Automatic tumour segmentation in brain MR images

Moving towards clinical implementation

Emilia Gryska

Department of Medical Radiation Sciences

Institute of Clinical Sciences

Sahlgrenska Academy, University of Gothenburg

UNIVERSITY OF GOTHENBURG

Gothenburg 2022

Cover illustration by Emilia Gryska

Automatic tumour segmentation in brain MR images: moving towards clinical implementation

emilia.gryska@gu.se

ISBN 978-91-8009-935-6 (PRINT)
ISBN 978-91-8009-936-3 (PDF)
http://hdl.handle.net/2077/72056

Printed in Borås, Sweden 2022
Printed by Stema Specialtryck AB

"It is not the critic who counts; not the man who points out how the strong man stumbles, or where the doer of deeds could have done them better. The credit belongs to the man who is actually in the arena, whose face is marred by dust and sweat and blood; who strives valiantly; who errs, who comes short again and again, because there is no effort without error and shortcoming; but who does actually strive to do the deeds; who knows great enthusiasms, the great devotions; who spends himself in a worthy cause; who at the best knows in the end the triumph of high achievement, and who at the worst, if he fails, at least fails while daring greatly, so that his place shall never be with those cold and timid souls who neither know victory nor defeat."

Theodore Roosevelt

Automatic tumour segmentation in brain MR images

Moving towards clinical implementation

Emilia Gryska

Department of Medical Radiation Sciences, Institute of Clinical Sciences,
Sahlgrenska Academy, University of Gothenburg
Gothenburg, Sweden

ABSTRACT

The aim of this thesis was to examine and enhance the scientific groundwork for translating deep learning (DL) algorithms for brain tumour segmentation into clinical decision support tools. **Paper II** describes a scoping review conducted to map the field of automatic brain lesion segmentation on magnetic resonance (MR) images according to a predefined and peer-reviewed study protocol (**Paper I**). Insufficient preprocessing description was identified as one factor hindering clinical implementation of the reviewed algorithms. A reproducibility and replicability analysis of two algorithms was described in **Paper III**. The two algorithms and their validation studies were previously assessed as reproducible. In this experimental investigation, the original validation results were reproduced and replicated for one algorithm. Analysing the reasons for failure to reproduce validation of the second algorithm led to a suggested update to a commonly-used reproducibility checklist; the importance of a thorough description of preprocessing was highlighted. In **Paper IV**, radiologists' perception of DL-generated brain tumour labels in tumour volume growth assessment was examined. Ten radiologists participated in a reading/questionnaire session of 20 MR examination cases. The readers were confident that the label-derived volume change is more accurate than their visual assessment, even when the inter-rater agreement on the label quality was poor. In **Paper V**, the broad theme of trust in artificial intelligence (AI) in radiology was explored. A semi-structured interview study with twenty-six

AI implementation stakeholders ws conducted. Four requirements of the implemented tools and procedures were identified that promote trust in AI: reliability, quality control, transparency, and inter-organisational compatibility. The findings indicate that current strategies to validate DL algorithms do not suffice to assess their accuracy in a clinical setting. Despite the recognition from radiologists that DL algorithms can improve the accuracy of tumour volume assessment, implementation strategies require more work and the involvement of multiple stakeholders.

Keywords: brain tumour segmentation, implementation, deep learning, radiology

ISBN 978-91-8009-935-6 (PRINT)
ISBN 978-91-8009-936-3 (PDF)
http://hdl.handle.net/2077/72056

SAMMANFATTNING PÅ SVENSKA

I klinisk praxis används tumörvolym som ett kriterium för att bedöma sjukdomsstatus. Volymen mäts sällan i exakta mått, istället skattas den ofta visuellt. Artificiell intelligens (AI) algoritmer för hjärntumörsegmentering, som automatiskt avgränsar en tumör i en bild, skulle kunna ge neuroradiologer exakta mätningar av tumörvolymen. Trots ett ökande antal vetenskapliga studier som hävdar att algoritmerna har hög noggrannhet används de inte allmänt som kliniska verktyg. I min avhandling undersöker jag klyftan mellan forskningen och kliniken med följande frågor: Ger forskningen belägg för att algoritmerna kommer att fungera väl i klinisk rutin? Litar radiologer på AI-genererad information och anser de att den är till hjälp? Vad behöver radiologer för att kunna lita på och använda verktygen?

Mina resultat visar att de flesta vetenskapliga studier inte utvärderar algoritmerna tillräckligt noga avseende deras prestanda i ett kliniskt scenario. Många studier är inte reproducerbara: en oberoende forskare kan inte återskapa en algoritm, tillämpa den på samma bilder och få liknande resultat. För att lösa detta problem har jag föreslagit uppdaterade riktlinjer för att utforma studier så att de kan reproduceras. Utvärdering av noggrannhet och tillförlitlighet är bara det första steget mot ett kliniskt införande. I slutändan är det ändå läkarna som avgör om den AI-genererade informationen är trovärdig och användbar. Jag fann att radiologer tenderar att lita mer på noggrannheten hos den AI-beräknade tumörvolymen, trots en stor variation i den upplevda segmenteringskvaliteten, än på den egna visuella bedömningen. Denna positiva inställning till AI är dock inte tillräcklig för att säkerställa att de framtida kliniska verktygen faktiskt kommer att få förtroende och användas. Jag identifierade fyra specifika förtroenderelaterade villkor som måste uppfyllas för att framgångsrikt införa AI-verktyg inom radiologi. Verktygen och implementeringen måste vara tillförlitliga och transparenta avseende hur informationen genereras och hur den ska tolkas. Dessutom måste verktyget lätt kunna implementeras i och vara kompatibelt med den kliniska verksamheten.

Resultaten överbryggar delvis klyftan mellan forskningen och kliniken. Framtida forskning bör fokusera på att pröva algoritmer på ett sätt som

fastställer starka bevis för segmenteringsnoggrannhet i en klinisk miljö och hur radiologer använder den erhållna informationen i kliniska arbetsflöden.

LIST OF PAPERS

This thesis is based on the following papers, referred to in the text by their Roman numerals.

I. **Gryska, E. A.**, Schneiderman, J., & Heckemann, R. A. Automatic brain lesion segmentation on standard MRIs of the human head: a scoping review protocol. *BMJ open* 2019; *9*(2), e024824.

II. **Gryska, E.**, Schneiderman, J., Björkman-Burtscher, I. M., & Heckemann, R. A. Automatic brain lesion segmentation on standard magnetic resonance images: a scoping review. *BMJ open*, 2021; *11*(1), e042660.

III. **Gryska, E.**, Björkman-Burtscher, I. M., Jakola, A. S., Dunås, T., Schneiderman, J., & Heckemann, R. A. Deep learning for automatic brain tumour segmentation on MRI: evaluation of recommended reporting criteria via a reproduction and replication study. *BMJ open*, 2022; 12, e059000

IV. **Gryska, E.**, Hoefling, N., Laesser, M., Heckemann, R. A., Schneiderman, J., & Björkman-Burtscher, I. M. Evaluation of contrast-enhanced tumour volume increase in glioblastoma patients: radiologists' perception of tumour segmentation and volumetry. *Submitted to European Radiology*

V. Bergquist, M., Rolandsson, B., **Gryska, E.**, Laesser, M., Hoefling, N., Heckemann, R. A., Schneiderman, J., & Björkman-Burtscher, I. M. Trust and stakeholder perspectives on the implementation of AI tools in clinical radiology. *Manuscript*

CONTENT

ABBREVIATIONS

AI – artificial intelligence

BTS – brain tumour segmentation

CAD – computer-aided diagnosis

CDS – clinical decision support

CNN – convolutional neural network

DL – deep learning

DSC – Dice similarity coefficient

ET – enhancing tumour

GBM – glioblastoma

HGG – high grade glioma

LGG – low grade glioma

MR – magnetic resonance

PPV – positive predictive value

TCIA – The Cancer Imaging Archive

1 INTRODUCTION

The outstanding performance of deep learning (DL – which falls under the umbrella term of artificial intelligence (AI)) algorithms in scientific validation studies[1] generated high expectations that DL will revolutionise radiology. The early ideas about the role of DL in medicine, as envisioned by technologists, painted a futuristic picture of healthcare. A prominent venture capitalist, Vinod Khosla, boldly expressed the vision in 2012:

"Eventually, we won't need the average doctor and will have much better and cheaper care for 90-99% of our medical needs. We will still need to leverage the top 10 or 20% of doctors (at least for the next two decades) to help that bionic software get better at diagnosis. So a world mostly without doctors (at least average ones) is not only not reasonable, but also more likely than not. There will be exceptions, and plenty of stories around these exceptions, but what I am talking about will most likely be the rule and doctors may be the exception rather than the other way around."[2]

Understandably, clinicians became sceptical toward AI, fearing that the technology would make them redundant. Ten years later, we see a shift in the expectations of AI in healthcare. Rather than taking over the jobs, AI is now expected to take over specific tasks to assist clinicians in decision-making. Such support is welcome by radiologists[3] whose workload has been significantly increasing during the last two decades[4].

The growing expectations of DL to be implemented as clinical decision support (CDS) in radiology have been matched by the ever-increasing number of scientific studies that propose and validate DL algorithms for radiological image processing. However, scientific studies on the implementation of the algorithms as clinical tools remain scarce; the expansion of the field has only recently led to the availability of a few DL tools in specific clinical workflows.

Moving away from proof-of-concept studies to validating the clinical suitability of the tools is a complex and multifaceted challenge[5]. Recently, the new European Medical Device Regulation (EU) 2017/745 became another obstacle for researchers in the field; MDR practically limits the development and implementation of DL tools for clinical utilisation to the commercial sector. As a result, non-commercial, in-house validation and implementation of such tools are practically infeasible. Despite this new obstacle, research on the implementation and validation of DL CDS should not come to a halt. Quite the opposite – many critical pieces of the implementation puzzle, which will facilitate a sustainable uptake of AI tools in clinical radiology, remain to be uncovered.

In my project, I focused on brain tumour segmentation (BTS) on magnetic resonance (MR) imaging, one of the most common radiological applications of DL addressed in research[6,7]. I used this sample application to identify some missing pieces of the abovementioned puzzle. The abundance of research that proposes and validates DL BTS algorithms contrasted with the scarcity of available clinical tools determined the two main lines of inquiry of this doctoral project: the scientific and clinical validity of the technical developments in the field of DL BTS, and the tool users' perception and trust in tumour labels generated by AI.

In the following sections, I will introduce relevant topics that comprise the background for this research, and also provide the reader with the necessary understanding of the field and the research gaps I addressed. First, I will briefly describe what BTS is, and its place and benefits for neuroradiological tumour status evaluation. Then I will provide a basic description of the DL technology for BTS, as well as its strengths and limitations from the clinical implementation perspective. Then I will move on to introducing the topic of technology assessment in diagnostic imaging, users' perception of AI-generated tumour labels, and trust in the technology.

1.1 BTS & Neuroradiological Tumour Status Evaluation

Image segmentation is the process of finding and labelling pixels/voxels that belong to semantically distinct regions in an image[8]. BTS generates a tumour label in a cranial image (Figure 1). MR imaging is one of the most commonly used modalities for the neuroradiological assessment of patients suspected of having, or diagnosed with, a brain tumour. This modality enables visualisation of different tissue types, including tumour components, which are relevant for diagnosing and monitoring the disease and evaluating treatment effects. MR images are, therefore, most often used as input for BTS algorithms.

