university of groningen / faculty of science and engineering

Designing a User Interface for Semi-Automatic Tumor Segmentation

Using Certainty Visualization to Promote Explainability

Liv Ziegfeld
S3369390

MSc in Computational Cognitive Science
MASTER'S THESIS (45 ECTS)

Internal Supervisor:
Dr. Fokie Cnossen (Artificial Intelligence, University of Groningen)

External Supervisors:
Dr. Peter van Ooijen (Faculty of Medical Sciences, University Medical Center Groningen)
Alessia de Biase (Department of Radiotherapy, University Medical Center Groningen; PhD)

August 31st, 2022

Abstract

Automatic tumor segmentation using deep learning models is a promising avenue for decreasing the inter-observer variability and time needed for manual tumor segmentation. Since tumor segmentation is still carried out manually at the Universitair Medisch Centrum Groningen (UMCG), the first aim of this project was to design a user interface for the computer-aided segmentation of head and neck tumors. To promote appropriate trust and optimal usage of the tool, it was also investigated how to incorporate visualizations of the model's confidence in its predictions. This should increase the explainability of the model's output, which has been lacking in similar tools that give binary outputs. Lastly, the project focused on exploring whether a semi-automatic segmentation tool is desired and is clinically feasible at the UMCG.

From interviews and user tests with clinicians it was determined that a first-reader tool for tumor segmentation would be most useful. A user interface was thus designed that supports radiation oncologists in creating, reviewing and editing automatic tumor contour predictions. The predictions were visualized in the form of an interactive certainty map containing colored, semi-transparent contours. This allows clinicians to compare the contours at different certainty thresholds and to select the one that best matches their own perception of the tumor's boundaries. The interface also contains an all-in-one map of all available prediction thresholds for a given slice that can be used as a general overview.

Results from the user testing indicated that radiation oncologists were optimistic about introducing a tool for computer-aided segmentation of tumors to the UMCG. Participants reported that the user interface is pleasant and easy to use and that the certainty maps generally aided their understanding of the model's predictions. Thus, the present research offers a promising outlook for visualizing the model's prediction certainty to make the outputs more explainable. Further, valuable suggestions for the improvement of the interface were gathered as well as suggestions for the implementation of this tool.

Acknowledgements

First of all, I would like to thank my internal supervisor Dr. Fokie Cnossen, who has not only taught me a lot, but also helped me find my place in the world of Cognitive Engineering. Thank you to Dr. Peter van Ooijen, my UMCG supervisor, for providing guidance during my project and for connecting me with the right people. Thank you to all the clinicians and researchers from the UMCG who have provided invaluable input for this project. I would also like to thank Alessia de Biase, my daily supervisor. The amount of time you spent teaching me and discussing our ideas is not taken for granted. Thanks for making this time a lot more enjoyable and for being available whenever I needed your help. Lastly, thank you so much to my family and friends who have supported me throughout this whole process.

Table of Contents

1. Introduction

1.1 Background Information

After heart disease, cancer is the most common cause of death (Cancer, 2021). According to the WHO, malignant tumors resulted in approximately 10 million deaths worldwide in 2020 (Cancer, 2021). Head and neck (H&N) cancer, which will be the focus of this paper, is an especially complex cancer type due to this area's intricate anatomy (Schutte et al., 2020). More than 830,000 cases of head and neck cancer are diagnosed each year worldwide (as of 2019), with a mortality rate above 50% each year (Cramer et al., 2019; Schutte et al., 2020). According to Schantz and Yu (2002), the prevalence of H&N cancer is on the rise, especially in younger populations.

Surgery and radiation therapy are the two most common treatments for H&N cancer (Marur & Forastiere, 2008). Whilst in the past treatment success was mainly measured by the extent to which the tumor could be removed and by general survival rates post-treatment, the requirements for successful treatment are becoming more nuanced (Semple et al., 2008). Nowadays, an additional important indicator of treatment success is the patient's quality-of-life post-treatment (Semple et al., 2008). Many of the advances in radiotherapy over the last two decades also led to an improvement in this post-treatment well-being of a patient (Morgan & Sher, 2020).

One of the main ways radiotherapy may reduce a patient's quality of life is by causing toxicity. Trotti (1997, p. 570) defines toxicity as "any temporary or permanent change in normal tissues and/or related symptoms from cancer treatment". Toxicities resulting from radiotherapy treatment usually appear in the first 90 days after treatment beginning (Cox et al., 1995; Trotti, 2000). These occur since ionizing radiation kills tumor cells and controls their growth, but also damages healthy cells and tissue surrounding the tumor (Baskar et al., 2014). In other words, toxicity can result in failure of the organs surrounding the tumor, also known as organs-at-risk (OARs). In the head and neck region, this may for instance lead to problems with swallowing or with speech (Semple et al., 2008). Around 15%-40% of H&N cancer survivors suffer from pain, which oftentime remains until well after treatment completion (Cramer et al., 2018). Apart from physical difficulties and pain following H&N cancer treatment, the patient's psychological well-being may also be compromised as a result of toxicity build-up. This may for instance occur due to feelings of shame caused by a change in speech (Semple et al., 2008).

Reducing the amount of radiation the organs-at-risk receive can strongly limit the build up of toxicity in the tumor-free regions. Hence, it is of utmost importance to attempt irradiating only the cancerous areas, while leaving surrounding healthy organs and tissue as unaffected as possible. However, if the area the radiotherapy is focused on is too small and some part of the tumor will not be targeted by radiotherapy, the tumor may grow or metastasize in the future (Foster et al., 2014). Thus, administering the majority of the radiation dose to the exact tumor is important. To do so, precisely defining the tumor's boundaries is required before starting the treatment, which is also known as tumor segmentation. The terms 'tumor segmentation',

'delineation', and 'contouring' will be used interchangeably throughout this paper. In this process, a radiation oncologist usually outlines the tumor on a medical imaging scan using a pen tool in a delineation software. This is frequently done on computed tomography (CT) scans, while also consulting positron emission tomography (PET) and magnetic resonance imaging (MRI) scans. Apart from helping to determine the precise area to irradiate, delineation is necessary to stage the tumor, create treatment plans, and to evaluate the effectiveness of the treatment while monitoring the progression of the cancer (Wong, 2005, as cited in Gordillo et al., 2013).

To date, much of the tumor segmentation is carried out manually by radiation oncologists. However, manual tumor segmentation is a labor-intensive process with high inter- and intra-observer variability (Andrearczyk et al., 2022; Foster et al., 2014). Such variability may be caused by differences in human perception, varying understanding of delineation guidelines, individual bias or differences in the experience of radiation oncologists (Sorantin et al., 2021; Njeh, 2008; Sadeghi et al., 2021). Furthermore, the process is error-prone (Andrearczyk et al., 2022; Foster et al., 2014). Tumor delineation can be ambiguous due to irregular and fuzzy boundaries of the tumors, as well as low image resolution and high noise in PET images (Foster et al., 2014). Such difficulties may be magnified for tumors in the head and neck region, due to the region's complexity and small structures (Van Dijk et al., 2020). Thus, delineations are often subjective, limiting the reproducibility of the segmentation results and increasing the possibility for suboptimal irradiation (Foster et al., 2014). Tumor segmentation is also very time consuming (Andrearczyk et al., 2022; Foster et al., 2014). Radiation oncologists specialized in H&N cancer can take up to several hours delineating a single patient's tumor. After delineating a tumor, radiation oncologists often consult one another to discuss the accuracy of their delineations due to the high inter-observer variability, further adding to the time needed for delineation. This increases the difficulty for radiation oncologists to keep up with their workload and to ensure rapid treatment for each patient.

Recent scientific advancements have shown that automatic tumor segmentation may increase the reproducibility, accuracy, and speed of the segmentation process (Andrearczyk et al., 2022). Deep learning (DL) algorithms, which are artificial intelligence models that are inspired by human neural networks, have shown enormous potential in predicting tumor boundaries on imaging scans (Andrearczyk et al., 2022). Such models can be used as decision-support tools to guide and assist the clinician in their task.

Despite the promising nature of DL models for tumor segmentation, their adoption in the clinic is still sparse (Gulum et al., 2021). This is mainly due to the fact that most current deep learning tools lack transparency and understanding by their users due to their black-box nature (Sorentin et al., 2021). While research efforts are increasing to make DL tools more explainable in general, the progress with making AI tools more transparent for medical imaging specifically have been limited (Natekar et al., 2020). This is also reflected by the nature of current deep learning segmentation tools: The user generally only sees the output of the model, but does not receive any information on how the model came up with these predictions.

One way to increase the explainability of DL models is to communicate their certainty to the user. This involves indicating the model's confidence in its output so the user gets more insight into why the model made a certain prediction. As suggested by Natekar et al. (2020), clear visualizations of the working of a deep learning model allow users to make better judgements on the veracity of the predictions. For instance, if the model has a low certainty of being accurate in a specific instance, this could prompt the user to review this region or data point more closely. Most current automatic segmentation methods lack this indication of the predictions' uncertainty (Wang et al., 2019). This may promote false assumptions about the results' reliability, causing areas that need expert review to be overlooked (Shi et al., 2011; Wang et al., 2019). Over-reliance on the automatic tool and errors may occur as a result. Alternatively, the user's trust in the system can be compromised if the model shows an output the user believes to be wrong and it does not communicate that it only has low confidence in this prediction. This could eventually lead to users stop using the tool.

Research is currently being conducted at the Universitair Medisch Centrum Groningen (UMCG) on this topic. Researchers at the UMCG developed a deep learning model that generates automatic predictions of head and neck tumor contours on CT and PET images and it is now being investigated how to make this tool more explainable and trustworthy. Instead of only displaying a single predicted contour, the model outputs probability maps indicating the degree of certainty of each pixel to be classified as tumor or not. This model is not used in the clinic yet and it does not currently have a user interface to present its results to the clinicians. Further, no research has been carried out in the UMCG on how to display the model's uncertainty in the user interface to date.

1.2 Research Goal

The aim of the current research is to design an interactive user interface for a decision support tool for automated tumor segmentation. It should display the outputs of the tumor segmentation model along with a visualization of the model's certainty. This interface should be part of a tool that aids radiation oncologists in their tumor segmentation of oropharyngeal cancer patients receiving radiotherapy as their primary treatment. It will be investigated how to best communicate the model's certainty in the form of a certainty map. Since the map should give users more insight into the decision process of the model, it will be examined whether it aids clinicians in their understanding of the predictions and whether more appropriate confidence in the system's output could be achieved. Further, the satisfaction with the design of the user interface and the clinical feasibility of such a system at the UMCG will be evaluated during user tests. This research is being conducted to establish how an automatic tumor segmentation tool may eventually be introduced as a decision support tool to the UMCG in the most useful manner. Identifying what radiation oncologists require from such a tool and examining whether certainty visualizations are helpful and desired is crucial before implementing a similar system in practice.

Thus, the current paper addresses the following research questions:

1. What are the user requirements for an automatic tumor segmentation interface?
2. How is the model's certainty best visualized using certainty/probability maps?
3. What is the perceived clinical feasibility and utility of such a system at the UMCG?

1.3 Paper's Structure

The topic's theoretical framework will first be outlined by providing background information on head and neck cancer and how tumors are segmented at the UMCG. Next, an overview of current automatic segmentation tools and of explainable AI in medical imaging will be given. Lastly, human factors considerations and cognitive engineering principles will be discussed. Following this, the design of the prototype of the user interface will be outlined. Findings from interviews and the shadowing of UMCG radiation oncologists and researchers will be presented that resulted in user requirements used for the interface design. Next, results of the user testing will be discussed along with recommendations for the improvement of the original prototype design. In the last part, suggestions will be made for future research and for clinical implementation.

PART 1: THEORETICAL FRAMEWORK

Since the automatic segmentation tool is focused on head and neck cancer, the following section will provide background information on this type of cancer. Following this, the manual tumor delineation process will be outlined in a general manner to provide an understanding of the tasks involved.

2. Head and Neck Cancer

The majority of H&N cancers, approximately 90%, are rapidly growing squamous cell carcinomas (HNSCC) (Marur & Forastiere, 2008; Schutte et al., 2020). HNSCCs get their name from the cells this cancer originates in, namely squamous cells, which form the mucus layers of the head and neck (Head and Neck Cancers, n.d.). H&N cancer may also originate in the salivary glands, sinuses, muscles, or nerves, but these cancer types have a lower prevalence rate than HNSCCs (Chow, 2020; Head and Neck Cancers, n.d.).

H&N cancer can occur in different regions within the head and neck, as depicted in Figure 1 below (Waqar et al., 2019). These regions include the oral cavity, the throat (also known as the pharynx, containing the sub-parts of the nasopharynx, the oropharynx, and the hypopharynx), the voice box (larynx), the paranasal sinuses and the nasal cavity, or in the salivary glands (Waqar et al., 2019; Head and Neck Cancers, n.d.). The highest risk factors for developing H&N cancer include smoking and drinking alcohol, but dietary deficiencies including a vitamin A and iron deficiency have also been associated with H&N cancer (Marur & Forastiere, 2008). Furthermore, around one fourth of HNSCCs are linked with the human papillomavirus (HPV) (Marur & Forastiere, 2008).

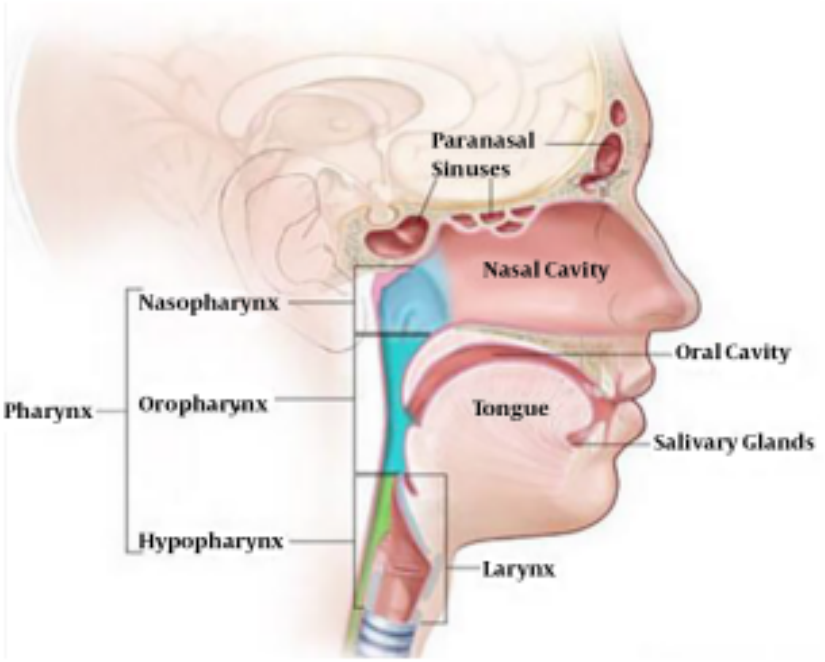


Figure 1. Head and Neck Cancer Regions
Note. Figure from Waqar et al. (2019). Copyright 2019 by Waqar et al.

The symptomatology of H&N cancer depends on the stage of the cancer and on its primary location. Some of the most common symptoms of H&N cancer include sores and ulcers that don't heal, a sore throat, difficulty and pain while swallowing, and ear pain (Marur & Forastiere, 2008). Neck masses or lumps in the mouth region can also be observed in some patients (Marur & Forastiere, 2008).

The following section will describe how H&N cancer is usually detected at the Universitair Medisch Centrum Groningen (UMCG), the hospital this research is based on.

3. Treatment of Head and Neck Cancer at the UMCG

To design an effective user interface that can be implemented as seamlessly as possible, it is crucial to understand the context the interface should be implemented in. Hence, the following sections briefly describe the workflow of radiation oncologists at UMCG. This is also depicted in Figure 2. Most of the information on this workflow was gathered from interviews with UMCG radiation oncologists. Where required, this was supplemented with findings from literature. A more detailed task analysis can be found in section 7 "Requirements Analysis".

3.1 H&N Cancer Detection

Tumors can be detected in different manners in the UMCG. Patients may present to the clinic with some of the aforementioned symptoms or complaints (see section 2), or patients might be referred to the UMCG for treatment by another hospital which identified a tumor. Lastly, some

patients are asymptomatic but coincidentally find out they have a tumor due to another health check-up. This may for instance involve a dental appointment or another condition for which an MRI (magnetic resonance imaging), CT or PET scan is made. Once a tumor has been detected, it will be diagnosed and the extent of the tumor will be determined to appropriately stage the tumor (Marur & Forastiere, 2008). Further, the patient will be checked for metastases (Marur & Forastiere, 2008).

The radiation oncologists at UMCG treat patients that have never received any treatment for H&N cancer before, but also patients who have previously been operated on to remove the tumor. Such patients may then need radiotherapy due to re-growths or since parts of the tumor were missed or could not be removed in the operation.

3.2 Manual Delineation Overview

Once the patient presents to the hospital, radiation oncologists at the UMCG meet the patient to discuss their symptoms, do physical examinations, and explain the process of radiotherapy. Following this, imaging scans will be made. These can include CT, PET and MRI scans and are also known as simulation scans. Before generating the treatment plan, the organs at risk and the target volumes will be segmented from the imaging scans by various clinicians, as described in the following sections.

3.2.1 Organ at risk segmentation.

Before the radiation oncologist starts segmenting the tumor, a radiation technologist delineates all organs at risk (OARs). Generally, about 25 OARs are segmented per patient in the H&N region. These will be used for planning of the treatment, since the amount of radiation that will hit the OARs is a decisive factor for which therapy to administer.

Organ at risk segmentation is already done partially automatically at the UMCG. Automatic OAR segmentation is a less complex task than automatic tumor segmentation for a variety of reasons. First, organs usually do not grow or spread, unlike tumors. Second, the location of OARs is always the same, while tumors can be located in different areas of the head and neck. Moreover, OARs are delineated so they can be avoided during radiotherapy, while tumors are delineated for the opposing reason: they are the target of radiotherapy. Thus, tumors have to be delineated in an extremely precise manner to ensure that the entire tumor volume is targeted, while simultaneously irradiating as little surrounding tissue as possible. The precision of OAR segmentation does not have to be exactly as high. This lower complexity and slightly lower required precision of OAR segmentation explains why it was successfully introduced to UMCG earlier than automatic tumor segmentation.

3.2.2 Primary tumor segmentation.

After the OARs are segmented, the delineation of the gross tumor volume (GTV) can begin. This is the extension of the tumor itself that is visible on the imaging scans (Leer, 2005). During GTV segmentation, a radiation oncologist examines medical imaging scans to identify the tumor(s)

and outlines them using a delineation software. Scans from different imaging modalities may be used for this. For H&N tumors, (planning) CT and PET scans are used most commonly in the segmentation procedure. These modalities are frequently used due to the complementary information they offer. CT scans provide anatomical information, while PET scans give insights into the functional and metabolic processes of volumes of interest (VOIs), which are the structures to be delineated and targeted by treatment (Foster et al., 2014). In some cases radiation oncologists may use overlapped CT/PET images, meaning that structures of the H&N region can be seen on the same image as the metabolic activity in that area. However, CT and PET scans are taken at different time points and possibly with a different positioning of the patient, which in some cases can create a mismatch between the two scans in terms of overlap. In such cases the clinician may choose to view them separately in their software. In addition, MRI images may be used which are made in 3D. Consulting these three imaging modalities can be useful to check the extension of the tumor and whether the delineations have been made correctly. Nevertheless, the availability of scans from different modalities may vary. For instance, when patients are referred from other hospitals to the UMCG, the radiation oncologists at UMCG often still use the scans from the referring hospital, where MRI scans are sometimes not made.

The frequency of taking imaging scans during the radiotherapy course may also vary from patient to patient. Around two weeks prior to treatment beginning, the PET, CT and MRI scans are made for all patients. The scans will be made in the same position the treatment will be administered and a fixation mask is used. With specific therapy types or with certain patients, imaging scans may be taken again throughout the treatment. If a patient changes a lot during the course of the therapy, for instance by losing a lot of weight, if the tumor shrinks very rapidly, or when having a lot of edema (swelling), cone-beam imaging is often used throughout the treatment. This ensures the radiation is still appropriate and allows making adjustments in the treatment plan if necessary.

When delineating the tumors, the radiation oncologists always take patient information into account. Sometimes they review the patient data first and then proceed to the scans, while on other occasions the scans are first analyzed after which the patient chart is consulted. Patient data in the chart may for instance include age, gender, prior diseases and injuries, teeth removals, and a description of the tumor's location.

3.2.3 Segmentation Review Meetings.

Due to the high interobserver variability and the complexity of delineating tumors in the head and neck region, clinicians at UMCG usually consult each other to discuss their work. After the radiation oncologist who treats the patient in question has segmented their tumor, a meeting will be held in which the segmentation will be discussed. A head and neck radiologist usually first shares findings from an MRI examination done in the software SECTRA, if available. Here any areas that might contain tumors are discussed and it will be checked whether the lymph nodes look pathological. Next, the team switches to the software RayStation and the radiation oncologist takes over to share the delineation they have made prior to the meeting. Other

clinicians will then state any disagreements or areas that need closer review. All tumor segmentations will also be examined by H&N nuclear medicine physicians. After the meeting, the radiation oncologist goes back to make adjustments to their segmentation if necessary, which is relatively common. Following this, the treatment plan will be generated after which the actual treatment may commence.

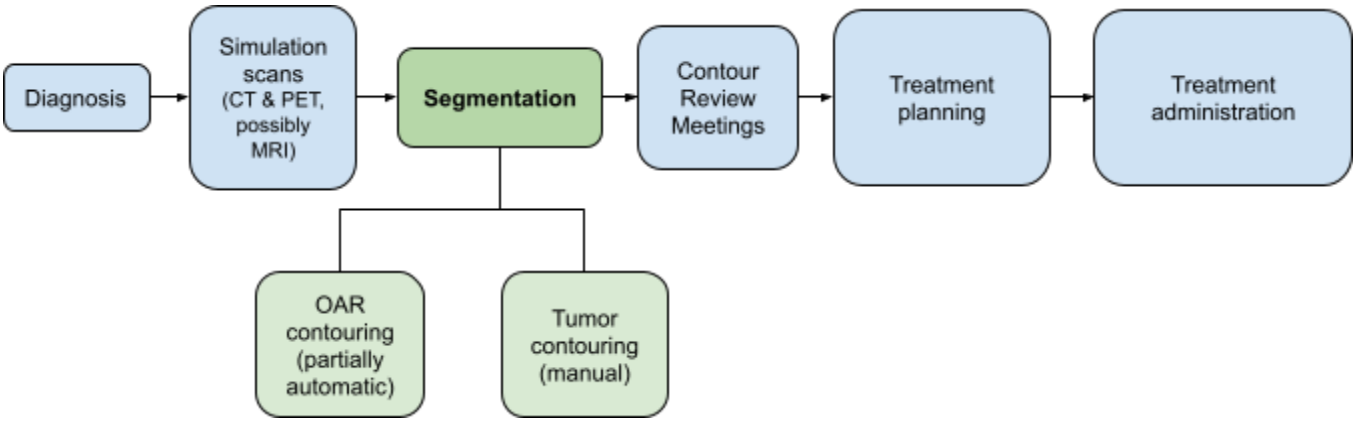


Figure 2. Radiotherapy Workflow at UMCG

Now that a general overview of the manual segmentation process of radiation oncologists has been provided, the following section will describe why automation has been introduced into this workflow, what the promise of this is and what hurdles still exist in successfully implementing auto-contouring tools on large scales.

4. Automatic Tumor Segmentation

The number of AI applications in medical imaging has shown a sharp increase over the last years (Suzuki, 2017). One of the main reasons for this rapid advancement is that AI tools can contribute to better clinical outcomes and make the workload of radiologists more manageable by increasing the speed of work (Alexander et al., 2020; Andrearczyk et al., 2022). Alexander et al. (2020) found that 90% of interviewed US radiologists perceived an increase in their workload between 2016 and 2019, a trend that reflects the increased demands for efficiency in radiology in the Netherlands as well (Strohm et al., 2020). Furthermore, pressures to reduce the cost of healthcare in the Netherlands point to AI tools as a promising avenue (Strohm et al., 2020).

4.1 Computer-aided Diagnosis (CAD)

Machine learning (ML) is a subset of AI which is commonly used in automated systems in the medical domain, as depicted in Figure 3 (Suzuki, 2017; Abdellah & Koucheryavy, 2020). ML models learn from data by automatically identifying patterns in datasets which can be used to

make predictions about new cases (Lee et al., 2017). ML is now being employed for computer-aided diagnosis (CAD). CAD is the general term to describe the process of a physician detecting, diagnosing or classifying abnormalities from medical imaging scans using AI systems (Jorritsma et al., 2015; Doi et al., 1999). Such systems work in a semi-automatic manner (not fully automatically), since they only assist the professional in their task by promoting diagnostic accuracy, reliability and efficiency while still requiring human verification (Doi et al., 1999; Sorentin et al., 2021).

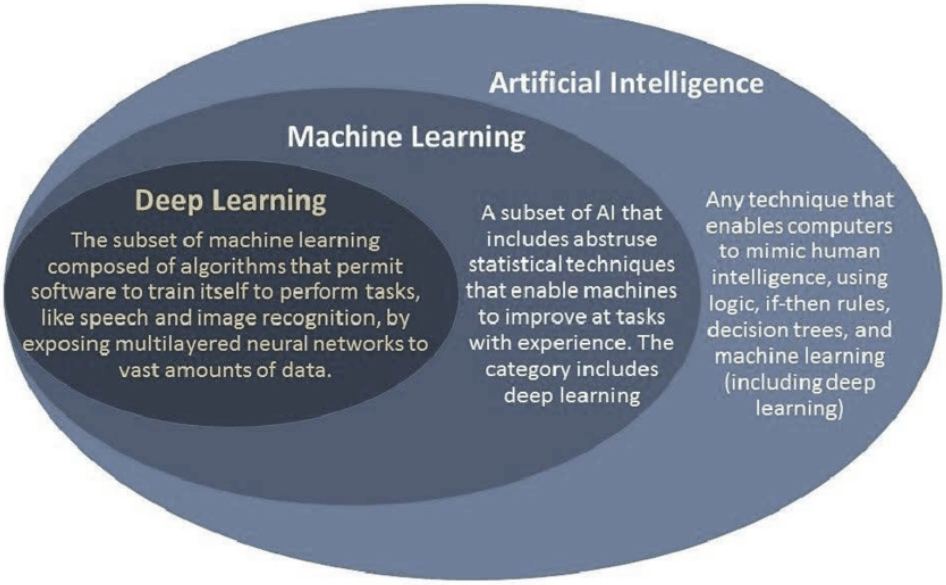


Figure 3. Relationship between AI, ML and DL
Note. Figure from Abdellah and Koucheryavy (2020).

