes an acceptable segmentation and one that requires

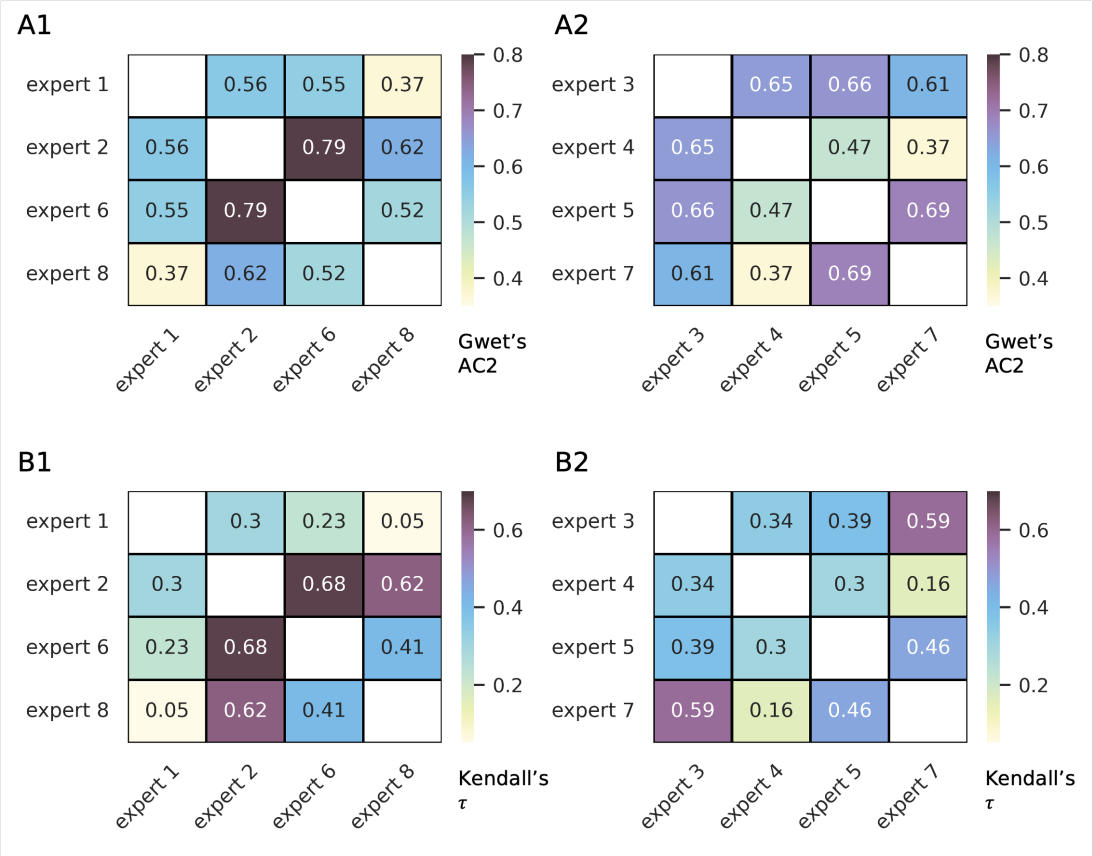

Figure 5-7: **Pairwise agreement and correlation between experts.** Pairwise agreement and correlation were computed between all pairs of experts on the two sets of cases. Each panel (A1/A2 and B1/B2) represents one set of 30 cases that were rated by the same group of experts without overlap between the sets. A1/A2: pairwise agreement between experts' ratings using Gwet's AC2; B1/B2: pairwise correlation between experts' ratings using Kendall's τ.

minor changes varies between the experts.

However, as indicated by the variability in the pairwise correlations between experts' ratings, depicted in panels B1 and B2 of Figure 5-7, these individual thresholds do not account for the total variability we observed in the quality ratings. Therefore, we assessed whether there were additional factors that influenced these differences. For this analysis, we separated all cases into two groups according to the difference between their lowest and highest rating: a high agreement group (rating difference ≤ 1) consisting of 40 cases and a low agreement group (rating difference ≥ 1) consisting of 20 cases.

The following factors showed statistically significant associations with a lower agreement between experts: higher segmentation volume of the automatic segmentation (p-value = 0.04), lower Dice score (p-value < 0.001), higher 95th percentile Hausdorff distance (p-value = 0.046) between the automatic and the manual ground truth segmentation, and a higher segmentation uncertainty of the segmentation model (p-value < 0.001). The different distributions of these metrics in the low and high agreement groups are illustrated in Figure 5-8B. We found no differences between the low and high agreement groups' surface area (p-value = 0.12), sphericity (p-value = 0.62), and volume similarity (p-value = 0.51).

Differences between commonly used metrics and expert quality perception Lastly, we compared the correlation between the ratings of each expert and the most used quantitative segmentation quality metrics. The Kendall's τ correlation coefficients between seven selected quantitative metrics and the ratings of all eight raters are presented in Figure 5-9. Overall, we observed a high variability between the different metrics (rows) and among the experts (columns). These findings indicate that some metrics agree better with expert quality perception and that there are differences in the agreement between experts.

The highest correlations and lowest variability among raters were found with the 95th percentile Hausdorff distance, followed by sensitivity and surface Dice. Overlap and volume-based metrics, like the Dice score, volume similarity, and relative volume

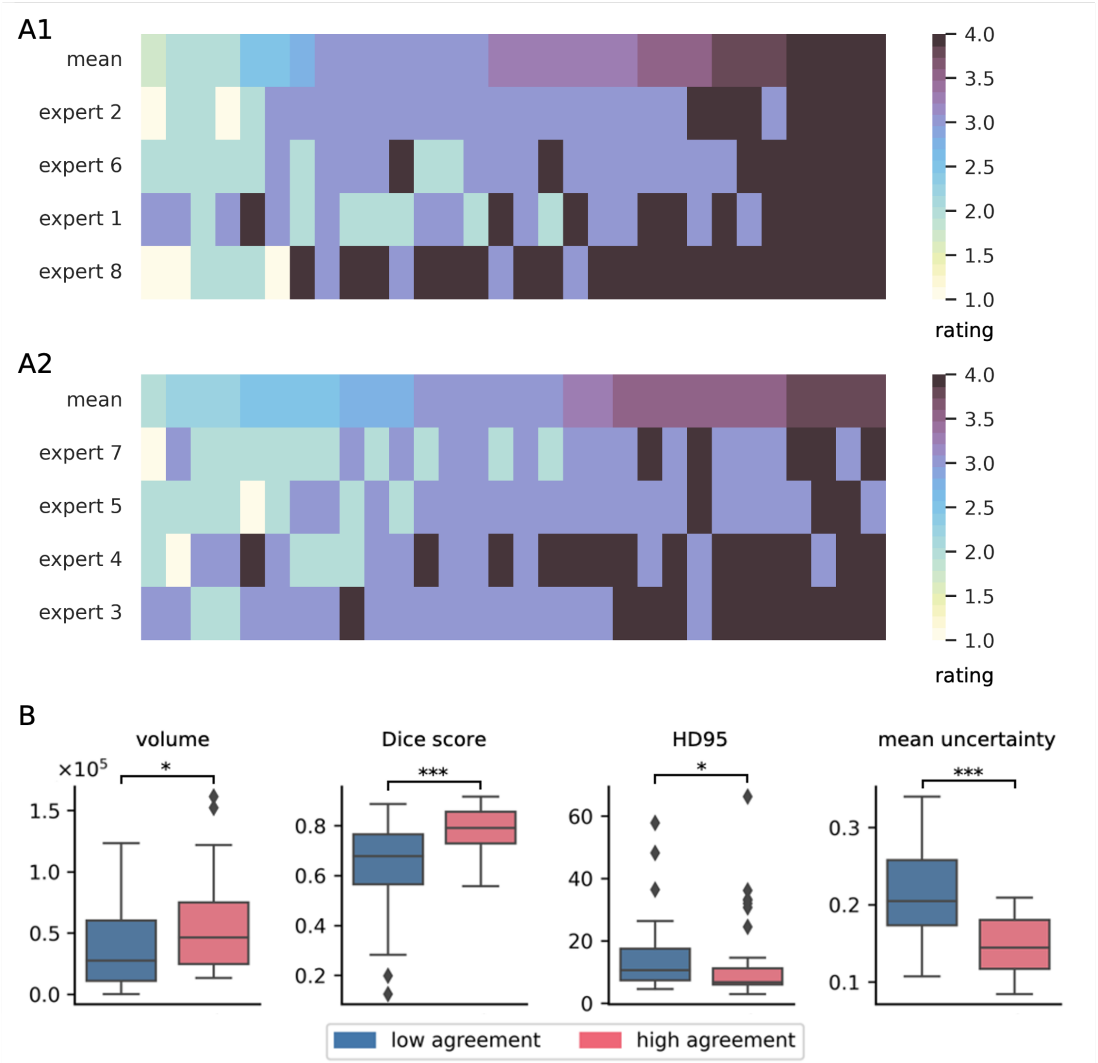

Figure 5-8: **Factors influencing disagreement between experts.** A: Agreement between raters for single cases. Each column represents the ratings for one case and each row one expert. The cases within both of subsets of data were ranked based on their average rating. Experts were ordered based on the average rating assigned to all cases from lowest (top) to highest (bottom) average rating. Each panel (A1/A2) represents one set of 30 cases that were rated by the same group of experts without overlap. B: Distributions of rater independent characteristics of cases for the group with low (blue) and high agreement between experts (red). Statistically significant differences between the distributions are indicated by brackets above the box plots. *: p-value ≤ 0.05; **: p-value≤ 0.01; ***: p-value ≤ 0.001; HD95: 95th percentile Hausdorff distance.

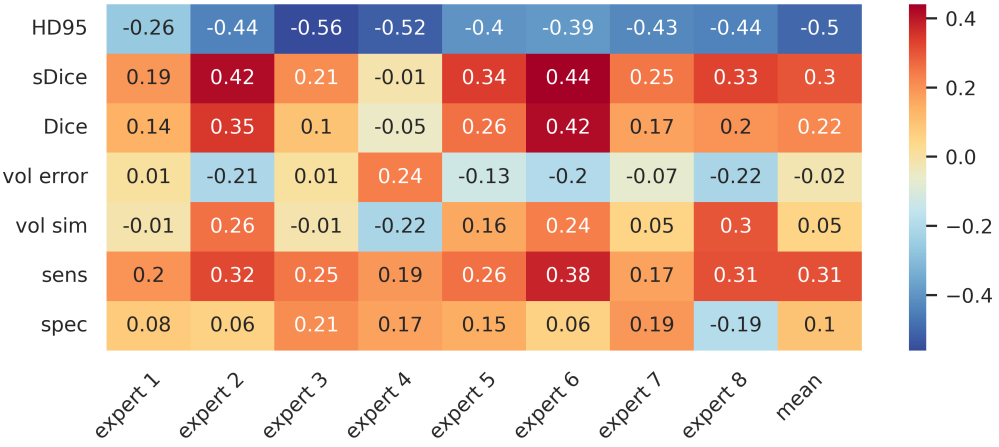

Figure 5-9: **Correlation between segmentation metrics and ratings.** Kendall's τ correlation coefficient between the ratings provided by each expert and selected segmentation quality metrics: 95th percentile Hausdorff distance (HD95th), surface Dice score (sDice), Dice score, relative volume error (vol error), volume similarity (vol sim), sensitivity (sens), and specificity (spec).

error, showed lower correlations than distance metrics.

5.3.2 Discussion

In this section, we presented results from a study on expert-centered segmentation quality evaluation. First, we identified how articles reporting on deep learning brain tumor segmentation evaluate those models. Second, we performed an experimental study on clinical experts' perception of brain tumor segmentation and showed how current evaluation practices identified in the literature review relate to the experts' segmentation quality assessment.

Evaluation of segmentation quality primarily relies on the same quantitative metrics. The Dice score was the most popular metric and 23.3% of the analyzed articles relied on the Dice score as the only metric for segmentation performance evaluation. However, overlap metrics have low correlations with the human quality perception [50]. Our experimental data on the quality perception of post-operative brain tumor segmentation confirms this finding. Correlations between the most frequently used metric, the Dice score, and experts' quality perception were low, as

shown in Figure 5-9. Additionally, we observed high variability in the correlations between individual experts. Distance-based metrics that showed the highest agreement and lowest variability among experts were never used as the primary performance metric for segmentation quality (Figure 5-5). Unlike overlap- and confusion-matrix-based metrics, most distance metrics, like the Hausdorff distance, are not bounded between 0 and 1. Therefore, they are harder to interpret and compare between studies using different datasets. Our findings suggest that the surface Dice score can be a promising alternative, as it had a higher agreement with experts' ratings than the Dice score and is constrained to values between 0 and 1.

Expert-centered evaluation suffers from high subjectivity in segmentation quality perception Even though DL-assisted segmentation algorithms can increase inter-rater agreement of segmentations [279], the inter-rater reliability in segmentation quality perception has not been studied to date. Given the known low inter-rater agreement in the manual segmentations for challenging targets such as tumors [269, 271, 270], we expected variability in the segmentation quality perception of the participating experts. The intra-rater reliability in our experiments was high, with a median of 0.88 (Gwet's AC2) for the six experts. This finding contrasts the low inter-rater agreement between all raters (Krippendorff's alpha: 0.39). In part, individually varying thresholds between adjacent segmentation quality grades can account for the observed variability (Figure 5-8, A1 and A2). Similar variability in individual thresholds between ordinal rating categories has been observed in disease severity classification [41, 300]. In contrast, experts are more consistent in their assessment of disease severity when they are comparing images rather than assigning absolute ratings [47]. Therefore, alternative evaluation techniques based on comparisons between segmentations and a defined segmentation quality standard may warrant higher agreement between raters.

Furthermore, we identified several factors significantly associated with a lower agreement between raters (Figure5-8B). Among these factors were smaller volumes of the automated segmentation, indicating that the experts saw aberrations in these cases as of different importance. Additionally, cases with a high disagreement in the ratings

were associated with higher model uncertainty. Epistemic uncertainty is expected to be higher for cases dissimilar to those seen during model training [301]. Model uncertainty, which can be computed independent of any manual ground truth, is highly correlated with the segmentation Dice score [302, 303]. As a result of our finding that epistemic uncertainty is higher for samples with a lower agreement between raters, cases with a high uncertainty could be automatically routed for review by multiple experts, e.g., during a peer review session.

Low consensus in the performance of expert-centered evaluation of segmentation models In the surveyed articles, expert-centered evaluation, if performed, played a supplementary role to a primarily quantitative assessment of the test data using established quantitative metrics. Among the five studies that involved clinical experts in the performance evaluation, we observed high variability in the applied process (see Table 5.1). This variability may reflect the diversity of use cases for segmentation, e.g., treatment monitoring or planning. Solely relying on quantitative evaluation is insufficient to assess the readiness of DL segmentation models for downstream clinical tasks [165]. As a solution, Jha et al. advocated that artificial intelligence algorithms should be evaluated with the involvement of clinical experts and their clinical context in mind [304].

5.3.3 Limitations

There were some limitations to this study. The assessment of T2-weighted FLAIR abnormality segmentations was based only on T2-weighted FLAIR images and no additional sequences. In clinical settings, radiologists and neuro-oncologists may use other sequences such as T2-weighted and T1-weighted with and without contrast for additional information. Therefore, the simplified study setup did not fully mimic segmentation quality assessment as it would be performed in a clinical workflow. Furthermore, the experiments were limited to the segmentation quality perception of postoperative brain tumors on T2-weighted FLAIR, a highly complex and ambiguous segmentation target. Our findings may not generalize to other segmentation

targets. Future studies should evaluate whether the observed disagreement between segmentation quality metrics and the quality perception of experts can be observed for other segmentation targets as well. Based on our findings, we suggest that the performance of segmentation models should include a use-case-focused assessment performed by clinical experts. If this is not feasible, a purely quantitative analysis should utilize selected segmentation quality metrics that correlate with their usefulness for the desired clinical application.