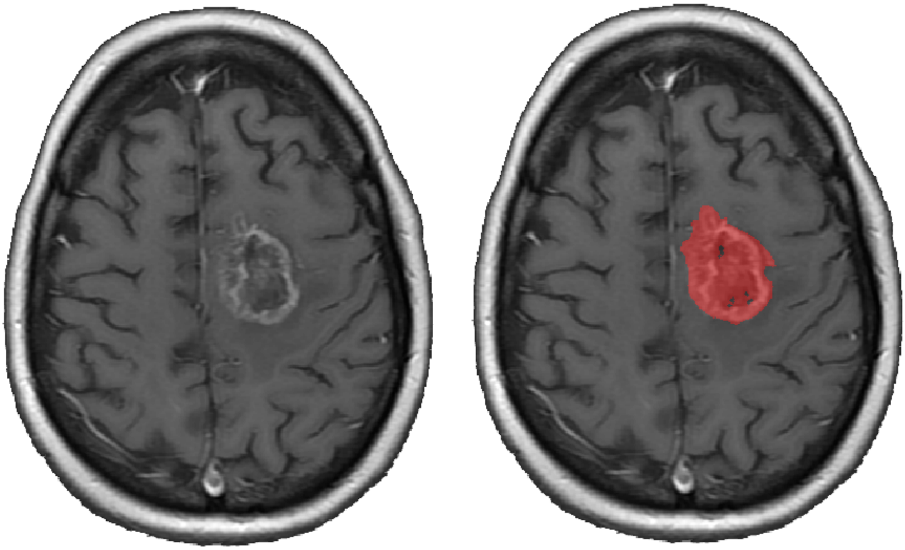

Figure 1. An example of a T1w-Gd brain MRI (left) with a contrast-enhancing tumour label overlaid on the image (right).

BTS is clinically helpful for estimating tumour volume and planning radiation therapy. Radiological tumour evaluation, including its size, is one criterion that determines the disease status (regression, stable disease, or progression). Accurately determined disease status is important not only for the prognosis and therapeutic decisions but also the mental and emotional state of the patient and their family.

In clinical practice, a tumour label would be typically acquired by a trained rater via manually delineating tumour boundaries on each slice of a 3D image. More recently, semi automatic tools that use image processing algorithms became available for clinical use to improve the efficiency of the process (e.g. “SmartBrush,” developed by Brainlab, Feldkirchen, Germany). These tools, however, require user supervision, and the generated labels often need editing. Whether manually or with the support of algorithms, the boundaries of some brain tumours can be difficult to determine because of their heterogeneous appearance on MRI. Even in the case of homogeneous tumours with clear boundaries, manual or semi automatic delineation on multiple slices of an image is too time-consuming to be used routinely in clinical practice. Furthermore, manual segmentation is subject to inter-rater variability. Even findings of studies that assess inter-rater agreement are variable. For gross tumours, the mean largest to smallest tumour volume ratio calculated for nine physicians of 2.04 was recorded[9] compared to Dice similarity coefficient (DSC) of around 0.8 (indicating much better overlap and therefore volume ratio)[10]. For enhancing tumour (ET) inter-rater DSC of less than 0.6 to more than 0.8 was reported[10,11].

Instead, established tumour size evaluation presently relies on visual (qualitative) assessment and approximate quantitative estimations[12]. A few metrics that estimate the size of a tumour have been developed to standardise the reporting of treatment effects on tumours in clinical trials. Subsequently, they were adopted in radiological practice[13]. These metrics include 1D and 2D measurements of the ET: the length of the largest diameter – RECIST[14], a product of the two largest, perpendicular diameters – MacDonald[15]. The RANO[16] criteria include the 2D measurement of ET and non-enhancing lesions. Each metric indicates disease status according to established criteria (Table 1).

Table 1. A comparison of the RECIST[14], MacDonald[15], and RANO[16] response criteria for brain tumour response evaluation. ET – enhancing tumour. T2/FLAIR – transverse relaxation time/fluid attenuated inversion recovery. Excerpt adapted from Chukwueke and Wen (2019)[20] available under CC BY-NC-ND 4.0[21].

Criterion	**RECIST**	**MacDonald**	**RANO**
Measurement	1D ET: the largest diameter/sum of the largest diameters for multiple lesions	2D ET: the product of the largest perpendicular diameters	2D ET + T2/FLAIR: the product of the largest perpendicular diameters
Progression	≥ 20% increase	≥ 25% increase in	≥ 25% increase
Response	≥ 30% decrease	≥ 50% decrease	≥ 50% decrease

In Sweden, the guidelines for neuroradiological tumour evaluation[17] do not specify how to measure tumour size, but the tumour response is evaluated according to the RANO criteria. The guidelines contain several items that help the assessment; however, they do not impose any structure on the report[17]. As a result, the reporting remains subjective and variable, as no criteria are imposed in the national and international care program[12,18]. A recent initiative, BT-RADS[19], aims to standardise and simplify reporting and evaluation of brain tumours by providing a simple list of criteria for the clinician to fill and a resultant management recommendation. BT-RADS adopted the tumour size measurement from RANO as well. Despite more and more available algorithms for semi- or fully automatic tumour segmentation, the change in tumour size is still measured using approximations.

Limited evidence supports the use of volumetric tumour size evaluations to improve patient outcomes[22–24]. A substantial body of literature, however, indicates that volumetric tumour size evaluation is more representative of actual tumour growth (and thus progression) and less sensitive to subjective factors. Fathallah-Shaykh et al.[25] showed that the availability of automatically generated tumour labels and the derived tumour volume resulted in earlier detection of low grade glioma (LGG) progression (14 vs 44 months) at smaller tumour sizes (57% vs 174% volume increase), as compared to visual estimation[25]. Berntsen et al.[26] compared the visual and 2D tumour volume measurements to label-derived volume in patients suffering from glioblastoma (GBM). They found that visual and 2D evaluation accuracy was moderate – within the 66 to 68% range. Gui et al.[27] and Jakol et al.[28] indicated that radiological reports still determine a stable disease when an ~11% increase in tumour volume was present in patients with LGG. According to RANO, this volume increase corresponds to stable disease. However, stable disease, as determined on two consecutive examinations, may be, and often is[27,28], progressive disease when comparing the tumour volume in the latest scan with the baseline examination.

The scanning conditions and image parameters may also impact accurate tumour volume measurement. Studies by Schmitt et al.[29] and Reuter at al.[30] explored the impact of image slice thickness, head rotation, and placement in the scanner on the accuracy of tumour response assessment. The authors showed that measuring tumour size with 2D measurements was subject to more variability and less reliability than volumetric segmentation when analysing various scanning conditions.

To sum up, volumetric tumour measurement is expected to find its way into clinical workflow through automatic BTS tools. From a practical perspective, providing clinicians with accurate measures of volume change will reduce uncertainty in reporting. It will also provide a less variable basis for evaluating the impact of volumetric measurements on patient outcomes.

1.2 Deep Learning for BTS on MR Images

As in many other disciplines, supervised DL methods (in particular convolutional neural networks – CNNs) have become the default choice of algorithm for medical image segmentation[31,32]. The number of articles that describe or validate a DL algorithm for BTS has been increasing at a substantial rate in recent years (Figure 2).

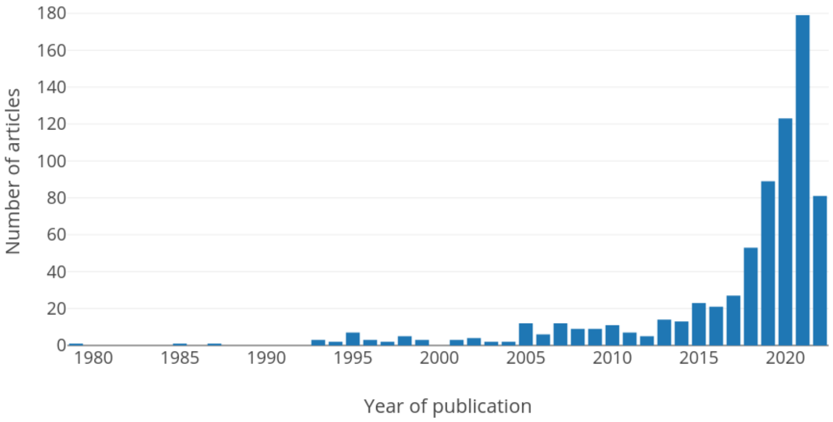

Figure 2. PubMed search result of "brain AND (tumor or tumour) AND segmentation AND (deep learning OR DL OR artificial intelligence OR AI)" on 6/05/2022.

One factor that facilitates the growth of the field is the availability of training and testing images. The Brain Tumour Segmentation challenge BraTS[33], held annually since 2012, offers a publicly available, expanding database of annotated training and testing images. In 2021, the database consisted of images of 2000 patients[34]. Over the years, alongside the increasing number of available images in the database, the performance of the best algorithms has also been increasing[1] (Figure 3).

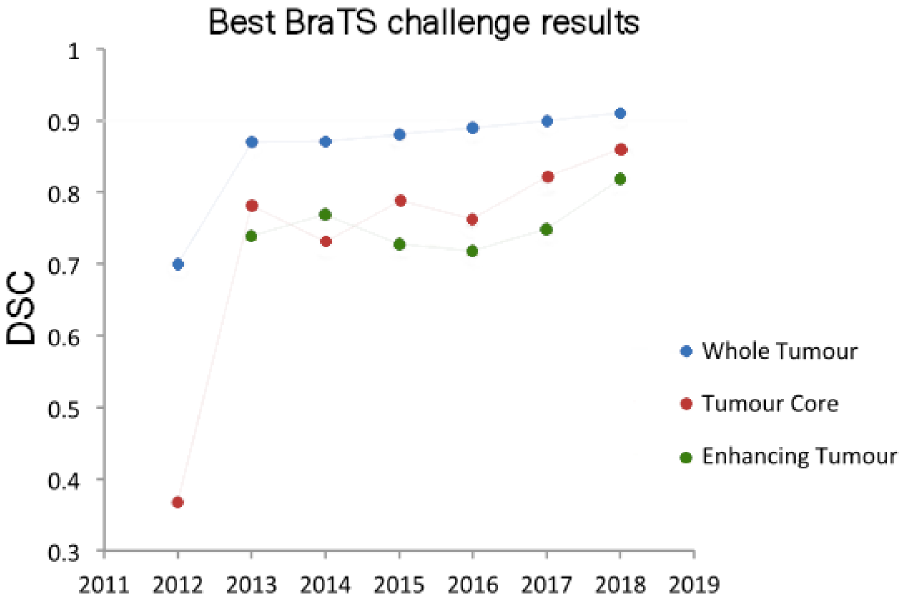

Figure 3. The best DSC scores achieved in the BraTS challenge in years 2012 – 2018. Adapted from Ghaffari et al.[1] © [2020] IEEE.

While BraTS is a crucial contribution to benchmarking DL BTS algorithms, the results achieved on BraTS images themselves do not automatically translate to the expected clinical performance. Before I describe the process of evaluating the diagnostic efficacy of a DL BTS method[7,35] in the next chapter, I will provide basic information on how supervised artificial neural networks work, which elements determine segmentation model performance, as well as the strengths and limitations of DL BTS in clinical applications.

1.2.1 Artificial Neural Networks

An artificial neuron models the most basic properties of a biological one. Just like in a biological neuron, it takes multiple *inputs*, *sums* up the incoming weighted signals, and passes the signal on if the sum exceeds a threshold (indicated by an *activation function*). Figure 4 shows a schematic representation of an artificial neuron.

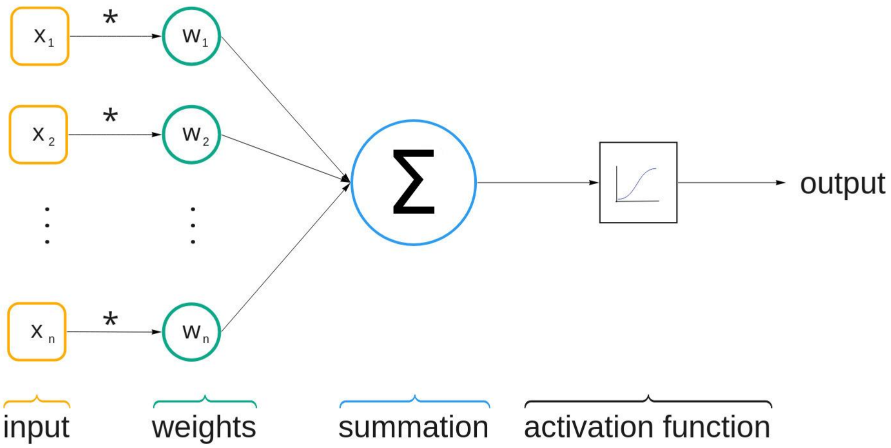

Figure 4. A schematic representation of an artificial neuron (perceptron).