CAD systems can be employed in three different manners, as depicted in Figure 4. Frequently, CAD is used as a second reader whereby the expert first evaluates medical images in their routine way, after which the system prompts the physician to reevaluate their work in case of inaccuracies or missed targets (Jorritsma et al., 2015; Fujita, 2020). This increases the time needed for image analysis (Fujita, 2020). When using CAD as a time-saving concurrent-reader, the physician has access to the model's prediction from the onset, and hence will use the prediction while simultaneously forming their own judgment (Fujita, 2020). Lastly, first-reader CAD systems fully evaluate the medical images initially and the physician only interprets and corrects the system's predictions afterwards (Fujita, 2020). This CAD type saves the most time (Fujita, 2020).

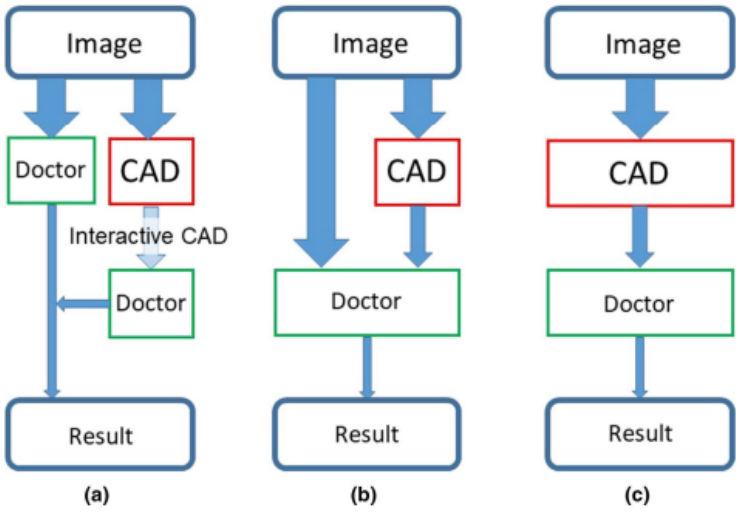


Figure 4. Different Uses of CAD Systems; A) Second-reader, B) Concurrent reader, C) First-reader

Note. Figure from Fujita (2020).

Besides for detecting tumors, CAD can also be used for object classification, which is the focus of this paper (Suzuki, 2017). Classification refers to making a decision on which class a certain object belongs to, for instance if it is cancerous or not (Suzuki, 2017). This involves analyzing so-called input features from medical images (Suzuki, 2017). Input features may vary, but include characteristics such as darkness, size, contrast, shape and texture (Suzuki, 2017; Sorentin et al., 2021). In ML, such input features must be determined by the human operator, making this a laborious task (Holzinger et al., 2017).

However, in the current state of the art technique for medical image segmentation, human intervention is not required to select such features (Suzuki, 2017; van der Velden et al., 2022). This involves a subclass of ML, known as deep learning (DL). As illustrated in Figure 5, instead of using predetermined features as model input, deep learning models take large datasets of medical images as input to analyze the images' pixel values directly to make classification decisions (Suzuki, 2017). Since human feature identification is not required, DL models are also known as 'end-to-end ML' models since they can run autonomously once fed with the training data (Suzuki, 2017). This minimal human involvement is a major advantage of DL models (Van Dijk et al., 2020). Figure 6 depicts the steps of hand-crafted feature selection in ML models that are taken over by the model in DL (Suzuki, 2017).

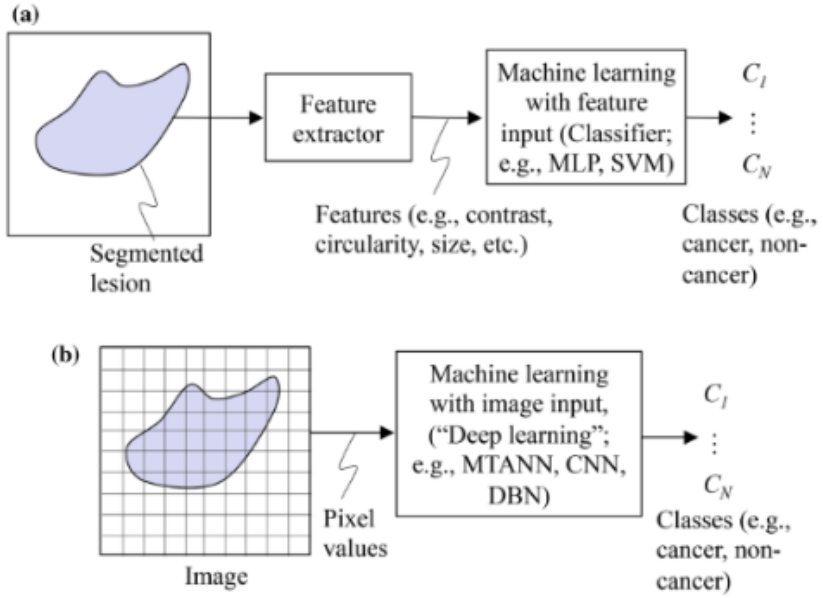


Figure 5. Machine Learning (A) Trains on Predetermined Image Features, Whereas Deep Learning (B) Directly Makes Predictions Based on Automatically Extracted Features
Note. Figure from Suzuki (2017).

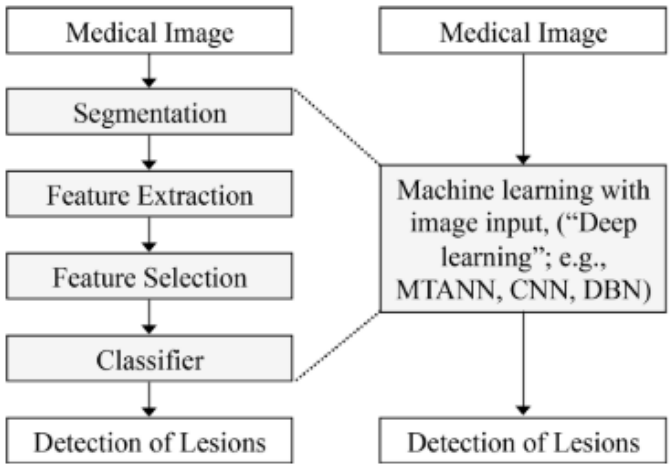


Figure 6. Eliminated Steps of Human Intervention in DL Models (Right Hand Side) Compared to Hand-Crafted Feature Extraction in ML Models (Left Hand Side)
Note. Figure from Suzuki (2017).

The current leading standard DL algorithm for automatic tumor segmentation are convolutional neural networks (CNNs) (Jiang et al., 2021; Muhammad et al., 2021). These will be briefly explained in the following.

4.1.1 Convolutional Neural Networks (CNNs) for Automatic Tumor Segmentation.

CNNs get their name from their properties that mimic a human neural network. They are structured in a hierarchical manner where data flows through different layers consisting of 'neurons' (Holzinger et al., 2017). The structural similarity between human and artificial neurons is displayed in Figure 7 (Lee et al., 2017). When used for automatic tumor segmentation, such CNNs determine whether each pixel or voxel in a medical image (a voxel can be thought of as a 3D pixel) is part of a tumor or not (Vandewinckele et al., 2020).

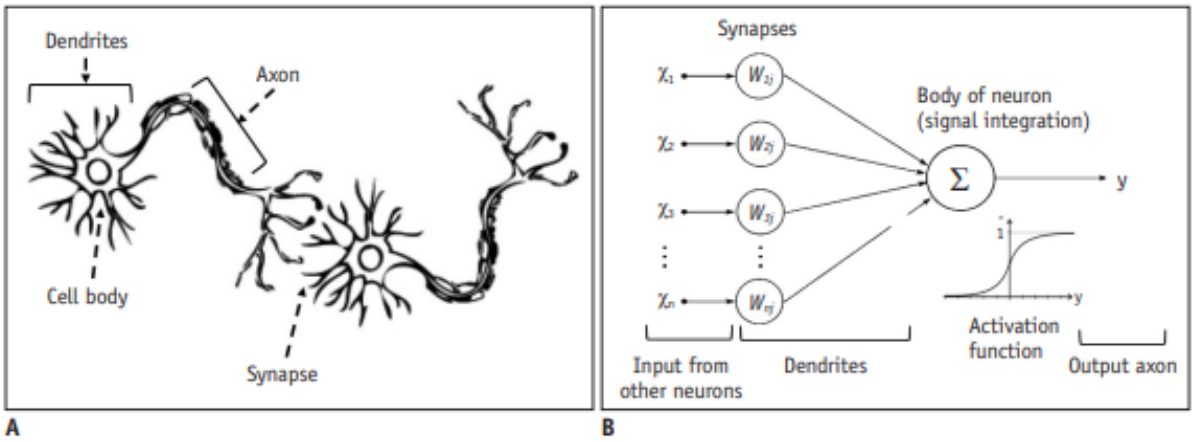


Figure 7. Structural Similarity of Human (left) and Artificial Neurons (right)
Note. Figure from Lee et al. (2017). Copyrighted 2017 by Lee et al.

Before using a CNN for automatic tumor segmentation, a training set of medical images must first be generated. This involves collecting images (frequently CT or PET) containing the tumor type of interest (Begoli et al., 2019). Next, labeling of the data will occur in which a clinician manually segments the tumor, which will be seen as the 'ground truth'. Sometimes pre-segmented scans are used to construct a training set, while on other occasions the scans will be marked specifically for the purpose of developing a DL model. The required size of the training set differs for each model since it is dependent on the variability within the dataset and the quality of the labels. Nevertheless, datasets of top-performing CNN auto-segmentation models contain more than 100 patients (Vandewinckele et al., 2020; Men et al., 2017; van Dijk et al., 2020). The model will use the ground truth scans as training input by feeding the pixels of the image to separate neurons in the first layer of the neural network.

As depicted in Figure 8, the input layer relays the information across different convolutional and pooling layers with varying activations and biases. This creates an iterative training where so-called hidden layers receive activations from the previous neuron layer and compute metrics such as the weighted sum of these inputs (Holzinger et al., 2017). Learning

parameters are continually updated and optimized until reaching the performance metric in question, which frequently is segmentation accuracy (Begoli et al., 2019). Ultimately, the hidden layers carry valuable information about the task features, which are finally passed to the output layer (Holzinger et al., 2017). Here model takes all information into account to make a decision on which class (tumor or not) the input belongs to.

Following the training of the model, the performance of the algorithm will be evaluated on a test set. This is a dataset that has not been seen in training and that has been preserved to examine how the model generalizes to new instances (Begoli et al., 2019).

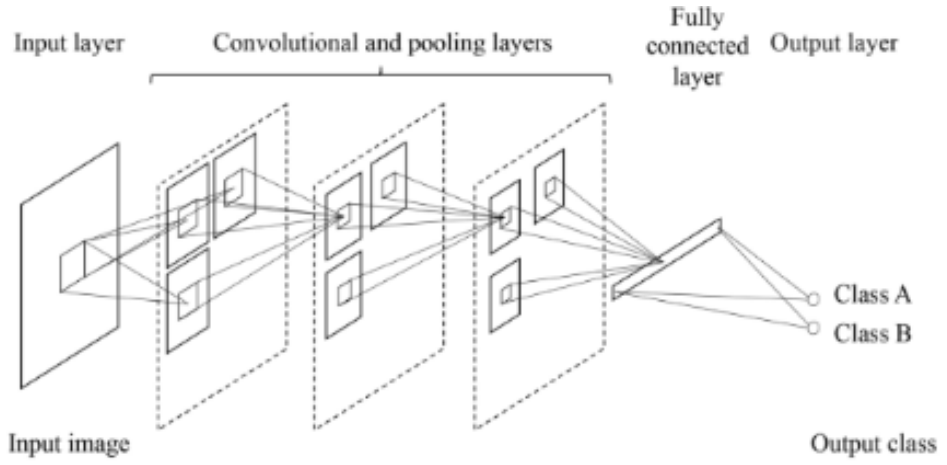


Figure 8. Structure of a Convolutional Neural Network (CNN)
Note. Figure from Suzuki (2017).

4.1.2 Performance of Current Automatic Segmentation Models.

The standard for current auto-contouring CNNs is to output a classification of whether each pixel/voxel contains tumor or not (Oreiller et al., 2022). Hence, such models predict a tumor's boundary in a binary way.

Autocontouring CNNs have achieved promising accuracies, with their performance being on par with manual segmentation by medical professionals or even higher than that of a single radiologist (Savenije et al., 2020; Vandewinckele et al., 2020; Jorritsma et al., 2015). Badrigilan et al. (2021) conducted a meta-analysis of 17 H&N DL segmentation studies based on MRI in which they found that the pooled DICE score was 0.8965. The DICE similarity coefficient measures segmentation accuracy by calculating the overlap between an autocontour and the ground truth, whereby a score of 1 represents perfect overlap. A dice score of 0.7785 was also achieved for GTV auto-segmentation on PET/CT images (Xie & Peng, 2022; Andrearczyk et al., 2022). Hence, these results suggest that auto-contouring models can already achieve very high accuracy.

Reviews on the clinical implementation of autocontouring CNNs for H&N cancer are lacking in the current literature, which is reflective of the low level of clinical adoption (Vandewinckele et al., 2020; Tang et al., 2018). Despite the outstanding results in terms of accuracy, reliability and speed of autocontouring CNNs, many clinics are hesitant to integrate such systems into their workload (Gerlings et al., 2021; Yang et al., 2022). Strohm et al. (2020) suggest a multitude of reasons for this, including lacking funding for AI adoption in the clinic, an unstructured implementation strategy and inadequate trust in AI predictions. It has been suggested that lacking trust is largely caused by the fact that CNNs are often not interpretable by their users (Gerlings et al., 2021; Yang et al., 2022). As a consequence, many intricate AI models are never actually employed in practice and get stuck in the development process (Tjoa & Guan, 2021; Gerlings et al., 2021). The causes for the lack of interpretability will now be discussed, along with their effect on users and the consequences for clinical implementation.

4.2 Problems with DL Models

4.2.1 Black-box Nature.

The widespread lack of understanding of DL models combined with the significance of clinical decisions has led to hesitation for large-scale employment of such algorithms in the clinic (Gulum et al., 2021; Begoli et al., 2019). A big part of this is due to the low intervention of humans in DL models, causing the user to be less aware of the model's workings. In DL, the user cannot examine the features extracted in the training procedure which will be decisive for the model's output (Wong et al., 2020). This lacking association between the model's output and the features extracted makes such models less explainable (Reyes et al., 2020). As Gulum et al. (2011) point out, it is usually models with high accuracies that are the least explainable due to their high complexity. This creates a tradeoff between accuracy and transparency, as depicted in Figure 9 (Yang et al., 2022). Such highly complex deep learning models with low interpretability are often referred to as 'black-box models.'

These undesirable effects of black-box models on the user are rooted in human factors principles. When designing new tools, especially ones involving automation, it is of great importance to consider these cognitive and usability aspects. Hence, some of the main human factors concepts relevant to the current project will be outlined below. After this, some implications for the design of the new interface will be drawn from these findings.

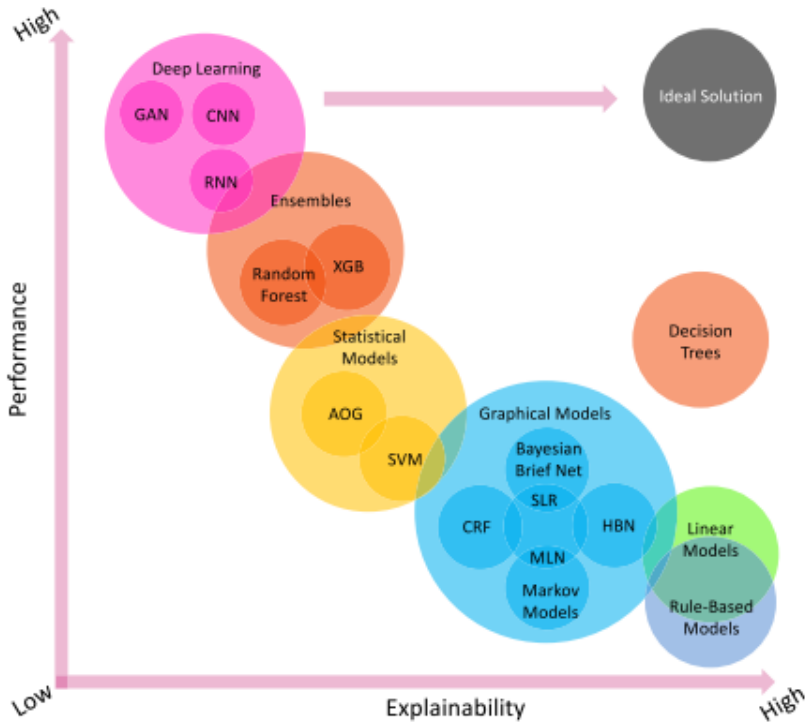


Figure 9. Performance-explainability Tradeoff; Deep Learning Models (e.g. CNNs) have high Accuracies but low Interpretability
Note. Figure from Yang et al. (2022)

5. Human Factors in Automatic Tumor Segmentation

Technological aspects have long received the most research attention in the development of automated tools (Parasuraman & Riley, 1997). Psychosocial phenomena have however been neglected by CAD researchers in the past (Strohm et al., 2020; Nishikawa & Bae, 2018). Considering human factors is especially important since the workflow of the user may fundamentally change when introducing automation (Lee & Seppelt, 2012). New tasks may get added to the user's workflow and old tasks may require modifications (Lee & Seppelt, 2012). When these changes are not communicated to the user or are misunderstood, severe errors may occur (Lee & Seppelt, 2012). The following section describes how failing to make human factors considerations when introducing automated segmentation tools into the workflow can lead to inappropriate system usage and negative clinical outcomes.

5.1 Trust

Jorritsma et al. (2015) point out that the current radiologist-CAD interaction is suboptimal. In big part this is attributable to the inappropriate trust radiologists have in CAD systems. Trust is a complex concept that is affected by one's experience, personality, and ability (von Eschenbach, 2021). Improper trust can lead to an inappropriate reliance on automation, and is therefore a crucial factor when examining the cooperation and delegation of tasks from humans to machines (von Eschenbach, 2021; Jorritsma et al., 2015). When trust is inappropriate, two main problems can occur, as described below.

5.1.1 Under-reliance

Lacking trust can lead to an under-reliance on AI tools, which could present itself as clinicians not taking the predictions seriously (Jorritsma et al., 2015). Jorritsma et al. (2015) suggest that true-positive predictions may be overlooked as a result, which may be detrimental to patients' well-being. Further, complete disuse of the tool may also occur (Jorritsma et al., 2015). When radiologists lack trust in an automated decision-support system, they frequently revert back to their manual segmentation methods. This may mean that optimal accuracies are not achieved which could negatively impact patient treatment, the procedure will be more time-intensive and spending on AI tools may be unnecessary if they are not used in practice.

5.1.2 Over-reliance

On the other hand, automation-induced complacency can occur when a user excessively relies on the system, causing misuse (Sujan et al., 2019; Parasuraman & Riley, 1997; Jorritsma et al., 2015). This can occur if the user blindly trusts the system and is not aware that the predictions are not 100% accurate. Over-reliance can lead to two types of errors. Errors of commission occur in the case of false-positives, where an incorrect prediction is accepted by the clinician, thus decreasing the accuracy (Mosier et al., 1998; Sujan et al., 2019). Opposingly, errors of omission describe false-negatives that occur when the CAD system does not identify a present abnormality and the user hence overlooks these (Jorritsma et al., 2015).

5.2 Out-of-the-loop Syndrome

Apart from inappropriate trust and therefore ineffective usafe, one other main risk with automation is that it creates a distance between the user and the task (Lee & Seppelt, 2012). When the entire task can be completed without intervention by the human operator, their monitoring abilities and their situation awareness strongly decrease as well (Lee & Seppelt, 2012; Endsley, 2012). Even when instructing a human to oversee the workings of the automatic system, it can be more challenging to regain control in emergency situations or to identify anomalies (Lee & Seppelt, 2012). This describes the out-of-the-loop syndrome (Endsley, 2012). In the case of automatic tumor segmentation, this could mean that the user does not identify when errors are made by the system. The clinician may also be less involved with the treatment

administration when completely relying on fully automatic systems for the segmentation, which could lead to unpleasant effects on the relationship with the patient.

5.3 User-Centered Design

To minimize the possibility of inappropriate trust in and usage of technologies, a user-centered design process is vital. Human factors should be considered at every stage of development to achieve optimal usability. Usability describes "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" (ISO 9241-11, 1998; as cited in Dias et al., 2017, p.20).

The interaction design life cycle by Rogers et al. (2013) in Figure 10 describes the steps of a user-centered design process. This involves first establishing the requirements of the future users for the to-be-designed product or service, after which first designs are created. Next interactive prototypes are made, which are then evaluated. This process is iterative, meaning that adjustments to the design are continuously made as new requirements are identified. This can occur through prototype evaluations or through new information from interviews or literature search. Only when the product has been extensively tested and user satisfaction has been assured, is the design life cycle complete. Involving the user in this manner increases the possibility of identifying poor design choices early on and ensures that the tools are appropriate for their implementation context.

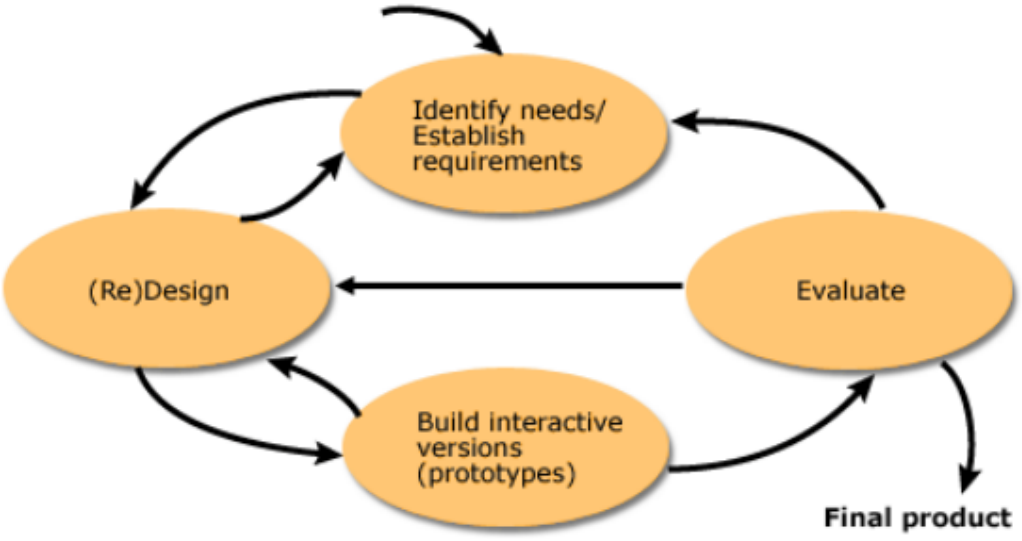


Figure 10. Interaction Design Life Cycle by Rogers et al. (2013)

As such human factors principles are gaining more research attention, there are more efforts at incorporating them while designing new technologies. The current movement of eXplainable AI aims to develop more transparent AI tools in which optimizing the understanding

of predictions and designing for more appropriate trust are priorities. This will be described in more detail in the following section.

6. Explainable AI

The potential of DL models and the lack of their interpretability has sparked researchers all across the globe to work towards so-called eXplainable AI (xAI). Miller (2019) defines explainability as the extent of understanding of the cause of a decision. This means that users want more insight into why an AI model reached a particular decision (Molnar, 2020; Natekar et al., 2020). The terms explainability, transparency, and interpretability will be used interchangeably throughout this paper to describe this concept.

Identifying every step the model takes to reach a conclusion is not necessary to increase explainability. Instead, Holzinger et al. (2017) suggest that a 'functional understanding' is needed, instead of detailed algorithmic knowledge of how the model works. Moreover, XAI should make the user aware of what the model is capable of and where its limitations are, as well as providing the user with an understanding of its workings in the future (Yang et al., 2022).

Ultimately, increasing the explainability of AI models can promote their acceptance and increase the trust of medical staff (Natekar et al., 2020; Holzinger et al., 2017). Further, an understanding of the model's output may decrease the possibility of misinformed decisions which could reduce errors (Gulum et al., 2021). Apart from increasing the trust of clinicians, the demand for xAI in the medical field is increasing due to legal reasons, where clinicians may be asked to justify their reason for reaching a certain decision (Gerlings et al., 2021; Gulum et al., 2021).

While there is a movement towards more transparent AI tools in all fields, the necessity in the medical field is especially urgent. In the medical field, decisions made using AI systems have high stakes which have profound effects on a patient's well-being and treatment outcomes, sometimes even being decisive for a patient's life or death (Patrício et al., 2022). Yet, little research attention has been devoted to date to making deep learning models for medical imaging more interpretable (Natekar et al., 2020; Holzinger et al., 2017). Hence, the following sections will describe methods to promote XAI in radiology, with a specific focus on automatic segmentation tools.

6.1 Increasing the Explainability of Tumor Segmentation Models

According to Cai et al. (2019), medical professionals demand local and global explanations for AI models. Local explanations refer to case-by-case information on why a particular decision was reached by the model, whereas global reasoning pertains to general information on the model (Cai et al., 2019).

The explainability of models for medical image analysis can be increased through visual information, texts, or via examples (van der Velden et al., 2022). To date, visual information is most frequently used in this domain (van der Velden et al., 2022). Saliency mapping is the leading form of visual explanation in which the areas on a medical scan that were most decisive

for reaching the model's decision are highlighted (van der Velden et al., 2022). Textual explanations include image captions generated by AI models or reports outlining the model's results (van der Velden et al., 2022). Lastly, example-based explanations make references to prior clinical decisions that are similar to the task at hand in order to promote generalizing from examples (van der Velden et al., 2022).

To increase the adoption of AI systems in the clinic, it has further been suggested that uncertainty quantification is necessary (Begoli et al., 2019; Lim et al., 2019). Gillman et al. (2021, p. 669) define uncertainty as the "quantification of the doubt about the measurement result." The potential for uncertainty visualization in an automatic tumor segmentation interface is two-fold. Firstly, insight into the model's prediction confidence may increase the user's trust in the system. This allows identifying whether the model's predictions are reliable or not (Lim et al., 2019). Not communicating uncertainty measures clearly can cause misinterpretation by clinicians (Gillmann et al., 2021). Having an uncertainty quantification may therefore evoke a more nuanced information uptake by clinicians in which they take all available information into account to make a critical judgment of whether the model is correct or not. The second way in which insight into the model's ambiguity can be useful is that it could point out regions with especially low certainties (Jungo et al., 2018). These could then be examined more closely by medical professionals and manually corrected. The following section provides more detail on displaying uncertainties.

6.2 Uncertainty Quantification

Uncertainty in tumor segmentation models originates from a variety of sources, for instance from image acquisition, processing and reconstruction (Gillmann et al., 2021). Incorrect setup of the patient or movement during the CT/MRI/PET scan can also cause artifacts (Njeh, 2008; Gillmann et al., 2021). Since DL models learn from the segmentations made by human radiation oncologists, the prediction quality also depends on the accuracy of the ground truth data (Begoli et al., 2019). Further, it could be that the training set is not representative of all possible tumor types, creating inaccuracies in the prediction of boundaries for tumors that have not been trained on (Begoli et al., 2019). Hence, training sets should be as large as possible so that the model can learn all types of delineations, increasing the accuracy of its predictions (Sorantin et al., 2021).