5.3.4 Conclusions

In summary, in this section we presented a literature review and experimental study on the performance evaluation of deep learning brain tumor segmentation models. We found that most published studies relied on very similar sets of quantitative evaluation metrics. Only a few studies involved clinical experts in the evaluation of the segmentation models. However, our experimental results elucidated the disconnect between quantitative quality metrics and the qualitative perception of segmentation quality by domain experts. Consequently, next we aim to develop a better understanding of the quality perception of clinical experts to catalyze the development of tailored quantitative metrics to develop clinically helpful segmentation models.

5.4 Role of context in experts' perception of brain tumor segmentation quality

5.4.1 Results

Study participants

We recruited ten clinical experts experienced in working with the imaging of brain tumors. Four participants were neuro-oncologists (one with more than three, one with more than six, and two with more than eight years of experience). The remaining six participants were neuroradiologists (two with more than three, one with more than

six, and three with more than eight years of experience).

Segmentation quality criteria

We collected responses to 130 questionnaires (13 comparisons from each of the ten participants). We received a total of 122 justifications describing why experts favored one segmentation over the other or considered them equal quality.

Participants were equivalent to random chance in selecting the segmentation quantitatively identified as better, either by Dice score or 95th percentile Hausdorff distance (Binomial test, Null hypothesis: $p>1/2$). We aggregated the responses of all participants through a majority vote. It is important to keep in mind that these evaluations were performed based on illustrations of segmentations that provided only minimal context.

Based on the justifications the participants gave for why they favored one imperfect segmentation over another or considered them equal quality, we identified five segmentation quality criteria. Since most justifications listed one quality criterion as the reason for their decision, we first identified co-occurrences of codes used to justify the same decision. Themes were determined based on co-occurring codes in combination with the context of each comparison, i.e., the difference between the two depicted segmentations as illustrated in Figure 5-10.

Over-segmentation is better than under-segmentation Experts generally preferred over- to under-segmentation. Generous segmentations that may include potentially healthy tissue surrounding the tumor are preferred over conservative segmentations that may miss tumorous areas at the tumor margins. Figure 5-10A presents an example for over (left) and under-segmentation (right). This quality criterion is mainly based on knowledge about the growth behavior of highly invasive high-grade gliomas. As one participant explained: *"it's [an] infiltrative tumor, so there probably is tumor beyond that margin [...]."*

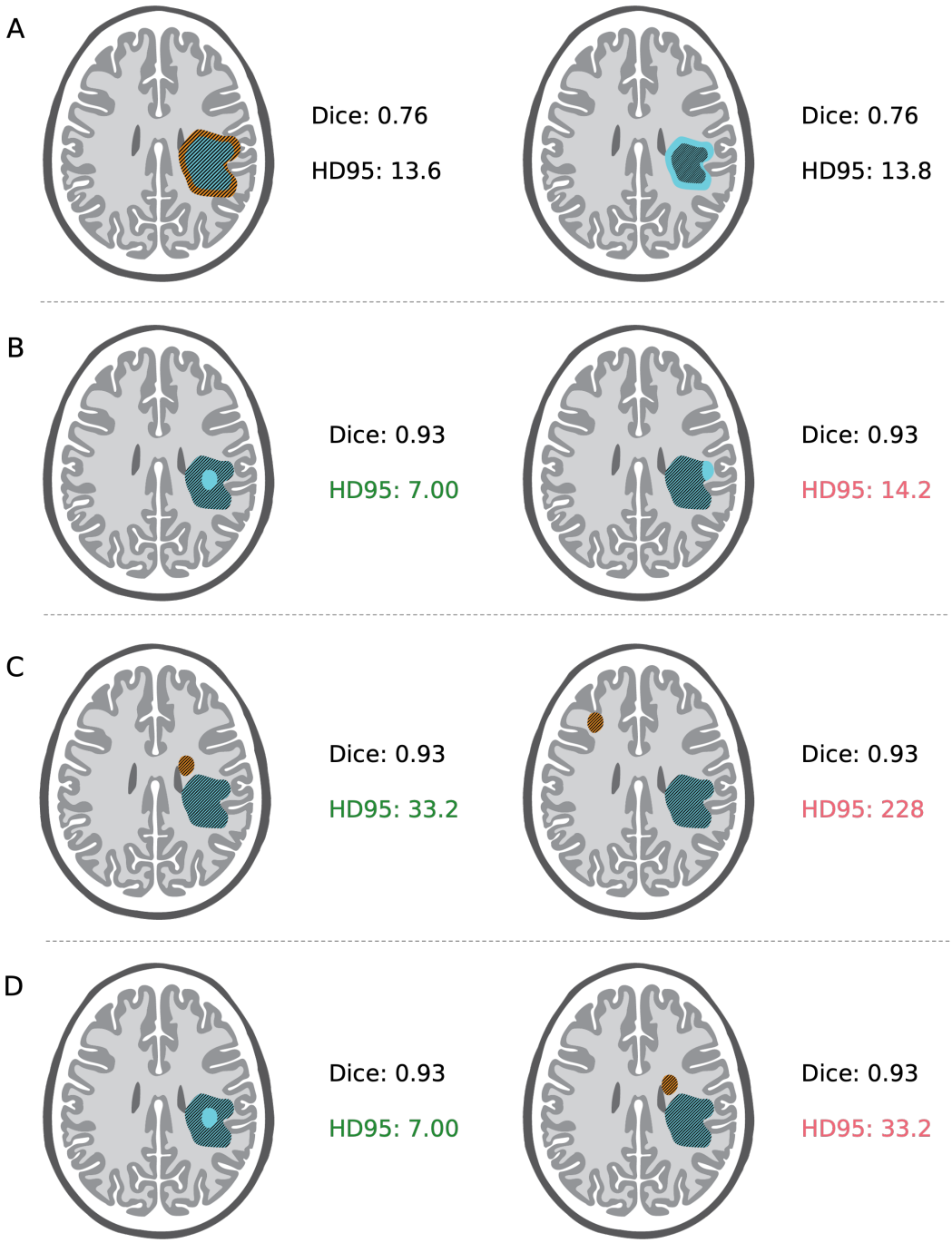

Figure 5-10: **Categories of questionnaire comparisons.** Each of the comparisons participants had to make in the questionnaire could be categorized as one of four categories. A: Over-segmentation (left) vs under-segmentation (right); B: Position of false negatives (holes) within the segmentation; C: Distance between a false positive and the primary tumor; D: Segmentation containing false a negative area (left) vs one containing a false positive area (right).

Capturing the tumor margins The second quality criterion is the extent to which the segmentation captures the margins of the tumor, as the margins were generally perceived as more important than the central areas. Holes in the middle of a segmentation, as depicted on the left in Figure 5-10B, were perceived as acceptable errors, particularly if the segmentations were intended for use in surgery or radiation treatment planning: *"The extent of tumor is known and if resected, the central part [that the segmentation missed] will be taken out as well."* However, if parts of the tumor margins were not included in the segmentation and therefore not treated, this would likely lead to tumor recurrence.

Distance between false positive areas and the primary lesion If a segmentation wrongly includes tissue considered healthy (false positive), the distance between this false positive area and the primary lesion is an important quality criterion. Based on the respondent's knowledge of tumor behavior and experience, high-grade gliomas are more likely to spread locally than in more distant regions. One expert explained: *"[the segmentation] captures an area that is close to the original tumor ROI [region of interest], so it might be real."* An extreme example is a false positive area contralateral from the primary lesion as *"[A] false positive [in the] contralateral [hemisphere] suggests even wider spread of tumor."* The diagnostic error based on this erroneous segmentation would have severe consequences for the patient, leading to the wrong assumption that the patient has bilateral disease.

Indication of the right number of lesions While evaluating the quality of a segmentation, participants considered clinical errors that could be made based on an erroneous segmentation. If a segmentation contained a separate false positive area, it would wrongly indicate that the patient has two tumorous lesions instead of one. This error has important clinical implications for the patient's therapeutic opportunities.

Effort it takes to correct a lesion Participants showed a preference for lesions they perceived as easier to correct. The perception of correction effort was influenced by the software that participants had previous experience with, e.g., software that

would require them to paint over the whole tumor volume versus outlining the tumor. In accordance with the importance of capturing the tumor margins, we observed a higher emphasis on the tumor margins compared to the full tumor volume.

Experts' personal preferences for segmentation quality criteria We found that some experts repeatedly used the same reason to justify their decisions. Some reasons were overwhelmingly brought up by only one or two experts while other justifications were used more broadly. The importance that a segmentation captures the tumor margins is mentioned disproportionately often by two experts. They used this justification for five and nine of their decisions. The effort to correct a segmentation was only brought up by two experts. However, one of them used this criterion as a justification in seven out of the 13 cases. Lastly, the preference for over- versus under-segmentation was a widely used quality criterion, and was brought up by seven of the ten experts.

Agreement with quantitative segmentation quality metrics Some criteria we identified agreed with quantitative metrics that are commonly used to evaluate segmentation quality: We found that justifications that emphasized the importance of the margins and the distance between a false positive area and the primary lesion often agreed with the 95th percentile Hausdorff distance. When participants expressed their preference for over-segmentation, these decisions mostly disagreed with both the Hausdorff distance and the Dice score. The most crucial concept that did not agree with any quantitative metric was the preference for a segmentation because it was seen as biologically more plausible.

Experts' thought process while correcting segmentations

For the second part of our study, we asked each expert to manually correct four imperfect segmentations of post-operative brain glioblastomas during an interactive session with a semi-structured interview. The segmentations were generated using a deep learning algorithm. The quality of the proposed segmentations based on the

Table 5.2: Segmentation model performance based on the manual ground truth for the four cases study participants corrected during the interview

Case number	Dice score	HD95 [voxel]
1	0.81	4.57
2	0.68	66.24
3	0.73	7.00
4	0.85	6.00

available manual ground truth is listed in Table 5.2.

The thematic analysis of the interview transcripts was performed to identify the thought process during the correction process. A total of ten interviews were performed between September 30th and December 10th, 2021, resulting in 12.4 hours of recordings and 39 manually corrected segmentations. Additionally, we performed a quantitative analysis of the final corrected segmentations to determine the extent of the performed changes and the agreement between the experts' corrected segmentations.

We identified the following five themes with respect to correcting segmentations of post-operative brain tumors:

Postoperative brain tumors are highly ambiguous segmentation targets Given their biological background as a heavily infiltrative disease, the boundaries of GBMs are often unclear. All experts expressed uncertainty about where to draw the line between normal and abnormal intensity. One expert described it as follows: *"[...] GBM is a heavily infiltrative disease. So, the [...] imaging segmentation is a little bit of an ... I don't want to say artifact, but it's not exact. Yeah, it is a line in the sand that people draw."* Experts often referred to a "gray area" around the expected tumor margin.

Contextual information helps experts interpret the ambiguous presentation of the tumors on imaging Experts rely heavily on context to limit the ambiguity of the tumor margins described above. Contextual information is used to distinguish between abnormal findings on imaging that could be attributed to the brain tumor or other causes, e.g., white matter changes that could also be caused by hypertension or

radiation treatment. We identified three categories of context relevant to brain tumor segmentations: patient-related, application-related, and general context.

Patient-related context includes, for example, the patient's age, information about comorbidities, and treatment status. Additionally, access to a baseline image before the tumor diagnosis or before the start of treatment is thought to offer valuable information about the general brain health of a patient. Information about the clinical application a segmentation will be used for will influence certain corrections. For example, while correcting a segmentation intended for radiation treatment planning, experts would choose to be more conservative in an area close to radiation-sensitive structure. Other helpful contextual information participants identified are the patient's anatomy, knowledge about imaging artifacts, and the integration of information from different MRI sequences.

The final segmentation is shaped by personal beliefs and preferences However, not all ambiguities can be resolved through contextual information. Therefore, some subjective decisions must be made to finalize the segmentation. We found that each expert had their own philosophy influencing whether they tended to be more inclusive or exclusive in their segmentation corrections. We considered an expert who included voxels that are possibly but not certainly tumorous in the segmentation to display an inclusive philosophy. Experts who only selected voxels that they were certain to be tumorous were considered to be more exclusive in their segmentation approach. Participants also acknowledged that starting from a proposed segmentation possibly biased their decisions and the final corrected segmentation.

Due to the absence of objective quality criteria, consistency, within a case and between repeated imaging of the same patient, was often referred to as a proxy for segmentation quality. For longitudinal monitoring, ensuring that the segmentation is consistent between time points, is crucial for accurate treatment monitoring. Lastly, the amount of time and effort a physician invests into editing a segmentation will influence the outcome. One expert said that *"[. . .] you could drive yourself crazy, kind of fine-tuning these [segmentations]. And so, [. . .] picking and choosing your*

battles is the bottom line."

Role of deep learning algorithms in brain tumor segmentation Participants saw two potential advantages of the clinical deployment of deep learning algorithms: First, they expected segmentations generated by deep learning algorithms to be more consistent, as algorithms are not influenced by the time of the day, are not affected by time constraints, and cannot get distracted easily. Second, even if only a small routine step within a complicated task can get automated, this would free up healthcare workers to spend their energy on more challenging problems. However, our interview partners agreed that the technology is not yet fully ready to be clinically implemented. Particularly limitations in the ability to adapt to new and unseen scenarios and fringe cases are seen as a hurdle. In the near future, deep learning algorithms are seen as useful for narrowly defined tasks. *"I think that they're probably 80% there. [...] there still is work that needs to be done. But [...] getting that last 20% is gonna be really hard."*

Working with the output of deep learning algorithms during the interview, experts were highly aware of their potential bias toward the model's performance. One expert indicates that they are more cautious since, in their personal experience, deep learning algorithms for brain tumor segmentation are not very reliable: *"I know that the model output is not the greatest. [...]. It makes me rely more on myself than on the model"* However, others, given the highly ambiguous decisions they are facing, rely on the model's decisions when insecure about a particular area: *"I don't think, I would have chosen that, but I don't feel strongly enough against it to undo it considering that there maybe there was something in the algorithm making it choose that. This also happens with over-reading fellows or residents."*

Analysis of the corrected segmentations The quantitative analysis of the segmentations, corrected by the participants, showed large variability in the volume of changes each participant performed. Figure 5-11A illustrates the total changes that each participant performed during the interview. We also found that the agreement

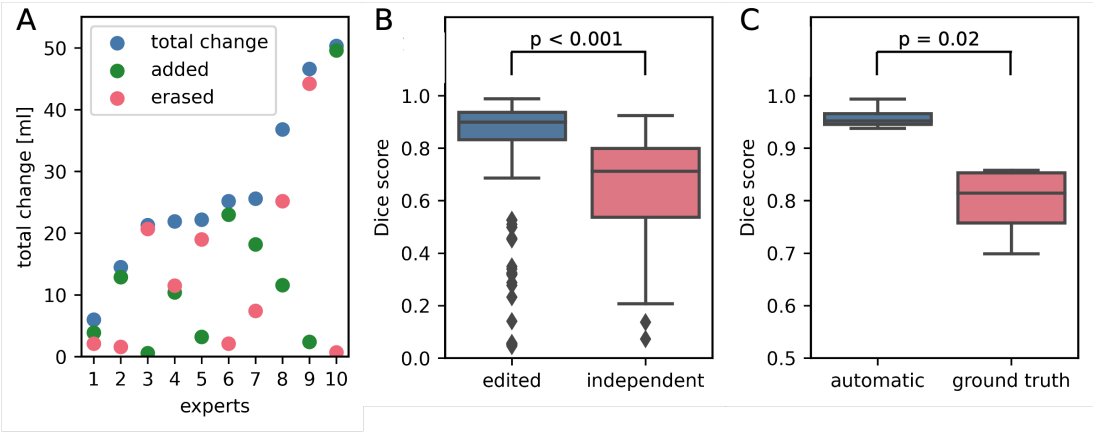

Figure 5-11: **Quantitative analysis of corrected segmentations.** A: Changes each expert performed during the interview: total volume of changes for all four cases (added to and erased from the proposed segmentation) (blue), added volume (green), and erased volume (red); B: Comparison of the agreement between all corrected segmentations and agreement between independently drawn manual segmentations on the same dataset. Blue: Pairwise Dice scores between the corrected segmentations (for each of the four cases); red: agreement between two independently drawn manual segmentations for a subset of the same post-operative brain tumor dataset; C: Comparison between the agreement of the corrected segmentations (aggregated using the STAPLE algorithm) with the automatic segmentation (blue) and the manual ground truth (red).

between the corrected segmentations of the participants was significantly higher than the inter-reader agreement previously identified for this dataset (Figure 5-11B, p-value < 0.001, Kruskal-Wallis test). Furthermore, the agreement between the corrected segmentations and the initial segmentation is significantly higher than between the corrected segmentations and the independently generated manual ground truth (Figure 5-11C, p-value $= 0.02$, Kruskall-Wallis test).

