The influence of automation on tumor contouring

Aselmaa, Anet; van Herk, Marcel; Song, Yu; Goossens, Richard H.M.; Laprie, Anne

DOI
10.1007/s10111-017-0436-0

Publication date
2017

Document Version
Final published version

Published in
Cognition, Technology and Work

Citation (APA)
Aselmaa, A., van Herk, M., Song, Y., Goossens, R. H. M., & Laprie, A. (2017). The influence of automation on tumor contouring. *Cognition, Technology and Work*, *19*(4), 795-808. https://doi.org/10.1007/s10111-017-0436-0

Cogn Tech Work
DOI 10.1007/s10111-017-0436-0

ORIGINAL ARTICLE

The influence of automation on tumor contouring

Anet Aselmaa[1] · Marcel van Herk[2] · Yu Song[1] · Richard H. M. Goossens[1] · Anne Laprie[3]

Received: 4 June 2017 / Accepted: 31 August 2017

Abstract Fully or semi-automatic contouring tools are increasingly being used in the tumor contouring task for radiotherapy. While the fully automatic contouring tools have not reached sufficient efficiency, the semi-automatic contouring tools balance more effectively between the human interaction and automation. This study evaluates the influences of a semi-automation contouring tool, called between-slice interpolation, on the resulting contours and the contouring process. The tumor contouring study was conducted on three patient cases with five physicians in a naturalistic setting. The contouring task consisted of initiating the 2D contour manually or with the interpolation tool and correcting that initial contour. The similarity of the resulting contours was pairwise measured within the manual or the interpolated category. Interactions with the software were recorded, and variations in the contouring workflows steps were compared. Results indicated that using the between-slice interpolation tool for creating the initial contour, instead of initiating it manually, influenced both the contouring process and outcomes. First, it was identified that contours initiated by the interpolation tool showed an increased similarity among themselves compared to the manually initiated contours. At the same time, influences to the resulting contours were below clinical relevance, and toward the desired direction—improved consistency of contours. Second, when interpolation was used, in two cases out of three, the average contouring time also decreased significantly. Therefore, the use of such an automation tool can be encouraged.

Keywords Tumor · Contouring · Automatic contouring · Interpolation · Automation bias · Radiotherapy

1 Introduction

1.1 Background

Radiotherapy is one of the most effective methods for the treatment of cancer (Njeh 2008) with an estimate of 52% of cancer patients benefitting from it (Delaney et al. 2005). With the aging population, cancer incidence and mortality are expected to increase (Yancik and Ries 2004). There is an increasing need to optimize the radiotherapy workflows as well as to automate different (parts of) tasks in order to improve the efficiency of radiotherapy and to reduce the workload of the physicians (e.g., Olsen et al. 2014; Kirrmann et al. 2015; Winkel et al. 2016).

One of the tasks within radiotherapy planning where automation has been increasingly introduced is the

✉ Yu Song
y.song@tudelft.nl

Anet Aselmaa
a.aselmaa@tudelft.nl

Marcel van Herk
marcel.vanherk@manchester.ac.uk

Richard H. M. Goossens
r.h.m.Goossens@tudelft.nl

Anne Laprie
laprie.anne@iuct-oncopole.fr

[1] Faculty of Industrial Design Engineering, Delft University of Technology, Landbergstraat 15, 2628 CE Delft, The Netherlands

[2] Division of Cancer Sciences, Manchester Academic Health Sciences, University of Manchester, 46 Grafton Street, Manchester M13 9NT, UK

[3] Département de Radiothérapie, Institut Claudius-Regaud, Institut Universitaire du Cancer de Toulouse-Oncopole, 1, avenue Irène-Joliot-Curie, 31059 Toulouse, France

Published online: 11 September 2017

contouring task. During this task, the tumor and the surrounding organs at risk are contoured on the medical images by a physician (Vieira et al. 2016). Manual contouring can be tedious and time-consuming (Dowsett et al. 1992; Vorwerk et al. 2014), and introducing automatic contouring tools (i.e., contouring with the support of automation) is generally expected to decrease the task duration (Lim and Leech 2016). However, automation may influence physicians' decision-making process, i.e., introduce bias. Automation bias is the phenomenon that appears when the automatically generated decision aids are used as a replacement for a more vigilant system monitoring or decision-making (Skitka et al. 1999). This may result in decisions that are strongly guided by those automatically generated advices (Parasuraman and Manzey 2010). Regarding tumor contouring, automation bias may result in the errors of omission and the errors of commission (Skitka et al. 1999). Here, the errors of omission indicate that the automatic contour did not include all the relevant regions, but was still accepted by the physician. The errors of commission, at the same time, would mean that an automatically suggested and accepted tumor contour included also healthy tissue. These errors could lead to missing tumorous tissue during radiation, or irradiating healthy tissue unnecessarily. Therefore, automation bias should be taken into account when introducing automatic contouring tools to radiotherapy planning software (Wesley and Dau 2017).

Using a relatively basic automatic contouring tool, named *between-slice interpolation*, this paper aims at studying the influences of an automatically generated initial contour(s) on the resulting contours and the contouring process. Three aspects will be explored regarding this topic: (1) The variations among the contours created by physicians with and without the automation tool, as previous research indicated that variability among physicians in manual contouring is a large concern in radiotherapy (van Herk 2004; Fitton et al. 2011); (2) The duration of the contouring task, i.e., the efficiency; and (3) Changes in the contouring workflow introduced by using the automation tool.

The remainder of the paper is arranged as follows: Sect. 1.2 gives an overview of the workflow in radiotherapy with the emphasis on the contouring task. In Sect. 2, the materials and methods of the proposed research are described. The results regarding the influence of automation on the contours and the contouring process are given in Sect. 3. Finally, the findings are discussed in Sect. 4 and conclusions are given in Sect. 5.

1.2 Contouring in radiotherapy planning

The planning of radiotherapy involves a number of clinicians and tasks (Vieira et al. 2016). Once radiotherapy is suggested based on the diagnosis and is discussed with the patient, the necessary data for the treatment planning, such as medical image datasets of different modalities (Batumalai et al. 2016), are acquired. Those images may consist of Computed Tomography (CT) images, various sequences of Magnetic Resonance Imaging (MRI) images, and/or Positron Emission Tomography (PET) images, depending on the type of the tumor (Batumalai et al. 2016). All acquired medical image datasets are then co-registered, i.e., aligned to the same coordinate space for inclusion in the planning process (Weersink 2016).

The image co-registration step is then followed by the contouring task, during which the various treatment volumes, i.e., the tumor, as well as the surrounding healthy tissues are contoured by a physician(s) (Vieira et al. 2016). One of the axioms of radiotherapy is to maximize the prescribed radiation dose to the tumor while sparing surrounding organs at risk (Burnet 2004). For this, accurately identifying the location and the shape of the tumor is a prerequisite. This is especially true, as with the technological advancements in image-guided radiotherapy, it is possible to precisely deliver the radiation to complicated 3D volumes (Nutting et al. 2000; Xing et al. 2006).

Different types of volumes are used for the treatment planning as recommended by the International Commission on Radiation Units and Measurements (ICRU) in report 62 (ICRU 1999). The Gross Tumor Volume (GTV), which represents the visible (on medical image datasets) and/or palpable tumor, is the basis of other relevant tumor volumes, such as the Clinical Target Volume (CTV) (Burnet 2004). During the contouring process of the GTV, medical image datasets are presented on the computer screen as 2D images, each of them representing a section of the human body (i.e., "slice"). The physician then draws the visually seen borders of the tumor on a number of slices (Dowsett et al. 1992), resulting in a set of 2D contours representing the 3D volume of the GTV. Once all the relevant contours are created, different 3D volumes are constructed, e.g., by lofting those contours. Radiation dose is then planned and validated (e.g., Winkel et al. 2016) based on the dose constraints on these volumes. Among different contouring tasks, the GTV contouring task is especially important in radiotherapy planning since GTV is the basis for defining other volumes for the treatment planning and consequently, uncertainties in this step may introduce a systematic error for the complete treatment planning (van Herk 2004).

The GTV contouring task can be divided into three phases: familiarization, contouring (i.e., action), and evaluation (Aselmaa et al. 2017). Prior to creating any contour, the physician usually spends some amount of time exploring the information presented on the graphical user interface getting familiar with the data. The contouring action itself can be further divided into creating an initial contour(s) and correcting this contour(s), either immediately or later. Then, the contour(s) is iteratively evaluated and modified as needed throughout the contouring process. For example, a physician may first create the initial contours for a few neighboring slices and then continue with iteratively modifying these contours or creating contours on further slices.