6.2.1 Certainty Maps to Display Model's Confidence

With the trend towards visual explanations in medical image analysis and the advantages of depicting a model's confidence in its predictions, a powerful way of combining these may be through a certainty map. Certainty maps are visual representations of the varying confidence levels the model has in its output across an imaging scan. In the case of tumor segmentation, the map would therefore represent the certainty in its classification decision for each pixel of whether it contains tumor or not.

Certainty maps are advantageous for a variety of reasons. Firstly, since tumor segmentation is a very visual process, it is beneficial to use this same modality to promote

insights into the model's reasoning. Using a different modality, such as having clinicians read a text might create a feeling of detachment towards the task at hand. Visual maps can be integrated with imaging scans, hence allowing a seamless integration into the tasks of radiologists (Reyes et al., 2020). Further, certainty maps give case-specific information, as recommended by Cai et al. (2019), since the model produces a different certainty map for each patient. This can be useful since it is akin to how clinicians discuss cases with other colleagues during segmentation review meetings (Cai et al., 2019; Gerlings et al., 2021). Like in review meetings, areas of ambiguity are highlighted in certainty maps that might need further review. Hence, certainty maps could simulate the discussion with a colleague, making this a technique that could fit well into the workflow of radiation oncologists. Lim et al. (2019) confirmed this by suggesting that a system is expected to be trustworthy if the method of communicating uncertainty is similar to how humans would share their uncertainty with others. For these reasons, the current research is based on a DL model containing an certainty map as its output, as specified in the following section.

Given the information on CAD tumor segmentation systems, black-box models, XAI and on evoking more appropriate trust in autocontours discussed thus far, the design of an automated tumor segmentation model for the UMCG will now be outlined.

PART 2: DESIGN

The aim of the current project was to design a user interface for an automatic tumor segmentation tool for radiation oncologists at the UMCG. Furthermore, it was investigated how to best incorporate probability maps reflecting the model's certainty into the interface and how optimal interaction with them could be ensured. The design was based on the previously discussed literature on developing tools to increase the usability and explainability of AI models, in order to evoke more adequate trust in the predictions.

To structure the design process and to take the future user into account, the interaction design lifecycle by Rogers et al. (2013) was followed, as was described previously (Figure 10). The following section describes how the first step of the interaction lifecycle was tackled, namely the identification of user requirements.

7. Requirements Analysis

In order to understand what is expected of an automatic tumor segmentation interface at the UMCG, the current tasks of radiation oncologists first had to be analyzed in more detail.

7.1 Shadowing and Semi-structured Interviews

Participants

To understand their workflow, two radiation oncologists were shadowed while performing a live segmentation of tumors. These radiation oncologists were already involved in the project of Alessia de Biase, who is the developer of the automatic segmentation model this interface is

based on. One of them is specialized in head and neck tumors (shadowed online via Microsoft Teams), while the other focuses on lung tumors (shadowed in person at the UMCG). Despite the focus of this research being on head and neck tumors, a lung specialist was observed for availability reasons and since there is a lot of overlap with the general steps required for delineating tumors in different regions. Two radiation oncologists were shadowed to minimize the possibility of missing important information when watching only one of them.

Apart from the radiation oncologists, several other UMCG staff members from the radiology department were interviewed. Two PhD candidates were interviewed, one with experience in automatic organ-at-risk segmentation, and one with experience with AI tools for medical imaging. Further, three UMCG researchers who were present at monthly meetings with the head and neck tumor segmentation team with experience in deep learning and tumor delineation provided valuable insights into the requirements for this interface.

Method

The radiation oncologists segmented tumors while explaining their procedure throughout the process. The goal of this requirements analysis was to gain more insight into tumor segmentation, including an overview of tools the radiation oncologists need in a segmentation software, any issues they currently face, and wishes and needs for the to-be-designed interface. Semi-structured interviews were also carried out in these sessions to gain more specific information on their tasks. Some of the main questions that were asked during these meetings can be found in Appendix A.

Results

The findings from all interviewees will now be presented collectively.

Equipment and environment. The radiation oncologists at the UMCG currently use the software RayStation 10B (RaySearch Laboratories) on a Windows PC to manually segment tumors (as of April 2022). They have two PC monitors, one frequently used to review patient details and other information relevant to their segmentations, while the second monitor is used to actually delineate the tumors. A screenshot of the interface of RayStation can be seen in Figure 11 below. Radiation oncologists at the UMCG segment tumors in their offices. No major nuisances were observed in this work environment, although the observees could be subject to distractions such as phone calls or other people coming into their office.

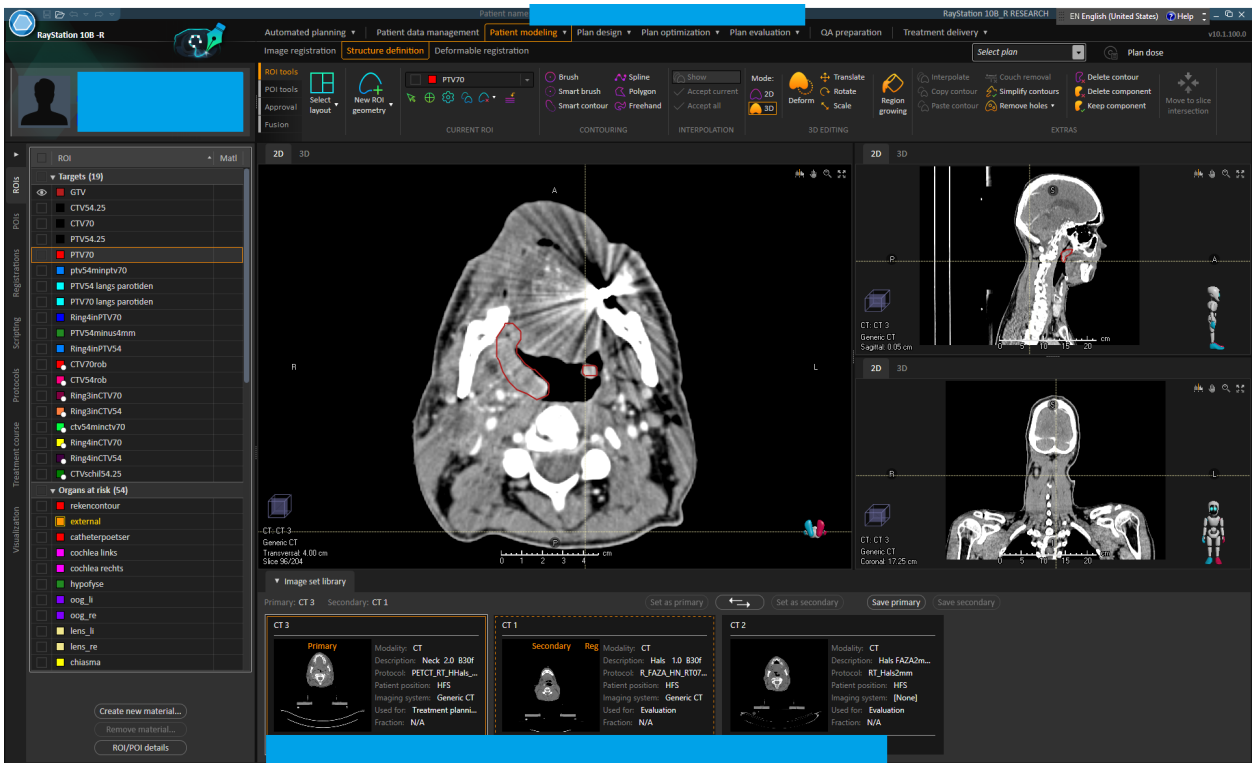


Figure 11. Screenshot of the Interface of the Software RayStation (RaySearch Laboratories), used by UMCG Radiation Oncologists for Manual Tumor Delineation

Hierarchical Task Analysis of Manual Tumor Segmentation. A hierarchical task analysis (HTA) was constructed from the semi-structured interviews with the radiation oncologists and from shadowing them during manual tumor segmentation. This was done to explore the steps that are currently necessary to segment a tumor manually. Identifying the functionalities that are currently being used and the steps necessary for manual tumor segmentation can provide insight into the requirements for an automatic tumor segmentation interface. Further, any problems or inefficiencies with the current software could inspire improvements for the interface prototype that will be designed.

The hierarchical task analysis can be seen in Figure 12. It includes a more general overview of the steps taken in the workflow of manual tumor segmentation. Details regarding the exact clicks made in the manual tumor segmentation software were omitted, since the aim of this project was not to redesign the current software of radiation oncologists at UMCG. Instead, the goal of the HTA was to provide a structured overview of the subtasks involved and which order they are carried out in. The steps of the HTA will briefly be described hereunder.

First, radiation oncologists review the patient information. This includes the patient's demographics and information about the location and extension of the tumor. Other relevant information may include prior treatment histories such as operations. Next, the radiation

oncologist opens RayStation and loads the patient's imaging scans (CT, PET, possibly MRI). The axial (transverse) view is usually preferred for primary tumor segmentation in the H&N region. They will then scroll through the slices of the imaging scans to get a rough idea of where the tumor is located. Sometimes this is also done as the first step, while patient information is consulted in the second step. This depends on the personal preferences of the radiation oncologist. Next the tumor segmentation is started. For this, clinicians frequently scroll to a CT slice in which the tumor is well-visible and will start using the brush tool in RayStation to trace the outline of the tumor they have identified. The segmentation is then often turned on and off, so its accuracy can be examined without having the segmentation occlude a part of the tumor. If the segmentation needs to be altered, the brush tool can again be used to enlarge the delineation or it can be made smaller using the erase tool. Once the radiation oncologist is happy with their segmentation on their first slice, they will move on to the next slice. Every slice can be individually segmented in this manner, but RayStation's interpolation tool is frequently used to save time. This involves delineating one slice, skipping a few slices, and then delineating a second slice. This will give 'end' delineations for a certain region between which the remaining delineations can be interpolated, or created automatically based on these 'end' delineations. The radiation oncologist will then review the interpolated slices and edit them if necessary. While the delineations are being made on the CT, the PET scan of the same slice is frequently consulted since the uptake can provide valuable information on the tumor's extension. Once all slices have been delineated, a final review is made and the MRI may be consulted to check the segmentation accuracy. Some clinicians also choose to use the MRI at the beginning of their segmentation or throughout their segmentation procedure. Once the radiation oncologist is satisfied with their delineation they can save the delineation by pressing the button labeled 'Create new ROI', entering a name (usually naming conventions are used, such as GTVp for gross primary tumor volume), and selecting a conventional color for the delineation type (for the GTVp the UMCG usually uses yellow). After the GTVp has been created, the other important target volumes are created by expanding the GTV by certain margins, which is usually done with a preset in RayStation. These include the clinical target volume (CTV), the planning target volume (PTV) and the internal target volume (ITV).

This hierarchical task analysis already captures most of the basic tool and design requirements that are needed in such a system and that will also be used in the design of the new interface, which are listed in Appendix B.

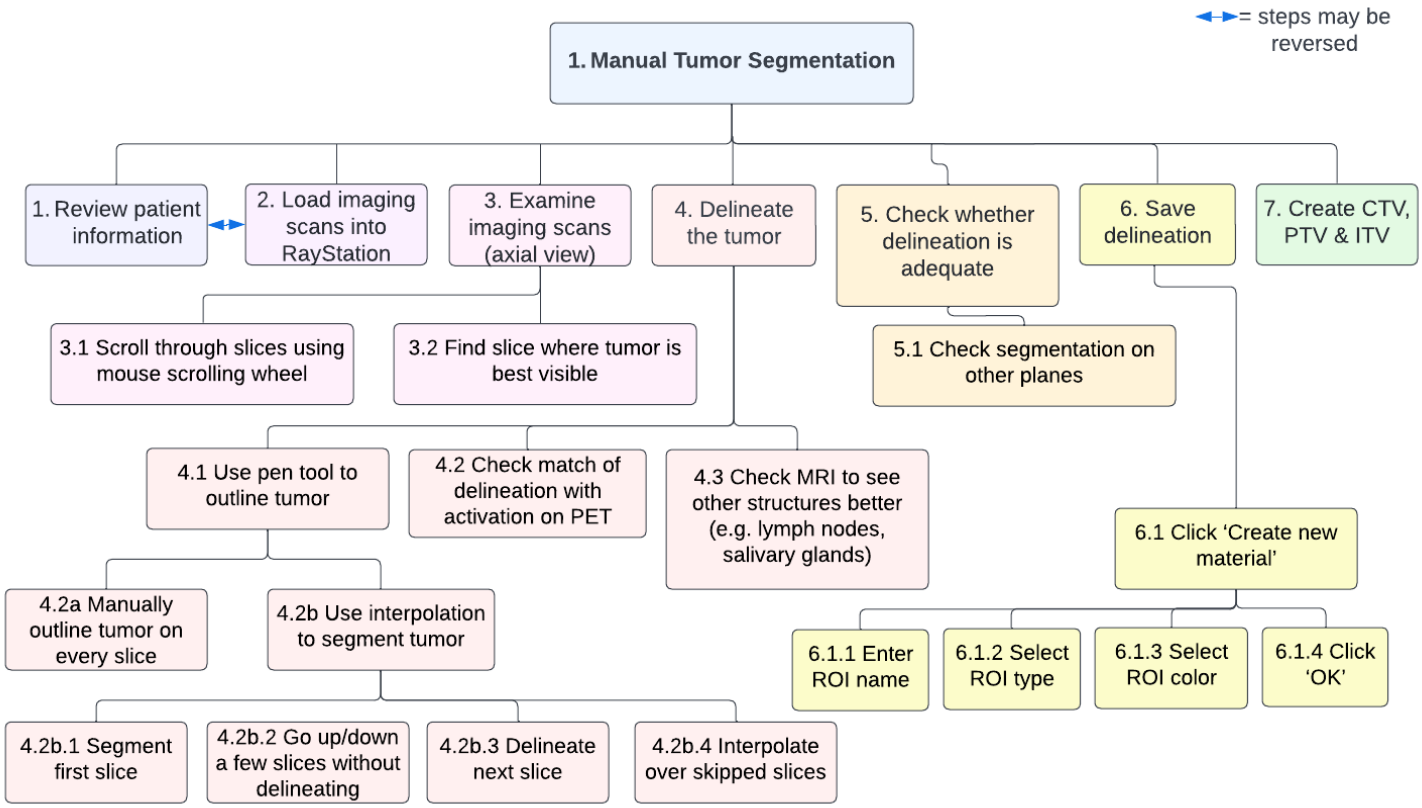


Figure 12. Hierarchical Task Analysis: Steps involved in Manual Segmentation

Tool Usage. The interviewees stated that they want a 'first-reader' decision-support system, as was defined in section '4.1 Computer-aided Diagnosis (CAD)'. Thus, the to-be-designed interface will be for a tool that predicts the initial tumor boundary automatically. The radiation oncologist's role will then be to review the model's predictions and to edit them if any corrections are necessary. The interviewees preferred this mode of usage since they mentioned that this would save the most time which could reduce their workload.

However, the previously discussed information on out-of-the-loop syndromes and inappropriate trust in CAD systems showed the dangers of designing intransparent systems with high levels of automation, which can especially be problematic for first-reader tools. To reduce this risk of blindly trusting the system and becoming complacent, the tool will be designed so that initial contour predictions are made by the models autonomously, but so that the user still has to engage with the predictions to decide on the most fitting prediction. Adding an interactive element to the model's output will increase the likelihood that users fully engage with the predictions which may make them more alert and critical. It will now be discussed how this interactive element was designed and incorporated into the interface.

8. Certainty Visualization

This interface and the included certainty visualization was built around an AI model by de Biase et al. (2022). Before stating the requirements for the certainty visualization, the model will briefly be described below to convey an understanding of the idea.

8.1 Deep Learning Model the Interface is Built Around

The automatic tumor contour predictions that were visualized in the interface prototype stem from a 2D deep-learning neural network for oropharyngeal cancer by de Biase et al. (2022). It determines the degree of classification certainty of each pixel being tumorous or not (de Biase et al., 2022). The model gives a probability map as its output representing its certainty. Hence, the idea was to design an interface that shows a visual representation of the different prediction certainties for given regions on a CT scan.

The DL model was trained on planning CT and PET images of 241 oropharyngeal cancer patients that received radiation therapy at the UMCG between 2014 and 2022. Testing was done on 61 patients. Bounding boxes of size 144x144x144 were used for training and testing the model. The training data set contained labeled ground-truth data in the form of manual delineations of the primary gross tumor volume (GTVp) by radiation oncologists of the UMCG. For further information on the training of the DL model, please refer to de Biase et al. (2022), which contains a description of a similar network trained on fewer patients.

8.2 Design of the Certainty Map

Next, it had to be established how the idea of the certainty visualization would be translated into the interface. In the semi-structured interviews with the radiation oncologists and UMCG researchers, first mock-ups of possible certainty visualizations were presented to understand their preferences. Three types of certainty visualizations were initially presented to them, as shown in Figure 13. These included a solid colored map, a solid gradient map, and a single solid contour. The participants indicated that they preferred the colored map as in option A) (Figure 13) due to better visibility of the differences between the probability thresholds. However, it was mentioned that it was problematic that the tumor could now no longer be seen due to the certainty map's opacity. Hence, it was agreed to keep the colors but instead of using solid colored areas, only colored contours would be used that have lower opacity, as seen in Figure 14.

Moreover, it was mentioned that it is important to consider whether the autocontours have their border outside or inside of the predicted tumor, as shown in Figure 15. The interviewed clinicians reported that outside borders would be more advantageous as they do not occlude any part of the tumor and hence allow for better visualization of the tumor.

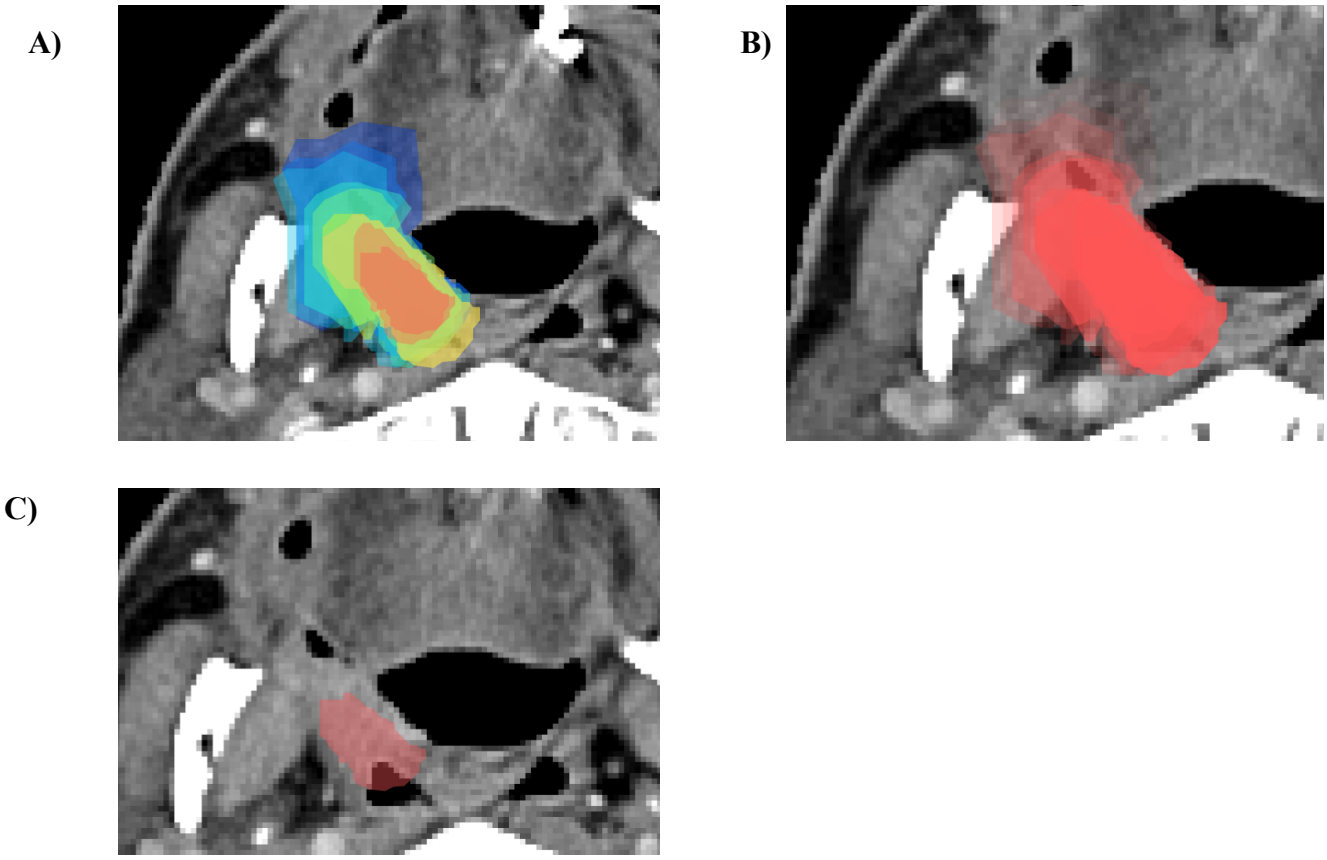


Figure 13. Initial Certainty Visualization Ideas on Cropped CT Scans: A) Colored solid map, B) Gradient solid map. C) Single solid contour

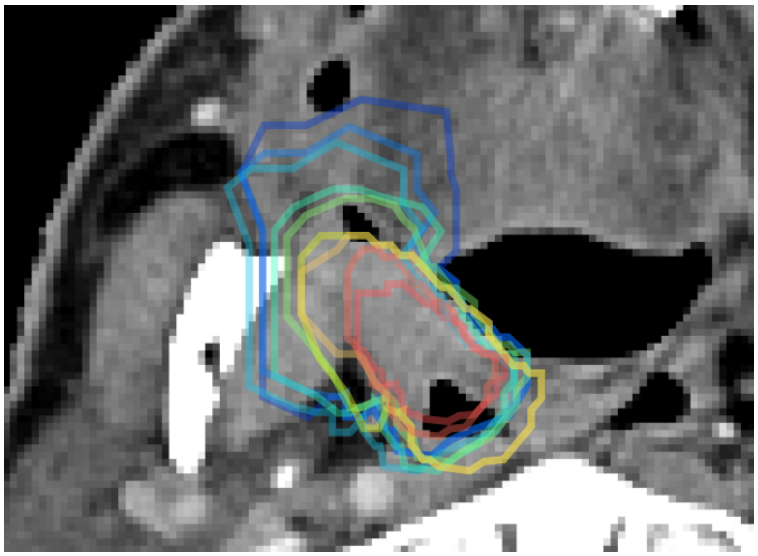


Figure 14. Updated Certainty Visualization: Colored Contours

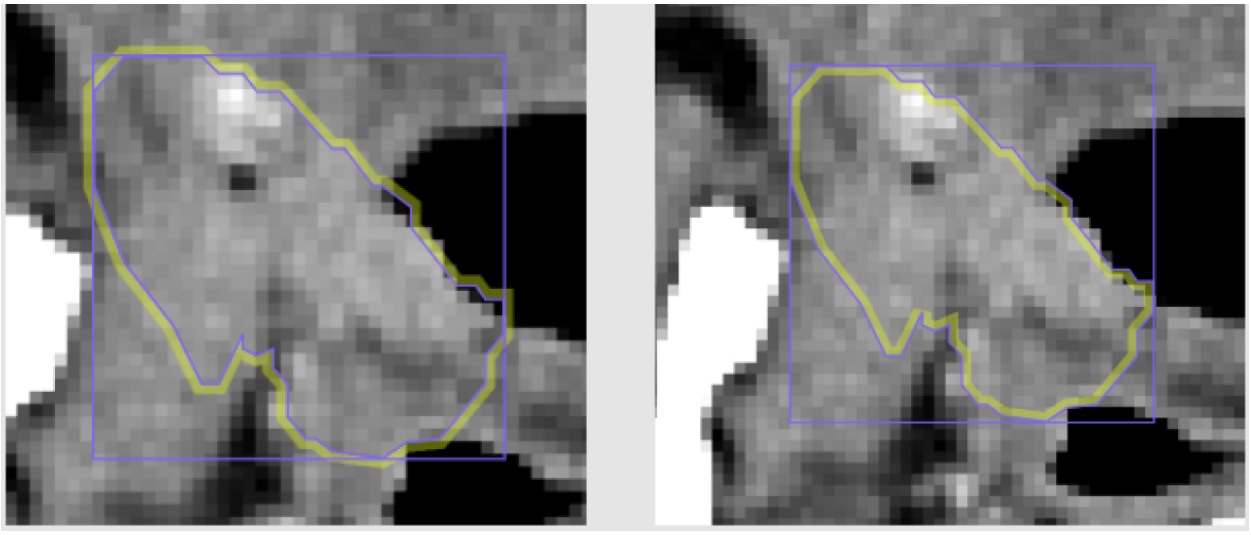


Figure 15. Autocontour Border Options: A) Outside border, B) Inside border

After having established the main requirements for the general functionalities of the tool and the certainty map, the visual design of the interface prototype was then tackled, as described below.

9. Prototype Design

Taking the information into account that has been gathered thus far, first prototypes were then designed for the segmentation interface. This was done using the prototyping software Figma.

9.1 Layout

Since the shadowing and unstructured interviews did not reveal any major issues with the general layout of RayStation, a similar layout was adopted for the new interface prototype. This should make adjustment to the new interface easier and interaction more intuitive since the radiation oncologists are already used to working with RayStation. Hence, the wireframe as seen in Figure 16 contains the primary view in which the biggest imaging scan could be viewed, as well as two smaller scans in the secondary and tertiary view to the right. The view selection at the bottom allows switching the images used for the primary, secondary and tertiary views and provides an overview of the available scans for the patient. At the top of the wireframe is a panel for the patient information and the controls, which will include buttons such as creating a contour, brush, eraser etc. The program controls at the very top would include buttons like open and save. The left panel would contain the delineations. As in RayStation, these should include the organs at risk which can be turned on and off so they can be used when delineating the primary tumor. Further, the actual targets should be present here.