5.4.2 Discussion

In this section, we investigated how clinical experts perceive the quality of brain tumor segmentations from two angles: by identifying criteria that experts apply to evaluate segmentation quality and by elucidating experts' thought processes while correcting segmentations generated by a deep learning algorithm.

Criteria used to evaluate brain tumor segmentation quality Previous research has shown that the segmentation quality perception of experts does not agree with currently used quantitative metrics, like the Dice score and Hausdorff distance [50] (see Section 5.3). Through thematic analysis of the responses to the questionnaire, we have identified five quality criteria that influence how experts perceive the quality of brain tumor segmentation. While some quality criteria partially agree with the Dice score and Hausdorff distance, they depend more on the context of the segmentation error. For example, false positive errors should be treated differently to false negative ones.

Quality criteria can be contradictory Some of the criteria we identified here can be conflicting in certain cases. As depicted in Figure 5-3B, for example, a false positive connected to the primary lesion was perceived as more challenging to correct than a false positive of the same size in the contralateral hemisphere (correction effort) by some participants. At the same time, other participants considered the false positive area in the contralateral hemisphere a more consequential clinical segmentation error (distance). In these cases, individual preferences and application-specific criteria may come into play. This may also explain differences between the quality ratings of experts reported in Section 5.3.1.

Experts' thought processes while editing a segmentation Through an interview in which we tasked experts to correct segmentations generated by a deep learning algorithm, we identified four themes that describe essential characteristics of the correction process. Due to the ambiguous appearance of the tumor margins and the limited amount of information that can be derived from macroscopic imaging, the task of outlining brain tumors from MRI has several potentially acceptable solutions. The participating physicians extensively used context and their knowledge about tumor growth behavior, information that is not present in the imaging itself, to make decisions in uncertain situations. They also showed a high awareness of the ambiguity of the task and, therefore, would accept a reasonable segmentation even if they would

outline the tumor differently.

Influence of the proposed segmentation Furthermore, we found that due to the high ambiguity of the tumor margins, the proposed segmentation influences the outcome of the correction process. Our data indicates that in the absence of clear boundaries and indicators whether an area is part of the tumor or healthy tissue, experts tend to accept the proposed segmentation if it appeared reasonable to them. This finding explains the increased agreement between observers when they correct a proposed segmentation instead of perform a segmentation *de novo*. Due to the absence of an objective ground truth, it remains to be seen whether the improved agreement between experts will translate into improved clinical outcomes. The volume of changes that participants performed were additionally influenced by their perception of the capabilities of deep learning models. More research is needed to develop a detailed understanding of how editing decisions are influenced by a proposed segmentation and providers' inherent biases regarding a segmentation model's performance.

Importance of the clinical application Both parts of the present study reflected the importance of the downstream application for which a brain tumor segmentation is intended, as quality criteria highly dependent on it. However, this dependency is not reflected in current practices of segmentation quality evaluation. Participants in our study considered the clinical consequences of potential segmentation errors as they evaluated the quality of a segmentation, and these consequences differ between clinical applications [276, 305].

Additionally, there was consensus among our interview participants that artificial intelligence segmentation tools should be developed for narrowly-defined application areas, e.g., exclusively for use in radiation treatment planning. Based on our findings, we also recommend developing and clinically validating application-specific quantitative metrics. Those metrics should reflect the importance of the clinical consequences of specific errors. Furthermore, domain experts' evaluation of segmentation quality should be standardized through guidelines.

Generalizability Our study was focused explicitly on segmentations of glioblastomas, tumors known to be particularly challenging to manually segment on MRI due to their invasive growth behavior. Our findings may not generalize to other, less ambiguous segmentation structures. For cardiac segmentation, for example, it has been shown that experts' quality perception correlates well with conventional segmentation quality metrics, like the Dice score, Hausdorff distance, and sensitivity [306]. Therefore, there is no need for more contextual quality criteria for these segmentation tasks. However, we expect that the quality criteria and the thought processes we have identified here will partially generalize to other similarly challenging segmentation tasks, particularly other tumors.

Based on our findings presented in this section and Section 5.3, we conclude that the successful clinical deployment of deep learning-based brain tumors segmentation algorithm will require the following: To ensure that segmentation models' performance reflects clinical utility, improved quantitative metrics need to be developed. These metrics should reflect the context-dependent segmentation quality perception of experts. Second, the definition of a specific use case for each deep learning algorithm will limit the amount of ambiguity that physicians are facing while evaluating and correcting segmentations. Lastly, a more detailed understanding of why and how much physicians rely on the provided segmentation that they are tasked to correct will be needed.

5.4.3 Limitations

There are some limitations to this study. First, all participants who volunteered for the study were from one hospital system, and some were trainees of other participants. Experts from other institutions may value other segmentation quality criteria or show an approach to correcting segmentations that is not reflected in our findings. Furthermore, in clinical settings, tumor segmentation is a collaborative process. Physicians can ask their colleagues for advice if uncertain about specific aspects of a case [307]. We did not mimic this process in our interview study and did not provide patient- and treatment-related information about the cases. Therefore, participants faced more uncertainty, potentially forcing them to make unusual decisions.

5.4.4 Conclusions

Outlining high-grade brain tumors is highly challenging even for experts due to ambiguities in the tumor margins. Experts heavily rely on contextual information to limit uncertainties in segmenting brain tumors. Sources of contextual information can be information about the patient, e.g., their age, anatomical structures, or the application a segmentation is intended for. Furthermore, experts' decisions are influenced by their personal beliefs about whether ambiguous areas should be included or excluded from a segmentation and individual biases toward the performance of segmentation algorithms. This work highlights the need for specifying the application area for automatic segmentation algorithms during development and particularly during the evaluation of its performance. Lastly, new application-specific segmentation quality metrics that reflect clinical usefulness are needed.

5.5 Perspective on the evaluation of DL models for brain tumor segmentation

5.5.1 Efficient evaluation frameworks for segmentation models

We find a disconnect between the published research and clinical needs based on our extensive review of the literature on deep learning brain tumor segmentation and the quantitative and qualitative findings from our studies on postoperative brain tumor segmentation. Involving stakeholders in the evaluation and augmenting quantitative evaluation through qualitative accounts is expected to lead to the development of AI systems that fulfill their goal beyond meeting one target metric [218]. An evaluation process that involves multiple perspectives will help to align the optimization goals with the clinical needs. However, this process will require systematic support from scientific journals and regulatory bodies like the FDA.

Reporting checklists are an efficient tool to improve the completeness of scientific reports [308, 309]. While the CLAIM checklist asks for a failure case analysis, it does not require the use of multiple and complementary evaluation metrics and the

involvement of clinicians in the evaluation process [217] and should therefore be extended. Additionally, guidelines and checklists should be used early in the planning and execution of a study, not just in the final writing process [310]. Furthermore, we would welcome the FDA requiring the use of comprehensive metrics and the involvement of stakeholders in the evaluation process. For example, the first deep learning segmentation tool for high-grade glioma recently received FDA pre-market approval based on a performance analysis using the Dice score and volume as the only quality metrics [311].

5.5.2 Future directions

We envision two different lines of research resulting from our work on segmentation quality perception.

DL model development The first area concerns the development of improved deep-learning models for brain tumor segmentation. Based on our findings, we expect that the use of anatomical priors will improve the performance of brain tumor segmentation models. Due to GPU memory limitations, most 3D segmentation models are trained on small patches of the entire input image, leading to a loss of anatomical context from the models. The explicit use of anatomical context has improved segmentation for classical ML [312] and DL segmentation [313]. It remains to be seen if incorporating anatomical context can alleviate some error patterns our study participants have identified.

Furthermore, we anticipate an improved longitudinal segmentation consistency from the incorporation of inter-visit dependencies in the segmentation pipeline. We found that consistency between visits was highly valued for longitudinal follow-up, even beyond personal perceptions of how the "right" outline should look in highly ambiguous areas. Lastly, until new metrics have been developed and validated, a simple yet efficient strategy could be the combination of several quality metrics with different weights. For example, omissions at the margins should be weighted higher than in the middle of a segmentation.

Expert- and application-centered segmentation quality metrics In addition to the approaches that concern model development, we propose further research into segmentation quality metrics. These metrics should be centered around the needs of experts and focused on specific applications rather than represent general all-purpose metrics. Large-scale quantitative surveys can provide insight into the importance of the segmentation quality criteria identified in Section 5.4.1. Mainly, these metrics should be evaluated in the context of their application and using authentic images instead of illustrations to investigate their generalizability. Furthermore, we recommend extending the research approaches outlined in this thesis to other similarly ambiguous segmentation targets to identify synergies.

In conclusion, in this chapter we have presented how experts' qualitative segmentation quality perception differs from currently available segmentation quality metrics. We envision that our findings and the described research approaches will lead to efficient evaluation protocols for brain tumor segmentation models. These metrics will benefit the clinical translation of DL brain tumor segmentation models by supporting the development of segmentation models that generate predictions that satisfy the expectations of expert users.

Chapter 6

Conclusions

The research in this thesis is centered around three fundamental challenges to translating ML algorithms into clinical care settings: Robustness of ML in routine clinical settings, choosing appropriate modeling approaches, and clinically meaningful evaluation of model performance.

In Chapter 3, we presented how small decisions within a radiomics pipeline affect the test-retest repeatability of radiomics features. Furthermore, we highlighted research on methodological errors in the radiomics literature that inflate model performance, improving the repeatability of deep learning algorithms, and automatic estimation of segmentation quality. Algorithms and features alike are susceptible to design choices made in the process of an ML project. Our research on the stability of ML algorithms serves as a guide to best practices in the development of a clinically deployable algorithms. The development of stable algorithms will be essential in gaining the trust of physicians and patients for the clinical deployment of ML algorithms. Our research was retrospective and *in silico*; future research should focus on evaluating and, if required, improving the stability of ML algorithms prospectively and under real-world clinical conditions.

In Chapter 4, we demonstrated that the predictions of a deep learning algorithm reflect the actual distribution of the variable of interest only if the selected network

design is appropriate for the given task. In Section 4.4, we introduced a generalizable framework that can recover information lost by discretizing continuous variables. Our framework is closing the gap between the ordinal labels and the underlying continuous variable. This framework could support efforts to automatically identify patients at risk of deterioration by providing granular information about the dynamic of a disease.

Based on the concept of a latent continuous distribution that underlies discrete ordinal labels, Section 4.5 presented the first two methods for joint learning of ordinal classification and annotators' individual biases. Previously described approaches proved not appropriate for identifying differences between annotators' labeling behavior for a latent continuous variable. In the future, these methodologies can support the development of a complete model that characterizes the noise patterns observed in discrete ordinal labels.

In Chapter 5, we illustrated how performance evaluation of brain tumor segmentation models can be designed to reflect clincal usefulness.

In Section 5.3, we described the discrepancies between the subjective quality perception of clinical experts and the metrics typically used to evaluate performance. We found only a low to moderate agreement between the quantitative metrics and experts' ratings of segmentation quality. Furthermore, we demonstrated a high inter-rater variability in the perception of quality perception between experts, similar to the variability observed in manual segmentations. Various factors, such as the size of the segmentation target, influence these differences.

Through a multiple-method study focusing on qualitative accounts of experts' perception of segmentation quality, we identified criteria that experts use to evaluate the quality of automatically generated segmentations and describe their thought processes as they correct them. These findings are presented in Section 5.4. Experts assess the quality of brain tumor segmentation given the medical context. In particular, knowledge about the invasive growth behavior of high-grade brain tumors and the intended clinical use of the segmentation influence the quality assessment. Due to the high ambiguity of brain tumor outlines, experts extensively use contextual information

to decrease the amount of ambiguity. Personal preferences and the perception of the performance level of deep learning segmentation algorithms further influence decisions. As deep learning algorithms will be considered for clinical translation, new segmentation quality metrics targeting specific clinical needs will be needed.

Recommendations for medical ML practitioners

Despite the impressive number of publications on ML image analysis algorithms, it has yet to be shown that these algorithms can improve clinical practice. Based on the work described in this thesis on three fundamental challenges for the translation of ML algorithms into clinical care, we deduce the following recommendations for ML practitioners:

1. Developing practical ML tools for image analysis requires a detailed understanding of the clinical need. This understanding will inform the choice of the right ML approach, image preprocessing, annotation, and algorithm evaluation.

2. Best practices and guidelines should be used throughout a project. Research is rarely a linear process, and while going back and forth between different approaches, researchers should remind themselves to adhere to a fixed set of best practices.

3. Before embarking on the model development process, it is advisable to consider whether simply reproducing a task as humans perform it is the right approach. There may be other ways to solve the problem at hand in a way humans would not be able to. For example, as described in Sections 3.7.2 and 4.4, nominal classification approaches can be successfully used to predict discrete ordinal labels. However, they are characterized by low repeatability and do not accurately reflect the underlying continuous distribution of the target variable.

4. A good understanding and evaluation of the noise patterns in the data will help in the training and evaluation process. As we have demonstrated in Section 4.5,

noise patterns can be learned and reveal critical information about the annotators who generated the training labels.

5. The influence of every design choice on model stability should be tested. Even if an algorithm achieves good classification performance, this does not necessarily correspond to robustness towards small perturbations in the data. In the absence of test-retest data, data augmentation, e.g., horizontal flips or small rotations, can serve as an instrument to test the repeatability of an algorithm's predictions.

6. Clinical experts should be involved in the evaluation. As described in Section 5.4, qualitative accounts can provide valuable information for further evaluation of algorithms. Algorithms should be evaluated with a focus on the intended clinical application and, even if retrospectively, under conditions as close to a realistic clinical setting as possible.

I want to close with the words of one of the participants in the interview study on the quality perception of brain tumor segmentations: *"I think that [AI algorithms are] probably 80% there. [...] there still is work that needs to be done. But [...] getting that last 20% is gonna be really hard."*

Appendix A

Continuous scores - Appendix

A.1 Dataset label distributions

List of label distributions for each dataset.

Retinopathy of prematurity Dataset size: 5511 images

- Normal: 4535 images (82·3%)
- Pre-plus disease: 804 images (14·6%)
- Plus disease: 172 images (3·1%)

Knee osteoarthritis (OA) Dataset size: 14173 images

- No OA (KL 0): 5793 images (40.9·%)
- Doubtful OA (KL 1): 2156 images (15·2%)
- Mild OA (KL 2): 2355 images (16·6%)
- Moderate OA (KL 3): 2604 images (18·4%)
- Severe OA (KL 4): 1265 images (8·9%)

Breast density Dataset size: 108230 images

- Fatty: 12428 images (11·5%)
- Scattered: 47909 images (44·2%)
- Heterogeneously dense: 41325 images (38·2%)
- Dense: 6568 images (6·1%)

A.2 Predicted rank vs. ground truth rank

Figure A-1 contains the same data from Figure 4-6 presented in Section 4.4.1. The predicted scores were ordered to determine a rank and were plotted against the expert's ranks. The MSE displayed in Table 4.3 was computed on these two variables. Since a linear correlation is expected on the rank-to-rank analysis, the Pearson coefficient was used.

A.3 Pair-wise statistical comparisons

Only the metrics showing no statistical differences between two metrics were included (only some metrics from Table 4.3). If no figure for a specific metric and dataset is present, it means all the pair-wise comparisons showed a statistical difference. All metrics presented in Table 4.3, Figure 4-7, Figure 4-8, and Figure 4-9 were analysed.

