Radiotherapy and Oncology 201 (2024) 110545

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

# Quantifying and visualising uncertainty in deep learning-based segmentation for radiation therapy treatment planning: What do radiation oncologists and therapists want?

M. Huet-Dastarac [a,*], N.M.C. van Acht [b,c,1], F.C. Maruccio [d,1], J.E. van Aalst [e,f], J.C.J. van Oorschodt [b,c], F. Cnossen [g], T.M. Janssen [d], C.L. Brouwer [e], A. Barragan Montero [a], C.W. Hurkmans [b,c]

[a] *Molecular Imaging, Radiation and Oncology lab (MIRO), UCLouvain, Brussels, Belgium*
[b] *Catharina Hospital Eindhoven - department of radiation oncology, Eindhoven, The Netherlands*
[c] *Eindhoven University of Technology - Department of Electrical Engineering and Department of Applied Physics and Science Education, Eindhoven, The Netherlands*
[d] *The Netherlands Cancer Institute (NKI), Department of Radiation Oncology, Amsterdam, The Netherlands*
[e] *University of Groningen, University Medical Center Groningen, Department of Radiation Oncology, Groningen, The Netherlands*
[f] *University of Twente, Department of Technical Medicine, Enschede, The Netherlands*
[g] *University of Groningen, Department of Artificial Intelligence, Groningen, The Netherlands*

ARTICLE INFO

*Keywords:*
Contouring/segmentation
Patient specific quality assurance
Artificial intelligence/deep learning

ABSTRACT

*Background and purpose:* During the ESTRO 2023 physics workshop on "AI for the fully automated radiotherapy treatment chain", the topic of deep learning (DL) segmentation was discussed. Despite its widespread use in radiotherapy, the time needed to evaluate and correct DL segmentations remains burdensome. While segmentation uncertainty could be beneficial for clinicians, there is a lack of understanding on what information should be presented to ease their task. This study aimed to gather insights from clinicians on uncertainty visualisation options.

*Materials and methods:* Two sessions of structured interviews were conducted across four institutions already using DL segmentation clinically. The first session focused on the main problems hindering the clinical use of DL. In the second session, ten visualisation options displaying uncertainty information at different levels (structure, slice, or voxel) with binary or continuous values were presented. Dosimetric information was also present in some visualisations. For each case, sixteen clinicians (radiation oncologists and radiation therapists) were asked to grade an overall score, the usability, the training required, and the expected time gain.

*Results:* Participants preferred the binary voxel-level uncertainty visualisation, followed by binary structure-level uncertainty visualisation. Combining structure-level and voxel-level visualisation methods has been proposed as a promising approach. The benefits of dosimetric information were perceived diversely among participants since it complexifies the display but could be useful for the online adaptive workflow.

*Conclusion:* Preferences for uncertainty visualisation methods were assessed within a multi-institutional experienced group of clinicians. Further refinement of preferences may help in selecting the best options for clinical implementation.

## Introduction

In October 2023, the sixth ESTRO physics workshop was held in Turin. One of the topics was "AI for the fully automated radiotherapy treatment chain". During the workshop, various discussions took place, among others, about which deep learning (DL) models are now establishing their position in clinical use and what hampers further clinical use of DL models in the radiotherapy treatment process. Although it was clear that DL is used in many parts of the radiotherapy treatment chain, DL segmentation was the most prominent and clinically most widely

* Corresponding author at: Molecular Imaging, Radiotherapy and Oncology (MIRO), Avenue Hippocrate 54, B-1200 Bruxelles, Belgium.
*E-mail address:* margerie.huet@uclouvain.be (M. Huet-Dastarac).
[1] These two authors contributed equally to the work.

https://doi.org/10.1016/j.radonc.2024.110545
Received 23 May 2024; Received in revised form 18 September 2024; Accepted 20 September 2024
Available online 24 September 2024

implemented [1]. The use of DL segmentation models has shown the potential to decrease inter- and intra-observer variation and considerably speed up the delineation process [2–5]. However, these automatic segmentations remain to be carefully reviewed and, when required, edited by radiation oncologists (ROs), radiation therapists (RTTs), or dosimetrists [6]. It was suggested that one of the factors that hamper clinical use of DL segmentation and additional time gain is the absence of information about the uncertainty in the patient-specific segmentations in current (commercial) models.