Artificial neural networks are built by stacking neurons in a layer and stacking layers in a network. The number of neurons in a layer corresponds to the number of features that this layer can learn; stacking multiple layers together results in the network's ability to learn more complex and abstract features which best predict the correct output (discriminative features).

A CNN is a particular artificial neural network inspired by biological visual processing[36–38]. The first computational model that mimics a cat's visual cortex – neocognitron – was proposed by Fukushima in 1980[36]. The biological model contains two types of hierarchically stacked cells: simple and complex. The simple cell recognises simple features, e.g. lines, in a specific orientation and location in the visual field. A simple cell layer, therefore, has a cell for each location and orientation of the feature. The complex cell takes input from multiple single cells of specific orientation at different locations. As a result, a complex cell layer recognises features regardless of its location in the visual field. This process was reconstructed in the neocognitron, where layers of simple and complex cells were stacked alternately. Stacking simple cell layers leads to recognition of global, complex features in the deeper layers, while the complex cell layers impose location-invariance of the detected features[36]. Computationally, location invariant feature detection is equivalent to convolution with small kernels followed by pooling[38]. Convolutional and pooling layers are the building blocks of CNNs (Figure 5). The values of the kernels are the network parameters that are determined during training a model.

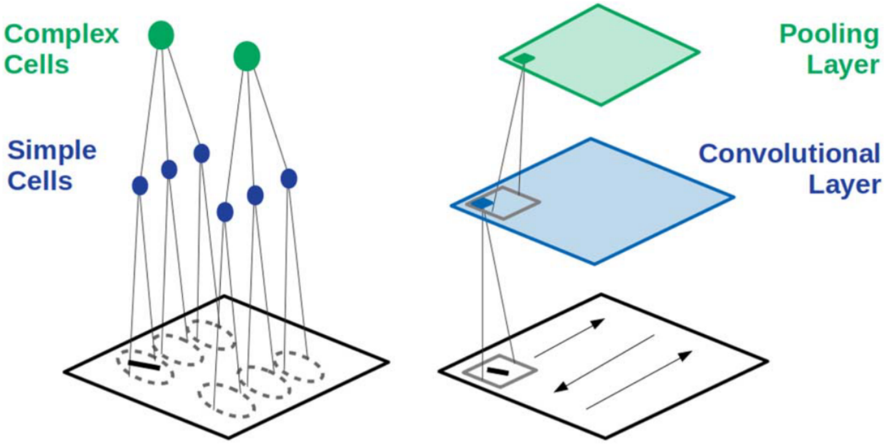

Figure 5. Comparison between the basic visual cortex structure[37] and CNN operation. Simple cells in cats' visual cortex (left, blue) respond to simple features in a specific orientation. A complex cell (green) receives input from many simple cells making their responses more spatially invariant. In the right image, CNN replicates these processes. In the first convolutional layer (blue) image is filtered with a small filter (grey box) at every location creating a simple feature map. The feature map is downsampled in a max-pooling operation, which introduces a certain level of rotation and shape invariance. Reprinted under CC BY 4.0[21] licence from Lindsay et al.[39]

Deep CNNs are built of multiple hidden layers. A hidden layer is any layer other than the network's input or output layer. Classical deep networks are feed-forward models: they learn and recognise features hierarchically because the output of one layer becomes an input to the next layer in the network; the more layers in a network, the more complex and abstract features can be recognised[40]. For example, suppose the outputs of the first convolutional layer with two kernels are maps for vertical and horizontal lines in an image. In that case, the second convolutional layer will result in feature maps that combine the vertical and horizontal edges, e.g. crosses, corners or skewed lines. The third layer would extract features that combine the crosses, corners, lines, etc. The pooling layers in between, next to reducing the size of the feature maps, reduce the network's sensitivity to feature details, such as translation or rotation, making it more robust.

The first CNN was proposed by LeCun et al. in 1989[38]. The CNN architecture resembled the model proposed by Fukushima[36], but the training procedure relied on backpropagation[41]. Backpropagation became a crucial component of training feed-forward networks as it efficiently calculates gradients of an error between the predicted output and the true output for each parameter in a network. Finding such network parameters that the error is minimised is at the core of DL.

Overfitting is one of the pitfalls of DL models. A model that shows high accuracy on training images but performs poorly on images not included in the training set is overfitting. Such a model has learned features that are specific to the training data but do not generalise well to the whole possible distribution of best discriminating features. Several strategies have been developed to minimise overfitting. If available, large and varied training sets are advantageous. Otherwise, the model can be modified with regularisation (penalising learning complex patterns) or dropout (setting certain parameters to zero).

1.2.2 TRAINING

In an untrained network, the parameters (weights/kernels) are initialised as random numbers. The training process adjusts the parameters to learn the most discriminative features of the training data that best predict the output. In our case, the output is labelled image voxels (e.g., tumour or not-tumour). The training occurs in three steps that are iterated until the network can perform the task with satisfactory accuracy:

> I) the network processes a training image (i.e., a pre-labeled image, but the labels are not fed to the network) and produces an output based on current parameters,
>
> II) the current output is compared to the true labels of the training image by calculating an error between the output and the true labels,
>
> III) the parameters are updated through backpropagation[41] and an optimisation algorithm so that the error between the output and the true label is minimised.

The training procedure depends on several hyperparameters that determine how effective and efficient the training will be. First, the function that calculates the error between the current output and the true label (cost function) must be determined. Then, we must specify an optimisation algorithm and learning rate to find the optimal parameters in each training run to minimise the cost function. Other hyperparameters include the number of training cycles to be run with all of the data (epochs) or whether all or a sub-set of training images are passed in each training iteration (batch size).

1.2.3 Data

The performance of DL models is directly connected to the quantity and quality of the data used for training and validation; of particular importance is how representative of the population the training data is. Therefore, data preparation is one of the most critical and time-consuming tasks in developing DL models[42]. A fundamental aspect of data preparation for DL applications is preprocessing. Each feature in a processed image must be represented by a similar value distribution across the whole data sets[40]. Preprocessing is particularly important for segmenting MR images. For example, tissue properties are typically discriminated based upon the contrast between pixels/voxels rather than the value of pixel/voxel intensity. The intensity profiles of several MR images can vary depending on the scanner manufacturer, sequence, and acquisition parameters. They are also often influenced by artefacts. Signal intensity, therefore, must be normalised across images fed to a DL model for optimal performance.

In supervised learning, the images used for model training and validation must be annotated. In BTS, high-quality reference segmentation labels must be provided. As mentioned in Section 1.1, manual delineation of tumours is resource-expensive, especially in large data sets that are usually required for DL studies. The reference labels are furthermore subjective, even if done by experienced professionals. Access to large data sets with reference annotations may be particularly important for the segmentation of regions of interest that vary in visual appearance, shape, size and location – such as brain tumours. Therefore, the availability of high-quality data may be an issue for model developers. Ethical and legal considerations of data acquisition, use, and storage are other essential aspects that affect data accessibility in DL applications for healthcare. Publicly available datasets have been a significant catalyst for advancing the DL BTS field.

1.2.4 Advantages and Limitations

One of the most significant advantages of DL-based methods is their ability to process high volumes of images quickly. Once trained, they can fully process an individual input image in a few minutes or seconds[32,33], compared to average 10 minutes or more for manual segmentation[43,44]. Thanks to their pattern-learning ability, these methods are expected to relieve clinicians from repetitive and often laborious tasks. For medical image and brain tumour segmentation, the DL models achieve expert-level accuracy, at least in proof-of-concept and standalone diagnostic accuracy validation studies[1,11]. Automatic segmentation algorithms are furthermore presumed to decrease interrater variability, even when the segmentation labels need editing[45].

The fundamental elements of DL design, which can lead to its exceptional performance, are also the reason for its limitations. Data fed to an algorithm is one such element. In Section 1.2.3, I described how the data is a part of the model and determines how well the model will perform. A big concern with DL models for MR image processing is the sensitivity of the model to the variability in the data. The variability arises due to different image acquisition parameters and protocols, scanner manufacturers, artefacts, etc.. Therefore, comprehensive and efficient preprocessing is key to a successful segmentation, although it can be challenging[32,46].

The best solution to the problem is training a model with a large data set that contains a representative sample of lesions, scanners, acquisition parameters and clinical settings. Publicly available data sets, while providing an ever-increasing number of available images, are curated and may not include a representative sample of the whole population, potentially incurring biassed outcomes. On the other hand, personal data protection laws heavily regulate acquiring clinical images independently, which limits the accessibility of clinical data. Independently collected data must also be labelled – another previously described resource issue. The perceived quality of the annotations may also, to some extent, be dependent on the local guidelines and, therefore, may not be generalisable globally.

The potential for high accuracy of DL models comes at a price of complexity of these algorithms. Currently, DL tools for healthcare applications do not support domain experts in understanding the tool outcomes in a way that is compatible with expert reasoning[47]. DL models are still “black boxes” to physicians[48]. In high-stakes domains like healthcare – where the decisions are supposed to be based on sound and transparent reasoning – the correlational principle of DL algorithms' operation makes end-users reluctant to trust the technology[49,50]. The sustainable uptake of DL CDS tools will depend on the users' acceptance of the tools and willingness to incorporate them into their workflow.

1.3 DL Image Analysis Technology Assessment in Radiology

Neuroradiology as an evidence-based practice[51,52] requires that diagnostic imaging tools are thoroughly validated; the evidence of the tools' accuracy must be convincing. The tool's users must know how to use it and interpret the information it provides[53]. Crucially, they must sufficiently trust that the tool works accurately. Introducing AI/DL tools in clinical practice for CDS poses a challenge not only from the technical point of view but also due to the DL explainability challenge mentioned in the previous section. While domain-tailored explanations and interpretations of DL tool outcomes may presently be lacking, other solutions can be proposed to promote users' acceptance and trust in the technology as important first steps.

Regardless of a clinical tool's underlying technology or decision-making principles, its validation must follow an evidence-based approach. The most commonly referred to and used framework for evaluating diagnostic imaging efficacy has been described by Fryback & Thornbury[35]. It proposes six levels of efficacy: technical, diagnostic accuracy, diagnostic thinking, therapeutic, patient outcome, and societal. Efficacy evidence at each level is meaningful only when the previous levels were achieved. Van Leeuwen et al.[7] refined this hierarchical framework to accommodate diagnostic efficacy assessment of AI tools for medical image analysis (Table 2).

Despite the explosion of the number of articles validating AI algorithms for medical image processing and analysis, the evidence for their efficacy as CDS tools is scant. Van Leeuwen et al.[7] reviewed 100 CE-marked AI tools for clinical radiology available at the time at the AI for Radiology website (www.aiforradiology.com). The evidence for Level II diagnostic efficacy was available for 36 products. Only 18 of the 100 products had published evidence at the third level or higher[7]. None of these tools, however, were automatic BTS tools.

Table 2. The hierarchical model of diagnostic imaging efficacy proposed by Fryback and Thornbury[35] and refined to evaluate the impact of AI on diagnostic imaging procedures. Adapted from van Leeuwen et al[7] available under CC BY-NC-ND 4.0[21].