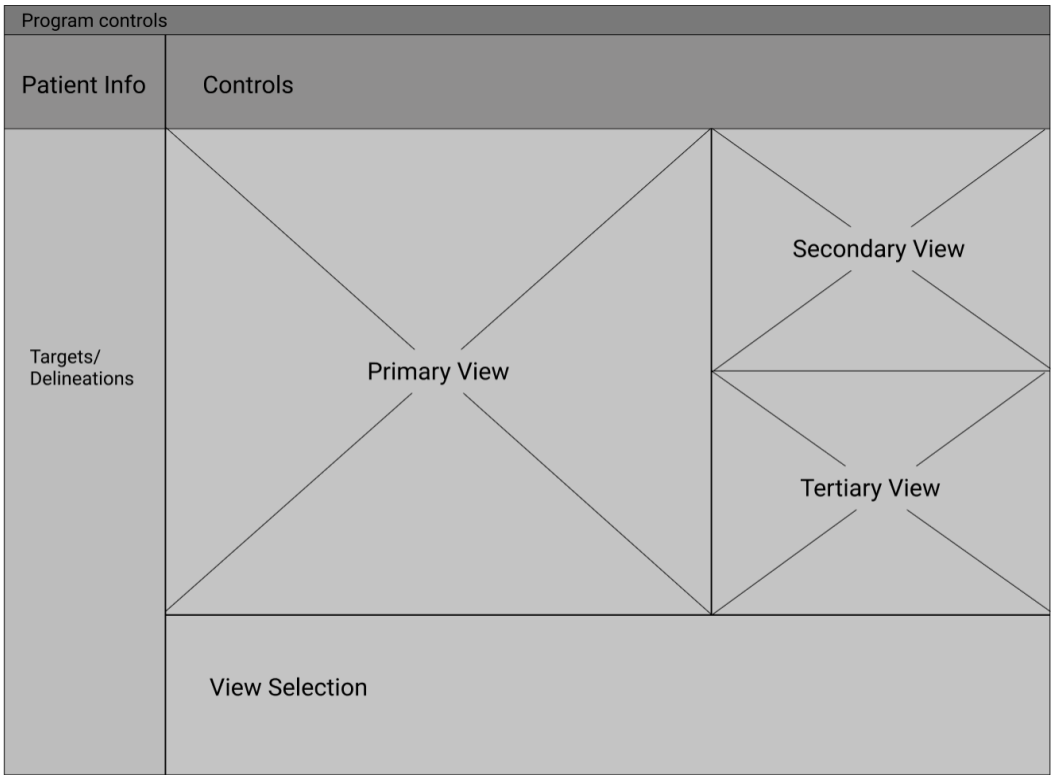


Figure 16. Wireframe showing Prototype Layout

9.2 Contents

It was identified that the interface would require two main pages or modes, as described below. Low-fidelity prototypes of these pages can be seen in Figure 17. Larger screenshots can be found in Appendix C.

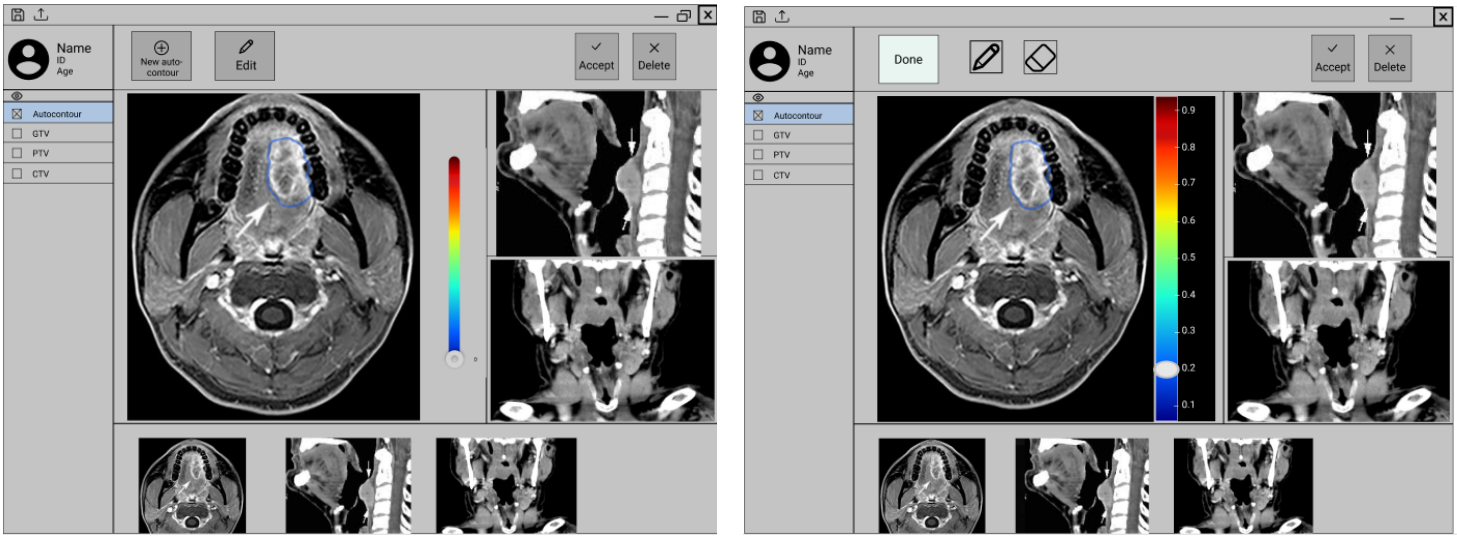
9.2.1 Pages

Homepage. The user would see the homepage when first opening the program. Icons were added to the program controls panel to open patient files, to save the file, and to minimize, maximize and close the program. Further, a patient icon was added to the patient information tab and the controls panel now has buttons to interact with and modify the autocontours. This includes a button to create a new autocontour, a button for editing the contour, and buttons to accept (save) and delete it. The targets/delineations tab has some sample targets and mock up images were added for the sake of prototyping (these images do not represent an actual H&N cancer case from the UMCG). Further, a first slider prototype was designed intended for changing the certainty threshold of the model's prediction. The commonly used RGB color range from red (very certain) to blue (very uncertain) was chosen for the slider to have a wide range of colors, increasing the ease of distinguishing them from each other.

Edit page. Apart from the homepage, a first prototype of the 'edit' page was designed. This would be visible after creating the autocontour and after clicking on 'Edit' in the control panel. This would reveal the brush and erase buttons using which the autocontour could be adjusted. 'Done' could be pressed once satisfied with the manual edits and when wanting to return to the homepage.

9.2.2 Interactive Certainty Visualization

As discussed previously, colored contours were chosen to represent the different prediction certainties of the DL model. It was agreed upon with the interviewees that an interactive certainty map would be useful that allows selecting a given certainty threshold and then switching between different thresholds to understand where the model is less and more confident about the presence of a tumor. To allow for this interactivity, a slider was added to the primary view with certainty thresholds from 0 percent to 100%, as seen in Figure 17. For instance, if the user would select a prediction certainty threshold of 90%, the interface would show an outlined region for which the model is 90% confident that it contains the tumor. In addition to the slider labels, the colors on the slider correspond to the different colors of the probability map, so the user can easily understand which threshold they are currently viewing on the CT scan.



A) **B)**

Figure 17. Prototype Version 1 of A) Homepage and B) Edit page (right)
Note. Secondary and tertiary view figures from van den Brekel & Castelijns (2005) and Hennessy (2015).

Next, more details were added to the prototypes.

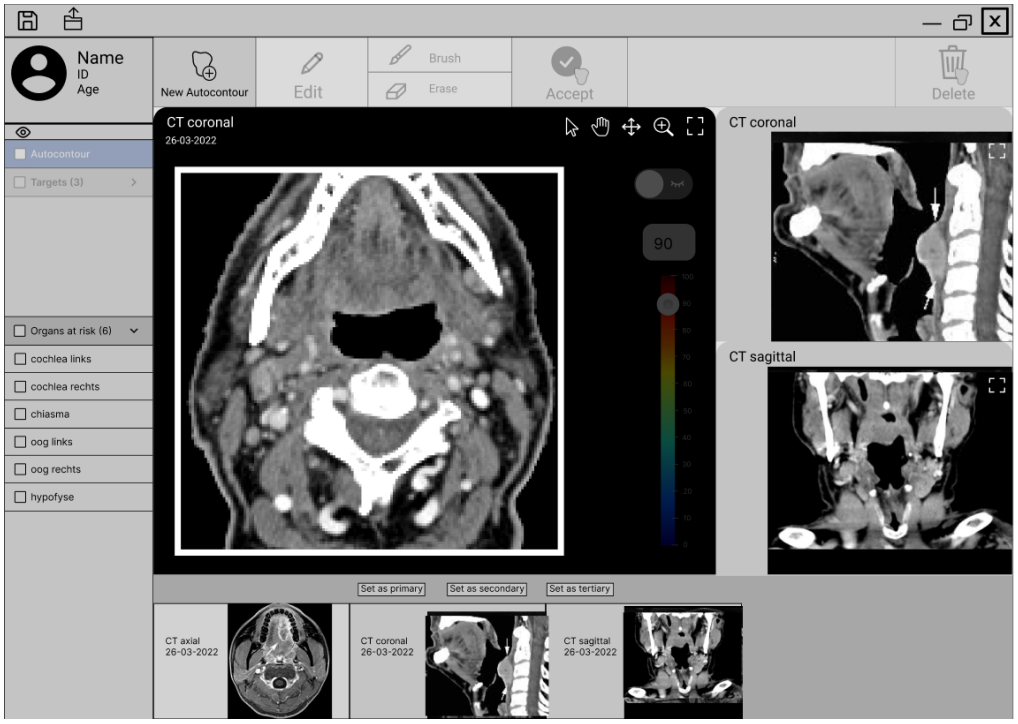
9.3 Buttons

Instead of having pages with different button locations for the homepage and the 'Edit' page as in Figure 17, it was decided that the button locations should remain consistent to reduce confusion. Hence, all main buttons were designed to have an active and an inactive state. The clickable active buttons have full opacity, while the inactive buttons appear less opaque to communicate that they cannot be clicked on a given page.

For instance, Figure 18 A) shows the screen that would be visible after the patient's scans have been loaded into the tool and before the autocontours have been created. At this point, the only actions the user can take is to switch the organs-at-risk segmentations on and off, to change the images in the gallery, or to create the autocontour. Hence, these are the only buttons that are active and appear clickable. The updated 'Edit' page in Figure 18 B) would appear after the contours have been created and the user clicks 'Edit'. Here it has been made clear that the 'New Autocontour' button is no longer relevant since the autocontour has already been created, while the 'Edit' button is now darker to show that the interface is in edit mode. The buttons brush, erase, accept and delete are now in their active form, as well as the slider, so the user can perform manual edits of the predicted contours.

Further, icons were added to the buttons for easier navigation and more clarity. Tool icons were also placed in the primary view to be able to select, scroll, move the scan, zoom in and out, and maximize the scan. Lastly, a switch was added to toggle the autocontour on and off as well as labels for the slider indicating the certainty thresholds and an indication of the threshold the slider is currently set to for easy usage. Additionally, the image views were labeled so the user can identify which modality they are looking at and when the imaging scan was taken.

A)



B)

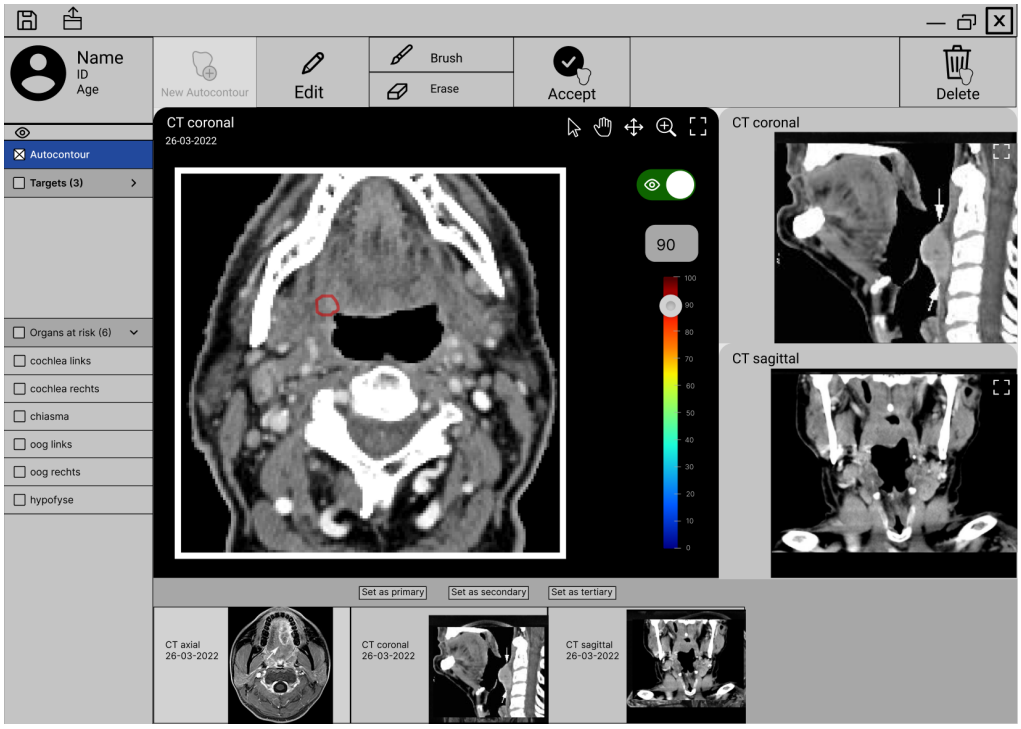


Figure 18. Prototype Version 2- Added Details and Active/Inactive Buttons on A) Homepage and B) Edit Page

Note. Secondary and tertiary view figures from van den Brekel & Castelijns (2005) and Hennessy (2015).

9.4 Colors

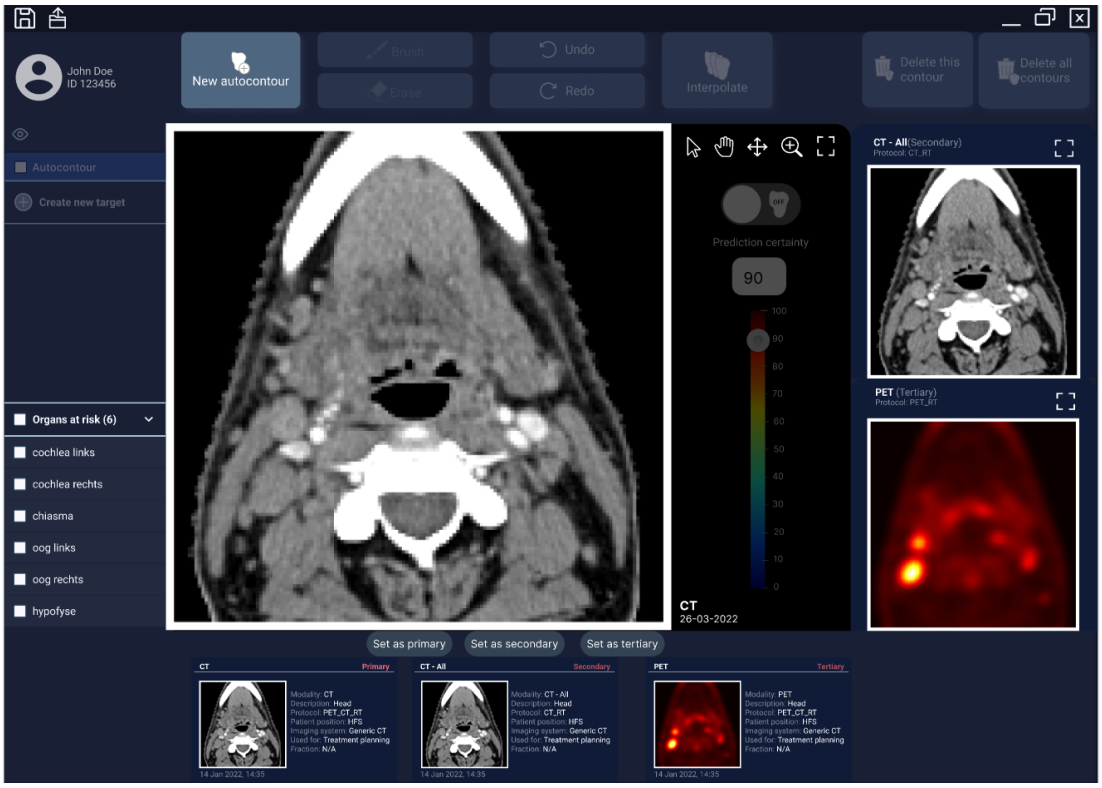
Since the manual segmentation software RayStation used at UMCG has a very dark theme with gray and black being the primary colors, a dark color scheme was also opted for in this prototype. This was done to make the new interface more similar to RayStation to avoid too big adjustments and dark colors can also be easier on the eyes than light ones, which could decrease fatigue when radiation oncologists spend several hours on their computers segmenting tumors. However, some color differences were desired to make the interface stand out and to make it more clear that it is separate from RayStation. Hence, a dark blue color theme was opted for, as seen in the mood board used for selecting the interface colors in Figure 19. Dark blue is a calming color that is frequently used in the medical domain and was hence deemed appropriate for this interface.

Figure 19. Mood Board for Color Choice of the Prototype.
Note. Images from SergeyBitos (n.d.), K. (n.d.), ZinetroN (n.d.), and Bureau Oberhauser (n.d.).

The selected colors were then added to the prototype along with further details and refinements to result in the final prototype, as depicted in Figure 20. The button ‘Edit’ was removed as it was deemed an unnecessary intermediate step upon further consideration. Instead, the brush and erase buttons were placed onto the main screen directly. Furthermore, an ‘Interpolate’ button was added, as the requirements analysis revealed that the radiation oncologists frequently speed up their manual segmentation procedure by using automatic interpolation between two end slices. In the automatic tool, the predicted contours of two end slices could also be manually edited, after which the user may decide to interpolate between the manually edited slices. Moreover, instead of one ‘Delete’ button, the updated interface now has two deletion buttons. Firstly, the ‘Delete this contour’ allows the user to delete the contour on the slice they are currently viewing in the main view in case they deem it as inaccurate and would rather wish to manually segment the tumor on a given single slice. ‘Delete all contours’ should be used if the user wants to remove the entire prediction on all slices at once. Since radiation oncologists frequently use CT and PET scans simultaneously, the PET images were now also included in the tertiary view.

The secondary view now contains an ‘all-in-one’ probability map which presents the predictions at every probability threshold available, with the colors corresponding to the colors on the slider. This is mainly intended to offer an overview of how the predicted contours change with the different probability thresholds. It allows comparing different thresholds and could motivate the user to examine a single probability more closely in the main view in case of any prediction thresholds that need a closer inspection.

A)



B)

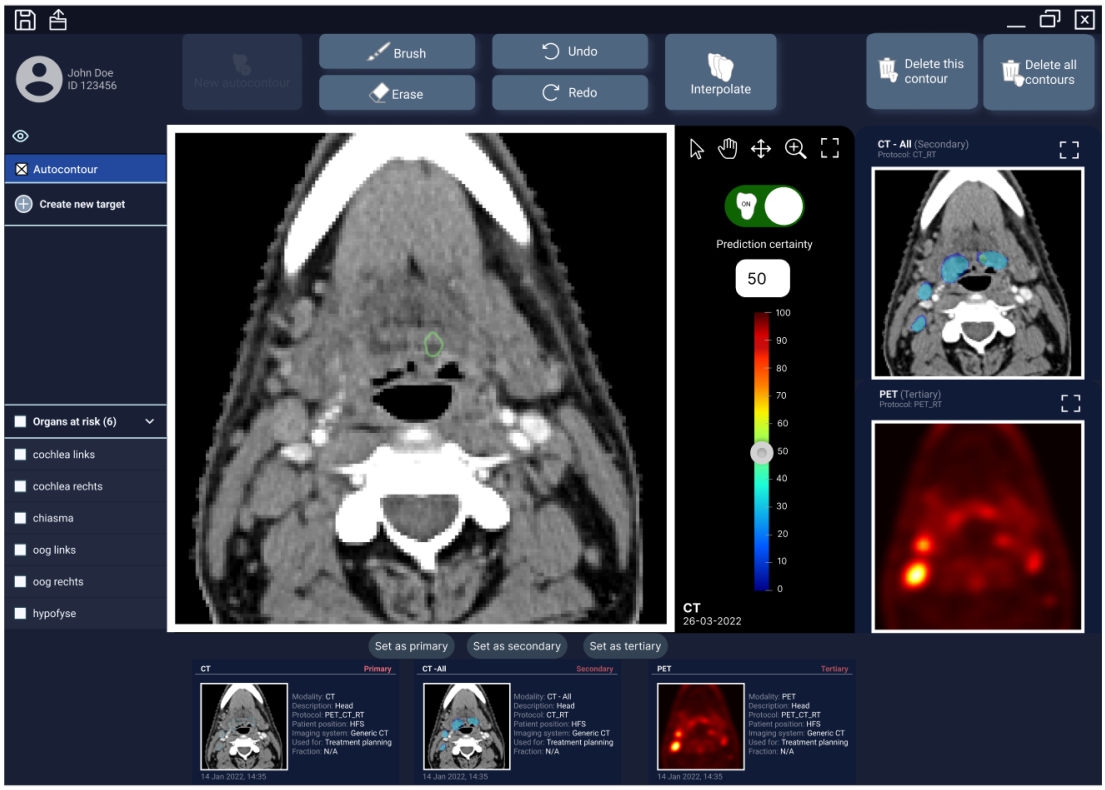


Figure 20. Final Prototype with A) Home Page and B) Edit Page

9.5 State Awareness

Since this is an interface prototype for a semi-automatic tool, several considerations had to be made so the user knows what is going on at every point and does not get confused by the state of the tool.

Progress Indication

Firstly, when the user prompts the tool to create a new set of autocontours, the interface will display a progress bar as seen in Figure 21. Since it is expected that the tool would take a few minutes to create its predictions for new patients, the user should be informed that it is currently processing the images. If this is not done, frustration and confusion could arise since the user may believe that the tool is not reacting as expected.

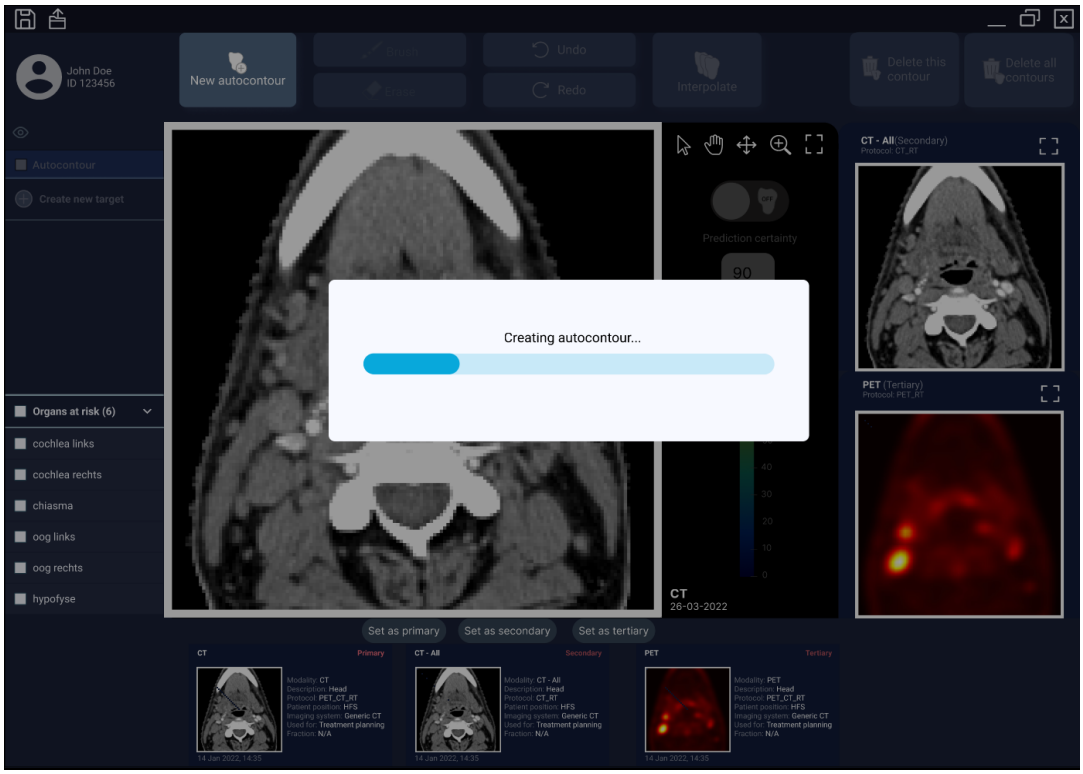


Figure 21. Progress Bar- Creating Autocontour

Warnings

On some slices, the DL model was not able to make predictions with high certainty, for instance when the tumor was very small or barely visible on a given slice. In Figure 22 it can be seen that the all-in-one map only contains probability certainties up to about 50%. When the user would select a higher percentage such as 90%, there would be no predicted contour that can be displayed. To inform the user of this, a warning was included in the interface that reads “No predicted probability available at this threshold.” This was done to prevent the user from thinking that there is a general error with the tool, and instead promote their understanding that for certain slices the model cannot reach a certainty above a specific threshold.

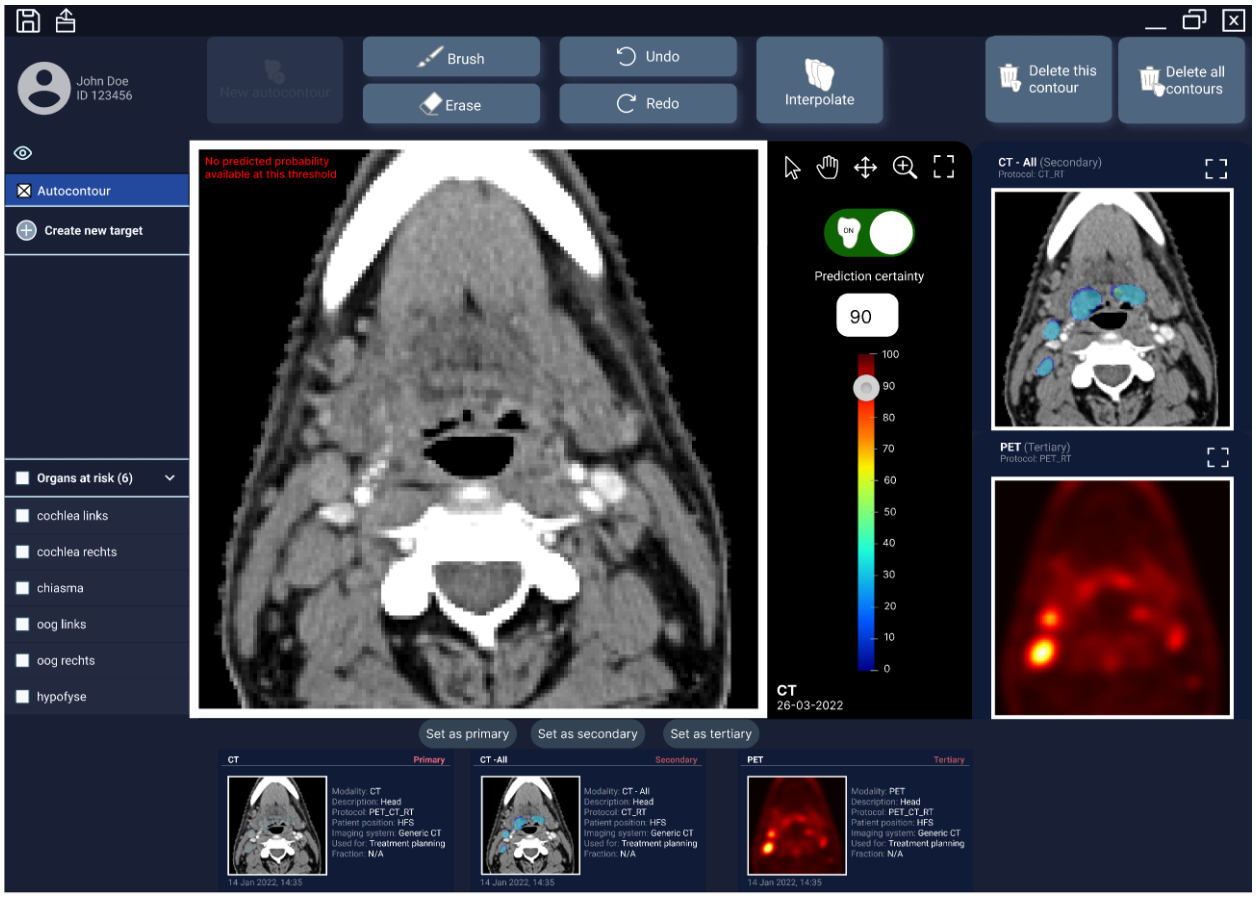


Figure 22. Warning When no Prediction is Available for a Selected Certainty Threshold

PART 3: EVALUATION

After the prototype design for the automatic tumor segmentation interface was completed, a user evaluation was conducted which will be discussed in this chapter. The user evaluation was carried out to examine whether the radiation oncologists at UMCG deem the general idea of such an automatic tumor segmentation study useful and clinically feasible. Further, it was investigated whether the current prototype interface design is appropriate for supporting interactions with the deep learning model's certainty map of the predicted tumor contour. Since the UMCG currently does not use a similar automatic tool for tumor segmentation yet, this study's goal was to identify any other requirements that were missed when designing the prototype but that should be incorporated into the interface when expanding on this design in the future. Due to the exploratory nature of this project, a fully functional system was not yet built, but instead the interactive prototype was evaluated and recommendations for improvements were made which will be discussed later on. The design of the user testing will now be discussed along with the results, followed by recommendations for changes.