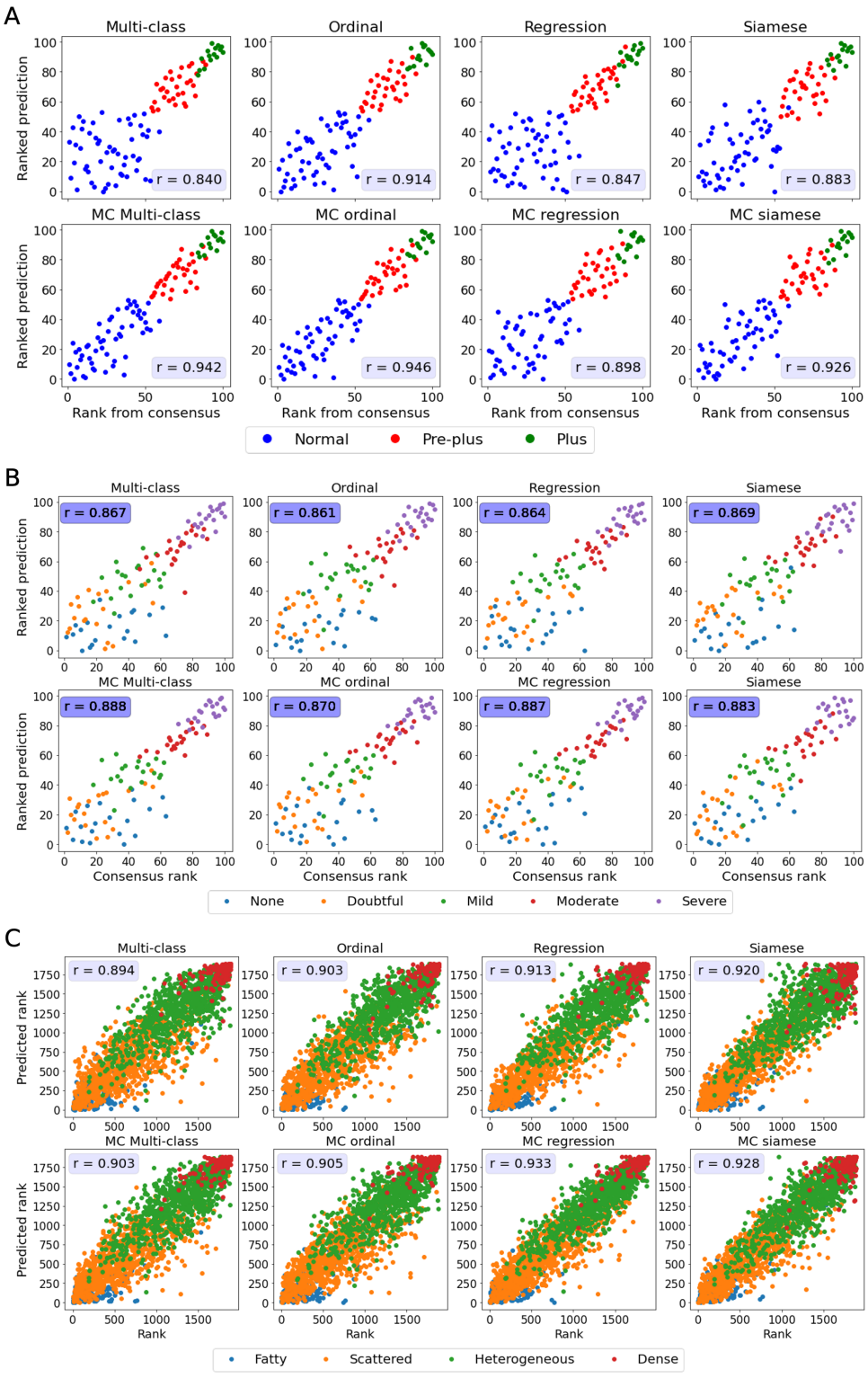

Figure A-1: **Correspondence between model predicted rank and true severity rank.** A: Retinopathy of prematurity; B: Knee osteoarthritis; C: Breast density. For each model the Pearson correlation coefficient (r) is displayed and indicate the strength of the linear correlation where 1 is a perfectly positive linear correlation and -1 a perfectly negative linear correlation.

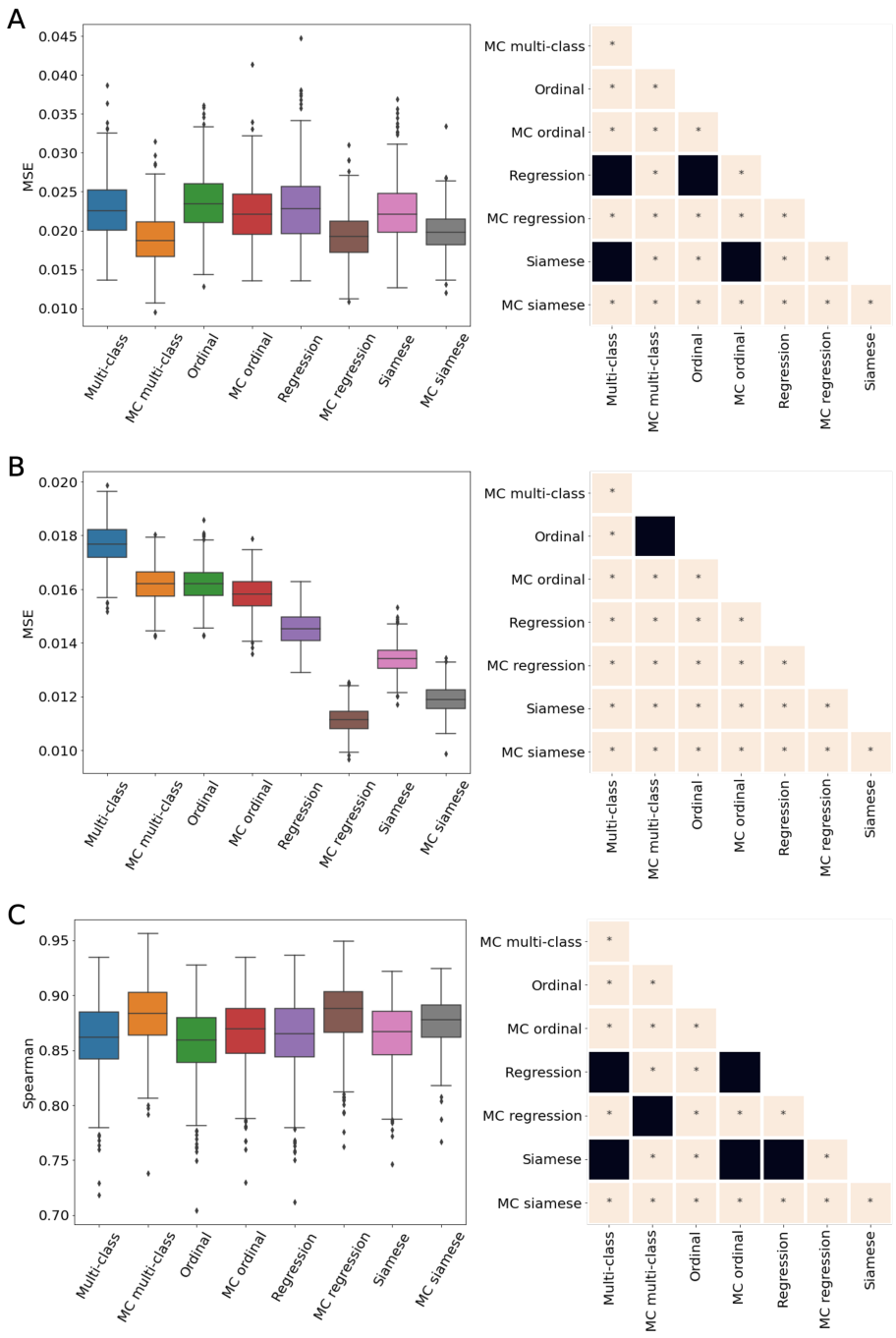

Figure A-2: **Pair-wise statistical comparisons for MSE, Spearman, and AUC metrics (metric - dataset).** A/C: Knee osteoarthritis (A: MSE, C: Spearman correlation coefficinet); B: Breast density. For each metric on a given test set, each pairs of models (MC and non-MC multi-class, ordinal, regression, Siamese) was compared. The box plots on the left side displays the value range of the MSE and Spearman correlation coefficient, respectively, obtained through 500 bootstraps. The grid on the right side includes the 28 pair-wise comparisons. * means that a statistical difference ($p - value < 0.05$ on a two-sided t-test) was reached while a black square indicates no statistical differences. Only metrics where at least one pair had no statistical difference was presented.

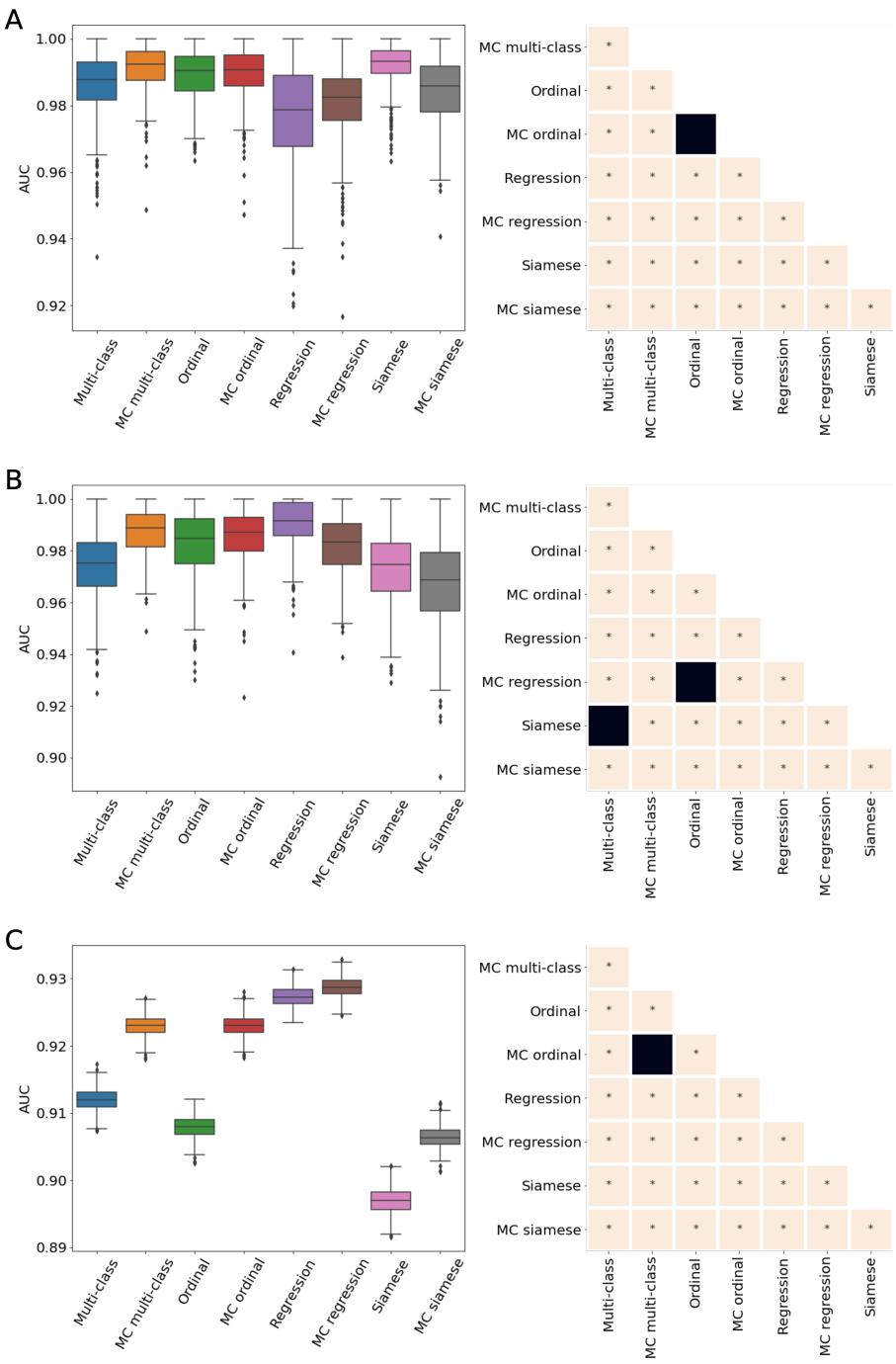

Figure A-3: **Pair-wise statistical comparisons for MSE, Spearman, and AUC metrics (metric - dataset).** A: Retinopathy of prematurity; B: Knee osteoarthritis ; C: Breast density. For each metric on a given test set, each pairs of models (MC and non-MC multi-class, ordinal, regression, Siamese) was compared. The box plots on the left side displays the value range of the AUROC, respectively, obtained through 500 bootstraps. The grid on the right side includes the 28 pair-wise comparisons. * means that a statistical difference ($p - value < 0.05$ on a two-sided t-test) was reached while a black square indicates no statistical differences. Only metrics where at least one pair had no statistical difference was presented.

Appendix B

Learning the bias of individual annotators from single ordinal labels - Appendix

B.1 Generation of the synthetic dataset

Each image was 400 x 400 pixels in size. The background and foreground colors were independently sampled as RGB tuples from a uniform distribution ranging from 0 to 255. We defined the length of each square, the radius of its corners, and its rotation by sampling from uniform distributions with the following ranges:

- length: $[10, 380]$ (pixels)
- radius: $[0.001, \frac{\text{length}}{2}]$ (pixels)
- rotation: $[0, 89]$ (degrees)

Furthermore, the squares were placed within the image such that the full square fit and no parts were cut off. We generated 10000 images with rounded rectangles. The distribution of the continuously valued ground truth, the roundness of the squares, is depicted in Figure

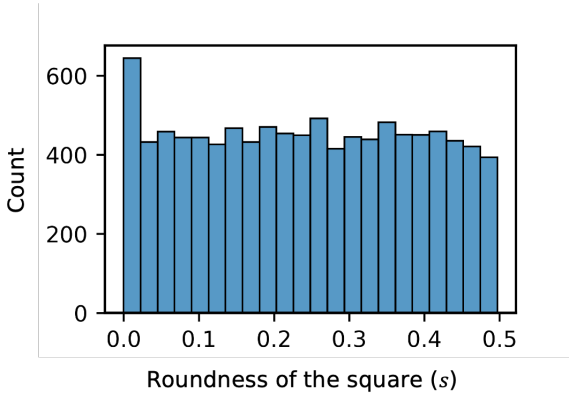

Figure B-1: **Distribution of the synthetic dataset.** Histogram of the continuously valued ground truth, the roundness of the squares, in the synthetic dataset.

Bibliography

[1] Wald, C. *et al.* AI Central (2022). URL https://aicentral.acrdsi.org/.

[2] Tai, M. C. T. The impact of artificial intelligence on human society and bioethics. *Tzu-Chi Medical Journal* **32**, 339 (2020). URL /pmc/articles/PMC7605294//pmc/articles/PMC7605294/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7605294/.

[3] Lynch, S. Andrew Ng: Why AI Is the New Electricity. URL https://www.gsb.stanford.edu/insights/andrew-ng-why-ai-new-electricity.

[4] The world's most valuable resource is no longer oil, but data. URL https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data.

[5] Copeland, B. artificial intelligence (2021). URL https://www.britannica.com/technology/artificial-intelligence.

[6] Wallis, C. How Artificial Intelligence Will Change Medicine. *Nature* **576**, S48 (2019).

[7] Beriault, D. R., Gilmour, J. A. & Hicks, L. K. Overutilization in laboratory medicine: tackling the problem with quality improvement science. *Critical Reviews in Clinical Laboratory Sciences* **58**, 430–446 (2021).

[8] Smith-Bindman, R. *et al.* Trends in Use of Medical Imaging in US Health Care Systems and in Ontario, Canada, 2000-2016. *JAMA - Journal of the American Medical Association* **322**, 843–856 (2019). URL http://www.ncbi.nlm.nih.gov/pubmed/31479136http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6724186.

[9] Dall, T., Ryenolds, R., Chakrabarti, R., Jones, K. & Iacobucci, W. The Complexities of Physician Supply and Demand: Projections from 2018 to 2033 Final Report. *Association of American Medical Colleges* 1–59 (2020).