Efficacy level	**Description**	**Measures of efficacy**
I_t – technical	Proof-of-concept validation shows technical feasibility of the tool/algorithm	Reproducibility, error rate
I_c – potential clinical	The tool/algorithm demonstrates the feasibility of clinical application	Correlation to established/other clinical processes and examinations
II – diagnostic accuracy	The standalone performance of the tool/algorithm is evaluated	Standalone DSC, sensitivity, specificity, ROC analysis
III – diagnostic thinking	The added value of the tool/algorithm to the diagnostic thinking process is demonstrated	Impact on radiologists performance and judgement
IV – therapeutic	The added value of the tool/algorithm to the patient management process is demonstrated	Effect on treatment and further examinations
V – patient outcome	The added value of the tool/algorithm to the patient outcome is demonstrated	Effect on life quality, survival, or morbidity
VI – societal	The economic impact of the tool/algorithm on society is evaluated	Effect on cost adjusted for quality of life

Another review by Ebrahimian et al.[54] assessed 118 FDA-approved AI/ML-based algorithms. The overview of the algorithms' validation studies focused on attributes of the technical (Level I) or diagnostic accuracy (Level II) efficacy, such as the number of patients' images used for training and validation or the number of readers assessing the results. In 52/118 algorithms, the source or the number of images used for training and validation were not described sufficiently – even when a study was referred to as being a clinical validation (n=35). Furthermore, most of the reviewed algorithms' validation studies (n=99) did not provide sufficient information regarding the label annotators involved in the study[54]. Only one reviewed algorithm was intended for BTS (VBrain by VYSIONEER[55]). The validation studies available for that algorithm evaluate its diagnostic efficacy at Level II[56,57].

A recent (31st of May 2022) visit to the AI for Radiology portal revealed that the number of available certified products nearly doubled (n=198) since van Leeuwen et al. accessed the resource. Forty-seven of the listed products were tools for processing neurological MR images, and two were for BTS (Sens.ai by Graylight Imaging[58] and Brain Tumours Application by BioMind[59]). Of these two products, a peer-reviewed article presenting diagnostic accuracy (Level II) validation was available only for the Sens.ai[60]. A search of FDA-approved algorithms on the website used by Ebrahimian et al.[54] did not reveal any new BTS algorithms.

A crucial aspect that facilitates assessing the efficacy of the DL BTS algorithms that was not explicitly named by Fryback and Thornbury,[35] but was recognised by van Leeuwen et al.[7] is the reproducibility of the validation studies. While various understandings of the term reproducibility occur in different fields, the most common definition, which I will also follow, has been proposed by the National Academies of Science, Engineering, and Medicine:

"reproducibility is obtaining consistent results using the same input data; computational steps, methods, and code; and conditions of analysis"[61].

It may not always be possible to evaluate the reproducibility of commercially available algorithms independently. As described above, diagnostic efficacy validation studies are not always available for the products. When the validation studies are available, the tool most likely must be purchased for independent testing, and the training and validation images may not be available. Still, the commercial products are likely based on scientifically developed algorithms[55,56]. The reproducibility of scientifically validated algorithms is imperative for producing sufficient evidence for their accuracy in a clinical setup.

Reproducing DL algorithm validation studies is inherently challenging due to the complex, multiparametric, and indeterminate nature of the algorithms[31,62].

Pineau et al.[62] further pointed out that insufficient documentation of the models and validation procedures hinders reproducibility. The authors[63] proposed a checklist to aid the description of the DL models and validation procedure to facilitate reproducibility. Pineau et al.'s checklist has been adopted by one of the most influential societies in the field, the Medical Image Computing & Computer Aided Intervention Society (MICCAI[64]), which urges authors to fill the checklist upon submission to their conferences.

The topic has also been explored by Renard et al.[31] in DL applications for medical image segmentation. The authors explore the sources of variability in DL models and performance assessment practices in a literature review. They also propose recommendations that facilitate reproducibility. The recommendations include items that should be reported in DL validation studies. These items coincide with those proposed by Pineau et al. but are grouped differently. Through the literature review, the authors identified only three articles that provide sufficient description of the algorithm and validation procedure (according to their recommendations) for reproducing the results. Two algorithms[65,66] were developed and validated for segmenting brain tumours on MR images.

This last theoretical investigation raises questions regarding the documented evidence for the diagnostic efficacy of the many DL algorithms proposed and validated for BTS. The immense developments in the field have generated a breadth of proof-of-concept studies that possibly lack the scientific foundations required to assess their clinical suitability simply because they are not reproducible.

1.4 Clinical Relevance and Stakeholder Perception of DL BTS

Proof of standalone accuracy of a DL BTS tool (Level II of the diagnostic efficacy model) alone does not guarantee trust and sustained use by clinicians[67,68]; it is necessary to evaluate the impact of the tools on diagnostic thinking or diagnostic processes (Level III)[69]. Such investigation will improve the chances of a successful implementation and build the foundation for evaluating the usefulness of volumetric tumour assessment for patient outcomes.

The number of studies that evaluate radiologists' interaction with DL tools in a clinical scenario or a clinical setup is scant[70,71]. van Garderen et al.[72] assessed the accuracy and usefulness of LGG labels generated by a DL tool[73] integrated into the clinical workflow – EASE. Of the 55 patient cases evaluated with EASE, tumour labels in 36 cases were of acceptable quality to be used in the disease status (stable or progressive) evaluation. These labels, however, did not change the radiologist's decision about the disease status compared to the visual assessment (32/36 patients were assessed as having stable disease, which could explain why there was no difference in the assessment between the visual and segmentation-based evaluation). While the impact of the results themselves may be limited, the importance of carrying out such work in order to increase utilisation is nevertheless paramount.

Assessing the accuracy of DL-generated lesion labels can be considered a relatively simple case, as labels and source images are always available to radiologists for quality evaluation[71]. In other applications, when an algorithm makes a prediction based on a large amount of data that is beyond human capacity to process, assessing the validity of DL output is not easily accessible. In either case, an appropriate level of trust in the accuracy of the prediction is imperative for sustained use of the tools. A well-known concern in radiology that could influence the trust in DL BTS is inter- and intra-rater variability[74–76]. The resulting variability in the "gold standard" reference limits the accuracy with which an algorithm can be assessed. Similarly, there is a risk that the same segmentation will be assessed differently by different readers.

As described above, volumetric measurements have not been commonly adopted for radiological tumour status assessment[19]. The availability of DL tools for BTS and volume estimation in the diagnostic workflow will provide a new metric for the radiologists in the decision-making process. Appropriate levels of trust in the accuracy of the new information will be crucial for the sustainable and beneficial adoption of AI in medical image assessment workflow.

An appropriate level of users' trust in computer-aided diagnosis (CAD) was identified as an essential requirement for successfully implementing CAD tools[53]. In healthcare, the decisions are expected to be algorithmic i.e., based on well-defined, evidence-based rules and traceable professional reasoning. However, because AI, and especially artificial neural network-based tools, generate results in a non-transparent way, future AI users will likely have to develop appropriate levels of trust in AI without access to the same reasoning for a decision as with another medical expert or rule-based CAD system[77]. Still, specific demands that foster trust must be fulfilled to successfully implement AI tools in radiological workflows[78]. What these demands are and what trust entails in that context have not been agreed on[79]. It has also been clear that a successful design and evaluation of the CDS tools requires collaboration between the users and the developers[80,81]. Despite being the key actors in the implementation process, the users are not the only ones that should be involved. Developing successful implementation solutions that foster building trust in AI requires identification of crucial stakeholders, their engagement in the processes[82] and their specific role in it.

2 THESIS AT A GLANCE

In this doctoral project, I aimed to assess and advance the scientific foundations and the evidence supporting the translation of DL algorithms into clinically beneficial and acceptable tools for BTS on MR images. The scientific work presented in this thesis follows four themes that address gaps and issues described in the introductory chapter. The two angles of my investigation included: an evidence-based technology assessment of BTS algorithms (**Themes I and II**), and the stakeholder's perspective on the information generated by BTS algorithms and a broad understanding of trust in AI in radiology (**Themes III & IV)**. Given the multidisciplinary nature of the investigated problem, each theme follows a different methodological approach. The themes, aims of each investigation, methodological approach, and the results are presented in Table 3.

Table 3. The summary of the scientific work included in the thesis. P – paper number in which the investigation was described.

Theme	Aim	Methodology	Results	P
I: Mapping the field	To map the field of BTS on MR images through the lens of clinical relevance and suitability.	A scoping review conducted according to a published protocol.	The prevalent study design in the field provides evidence for the technical accuracy of the methods. Preprocessing descriptions are, however, insufficient to validate clinical accuracy.	I & II
II: Reproducibility & replicability of BTS algorithms	To reproduce and replicate two theoretically reproducible BTS methods; to evaluate whether established reproducibility criteria are sufficient.	Independent implementation and validation of the methods using the original data set and in-house collected images.	Only one method was reproducible and replicable. Preprocessing of the other one was not sufficiently described. The reproducibility and replicability criteria were updated.	III
III: Radiologists' perception of	To evaluate: (a) radiologists' accuracy and confidence in	A questionnaire with 10 respondents who	Tumour labels increase the accuracy of the estimation; the	IV

BTS labels	estimating visual tumour volume increase, and (b) the impact of tumour labels on diagnostic thinking (level III of diagnostic efficacy).	assessed 20 unlabeled/ labelled patient cases / 40 MR examinations.	readers are more confident that label-derived volume increase is more correct than their visual estimation. There was a poor inter-reader agreement regarding the label quality.	
IV: **Trust in AI in radiology**	To identify the knowledge gaps in stakeholders' perspectives on trust in AI in radiology and how to facilitate building the trust.	Semi-structured interviews were conducted with twenty-six stakeholders and analysed in an iterative coding process.	Four areas of trust emerged that relate to: the demand for reliability, transparency, quality verification, and inter-organizational compatibility of AI solutions.	V

3 THEME I: MAPPING THE FIELD

3.1 Knowledge Gap & Aims

While many literature surveys of automatic brain lesion segmentation methods have been published[83–88], a comprehensive and systematically conducted review of the published and validated methods had been missing. More importantly, assessment of the clinical suitability of the methods had often been neglected. The aim of this investigation, therefore, was to: (1) understand the clinical suitability of published segmentation algorithms and their validation methods, (2) outline limitations, gaps, and challenges in the field, and (3) suggest ways of facilitating translation of the research for clinical use.

3.2 Materials & Methods

A scoping review methodology has been actively developed to facilitate comprehensive and systematic mapping of a research field, to summarise and synthesise evidence available in the field, and to identify research gaps[89–92]. The standardised approach to conducting scoping reviews assures high quality of the findings, which are often used to inform policies and practice[92]. The scoping review methodology follows five (optionally six) well-defined stages[89,90,93] that should be defined in the protocol.

Stage 1: identifying research questions – at this stage, broad research questions pertaining to the aim of a scoping review are defined.

Stage 2: identifying relevant studies – a search strategy, including databases and query phrases, is proposed to identify all relevant literature; a procedure and criteria that identify relevant studies are established.

Stage 3: selecting relevant studies – final exclusion criteria are identified according to which non-eligible articles are excluded from the scoping review.

Stage 4: data charting – categories of information relevant to the study aims and research questions to be extracted from eligible articles are proposed.

Stage 5: collating, summarising, and reporting the results – a strategy is proposed to present the results so the research questions are answered.

Stage 6: consultation – this is an optional phase where relevant stakeholders are consulted regarding the findings of previous stages to ensure the applicability of the results in the scoped field.

Publishing a protocol prior to conducting the study can enhance the scientific quality of the work. A peer-reviewed study protocol is exposed to scrutiny and ideas of fellow researchers, which can further add to the value of conducted research. Good research practice demands that the protocol is designed separately and followed carefully during the actual literature review. This separation also provides a basis for detailed documentation of any divergence from the original protocol. As a result, any epistemic drift that may occur has to be justified and thoroughly accounted for.

The epistemic drift in my work resulted from a discrepancy between my initial method-oriented and quantitative understanding of the brain lesion segmentation problem and what is necessary to advance the implementation of the research findings into clinical practice. Publishing a study protocol proved valuable in this situation; I followed the original protocol closely and answered the original research questions to the extent possible while having grounds to justify the changes meticulously.

The original research questions I posed concerned identifying:

1) common image processing steps in a segmentation framework
2) underlying mathematical and computational theories
3) efficacy of the algorithms
4) limitations of the methods concerning clinical use
5) commonly used MR images for algorithm validation.

A preliminary screening of articles identified in Stage 2 prompted me to challenge the scientific and practical value of the answers to research questions 1, 2, and 3. Many published reviews already describe algorithmic approaches to medical imaging, specifically brain lesion segmentation[83–88]. The scientific novelty of answers to questions 1 and 2 is therefore limited.