Uncertainty estimation techniques play a crucial role in providing insights into the confidence levels and potential limitations of DL models when applied to medical image segmentation. In safety–critical environments such as medical applications, it is crucial to have accurate and fast uncertainty estimation methods, able to convey the confidence of the DL model output since wrong predictions can jeopardise patients' safety. Predictive uncertainty of DL models can be decomposed into two parts: epistemic and aleatoric [7]. Epistemic uncertainty refers to the model's lack of knowledge or information. It typically arises from data sampling problems (e.g. patients with certain characteristics were not included during training) and is reducible by increasing the size of the training set. Aleatoric uncertainty describes uncertainty or "noise" inherent to the data itself. In the context of models in radiation therapy, it can be due to the quality of imaging technique, inter-observer and intra-observer variability in the contours and plan optimization.

There is a wide range of uncertainty quantification methods proposed in the literature [8–10], but the prominent epistemic uncertainty methods which have made it to clinical research environments are Monte Carlo Dropout [11–21] and Deep Ensembles [15,20,22–27]. A pitfall of these methods is their high inference time (i.e. multiple inference passes) which impedes their integration into adaptive treatment workflows. Recently, direct (i.e. single inference pass) methods are being investigated, like Evidential Deep Learning [28–32] and Conformal prediction, which achieve promising results and have strong statistical basis. Aleatoric uncertainty can be modelled by estimating the output likelihood [13,33–36], sampling in the original data space (Test-Time Augmentation [35–39]) or in latent space (joint latent space methods [40–44]). The output of these methods can be 3D uncertainty maps or a single aggregated value [13]. Reliable uncertainty maps provide voxel-wise uncertainty and can be used to apprehend the regions of uncertainty. A single uncertainty value can be used to flag slices or organs with uncertain segmentation and require the attention of a clinician to proceed with the treatment.

Although considerable research has delved into uncertainty quantification methods, only a limited number of studies have explored techniques for visually representing uncertainty estimates and effectively conveying them to clinicians. In 2021, Gillmann et al. provided a state-of-the-art analysis of uncertainty-aware imaging encompassing the whole medical imaging pipeline [45]. Within the context of image segmentation, maps based on probability and uncertainty values emerged as the prevailing visualisation technique. These maps are presented in various ways such as continuous or discrete, thresholded, grayscale or colour-coded, and overlaid on intensity information or segmentation [17,46–52]. Additionally, alternative methods employed confidence bands or confidence contours to direct clinicians' attention only to areas with specific confidence values [16,18]. Moreover, the single aggregated output has been employed as either an overall structure uncertainty score or as an indicator of uncertainty levels on a slice-by-slice basis [47,52].

Recognising clinicians as the end-users of such tools, we emphasise the importance of involving them in the investigative process to find the tool(s) that are clinically more beneficial, achieving a good trade-off between informativeness, usability, and time gain. Finding such a trade-off is utterly important since it has been reported that clinicians may not always be eager to see uncertainties if they are perceived as confusing [53]. This work aims to bridge the gap between theoretical uncertainty quantification and practical visualisation of such uncertainty by providing clinicians with a range of visualisation examples and evaluating their feedback.

## Materials and methods

The study was a collaborative effort of four institutions which participated in the ESTRO workshop: the Netherlands Cancer Institute, the Catharina Cancer Institute, and University Medical Centre Groningen, in the Netherlands, and Cliniques Universitaires Saint Luc – UCLouvain in Belgium. Two questionnaire sessions were conducted in the four institutions. Each session has been split into a panel of ROs and a panel of RTTs with a total of 16 participants (10 ROs and 6 RTTs). The questions were presented to the participants during the group sessions which lasted approximately one hour. To standardise the sessions, the set of questions were put forward in written form, and therefore was the same for all institutions. The presenters could answer the participants' questions and clarify unfamiliar concepts. Table 1 summarises the participants' information: their institution and participant's identification letters or numbers in the questionnaires, years of experience in radiation therapy in general and with automatic segmentation tools, and the