Contouring without any computational support can be lengthy and tedious as it requires drawing the visually seen borders of the tumor on all intersecting slices (Dowsett et al. 1992; Vorwerk et al. 2014). In the past decades, extensive research has been conducted and various automatic contouring tools (i.e., segmentation methods) have been developed (Olabarriaga and Smeulders 2001). Some of these tools have been gradually introduced in commercial radiotherapy planning software solutions (Sykes 2014). The expected benefits of using automatic tools are the reduction of the overall amount of time taken to draw the contours, and potentially also increased reproducibility of the contours (i.e., reduced inter-observer variation).

Contouring tools can be categorized as fully automatic, semi-automatic or fully manual (Heckel et al. 2013; Ramkumar et al. 2016) based on the intended level of involvement of the physician and computation. Fully automatic contouring is potentially the most time efficient as it is designed to have little involvement of the physician. However, fully automatic contouring methods have shown limited success and often extensive post-processing is needed (Bauer et al. 2013; Sykes 2014). Automation may be introduced in different levels for semi-automatic methods: from automatically generated 3D volumes based on a few 2D contours [e.g., based on the foreground and background seeds (Dolz et al. 2016)] to computationally adjusting the contour while it is being drawn by the physician [e.g., live-wire tool (Barrett and Mortensen 1997)]. And different levels of automation may pose different influences (Bravo and Ostos 2017) on the physician's decision-making process. Among different semi-automatic methods, a commonly used category of tools in commercial software solutions is the contour interpolation [e.g., shape based interpolation (Prabhakar et al. 2011)]. One such tool is the *between-slice interpolation* which generates a 2D contour based on the nearest contours on the inferior and superior slices, and the physician is expected to make corrections until reaching a satisfactory result. The advantage of such a semi-automatic method is that it accelerates the contouring process by combining the power of computing and human expertise for the initial contours, based on the assumption of the continuity of the tumor shape while allowing physicians to control the outcomes.

Physicians play a central role in steering and correcting the outcomes of the contouring task (Heckel et al. 2013). However, their cognition can be influenced by those automatically generated or corrected contours, especially as there is no gold standard in GTV contouring (Weiss and Hess 2003; Aselmaa et al. 2017). A higher level of automation can introduce a higher level of bias (Manzey et al. 2012). On the other side, lower level of automation, which has a higher level of human involvement, may have a smaller gap between physicians' cognition and the data; thus, the influence of the automation can be expected to be smaller. The contouring task is an iterative process during which contours are being inspected multiple times. Therefore, it is expected that the gap narrows even further in this iterative process. However, literature study did not reveal to what extent such or similar interpolation may influence the physicians' decision-making process. The questions about the clinical relevance of such an automation bias and its effects on the inter-observer variation also remain to be answered.

2 Methods

2.1 Study setup

To evaluate the influence of using the between-slice interpolation tool on the resulting GTV contours and on the contouring process, a GTV contouring study was conducted in the radiotherapy department of Institut Claudius-Regaud, Institut Universitaire du Cancer de Toulouse-Oncopole, Toulouse, France with five physicians (three medical residents, two attending physicians) over the period of 5 days. The investigated task was the GTV contouring of the Glioblastoma Multiforme (GBM) tumor, a common type of primary brain tumor (Behin et al. 2003). Four patient cases (a sample case, C-1, C-2, and C-3) were used in the study. Subjective rating of the case difficulties (easy, medium difficulty or difficult) was given by an experienced physician independently from the present study. Similar to the clinical practice, eight image datasets were made available for the physicians for each case. Those eight images datasets were: radiotherapy planning CT, radiotherapy planning MRI T1-weighted with contrast enhancement, radiotherapy planning MRI T1-weighted without contrast enhancement, radiotherapy planning MRI T2-weighted, radiotherapy planning MRI FLAIR, pre-surgery MRI T1-weighted with contract enhancement, pre-surgery MRI T1-weighted without contrast enhancement, and pre-surgery MRI FLAIR. Prior to conducting the study, the MRI datasets were co-registered to the radiotherapy planning CT coordinate system. The distance between any two consecutive axial slices was 2.5 mm in the case C-1, and 1.25 mm for the other two cases.

The study was conducted using a modified and extended version of a research contouring software (Steenbakkers et al. 2005) (Fig. 1). The software allowed manual contouring (i.e., using the freehand and/or the nudge tools) and *between-slice interpolation* (i.e., using the interpolation tool) on any of the axial slices of any of the available datasets displayed on the computer screen. Using the interpolation tool, a contour could be generated on the displayed axial slice based on the contours on the nearest neighboring slices via linear interpolation where the point correspondences were obtained using a radial coordinate system. Within this study, the interpolation tool was used only for creating the initial contour, i.e., the method was only available when there was no existing contour on the slice. The interpolation tool was not available for correcting an existing contour, neither for the first and last slices, as the interpolation relied on the information of the neighboring contours. For the rest, to guarantee the "natural" performance of physicians as it is in a clinical setting, physicians were free to choose either the manual or the interpolation tool to create the initial 2D contours. We expected that such a "randomized" setup will minimize the cognitive and psychological difference in the selection of methods. Meanwhile, the software recorded all user interactions during the task into a log file together with timestamps.

In the beginning of the task, the physician was introduced to the software and a brief training was given with the sample case. In addition, the physician was allowed to explore the software further as they felt necessary. Then, the physician was asked to contour the GTV of the rest of the three GBM cases. Following the treatment protocol (Stupp et al. 2005), the GTV was instructed as "consisting of the resection cavity and any residual contrast-enhancing tumor". The order in which these three cases were presented varied among physicians in order to distribute the impact of learning effects. The researcher was observing the task during the study and was available for assistance with the use of the software at request.

After finishing each GTV contouring task, each physician filled the NASA Task Load Index (NASA-TLX) questionnaire (Hart and Staveland 1988), which was used for assessing the mental workload of physicians based on the subjective rating on six aspects: the physical demand, the mental demand, the temporal demand, the performance, the effort, and the frustration. The original NASA-TLX consists of two parts: rating each aspect, and comparing them pairwise based on their perceived importance. However, it has been shown that the unweighted and the weighted ratings have a high correlation (Noyes and Bruneau 2007). In this study, the outcome of the NASA-TLX was calculated based on the unweighted ratings.