10. Methods

Participants

Nine clinicians from the UMCG participated in this user testing by voluntarily signing up. Recruitment was done using flyers and emails that were sent around in the radiotherapy department of the UMCG. Seven participants were radiation oncologists, one was a medical doctor and radiation oncology researcher, and one was a radiotherapy researcher. Seven participants were specialized in head and neck tumors, while one was from the lung department and one was from the urology and palliative department. Even though the interface includes predictions for head and neck tumor contours, similar tools could also be designed for other departments in the future. Further, the focus of the user testing was not to judge the accuracy of the predictions, but rather to evaluate the utility and usability of the prototype in general, hence participants from other specializations were also recruited.

The experience in tumor segmentation ranged from approximately 6 months to 20 years. Two participants were already familiar with the purpose of the project and had seen previous versions of the prototype since they were shadowed and interviewed during the requirements analysis. Even though they were not completely new to the interface, they had not seen the final version of the prototype.

All participants had experience with RayStation, the current manual segmentation software. None of the participants had extensively worked with similar automatic segmentation tools before. One participant mentioned that they had tested a similar tool by a company in the clinic, but that this was very disappointing due to its low accuracy and due to the fact that it was only based on CT scans. Hence, this tool was never implemented at the UMCG.

Materials

The user testing was conducted in a room in the radiotherapy department of the UMCG. The prototype interactions were carried out using the prototyping software Figma, run in the browser version using Google Chrome. The mouse and the keyboard were used for interaction with the prototype. The setup consisted of two Windows PC monitors. One was used to display the actual prototype, while the other had a Google Sheets presentation containing variants of the main prototype that were shown after the exploration of the prototype. The variants were covered during the prototype interactions so participants wouldn't be distracted during the prototype testing phase. To record the screen during the prototype interaction, a Google Chrome browser plugin of the software Screen Recorder was used. An iPhone 10 was used for the voice recordings.

Design & Procedure

Ethical approval was obtained for this user evaluation study from the Research Ethics Committee (CETO) of the Faculty of Arts, University of Groningen. Each user evaluation lasted for approximately 30 minutes. Some participants stayed a bit longer depending on their availability and whether or not they still had feedback to discuss. Upon arrival, the participants were

welcomed, after which they read an information brochure about the purpose of the study and how their data will be processed and stored (see Appendix D), as well as an information sheet with details about the interface and the deep learning model (see Appendix E). The most important points were then repeated verbally to them in order to ensure everyone understood the information in the sheets and had sufficient knowledge for the user evaluation. The participants then gave their informed consent (see Appendix F), after which the screen and voice recordings were started.

After this, demographic data were collected. This included how much experience they have in radiation oncology/radiotherapy and tumor segmentation, if they use RayStation in their everyday tasks, and if they have used an automatic tumor segmentation tool before. During the sign up procedure, they were already asked for their job title, their department, their specialization, and for any additional demographic information they think would be beneficial for us to know.

Next, the free exploration phase started. Since the participants knew the purpose of the interface, they were asked to try to simulate creating a real automatic tumor segmentation with this prototype. They were asked to investigate the prototype and to simply do whatever comes naturally to them. Limited instructions were given to avoid leading the participants too much to examine if the interface is explainable by itself. The aim was to examine whether all buttons were in intuitive locations, or if any navigation issues or problems with the functionalities would arise. Furthermore, the steps taken during tumor segmentation can show quite a lot of variability between different clinicians. The goal was to analyze each participant's way of working in order to review whether the interface supports all interaction methods. Hence, the instructions during the prototype exploration were kept to a minimum and they were only guided if they encountered any issues or if they did not know how to proceed. Participants were asked to think aloud during this stage, according to the think aloud method (Ericsson & Simon, 1993, as cited in Ericsson & Simon, 1998). This involves verbalizing what the user is looking for, what they are currently doing, what they like and dislike about the interface and what their next steps are.

While participants were thinking out loud, they were sometimes asked follow-up questions to make them elaborate on their thoughts. For instance, it was sometimes asked "What do you think of the certainty map and why?". The researchers were also open to questions throughout the user testing and responded or corrected participants when misunderstandings occurred while trying not to guide or bias them.

After exploring the interface for about 10 minutes, the participants were indirectly pointed to certain functionalities they might have missed. For instance, they were told "There is a way to turn the delineation prediction on and off. Could you try finding this?". This method was used to ensure that all participants were aware of all the functionalities, but it still allowed examining whether the remaining buttons were placed intuitively once told about their existence.

Following the exploration phase, participants were asked about their general impressions of the interface. They were asked how the interaction went, if they think the interface is pleasant, if they have other general remarks and what they think of the utility and feasibility of this idea.

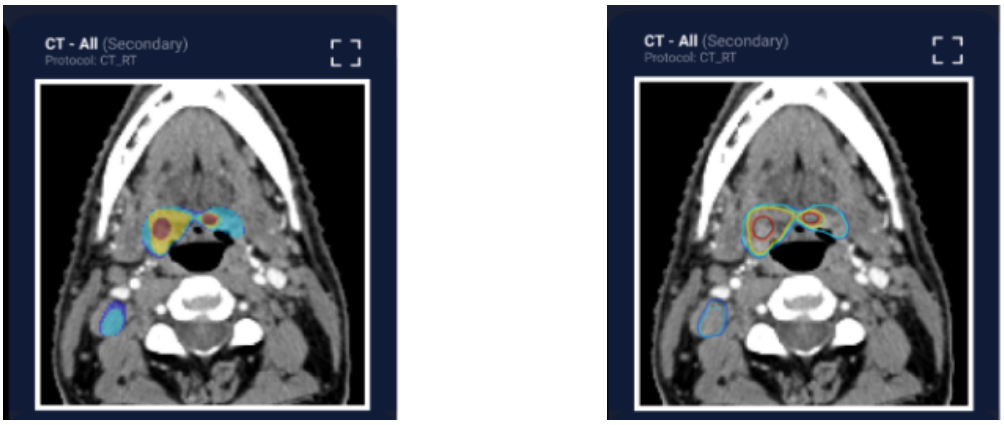
This was done verbally to allow them to 'vent' and to give them the opportunity to share anything still on their mind that hadn't been mentioned before taking the questionnaire.

Then the participants responded to a questionnaire as seen in Appendix G. This consisted of rating statements on a Likert Scale ranging from 1 (strongly agree) to 5 or 7 (strongly disagree), depending on the scale. While the participants were responding to the questionnaire they were told to elaborate on any responses they may wish to and that they may ask for clarifications when there was doubt about the meaning of a certain question. The questionnaire was created from a variety of validated questionnaires and subscales. The included scales were:

- *System Usability Scale (SUS)* (Brooke, 1996): To measure general satisfaction and usability of the prototype. The complete questionnaire was included.
- *Computer System Usability Questionnaire* (Lewis, 1995): To measure satisfaction with the user interface. Only the 'Interface Quality' subscale was included. The response option NA (not applicable) was removed since all statements were deemed relevant for this research and this was done to encourage responses.
- *7 own questions:* 7 questions developed by the researchers about automatic segmentation and certainty maps were asked.
- *Human-Computer Trust (HCT) scale* (Madsen & Gregor, 2000): To measure the understanding the participants have of the system and the extent of (over) trust. Only the most relevant items were selected from the subscales of Perceived Technical Competence, Perceived Understandability and Faith. Limited items were included for time reasons and as to not confuse participants with irrelevant questions.
- *Trust in Automation Questionnaire* (Körber, 2019): To examine general attitudes towards automation and the trust in this system. Again only included the most relevant items.

After responding to the questionnaire, the last step was to review different variants of the interface prototype. This was done to examine the participants' preference regarding the layout and the contents of the interface. For ease of use and to speed up the user testing, these variants were presented in a Google Slides presentation and participants verbally commented on their preferences. First, it was asked whether the participants prefer having a solid (filled out) map in the all-in-1 scan (secondary view, as seen in option A in Figure 23) or only having the outlines in the map (option B). Option B was more akin to the contours they saw in the primary view and has the advantage of showing more of the tumor region itself, while the certainty map is better visible in option A. In the second variant (Figure 24) the overlapped PET/CT scan was included and the participants were asked if this would be helpful or if it is not needed. Third, two buttons were added to the Controls panel (Figure 25). A button to change the level/window was added, since the radiation oncologists already use this in their manual segmentation tool and it wasn't yet included in the main prototype. This can be used to obtain a better visualization of the tumor by changing the brightness or contrast of certain regions Further, a button for customizing the thickness and opacity of the automatic contour was prototyped. It was investigated whether these customization options would be good to have. Lastly, an alternative layout was shown to the participants. Here the primary view allowed viewing multiple contours of different probability

thresholds simultaneously by clicking the respective checkboxes (Figure 26). In the original prototype only one contour could be observed at a time using the slider.



A) **B)**

Figure 23. Variant 1: Secondary View Certainty Map - A) solid vs. B) map outlines

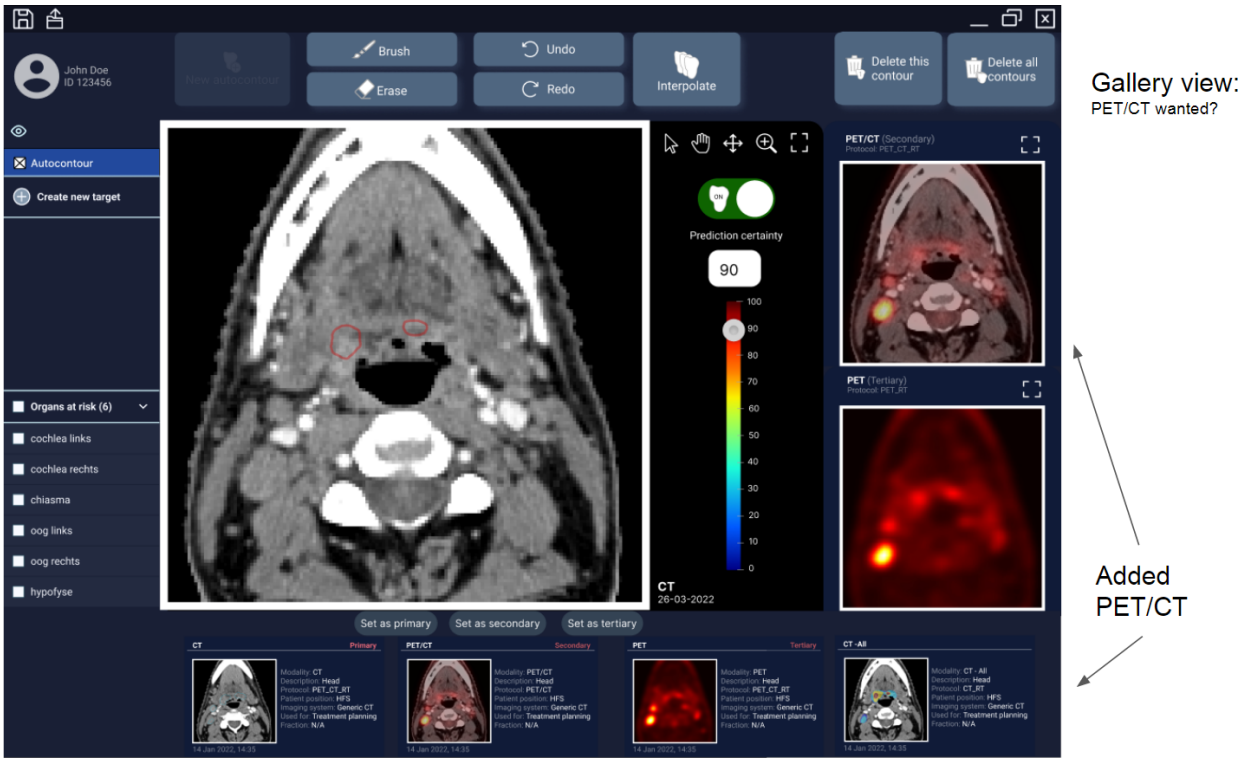


Figure 24. Variant 2: Overlapped PET/CT included in View Selection

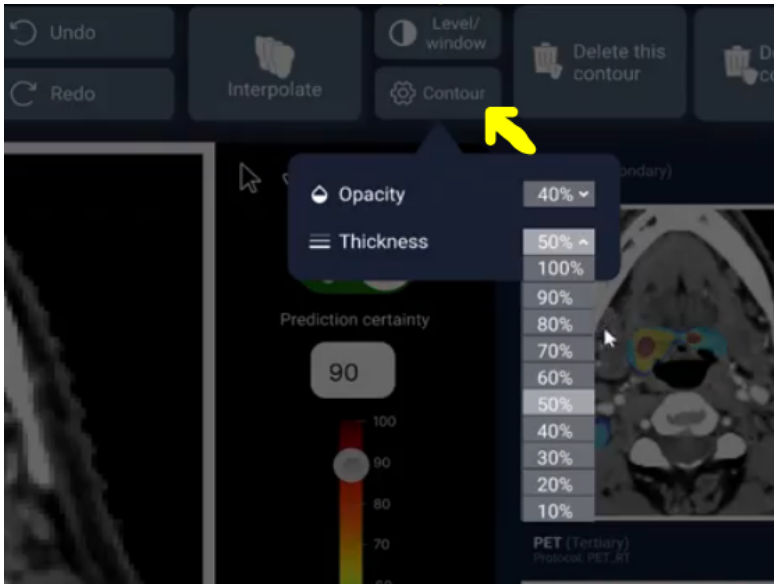


Figure 25. Variant 3: Buttons Added to Controls Panel to Change Level/Window and Contour Opacity and Thickness

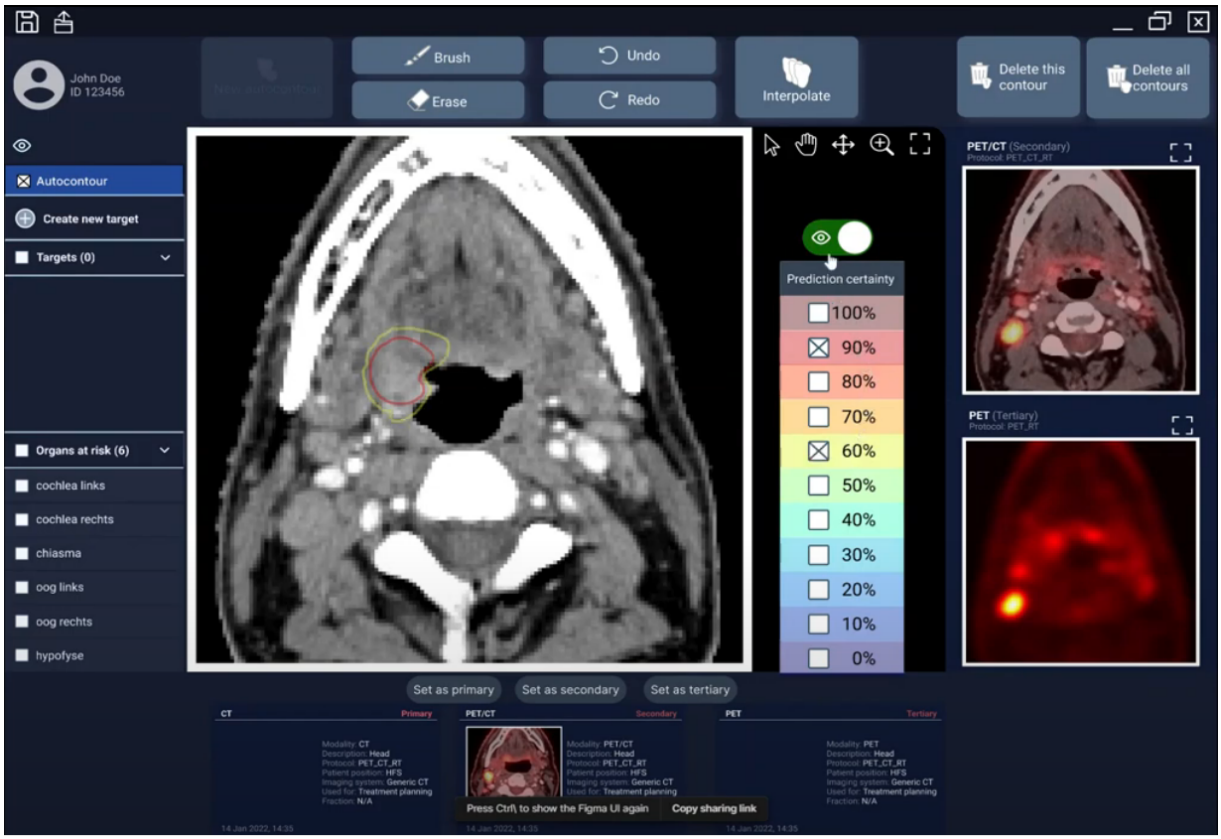


Figure 26. Variant 4: Layout Change - View Multiple Predictions Simultaneously

Following the variant review, participants were asked whether they had any more remarks and were then thanked for their participation. The voice and screen recordings were then stopped.

11. Results

The results from the interface evaluation will now be discussed along with suggested improvements for how to optimize the interface.

11.1 General Remarks

Several participants reported that they would ideally like to see such a tool be integrated with their current delineation software, RayStation. They liked the tool's functionalities and design, but mentioned that it would save time and effort if they did not have to switch between softwares. This also would make getting used to the new functionalities easier. While RayStation should investigate the possibility of adding an automatic tumor segmentation feature, the further results from this study will be discussed in the form of a separate interface. The final form of such a tool is less crucial, but what is of importance here is the functionalities included and how to best visualize the model's certainties.

Despite this preference for the tool to be integrated with RayStation, participants quickly got accustomed to the tool. Generally, the location of the buttons seemed to be appropriate. The participants easily found the respective buttons when they were looking for something or when prompted to complete a certain action.

The participants also made some remarks regarding the deep learning model this tool was based on. Firstly, participants requested automatic delineations of lymph nodes in addition to the primary tumor predictions. Further, as mentioned earlier, MRI images should be included in the training and testing of the model so the predictions also include information about the MRI and so that the predicted contour can be visualized on this modality as well. Lastly, several participants noticed that the CT scans used in this interface looked slightly 'rougher' than those in RayStation. They assumed that RayStation adds a filter on top of the scans to smoothen them. For the optimal transition to this system, the CT quality should be the same as that in RayStation.

11.2 Questionnaire Results

The results from the System Usability Scale (SUS) reflect that the interface had an excellent usability (Mean = 84.72/100). The current average score across participants of 84.72 is well above this scale's average usability score of 68 (Sauro, 2011). The main takeaways from these results are that the participants would like to use such a system frequently, that the system was easy to learn and to use, and that they felt confident while using the system. Further, not a lot of technical assistance is required to use the system.

The Interface Quality subscale from the Computer System Usability Questionnaire (CSUQ) revealed that the participants were also satisfied with the interface (Subscale Mean = 5.61/7). The responses suggest that the designed interface is pleasant, that most functionalities

expected of a system like this were also included in the prototype, and that there is general satisfaction with the system.

Next, results of the questionnaire items will be discussed that were specifically created for this study. Firstly, participants felt like they would be more confident in their delineations if they would use automatic segmentation as a starting point (Mean = 5.78/7). Second, participants perceived that their understanding of the model's predictions was augmented by the probability maps (Mean = 5.89/7). Seeing the certainties of the predicted outputs was preferred over seeing a single (binary) prediction (Mean = 5.78/7) and the participants did not feel very confused by the different probabilities (Mean (negatively worded) = 1.56). Further, time decreases were expected when using this system (Mean = 5.78/7) and the participants thought it would be feasible to use such a system in the clinic (Mean = 6.33/7). Lastly, participants did not prefer their usual manual segmentation method over using a tool like this (Mean (negatively worded) = 2.78). This suggests that the idea of introducing certainty maps into the interface was successful.

Results from the selected items of the Human-Computer Trust Questionnaire reveal that the users understand how the system will aid their decision making (Mean = 5.56/7), and that they perceive the tool to be helpful for decision making even if they do not fully understand the system's workings (Mean = 5.89/7). Moreover, participants indicated that they would not be totally confident that the system is correct when unusual advice is given (Mean = 3.11/7 (negatively worded)).

Lastly, the Trust in Automation Questionnaire reflects that users had a rather high level of trust in the system (Mean = 3.78/5). Participants were confident about the system's capabilities (Mean = 4/ 5) and felt like it works reliably (Mean = 4/ 5). Moreover, the system state was usually clear to participants (Mean = 4.11/5), which is important to avoid confusion and to prevent the user from getting lost in the system. Further, the participants felt like automated systems generally work reasonably well (Mean = 3.67/5).

11.3 Variant Results

The results from presenting the variants to the participants inspired some changes to the layout of the interface. Variant 1 revealed that 6 out of 9 participants preferred the color wash (solid) all-in-one map, while 2 participants preferred the contours and one person said to keep both as options. The latter is what was eventually decided upon to accommodate everyone's preferences. As seen in Figure 27, selection tabs were added to the top of each view allowing the users to switch between color washes and outlines. The colors of the tabs allow users to easily identify which option they have currently selected.

Feedback on variant 2 revealed that participants would also like to have the overlapped PET-CT available in the tool. Further, variant 3 showed that the buttons for adjusting the level/window of the imaging scans should be included in the final interface. Some participants also mentioned that this button should include different presets, like in RayStation. These may for instance include a bone, a brain and a larynx setting a quick optimal visualization of the respective structures. The button for customizing the contour's thickness and opacity also

presented in variant 3 was generally liked, but not seen as completely necessary. Most participants mentioned that it is good to keep in case someone wanted to change it, but that they probably would not use this frequently since they thought the contours were good as they were in the prototype. One participant mentioned that changing the opacity would especially be useful for the all-in-one map, since that included a color wash in the prototype which could occlude the tumor. Making this more transparent could allow for a better visualization of the tumor while keeping the color wash. These findings have been taken into consideration for the revised interface design, as shown in Figure 29.

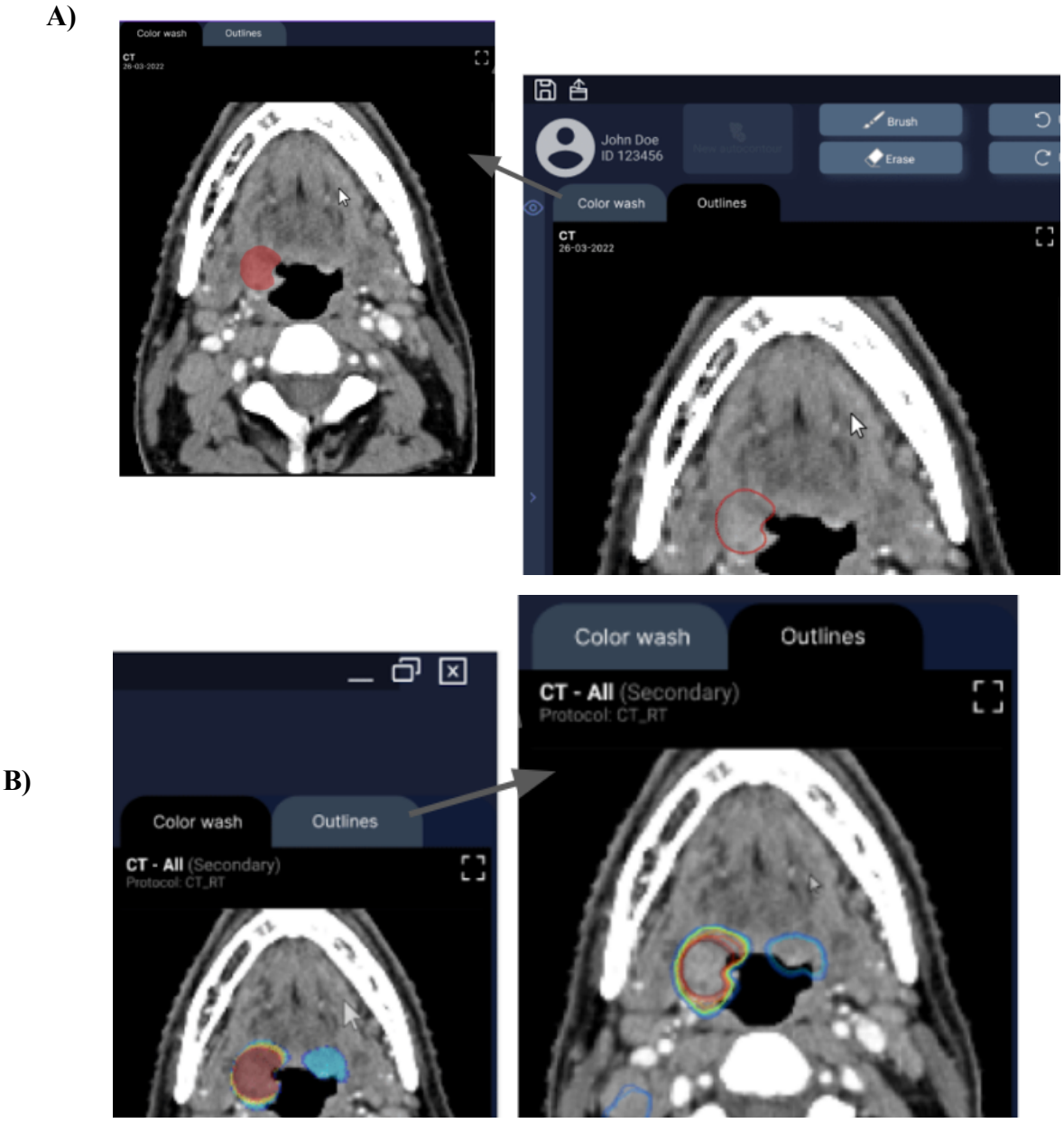


Figure 27. Added Selection Tabs to Interface to Switch Between Seeing the Predictions as a Color Wash and an Outline: A) Primary view (CT) with tabs ; B) All-in-one view with tabs

Furthermore, all participants agreed that it was a good idea to view multiple predictions simultaneously, as prototyped in variant 4. Since the participants preferred the slider design over the checkboxes, the two concepts were integrated to design a slider that allows selecting one main probability (the big handle) which is always needed when the slider is turned on, and as many sub-probabilities as the user wants to view (the smaller handle dots on the slider). To promote a better understanding of which probabilities have been selected, the color of the selected certainties on the slider labels differs from those not selected (Figure 28).