Furthermore, a comparative evaluation of the efficacy of the proposed methods (question 3) is complex and could not be done in the scoping review.

To harness the scientific value of the scoping review and ensure the novelty of my work, I consulted clinicians at this stage. Doing so prompted me to focus on the clinical relevance of the published studies – a perspective largely overlooked in the scientific literature. I wanted to understand the suitability of the proposed segmentation methods to be developed into clinical tools as well as challenges and limitations that hinder clinical implementation, and propose a way forward.

In Stage 2, the search strategy designed in the protocol was applied to three databases (PubMed, IEEE Xplore, Scopus). To deal with many returned articles efficiently, I screened the articles' titles, abstracts, and methods to determine their eligibility in Stage 3. Next to a hierarchical approach to identifying eligible studies described by Arksey and O'Malley (2005)[89] and Levac et al. (2010)[90], the authors also propose a strategy for identifying the most relevant eligibility criteria by screening the whole sample of articles iteratively. This approach was not feasible for me, given the high number of articles (n=2500) identified in Stage 2. Instead, I relied on the information gleaned during the consultation. During the abstract screening phase, I also refined the study selection process by identifying prevalent themes in a randomly selected subsample of 100 abstracts. The whole sample was subsequently screened according to the refined criteria. Data charting in stage 4 and data reporting in stage 5 were adjusted from the protocol to account for the shift to a clinical focus of the review.

3.3 Results

The following study design attributes gleaned from the eligible sample (n=441) were the most common. A brain tumour segmentation (n=216) algorithm that uses artificial neural networks (n=85) to perform the segmentation is evaluated on data with reference segmentations available either non-publicly (n=254) or publicly (n=217), or both. The data used to validate the algorithm include multi-sequence (n=307) scans of 50 patients or fewer. One or two expert raters provide the reference segmentations. The images undergo the following preprocessing steps: intensity normalisation (n=224), bias field correction (n=192), brain extraction (n=190), and image (co-)registration (n=179); however, it is not commonly stated whether these steps are integrated with the automatic segmentation procedure or performed independently. The automatically generated segmentation is evaluated by comparison to the reference segmentation using overlap measures: DSC, Jaccard coefficient, sensitivity, specificity, and positive predictive value (PPV). The information regarding the average processing time performed on the specified computational system is given (n=233), but the algorithm is not available to download for independent testing (n=417).

The most significant limitation of the prevalent study design is an insufficient description of methods (particularly preprocessing) to enable reproducibility, and scarce online access to the whole processing chain to validate a method independently.

4 THEME II: REPRODUCIBILITY & REPLICABILITY OF BTS ALGORITHMS

4.1 Knowledge Gap & Aims

Renard et al.'s work[31] was the basis for my pursuit of an existing algorithm that can be used as a prototype of a clinical tool. The study by Renard et al.[31] provides recommendations for describing DL frameworks to facilitate reproducibility. The authors also identified three studies that fulfil the recommendations; two of those validated BTS methods on MR images[65,66].

The proposed recommendations mention preprocessing; however, they only highlight the need to report reasons for data exclusion and data augmentation procedures. My scoping review[32] indicated that a sufficient description of the preprocessing does not receive due emphasis. An attempt to reproduce and replicate the validation results of the two DL BTS algorithms[65,66] allowed me to:

> a) substantiate the technical efficacy (Level I) and diagnostic accuracy efficacy (Level II in the diagnostic imaging efficacy model[7,35]) of the segmentation methods;
>
> b) to verify whether the proposed reproducibility criteria are sufficient in terms of preprocessing description.

4.2 Materials & Methods

The reproducibility of the two DL BTS algorithms: DeepMedic[65] and an algorithm developed by Pereira et al.[66] (that I will refer to as ProfoundDoc) was determined by implementing the algorithms and the preprocessing pipelines according to the description provided in the original validation articles. In my reproducibility analysis, both algorithms were trained and tested on the BraTS[33] 2015 data set that was used in the original studies. A successful reproduction of the results meant that the mean values of DSC, PPV, and sensitivity of the test set tumour labels generated by our implementation were comparable to those reported originally. The unsuccessful reproduction was investigated to identify the processing step(s) that were not sufficiently described for reproduction.

Following a successful reproduction, the replicability of DeepMedic was assessed by testing the model trained for the reproducibility analysis on an in-house collected set of images. We then compared the segmentation accuracy on the in-house set to the accuracy achieved on the BraTS testing set. The in-house data were collected and handled according to the Swedish Ethical Review Authority's approval (DNR 702-18). The requirement for informed consent for this secondary-use study was waived by the approval. No other personal data than the images and tumour grade diagnosis were used in this study.

4.3 Results

The attempt to reproduce ProfoundDoc results did not succeed; I could not implement the preprocessing procedure, even after consultation with the leading developer of ProfoundDoc, since specific parameters were not originally specified and could not be retrieved. This failed reproduction of the preprocessing chain was the basis for updating the reproducibility and replicability checklist for medical image segmentation. Because such important information was available for DeepMedic, I successfully reproduced and replicated the validation results originally reported with it (Table 4).

Table 4. Reproducibility and replicability analysis results for DeepMedic[65]. DSC – Dice similarity coefficient, PPV – positive predictive value, WT – whole tumour, TC – tumour core, ET – enhancing tumour, HGG – high grade glioma, LGG – low grade glioma.

	DSC			**PPV**			**Sensitivity**		
	WT	**TC**	**ET**	**WT**	**TC**	**ET**	**WT**	**TC**	**ET**
Reproduction	0.85	0.68	0.64	0.85	0.83	0.62	0.88	0.64	0.70
Replication HGG	–	0.77	–	–	0.72	–	–	0.88	–
Replication LGG	0.73	–	–	0.83	–	–	0.67	–	–

5 THEME III: RADIOLOGISTS' PERCEPTION OF BTS LABELS

5.1 Knowledge Gap & Aims

The evidence indicating that volumetric tumour growth assessment directly translates to therapeutic benefits (Level IV of diagnostic efficacy) for patients has not been established. An important intermediate step that needs to be fulfilled for the therapeutic benefit to happen is a change in diagnostic thinking[35]. If quantitative tumour growth evaluation results in a more accurate assessment or more confidence in the tumour status assessment, it benefits the diagnostic thinking process (Level III in the diagnostic efficacy model[7,35]) and builds a foundation for evaluating therapeutic benefit.

In this line of investigation, we aimed to evaluate radiologists' confidence and accuracy in visual identification and quantification of ET growth. We further assessed whether the availability of ET labels and their volumetric measurements change the diagnostic thinking (Level III) efficacy.

5.2 Materials & Methods

In a structured reading/interview session, we asked ten radiologists to identify and quantify ET volume increase in twenty pairs of GBM patient MR examinations and assess their confidence in the accuracy of their assessment. We then compared the accuracy of their assessment to their confidence in the accuracy of automatically generated ET volume increase derived from ET labels. The study design was refined following a pilot reading/interview session with three certified neuroradiologists who were excluded from participation. T1w images with contrast enhancement and corresponding ET labels were retrieved from Brain-Tumor-Progression collection in The Cancer Imaging Archive (TCIA)[94].

Since the clinical quality of the tumour labels available in the TCIA collection was evaluated as insufficient in the pilot, a new set of ET labels for the MR images was manually generated by one of the certified neuroradiologists that participated in the pilot. The new set of labels was used for the main data collection part of the study, and the labels were presented as generated automatically by a DL algorithm. By doing so, I wanted to avoid bias against the segmentation quality by simulating a scenario wherein radiologists have access to an automatically generated tumour label of expert-level quality. In a follow-up session, I asked the radiologists to rate the quality of the same labels, which were then presented as generated manually.

We collected all interview data anonymously, so informed consent from the participants was not needed. The images used in this study have been made publicly available by TCIA[94] in compliance with the United States Health Insurance Portability and Accountability Act of 1996[95]; therefore, no ethical permit was needed for this study.

5.3 Results

The results indicated that visual estimation of ET growth is difficult; the readers tended to underestimate the percentage of ET growth, and their confidence in the task was moderate. The readers had more confidence in the accuracy of the automatically quantified ET volume increase than in their assessment. Overall, we observed a positive impact of ET labels on MR images on diagnostic thinking (Figure 6); however, not unanimously. The quality assessment of the ET labels, presented as automatically generated, showed poor inter-rater agreement. Re-evaluation of the same ET labels' quality, but when the labels were presented as generated manually, did not improve the inter-rater agreement.

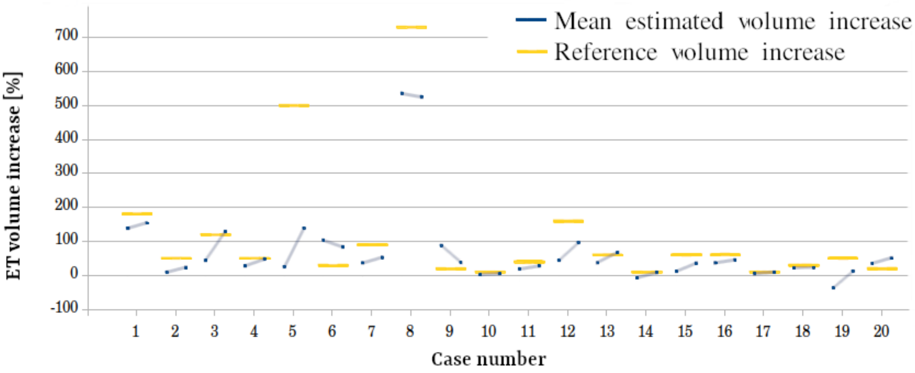

Figure 6. Results of visual estimation of ET volume increase for each case. The first point of the blue dash shows a mean ET volume increase estimated for a given case based on MR images only, while the second point shows the mean value estimated when ET label was displayed.

6 THEME IV: TRUST IN AI IN RADIOLOGY

6.1 Knowledge Gap & Aims

Appropriate levels of trust in AI are fundamental for successful implementation and sustained use of the technology for CDS. Trust in AI is often approached from the algorithm explainability and interpretability angle, as a way to provide the users with some reasoning for the output. While the users' attitude towards CDS tools is crucial, building trust in new technology requires a much more comprehensive approach and involvement of all relevant stakeholders[82]. While the implementation challenges[96,97] and opinions of AI in healthcare among both the users and other stakeholders have been investigated[98,99], solutions and particular conditions that facilitate trust need to be defined. In this study, therefore, we identified knowledge gaps in the intersection of AI, trust, and stakeholders' perspectives on AI, and how trust in AI for radiological decision support can be improved.

6.2 Materials & Methods

A semi-structured interview study with a hand-selected sample of twenty-six respondents was conducted to elicit stakeholders' perspectives on trust in AI tools for clinical radiology and healthcare. Nineteen interviewees participated in implementation of AI solutions in their radiological practice while six respondents were involved in deploying AI solutions in other healthcare domains.

The interview questions explored the following three themes, while leaving space for follow up questions tailored to the interviewees' responses:

- current practices and how they will change with DL CDS,
- the relationship between decision-making and normative responsibilities that guide the moral basis of professional thinking,

- the role of management and organisational procedures in the implementation of DL CDS in clinical practice.

The interviews were recorded and transcribed. The transcripts were analysed in a three-stage, iterative coding procedure. The stages consisted of open coding (identifying significant parts of the transcripts), thematic coding (existing codes were aggregated to identify prevalent themes), and theoretically informed coding (the codes were interpreted through the lens of theoretical reflection).

6.3 Results

The theme most commonly mentioned in the interviews concerned the various demands of a given tool and its implementation procedure that facilitate trust in AI. Four theoretically informed dimensions of trust emerged that pertain to the demands for reliability, transparency, quality verification, and inter-organizational compatibility. These dimensions can be divided into substantial and procedural requirements. Within each of these dimensions, specific aspects and conditions need to be fulfilled (Table 5)

Table 5. The four demands to build trust with specific conditions that need to be fulfilled.