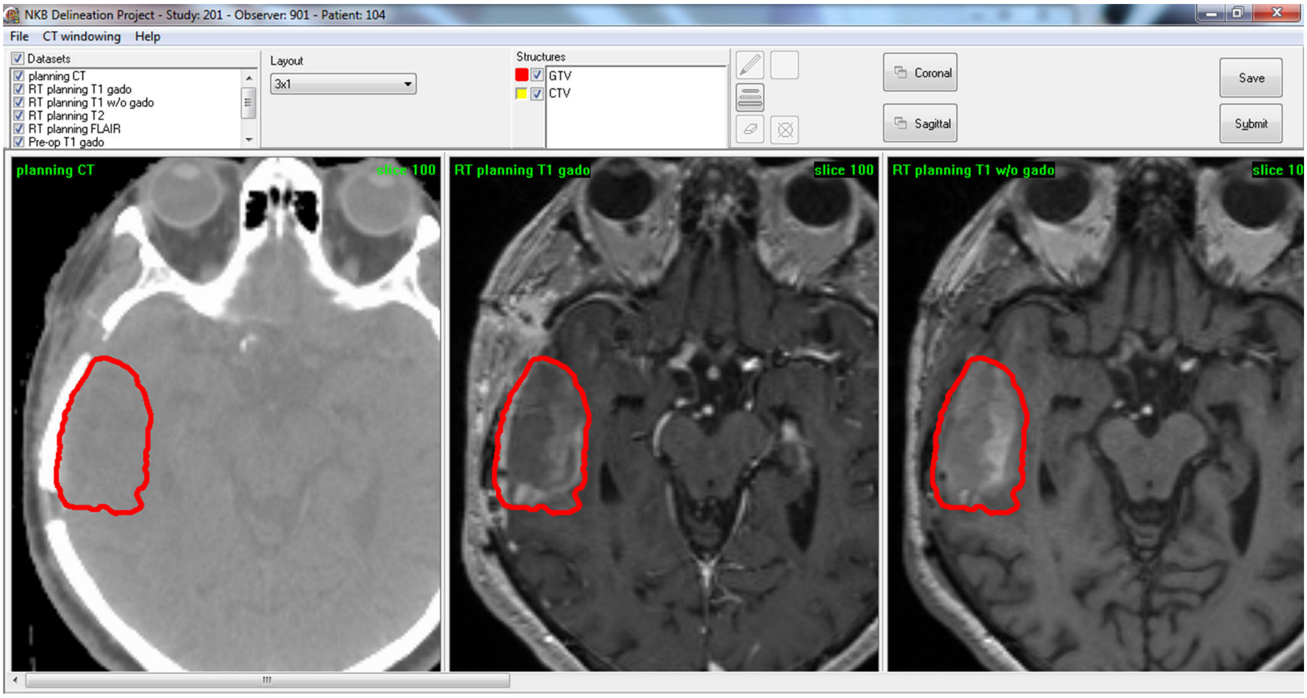


Fig. 1 A screenshot of the software prototype used in the study. The contouring tools available in the study are in the top part of the graphical user interface. The image datasets are presented in the middle-bottom region. An illustrative 2D contour of the GTV is shown in red color, overlaid on the radiotherapy planning CT (left image), radiotherapy planning MRI T1-weighted with contrast enhancement (middle image), and radiotherapy planning MRI T1-weighted (right image)

2.2 Data inclusion and analysis

Figure 2 illustrates different steps in the data inclusion and analysis process. In the proposed research, contouring was made possible on axial slices only. The slices toward the superior and inferior boundary of the tumor typically have a larger level of variation than the central slices. For instance, given a boundary slice, it was often that not all physicians contoured, i.e., not all physicians agreed that there was tumorous tissue. Such a cognitive difference often leads to large deviations among the boundary slices. The influence of automation, if any, was expected to be smaller than other influences. Therefore, two criteria were applied to eliminate "boundary slices": (1) the slices on which not all physicians contoured ($N < 5$) were excluded from the analysis; (2) for the remaining slices, the mean enclosed areas of each contour ($\mathrm{Mean}_{\mathrm{area}}$) and the standard deviation ($\mathrm{SD}_{\mathrm{area}}$) among them were calculated for each slice over the observers. Then, the coefficient of variation (i.e., relative standard deviation, $\mathrm{CV}_{\mathrm{area}}$) within each slice was calculated as the ratio between $\mathrm{SD}_{\mathrm{area}}$ and $\mathrm{Mean}_{\mathrm{area}}$. Contours that would be included in the further analysis were defined by its $\mathrm{CV}_{\mathrm{area}}$ being less than the mean of $\mathrm{CV}_{\mathrm{area}} + 1\ \mathrm{SD}_{\mathrm{area}}$ of the given case.

Then contours on the included slices were categorized as being *manual* or *interpolated* based on whether interpolation was used to generate the initial 2D contour or not. The 2D slices on which at least one contour was interpolated or manually created, remained for the further analysis. All contours were resampled to increase the point density—the maximum distance between two neighboring points of a resampled contour was 0.01 mm as we wanted to achieve a 0.1 mm measurement accuracy.

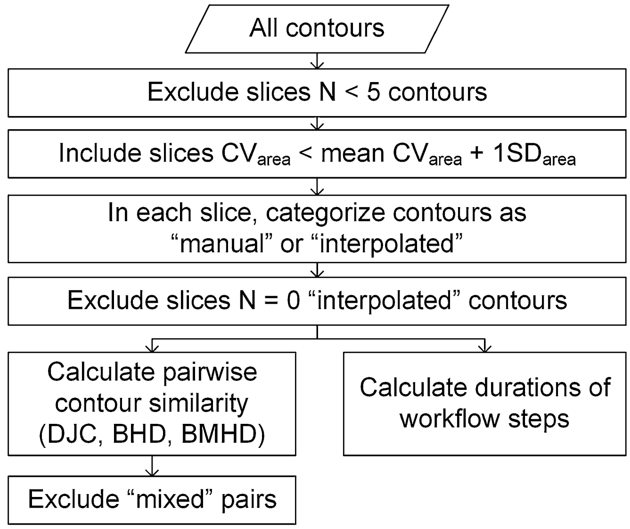

All contours
Exclude slices N < 5 contours
Include slices CV$_{area}$ < mean CV$_{area}$ + 1SD$_{area}$
In each slice, categorize contours as "manual" or "interpolated"
Exclude slices N = 0 "interpolated" contours
Calculate pairwise contour similarity (DJC, BHD, BMHD)
Calculate durations of workflow steps
Exclude "mixed" pairs

Fig. 2 Data inclusion and analysis process

Using three different measures, the similarity of contours was evaluated pairwise by a program developed based on the MevisLab® (MeVis Medical Solutions AG 2016; Kuijf 2015). The Dice–Jaccard coefficient (DJC) (Fotina et al. 2012) was introduced as a measure of the overlap of the enclosed areas between two contours where 1 indicates complete overlap and 0 indicates no overlap. The Bidirectional Hausdorff Distance (BHD) (Huttenlocher et al. 1993) was used to measure the largest variation between shapes of two contours. BHD is defined based on Direct Hausdorff Distance (DHD). Given two contours $C1$ and $C2$, DHD delivers the distance from $C1$ to $C2$ and it can be defined as $\mathrm{DHD}(C1, C2) = \sup_{r\in C1} (\inf_{s\in C2}|r - s|)$. In a generalized discreet form, contours $C1$ and $C2$ are available as the point sets P_{C1} and P_{C2}, where $P_{C1} = \{P^i_{C1} \in C1|i = 1, m\}$ and $P_{C2} = \{P^i_{C2} \in C2|i = 1, n\}$, representing contour $C1$ and $C2$, respectively. Thus the DHD from P_{C1} to P_{C2} is $\mathrm{DHD}(C1, C2) = \max_{i=1,m} \min_{j=1,n}\left|P^i_{C1} - P^j_{C2}\right|$. Though DHD is able find the largest shape variation from contour $C1$ to $C2$, it is directional, i.e., $\mathrm{DHD}(C1, C2)$ is not always same as $\mathrm{DHD}(C2, C1)$. Therefore, we introduced BHD which is defined as $\mathrm{BHD}(C1, C2) = (\mathrm{DHD}(C1, C2)+ \mathrm{DHD}(C2, C1))/2$. Similar to the concept of BHD, to measure the average deviation between contour $C1$ to $C2$, we introduced Bidirectional Mean Hausdorff Distance (BMHD), which is defined as $\mathrm{BMHD}(P_M, P_E) = \frac{1}{2}\left(\frac{1}{m}\sum_{i=1}^{m} \min_{j=1,n}\left|P^i_{C1} - P^j_{C2}\right| + \frac{1}{n}\sum_{i=1}^{n} \min_{j=1,m}\left|P^i_{C2} - P^j_{C1}\right|\right)$, as the overall shape similarity measure (Song et al. 2017). BMHD is nondirectional regarding contours and comparing to BHD, it is able to reduce the sensitivity to noise and represents the overall shape similarity between contours $C1$ and $C2$.

Measures of the contours were calculated for $C^2_5 = 10$ pairs of contours in each slice. Those pairwise measures were then categorized as being *manual* (both physicians contoured manually), *mixed* (one physician contoured manually, the other used interpolation), or *interpolated* (both physicians used the interpolation tool). The *mixed* pairs were not further analyzed. Independent samples *t* test was conducted to evaluate the significance of variation in the mean values using SPSS (version 22).

The details of the software interactions within each slice were extracted from the interaction log files. Each interaction was categorized according to the moment it happened within the steps of the workflow: familiarization, initial contouring, immediate correction, evaluation, and additional corrections. The duration of each of the contouring workflow step was calculated as a sum of the durations of the interactions occurring within this step. Since not all physicians had interactions within each of the five workflow steps, the overall occurrence rate was calculated as a percentage of the total number of engagements in the step over the total number of contouring workflows of the given case. Independent samples *t* test was conducted to evaluate the significance of variation among the durations of the workflow steps in the manual and interpolated contour using SPSS.