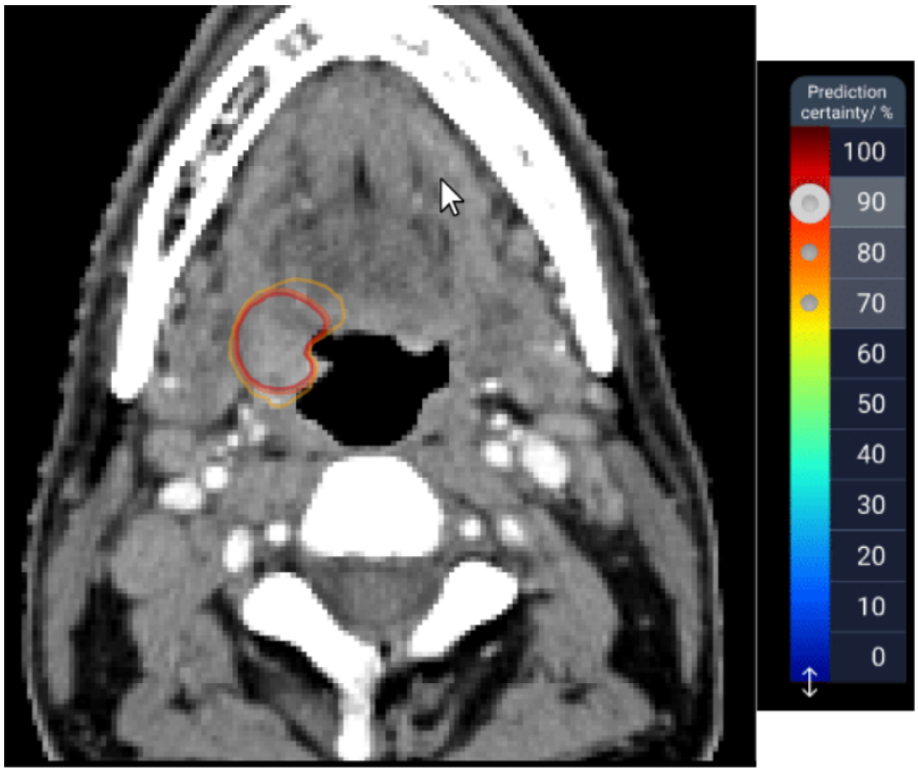


Figure 28. New Slider Design to Allow Viewing Multiple Prediction Thresholds at Once

11.4 Other Remarks

During the user evaluations the participants shared other valuable information on how the interface could further be improved to suit their needs better. These suggestions were accommodated in a revised interface design, which can be seen in Figure 29.

As the first improvement suggestion, participants mentioned that they would like the option to view the CT and PET scans as two big images next to each other, instead of only viewing the CT scan in a large format. To make more room so the scans can be seen in more detail, the region of interest (ROI) sidebar to the left of the interface was minimized. When the user clicks the minimized sidebar, an overlay will open that reveals all necessary information.

Second, participants requested the predicted contour to also be visible on the PET scan. This visualizes where the prediction would be located in the PET scan and having the predictions on multiple modalities facilitates judging their accuracy.

Moreover, the participants asked for the MRI to also be included in the interface, since this is also frequently used while segmenting tumors. The current deep learning model was only trained on CT and PET scans, but to further improve the model accuracy, MRI scans should also be used for the training in the future. This would also allow displaying the predictions on the MRI scans, as shown in Figure 29. Apart from the MRI, the participants requested having the option of loading all available scans and planes into the tool. This was accommodated by adding a horizontal sliding gallery at the bottom of the interface containing the overlapped PET-CT as well as all planes available.

Furthermore, a participant requested showing the mouse cursor on all imaging scans. While the user hovers over the CT scan with their mouse, a copy of this cursor will appear in the other scans as well, to show the user which area in the other scan they are currently looking at. This can be especially useful when the different modalities do not match up perfectly due to different patient positioning. The copies of the cursors have a lower opacity so the user can still easily identify which cursor is their main one.

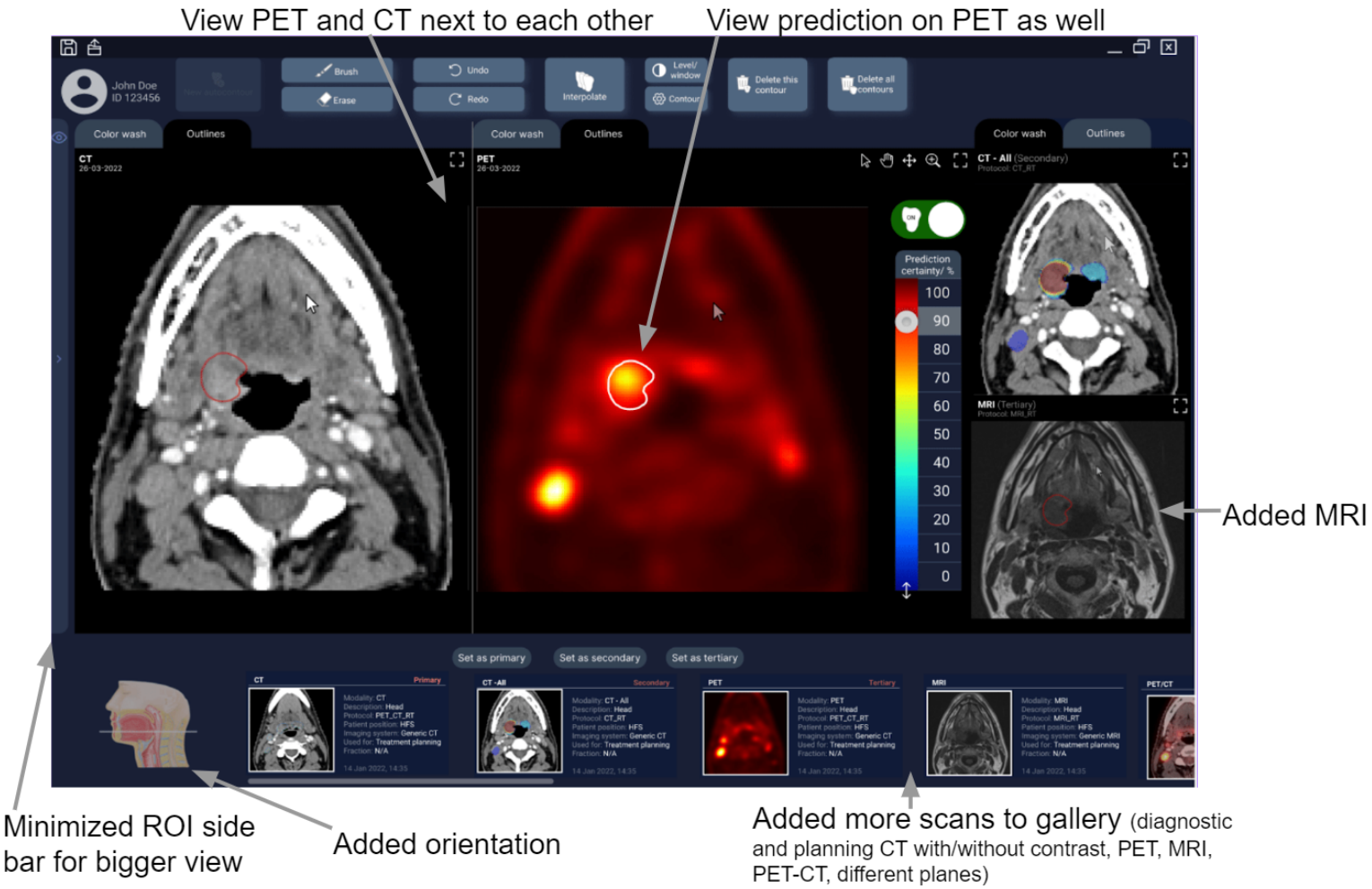


Figure 29. Revised Interface accommodating Participant's Feedback

Since some participants felt some confusion regarding which slice of the oropharynx they were looking at, an orientation aid was added to the interface. This is a head icon in the bottom left corner (Figure 29) that has a transverse plane through it, corresponding to the slice that is currently being viewed.

Next, a suggestion to improve dealing with false positives was given. Figure 30 shows an example of where the model predicts a false positive. Here two contours are predicted, one bigger one for the primary tumor and the smaller one is the false positive. This false positive prediction occurred because the PET scan shows some normal uptake in the tonsil area, which is interpreted as a tumor by the model. The all-in-one map helps detect that this is probably a false positive, since it only shows very low probabilities (blue colors) for this area and the high probabilities (red colors) are only visible in the area of the primary tumor. Thus, the participants generally were not very confused by the false positives, but they requested an easy way to only remove the false positive contour. This was prototyped by clicking on the false positive, after which a selection box appears. Then the user can press the 'Delete' button on their keyboard or use the 'Delete this contour' button in the interface. Such false positives should automatically be deleted across all slices so the user does not have to repeat this for each slice.

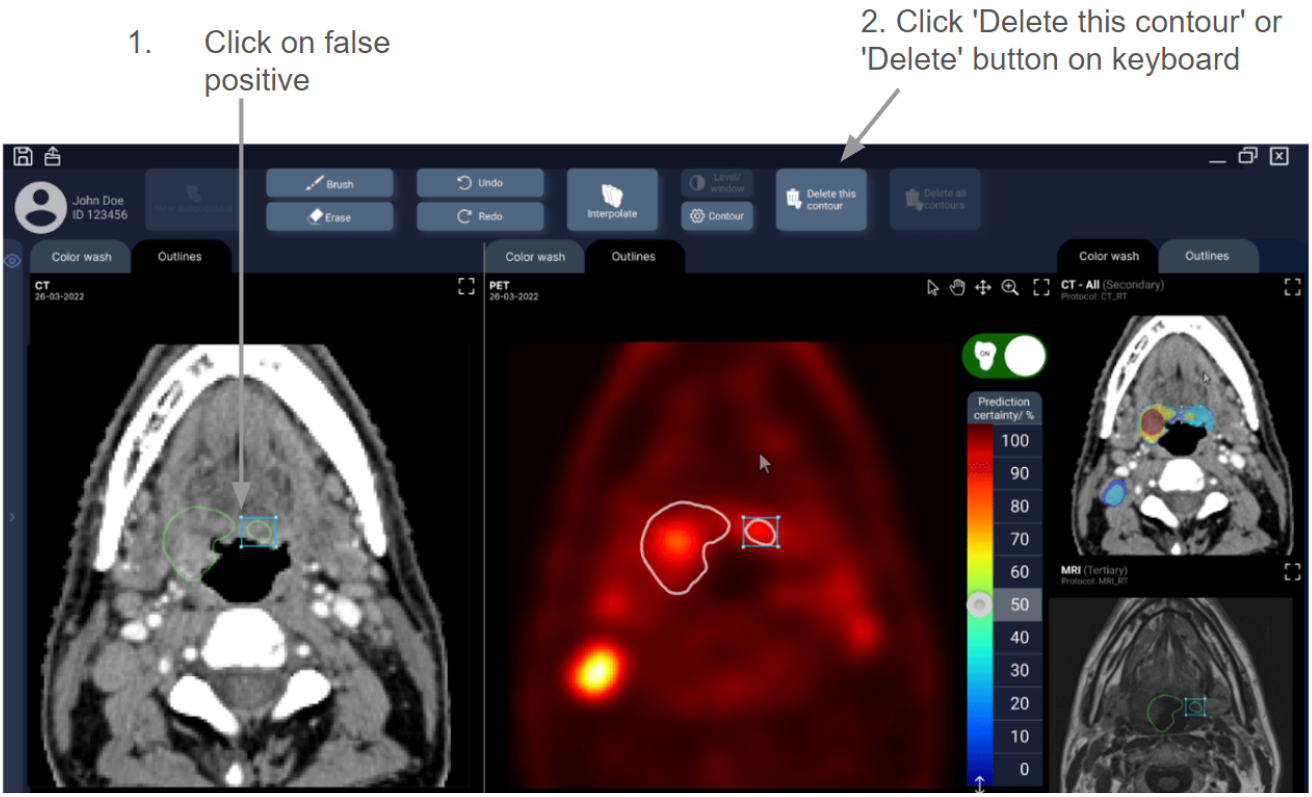


Figure 30. Added Option To Delete False Positives Across Slices

The next improvement is with regards to slices for which the model was not able to make predictions with high certainty, for instance because the tumor was very small or barely visible

on these slices. In Figure 31 it can be seen that the all-in-one map only contains probability certainties up to about 50%. When participants in the user testing selected higher percentages, such as 90%, a warning was shown that read "No predicted probability available at this threshold." Although users understood this message pretty rapidly, they suggested adding a prompt telling users what to do next. Hence, the revised interface presents the warning "No predicted probability available at this threshold. Change slice or probability threshold."

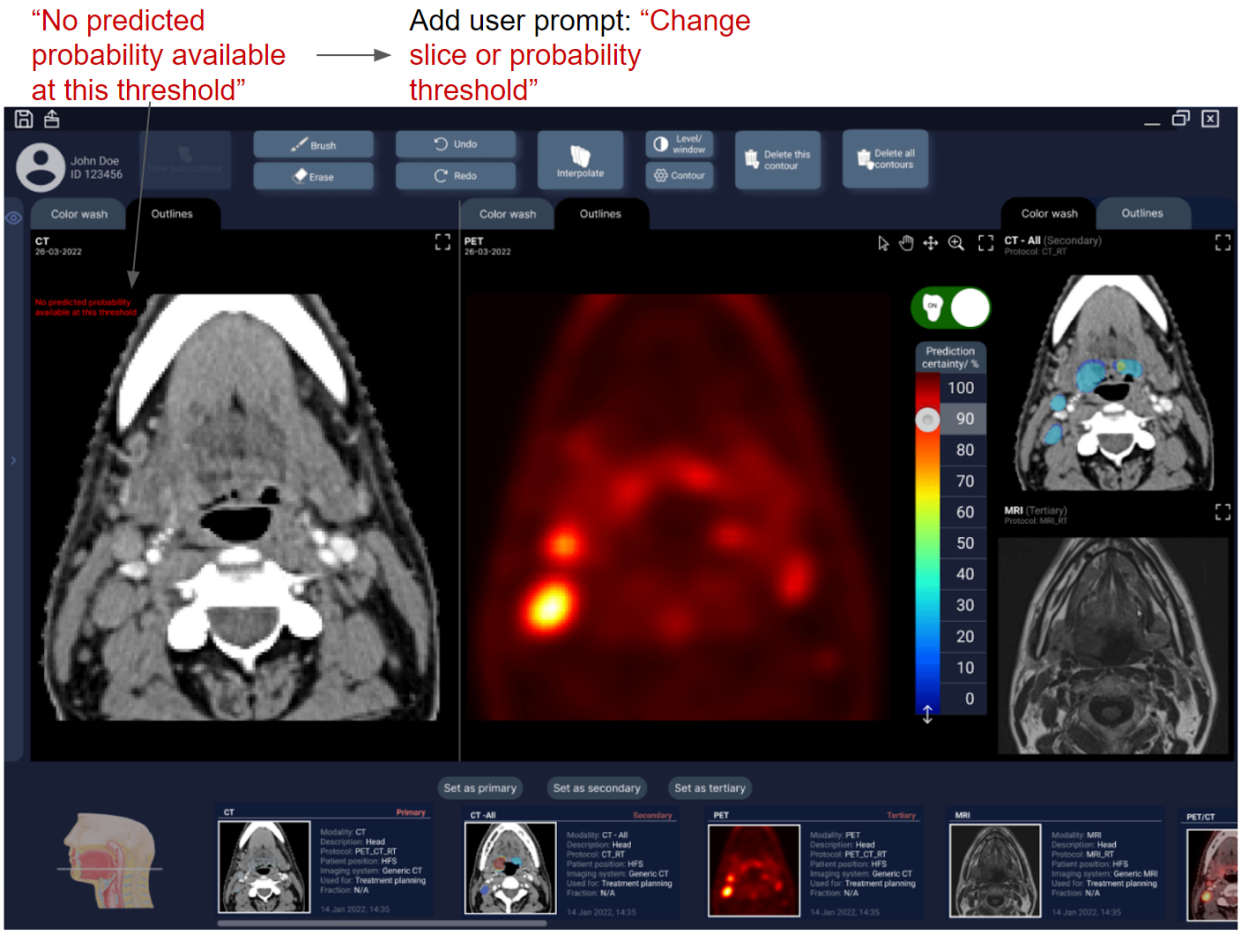


Figure 31. Warning Message that no Prediction is Possible at a Threshold of 90%. Users Suggested Adding a Prompt.

Some participants mentioned that they like the probability map, but that they probably would not use the low probabilities. Even though they can be useful for gaining more insight into the model's decisions or for detecting false positives as mentioned previously, users should still have the option to hide these low probabilities. Hence, a solution could be to have a handle at the bottom of the slider which users can drag up. When set to 50% for example, as shown in Figure 32, the lower percentages will be removed from the all-in-one map and a 'smaller' slider will be visible.

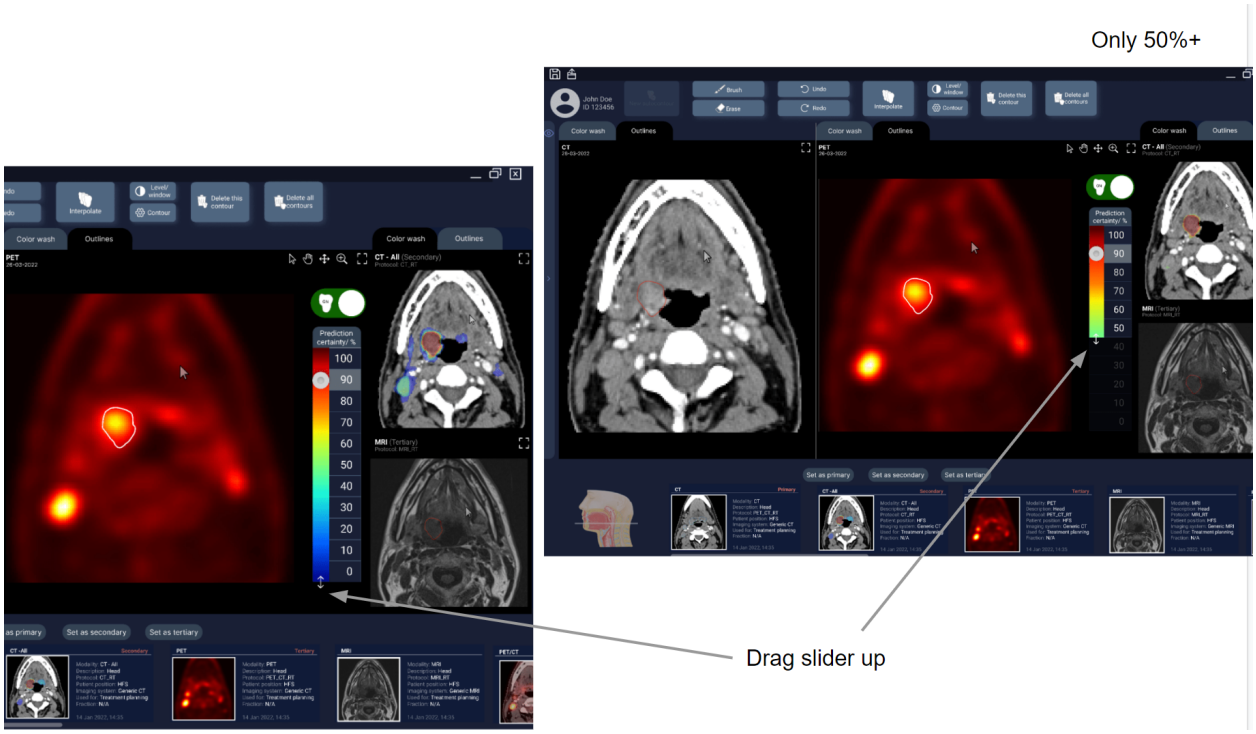


Figure 32. Added Option to Hide Low Probabilities from All-in-one Map

Moreover, it was suggested that the model's predictions should be differentiated from manual edits the user made. This has the advantage that changes can be tracked and it will remain clear what the model predicted and what the user decided to change. This could be useful to justify one's decisions when reviewing segmentations with other radiation oncologists. The contours could be differentiated by color or by dashed and solid lines, as seen in Figure 33. The sidebar should also include a clear legend so the user does not get confused about which line corresponds to which contour.

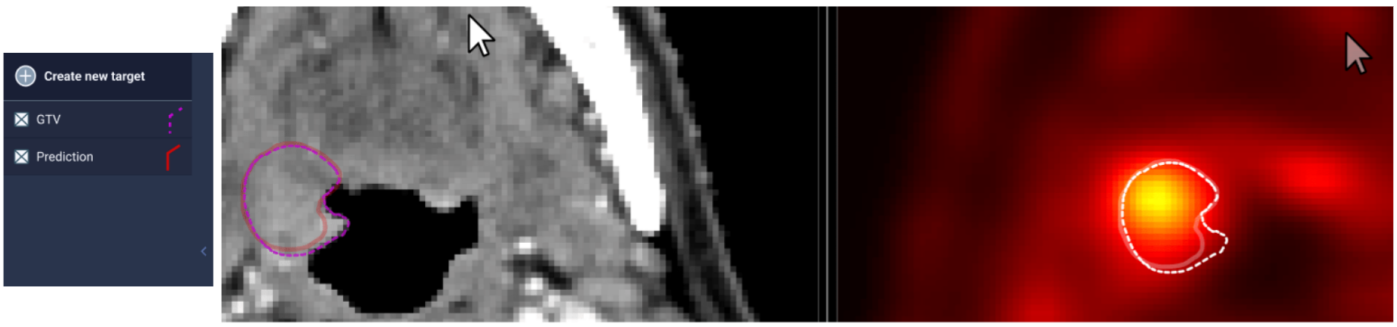


Figure 33. Using Different Colored or Dashed Lines to Distinguish Manual Edits From Model's Predictions

Lastly, one participant suggested that there should be an option to accept contours. If the user is satisfied with the automatic contour or with their manual edit, they should be able to press the 'Accept contour' button in the interface or press the 'Enter' key on their keyboard so save the contour for this specific slice. The interface will then show a signal, such as the green check mark in Figure 34 to communicate that this slice has already been reviewed. This allows for a clear overview of which slices still need review and avoids users overlooking any slices. It should only then be able to save or export the segmentations once every slice has been accepted.

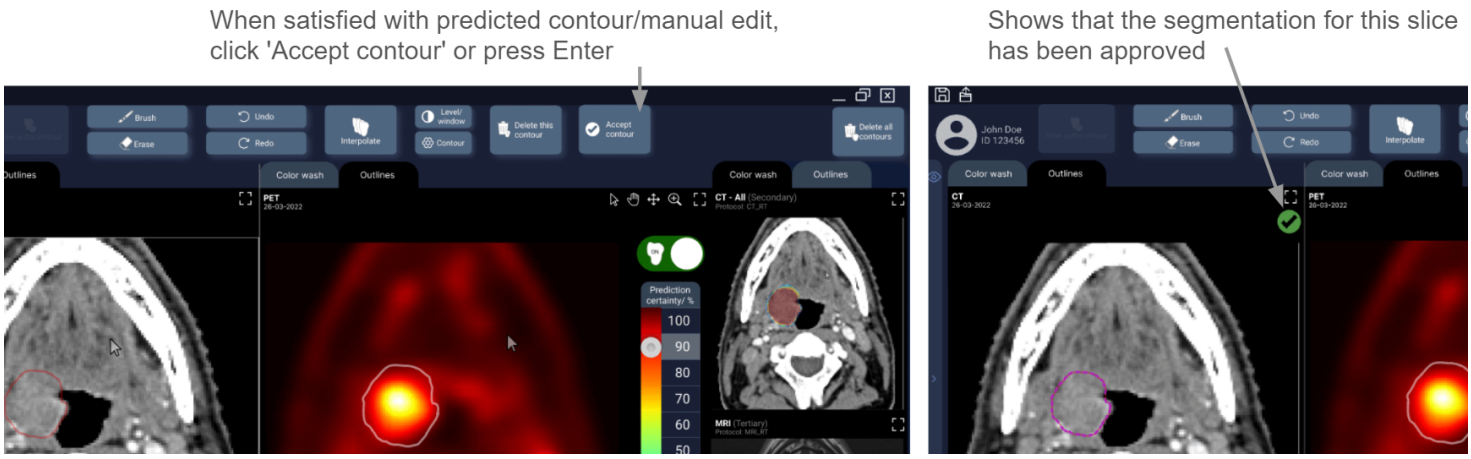


Figure 34. Added Button to Accept Contours – Allows Keeping Track of Reviewed Slices

PART 4: DISCUSSION

This project was aimed at designing a user interface for a tool for the computer-aided diagnosis of head and neck tumors. This was done to investigate the possibility of eventually introducing a semi-automatic tumor segmentation tool to the Universitair Medisch Centrum Groningen (UMCG). The expected benefit of such a tool is to decrease the inter-observer variability and the time needed for manual tumor segmentation. The main goal of the interface that was designed in this project was to allow the users to create, review and edit automatic tumor predictions from a deep learning model developed at the UMCG. Apart from this, it was investigated how to use findings from the field of explainable AI to make the outputs of the deep learning model more intuitive and understandable. This was examined since the black-box nature of similar AI models has led to low clinical implementations and a lack of trust in these diagnostic aids. Hence, this project focused on visualizing the model's certainty in its predictions to promote a better understanding of the model's decisions and ultimately more appropriate trust.

Interviews and user tests with clinicians were used to establish the optimal interface design. It was identified that the interviewees want a 'first-reader' tool which directly predicts a tumor's boundaries. The expert's role would then be to review the model's output and to edit the predicted contour if necessary. Further, it was identified that the users want a probability map representing the model's confidence in its predictions that consists of colored, semi-transparent

contours. The different colors correspond to the different prediction certainty thresholds which are visualized on a slider that was added to the interface. This slider allows the user to switch between different certainty thresholds with matching contour predictions and to select the one the user deems as most accurate. This adds an interactive element to the model's output which may promote more engagement with the predictions.