	Demand	**Aspects**
Substantial requirements	Reliability	- Volume - Granularity - Bias
	Quality verification	- Methodological rigour - Local validation
Procedural requirements	Transparency	- Standards - Traceability - Explainability
	Inter-organizational compatibility	- Capacity - Control

7 DISCUSSION

In this thesis, I examined two dimensions of the scientific foundation that facilitate the translation of scientifically validated DL BTS algorithms as CDS tools. First, I assessed whether the standard BTS algorithm validation procedure is sufficient to appraise the technology's suitability for clinical applications. Mapping the field in the scoping review (**Theme I)** based on 441 articles revealed that insufficient image preprocessing description and scant availability of the algorithms for independent validation pose a severe obstacle to assessing the clinical suitability of most available BTS algorithms. Experimental investigation of the reproducibility and replicability of DL BTS algorithms described according to established reproducibility checklists (**Theme II**) corroborated the suspicion: the requirements for preprocessing description are insufficient. In this line of investigation, I identified one reproducible and replicable DL algorithm for BTS.

The second dimension took the perspective of the users of the tools – radiologists – on whether and how BTS labels change their diagnostic thinking (**Theme III**), and a wider group of stakeholders on the conditions that lead to trust in AI (**Theme IV**). We found that just the availability of the tumour labels in general increases the accuracy of tumour volume increase estimation. The radiologists have only moderate confidence in visual tumour volume change assessment and tend to trust the label-derived volume measurements more, even when there was a poor interrater agreement on the quality of the labels. While quality verification will have to be a crucial step in incorporating DL CDS in clinical workflow, other requirements of the tools and the implementation infrastructure will have to be met to ensure appropriate levels of trust in the tool, while minimising the inter-rater variability effect and personal biases against DL. These requirements include reliability of the tool, transparency of the decisions, and inter organisational compatibility with the new workflows.

In the following sections, I will discuss these two dimensions in depth and the results of my investigation.

7.1 Technical and Clinical Validity of DL BTS Validation Studies

In **Theme I**, I found that the prevalent design of BTS algorithm validation studies provides a proof-of-concept level of evidence for diagnostic efficacy (level I in the diagnostic efficacy model[7,35]). Even when an algorithm is tested on unseen images acquired from sources different from the training images (external validation, Level II of the diagnostic efficacy model[7,35]), it is not sufficient to assess the performance of an algorithm in a clinical setting. The scoping review findings highlighted a potential issue that limits the clinical relevance of many studies: very few methods are sufficiently described[31,32] and/or are available to download[32] for independent reproducibility analysis. Successful reproduction of the technical validation of an algorithm is a cornerstone for evaluating its diagnostic efficacy[7]. Preprocessing is the particular element of a segmentation pipeline that often does not receive enough attention.

The need for improved reporting to facilitate reproducibility has been recognised by Pineau et al.[62] regarding ML studies in general. Renard et al.[31] further proposed reproducibility criteria for studies that describe and validate DL methods for medical image segmentation. However, in the investigation described in Theme III, I showed that the established reproducibility criteria neglected the preprocessing description. As a result, the criteria are insufficient to facilitate clinical validation of DL BTS algorithms.

It can be speculated that the advances of automated BTS on MRI have been driven by the BraTS[33] challenges that, starting from 2012, have provided a growing database of annotated MRI images. For many scientists and algorithm developers, publicly available image repositories with reference annotations are the only source of images to train and test their algorithms. Even for researchers with close collaborations with hospitals, acquiring clinical data is resource-expensive. BraTS, therefore, has become a standard for evaluating DL BTS algorithms. The 2021 BraTS challenge provided the participants with more than 2000 patients' MR examinations[100] – datasets of this size are otherwise unachievable for most scientists. The more images are available to test an algorithm, the better we can predict how well it will perform on clinical images. There is, however, a caveat. To provide a basis for comparing various segmentation algorithms, the images in BraTS are curated and already preprocessed to a certain extent[33,100]. A clinical image, therefore, must undergo the BraTS preprocessing procedure as a first step to assess the clinical performance of an algorithm trained and validated on BraTS images. The preprocessing procedure is known and recently available through online tools, such as BraTS Toolkit[101] and Cancer Imaging Phenomics Toolkit[102–104] (https://cbica.github.io/CaPTk/). In the replicability analysis described in **Theme III**, I uncovered potential problems with the skull-stripping step in the BraTS Toolkit that affect the segmentation algorithm. This issue would not be detectable in the BraTS image set as all images available are screened for quality.

According to **Papers II & III**, many BTS methods require additional preprocessing steps next to the BraTS pipeline for optimal algorithm performance[32,66]. In the scoping review, I revealed that while the necessary preprocessing steps tend to be mentioned, implementation parameters and whether the pipeline is automatic and integrated with the segmentation method are not commonly stated[32]. It appears that the focus in the field has been on automating, innovating, and evaluating the segmentation algorithms rather than the whole processing chain of steps applied to a raw image. While this is acceptable and common in scientific validation studies, clinical validation requires that the whole processing chain is described sufficiently to be reproducible and replicable technical validation results on raw clinical images acquired externally.

A successful reproducibility of the DeepMedic[65] validation study and satisfactory reproduction of the results on an independent clinical image set indicate that DeepMedic is suitable for clinical testing. It was also, in fact, also used in a commercial BTS tool[55,105]. My findings, however, do not imply that there are no other reproducible DL BTS methods. For example, van Garderen et al.[71] successfully implemented a U-Net developed by Isensee et al.[106] in a clinical workflow. In the reproducibility study, I also relied on the review of relevant and theoretically reproducible methods conducted by Renard et al.[31] and published in 2020. An independent review of the literature to identify theoretically reproducible BTS algorithms might return more studies, including the most recent ones. This, however, was outside of the scope of that study and would not likely provide more insight to answer the research question of whether the established reproducibility checklists sufficiently address the description of preprocessing. Still, the number of theoretically reproducible algorithms for BTS identified by Renard et al.[31]

(n=2), of which only one was experimentally reproducible, contrasted with the number of BTS algorithms included in the scoping review (n=214), gives an idea of the fraction of reproducible studies in the field. It also could be one of the factors contributing to the implementation gap. Even though the best segmentation solutions may remain unpublished (the developers may want to secure a chance of winning competitions that bring recognition and funding), and clinical tools must be commercial products, widespread reproducibility will likely contribute to a quicker translation of the scientific findings into clinical benefits. Furthermore, the wider availability of reproducible algorithms and reproducibility studies could also improve the attitudes towards the algorithms for CDS.

7.2 User's Perceptions and Trust in BTS

Even though estimating tumour volume change using RANO[16] measurements has become the proposed standard in clinical practice[19], visual estimation is still prevalent in clinical reports[12,25,27,107]. In LGG, a stable lesion tends to be determined when a tumour has grown by approximately 11%, while a progressive lesion becomes detectable at ~20% tumour volume increase[27,28]. Even though RANO criteria require a 40% volume increase (corresponding to a 25% increase of the diameter product)[27] to determine a progressive lesion, underestimated volume increase may delay the detection of progressive disease and beneficial resection surgery[108]. In another study, an analysis of retrospective radiological reports of imaging examinations of patients diagnosed with LGG revealed that the median tumour volume increase of 174% was needed for the radiologists to determine a progressive disease. This number decreased to a 57% volume increase when tumour segmentation and quantitative measures were available. In **Paper IV**, I evaluated how much the visual perception of tumour growth differs from the reference value in GBM patients treated surgically and with chemoradiotherapy. My results are in line with the previous findings. Visual volume increase estimations by radiologists differ on average by 80 percentage points from the value derived from manual segmentation. Even though GBM lesions, especially post resection, have a much more heterogeneous appearance resulting in even more difficult visual estimation, tumour volume change assessment is not accurate based on standard evaluation methods.

The findings from **Theme IV** also indicate that just overlapping lesion labels on MR images increases radiologists' accuracy in visual assessment of the direction and extent of tumour volume increase (Figure 5). The volume increase estimations are, on average, closer to the reference values (65 percentage points difference) and present better agreement. Even though the reader's confidence in the tasks is rather moderate and generally does not change when the ET label is displayed, the radiologists tend to trust the accuracy of label-derived values more than their assessment. This finding can be a moderate indicator that the radiologists will trust the label-derived volumetric information in a clinical workflow. In this study, however, the labels were generated manually by a trained neuroradiologist. We decided to do so to potentially limit inducing negative bias against DL-generated labels, were they of insufficient quality. Still, some labels were assessed as having limited diagnostic quality. Furthermore, there was a poor agreement regarding segmentation quality among the radiologists.

These findings point to a potentially big challenge for sustained use and acceptance of DL tools. On the one hand, the acceptance and trust in the BTS tools will depend on the quality of the automatic segmentation[71]. On the other hand, as shown in **Paper IV**, quality evaluation is very subjective. While it is unlikely to effectively address the subjectivity of radiologists' perceptions, providing them with appropriate evidence for the accuracy of the tools and benefits of using the volumetric estimates could improve their acceptance of DL BTS tools and result in long-term use of the tools. As already discussed, the evidence for accuracy efficacy is lacking for the majority of methods. The lack of clinically implemented and evaluated tools also resulted in scarce evidence that would indicate the impact of label-derived tumour volume on diagnostic thinking efficacy. It is likely that only once DL BTS tools are implemented and routinely used in radiological workflow, will we be able to evaluate their impact on the therapeutic efficacy and higher levels of diagnostic efficacy of such tools.

In **Theme IV**, we identified substantial and procedural requirements of trustworthy AI tools and implementation procedures for radiological applications. Fulfilling these conditions will require involvement of all stakeholders. We found that trust in AI CDS tools is not based solely on the consistent accuracy (cf. 6.3, demand for *reliability* and demand for *quality control*) of the AI generated information; it requires the information is clinically relevant, and can be compared to previous findings. Furthermore, the way the information if generated and analysed must also be transparent. Here, the term transparency includes a broader spectrum of requirements compared to the technical explainability of the decisions made by the tool. The understanding of the "why" in a particular patient case and from a broad perspective is crucial for developing appropriate levels of trust[109,110]. Furthermore, explainability is built on radiologists' expertise enhanced with the CDS tools and evidence-based references. All those requirements must be backed up by inter organisational compatibility framework – appropriate infrastructure must be developed to govern data handling and provide control over data and monitor how the data and the CDS tools are used.

Our findings overlap with some aspects of the theoretically derived AI implementation framework proposed by Toreini et al. that supports trust in AI[110]. They identified four features of technologies that constitute trustworthy AI tools: fairness, explainability, auditability, and safety. Another implementation framework that was evaluated in a use case was proposed by Juluru et al. The authors identified the two elements also found in our investigation: quality control and results database as a necessary component of AI software. Our findings, however, take a broader look at the implementation problem that involves not only trustworthiness of the tools but the whole implementation framework. While particular solutions to fulfil these requirements may not be generalizable, our work provides general guidelines for successful implementation of AI tools in radiology that can serve as a base for developing implementation frameworks tailored to specific settings.

8 CONCLUSIONS

The abundance of studies that validate DL BTS algorithms has not yet lead to the abundance of DL tools in clinical practice. Commonly, these studies do not publish the proposed and validated algorithms for independent evaluation. Furthermore, the description preprocessing of images fed to the algorithms tends to be insufficient to facilitate reproducibility. Prevalent validation study design, therefore, does not facilitate clinical validation of proposed algorithms. Thorough validation of the algorithms on routine clinical images is necessary to build grounds for developing the algorithms into clinically beneficial tools. Availability of tumour labels and derived volumetric measurements will likely lead to a more accurate and confident tumour growth assessment. In my work, I found that radiologists trust that the AI-generated tumour volume information is more accurate than their visual assessment and find it clinically helpful when the quality of the labels is erceived as sufficient.

Quality assessment will be a crucial step in implementing DL CDS tools. However, the quality assessment will be subject to inter-rater variability and may also be influenced by a bias towards the technology. Strong evidence for their reproducibility and replicability will likely increase trust in the DL technology. It will provide knowledge of when the methods perform well and when not, strengthen the evidence for diagnostic thinking efficacy and provide a basis for evaluating the therapeutic efficacy of the availability of CDS in tumour status assessment.