**Table 1**
Information of the participants in the study.

| Institution ID | Participant ID | Clinical experience [years] | Automatic segmentation experience [years] | Tumour sites for which auto-segmentation tools were used |
|---|---|---|---|---|
| A | RO1 | 10 | 3 | Prostate; bladder |
| A | RO2 | 5 | 2 | Liver; rectum |
| A | RO3 | 10 | 2 | Prostate, Bladder, Stomach, Neuro, Paediatric |
| A | RTT1 | 23 | 3 | Pelvis, Abdomen, Stomach |
| A | RTT2 | 9 | 3 | Pelvis, Abdomen, Stomach |
| A | RTT3 | 10 | 3 | Pelvis, Abdomen, Stomach |
| B | RO1 | 10 | 2 | Cervix, brachy |
| B | RO2 | 5 | 2 | Rectum CTV |
| B | RTT1 | 15 | 1.5 | Rectum CTV |
| B | RTT2 | 15 | 1 | Rectum CTV |
| C | RO1 | >15 | 2 | All OARs, breast (nodes + CTVs) |
| C | RO2 | 8 | 2 | All OARs, breast (nodes + CTVs) |
| C | RO3 | 9 | 1.5 | All OARs, breast (nodes + CTVs) |
| C | RO4 | 15 | 1 | All OARs, breast (nodes + CTVs) |
| C | RTT1 | 10 | 2 | All OARs, breast (nodes + CTVs) |
| C | RTT2 | >15 | 2 | All OARs, breast (nodes + CTVs) |
| C | RTT3 | 15 | 1 | All OARs, breast (nodes + CTVs) |
| D | RO1 | 11 | 5 | Head and neck |
| D | RTT1 | 6 | 5 | Head and neck |

tumour sites they have experience with.

*First session*

The goal of the first session was to grasp the current level of understanding of DL models within the clinic and learn their expectations and preferences regarding the integration of uncertainty estimation tools. We provided the following 3-item open-ended questionnaire:

Q1. What do you think are the main problems of using the DL segmentation models we now have for organs at risk (OARs) and target volumes in the clinic? And what do you think could solve this?

Q2. Do you think that information on the uncertainty of the segmentation would help you in the assessment of the DL segmentation?

Q3. What information would that be?

(a) Local uncertainty and/or global uncertainty? During the sessions, the terms local and global were described as follows: global refers to one value for the structure or the slice (whichever is displayed), whereas local provides a value of uncertainty for all individual voxels or a subset of voxels. Thereafter, we refer to these notions as structure-level, slice-level, and voxel-level uncertainty, respectively.

(b) Geometric and/or dosimetric?

*Second session*

Based on the preceding literature review and first session discussion outputs, we developed 10 visualisation options, referred to as “Cases”, of the uncertainty information, see Table 2. The uncertainty information referred to different levels (structure, slice, or voxel) and the values were binary or continuous. Cases 1 and 3 provided information on the structure level, binary and continuous, respectively. Cases 2, 4 and 9 provided slice-level uncertainty. Cases 5 and 6 provided voxel-level uncertainty information on the segmented contour itself, while Cases 7 and 8 included the surrounding regions. The binary envelope (Case 7) is a thresholded version of Case 8. The last two options provided not only the uncertainty of the segmentation model but also dosimetric information, represented in the form of isodose lines. We explicitly mentioned that the colours used in the visualisation Cases were just random examples and not part of the evaluation. The underlying image data (i.e. CT, CBCT, MR, …) would also still be visible.

The Cases were presented to the panels in written form with descriptions of the visualisation (See Table 2). Each individual participant was invited to evaluate the usability/interpretability, level of training required, and expected time gain, and provide an overall score (See evaluation grid Figure S.1 in the Supplementary materials). A distinction was made between using the proposed Cases during regular treatment planning or online adaptive workflows. The participants scored each criterion between 0 (difficult, requires training, no time gain expected) and 5 (easy, intuitive, fast) individually. Finally, the following binary questions were asked:

Q1. Would you like the system to present, on top of the flagging/highlighted region, an alternative segmentation (e.g. from another model) or would you prefer to edit what you already have?