Further, selecting multiple thresholds was made possible to allow comparing the segmentations at several certainty levels on one imaging scan. Moreover, buttons for manual edits of the autocontours were included and it was established that the manual edits should always remain differentiable from the autocontours to easily backtrack one's changes, for instance when reviewing segmentations with colleagues. In addition to the interactive probability map, a smaller CT scan was added to the interface containing an all-in-one probability map. This shows all certainty thresholds at once, allowing the user to quickly get an overview of how the prediction changes for a given slice when changing the threshold.

The user evaluation of the interactive prototype built in this project has shown that the participants were optimistic about introducing a semi-automatic tumor segmentation tool to the UMCG. The participants generally thought the designed interface was pleasant and intuitive, and an excellent usability score was achieved. However, it was mentioned that a similar tool should ideally be integrated into the software currently used for manual tumor segmentation at the UMCG, RayStation. This was desired as it would save time needed for switching between softwares and since it might be faster to get accustomed to a few new functionalities over an entire new programme.

There was a general consensus that the probability maps helped in understanding the model's predictions. Participants reported that they preferred being able to view and interact with different certainty thresholds, instead of only viewing a single binary prediction, like most other tumor segmentation tools offer. Hence, the certainty maps designed here were a successful addition to the interface.

Further questionnaire results revealed that the users understood how the tool could aid them in their decision making even if they do not fully understand the model's inner workings. The trust the users reported to have in the system was adequate and they did not give the impression that they were overly reliant on the predictions. This would be important to ensure that the users remain critical of the model's outputs.

11.1 Limitations

Nevertheless, the current findings should not be interpreted without taking the limitations of this research into account. Firstly, our sample was a convenience sample consisting of clinicians who voluntarily signed up for the user test. It is not a random sample of radiation oncologists and radiotherapists, and is thus not necessarily representative for the UMCG as whole. It could be that the participants had a higher interest in AI tools and were more favorable of an automatic segmentation tool compared to their colleagues who did not participate.

Moreover, the evaluation only consisted of self-report measures. The subjective findings described in this paper, such as the potential of the tool to decrease delineation time and the finding that clinicians mentioned they would trust the predictions appropriately, should be examined in a more objective manner in the future. The entire tool should also undergo more extensive user testing with a larger array of sample cases before being implemented in the clinic. This should include difficult cases, for instance ones where the model makes predictions that are incongruent with the beliefs of the user.

Further, there are some shortcomings regarding the scales used in the questionnaire. Several questions were developed for this study and thus do not represent validated questionnaire items. It has not been investigated whether the items accurately reflect the constructs they are measuring. Nevertheless, for this exploratory study these items still provided valuable inputs. Moreover, only the most relevant items were selected from the Human-Computer Trust Questionnaire and from the Trust in Automation Questionnaire in an effort to keep the questionnaire short. This however makes interpreting the results from individual items difficult, and the findings from these questionnaires should be interpreted critically. In the future the entire scale should be used or at least complete subscales.

11.2 Future Research

This project also sparked some suggestions for future research. Apart from testing the automatic contours in a more objective manner, future investigations should be carried out into whether the trust in the predictions is associated with segmentation experience. During the user evaluations it appeared as though participants with less experience in manual segmentation would trust the model's predictions more easily. More experienced clinicians repeatedly make statements such as "I would trust myself over the system", while less experienced individuals for instance said "I can rely on the predictions when I don't know where the tumor is". If a link between experience and trust could be established, this could give pointers for the implementation of such a tool. For instance, novices may be given additional training so they are more critical of the system. This could be beneficial as their lower experience might mean they could be biased more easily by the predictions and potentially identify less errors than more experienced clinicians.

Further, future research should examine what kind of briefing should be given to users of the tool for an optimal understanding of the model. The focus here should especially be on investigating how much knowledge of AI and the workings of deep-learning models clinicians should have. Gerlings et al. (2021) note that insight into this topic is still scarce for AI decision-making tools.

11.3 Suggestions for Implementation

Taking the results of this project together with its limitations, some suggestions for implementing a tool for semi-automatic tumor segmentation in the clinic will now be outlined. Firstly, a slow clinical implementation is suggested so that users can get to know the strengths and weaknesses

of the model in predicting tumor boundaries. Initially the tool could be used as a second-reader, where the user manually delineates the first few patients and only then compares their delineations to the autocontours. This could pinpoint areas that may generally need closer review and allows the user to judge how close the model is to their personal delineations. When accustomed to the tool and confident in its outputs, the user could then proceed to critically using it as a first-reader tool.

Second, it is recommended that the user always form their own general judgment of the tumor boundaries before looking at the model's results. This can be done by scrolling through the slices in the interface before generating the autocontours. Having an idea of where the delineation is believed to be can help to identify regions about which the user is not in agreement with the model and it may reduce being biased by the model's predictions.

Further, clinicians should be properly trained to use this new tool. A short user guide or instructional video should be made that explains the tool's main functionalities. The model's limitations should also be mentioned here, for instance that the model cannot take any patient information into account that may be relevant for the radiotherapy treatment and that the predictions are generated based on the idea that the ground truths used for the training of the model were accurate. Moreover, users should be made aware of certain conditions that could promote their susceptibility to overreliance on automation, such as high task complexity and a high workload (Sujan et al., 2019).

Lastly, Sujan et al. (2019) raised an interesting question of whether upcoming generations who start their training with automatic tools already in place will have the same manual (segmentation) skills as older generations. A solid manual segmentation foundation is needed to assess the accuracy of the predictions and to revert to manual methods in case of potential system failures. Hence, it should be ensured that an automatic tumor segmentation tool is only an aid, not a necessity for this task.

11.4 Conclusion

In conclusion, this study has suggested that a tool for semi-automatic tumor segmentation may be useful for radiation oncologists at the Universitair Medisch Centrum Groningen. User tests with radiation oncologists showed that they were satisfied with the user interface designed in this project and that they were optimistic about collaborating with an artificial intelligence tool in the segmentation of tumors. Furthermore, introducing interactive probability maps representing the model's prediction confidence to the interface offers a promising avenue to increase the user's understanding of the model's decisions. Hence, insights from explainable AI have been successfully used in this project to design a tool for computer-aided diagnosis. If more objective tests confirm that the inter-observer variability of manual segmentation and the time required for this task can be reduced, the workflow of radiation oncologists may be severely improved by introducing such a tool, which could ultimately result in more efficient and reproducible patient treatment.

References

Abdellah, A., & Koucheryavy, A. (2020). Survey on artificial intelligence techniques in 5G networks. *Telecom IT*, *8*(1), 1–10. https://doi.org/10.31854/2307-1303-2020-8-1-1-10

Alexander, A., Jiang, A., Ferreira, C., & Zurkiya, D. (2020). An intelligent future for medical imaging: A market outlook on artificial intelligence for medical imaging. *Journal of the American College of Radiology*, *17*(1), 165–170. https://doi.org/10.1016/J.JACR.2019.07.019

Andrearczyk, V., Oreiller, V., Boughdad, S., Rest, C. C. Le, Elhalawani, H., Jreige, M., Prior, J. O., Vallières, M., Visvikis, D., Hatt, M., & Depeursinge, A. (2022). Overview of the HECKTOR challenge at MICCAI 2021: Automatic head and neck tumor segmentation and outcome prediction in PET/CT images. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13209 LNCS*, 1–37. https://doi.org/10.48550/arxiv.2201.04138

Badrigilan, S., Nabavi, S., Abin, A. A., Rostampour, N., Abedi, I., Shirvani, A., & Ebrahimi Moghaddam, M. (2021). Deep learning approaches for automated classification and segmentation of head and neck cancers and brain tumors in magnetic resonance images: A meta-analysis study. *International Journal of Computer Assisted Radiology and Surgery*, *16*(4), 529–542. https://doi.org/10.1007/s11548-021-02326-z

Baskar, R., Dai, J., Wenlong, N., Yeo, R., & Yeoh, K. W. (2014). Biological response of cancer cells to radiation treatment. *Frontiers in Molecular Biosciences*, *1*(NOV). https://doi.org/10.3389/FMOLB.2014.00024

Begoli, E., Bhattacharya, T., & Kusnezov, D. (2019). The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence 2019 1:1*, *1*(1), 20–23. https://doi.org/10.1038/s42256-018-0004-1

Brooke, J. (1996). *SUS-A quick and dirty usability scale*. Retrieved July 3, 2022, from www.TBIStaffTraining.info

Bureau Oberhauser. (n.d.). *Q bio website*. Behance. Retrieved August 31, 2022, from https://www.behance.net/gallery/136409819/Q-Bio-Website

Cai, C. J., Research, G., Team, B., Samantha Winter, U., Health, G., David Steiner, U., Lauren Wilcox, U., Michael Terry, U., Winter, S., Steiner, D., Wilcox, L., & Terry, M. (2019). "Hello AI": Uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW). https://doi.org/10.1145/3359206

Cox, J. D., Stetz, J. A., & Pajak, T. F. (1995). Toxicity criteria of the radiation therapy oncology group (RTOG) and the European organization for research and treatment of cancer (EORTC). *International Journal of Radiation Oncology, Biology, Physics*, *31*(5), 1341–1346. https://doi.org/10.1016/0360-3016(95)00060-C

Cramer, J. D., Burtness, B., Le, Q. T., & Ferris, R. L. (2019). The changing therapeutic landscape of head and neck cancer. *Nature Reviews Clinical Oncology*, *16*(11), 669–683. https://doi.org/10.1038/s41571-019-0227-z

Cramer, J. D., Johnson, J. T., & Nilsen, M. L. (2018). Pain in head and neck cancer survivors: Prevalence, predictors, and quality-of-life impact. *Otolaryngology - Head and Neck Surgery (United States)*, *159*(5), 853–858. https://doi.org/10.1177/0194599818783964

De Biase, A., Sijtsema, N. M., van Dijk, L., Langendijk, J. A., & van Ooijen, P. (2022). Slice-by-slice deep learning aided oropharyngeal cancer segmentation with adaptive thresholding for spatial uncertainty on FDG PET and CT images. *Radiotherapy and Oncology*, *170*, S1392–S1394. https://doi.org/10.48550/arxiv.2207.01623

Dias, C. R., Pereira, M. R., & Freire, A. P. (2017). Qualitative review of usability problems in health information systems for radiology. *Journal of Biomedical Informatics*, *76*, 19–33. https://doi.org/10.1016/J.JBI.2017.10.004

Doi, K., MacMahon, H., Katsuragawa, S., Nishikawa, R. M., & Jiang, Y. (1999). Computer-aided diagnosis in radiology: Potential and pitfalls. *European Journal of Radiology*, *31*(2), 97–109. https://doi.org/10.1016/S0720-048X(99)00016-9

Endsley, M. R. (2012). Situation Awareness. *Handbook of Human Factors and Ergonomics: Fourth Edition*, 553–568. https://doi.org/10.1002/9781118131350.CH19

Ericsson, K. A., & Simon, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *CULTURE AND ACTIVITY*, *5*(3), 178–186.

Foster, B., Bagci, U., Mansoor, A., Xu, Z., & Mollura, D. J. (2014). A review on segmentation of positron emission tomography images. *Computers in Biology and Medicine*, *50*, 76–96. https://doi.org/10.1016/J.COMPBIOMED.2014.04.014

Fujita, H. (2020). AI-based computer-aided diagnosis (AI-CAD): the latest review to read first. *Radiological Physics and Technology, 13*(1), 6–19. https://doi.org/10.1007/S12194-019-00552-4/FIGURES/11

Gerlings, J., Jensen, M. S., & Shollo, A. (2021). Explainable AI, but explainable to whom?. *arXiv preprint arXiv:2106.05568*.

Gillmann, C., Saur, D., Wischgoll, T., & Scheuermann, G. (2021). Uncertainty-aware Visualization in Medical Imaging - A Survey. *Computer Graphics Forum*, *40*(3), 665–689. https://doi.org/10.1111/CGF.14333

Gordillo, N., Montseny, E., & Sobrevilla, P. (2013). State of the art survey on MRI brain tumor segmentation. *Magnetic Resonance Imaging*, *31*(8), 1426–1438. https://doi.org/10.1016/J.MRI.2013.05.002

Gulum, M. A., Trombley, C. M., & Kantardzic, M. (2021). A review of explainable deep learning cancer detection models in medical imaging. *Applied Sciences 2021, Vol. 11, Page 4573*, *11*(10), 4573. https://doi.org/10.3390/APP11104573

Head and Neck Cancers - NCI. (n.d.). Retrieved May 24, 2022, from https://www.cancer.gov/types/head-and-neck/head-neck-fact-sheet

Hennessy, O. (2015). Lymphoma of pharynx and both orbits. *Radiopaedia.Org*. https://doi.org/10.53347/RID-33516

Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain?. *arXiv preprint arXiv:1712.09923*.

ISO 9241-11, Ergonomic requirements for office work with visual display terminals (VDTs) – Part 11: Guidance on usability, 1998.

Jiang, H., Diao, Z., & Yao, Y.-D. (2021). Deep learning techniques for tumor segmentation: a review. *The Journal of Supercomputing*, *78*, 1807–1851. https://doi.org/10.1007/s11227-021-03901-6

Jorritsma, W., Cnossen, F., & van Ooijen, P. M. A. (2015). Improving the radiologist-CAD interaction: Designing for appropriate trust. *Clinical Radiology*, *70*(2), 115–122. https://doi.org/10.1016/J.CRAD.2014.09.017

Jungo, A., McKinley, R., Meier, R., Knecht, U., Vera, L., Pérez-Beteta, J., Molina-García, D., Pérez-García, V. M., Wiest, R., & Reyes, M. (2018). Towards uncertainty-assisted brain tumor segmentation and survival prediction. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10670 LNCS*, 474–485. https://doi.org/10.1007/978-3-319-75238-9_40/TABLES/3

K., D. (n.d.). *UI/UX for Broadband Internet Provider Webpage*. Behance. Retrieved August 31, 2022, from https://www.behance.net/gallery/29459013/UIUX-for-Broadband-Internet-Provider-Webpage?tracking_source=search_projects%7Ctech+website

Körber, M. (2019). Theoretical considerations and development of a questionnaire to measure trust in automation. *Advances in Intelligent Systems and Computing*, *823*, 13–30. https://doi.org/10.1007/978-3-319-96074-6_2/TABLES/4

Lee, J. D., & Seppelt, B. D. (2012). Human factors and ergonomics in automation design. *Handbook of Human Factors and Ergonomics: Fourth Edition*, 1615–1642. https://doi.org/10.1002/9781118131350.CH59

Lee, J. G., Jun, S., Cho, Y. W., Lee, H., Kim, G. B., Seo, J. B., & Kim, N. (2017). Deep learning in medical imaging: General overview. *Korean Journal of Radiology, 18*(4), 570–584. https://doi.org/10.3348/KJR.2017.18.4.570

Leer, J. W. H. (2005). What the clinician wants to know: radiation oncology perspective. *Cancer Imaging*, *5*, 1–2. https://doi.org/10.1102/1470-7330.2005.0027

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *Https://Doi.Org/10.1080/10447319509526110*, *7*(1), 57–78. https://doi.org/10.1080/10447319509526110

Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. In *11th australasian conference on information systems* (Vol. 53, pp. 6-8). Brisbane, Australia: Australasian Association for Information Systems.

Marur, S., & Forastiere, A. A. (2008). Head and neck cancer: Changing epidemiology, diagnosis, and treatment. *Mayo Clinic Proceedings*, *83*(4), 489–501. https://doi.org/10.4065/83.4.489

Men, K., Dai, J., & Li, Y. (2017). Automatic segmentation of the clinical target volume and organs at risk in the planning CT for rectal cancer using deep dilated convolutional neural networks. *Medical Physics*, *44*(12), 6377–6389. https://doi.org/10.1002/MP.12602

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/J.ARTINT.2018.07.007

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Morgan, H. E., & Sher, D. J. (2020). Adaptive radiotherapy for head and neck cancer. *Cancers of the head & neck*, *5*(1), 1-16. https://doi.org/10.1186/s41199-019-0046-z

Mosier, K. L., Skitka, L. J., Heers, S., & Burdick, M. (1998). Automation bias: Decision making and performance in high-tech cockpits, *8*(1), 47–63. https://doi.org/10.1207/S15327108IJAP0801_3

Muhammad, K., Khan, S., Member, S., Del Ser, J., Member, S., Hugo de Albuquerque, V. C., & Muhammad Khan Muhammad, K. (2021). Deep learning for multigrade brain tumor classification in smart healthcare systems: A Prospective Survey. *IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS*, *32*(2), 507. https://doi.org/10.1109/TNNLS.2020.2995800

Natekar, P., Kori, A., & Krishnamurthi, G. (2020). Demystifying brain tumor segmentation networks: Interpretability and uncertainty analysis. *Frontiers in Computational Neuroscience*, *14*, 6. https://doi.org/10.3389/FNCOM.2020.00006/BIBTEX

Nishikawa, R. M., & Bae, K. T. (2018.). Importance of better human-computer interaction in the era of deep learning: Mammography computer-aided diagnosis as a use case. *Journal of the American College of Radiology*, *15*(1), 49-52.

Njeh, C. F. (2008). Tumor delineation: The weakest link in the search for accuracy in radiotherapy. *Journal of Medical Physics / Association of Medical Physicists of India*, *33*(4), 136. https://doi.org/10.4103/0971-6203.44472

Oreiller, V., Andrearczyk, V., Jreige, M., Boughdad, S., Elhalawani, H., Castelli, J., Vallières, M., Zhu, S., Xie, J., Peng, Y., Iantsen, A., Hatt, M., Yuan, Y., Ma, J., Yang, X., Rao, C., Pai, S., Ghimire, K., Feng, X., … Depeursinge, A. (2022). Head and neck tumor segmentation in PET/CT: The HECKTOR challenge. *Medical Image Analysis*, *77*. https://doi.org/10.1016/J.MEDIA.2021.102336

Patrício, C., Neves, J. C., & Teixeira, L. F. (2022). Explainable deep learning methods in medical diagnosis: A Survey. *arXiv preprint arXiv:2205.04766*.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human factors*, *39*(2), 230-253.

Reyes, M., Meier, R., Pereira, S., Silva, C. A., Dahlweid, F. M., Tengg-Kobligk, H. V., ... & Wiest, R. (2020). On the interpretability of artificial intelligence in radiology: challenges and opportunities. *Radiology: artificial intelligence*, *2*(3), https://doi.org/10.1148/ryai.2020190043

Rogers, Y., Sharp, H., & Preece, J. (2013). Interaction design: Beyond human-computer interaction. *Wiley.*

Sadeghi, S., Siavashpour, Z., Sadr, A. V., Farzin, M., Sharp, R., & Gholami, S. (2021). A rapid review of influential factors and appraised solutions on organ delineation uncertainties reduction in radiotherapy. *Biomedical Physics & Engineering Express*, *7*(5), 052001. https://doi.org/10.1088/2057-1976/AC14D0

Sauro, J. (2011). *Measuring usability with the system usability scale (SUS)*. MeasuringU. Retrieved August 26, 2022, from https://measuringu.com/sus/

Savenije, M. H. F., Maspero, M., Sikkes, G. G., Van Der Voort Van Zyp, J. R. N., Alexis, A. N., Bol, G. H., & Cornelis, C. A. (2020). Clinical implementation of MRI-based organs-at-risk auto-segmentation with convolutional networks for prostate radiotherapy.

Schantz, S. P., & Yu, G. P. (2002). Head and neck cancer incidence trends in young Americans, 1973-1997, With a Special Analysis for Tongue Cancer. *Archives of Otolaryngology–Head & Neck Surgery*, *128*(3), 268–274. https://doi.org/10.1001/ARCHOTOL.128.3.268

Schutte, H. W., Heutink, F., Wellenstein, D. J., van den Broek, G. B., van den Hoogen, F. J., Marres, H. A., van Herpen, C. M., Kaanders, J. H., Merkx, T. M., & Takes, R. P. (2020). Impact of time to diagnosis and treatment in head and neck cancer: A systematic review. *Otolaryngology–Head and Neck Surgery*, *162*(4), 446–457. https://doi.org/10.1177/0194599820906387

Semple, C. J., Dunwoody, L., George Kernohan, W., McCaughan, E., & Sullivan, K. (2008). Changes and challenges to patients' lifestyle patterns following treatment for head and neck cancer. *Journal of Advanced Nursing*, *63*(1), 85–93. https://doi.org/10.1111/j.1365-2648.2008.04698.x

SergeyBitos. (n.d.). *Modern medical examination in the style of HUD. ultrasound and Cardiogram. A futuristic medical interface, a virtual body scanning interface with heart, human body and electrocardiogram illustrations. stock vector*. Adobe Stock. Retrieved August 31, 2022, from https://stock.adobe.com/191779036?as_channel=adobe_com&as_campclass=brand&as_campaign=srp-raill&as_source=behance_net&as_camptype=acquisition&as_audience=users&as_content=thumbnail-click&promoid=J7XBWPPS&mv=other

Sorantin, E., Grasser, M. G., Hemmelmayr, A., Tschauner, S., Hrzic, F., Weiss, V., ... & Holzinger, A. (2021). The augmented radiologist: Artificial intelligence in the practice of radiology. *Pediatric Radiology*, 1-13.

Strohm, L., Hehakaya, C., Ranschaert, E. R., Boon, W. P., & Moors, E. H. (2020). Implementation of artificial intelligence (AI) applications in radiology: hindering and facilitating factors. *European radiology*, *30*(10), 5525-5532.

Sujan, M., Furniss, D., Grundy, K., Grundy, H., Nelson, D., Elliott, M., White, S., Habli, I., & Reynolds, N. (2019). Human factors challenges for the safe use of artificial intelligence

in patient care. *BMJ Health & Care Informatics*, *26*(1), 100081. https://doi.org/10.1136/BMJHCI-2019-100081

Suzuki, K. (2017). Overview of deep learning in medical imaging. *Radiological Physics and Technology*, *10*. https://doi.org/10.1007/s12194-017-0406-5

Tang, X., Wang, B., & Rong, Y. (2018). Artificial intelligence will reduce the need for clinical medical physicists. *Journal of Applied Clinical Medical Physics*, *19*(1), 6. https://doi.org/10.1002/ACM2.12244

Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (xai): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(11), 4793–4813. https://doi.org/10.1109/TNNLS.2020.3027314

Trotti, A. (1997). Toxicity antagonists in cancer therapy. *Current opinion in oncology*, *9*(6), 569-578.

Trotti, A. (2000). Toxicity in head and neck cancer: a review of trends and issues. *International Journal of Radiation Oncology*Biology*Physics*, *47*(1), 1–12. https://doi.org/10.1016/S0360-3016(99)00558-1

van den Brekel, M. W. M., & Castelijns, J. A. (2005). What the clinician wants to know: surgical perspective and ultrasound for lymph node imaging of the neck. *Cancer Imaging : The Official Publication of the International Cancer Imaging Society*, *5 Spec No A*. https://doi.org/10.1102/1470-7330.2005.0028

van der Velden, B. H. M., Kuijf, H. J., Gilhuijs, K. G. A., & Viergever, M. A. (2022). Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Analysis*, *79*, 102470. https://doi.org/10.1016/J.MEDIA.2022.102470

van Dijk, L. V., Van den Bosch, L., Aljabar, P., Peressutti, D., Both, S., Steenbakkers Roel, J. H. M., Langendijk, J. A., Gooding, M. J., & Brouwer, C. L. (2020). Improving automatic delineation for head and neck organs at risk by Deep Learning Contouring. *Radiotherapy and Oncology*, *142*, 115–123. https://doi.org/10.1016/J.RADONC.2019.09.022

Waqar, M., Nawaz Abro, M., Soomro, Q., Shahban, M., & Khatoon, S. (2019). Retrospective incidence analysis of head and neck cancer patients in rural areas of Sindh, Pakistan. *Jundishapur Journal of Chronic Disease Care*, *8*(4). https://doi.org/10.5812/JJCDC.95530

Wong, K.-P. (2005). Medical image segmentation: Methods and applications in functional imaging. *Handbook of Biomedical Image Analysis*, 111–182. https://doi.org/10.1007/0-306-48606-7_3

Wong, J., Fong, A., McVicar, N., Smith, S., Giambattista, J., Wells, D., Kolbeck, C., Giambattista, J., Gondara, L., & Alexander, A. (2020). Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning. *Radiotherapy and Oncology*, *144*, 152–158. https://doi.org/10.1016/J.RADONC.2019.10.019

Xie, J., & Peng, Y. (2022). The head and neck tumor segmentation based on 3D U-Net. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13209 LNCS*, 92–98. https://doi.org/10.1007/978-3-030-98253-9_8/TABLES/4

Yang, G., Ye, Q., & Xia, J. (2022). Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Information Fusion*, *77*, 29–52. https://doi.org/10.1016/J.INFFUS.2021.07.016

ZinetroN. (n.d.). *Futuristic user interface and infographic elements . medical infographic (lungs, liver and DNA human) body scanning (Sci Fi, UI, Hud elements). Modern Medical Examination hud style. human body scan. stock vector*. Adobe Stock. Retrieved August 31, 2022, from https://stock.adobe.com/272220214?as_channel=adobe_com&as_campclass=brand&as_campaign=srp-raill&as_source=behance_net&as_camptype=acquisition&as_audience=users&as_content=thumbnail-click&promoid=J7XBWPPS&mv=other

Appendix A - Interview Questions

Main questions asked to radiation oncologists during the semi-structured interviews to gather requirements.

1. Which software do you currently use for manual tumor segmentation?
2. Which modalities do you use (CT, PET, MRI, Xray)?
3. Which planes do you use (axial, sagittal, coronal)?
4. How many slices do you delineate?
5. Do you consider patient information while delineating? If yes, which?
6. How frequently do you delineate a single patient?
7. How is radiotherapy applied?
8. Which functionalities and tools of the segmentation software do you use?
9. Are there any problems, inefficiencies, or missing functionalities with the current software?
10. Do you use certain strategies/guidelines to identify tumors?
11. How long does the delineation take on average?
12. What causes uncertainty/doubt in your delineations? How do you solve this?
13. Do you discuss your delineation with other colleagues?