How the implementation process is designed and conducted on the organisational level will play a crucial role in the implementation success. To develop appropriate levels of trust towards DL tools for CDS, several demands for the implementation framework and the tool itself will have to be fulfilled. These demands concern not only the evidence for the tool's accuracy and the users' attitudes but also the involvement of other relevant stakeholders.

9 FUTURE PERSPECTIVES

Through the findings of my work, a path leading to sustainable implementation of DL BTS tools in clinical practice has emerged. Although not as alluring as developing a state-of-the-art DL BTS algorithm, reproducibility and replicability studies should be more common and encouraged if we want the clinicians and the public to develop appropriate levels of trust in DL CDS tools. Furthermore, extensive studies that assess diagnostic thinking efficacy of DL BTS in a clinical scenario are needed.

While industry may be the crucial player in the final steps of the implementation path, there are enormous opportunities for research to contribute to safe, efficient, and sustainable implementation and use of DL BTS tools. Collaboration between research and clinical units could lead to development of pre-clinical testing “sand boxes” of DL algorithms. Such sand boxes should allow us to assess accuracy of an algorithm in a simulated clinical setting on a variety of cases. Furthermore, the sand boxes should allow us to develop a framework for integrating the information in existing workflow, accounting for quality assessment, and other requirements discovered in my doctoral project. Such comprehensive evaluation should prepare an algorithm for clinical validation and certification. Furthermore, it could enhance the chances of a commercial product to be successfully integrated in clinical workflows. Scientific studies can therefore diminish the risk that enormous resources devoted by the commercial sector will be wasted because an algorithm will not perform well enough or it will not be accepted by users.

Another line of research that will facilitate tumour volumetry as a routine measurement in neuroradiology is refining the response assessment criteria. New criteria should be developed using BTS algorithms, not merely approximating the correspondence of a 2D measurement to a volume. More accurate tumour volume change estimation could also lead to more refined treatment recommendations, improving patient care and quality of life in the face of a serious, often terminal disease.

ACKNOWLEDGEMENTS

I read somewhere that being awarded a PhD title is not so much about the destination but who you become along the way. It resonated with me a lot as this journey – next to the academic endeavour – coincided with immense personal growth. And who I have become along the way is thanks to the people I had the privilege and luck to have by my side.

"Which is more important," asked Big Panda, "the journey or the destination?"

"The company." said Tiny Dragon.

llustration from Big Panda and Tiny Dragon by James Norbury 2021 redistributed under fair use case. Text and illustration copyright © James Norbury.

First and foremost, I would like to thank my principal supervisor, Rolf Heckemann, for embarking on this project with me. Thank you for your support, reassurance, patience, and guidance. One of the most remarkable characteristics of a great mentor is when a lost and hopeless student leaves nearly every supervisory meeting with hope and belief they have what it takes to carry on. Thank you for being that person for me and for trusting that I can do it long before I believed it myself! It means a lot to know that you have been willing to wander with me into the unpopular area of interdisciplinary research, challenging the status quo, and always encouraging me to follow my path.

I would like to thank my co-supervisor Justin Schneiderman. You have always helped me improve the quality of my work, often by playing the role of the devil's advocate. It's invaluable to have this example which encourages critical thinking and sacrificing good ideas to come up with great ones. Thank you for all your encouragement and always being available when I needed support.

Special thanks also go to Isabella Björkman-Burtsher for joining as my co-supervisor later on in the project. Your genuine interest in my work, insights and constructive critique shared during many conversations (which I always looked forward to!) meant a lot to me and improved the quality of my work immensely! Thanks for sharing the insiders' experience of radiological work, which was fundamental to the success of the trajectory my PhD project took on.

But most of all, I wanted to thank you, Rolf, Justin, and Isabella, for being fellow human beings, not only academic supervisors, and understanding the messiness of life we all have to face every once in a while. I wouldn't be where I am, and who I am now without your support and encouragement.

I also want to express my gratitude to my half-time seminar committee: Kerstin Lagerstrand, Rodrigo Moreno, Lasse Riis Østergaard. The discussion we had during the seminar encouraged me to follow the emerging, challenging path of my project. Thank you to my coauthors: Mats, Nickoleta, Bertil, Magnus, Asgeir, and Tora for your support and involvement in my work; to my colleagues and ex-colleagues for sharing experiences and perspectives.

Next, I would like thank Katka and Isidora, my best friends, my chosen family, catalysts, witnesses, and incredible supporters of my growth during this academic journey and outside of it. I hope I'll always have you in my life as you make it (and me) 10000 times better.

Isi, I would not be able to do it without you. I will be forever grateful to have you in my life, to have the moments we shared in our apartment in Studiegången, to be able to learn from you how to be a better human and friend. Most of all, thank you for all your patience with me, unwavering presence, and support – you know it means a life.

Katka, not only have you been my best friend, partner in many crimes and adventures, a shoulder to cry and laugh on hysterically – but also a mentor in this academic endeavor. I've always looked up to you and have learned so much from you in my personal and academic journey. I'm immensely grateful for you and your never ending support, love, encouragement, jokes, and discussions. I consider myself very lucky to call you my best friend.

I would also like to thank the lovely people I have the privilege to call my close friends: Anton, Henni, Tina, Hannah and Tobi for all the fun times, trips, conversations, bonfires, swims, and shared moments in life that I hold dear in my heart. Thanks to my volleyball team and volleyball friends – you are the best!

And last but not least, thank you to my parents and my brother. Dziekuję Wam za wsparcie w dążeniu do tego i wielu innych celów, nawet jeśli te cele i droga, ktorą obieram, jest wam nieznana.

REFERENCES

1. Ghaffari, M., Sowmya, A. & Oliver, R. Automated Brain Tumor Segmentation Using Multimodal Brain Scans: A Survey Based on Models Submitted to the BraTS 2012–2018 Challenges. *IEEE Reviews in Biomedical Engineering* **13**, 156–168 (2020).
2. Do We Need Doctors Or Algorithms? *TechCrunch* https://social.techcrunch.com/2012/01/10/doctors-or-algorithms/.
3. European Society of Radiology (ESR). Impact of artificial intelligence on radiology: a EuroAIM survey among members of the European Society of Radiology. *Insights Imaging* **10**, 105 (2019).
4. Bruls, R. J. M. & Kwee, R. M. Workload for radiologists during on-call hours: dramatic increase in the past 15 years. *Insights Imaging* **11**, 121 (2020).
5. Giansanti, D. & Di Basilio, F. The Artificial Intelligence in Digital Radiology: Part 1: The Challenges, Acceptance and Consensus. *Healthcare (Basel)* **10**, 509 (2022).
6. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Medical Image Analysis* **42**, 60–88 (2017).
7. van Leeuwen, K. G., Schalekamp, S., Rutten, M. J. C. M., van Ginneken, B. & de Rooij, M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *Eur Radiol* **31**,

3797–3804 (2021).

8. *Handbook of Medical Imaging: Processing and Analysis Management*. (Academic Press, 2000).

9. Weltens, C. *et al.* Interobserver variations in gross tumor volume delineation of brain tumors on computed tomography and impact of magnetic resonance imaging. *Radiotherapy and Oncology* **60**, 49–59 (2001).

10. Porz, N. *et al.* Multi-Modal Glioblastoma Segmentation: Man versus Machine. *PLOS ONE* **9**, e96873 (2014).

11. Visser, M. *et al.* Inter-rater agreement in glioma segmentations on longitudinal MRI. *Neuroimage Clin* **22**, 101727 (2019).

12. Abramson, R. G., Su, P.-F. & Shyr, Y. Quantitative metrics in clinical radiology reporting: a snapshot perspective from a single mixed academic-community practice. *Magn Reson Imaging* **30**, 1357–1366 (2012).

13. Sharma, M., Juthani, R. G. & Vogelbaum, M. A. Updated response assessment criteria for high-grade glioma: beyond the MacDonald criteria. *Chinese Clinical Oncology* **6**, 4–4 (2017).

14. RECIST. https://recist.eortc.org/.

15. Macdonald, D. R., Cascino, T. L., Schold, S. C. & Cairncross, J. G. Response criteria for phase II studies of supratentorial malignant glioma.

J Clin Oncol **8**, 1277–1280 (1990).

16. van den Bent, M. *et al.* Response assessment in neuro-oncology (a report of the RANO group): assessment of outcome in trials of diffuse low-grade gliomas. *The Lancet Oncology* **12**, 583–593 (2011).
17. Radiologisk diagnostik - RCC Kunskapsbanken. https://kunskapsbanken.cancercentrum.se/diagnoser/hjarna/vardprogram/bilaga-1.-radiologisk-diagnostik/.
18. Diagnostik - RCC Kunskapsbanken. https://kunskapsbanken.cancercentrum.se/diagnoser/hjarna/vardprogram/diagnostik/.
19. Zhang, J. Y. *et al.* Quantitative Improvement in Brain Tumor MRI Through Structured Reporting (BT-RADS). *Acad Radiol* **27**, 780–784 (2020).
20. Chukwueke, U. N. & Wen, P. Y. Use of the Response Assessment in Neuro-Oncology (RANO) criteria in clinical trials and clinical practice. *CNS Oncol* **8**, CNS28 (2019).
21. Creative Commons — Attribution-NonCommercial-NoDerivatives 4.0 International — CC BY-NC-ND 4.0. https://creativecommons.org/licenses/by-nc-nd/4.0/.
22. Auer, T. *et al.* Quantitative volumetric assessment of baseline enhancing tumor volume as an imaging biomarker predicts overall survival in patients with glioblastoma. *Acta radiologica* (2020)

doi:10.1177/0284185120953796.

23. Huber, T. *et al.* Progressive disease in glioblastoma: Benefits and limitations of semi-automated volumetry. *PLoS One* **12**, e0173112 (2017).
24. Fouke, S. J. *et al.* The role of imaging in the management of adults with diffuse low grade glioma. *J Neurooncol* **125**, 457–479 (2015).
25. Fathallah-Shaykh, H. M. *et al.* Diagnosing growth in low-grade gliomas with and without longitudinal volume measurements: A retrospective observational study. *PLOS Medicine* **16**, e1002810 (2019).
26. Berntsen, E. M. *et al.* Volumetric segmentation of glioblastoma progression compared to bidimensional products and clinical radiological reports. *Acta Neurochir* **162**, 379–387 (2020).
27. Gui, C., Lau, J., Kosteniuk, S., Lee, D. H. & Megyesi, J. Radiology reporting of low-grade glioma growth underestimates tumor expansion. *Acta Neurochirurgica* (2019) doi:10.1007/s00701-018-03783-3.
28. Jakola, A. S., Moen, K. G., Solheim, O. & Kvistad, K.-A. “No growth” on serial MRI scans of a low grade glioma? *Acta Neurochir* **155**, 2243–2244 (2013).
29. Schmitt, P., Mandonnet, E., Perdreau, A. & Angelini, E. D. Effects of slice thickness and head rotation when measuring glioma sizes on MRI: in support of volume segmentation versus two largest diameters methods.

J Neurooncol **112**, 165–172 (2013).

30. Reuter, M. *et al.* Impact of MRI head placement on glioma response assessment. *J Neurooncol* **118**, 123–129 (2014).
31. Renard, F., Guedria, S., Palma, N. D. & Vuillerme, N. Variability and reproducibility in deep learning for medical image segmentation. *Sci Rep* **10**, (2020).
32. Gryska, E., Schneiderman, J., Björkman-Burtscher, I. & Heckemann, R. A. Automatic brain lesion segmentation on standard magnetic resonance images: a scoping review. *BMJ Open* **11**, e042660 (2021).
33. Menze, B. H. *et al.* The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans Med Imaging* **34**, 1993–2024 (2015).
34. MICCAI BRATS - The Multimodal Brain Tumor Segmentation Challenge. http://braintumorsegmentation.org/.
35. Fryback, D. G. & Thornbury, J. R. The efficacy of diagnostic imaging. *Med Decis Making* **11**, 88–94 (1991).
36. Fukushima, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybernetics* **36**, 193–202 (1980).
37. Hubel, D. H. & Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology* **160**, 106–154 (1962).
38. LeCun, Y. *et al.* Handwritten Digit Recognition with a Back-Propagation

Network. in *Advances in Neural Information Processing Systems* vol. 2 (Morgan-Kaufmann, 1989).