[10] Shanafelt, T. D. *et al.* Changes in Burnout and Satisfaction With Work-Life Integration in Physicians Over the First 2 Years of the COVID-19 Pandemic. *Mayo Clinic Proceedings* 1–11 (2022). URL https://doi.org/10.1016/j.mayocp.2022.09.002.

[11] Willis, M., Duckworth, P., Coulter, A., Meyer, E. T. & Osborne, M. The Future of Health Care: Protocol for Measuring the Potential of Task Automation Grounded in the National Health Service Primary Care System. *JMIR Research Protocols* **8** (2019).

[12] Berwick, D. M., Nolan, T. W. & Whittington, J. The triple aim: Care, health, and cost (2008). URL `http://www.healthaffairs.org/doi/10.1377/hlthaff.27.3.759`.

[13] Beam, A. L. & Kohane, I. S. Translating artificial intelligence into clinical care. *JAMA - Journal of the American Medical Association* **316**, 2368–2369 (2016).

[14] Gunasekeran, D. V., Ting, D. S., Tan, G. S. & Wong, T. Y. Artificial intelligence for diabetic retinopathy screening, prediction and management. *Current Opinion in Ophthalmology* **31**, 357–365 (2020).

[15] Williams, D., Hornung, H., Nadimpalli, A. & Peery, A. Deep Learning and its Application for Healthcare Delivery in Low and Middle Income Countries. *Frontiers in Artificial Intelligence* **4**, 30 (2021). URL `https://www.frontiersin.org/articles/10.3389/frai.2021.553987/full`.

[16] World Health Organisation. *Ethics and governance of artificial intelligence for health: WHO guidance. Geneva:*. October (2021).

[17] Liu, J. *et al.* A review of arrhythmia detection based on electrocardiogram with artificial intelligence. *Expert Review of Medical Devices* **19**, 549–560 (2022). URL `https://doi.org/10.1080/17434440.2022.2115887`.

[18] Alhanai, T., Ghassemi, M. & Glass, J. Detecting depression with audio/text sequence modeling of interviews. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* **2018-Septe**, 1716–1720 (2018).

[19] Yang, Y. *et al.* Artificial intelligence-enabled detection and assessment of Parkinson's disease using nocturnal breathing signals. *Nature Medicine* **28** (2022).

[20] Mun, S. K., Freedman, M. & Kapur, R. Image management and communications for radiology. *IEEE Engineering in Medicine and Biology Magazine* **12**, 70–80 (1993).

[21] Alhajeri, M., Aldosari, H. & Aldosari, B. Evaluating latest developments in PACS and their impact on radiology practices: a systematic literature review. *Informatics in Medicine Unlocked* **181-190** (9).

[22] Nishikawa, R. M. *et al.* Initial experience with a prototype clinical intelligent mammography workstation for computer-aided diagnosis. In *SPIE Medical Imaging 1995: Image Processing*, 65–71 (Society of Photo-Optical Instrumentation Engineers, 1995).

[23] Mayo, R. C. & Leung, J. Artificial intelligence and deep learning – Radiology's next frontier? *Clinical Imaging* **49**, 87–88 (2018). URL https://doi.org/10.1016/j.clinimag.2017.11.007.

[24] Ranschaert, E., Topff, L. & Pianykh, O. Optimization of Radiology Workflow with Artificial Intelligence. *Radiologic Clinics of North America* **59**, 955–966 (2021). URL https://doi.org/10.1016/j.rcl.2021.06.006.

[25] Liang, S. *et al.* Magnetic Resonance Imaging Sequence Identification Using a Metadata Learning Approach. *Frontiers in Neuroinformatics* **15**, 1–10 (2021).

[26] Zhu, B., Liu, J. Z., Cauley, S. F., Rosen, B. R. & Rosen, M. S. Image reconstruction by domain-transform manifold learning. *Nature* **555**, 487–492 (2018). URL http://www.nature.com/articles/nature25988.

[27] Yuan, J., Liao, H., Luo, R. & Luo, J. Automatic Radiology Report Generation Based on Multi-view Image Fusion and Medical Concept Enrichment. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **11769 LNCS**, 721–729 (2019).

[28] Raymond Geis, J. *et al.* Ethics of artificial intelligence in radiology: Summary of the joint European and North American multisociety statement. *Radiology* **293**, 436–440 (2019). URL https://pubs.rsna.org/doi/10.1148/radiol.2019191586.

[29] Mendelson, E. B. Artificial Intelligence in Breast Imaging: Potentials and Limitations. *https://doi.org/10.2214/AJR.18.20532* **212**, 293–299 (2018). URL www.ajronline.org.

[30] Nestor, B. *et al.* Rethinking clinical prediction: Why machine learning must consider year of care and feature aggregation 1–7 (2018). URL http://arxiv.org/abs/1811.12583.

[31] Subbaswamy, A. & Saria, S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics (Oxford, England)* **21**, 345–352 (2020).

[32] Barak-Corren, Y. *et al.* Prediction across healthcare settings: a case study in predicting emergency department disposition. *npj Digital Medicine* **4** (2021).

[33] Van Calster, B. *et al.* Calibration: The Achilles heel of predictive analytics. *BMC Medicine* **17**, 1–7 (2019).

[34] Shah, N. H., Milstein, A. & Bagley, S. C. Making Machine Learning Models Clinically Usefyl. *JAMA - Journal of the American Medical Association* **322**, 1351–1352 (2019).

[35] Keane, P. A. & Topol, E. J. With an eye to AI and autonomous diagnosis. *npj Digital Medicine* **1**, 10–12 (2018). URL http://dx.doi.org/10.1038/s41746-018-0048-y.

[36] Raunig, D. L. *et al.* Quantitative imaging biomarkers: A review of statistical methods for technical performance assessment (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/24919831http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5574197.

[37] Sameera, V., Bindra, A. & Rath, G. P. Human errors and their prevention in healthcare. *Journal of Anaesthesiology Clinical Pharmacology* **37**, 328–335 (2021).

[38] Merisaari, H. *et al.* Repeatability of radiomics and machine learning for Diffusion Weighted Imaging: Short-term repeatability study of 112 patients with prostate cancer. *Magn Reson Med.* **83**, 2293–2309 (2020).

[39] Kim, H., Park, C. M. & Goo, J. M. Test-retest reproducibility of a deep learning–based automatic detection algorithm for the chest radiograph. *European Radiology* **30**, 2346–2355 (2020).

[40] Raisi-Estabragh, Z. *et al.* Repeatability of Cardiac Magnetic Resonance Radiomics: A Multi-Centre Multi-Vendor Test-Retest Study. *Frontiers in Cardiovascular Medicine* **7**, 1–16 (2020).

[41] Campbell, J. P. *et al.* Plus Disease in Retinopathy of Prematurity: A Continuous Spectrum of Vascular Abnormality as a Basis of Diagnostic Variability. *Ophthalmology* **123**, 2338–2344 (2016). URL https://linkinghub.elsevier.com/retrieve/pii/S0161642016307321.

[42] Greenland, S. Invited Commentary: The Need for Cognitive Science in Methodology. *American Journal of Epidemiology* **186**, 639–645 (2017). URL https://academic.oup.com/aje/article/186/6/639/3886035.

[43] Amini, M., Pedram, M., Moradi, A. & Ouchani, M. Diagnosis of Alzheimer's Disease Severity with fMRI Images Using Robust Multitask Feature Extraction Method and Convolutional Neural Network (CNN). *Computational and mathematical methods in medicine* **2021**, 5514839 (2021). URL http://www.ncbi.nlm.nih.gov/pubmed/34007305http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC8100410.

[44] Grassmann, F. *et al.* A Deep Learning Algorithm for Prediction of Age-Related Eye Disease Study Severity Scale for Age-Related Macular Degeneration from Color Fundus Photography. *Ophthalmology* **125**, 1410–1420 (2018). URL http://www.ncbi.nlm.nih.gov/pubmed/29653860.

[45] Brown, J. M. *et al.* Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks. *JAMA Ophthalmology* **136**, 803–810 (2018).

[46] Chang, K. *et al.* Multi-Institutional Assessment and Crowdsourcing Evaluation of Deep Learning for Automated Classification of Breast Density. *Journal of the American College of Radiology* **17**, 1653–1662 (2020). URL https://doi.org/10.1016/j.jacr.2020.05.015https://www.jacr.org/action/showPdf?pii=S1546-1440%2820%2930539-1.

[47] Kalpathy-Cramer, J. *et al.* Plus Disease in Retinopathy of Prematurity: Improving Diagnosis by Ranking Disease Severity and Using Quantitative Image Analysis. *Ophthalmology* **123**, 2345–2351 (2016).

[48] Wang, Z., Wang, E. & Zhu, Y. Image segmentation evaluation: a survey of methods. *Artificial Intelligence Review* **53**, 5637–5674 (2020). URL https://doi.org/10.1007/s10462-020-09830-9.

[49] Shi, R. *et al.* Human perception-based evaluation criterion for ultra-high resolution cell membrane segmentation (2020).

[50] Kofler, F. *et al.* Are we using appropriate segmentation metrics? Identifying correlates of human expert perception for CNN training beyond rolling the DICE coefficient (2021). URL https://arxiv.org/pdf/2103.06205.pdf.

[51] Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Proceedings of the 25th International Conference on Neu- ral Information Processing Systems* 1097–1105 (2012). URL http://code.google.com/p/cuda-convnet/.

[52] Qureshi, T. A., Habib, M., Hunter, A. & Al-Diri, B. A manually-labeled, artery/vein classified benchmark for the DRIVE dataset. *Proceedings of CBMS 2013 - 26th IEEE International Symposium on Computer-Based Medical Systems* 485–488 (2013).

[53] Zbontar, J. *et al.* fastMRI: An Open Dataset and Benchmarks for Accelerated MRI .

[54] Arindra Adiyoso Setio, A. *et al.* Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge (2017). URL https://luna16.grand-challenge.org/.

[55] Zech, J. R. *et al.* Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Medicine* **15**, 1–17 (2018).

[56] Lambin, P. *et al.* Radiomics: Extracting more information from medical images using advanced feature analysis. *European Journal of Cancer* **48**, 441–446 (2012). URL http://www.ncbi.nlm.nih.gov/pubmed/22257792http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4533986.

[57] Kumar, V. *et al.* QIN "Radiomics: The Process and the Challenges". *Magn Reson Imaging* **30**, 1234–1248 (2012).

[58] Guiot, J. *et al.* A review in radiomics: Making personalized medicine a reality via routine imaging. *Medicinal Research Reviews* **42**, 426–440 (2022).

[59] Li, Z. C. *et al.* Multiregional radiomics profiling from multiparametric MRI: Identifying an imaging predictor of IDH1 mutation status in glioblastoma. *Cancer Medicine* **7**, 5999–6009 (2018). URL http://www.ncbi.nlm.nih.gov/pubmed/30426720http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6308047http://doi.wiley.com/10.1002/cam4.1863.

[60] Coroller, T. P. *et al.* Radiomic-Based Pathological Response Prediction from Primary Tumors and Lymph Nodes in NSCLC. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer* **12**, 467–476 (2017). URL https://pubmed.ncbi.nlm.nih.gov/27903462/.

[61] Papp, L. *et al.* Glioma Survival Prediction with Combined Analysis of In Vivo 11C-MET PET Features, Ex Vivo Features, and Patient Features by Supervised Machine Learning. *Journal of nuclear medicine : official publication, Society of Nuclear Medicine* **59**, 892–899 (2018). URL https://pubmed.ncbi.nlm.nih.gov/29175980/.

[62] Bleker, J. *et al.* A deep learning masked segmentation alternative to manual segmentation in biparametric MRI prostate cancer radiomics. *European radiology* **32**, 6526–6535 (2022). URL https://pubmed.ncbi.nlm.nih.gov/35420303/.

[63] Zwanenburg, A. *et al.* The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology* **295**, 328–338 (2020). URL https://pubs.rsna.org/doi/10.1148/radiol.2020191145.

[64] Laine, A. & Fan, J. Texture Classification by Wavelet Packet Signatures. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **15**, 1186–1191 (1993).

[65] Fusco, R. *et al.* Radiomics in medical imaging: pitfalls and challenges in clinical management. *Japanese Journal of Radiology* **40**, 919–929 (2022). URL https://doi.org/10.1007/s11604-022-01271-4.

[66] van Timmeren, J. E., Leijenaar, R. T., van Elmpt, W., Reymen, B. & Lambin, P. Feature selection methodology for longitudinal cone-beam CT radiomics. *Acta oncologica (Stockholm, Sweden)* **56**, 1537–1543 (2017). URL https://pubmed.ncbi.nlm.nih.gov/28826307/.

[67] Radovic, M., Ghalwash, M., Filipovic, N. & Obradovic, Z. Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC Bioinformatics* **18**, 1–14 (2017). URL http://dx.doi.org/10.1186/s12859-016-1423-9.

[68] Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001).

[69] Boser, B. E., Guyon, I. M. & Vapnik, V. N. Training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory* 144–152 (1992).

[70] Aerts, H. J. *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications* **5**, 4006 (2014). URL `http://www.nature.com/articles/ncomms5006`.

[71] Leijenaar, R. T. *et al.* Development and validation of a radiomic signature to predict HPV (p16) status from standard CT imaging: A multicenter study. *British Journal of Radiology* **91**, 1–8 (2018).

[72] Cheng, K., Lin, A., Yuvaraj, J., Nicholls, S. J. & Wong, D. T. Cardiac computed tomography radiomics for the non-invasive assessment of coronary inflammation. *Cells* **10**, 1–17 (2021).

[73] Elkilany, A. *et al.* A radiomics-based model to classify the etiology of liver cirrhosis using gadoxetic acid-enhanced MRI. *Scientific Reports* **11**, 1–13 (2021). URL `https://doi.org/10.1038/s41598-021-90257-9`.

[74] Tomaszewski, M. R. & Gillies, R. J. The biological meaning of radiomic features. *Radiology* **298**, 505–516 (2021).

[75] Dercle, L. *et al.* Radiomics Response Signature for Identification of Metastatic Colorectal Cancer Sensitive to Therapies Targeting EGFR Pathway. *JNCI Journal of the National Cancer Institute* **112**, 902 (2020). URL `/pmc/articles/PMC7492770//pmc/articles/PMC7492770/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7492770/`.

[76] Chen, H. *et al.* Reproducibility of radiomics features derived from intravoxel incoherent motion diffusion-weighted MRI of cervical cancer. *https://doi.org/10.1177/0284185120934471* **62**, 679–686 (2020). URL `https://journals.sagepub.com/doi/full/10.1177/0284185120934471`.

[77] McHugh, D. J. *et al.* Image Contrast, Image Pre-Processing, and T1 Mapping Affect MRI Radiomic Feature Repeatability in Patients with Colorectal Cancer Liver Metastases. *Cancers* **13**, 1–21 (2021). URL `https://pubmed.ncbi.nlm.nih.gov/33440685/`.

[78] Um, H. *et al.* Impact of image preprocessing on the scanner dependence of multi-parametric MRI radiomic features and covariate shift in multi-institutional glioblastoma datasets. *Physics in Medicine and Biology* **64** (2019). URL `https://pubmed.ncbi.nlm.nih.gov/31272093/`.

[79] Yip, S. S. & Aerts, H. J. Applications and limitations of radiomics (2016). URL `http://stacks.iop.org/0031-9155/61/i=13/a=R150?key=crossref.134478778713970aff90f16abe110608`.

[80] Liu, R. *et al.* Stability analysis of CT radiomic features with respect to segmentation variation in oropharyngeal cancer. *Clinical and Translational Radiation Oncology* **21**, 11–18 (2020). URL https://doi.org/10.1016/j.ctro.2019.11.005.

[81] Poirot, M. G. *et al.* Robustness of radiomics to variations in segmentation methods in multimodal brain MRI. *Scientific Reports 2022 12:1* **12**, 1–10 (2022). URL https://www.nature.com/articles/s41598-022-20703-9.