3 Results

The subjective ratings of the cases were given by a physician prior to the study as: case C-1 was identified as easy, case C-2 as difficult, and case C-3 as medium difficulty (see example Fig. 3). The calculated NASA-TLX indexes corresponded to the rated difficulty levels, though gaps among them were small: the individual NASA-TLX index values being 5.6 out of 20 in case C-1, 7.8 out of 20 in case C-2, and 6.7 out of 20 in case C-3 (Fig. 4).

In total, 377 GTV contours on 83 slices were created by five physicians over the three cases. Fifteen slices had less than five contours on them and were excluded from further analysis. On the remaining slices, the mean enclosed area of contours in a slice was 448 mm^2 (SD = 199 mm^2) in case C-1, 876 mm^2 (SD = 323 mm^2) in case C-2, and 596 mm^2 (SD = 269 mm^2) in case C-3. In boundary slices toward the superior and inferior directions, the mean enclosed areas $Mean_{area}$ were decreasing as expected.

The CV_{area} gives a comparable measure of variation of the contoured areas on each slice, with a value of 0 indicating no variation. The mean CV_{area} was 0.18 (SD = 0.17), 0.22 (SD = 0.25) and 0.15 (SD = 0.20) for the three cases C-1, C-2, and C-3, respectively. Based on the CV_{area} and the standard deviation of it, six slices were categorized as "outliers" and were excluded from further analysis. In addition, six slices were eliminated as only containing one type of contours (all manual). As a result, for the detailed analysis, contours on 56 slices remained: 8 slices in case C-1, 23 slices in case C-2, and 25 slices in case C-3, involving 280 individual contours (40 in C-1, 115 in C-2, and 125 in C-3). Among these 280 contours, 144 contours were initiated manually (manual group), and 136 were initiated using the interpolation tool (interpolation group).

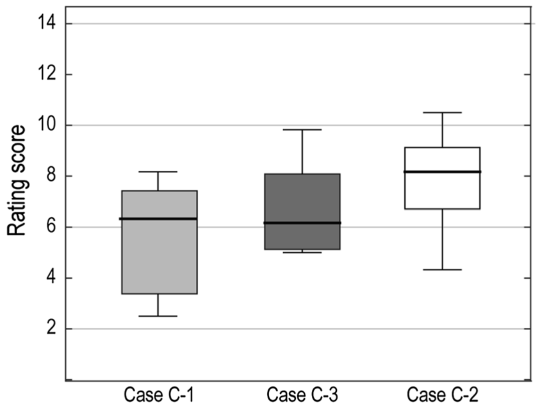


Fig. 4 Boxplot of the results of NASA-TLX regarding case C-1, C-2 and C-3, the sequence is adjusted according to the mean difficulty levels

3.1 Influence of automation to the contours

An overview of the calculated measures of the included contours is presented in Table 1.

The overlap between physicians' contours was generally high, with the overall mean DJC being 0.79 (min = 0.30, max = 0.94). In the interpolation group, the overall mean DJC was 0.81, thus being slightly higher than in the manual group where it was 0.77. In the studied cases, the DJC showed a tendency to be on average higher by 0.04–0.09 when the interpolation tool had been used. In two of the

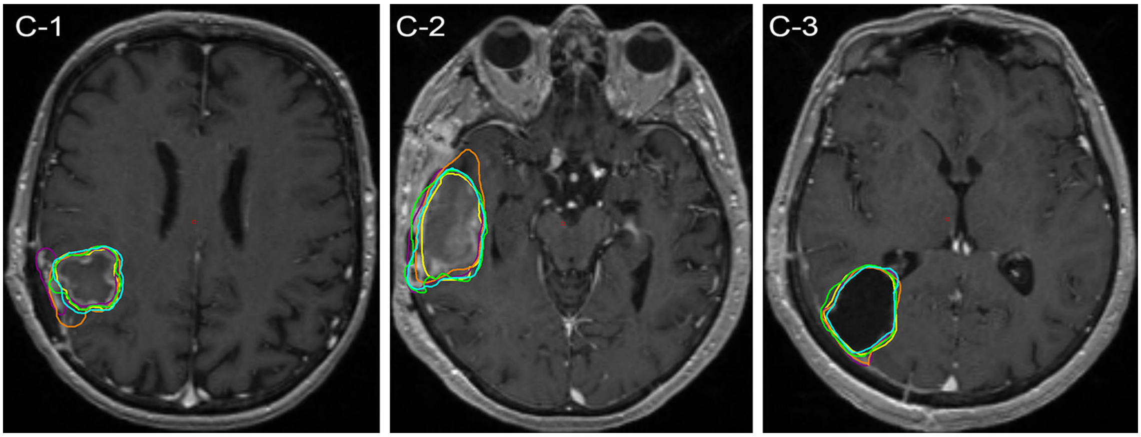


Fig. 3 Examples of the three cases and the resulting contours on 2D axial slices. The contours of five physicians (each in different color) are overlaid on MRI T1-weighted contrast enhanced image of case C-1, C-2 and C-3

Table 1 Overview of contour similarity measures

Measure Subjective rating	Grouping	Case C-1 Easy	Case C-2 Difficult	Case C-3 Medium
Number of contours	Manual	26	57	61
	Interpolated	14	58	64
Mean area	Manual	517 mm^2	947 mm^2	617 mm^2
	Interplolated	541 mm^2	908 mm^2	695 mm^2
	p value	0.607	0.489	0.060
Mean DJC	Manual	0.78	0.72	0.80
	Interpolated	0.87	0.76	0.85
	p value	0.002	0.155	0.011
Mean BHD	Manual	4.5 mm	10.5 mm	4.2 mm
	Interpolated	2.5 mm	7.8 mm	3.3 mm
	p value	0.003	0.005	0.005
Mean BMHD	Manual	1.4 mm	2.3 mm	1.3 mm
	Interpolated	0.9 mm	2.1 mm	1.1 mm
	p value	0.106	0.209	0.038

The following measures were calculated pairwise between two physicians in a slice within the group: *DJC* Dice–Jaccard coefficient; *BHD* Bidirectional Hausdorff Distance; *BMHD* Bidirectional Mean Hausdorff distance. Mean over these individual pairwise measures is presented in this table. *p* values are from the independent samples *t* test conducted between these two groups

three cases, the improvement also reached statistical significance ($p = 0.002$, $p = 0.011$ for cases C-1, and C-3, respectively). Such an increase indicated that contours initiated by the interpolation tool were more similar to each other within a slice.

The BHD on a slice was significantly smaller when interpolation was used for all three cases ($p = 0.003$, $p = 0.005$, and $p = 0.005$). The decrease was the highest in case C-2, where it was reduced by 2.7 mm, followed by C-1 where it was less by 1.5 mm and the smallest reduction was in case C-3 by 0.9 mm. In terms of the overall shape similarity as measured by BMHD, the average distance between the two contours, independent from its creation method, was 1.2 mm in the cases C-1 and C-3, and 2.2 mm in case C-2. Generally, the mean BMHD showed a tendency toward a decrease when the interpolation had been used but was only significant in case C-3 ($p = 0.038$).

3.2 Influence of automation on the contouring process

Detailed contouring workflow within a slice as was observed in the conducted study is depicted in Fig. 5. The initial contouring step (Step 2A or 2B) represented the action of creating the first (i.e., initial) closed loop boundary of the visible tumor, visually inspecting and perceiving the contour and/or the medical image(s) while contouring, as well as of deciding on the next action (i.e., to correct the contour or to navigate away). Contour corrections were categorized as immediate corrections and additional corrections. The immediate corrections (Step 3) accounted for the corrections of the contour until the first slice change (i.e., navigate away). These corrections were done, for example, to compensate for mouse inaccuracy (Zabramski 2011), or to adjust the contour based on the further inspection of the presented 2D medical image(s) as well as clinical reasoning. Returning to the contour for corrections after inspecting the neighbor slices or at any later moment, were identified as additional corrections (Step 5).

The mean durations of the workflow steps are presented in Fig. 6. Generally, a physician completed the contouring task faster when using the interpolation tool. In terms of specific workflow steps, when interpolation was used, physicians tended to spend more time on familiarizing (Step 1) and less time on evaluating (Step 4). Furthermore, some physicians tended to spend more time to complete the task compared to others.

The details of the workflow steps averaged over all physicians for each case are shown in Table 2. In addition, within each workflow, the total durations of contour corrections (sum of time spent on Step 3 and Step 5) were calculated. The total durations of the contouring process on a slice per physician were also summed. Furthermore, the average duration of each step was also calculated over the three cases.