39. Lindsay, G. W. Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. *Journal of Cognitive Neuroscience* **33**, 2017–2031 (2021).
40. Raschka, S. & Mirjalili, V. *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow , 2nd Edition*. (Packt Publishing, 2017).
41. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986).
42. Anaconda | State of Data Science 2020. *Anaconda* https://www.anaconda.com/state-of-data-science-2020.
43. Egger, J. *et al.* GBM Volumetry using the 3D Slicer Medical Image Computing Platform. *Sci Rep* **3**, 1364 (2013).
44. Deeley, M. A. *et al.* Segmentation editing improves efficiency while reducing inter-expert variation and maintaining accuracy for normal brain tissues in the presence of space-occupying lesions. *Phys. Med. Biol.* **58**, 4071–4097 (2013).
45. Vaassen, F. *et al.* Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. *Physics and Imaging in Radiation Oncology* **13**, 1–6 (2020).

46. Gryska, E. *et al.* Deep learning for automatic brain tumour segmentation on MRI: evaluation of recommended reporting criteria via a reproduction and replication study. *BMJ Open* **12**, e059000 (2022).

47. London, A. J. Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Cent Rep* **49**, 15–21 (2019).

48. Singh, A., Sengupta, S. & Lakshminarayanan, V. Explainable Deep Learning Models in Medical Image Analysis. *Journal of Imaging* **6**, 52 (2020).

49. Wadden, J. J. Defining the undefinable: the black box problem in healthcare artificial intelligence. *Journal of Medical Ethics* (2021) doi:10.1136/medethics-2021-107529.

50. Quinn, T. P., Jacobs, S., Senadeera, M., Le, V. & Coghlan, S. The Three Ghosts of Medical AI: Can the Black-Box Present Deliver? *arXiv:2012.06000 [cs]* (2020).

51. The Evidence-Based Radiology Working Group, T. Evidence-based Radiology: A New Approach to the Practice of Radiology. *Radiology* **220**, 566–575 (2001).

52. Lavelle, L. P., Dunne, R. M., Carroll, A. G. & Malone, D. E. Evidence-based Practice of Radiology. *RadioGraphics* **35**, 1802–1813 (2015).

53. Jorritsma, W., Cnossen, F. & van Ooijen, P. M. A. Improving the radiologist-CAD interaction: designing for appropriate trust. *Clin Radiol*

70, 115–122 (2015).

54. Ebrahimian, S. *et al.* FDA-regulated AI Algorithms: Trends, Strengths, and Gaps of Validation Studies. *Academic Radiology* **29**, 559–566 (2022).
55. VYSIONEER. https://www.vysioneer.com/.
56. 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).
57. Wang, J.-Y. *et al.* RADI-12. Deep learning for automatic detection and contouring of metastatic brain tumors in stereotactic radiosurgery: a retrospective analysis with an FDA-cleared software algorithm. *Neuro-Oncology Advances* **3**, iii20 (2021).
58. Graylight Imaging – Medical Imaging Software. *Graylight* https://graylight-imaging.com/.
59. BioMind. https://biomind.ai/.
60. Nalepa, J. *et al.* Fully-automated deep learning-powered system for DCE-MRI analysis of brain tumors. *Artificial Intelligence in Medicine* **102**, 101769 (2020).
61. National Academies of Sciences, E. *Reproducibility and Replicability in Science*. (2019). doi:10.17226/25303.
62. Pineau, J. *et al.* Improving Reproducibility in Machine Learning

Research (A Report from the NeurIPS 2019 Reproducibility Program). *arXiv:2003.12206 [cs, stat]* (2020).

63. Pineau, J. Machine Learning Reproducibility Checklist. https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf.
64. MICCAI. http://www.miccai.org/.
65. Kamnitsas, K. *et al.* Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis* **36**, 61–78 (2017).
66. 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).
67. Asan, O., Bayrak, A. E. & Choudhury, A. Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. *J Med Internet Res* **22**, (2020).
68. Retson, T. A. *et al.* Reader Perceptions and Impact of AI on CT Assessment of Air Trapping. *Radiol Artif Intell* **4**, e210160 (2022).
69. Li, M. D. *et al.* Radiology Implementation Considerations for Artificial Intelligence (AI) Applied to COVID-19, From the AJR Special Series on AI Applications. *American Journal of Roentgenology* **219**, 15–23 (2022).
70. Wong, J. *et al.* Implementation of deep learning-based auto-segmentation for radiotherapy planning structures: a workflow study at two cancer centers. *Radiation Oncology* **16**, 101 (2021).

71. van Garderen, K. A. *et al.* EASE: Clinical Implementation of Automated Tumor Segmentation and Volume Quantification for Adult Low-Grade Glioma. *Front Med (Lausanne)* **8**, 738425 (2021).

72. Börjesson, S. *et al.* A software tool for increased efficiency in observer performance studies in radiology. *Radiation Protection Dosimetry* **114**, 45–52 (2005).

73. Kickingereder, P. *et al.* Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study. *The Lancet Oncology* **20**, 728–740 (2019).

74. Chlebus, G. *et al.* Reducing inter-observer variability and interaction time of MR liver volumetry by combining automatic CNN-based liver segmentation and manual corrections. *PLOS ONE* **14**, e0217228 (2019).

75. Bø, H. K. *et al.* Intra-rater variability in low-grade glioma segmentation. *J Neurooncol* **131**, 393–402 (2017).

76. Joskowicz, L., Cohen, D., Caplan, N. & Sosna, J. Inter-observer variability of manual contour delineation of structures in CT. *Eur Radiol* **29**, 1391–1399 (2019).

77. Lee, J. D. & See, K. A. Trust in Automation: Designing for Appropriate Reliance. *Hum Factors* **46**, 50–80 (2004).

78. Weiser, S. Requirements of Trustworthy AI. *FUTURIUM - European*

Commission https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1 (2019).

79. Gille, F., Jobin, A. & Ienca, M. What we talk about when we talk about trust: Theory of trust for AI in healthcare. *Intelligence-Based Medicine* **1–2**, 100001 (2020).
80. Filice, R. W. & Ratwani, R. M. The Case for User-Centered Artificial Intelligence in Radiology. *Radiol Artif Intell* **2**, (2020).
81. Kushniruk, A. & Nøhr, C. Participatory Design, User Involvement and Health IT Evaluation. *Stud Health Technol Inform* **222**, 139–151 (2016).
82. Hasani, N. *et al.* Trustworthy Artificial Intelligence in Medical Imaging. *PET Clin* **17**, 1–12 (2022).
83. Ali Işın, Cem Direkoğlu, & Melike Şah. Review of MRI-based Brain Tumor Image Segmentation Using Deep Learning Methods. *Procedia Computer Science* **102**, 317–324 (2016).
84. Goswami, A. & Dixit, M. An Analysis of Image Segmentation Methods for Brain Tumour Detection on MRI Images. in *2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT)* 318–322 (2020). doi:10.1109/CSNT48778.2020.9115791.
85. Fawzi, A., Achuthan, A. & Belaton, B. Brain Image Segmentation in Recent Years: A Narrative Review. *Brain Sci* **11**, 1055 (2021).

86. García-Lorenzo, D., Francis, S., Narayanan, S., Arnold, D. L. & Collins, D. L. Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. *Med Image Anal* **17**, 1–18 (2013).

87. Kumari, N. & Saxena, S. Review of Brain Tumor Segmentation and Classification. in *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)* 1–6 (2018). doi:10.1109/ICCTCT.2018.8551004.

88. Mortazavi, D., Kouzani, A. Z. & Soltanian-Zadeh, H. Segmentation of multiple sclerosis lesions in MR images: a review. *Neuroradiology* **54**, 299–320 (2012).

89. Arksey, H. & O'Malley, L. Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology* **8**, 19–32 (2005).

90. Levac, D., Colquhoun, H. & O'Brien, K. K. Scoping studies: advancing the methodology. *Implementation Science* **5**, 69 (2010).

91. Colquhoun, H. L. *et al.* Scoping reviews: time for clarity in definition, methods, and reporting. *J Clin Epidemiol* **67**, 1291–1294 (2014).

92. Tricco, A. C. *et al.* A scoping review on the conduct and reporting of scoping reviews. *BMC Medical Research Methodology* **16**, 15 (2016).

93. Sargeant, J. M. & O'Connor, A. M. Scoping Reviews, Systematic

Reviews, and Meta-Analysis: Applications in Veterinary Medicine. *Frontiers in Veterinary Science* **7**, (2020).

94. Clark, K. *et al.* The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *J Digit Imaging* **26**, 1045–1057 (2013).
95. Freymann, J. B., Kirby, J. S., Perry, J. H., Clunie, D. A. & Jaffe, C. C. Image Data Sharing for Biomedical Research—Meeting HIPAA Requirements for De-identification. *J Digit Imaging* **25**, 14–24 (2012).
96. Recht, M. P. *et al.* Integrating artificial intelligence into the clinical practice of radiology: challenges and recommendations. *Eur Radiol* **30**, 3576–3584 (2020).
97. Prevedello, L. M. *et al.* Challenges Related to Artificial Intelligence Research in Medical Imaging and the Importance of Image Analysis Competitions. *Radiology: Artificial Intelligence* **1**, e180031 (2019).
98. Yang, L. *et al.* Stakeholders' perspectives on the future of artificial intelligence in radiology: a scoping review. *Eur Radiol* **32**, 1477–1495 (2022).
99. Scott, I. A., Carter, S. M. & Coiera, E. Exploring stakeholder attitudes towards AI in clinical practice. *BMJ Health Care Inform* **28**, e100450 (2021).
100. Baid, U. *et al.* *The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on*

Brain Tumor Segmentation and Radiogenomic Classification. http://arxiv.org/abs/2107.02314 (2021) doi:10.48550/arXiv.2107.02314.

101.Kofler, F. *et al.* BraTS Toolkit: Translating BraTS Brain Tumor Segmentation Algorithms Into Clinical and Scientific Practice. *Front. Neurosci.* **14**, (2020).

102.Davatzikos, C. *et al.* Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. *J Med Imaging (Bellingham)* **5**, 011018 (2018).

103.Pati, S. *et al.* The Cancer Imaging Phenomics Toolkit (CaPTk): Technical Overview. *Brainlesion* **11993**, 380–394 (2020).

104.Rathore, S. *et al.* Brain Cancer Imaging Phenomics Toolkit (brain-CaPTk): An Interactive Platform for Quantitative Analysis of Glioblastoma. *Brainlesion* **10670**, 133–145 (2018).

105.Hu, S.-Y. *et al.* Multimodal Volume-Aware Detection and Segmentation for Brain Metastases Radiosurgery. in *Artificial Intelligence in Radiation Therapy* (eds. Nguyen, D., Xing, L. & Jiang, S.) 61–69 (Springer International Publishing, 2019). doi:10.1007/978-3-030-32486-5_8.

106.Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods* **18**, 203–211 (2021).

107.Jakola, A. S. & Reinertsen, I. Radiological evaluation of low-grade glioma: time to embrace quantitative data? *Acta Neurochir* **161**, 577–578 (2019).

108. Jakola, A. S. *et al.* Surgical resection versus watchful waiting in low-grade gliomas. *Annals of Oncology* **28**, 1942–1948 (2017).

109. Ferrario, A. & Loi, M. How Explainability Contributes to Trust in AI. in *2022 ACM Conference on Fairness, Accountability, and Transparency* 1457–1466 (Association for Computing Machinery, 2022). doi:10.1145/3531146.3533202.

110. Toreini, E. *et al.* The relationship between trust in AI and trustworthy machine learning technologies. in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* 272–283 (Association for Computing Machinery, 2020). doi:10.1145/3351095.3372834.