[82] Echegaray, S. *et al.* Core samples for radiomics features that are insensitive to tumor segmentation: method and pilot study using CT images of hepatocellular carcinoma. *Journal of Medical Imaging* **2**, 041011 (2015). URL /pmc/articles/PMC4650964//pmc/articles/PMC4650964/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4650964/.

[83] Kalpathy-Cramer, J. *et al.* Radiomics of Lung Nodules: A Multi-Institutional Study of Robustness and Agreement of Quantitative Imaging Features. *Tomography* **2**, 430–437 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/28149958http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5279995http://digitalpub.tomography.org/i/763956-vol-2-no-4-dec-2016/199.

[84] Yagis, E. *et al.* Effect of data leakage in brain MRI classification using 2D convolutional neural networks. *Scientific Reports* **11**, 1–13 (2021). URL https://doi.org/10.1038/s41598-021-01681-w.

[85] Sun, Q. *et al.* Deep Learning vs. Radiomics for Predicting Axillary Lymph Node Metastasis of Breast Cancer Using Ultrasound Images: Don't Forget the Peritumoral Region. *Frontiers in Oncology* **10**, 53 (2020). URL /pmc/articles/PMC7006026//pmc/articles/PMC7006026/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7006026/.

[86] Peng, Y. *et al.* Pretreatment DCE-MRI-Based Deep Learning Outperforms Radiomics Analysis in Predicting Pathologic Complete Response to Neoadjuvant Chemotherapy in Breast Cancer. *Frontiers in Oncology* **12**, 1 (2022). URL /pmc/articles/PMC8960929//pmc/articles/PMC8960929/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC8960929/.

[87] Truhn, D. *et al.* Radiomic versus Convolutional Neural Networks Analysis for Classification of Contrast-enhancing Lesions at Multiparametric Breast MRI. *Radiology* **290**, 290–297 (2019). URL https://pubs-rsna-org.ezp-prod1.hul.harvard.edu/doi/10.1148/radiol.2018181352.

[88] Fukushima, K. Biological Cybernetics Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. *Biol. Cybernetics* **36**, 202 (1980).

[89] LeCun, Y. *et al.* Backpropagation applied to digit recognition (1989). URL https://www.ics.uci.edu/~welling/teaching/273ASpring09/lecun-89e.pdf.

[90] Gulshan, V. *et al.* Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* **316**, 2402 (2016). URL http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2016.17216.

[91] Shankar, K. *et al.* Automated detection and classification of fundus diabetic retinopathy images using synergic deep learning model. *Pattern Recognition Letters* **133**, 210–216 (2020). URL https://www.sciencedirect.com/science/article/abs/pii/S0167865520300714.

[92] Tiulpin, A., Thevenot, J., Rahtu, E., Lehenkari, P. & Saarakkala, S. Automatic knee osteoarthritis diagnosis from plain radiographs: A deep learning-based approach. *Scientific Reports* **8**, 1727 (2018). URL http://www.nature.com/articles/s41598-018-20132-7.

[93] von Schacky, C. E. *et al.* Development and validation of a multitask deep learning model for severity grading of hip osteoarthritis features on radiographs. *Radiology* **295**, 139–145 (2020). URL https://doi.org/10.1148/radiol.2020190925.

[94] Kieu, S. T. H., Bade, A., Hijazi, M. H. A. & Kolivand, H. A Survey of Deep Learning for Lung Disease Detection on Medical Images: State-of-the-Art, Taxonomy, Issues and Future Directions. *Journal of Imaging* **6**, 131 (2020). URL https://www.mdpi.com/2313-433X/6/12/131.

[95] Miao, S., Wang, Z. J. & Liao, R. A CNN Regression Approach for Real-Time 2D/3D Registration. *IEEE Transactions on Medical Imaging* **35**, 1352–1363 (2016).

[96] Shin, H. C. *et al.* Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Transactions on Medical Imaging* **35**, 1285–1298 (2016).

[97] Menze, B. H. *et al.* The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE transactions on medical imaging* **34**, 1993–2024 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/25494501.

[98] Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J. & Dalca, A. V. VoxelMorph: A Learning Framework for Deformable Medical Image Registration. *IEEE Transactions on Medical Imaging* **38**, 1788–1800 (2019).

[99] Sutskever, I. *et al.* Learning phrase representation using RNN Encoder-Decoder. *Proceedings of the Empherical Methods in Natural Language Processing* **4**, 1724–1734 (2014).

[100] Ronneberger, O., Fischer, P. & Thomas, B. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, vol. 9351, 234–241 (2015). URL `http://lmb.informatik.uni-freiburg.de/http://arxiv.org/abs/1505.04597`.

[101] Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255 (Ieee, 2009).

[102] Krizhevsky, A. Learning Multiple Layers of Features from Tiny Images (2009).

[103] Power, A., Burda, Y., Edwards, H., Babuschkin, I. & Misra, V. Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets 1–10 (2022). URL `http://arxiv.org/abs/2201.02177`.

[104] Perez, L. & Wang, J. The Effectiveness of Data Augmentation in Image Classification using Deep Learning (2017). URL `https://arxiv.org/abs/1712.04621v1`.

[105] Shorten, C. & Khoshgoftaar, T. M. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data* **6**, 1–48 (2019). URL `https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0197-0`.

[106] Pereira, S., Pinto, A., Alves, V. & Silva, C. A. Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images. *IEEE Transactions on Medical Imaging* **35**, 1240–1251 (2016).

[107] Wang, P. & Cheng, J. Accelerating convolutional neural networks for mobile applications. *MM 2016 - Proceedings of the 2016 ACM Multimedia Conference* 541–545 (2016).

[108] Ghosh, A., Kumar, H. & Sastry, P. S. Robust loss functions under label noise for deep neural networks. *31st AAAI Conference on Artificial Intelligence, AAAI 2017* 1919–1925 (2017).

[109] You, X. x., Liang, Z. m., Wang, Y. q. & Zhang, H. A study on loss function against data imbalance in deep learning correction of precipitation forecasts. *Atmospheric Research* **281**, 106500 (2023). URL `https://doi.org/10.1016/j.atmosres.2022.106500`.

[110] Wang, F., Kaushal, R. & Khullar, D. Should Health Care Demand Interpretable Artificial Intelligence or Accept "Black Box" Medicine? *Annals of Internal Medicine* **172**, 59–60 (2020).

[111] Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).

[112] Krupinski, E. A. Current perspectives in medical image perception. *Attention, Perception, & Psychophysics* **72**, 1205–1217 (2010).

[113] Peabody, J. W., Luck, J., Glassman, P., Dresselhaus, T. R. & Lee, M. Comparison of Vignettes, Standardized Patients, and Chart Abstraction: A Prospective Validation Study of 3 Methods for Measuring Quality. *JAMA* **283**, 1715–1722 (2000). URL https://jamanetwork-com.ezp-prod1.hul.harvard.edu/journals/jama/fullarticle/192552.

[114] Clauser, P., Dietzel, M., Weber, M., Kaiser, C. G. & Baltzer, P. A. Motion artifacts, lesion type, and parenchymal enhancement in breast MRI: what does really influence diagnostic accuracy? *Acta Radiologica* **60**, 19–27 (2019). URL https://journals-sagepub-com.ezp-prod1.hul.harvard.edu/doi/full/10.1177/0284185118770918.

[115] Rajpurkar, P. *et al.* Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Medicine* **15** (2018).

[116] Johnson, A. E. *et al.* MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data* **6**, 1–8 (2019). URL http://dx.doi.org/10.1038/s41597-019-0322-0.

[117] Majkowska, A. *et al.* Chest Radiograph Interpretation with Deep Learning Models: Assessment with Radiologist-adjudicated Reference Standards and Population-adjusted Evaluation. *Radiology* **294**, 421–431 (2020). URL https://pubs.rsna.org/doi/pdf/10.1148/radiol.2019191293.

[118] Brenner, R. J., Lucey, L. L., Smith, J. J. & Saunders, R. Radiology and medical malpractice claims: A report on the practice standards claims survey of the Physician Insurer Association of America and the American College of Radiology (1998). URL www.ajronline.org.

[119] Pinto, A. *et al.* Learning From Errors in Radiology: A Comprehensive Review. *Seminars in Ultrasound, CT and MRI* **33**, 379–382 (2012). URL http://dx.doi.org/10.1053/j.sult.2012.01.015.

[120] Olatunji, T., Yao, L., Covington, B., Rhodes, A. & Upton, A. Caveats in Generating Medical Imaging Labels from Radiology Reports 1–4 (2019). URL http://arxiv.org/abs/1905.02283.

[121] Taylor-Phillips, S. *et al.* Double reading in breast cancer screening: Cohort evaluation in the CO-OPS trial. *Radiology* **287**, 749–757 (2018). URL https://pubs-rsna-org.ezp-prod1.hul.harvard.edu/doi/10.1148/radiol.2018171010.

[122] Barnett, M. L., Boddupalli, D., Nundy, S. & Bates, D. W. Comparative Accuracy of Diagnosis by Collective Intelligence of Multiple

Physicians vs Individual Physicians. *JAMA Network Open* **2**, e190096–e190096 (2019). URL https://jamanetwork-com.ezp-prod1.hul.harvard.edu/journals/jamanetworkopen/fullarticle/2726709.

[123] Warfield, S. K., Zou, K. H. & Wells, W. M. Simultaneous Truth and Performance Level Estimation (STAPLE): An Algorithm for the Validation of Image Segmentation. *IEEE TRANS. ON MEDICAL IMAGING* **23**, 903–921 (2004).

[124] Winzeck, S. *et al.* ISLES 2016 and 2017-benchmarking ischemic stroke lesion outcome prediction based on multispectral MRI. *Frontiers in Neurology* **9**, 679 (2018).

[125] Commowick, O. *et al.* Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure. *Scientific Reports 2018 8:1* **8**, 1–17 (2018). URL https://www.nature.com/articles/s41598-018-31911-7.

[126] Ju, L. *et al.* Improving Medical Images Classification with Label Noise Using Dual-Uncertainty Estimation. *IEEE Transactions on Medical Imaging* **41**, 1533–1546 (2022).

[127] Krause, J. *et al.* Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy. *Ophthalmology* **125**, 1264–1272 (2018).

[128] Ryan, M. C. *et al.* Development and Evaluation of Reference Standards for Image-based Telemedicine Diagnosis and Clinical Research Studies in Ophthalmology. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* **2014**, 1902–1910 (2014).

[129] Aljabri, M., AlAmir, M., AlGhamdi, M., Abdel-Mottaleb, M. & Collado-Mesa, F. Towards a better understanding of annotation tools for medical imaging: a survey. *Multimedia Tools and Applications* **81**, 25877 (2022). URL /pmc/articles/PMC8948453//pmc/articles/PMC8948453/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC8948453/.

[130] Singh, N. K. & Raza, K. Medical image generation using generative adversarial networks: A review. *Health informatics: A computational perspective in healthcare* 77–96 (2021).

[131] Jia, Z., Huang, X., Chang, E. I. & Xu, Y. Constrained Deep Weak Supervision for Histopathology Image Segmentation. *IEEE Transactions on Medical Imaging* **36**, 2376–2388 (2017).

[132] Otesteanu, C. F. *et al.* A weakly supervised deep learning approach for label-free imaging flow-cytometry-based blood diagnostics. *Cell Reports Methods* **1**, 100094 (2021). URL https://doi.org/10.1016/j.crmeth.2021.100094.

[133] Tiu, E. *et al.* Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nature Biomedical Engineering* (2022).

[134] Ahmed, M. N., Yamany, S. M., Mohamed, N., Farag, A. A. & Moriarty, T. A modified fuzzy C-means algorithm for bias field estimation and segmentation of MRI data. *IEEE Transactions on Medical Imaging* **21**, 193–199 (2002).

[135] Li, C., Xu, C., Anderson, A. & Gore, J. MRI tissue classification and bias field estimation based on coherent local intensity clustering: A unified energy minimization framework. *nformation processing in medical imaging : proceedings of the 21st conference* **21**, 288–299 (2009).

[136] Wells, W. M., Crimson, W. E., Kikinis, R. & Jolesz, F. A. Adaptive segmentation of mri data. *IEEE Transactions on Medical Imaging* **15**, 429–442 (1996).

[137] Guillemaud, R. & Brady, M. Estimating the bias field of MR images. *IEEE Transactions on Medical Imaging* **16**, 238–251 (1997).

[138] Narayana, P. A., Brey, W. W., Kulkarni, M. V. & Sievenpiper, C. L. Compensation for surface coil sensitivity variation in magnetic resonance imaging. *Magnetic resonance imaging* **6**, 271–274 (1988).

[139] Tustison, N. J. *et al.* N4ITK: Improved N3 bias correction. *IEEE Transactions on Medical Imaging* **29**, 1310–1320 (2010).

[140] Gorgolewski, K. *et al.* Nipype: A flexible, lightweight and extensible neuroimaging data processing framework in Python. *Frontiers in Neuroinformatics* **5**, 13 (2011). URL `http://www.ncbi.nlm.nih.gov/pubmed/21897815http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3159964http://journal.frontiersin.org/article/10.3389/fninf.2011.00013/abstract`.

[141] Prior, F. W. *et al.* Facial recognition from volume-rendered magnetic resonance imaging data. *IEEE Transactions on Information Technology in Biomedicine* **13**, 5–9 (2009).

[142] Schwarz, C. G. *et al.* Identification of Anonymous MRI Research Participants with Face-Recognition Software. *New England Journal of Medicine* **381**, 1684–1686 (2019).

[143] Park, J. G. & Lee, C. Skull stripping based on region growing for magnetic resonance brain images. *NeuroImage* **47**, 1394–1407 (2009). URL `http://dx.doi.org/10.1016/j.neuroimage.2009.04.047`.

[144] Smith, S. M. Fast Robust Automated Brain Extraction. *Human Brain Mapping* **17**, 143–155 (2002). URL `https://onlinelibrary.wiley.com/doi/10.1002/hbm.10062,`.

[145] Iglesias, J. E., Cheng-Yi Liu, C.-Y., Thompson, P. M. & Zhuowen Tu, Z. Robust Brain Extraction Across Datasets and Comparison With Publicly Available Methods. *IEEE Transactions on Medical Imaging* **30**, 1617–1634 (2011). URL http://www.ncbi.nlm.nih.gov/pubmed/21880566http://ieeexplore.ieee.org/document/5742706/.

[146] Fennema-Notestine, C. *et al.* Quantitative Evaluation of Automated Skull-Stripping Methods Applied to Contemporary and Legacy Images: Effects of Diagnosis, Bias Correction, and Slice Location URL www.nbirn.net.

[147] Kleesiek, J. *et al.* Deep MRI brain extraction: A 3D convolutional neural network for skull stripping. *NeuroImage* **129**, 460–469 (2016). URL http://dx.doi.org/10.1016/j.neuroimage.2016.01.024.

[148] Fedorov, A. *et al.* End-to-end learning of brain tissue segmentation from imperfect labeling. *Proceedings of the International Joint Conference on Neural Networks* **2017-May**, 3785–3792 (2017).

[149] Hwang, H., Ur Rehman, H. Z. & Lee, S. 3D U-Net for Skull Stripping in Brain MRI. *Applied Sciences 2019, Vol. 9, Page 569* **9**, 569 (2019). URL https://www.mdpi.com/2076-3417/9/3/569/htmhttps://www.mdpi.com/2076-3417/9/3/569.

[150] Rehman, H. Z. U., Hwang, H. & Lee, S. Conventional and deep learning methods for skull stripping in brain MRI. *Applied Sciences (Switzerland)* **10**, 1–30 (2020).

[151] Nyú, L. G. & Udupa, J. K. On standardizing the MR image intensity scale. *Magnetic Resonance in Medicine* **42**, 1072–1081 (1999).

[152] Pluim, J. P., Maintz, J. B. & Viergever, M. A. Mutual-information-based registration of medical images: A survey. *IEEE Transactions on Medical Imaging* **22**, 986–1004 (2003).

[153] Greve, D. N. & Fischl, B. Accurate and robust brain image alignment using boundary-based registration. *NeuroImage* **48**, 63–72 (2009). URL http://dx.doi.org/10.1016/j.neuroimage.2009.06.060.