When the initial contour was done manually, physicians always returned to the slice (i.e., 100% occurrence of the evaluation step). No separate evaluation step was recorded in 15% ($N = 20$) of all contours initiated by the interpolation tool, which indicated that after the initial contour was interpolated, and possibly corrected ($N = 2$), the

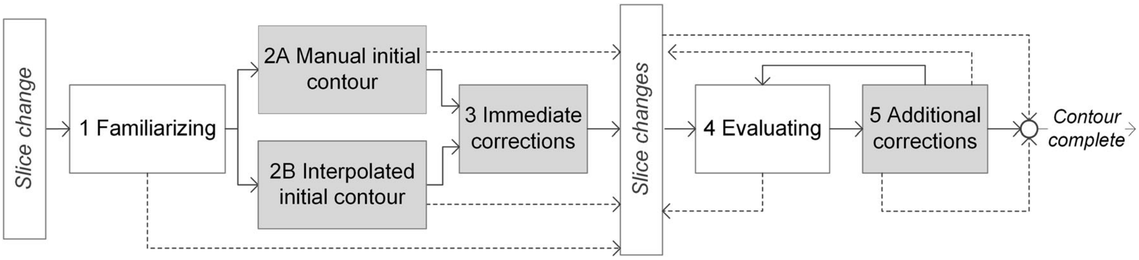
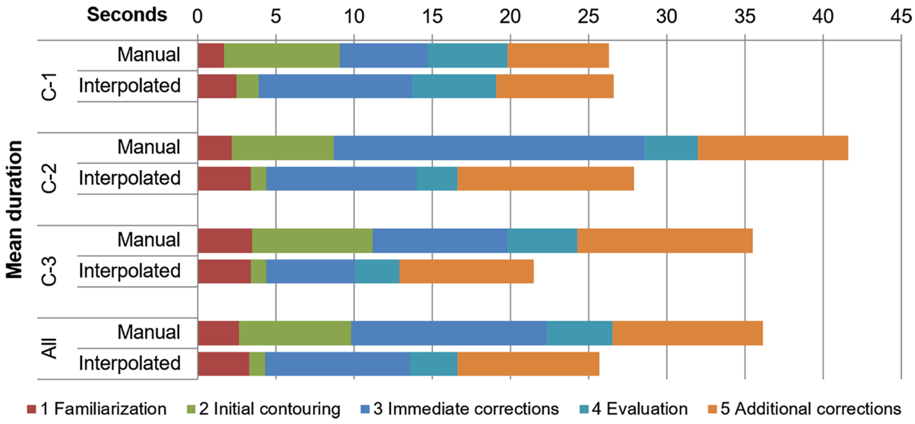


Fig. 5 Contouring workflows of this study with a variation in the initial contour creation step. In the manual workflow, as step 2A the physician manually contoured the boundary of the tumor. In the interpolated workflow, as step 2B the physician used the between contour interpolation tool. Data regarding the contouring process were extracted according to these workflow steps

Fig. 6 Mean durations of different workflow steps in case C-1, C-2, and C-3. The type of workflow is labeled as manual or interpolated. In addition, the average of each step over all cases is shown as "All"



physician did not revisit it. Regarding individual cases, such contouring workflows were present in two of the cases, being 22% ($N = 13$) in case C-2 and 10% ($N = 7$) in case C-3. Further analysis revealed that the average viewing time of those interpolated contours was 0.6 s (SD = 0.19 s), which is less than the overall average of 1.0 s. More than half of such contours ($N = 9$ in C-2, $N = 4$ in C-3) could be accounted for one physician.

4 Discussion

4.1 Automation bias

Automation may influence physicians' reasoning during contouring by providing an automatically generated contour. When such a contour is accepted without sufficient evaluation of the available data, automation bias occurs and errors might be introduced. Automation bias may have either negative or positive effect on the process and the outcomes of the contouring task, as in many steps of the contouring task, physicians must make a subjective decision based on their knowledge and experience.

The influences of automation on the reasoning process are more difficult to be categorized as being positive or negative. One of the challenges in evaluating the outcomes of a contouring task is that there is no gold standard in GTV contouring (Weiss and Hess 2003). There is general acknowledgement that less variation among physicians is desired, i.e., methods which lead to reduced inter-observer variation with improved consistency are preferred. However, categorizing variations to be erroneous is challenging due to the nature of task. Another aspect that can be measured is the amount of time spent on inspecting data as shown in the *Familiarizing* and *Evaluating* steps of the workflow. However, increased time does not necessarily correlate with the quality of contours as physicians are capable of detecting abnormalities rather rapidly (Drew et al. 2013).

In this paper, inter-observer variation of selected contours was used to evaluate effects of automation bias on the outcomes. On the negative side, the automation bias may lead to errors in the contours. On the positive side, it may increase consistency of contours. The inter-observer variation can be evaluated by different types of measures such as the DJC (area overlap), BHD (shape outliers) and BMHD (shape similarity), where smaller variation among physicians indicates higher confidence in having the "consistent" tumor contour. Regarding the process, the mean durations of different steps of the task were adopted as the measures of effects of automation bias.

Table 2 Details of the workflow steps of the three cases

Workflow		C-1		C-2		C-3		All	
Step	Type	Occurrence (%)	Mean duration (±1 SD)	Occurrence (%)	Mean duration (±1 SD)	Occurrence (%)	Mean duration (±1 SD)	Occurrence (%)	Mean duration (±1 SD)
1 Familiarizing	Manual	100	1.7 s (±2.7 s)	100	2.2 s (±1.4 s)	100	3.5 s (±5.2 s)	100	2.7 s (±3.7 s)
	Interpolated	100	2.5 s (±1.3 s)	100	3.4 s (±2.1 s)	100	3.4 s (±3.3 s)	100	3.3 s (±2.7 s)
	p value	–	0.303	–	0.001	–	0.937	–	0.100
2A/B Initial contour	Manual	100	7.4 s (±2.2 s)	100	6.5 s (±1.9 s)	100	7.7 s (±3.3 s)	100	7.2 s (±2.7 s)
	Interpolated	100	1.4 s (±0.9 s)	100	1.0 s (±0.8 s)	100	1.0 s (±1.3 s)	100	1.0 s (±1.0 s)
	p value	–	N/A	–	N/A	–	N/A	–	N/A
3 Immediate corrections	Manual	8	5.6 s (±3.8 s)	7	19.9 s (±14.0 s)	7	8.6 s (±6.5 s)	7	12.5 s (±11.0 s)
	Interpolated	50	9.8 s (±3.1 s)	28	9.6 s (±9.5 s)	5	5.7 s (±3.2 s)	19	9.2 s (±7.7 s)
	p value	–	0.142	–	0.095	–	0.468	–	0.316
4 Evaluating	Manual	100	5.1 s (±4.0 s)	100	3.4 s (±2.5 s)	100	4.5 s (±3.3 s)	100	4.2 s (±3.2 s)
	Interpolated	100	5.4 s (±7.0 s)	78	2.6 s (±4.8 s)	89	2.8 s (±3.2 s)	85	3.0 s (±4.5 s)
	p value	–	0.853	–	0.347	–	0.004	–	0.018
5 Additional corrections	Manual	46	6.5 s (±7.0 s)	46	9.6 s (±7.5 s)	39	11.2 s (±9.4 s)	43	9.6 s (±8.2 s)
	Interpolated	57	7.5 s (±5.3 s)	22	11.3 s (±9.8 s)	45	8.6 s (±6.6 s)	37	9.1 s (±7.4 s)
	p value	–	0.753	–	0.567	–	0.242	–	0.723
(3 + 5) Total correction	Manual	50	6.9 s (±7.0 s)	51	11.4 s (±10.4 s)	43	11.7 s (±9.2 s)	47	10.6 s (±9.4 s)
	Interpolated	79	11.7 s (±6.1 s)	40	13.1 s (±14.0 s)	48	8.6 s (±6.3 s)	48	10.7 s (±9.8 s)
	p value	–	0.092	–	0.623	–	0.144	–	0.969
Total	Manual	–	17.7 s (±9.9 s)	–	18.1 s (±12.0 s)	–	20.8 s (±13.5 s)	–	19.2 (±12.3 s)
	Interpolated	–	18.6 s (±14.3 s)	–	12.0 s (±15.9 s)	–	11.0 s (±8.9 s)	–	12.2 (±13.0 s)
	p value	–	0.818	–	0.022	–	<0.001	–	<0.001

Occurrence percentage, the mean duration (in seconds per slice), and standard deviation (SD) of the duration of each workflow step for both types of workflows are listed. Step 2A and 2B were not representing the same type of interactions; thus, statistical comparison was not suitable (N/A). s = second

4.2 Inter-observer variation among outcomes

In radiotherapy, 60% or more of the mis-administrations are due to human error (Duffey and Saull 2002). Lack of a gold standard, as well as the expected variation among physicians, increases the probability of human errors during the contouring task. For instance, Brundage et al. (1999) identified that insufficient target volumes were one of the common reasons for treatment plan modification. In order to tackle this, in clinical practice, peer review is the proposed approach to decrease the probability of such (and other) human errors (Marks et al. 2013; Mackenzie et al. 2016; Brunskill et al. 2017). In short, it is expected that the smaller the variations among physicians are, the fewer errors there are.