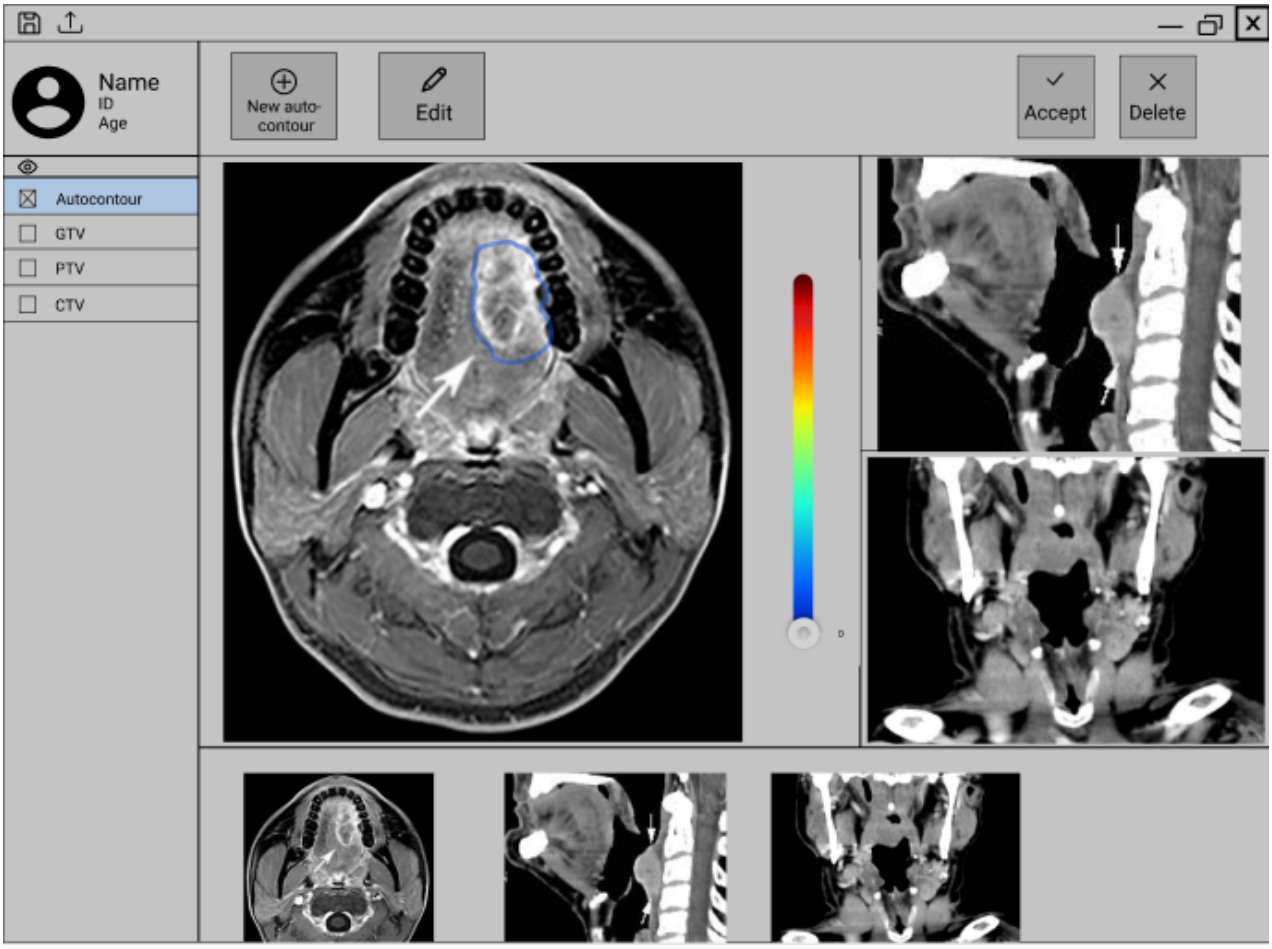
Appendix B - Requirements Table

Table 1. General Requirements for Automatic Segmentation Interface based on Hierarchical Task Analysis of Manual Segmentation

Requirement	**Description**
Standard software tools: open and close window, save	E,g, To open and save patient files
Case information	Include basic patient information (name, ID), and information on imaging scans (which modality, time stamps)
Zoom	To increase/decrease the size of the viewed area
Hand tool	To move to a different part of the scan.
Scrolling through slices	With mouse scroll wheel.
On/off button for segmentation	To view imaging scan with and without the segmentation
Brush tool	For segmenting the tumor. Should be able to change the size of the brush (bigger brush for larger areas and time saving, smaller brush for details), change the brush color and the brush should automatically fill in the outline when a closed shape is drawn.
Eraser tool	To edit or erase segmentations
Setting the different imaging modalities as primary, secondary or tertiary view	Gallery allows switching between CT, PET and MRI
Allow viewing multiple imaging scans at once	Using scans of different sizes
Interpolation button	To save time in delineations
Buttons to view/hide regions of interest and organs at risk	/
Button to save segmentation	With templates to save GTV, PTV, CTV, and ITV
Button to delete segmentation	/

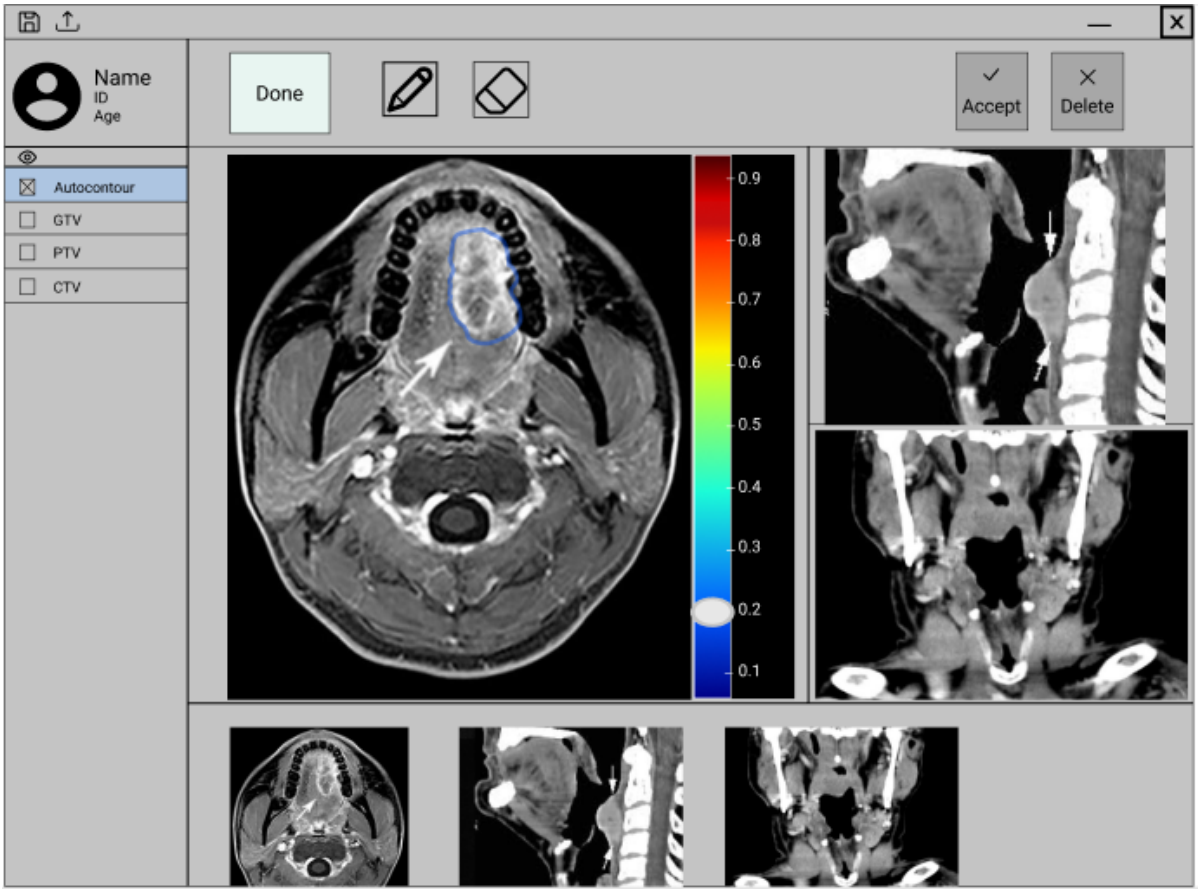
Appendix C - Prototype Design (Large Formats)

Larger versions of the prototypes:



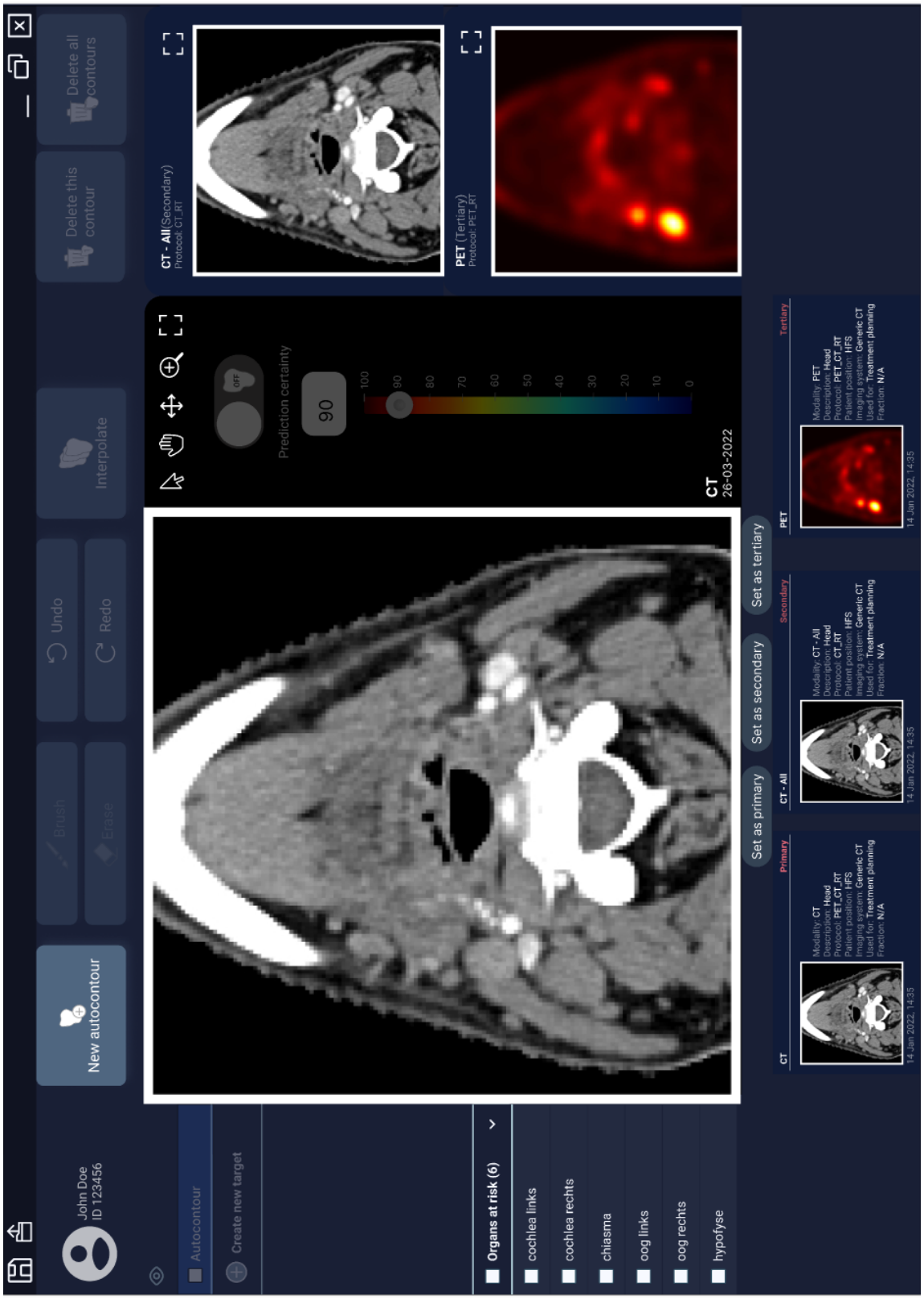
Prototype Version 1 of Homepage

Note. Secondary and tertiary view figures from van den Brekel & Castelijns (2005) and Hennessy (2015).

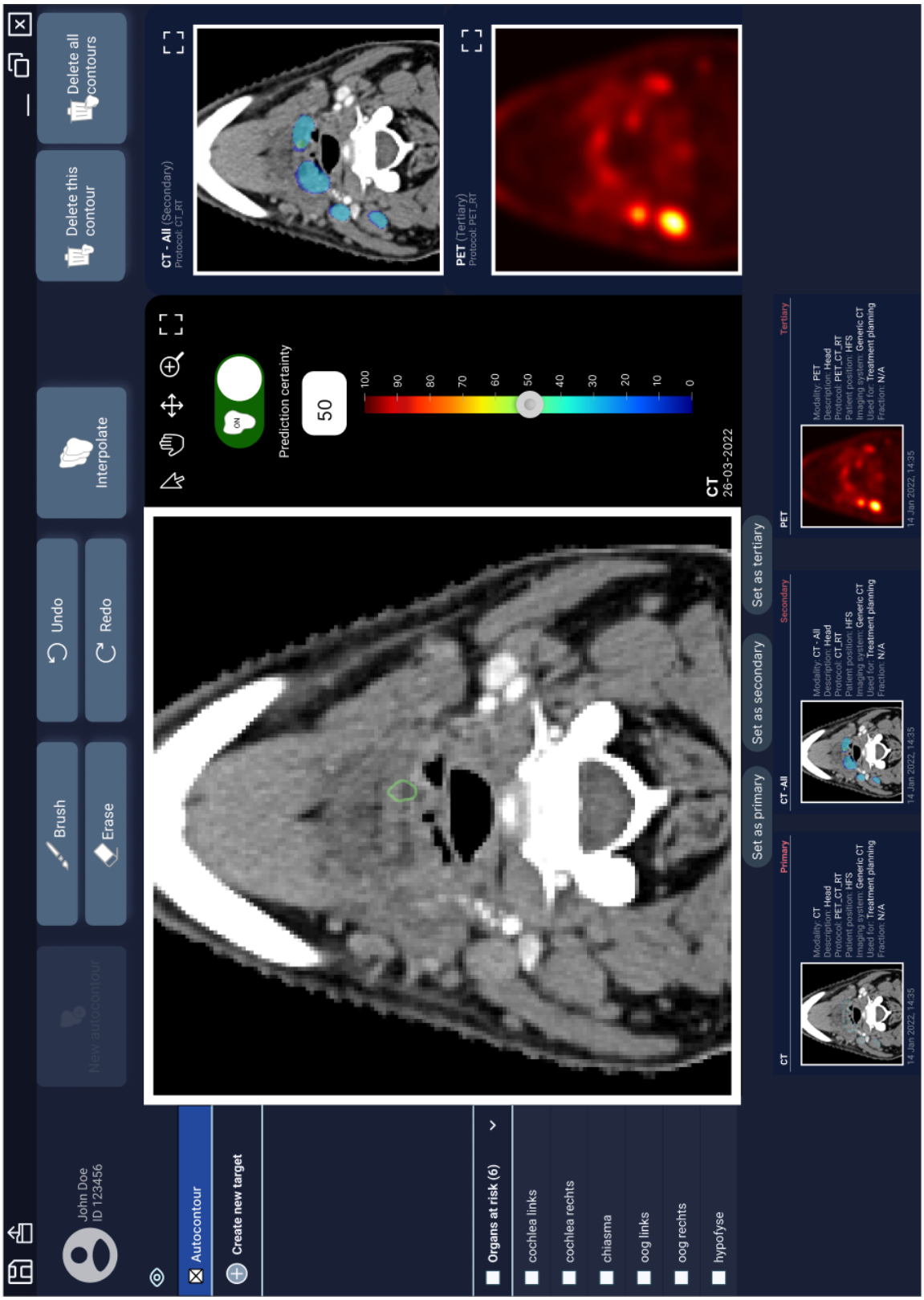


Prototype Version 1 of the Edit Page

Note. Secondary and tertiary view figures from van den Brekel & Castelijns (2005) and Hennessy (2015).



Final Prototype of the Homepage



Final Prototype of the Edit Page

Appendix D - Information Form containing Study Details

INFORMATION ABOUT THE RESEARCH

Version for participants

"Designing a User Interface for Automatic Tumor Segmentation"

· **Why do I receive this information?**

You recently indicated that you are willing to take part in this study investigating the best way to design an automatic tumor segmentation system. This document provides further information on this study and informs you of your tasks and rights.

· **Contact information**

The following researchers are involved in the study:

Liv Ziegfeld Alessia de Biase
MSc Student Computational Cognitive Science PhD Candidate
(RUG) Department of Radiotherapy (UMCG) Tel: 0616825078 E-mail:
a.de.biase@umcg.nl E-mail: l.u.ziegfeld@student.rug,nl

· **Do I have to participate in this research?**

Participation in the research is voluntary. However, your consent is needed. Therefore, please read this information carefully. Ask all the questions you might have, for example because you do not understand something. Only afterwards you decide if you want to participate. If you decide not to participate, you do not need to explain why, and there will be no negative consequences for you. You have this right at all times, including after you have consented to participate in the research.

· **Why this research?**

This study aims to investigate if and how an automatic tumor segmentation system can be useful and feasible at the UMCG. We want to examine whether automatic suggestions for tumor contours can speed up, facilitate and make manual contouring of tumors more accurate. This user test is mainly aimed at identifying your preferences for the design and functionalities of the interface for the automatic segmentation tool. Further, it will be investigated how to best represent the uncertainties of the automatic segmentations.

· **What do we ask of you during the research?**

After reading this form and providing informed consent, you will be asked to complete the following tasks during a user test study at UMCG:

- Interacting with a prototype of an automatic tumor segmentation system and completing tasks with it
- Answering questions on your preferences for the system
- Responding to short surveys on your experience with the system and elaborating on the responses if asked

· Having your voice audio-recorded. These recordings will only be shared with the research team and will allow us to take all your suggestions into account during the data evaluation.
· Having your screen recorded during the interaction with the interface. Mouse tracking will allow us to make the interface more intuitive by examining where the user first looks for certain functionalities. This data will only be shared with the research team.

· **What are the consequences of participation?**
  o There are no anticipated risks of participating in this study.
  o The study will help us understand how we can introduce a similar system most effectively in the UMCG. Taking your wishes and requirements into consideration will enable designing a system that is of maximum utility in assisting you in your delineations.

· **How will we treat your data?**
  o Before the study, contact data are collected to get in touch with you and for scheduling an appointment for the experiment
  o If you provide consent, personal contact data consisting of only your name, e mail address and speciality will be stored and retained in a file that is fully separated from the research data. These personal data will be used to send you a summary of the results if you would like that.
  o The answers that you provide on the questionnaires and during the interviews, as well as any remarks/suggestions you make about the system, will be collected as research data. The data will be pseudonymized via a participant number and will not be linkable to any personal data that you may provide to receive the results.
  o Your personal data will be stored until the study is completed. This will be September 31st, 2022, at the latest.
  o Pseudonymized research data will be retained for 10 years.
  o As a participant, you have the right to access, rectify, and erase your personal data until the data are made anonymous. You can contact one of the researchers up until the end of the study (September 31st, 2022) to ask for a copy of your personal data, have erroneous personal data corrected, and/or have your personal data withdrawn.

What else do you need to know?
You may always ask questions about the research: before, during the research, and after the end of the research. You can do so by speaking with one of the researchers present during the user testing or by emailing or phoning one of the researchers involved.

Do you have questions/concerns about your right as a research participant or about the conduct of the research? You may also contact the Ethics Committee of the Faculty of Behavioural and Social Sciences of the University of Groningen: ec-bss@rug.nl.

As a research participant, you have the right to a copy of this research information.

Appendix E - Information Sheet

Goal:
We are currently developing a system for automatic tumor delineation and would like to hear your thoughts and ideas in order to make this system as useful in the clinic as possible. The goal of the system is to save time in the delineation process. Further, the tool should increase the accuracy and reproducibility of manual contouring of tumors in the head and neck region.
How the system works:
The system will display predictions of the tumor contour on CT scans, which can then be used as a basis for the delineation. The system is therefore intended as the first step in the delineation process, where the programme can be used to extract contour predictions, which can then be edited manually if necessary. Thus, the user will still be required to check these predicted outputs and ensure their accuracy. It is expected that the editing time of the predicted contours is less than the time required to delineate a patient's tumor from scratch.

The system will predict tumor delineations based on registered CT and PET scans using a Deep Learning algorithm, which is an artificial intelligence model that is inspired by human neural networks. The model is trained and tested on bounding boxes of fixed size (144x144x144) extracted from the oropharyngeal cavity of 302 oropharyngeal cancer patients. The available scans were previously delineated by radiation oncologists of our institute between 2014 and 2022. The GTVp represents the "ground-truth" and the model learns to generate automatic contours by identifying features that are most important in leading to the ground truths. When the model receives new, unseen scans, it uses the rules it has established to make its best predictions on the contour of the new tumor.
Displaying the model's uncertainty:
Although this model is optimized to produce precise delineations, some inaccuracies are still possible in its output. These may for instance arise because of poor image quality, or because the model cannot take patient information into account. Hence, we developed a prototype of an interface that displays the uncertainty derived from the variability in contouring in the training set. It allows viewing different levels of prediction certainty and choosing the certainty level that best matches the user's judgment of the tumor boundaries. Selecting for instance a certainty threshold of 70% means that all pixels/voxels are selected for which the model is 70% certain or more that they are part of the tumor.

You will now get the chance to try out the prototype of this interface. This is only a simulation of an automatic tumor segmentation interface and hence not all functionalities work fully. We ask you to focus on the bigger picture to determine what you like and dislike about the interface.

The first view will be set on the first slice available for this user testing. To scroll through the slices, please use the up/down arrows instead of the scrolling wheel. Feel free to ask us questions related to the system at any point.

Appendix F- Informed Consent Form

“DESIGNING A USER INTERFACE FOR AUTOMATIC TUMOR SEGMENTATION”

- I have read the information about the research. I have had enough opportunity to ask questions about it.
- I understand what the research is about, what is being asked of me, which consequences participation can have, how my data will be handled, and what my rights as a participant are.
- I understand that participation in the research is voluntary. I myself choose to participate. I can stop participating at any moment. If I stop, I do not need to explain why. Stopping will have no negative consequences for me.
- Below I indicate what I am consenting to.

Consent to participate in the research:
[ ] Yes, I consent to participate; this consent is valid until 31-09-2022
[ ] No, I do not consent to participate

Consent to processing my personal data:
[ ] Yes, I consent to the processing of my personal data as mentioned in the research information. I know that until 31-09-2022 I can ask to have my data withdrawn and erased. I can also ask for this if I decide to stop participating in the research.
[ ] No, I do not consent to the processing of my personal data.

Participant’s full name:	Participant’s signature:	Date:

Full name of researcher present:	Researcher’s signature:	Date:

The researcher declares that the participant has received extensive information about the research. *You have*

the right to a copy of this consent form.

Appendix G - User Evaluation Protocol + Questionnaire

Demographics (verbal) - before using interface

1. What is your job title?
2. What is your specialization?
3. How many years of experience do you have in radiology/oncology?
4. Do you work with RayStation in your everyday tasks?
    a. If not, do you work with a different delineation software? Please name:
5. Have you ever worked with automatic segmentation tools?

First thoughts (verbal) - immediately after using interface

1. How do you think that went?
2. What did you like about the interface?
3. What did you dislike about the interface?
4. Are you missing any functionalities from the interface? If yes, which?
5. What do you think the probability map represents?
6. What do you think about the probability map?
7. What do you think about the layout?
8. What do you think of the idea of a (semi) automatic segmentation tool?
9. Other remarks?

Hand out the following scales on paper and ask them to elaborate on their answers/think aloud:

Please indicate to what extent you agree with the following statements:

System Usability Scale (SUS) (Brooke, 1996) (*The scale names were removed from the participant handouts):*

	Strongly Disagree				Strongly Agree
I think that I would like to use a system like this frequently	1	2	3	4	5
I found the system unnecessarily complex.	1	2	3	4	5
I thought the system was easy to use.	1	2	3	4	5
I think that I would need the support of a technical person to be able to use this system.	1	2	3	4	5
I found the various functions in this system were well integrated.	1	2	3	4	5
I thought there was too much inconsistency in this system.	1	2	3	4	5
I would imagine that most people would learn to use this system very quickly.	1	2	3	4	5
I found the system very awkward to use.	1	2	3	4	5
I felt very confident using the system.	1	2	3	4	5
I needed to learn a lot of things before I could get going with this system.	1	2	3	4	5

Please rate the following statements while taking the scale labels into account:

The "Interface Quality" subscale from the Computer System Usability Questionnaire (Lewis, 1995) (The "NA" answer option was removed since all questions should be applicable):

	Strongly Agree						Strongly Disagree
The interface of this system is pleasant.	1	2	3	4	5	6	7
I like using the interface of this system.	1	2	3	4	5	6	7

The system has all the functions and capabilities I expect it to have.	1	2	3	4	5	6	7
Overall, I am satisfied with this system.	1	2	3	4	5	6	7

Please indicate to what extent you agree with the following statements:

Own questions on automatic segmentation:

	Strongly Agree						Strongly Disagree
Using automatic segmentations as a starting point would make me more confident in my delineations	1	2	3	4	5	6	7
I feel like the probability map helps me understand the model's predictions better	1	2	3	4	5	6	7
I prefer seeing the uncertainties of the predicted outputs over seeing a single predicted contour	1	2	3	4	5	6	7
The different probabilities confuse me	1	2	3	4	5	6	7
I think using this system as a basis for my delineations could save time	1	2	3	4	5	6	7
I think it is feasible to use a system like this in the clinic	1	2	3	4	5	6	7
I prefer my usual method (manual segmentation) over using a tool like this	1	2	3	4	5	6	7

Human-Computer Trust (HCT) scale (Madsen & Gregor, 2000) (Only selected the most relevant items. The items are from the subscales of Perceived Technical Competence, Perceived Understandability, and Faith)

	Strongly Agree						Strongly Disagree
The system makes use of all the knowledge and information available to it to produce its solution to the problem	1	2	3	4	5	6	7
I understand how the system will assist me with decisions I have to make.	1	2	3	4	5	6	7
Although I may not know exactly how the system works, I know how to use it to make decisions about the problem.	1	2	3	4	5	6	7
It is easy to follow what the system does.	1	2	3	4	5	6	7
When the system gives unusual advice I am confident that the advice is correct.	1	2	3	4	5	6	7
I like using the system for decision making.	1	2	3	4	5	6	7

Please indicate to what extent you agree with the following statements:

Trust in automation Questionnaire (Körber, 2019) (Only selected most relevant items)

	Strongly disagree	Rather disagree	Neither disagree nor agree	Rather agree	Strongly agree	No response
The system state was always clear to me.	1	2	3	4	5	
I have already used similar systems.	1	2	3	4	5	
One should be careful with unfamiliar automated systems.	1	2	3	4	5	
The system works reliably.	1	2	3	4	5	
The system reacts unpredictably.	1	2	3	4	5	

I trust the system.	1	2	3	4	5	
A system malfunction is likely.	1	2	3	4	5	
I was able to understand why things happened.	1	2	3	4	5	
I rather trust a system than I mistrust it.	1	2	3	4	5	
I can rely on the system.	1	2	3	4	5	
The system might make sporadic errors.	1	2	3	4	5	
It is difficult to identify what the system will do next.	1	2	3	4	5	
Automated systems generally work well.	1	2	3	4	5	
I am confident about the system's capabilities.	1	2	3	4	5	

As the last step of this user testing, we will now present you with a few different options for the layout and the functionalities of the interface. Please tell us which options you prefer and why.