[154] Johnson, H., Harris, G. & Williams, K. BRAINSFit: mutual information rigid registrations of whole-brain 3D images, using the insight toolkit. *Insight J* **57**, 1–11 (2007).

[155] Fedorov, A. *et al.* 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magnetic Resonance Imaging* **30**, 1323–1341 (2012). URL http://www.ncbi.nlm.nih.gov/pubmed/22770690http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3466397https://linkinghub.elsevier.com/retrieve/pii/S0730725X12001816.

[156] Maier-Hein, L. *et al.* Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature Communications* **9** (2018). URL /pmc/articles/PMC6284017//pmc/articles/PMC6284017/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6284017/.

[157] Reinke, A. *et al.* Common Limitations of Image Processing Metrics: A Picture Story (2022). URL https://arxiv.org/pdf/2104.05642v4.pdfhttp://arxiv.org/abs/2104.05642.

[158] Dice, L. R. Measures of the Amount of Ecologic Association Between Species. *Ecology* **26**, 297–302 (1945). URL https://onlinelibrary.wiley.com/doi/full/10.2307/1932409https://onlinelibrary.wiley.com/doi/abs/10.2307/1932409https://esajournals.onlinelibrary.wiley.com/doi/10.2307/1932409.

[159] Sorensen, T. A. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Biol. Skar.* **5**, 1–34 (1948).

[160] Bertels, J. *et al.* Optimizing the Dice Score and Jaccard Index for Medical Image Segmentation: Theory and Practice. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **11765 LNCS**, 92–100 (2019).

[161] Zijdenbos, A. P., Dawant, B. M., Margolin, R. A. & Palmer, A. C. Morphometric Analysis of White Matter Lesions in MR Images: Method and Validation. *IEEE Transactions on Medical Imaging* **13**, 716–724 (1994).

[162] Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S. & Jorge Cardoso, M. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **10553 LNCS**, 240–248 (2017). URL https://link-springer-com.ezp-prod1.hul.harvard.edu/chapter/10.1007/978-3-319-67558-9_28.

[163] Taha, A. A. & Hanbury, A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Medical Imaging* **15** (2015). URL http://dx.doi.org/10.1186/s12880-015-0068-x.

[164] Yeghiazaryan, V. & Voiculescu, I. Family of boundary overlap metrics for the evaluation of medical image segmentation. *Journal of Medical Imaging* **5**, 1 (2018).

[165] Cha, E. *et al.* Clinical implementation of deep learning contour autosegmentation for prostate radiotherapy. *Radiotherapy and Oncology* **159**, 1–7 (2021). URL https://doi.org/10.1016/j.radonc.2021.02.040.

[166] Valentini, V., Boldrini, L., Damiani, A. & Muren, L. P. Recommendations on how to establish evidence from auto-segmentation software in radiotherapy (2014). URL http://dx.doi.org/10.1016/j.radonc.2014.09.014.

[167] Huttenlocher, D. P., Klanderman, G. A. & Rucklidge, W. J. Comparing Images Using the Hausdorff Distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **15**, 850–863 (1993).

[168] Ginneken, B. v., Heimann, T. & Styner, M. 3D Segmentation in the Clinic: A Grand Challenge. *MICCAIWorkshop on 3D Segmentation in the Clinic: A Grand Challenge* **1**, 7–15 (2007).

[169] Nikolov, S. *et al.* Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy (2018). URL https://arxiv.org/pdf/1809.04430.pdfhttp://arxiv.org/abs/1809.04430.

[170] Sharp, G. *et al.* Vision 20/20: Perspectives on automated image segmentation for radiotherapy. *Medical Physics* **41**, 050902 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24784366.

[171] Guo, X., Yin, Y., Dong, C., Yang, G. & Zhou, G. On the class imbalance problem. *Proceedings - 4th International Conference on Natural Computation, ICNC 2008* **4**, 192–201 (2008).

[172] Hicks, S. A. *et al.* On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports 2022 12:1* **12**, 1–9 (2022). URL https://www.nature.com/articles/s41598-022-09954-8.

[173] Aerts, H. J. *et al.* Defining a Radiomic Response Phenotype: A Pilot Study using targeted therapy in NSCLC. *Scientific Reports* **6**, 33860 (2016). URL http://www.nature.com/articles/srep33860.

[174] Rudie, J. D., Rauschecker, A. M., Bryan, R. N., Davatzikos, C. & Mohan, S. Emerging Applications of Artificial Intelligence in Neuro-Oncology. *Radiology* **290**, 607–618 (2019). URL http://pubs.rsna.org/doi/10.1148/radiol.2018181928.

[175] Aparicio, S. & Caldas, C. The Implications of Clonal Genome Evolution for Cancer Medicine. *New England Journal of Medicine* **368**, 842–851 (2013). URL http://www.nejm.org/doi/10.1056/NEJMra1204892.

[176] Traverso, A., Wee, L., Dekker, A. & Gillies, R. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. *International Journal of Radiation Oncology Biology Physics* **102**, 1143–1158 (2018). URL https://www.sciencedirect.com/science/article/pii/S0360301618309052?via%3Dihub.

[177] Van Griethuysen, J. J. *et al.* Computational radiomics system to decode the radiographic phenotype. *Cancer Research* **77**, e104–e107 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/29092951http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5672828.

[178] Garapati, S. S. *et al.* Urinary bladder cancer staging in CT urography using machine learning. *Medical Physics* **44**, 5814–5823 (2017).

[179] Madabhushi, A. & Udupa, J. K. New methods of MR image intensity standardization via generalized scale. *Medical Physics* **33**, 3426–3434 (2006). URL http://www.ncbi.nlm.nih.gov/pubmed/17022239http://doi.wiley.com/10.1118/1.2335487.

[180] Rizzo, S. *et al.* Radiomics: the facts and the challenges of image analysis. *European Radiology Experimental* **2**, 36 (2018). URL http://www.ncbi.nlm.nih.gov/pubmed/30426318http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6234198https://eurradiolexp.springeropen.com/articles/10.1186/s41747-018-0068-z.

[181] Schwier, M. *et al.* Repeatability of Selected Multiparametric Prostate MRI Radiomics Features. *arxiv* **ICC**, 18–20 (2018).

[182] Molina, D. *et al.* Influence of gray level and space discretization on brain tumor heterogeneity measures obtained from magnetic resonance images. *Computers in Biology and Medicine* **78**, 49–57 (2016). URL http://dx.doi.org/10.1016/j.compbiomed.2016.09.011.

[183] Duron, L. *et al.* Gray-level discretization impacts reproducible MRI radiomics texture features. *PLoS ONE* **14**, e0213459 (2019). URL http://www.ncbi.nlm.nih.gov/pubmed/30845221http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6405136.

[184] Batchelor, T. T. *et al.* Improved tumor oxygenation and survival in glioblastoma patients who show increased blood perfusion after cediranib and chemoradiation. *Proceedings of the National Academy of Sciences* **110**, 19059–19064 (2013). URL http://www.ncbi.nlm.nih.gov/pubmed/24190997http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3839699http://www.pnas.org/cgi/doi/10.1073/pnas.1318022110.

[185] Kalpathy-Cramer, J. & Gerstner, E. R. QIN GBM Treatment Response Dataset. URL https://wiki.cancerimagingarchive.net/display/Public/QIN+GBM+Treatment+Response.

[186] Chen, W., Liu, B., Peng, S., Sun, J. & Qiao, X. Computer-Aided Grading of Gliomas Combining Automatic Segmentation and Radiomics. *International Journal of Biomedical Imaging* **2018** (2018).

[187] Bartko, J. J. The intraclass correlation coefficient as a measure of reliability. *Psychological reports* **19**, 3–11 (1966).

[188] Zwanenburg, A., Leger, S., Vallières, M. & Löck, S. Image biomarker standardisation initiative (2016). URL http://arxiv.org/abs/1612.07003.

[189] Lin, J. Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory* **37**, 145–151 (1991).

[190] Kruskal, W. H. & Wallis, W. A. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association* **47**, 583–621 (1952).

[191] Dunn, O. J. Multiple comparisons using rank sums. *Technometrics* **6**, 241–252 (1964).

[192] Pavic, M. *et al.* Influence of inter-observer delineation variability on radiomics stability in different tumor sites. *Acta Oncologica* **57**, 1070–1074 (2018). URL https://doi.org/10.1080/0284186X.2018.1445283.

[193] Shinohara, R. T. *et al.* Statistical normalization techniques for magnetic resonance imaging. *NeuroImage. Clinical* **6**, 9–19 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/25379412http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4215426.

[194] Sun, X. *et al.* Histogram-based normalization technique on human brain magnetic resonance images from different acquisitions. *Biomedical engineering online* **14**, 73 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/26215471http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4517549.

[195] Chang, K. *et al.* Automatic assessment of glioma burden: a deep learning algorithm for fully automated volumetric and bidimensional measurement. *Neuro-Oncology* (2019). URL https://academic.oup.com/neuro-oncology/advance-article/doi/10.1093/neuonc/noz106/5514498.

[196] He, L. *et al.* Effects of contrast-enhancement, reconstruction slice thickness and convolution kernel on the diagnostic performance of radiomics signature in solitary pulmonary nodule. *Scientific reports* **6**, 34921 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/27721474http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5056507.

[197] Laws, K. I. Textured image segmentation. *University of Southern California, IPI Report* **940** (1980).

[198] Jain, A. K. & Farrokhnia, F. Unsupervised texture segmentation using Gabor filters. *Pattern Recognition* **24**, 1167–1186 (1991). URL https://www.sciencedirect.com/science/article/abs/pii/003132039190143S.

[199] Samala, R. K., Chan, H.-P., Hadjiiski, L. & Koneru, S. Hazards of data leakage in machine learning: a study on classification of breast cancer using deep neural networks. In *Medical Imaging 2020: Computer-Aided Diagnosis*, vol. 11314, 39 (2020).

[200] Tampu, I. E., Eklund, A. & Haj-Hosseini, N. Inflation of test accuracy due to data leakage in deep learning-based classification of OCT images. *Scientific Data* **9**, 1–8 (2022).

[201] Ragone, A., Mirylenka, K., Casati, F. & Marchese, M. A quantitative analysis of peer review. *Proceedings of ISSI 2011 - 13th Conference of the International Society for Scientometrics and Informetrics* **2**, 724–736 (2011).

[202] Godlee, F., Gale, C. R. & Martyn, C. N. Effect on the quality of peer review of blinding reviewers and asking them to sign their reports: a randomized controlled trial. *Jama* **280**, 237–240 (1998).

[203] Goodman, S. N., Berlin, J., Fletcher, S. W. & Fletcher, R. H. Manuscript quality before and after peer review and editing at Annals of Internal Medicine. *Annals of internal medicine* **121**, 11–21 (1994).

[204] Geirhos, R. *et al.* Shortcut learning in deep neural networks. *Nature Machine Intelligence* **2**, 665–673 (2020).

[205] Mahmood, U. *et al.* Detecting Spurious Correlations With Sanity Tests for Artificial Intelligence Guided Radiology Systems. *Frontiers in Digital Health* **3** (2021).

[206] Gal, Y. & Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, 1050–1059 (PMLR, 2016).

[207] Amodei, D. *et al.* Concrete Problems in AI Safety (2016). URL `https://arxiv.org/pdf/1606.06565.pdf`.

[208] Ridella, S. ODI RESUME MANUFACTURER & PRODUCT INFORMATION FAILURE REPORT SUMMARY. Tech. Rep. (2010). URL `https://www.safetyresearch.net/Library/PE16007_Closing.pdf`.

[209] Hendrycks, D. & Gimpel, K. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks 1–12 (2016). URL `http://arxiv.org/abs/1610.02136`.

[210] Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, vol. 2017-Decem, 6403–6414 (2017). URL `https://arxiv.org/abs/1612.01474`.

[211] Gal, Y. & Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning **48** (2015). URL http://arxiv.org/abs/1506.02142.

[212] Kendall, A., Badrinarayanan, V. & Cipolla, R. Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding (2015). URL http://arxiv.org/abs/1511.02680.

[213] Aberle, D. R. *et al.* Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine* **365**, 395–409 (2011).

[214] Baltruschat, I. *et al.* Smart chest X-ray worklist prioritization using artificial intelligence: a clinical workflow simulation. *European Radiology* (2020).

[215] Hanna, T. N., Lamoureux, C., Krupinski, E. A., Weber, S. & Johnson, J.-O. Effect of Shift, Schedule, and Volume on Interpretive Accuracy: A Retrospective Analysis of 2.9 Million Radiologic Examinations. *Radiology* **287**, 205–212 (2018). URL http://pubs.rsna.org/doi/10.1148/radiol.2017170555.

[216] Roberts, M. *et al.* Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence* **3**, 199–217 (2021).

[217] Mongan, J., Moy, L. & Kahn, C. E. Checklist for Artificial Intelligence and Medical Imaging (Claim). *Radiology: Artificial Intelligence* (2020).

[218] Thomas, R. L. & Uminsky, D. Reliance on metrics is a fundamental challenge for AI. *arXiv* (2020).

[219] Strathern, M. 'improving ratings': audit in the british university system. *European review* **5**, 305–321 (1997).

[220] Stidham, R. W. *et al.* Performance of a deep learning model vs human reviewers in grading endoscopic disease severity of patients with ulcerative colitis. *JAMA Network Open* **2**, e193963 (2019). URL https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2733432.

[221] De Fauw, J. *et al.* Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine* **24**, 1342–1350 (2018). URL http://www.nature.com/articles/s41591-018-0107-6.

[222] Leibig, C., Allken, V., Ayhan, M. S., Berens, P. & Wahl, S. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific Reports* **7**, 17816 (2017). URL http://www.nature.com/articles/s41598-017-17876-z.

[223] Antony, J., McGuinness, K., O'Connor, N. E. & Moran, K. Quantifying radiographic knee osteoarthritis severity using deep convolutional

neural networks. In *Proceedings - International Conference on Pattern Recognition*, 1195–1200 (2016). URL http://www.adamondemand.com/clinical-management-of-osteoarthritis/.

[224] Li, M. D. *et al.* Siamese neural networks for continuous disease severity evaluation and change detection in medical imaging. *npj Digital Medicine* **3**, 48 (2020). URL http://www.nature.com/articles/s41746-020-0255-1.

[225] Brown, J. M. *et al.* Fully automated disease severity assessment and treatment monitoring in retinopathy of prematurity using deep learning. *Proceedings of SPIE–the International Society for Optical Engineering* 22 (2018).

[226] Moleta, C. *et al.* Plus Disease in Retinopathy of Prematurity: Diagnostic Trends in 2016 Versus 2007. *American Journal of Ophthalmology* **176**, 70–76 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/28087400http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5376516.

[227] Dwork, C., Hardt, M., Pitassi, T., Reingold, O. & Zemel, R. Fairness through awareness. *ITCS 2012 - Innovations in Theoretical Computer Science Conference* 214–226 (2012).

[228] Redd, T. K. *et al.* Evaluation of a deep learning image assessment system for detecting severe retinopathy of prematurity. *British Journal of Ophthalmology* **103**, 580–584 (2019). URL http://www.ncbi.nlm.nih.gov/pubmed/30470715http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC7880608.

[229] Taylor, S. *et al.* Monitoring Disease Progression with a Quantitative Severity Scale for Retinopathy of Prematurity Using Deep Learning. *JAMA Ophthalmology* **137**, 1022–1028 (2019).

[230] Li, M. *et al.* Automated Assessment and Tracking of COVID-19 Pulmonary Disease Severity on Chest Radiographs using Covolutional Siamese Neural Networks. *Radiology: Artificial Intelligence* (2020).

[231] Zhang, Z. & Sabuncu, M. R. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in Neural Information Processing Systems* **2018-Decem**, 8778–8788 (2018).

[232] Almeida, M., Zhuang, Y., Ding, W., Crouter, S. & Chen, P. Mitigating class-boundary label uncertainty to reduce both model bias and variance. *arXiv* (2020).

[233] Ostyakov, P. *et al.* Label denoising with large ensembles of heterogeneous neural networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **11132 LNCS**, 250–261 (2019).