Q2. (for ROs only) If that tool was available and working, would you let the RTTs perform corrections without the supervision and approval of a Radiation Oncologist?

Differences between ROs and RTTs, as well as answers for regular treatment planning and online adaptive workflows, were tested for statistically significance with an independent *t*-test or Mann-Whitney *U* test depending on the validation of the normality condition. The condition of normality was assessed using the test of D’Agostino and Pearson [54]. The significance threshold considered was 0.05 for the p-values.

**Results**

Regarding the general question about DL segmentation (Q1), the participants expressed their favourable opinion on using automatic tools in their daily workflow, despite encountering some issues for patients with particularities, such as a stent, or prosthesis. The participants recognized the challenge for models to perform effectively on these patients due to the need for more similar cases in the dataset. They advocated for future models to be developed within a framework that facilitates regular updates to the training set.

Regarding the questions focussing on uncertainty (Q2, Q3), most participants, despite having several years of experience with DL segmentations, were unfamiliar with the concept of uncertainty estimation. After receiving some clarifications, they unanimously agreed that uncertainty estimation of DL models would be beneficial to their clinical workflow. Participants suggested that uncertainty representation should be easily understandable and interpretable for all users. The importance of ensuring that all staff members interpret the uncertainty in a similar manner was highlighted. They also noted that the representation should vary depending on its use in the planning stage or during online adaptive radiotherapy. Overall, participants agreed that voxel-level uncertainty can be more useful at the planning stage, where the time constraint is less strong. In contrast, structure-level uncertainty is particularly valuable for speeding up online adaptive workflows, by helping to identify and discard regions that do not require review.

As a future goal it was mentioned to prioritise areas where delineation errors would have clinically significant effects. However, it was generally agreed that a good starting point would be to focus on geometric uncertainty in the segmentation of OARs and target volumes. Most participants indicated that adding the dosimetric information could result in information overload. The complete report of this first session from each institution is provided in Supplementary materials Table S.1.

From the second questionnaire session, we obtained the evaluation grids of each participant containing their ratings of the proposed visualisation options. These results are provided in Table S.2 in the Supplementary materials.

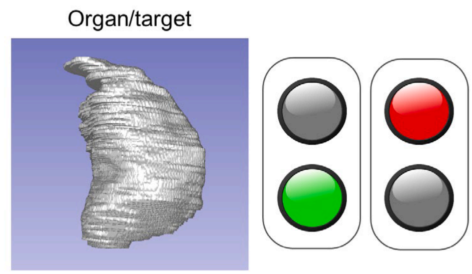
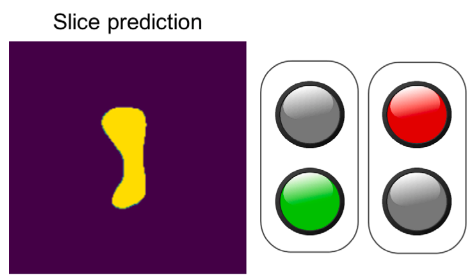
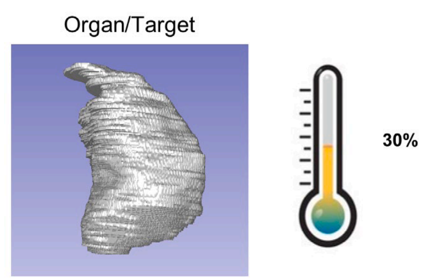
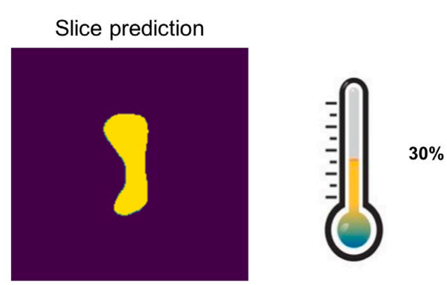
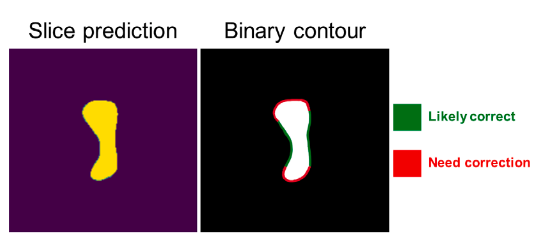
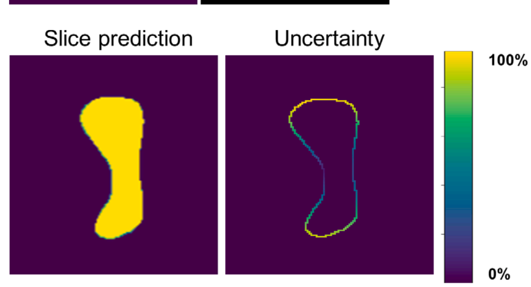
The hypothesis that the answers of RTTs and ROs followed the same distribution was tested with Mann-Whitney U tests and resulted in the acceptance of the hypothesis. Considering the answers of ROs and RTTs together, we tested if the answers provided for initial planning and online adaptive followed the same distribution with an independent *t*-test. P-values were above the threshold significance of 0.05 so there were no statistically significant differences. Bar chart plots for the different groups as well as the results of the statistical tests can be found in the Supplementary materials (Figures S.2-S.5 for the graphs and Table S.3 for the statistics).