Variation among physicians is well documented (e.g., van Herk 2004; Louie et al. 2010; Fotina et al. 2012; Dinkel et al. 2013; Whitfield et al. 2013). However, there is a lack of consensus on which measures to use for judging the variability (Fotina et al. 2012). Furthermore, there is no reproducible gold standard for evaluating the accuracy of contours due to many reasons [e.g., image quality, and subjectivity of physicians (Weiss and Hess 2003)]. In many studies, a manual contour done by an experienced physician (i.e., expert contour) is being used as a reference (Olabarriaga and Smeulders 2001). Such an approach can be sufficient to evaluate the reproducibility of an automatic contouring method, but the results are dependent on the contours provided by that expert. This study aimed to measure whether the manually initiated contours were more similar to each other than contours initiated by the interpolation tool. Thus, we incorporated pairwise contour similarity measures such as pairwise DJC, pairwise BHD, and pairwise BMHD. Based on the results presented in the results section, we observed a tendency that contours initiated by the interpolation tool were slightly more similar to each other among different physicians than manually initiated contours. In all three cases, the mean BHD and BMHD decreased, while also the mean DJC showed improvement. Statistical significance was reached for six of the nine pairwise calculated similarity measures. One of the sources for the increase is shape similarity might be that the computer is better in creating a smoother shape compared to the human, who must draw it manually with a mouse in this study.

Though the shape similarities of the GTV contours were improved, the improvements were below the current accuracy of radiotherapy. For instance, we observed a mean shape variation (measured by BMHD) decrease by 0.2–0.5 mm. In the treatment plan of GBM, the recommended margin to encompass possible treatment delivery uncertainties is between 3 and 5 mm, depending on the specific situation (Niyazi et al. 2016). Such margins are used to compensate the uncertainties in the GTV contouring as well as for shifts in patient positioning. For instance, Drabik et al. (2007) measured that on average there was an (up to) 0.5 mm positioning shift of a GBM patient in the treatment. Nevertheless, among multiple sources of uncertainty within the radiotherapy planning process (van Herk 2004), GTV contouring has been identified as the weakest link (Njeh 2008). Thus, decreasing variation in GTV contouring can be beneficial especially that the level of precision of dose delivery is increasing (Schaffner and Pedroni 1998).

The case difficulty could not be clearly associated with reduced variations of the contours initiated by the interpolation tool. The simplest case (C-1) showed the largest improvement, while the medium difficulty case (C-3) and difficult case (C-2) showed similar tendencies. Therefore, further studies with more cases of varying levels of difficulty are required to evaluate the correlation between the decrease of variation by utilizing automation and the difficulty of the case. At the same time, it was clear that the level of difficulty is related to the general level of variation among physicians. The more difficult case in the study (C-2) had the lowest DJC and the highest BHD. Besides, the BMHD in this case was nearly double compared to the other two cases.

4.3 The efficiency of and the influences on the contouring process

Detailed analysis of the contouring process reveals the impact of incorporating automatic initial contour creation (i.e., interpolation) to the overall process. The between-slice interpolation tool that was investigated in this study, changed the way the initial contour was created (click of a button or press of a key on keyboard, instead of drawing with the mouse). As expected, including automation generally decreased the overall contouring time. In the case C-1, the average duration was slightly higher, though not statistically significant. For this specific case, it might have been influenced by the small size of the tumor, larger slice thickness (2.5 mm instead of 1.25 mm), or being an easy case. In the case C-2, the overall duration was reduced due to the shorter initial contour drawing time. In the case C-3, the evaluation step was also significantly shorter when the interpolation tool had been used, resulting in a further reduction of the task completion time.

The availability of the interpolation tool for some physicians changed their contouring strategy. During this interpolation-influenced contouring strategy, the physician would first contour in a set of slices manually while skipping some in-between slices (i.e., seeing them but not contouring on them), and then return to the empty slices later in the process and utilize interpolation to fill in the missing contours. This type of contouring strategy is characterized by slight changes of the contouring workflow on the interpolated slices: longer time may be spent in familiarizing (step 1), fewer additional corrections (steps 3 and 5) on the interpolated contours, and there are fewer (or no) returns (step 4) to the slice once interpolation had been used.

The frequency of corrections gives a measure of the acceptance of the contour. Based on the presented three cases, it was observed that the frequency of corrections (on average 47.5% of the cases), as well as the duration of them, remained similar for both manually drawn initial contours (47%) and contours initiated by the interpolation tool (48%). This could indicate that if a contour is in a clinically acceptable range, then the likelihood of a manual correction is independent from its' original creation method. Eighty-four percent of these corrections occurred after returning to the contoured slice at a later point. One common motivation for correction, for example, is a comparison with neighboring slices (Aselmaa et al. 2017).

These later stage corrections can be assumed to correspond to the physicians updating their mental model (Varga et al. 2013) and then correcting the contours correspondingly.

In medical image related decision-making, the duration of one second is considered to be a significant allocation of visual attention for detecting an object of interest (Hillstrom 2000). In addition, it has been shown that a visual fixation time of one second is significantly correlated with correct detection of a lesion (Nodine et al. 2002). In our study, an interpolated contour was on average viewed for one second prior to an action, indicating that the level of evaluation for determining the correctness of a contour could be deemed sufficient. In the study, 15% of the interpolated contours were not revisited. In the contouring process of those contours, the physician spent on average 0.6 s viewing it prior to changing to another slice, being below the recognized sufficient level of visual attention allocation. However, this measure on its own is not sufficient for concluding whether this 0.6 s is a sufficient duration of visual inspection in such specific cases. At the same time, interpolated contours showed a slight improvement in the inter-observer variation. Thus, even though the automation bias seems to be present, it was leading toward more desirable results and reductions in the overall task completion times. Therefore, the use of interpolation can be encouraged.

4.4 Pros and cons of automation

The reasoning occurring during the contouring task is influenced by a number of variables, such as the type of treatment, whether there was a preceding surgery, the size and the location of the tumor, tumor characteristics, etc. (Aselmaa et al. 2014). Physicians need to weigh such various aspects against their past experiences in order to reach a decision. This process can be seen as case-based reasoning where individual knowledge captured from a very specific context (e.g., treating a particular patient with a particular disease) can be extrapolated to similar contexts (Pantazi et al. 2004).

The benefit of a (semi-)automatic contouring method strongly depends on its robustness. For example, during this study, in few instances, the interpolation generated a partially zig-zag contour instead of a smooth one which took physicians' more than average efforts to correct. Automatically generated contours, that are found unacceptable, result in unnecessary software interactions and thus could increase workplace frustrations. It has been reported that in general there is a rather high loss of work time due to frustrating experiences with software (Lazar et al. 2006) which in turn led to higher financial costs and possibly even impacts the outcomes of the treatment (Johnson 2006). Therefore, advances in improving the robustness and increasing the accuracy of contouring methods, together with improving the general usability of software solutions, are required.

In our study, it was identified that automation guided physicians toward more similar contours, which is a desired effect as there is no gold standard. We postulate that when the automation is used to provide contouring aids on 2D slices, the automation bias is more noticeable on the slices where the level of cognitive involvement is lower. At the same time, automation bias can be more prominent in the more cognitively demanding situation, but may be obfuscated by other variables influencing the physician's subjective reasoning process.