[234] Clough, P., Sanderson, M., Tang, J., Gollins, T. & Warner, A. Examining the limits of crowdsourcing for relevance assessment. *IEEE Internet Computing* **17**, 32–38 (2013).

[235] Beyer, L., Hénaff, O. J., Kolesnikov, A., Zhai, X. & Oord, A. v. d. Are we done with ImageNet? (2020). URL `https://arxiv.org/abs/2006.07159v1http://arxiv.org/abs/2006.07159`.

[236] Dawid, A. P. & Skene, A. M. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Applied Statistics* **28**, 20 (1979).

[237] Guan, M. Y., Gulshan, V., Dai, A. M. & Hinton, G. E. Who said what: Modeling individual labelers improves classification. In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, 3109–3118 (2018). URL `www.aaai.org`.

[238] Khetan, A., Lipton, Z. C. & Anandkumar, A. Learning From Noisy Singly-Labeled Data. *Conference Proceedings: International Conference on Learning Representations* 1–15 (2018).

[239] Tanno, R., Saeedi, A., Sankaranarayanan, S., Alexander, D. C. & Silberman, N. Learning from noisy labels by regularized estimation of annotator confusion. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, 11236–11245 (2019). URL `https://arxiv.org/pdf/1902.03680.pdf`.

[240] Shah, P. K. *et al.* Retinopathy of prematurity: Past, present and future. *World journal of clinical pediatrics* **5**, 35 (2016).

[241] Quinn, G. E. The international classification of retinopathy of prematurity revisited: An international committee for the classification of retinopathy of prematurity. *Archives of Ophthalmology* **123**, 991–999 (2005).

[242] Cui, A. *et al.* Global, regional prevalence, incidence and risk factors of knee osteoarthritis in population-based studies. *EClinicalMedicine* **29**, 100587 (2020).

[243] Kellgren, J. H. & Lawrence, J. Radiological assessment of osteo-arthrosis. *Annals of the rheumatic diseases* **16**, 494 (1957).

[244] Sung, H. *et al.* Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* **71**, 209–249 (2021).

[245] Liberman, L. & Menell, J. H. Breast imaging reporting and data system (BI-RADS) (2002). URL `https://pubmed.ncbi.nlm.nih.gov/12117184/`.

[246] Boyd, N. F. *et al.* Quantitative classification of mammographic densities and breast cancer risk: Results from the canadian national breast screening study. *Journal of the National Cancer Institute* **87**, 670–675 (1995).

[247] Bakker, M. F. *et al.* Supplemental MRI Screening for Women with Extremely Dense Breast Tissue. *New England Journal of Medicine* **381**, 2091–2102 (2019).

[248] Pisano, E. D. *et al.* Diagnostic Performance of Digital versus Film Mammography for Breast-Cancer Screening. *New England Journal of Medicine* **353**, 1773–1783 (2005). URL `http://www.nejm.org/doi/abs/10.1056/NEJMoa052911`.

[249] Highnam, R., Brady, M., Yaffe, M. J., Karssemeijer, N. & Harvey, J. Robust breast composition measurement - Volpara™. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6136 LNCS, 342–349 (Springer, Berlin, Heidelberg, 2010). URL `http://link.springer.com/10.1007/978-3-642-13666-5_46`.

[250] Wanders, J. O. *et al.* Volumetric breast density affects performance of digital screening mammography. *Breast cancer research and treatment* **162**, 95–103 (2017).

[251] Li, L. & Lin, H.-T. Ordinal regression by extended binary classification (2007).

[252] Cao, W., Mirjalili, V. & Raschka, S. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters* 325–331 (2020). URL `https://doi.org/10.1016/j.patrec.2020.11.008`.

[253] Bromley, J. *et al.* Signature verification using a "siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence* **7**, 669–688 (1993).

[254] Hadsell, R., Chopra, S. & LeCun, Y. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, 1735–1742 (IEEE, 2006).

[255] Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. R. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580* (2012).

[256] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* **15**, 1929–1958 (2014).

[257] He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778 (2016).

[258] Kingma, D. P. & Ba, J. L. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 1–15 (2015).

[259] Koch, G., Zemel, R. & Salakhutdinov, R. Siamese Neural Networks for One-shot Image Recognition. *International Conference on Machine Learning* **37** (2015). URL https://www.cs.cmu.edu/~rsalakhu/papers/oneshot1.pdf.

[260] Reijman, M., Hazes, J. M. W., Pols, H. A. P. & Bernsen, D. Validity and reliability of three definitions of hip osteoarthritis: cross sectional and longitudinal approach. *Ann Rheum Dis* **63**, 1427–1433 (2004). URL http://ard.bmj.com/.

[261] Redondo, A. *et al.* Inter-and intraradiologist variability in the BI-RADS assessment and breast density categories for screening mammograms URL www.metodo.uab.cat/macros].

[262] Goh, H. W., Tkachenko, U., Mueller, J., Cleanlab, J. A. & Cleanlab, C. Utilizing supervised models to infer consensus labels and their quality from data with multiple annotators (2022). URL https://arxiv.org/abs/2210.06812v1.

[263] Prabhakaran, V., Mostafazadeh Davani, A. & Diaz, M. On Releasing Annotator-Level Labels and Information in Datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, 133–138 (Association for Computational Linguistics (ACL), 2021). URL https://arxiv.org/abs/2110.05699v1.

[264] Chetoui, M. & Akhloufi, M. A. Explainable Diabetic Retinopathy using EfficientNET. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, vol. 2020-July, 1966–1969 (2020).

[265] Ostrom, Q. T. *et al.* CBTRUS Statistical Report: Primary brain and other central nervous system tumors diagnosed in the United States in 2010–2014. *Neuro-Oncology* **19**, 1–88 (2017). URL https://watermark.silverchair.com/nox158.pdf.

[266] Li, Y. *et al.* A systematic review of multifocal and multicentric glioblastoma. *Journal of Clinical Neuroscience* **83**, 71–76 (2021).

[267] Berntsen, E. M. *et al.* Volumetric segmentation of glioblastoma progression compared to bidimensional products and clinical radiological reports. *Acta Neurochirurgica* **162**, 379–387 (2020). URL https://pubmed.ncbi.nlm.nih.gov/31760532/.

[268] Fick, T. *et al.* Fully automatic brain tumor segmentation for 3D evaluation in augmented reality. *Neurosurgical Focus* **51**, 1–8 (2021).

[269] Chisholm, R. A., Stenning, S. & Hawkins, T. D. The accuracy of volumetric measurement of high-grade gliomas. *Clinical Radiology* **40**, 17–21 (1989).

[270] Moltz, J. H. *et al.* Analysis of variability in manual liver tumor delineation in CT scans. In *Proceedings - International Symposium on Biomedical Imaging*, 1974–1977 (IEEE, 2011). URL http://ieeexplore.ieee.org/document/5872797/.

[271] van der Veen, J., Gulyban, A. & Nuyts, S. Interobserver variability in delineation of target volumes in head and neck cancer. *Radiotherapy and Oncology* **137**, 9–15 (2019). URL https://doi.org/10.1016/j.radonc.2019.04.006.

[272] Joe, B. N. *et al.* Computer Applications Brain Tumor Volume Measurement : Comparison of Manual and Semiautomated. *Radiology* **m**, 811–816 (1999).

[273] Aselmaa, A., Goossens, R. H. M. & Freudenthal, A. What is Sensemaking in the Context of External Radiotherapy Treatment Planning? *DMD Europe 2013: Design of Medical Devices Conference - Europe Edition 2013* (2013).

[274] Riegel, A. C. *et al.* Variability of gross tumor volume delineation in head-and-neck cancer using CT and PET/CT fusion. *International Journal of Radiation Oncology Biology Physics* **65**, 726–732 (2006).

[275] Vinod, S. K., Jameson, M. G., Min, M. & Holloway, L. C. Uncertainties in volume delineation in radiation oncology: A systematic review and recommendations for future studies. *Radiotherapy and Oncology* **121**, 169–179 (2016). URL http://dx.doi.org/10.1016/j.radonc.2016.09.009.

[276] Weiss, E. *et al.* Conformal radiotherapy planning of cervix carcinoma: Differences in the delineation of the clinical target volume. A comparison between gynaecologic and radiation oncologists. *Radiotherapy and Oncology* **67**, 87–95 (2003).

[277] Egger, J. *et al.* GBM volumetry using the 3D slicer medical image computing platform. *Scientific Reports* **3**, 1–7 (2013).

[278] Porz, N. *et al.* Multi-modal glioblastoma segmentation: Man versus machine. *PLoS ONE* **9**, e96873 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24804720http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4013039.

[279] Bi, N. *et al.* Deep Learning Improved Clinical Target Volume Contouring Quality and Efficiency for Postoperative Radiation Therapy in Non-small Cell Lung Cancer. *Frontiers in Oncology* **9**, 1192 (2019). URL https://www.frontiersin.org/article/10.3389/fonc.2019.01192/full.

[280] Ma, C. Y. *et al.* Deep learning-based auto-segmentation of clinical target volumes for radiotherapy treatment of cervical cancer. *Journal of Applied Clinical Medical Physics* **23**, e13470 (2022). URL https://onlinelibrary.wiley.com/doi/10.1002/acm2.13470.

[281] Hoebel, K. V. *et al.* Do I know this? segmentation uncertainty under domain shift. *Medical Imaging 2022: Image Processing* **1203211**, 27 (2022).

[282] Asan, O., Bayrak, A. E. & Choudhury, A. Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. *Journal of Medical Internet Research* **22** (2020).

[283] Budd, S., Robinson, E. C. & Kainz, B. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis* **71**, 102062 (2021).

[284] Lo, A. C. *et al.* The impact of peer review of volume delineation in stereotactic body radiation therapy planning for primary lung cancer: A multicenter quality assurance study. *Journal of Thoracic Oncology* **9**, 527–533 (2014). URL `http://www.ncbi.nlm.nih.gov/pubmed/24736076`.

[285] Sherer, M. V. *et al.* Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review (2021). URL `https://doi.org/10.1016/j.radonc.2021.05.003`.

[286] Beers, A. *et al.* DeepNeuro: an open-source deep learning toolbox for neuroimaging. *Neuroinformatics* (2020).

[287] Jungo, A., Scheidegger, O., Reyes, M. & Balsiger, F. pymia: A Python package for data handling and evaluation in deep learning-based medical image analysis. *Computer Methods and Programs in Biomedicine* **198**, 105796 (2021). URL `https://www.sciencedirect.com/science/article/pii/S0169260720316291`.

[288] Roy, A. G., Conjeti, S., Navab, N. & Wachinger, C. Inherent brain segmentation quality control from fully convnet monte carlo sampling. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **11070 LNCS**, 664–672 (2018).

[289] Harris, P. A. *et al.* Research electronic data capture (REDCap)-A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics* **42**, 377–381 (2009). URL `http://dx.doi.org/10.1016/j.jbi.2008.08.010`.

[290] Harris, P. A. *et al.* The REDCap consortium: Building an international community of software platform partners. *Journal of Biomedical Informatics* **95**, 103208 (2019). URL `https://doi.org/10.1016/j.jbi.2019.103208`.

[291] Huang, R. RQDA: R-based Qualitative Data Analysis. (2016). URL `http://rqda.r-forge.r-project.org/`.

[292] Krippendorff, K. Computing Krippendorff's alpha-reliability (2011).

[293] Gwet, K. L. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology* **61**, 29–48 (2008).

[294] Quarfoot, D. & Levine, R. A. How Robust Are Multirater Interrater Reliability Indices to Changes in Frequency Distribution? *American Statistician* **70**, 373–384 (2016). URL `https://doi.org/10.1080/00031305.2016.1141708`.

[295] Lu, S. L. *et al.* Randomized multi-reader evaluation of automated detection and segmentation of brain tumors in stereotactic radiosurgery with deep neural networks. *Neuro-Oncology* **23**, 1560–1568 (2021).

[296] Conte, G. M., Weston, A. D., Vogelsang, D. C., Philbrick, K. A. & Cai, J. C. Generative Adversarial Networks to Synthesize Missing T1 and FLAIR MRI Sequences for Use in a Multisequence Brain Tumor Segmentation Model. *Radiology* **299**, 313–323 (2021).

[297] Di Ieva, A. *et al.* Application of deep learning for automatic segmentation of brain tumors on magnetic resonance imaging: a heuristic approach in the clinical scenario. *Neuroradiology* **63**, 1253–1262 (2021).

[298] Mitchell, J. R. *et al.* Deep neural network to locate and segment brain tumors outperformed the expert technicians who created the training data. *Journal of Medical Imaging* **7** (2020).

[299] Wang, G. *et al.* Interactive Medical Image Segmentation Using Deep Learning with Image-Specific Fine Tuning. *IEEE Transactions on Medical Imaging* **37**, 1562–1573 (2018).

[300] Li, M. D. *et al.* Multi-Radiologist User Study for Artificial Intelligence-Guided Grading of COVID-19 Lung Disease Severity on Chest Radiographs. *Academic Radiology* **28**, 572–576 (2021). URL `http://www.ncbi.nlm.nih.gov/pubmed/33485773http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC7813473`.

[301] Kiureghian, A. D. & Ditlevsen, O. Aleatory or epistemic? Does it matter? *Structural Safety* **31**, 105–112 (2009).

[302] Mehrtash, A., Wells, W. M., Tempany, C. M., Abolmaesumi, P. & Kapur, T. Confidence Calibration and Predictive Uncertainty Estimation for Deep Medical Image Segmentation (2019). URL `http://arxiv.org/abs/1911.13273`.

[303] Hoebel, K., Chang, K., Patel, J., Singh, P. & Kalpathy-Cramer, J. Give me (un)certainty – An exploration of parameters that affect segmentation uncertainty (2019). URL `https://arxiv.org/abs/1911.06357`.

[304] Jha, A. K. *et al.* Objective Task-Based Evaluation of Artificial Intelligence-Based Medical Imaging Methods:: Framework, Strategies, and Role of the Physician. *PET Clinics* **16**, 493–511 (2021). URL `https://doi.org/10.1016/j.cpet.2021.06.013`.

[305] Kruser, T. *et al.* NRG Brain Tumor Specialists Consensus Guidelines for Glioblastoma Contouring. *Journal of Neuro-Oncology* **143**, 157–166 (2019).

[306] van den Oever, L. B. *et al.* Qualitative Evaluation of Common Quantitative Metrics for Clinical Acceptance of Automatic Segmentation: a Case Study

on Heart Contouring from CT Images by Deep Learning Algorithms. *Journal of Digital Imaging* **35**, 240–247 (2022). URL https://doi.org/10.1007/s10278-021-00573-9.

[307] Aselmaa, A. *et al.* Medical Factors of Brain Tumor Delineation in Radiotherapy for Software Design. *Proceedings of the 5th International Conference on Applied Human Factors and Ergonomics* 4865–4876 (2014).

[308] Glasziou, P. *et al.* Reducing waste from incomplete or unusable reports of biomedical research. *The Lancet* **383**, 267–276 (2014). URL http://dx.doi.org/10.1016/S0140-6736(13)62228-X.

[309] Barnes, C. *et al.* Impact of an online writing aid tool for writing a randomized trial report: The COBWEB (Consort-based WEB tool) randomized controlled trial. *BMC Medicine* **13**, 1–10 (2015). URL http://dx.doi.org/10.1186/s12916-015-0460-y.

[310] Marušić, A. A tool to make reporting checklists work. *BMC Medicine* **13**, 10–12 (2015).

[311] Food, U., Drug Administration, C. f. D. & Health, R. Neosoma 510(k) Premarket Notification (2022). URL https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K221738.

[312] Dalca, A. V. *et al.* Segmentation of cerebrovascular pathologies in stroke patients with spatial and shape priors. In *International Conference on medical image computing and computer-assisted intervention*, 773–780 (Springer, 2014).

[313] Liu, C., Zeng, X., Liang, K., Yu, Y. & Ye, C. Improved Brain Lesion Segmentation with Anatomical Priors from Healthy Subjects. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 186–195 (Springer, 2021).