4.5 Limitations

The study presented was conducted on three different patient datasets. Conducting a study involving manual contouring is challenging due to the time requirement from the physicians. However, a larger sample size would be beneficial to have a deeper understanding the influence of automation bias in relation to other variables such as the size of the tumor, slice thickness, levels of case difficulties, or levels of physicians' experience.

The case difficulties were based on a subjective rating of a senior physician acquired independently from the present study. Those ratings were given in three-point scale (easy, medium, difficult). A more robust evaluation method for determining case difficulty could be beneficial, for example, objective description of the tumor based on image features (Gevaert et al. 2014).

The aim of this study was to investigate the automation bias in a naturalistic setting. While we found our findings valuable, a controlled study with fewer variables (e.g., predefined choice of the tool per physician) may reach stronger conclusions. In addition, though this study describes the relations between automation bias and the reasoning process based on the software interaction data, studies complemented with eye tracking might reveal more insights of the influence of automation on the reasoning process.

5 Conclusion

Automation is increasingly incorporated into the radiotherapy planning process. This paper presented a study of evaluating the impact of using a between-slice interpolation for initiating a contour on the resulting contours as well as on the contouring process in comparison with the fully manual contouring.

A GTV contouring study with five physicians on three patient cases was conducted, from which 280 individual 2D contours were analyzed. The contours obtained with and without the use of the interpolation tool were pairwise analyzed within each slice in terms of area overlap (DJC), shape outliers (BHD), and overall shape similarity (BMHD). In all measures, outcomes based on the use of the interpolation tool showed an increased agreement among physicians (DJC increase by 0.04–0.09; BHD decrease by 0.9–2.7 mm; BMHD decrease by 0.2–0.5 mm).

Influences to the contouring process were also identified. The efficiency was improved—the overall interaction time within a slice was reduced by 6.1 s ($p = 0.022$) and 9.8 s ($p < 0.001$) in two of the three cases, mainly due to the time-saving in creating the initial contour. In addition, interpolated contours were corrected at a similar rate as manually drawn contours, which indicated a similar level of evaluation. In a sub-set of contouring processes, an interpolation-influenced contouring strategy was identified. This contouring strategy consisted of first contouring in a set of slices manually and then used the interpolation tool to fill in the missing contours in the in-between slices. However, precaution is needed, as in our study 15% of interpolated contours were not revisited after initial creation and inspection.

Based on the presented findings, it can be concluded that using the between-slice interpolation tool influences the contouring outcomes in a desirable direction, as well as significantly decreases task completion time. Thus, the use of such automatic contouring tools can be encouraged in radiotherapy planning software.

Acknowledgements The authors would like to express their appreciations to other members of the SUMMER consortium for their valuable advices regarding the proposed research. The SUMMER project has received funding from the European Union Seventh Framework Programme (FP7-PEOPLE-2011-ITN) under grant agreement PITN-GA-2011-290148. We would also like to thank the physicians from Oncopole who participated in the study. The authors would also like to express their appreciations to Mr. Joop C. Duppen, who developed the original “Big Brother” software, and Mr. Lennert Ploeger, for technical support during software prototype development. The authors would also like to thank Mr. Peter Kun, for his constructive comments during the preparation of the manuscript.

Compliance with ethical standards

Ethical standards Ethical approval for the use of patient data for research purposes was obtained by Département de Radiothérapie, Institut Claudius-Regaud, Institut Universitaire du Cancer de Toulouse-Oncopole. The ethics committee of Delft University of Technology approved the study with physicians. Consents from participants were obtained before the experiment.

References

Aselmaa A, Goossens RHM, Rowland B et al (2014) Medical factors of brain tumor delineation in radiotherapy for software design. In: Ahram T, Karwowski W, Marek T (eds) 5th International conference on applied human factors and ergonomics (AHFE), pp 4865–4875

Aselmaa A, van Herk M, Laprie A et al (2017) Using a contextualized sensemaking model for interaction design: a case study of tumor contouring. J Biomed Inform 65:145–158. doi:10.1016/j.jbi.2016.12.001

Barrett WA, Mortensen EN (1997) Interactive live-wire boundary extraction. Med Image Anal 1:331–341. doi:10.1016/S1361-8415(97)85005-0

Batumalai V, Holloway LC, Kumar S et al (2016) Survey of image-guided radiotherapy use in Australia. J Med Imaging Radiat Oncol. doi:10.1111/1754-9485.12556

Bauer S, Wiest R, Nolte L-P, Reyes M (2013) A survey of MRI-based medical image analysis for brain tumor studies. Phys Med Biol 58:R97–R129. doi:10.1088/0031-9155/58/13/R97

Behin A, Hoang-Xuan K, Carpentier AF, Delattre J-Y (2003) Primary brain tumours in adults. Lancet 361:323–331. doi:10.1016/S0140-6736(03)12328-8

Bravo ER, Ostos J (2017) Performance in computer-mediated work: the moderating role of level of automation. Cognit Technol Work. doi:10.1007/s10111-017-0429-z

Brundage MD, Dixon PF, Mackillop WJ et al (1999) A real-time audit of radiation therapy in a regional cancer center. Int J Radiat Oncol 43:115–124. doi:10.1016/S0360-3016(98)00368-X

Brunskill K, Nguyen TK, Boldt RG et al (2017) Does peer review of radiation plans affect clinical care? A systematic review of the literature. Int J Radiat Oncol 97:27–34. doi:10.1016/j.ijrobp.2016.09.015

Burnet NG (2004) Defining the tumour and target volumes for radiotherapy. Cancer Imaging 4:153–161. doi:10.1102/1470-7330.2004.0054

Delaney G, Jacob S, Featherstone C, Barton M (2005) The role of radiotherapy in cancer treatment: estimating optimal utilization from a review of evidence-based clinical guidelines. Cancer 104:1129–1137. doi:10.1002/cncr.21324

Dinkel J, Khalilzadeh O, Hintze C et al (2013) Inter-observer reproducibility of semi-automatic tumor diameter measurement and volumetric analysis in patients with lung cancer. Lung Cancer 82:76–82. doi:10.1016/j.lungcan.2013.07.006

Dolz J, Kirişli HA, Fechter T et al (2016) Interactive contour delineation of organs at risk in radiotherapy: clinical evaluation on NSCLC patients. Med Phys 43:2569–2580. doi:10.1118/1.4947484

Dowsett RJ, Galvin JM, Cheng E et al (1992) Contouring structures for 3-dimensional treatment planning. Int J Radiat Oncol Biol Phys 22:1083–1088. doi:10.1016/0360-3016(92)90812-V

Drabik DM, MacKenzie MA, Fallone GB (2007) Quantifying appropriate PTV setup margins: analysis of patient setup fidelity and intrafraction motion using post-treatment megavoltage computed tomography scans. Int J Radiat Oncol 68:1222–1228. doi:10.1016/j.ijrobp.2007.04.007

Drew T, Evans K, Võ ML-H et al (2013) Informatics in radiology: what can you see in a single glance and how might this guide visual search in medical images? RadioGraphics 33:263–274. doi:10.1148/rg.331125023

Duffey RB, Saull JW (2002) Know the risk: learning from errors and accidents: safety and risk in today's technology. Butterworth-Heinemann, Oxford

Fitton I, Cornelissen SAP, Duppen JC et al (2011) Semi-automatic delineation using weighted CT-MRI registered images for radiotherapy of nasopharyngeal cancer. Med Phys 38:4662–4666. doi:10.1118/1.3611045

Fotina I, Lutgendorf-Caucig C, Stock M et al (2012) Critical discussion of evaluation parameters for inter-observer variability in target definition for radiation therapy. Strahlenther Onkol 188:160–167. doi:10.1007/s00066-011-0027-6

Gevaert O, Mitchell LA, Achrol AS et al (2014) Glioblastoma multiforme: exploratory radiogenomic analysis by using quantitative image features. Radiology 273:131731. doi:10.1148/radiol.14131731

Hart SG, Staveland LE (1988) Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. Adv Psychol 52:139–183. doi:10.1016/S0166-4115(08)62386-9

Heckel F, Moltz JH, Tietjen C, Hahn HK (2013) Sketch-based editing tools for tumour segmentation in 3D medical images. Comput Graph Forum 32:144–157. doi:10.1111/cgf.12193

Hillstrom AP (2000) Repetition effects in visual search. Percept Psychophys 62:800–817. doi:10.3758/BF03206924

Huttenlocher DP, Klanderman GA, Rucklidge WJ (1993) Comparing images using the Hausdorff distance. IEEE Trans Pattern Anal Mach Intell 15:850–863. doi:10.1109/34.232073

International Commission on Radiation Units and Measurements (1999) ICRU Report 50. Prescribing, recording, and reporting photon beam therapy. Bethesda, MD

Johnson CW (2006) Why did that happen? Exploring the proliferation of barely usable software in healthcare systems. Qual Saf Health Care 15:i76–i81. doi:10.1136/qshc.2005.016105

Kirrmann S, Gainey M, Röhner F et al (2015) Visualization of data in radiotherapy using web services for optimization of workflow. Radiat Oncol. doi:10.1186/s13014-014-0322-3

Kuijf HJ (2015) MeVisLab-Hausdorff-distance source code. https://github.com/hjkuijf/MeVisLab-Hausdorff-distance. Accessed 23 Mar 2017

Lazar J, Jones A, Shneiderman B (2006) Workplace user frustration with computers: an exploratory investigation of the causes and severity. Behav Inf Technol 25:239–251. doi:10.1080/01449290500196963

Lim JY, Leech M (2016) Use of auto-segmentation in the delineation of target volumes and organs at risk in head and neck. Acta Oncol (Madr) 55:799–806. doi:10.3109/0284186X.2016.1173723

Louie AV, Rodrigues G, Olsthoorn J et al (2010) Inter-observer and intra-observer reliability for lung cancer target volume delineation in the 4D-CT era. Radiother Oncol 95:166–171. doi:10.1016/j.radonc.2009.12.028

Mackenzie J, Graham G, Olivotto IA (2016) Peer review of radiotherapy planning: quantifying outcomes and a proposal for prospective data collection. Clin Oncol 28:e192–e198. doi:10.1016/j.clon.2016.08.012

Manzey D, Reichenbach J, Onnasch L (2012) Human performance consequences of automated decision aids. J Cogn Eng Decis Mak 6:57–87. doi:10.1177/1555343411433844

Marks LB, Adams RD, Pawlicki T et al (2013) Enhancing the role of case-oriented peer review to improve quality and safety in radiation oncology: executive summary. Pract Radiat Oncol 3:149–156. doi:10.1016/j.prro.2012.11.010

MeVis Medical Solutions AG (2016) MeVisLab—development environment for medical image processing and visualization. MeVis Medical Solutions AG, Bremen. Available at http://www.mevislab.de/mevislab/features/image-processing/. Accessed 28 Aug 2017

Niyazi M, Brada M, Chalmers AJ et al (2016) ESTRO-ACROP guideline "target delineation of glioblastomas". Radiother Oncol. doi:10.1016/j.radonc.2015.12.003

Njeh CF (2008) Tumor delineation: the weakest link in the search for accuracy in radiotherapy. J Med Phys 33:136–140. doi:10.4103/0971-6203.44472

Nodine CF, Mello-Thoms C, Kundel HL, Weinstein SP (2002) Time course of perception and decision making during mammographic interpretation. Am J Roentgenol 179:917–923. doi:10.2214/ajr.179.4.1790917

Noyes JM, Bruneau DPJ (2007) A self-analysis of the NASA-TLX workload measure. Ergonomics 50:514–519. doi:10.1080/00140130701235232

Nutting C, Dearnaley DP, Webb S (2000) Intensity modulated radiation therapy: a clinical review. Br J Radiol 73:459–469. doi:10.1259/bjr.73.869.10884741

Olabarriaga S, Smeulders AW (2001) Interaction in the segmentation of medical images: a survey. Med Image Anal 5:127–142. doi:10.1016/S1361-8415(00)00041-4

Olsen LA, Robinson CG, He GR et al (2014) Automated radiation therapy treatment plan workflow using a commercial application programming interface. Pract Radiat Oncol 4:358–367. doi:10.1016/j.prro.2013.11.007

Pantazi SV, Arocha JF, Moehr JR (2004) Case-based medical informatics. BMC Med Inform Decis Mak 4:19. doi:10.1186/1472-6947-4-19

Parasuraman R, Manzey DH (2010) Complacency and bias in human use of automation: an attentional integration. Hum Factors 52:381–410. doi:10.1177/0018720810376055

Prabhakar R, Haresh K, Laviraj M et al (2011) A study on the tumor volume computation between different 3D treatment planning systems in radiotherapy. J Cancer Res Ther 7:168. doi:10.4103/0973-1482.82917

Ramkumar A, Dolz J, Kirisli HA et al (2016) User interaction in semi-automatic segmentation of organs at risk: a case study in radiotherapy. J Digit Imaging 29(2):264–277. doi:10.1007/s10278-015-9839-8

Schaffner B, Pedroni E (1998) The precision of proton range calculations in proton radiotherapy treatment planning: experimental verification of the relation between CT-HU and proton stopping power. Phys Med Biol 43(6):1579–1592. doi:10.1088/0031-9155/43/6/016

Skitka LJ, Mosier KL, Burdick M et al (1999) Does automation bias decision-making? Int J Hum Comput Stud 51:991–1006. doi:10.1006/ijhc.1999.0252

Song Y, Hoeksema J, Ramkumar A, Molenbroek JFM (2017) A landmark based 3D parametric foot model for footwear customization. Int J Digital Human (**in press**)

Steenbakkers RJHM, Duppen JC, Fitton I et al (2005) Observer variation in target volume delineation of lung cancer related to radiation oncologist–computer interaction: a "Big Brother" evaluation. Radiother Oncol 77:182–190. doi:10.1016/j.radonc.2005.09.017

Stupp R, Mason WP, van den Bent MJ et al (2005) Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. N Engl J Med 352:987–996

Sykes J (2014) Reflections on the current status of commercial automated segmentation systems in clinical practice. J Med Radiat Sci 61:131–134. doi:10.1002/jmrs.65

van Herk M (2004) Errors and margins in radiotherapy. Semin Radiat Oncol 14:52–64. doi:10.1053/j.semradonc.2003.10.003

Varga E, Pattynama PMT, Freudenthal A (2013) Manipulation of mental models of anatomy in interventional radiology and its consequences for design of human–computer interaction. Cogn Technol Work 15(4):457–473. doi:10.1007/s10111-012-0227-6

Vieira B, Hans EW, van Vliet-Vroegindeweij C et al (2016) Operations research for resource planning and-use in radiotherapy: a literature review. BMC Med Inform Decis Mak 16:149. doi:10.1186/s12911-016-0390-4

Vorwerk H, Zink K, Schiller R et al (2014) Protection of quality and innovation in radiation oncology: the prospective multicenter trial the German Society of Radiation Oncology (DEGRO-QUIRO study). Strahlenther Onkol 190:433–443. doi:10.1007/s00066-014-0634-0

Weersink RA (2016) Chapter 3—image fusion and visualization. In: Farhat WA, Drake J (eds) Bioengineering for surgery, Elsevier, Amsterdam, pp 29–58

Weiss E, Hess CF (2003) The impact of gross tumor volume (GTV) and clinical target volume (CTV) definition on the total accuracy in radiotherapy. Strahlenther Onkol 179:21–30. doi:10.1007/s00066-003-0976-5

Wesley D, Dau LA (2017) Complacency and automation bias in the enbridge pipeline disaster. Ergon Des 25:17–22. doi:10.1177/1064804616652269

Whitfield GA, Price P, Price GJ, Moore CJ (2013) Automated delineation of radiotherapy volumes: are we going in the right direction? Br J Radiol 86:20110718. doi:10.1259/bjr.20110718

Winkel D, Bol GH, van Asselen B et al (2016) Development and clinical introduction of automated radiotherapy treatment planning for prostate cancer. Phys Med Biol 61:8587–8595. doi:10.1088/1361-6560/61/24/8587

Xing L, Thorndyke B, Schreibmann E et al (2006) Overview of image-guided radiation therapy. Med Dosim 31:91–112. doi:10.1016/j.meddos.2005.12.004

Yancik R, Ries LAG (2004) Cancer in older persons: an international issue in an aging world. Semin Oncol 31:128–136. doi:10.1053/j.seminoncol.2003.12.024

Zabramski S (2011) Careless touch. In: Proceedings of the 23rd Australian computer-human interaction conference on—OzCHI ’11. ACM Press, New York, New York, USA, pp 329–332