According to the overall score provided by each participant, Case 5, with the binary flag on the segmentation contour was the preferred option with a mean score of 3.87. The following options are Case 6, Case 1, and Case 8 (3.20, 3.23 and 3.11 respectively). On the other hand, Cases 3, 4 and 9 achieved the worst scores with respective means of 1.73, 1.77 and 1.79, as shown in Fig. 1.

Fig. 2 is a bar chart displaying the mean of the four criteria (overall, usability, required training and expected time gain) for each Case. For these metrics, Case 5 and Case 1 have the highest scores. Scores attributed to expected time gain seem to be closer to the overall scoring compared with usability and required training. Cases displaying continuous information (Cases 3, 4, 6 and 8) seem to require more training (respective means of 2.17, 2.23, 2.67, 2.5) compared to Cases presenting binary values (Cases 1, 2, 5 and 7 with respective means of 3.73, 3.60, 3.53 and 2.60). The mean usability is higher than the other criteria for most of the Cases.

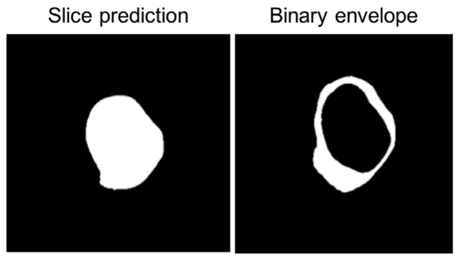
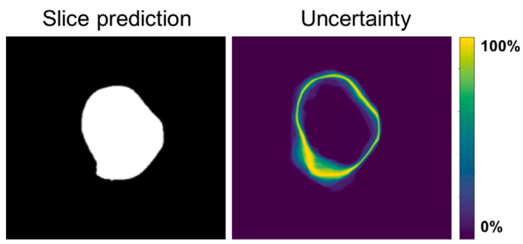
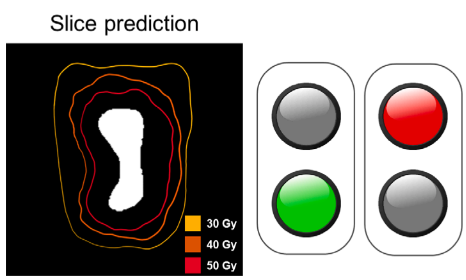
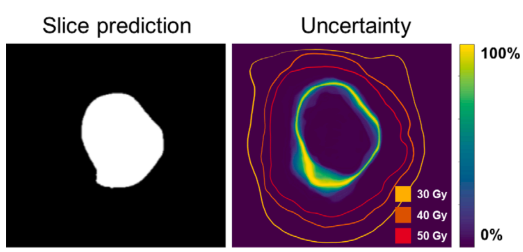
The topics discussed during this second session can be categorised into four main themes:

**Table 2**
Visualisation options of the uncertainty information presented to the panel of clinicians in the second session.

| Uncertainty visualisation option | Illustration | Description |
|---|---|---|
| C1: Binary traffic light (structure-level) |   | Outputs one binary value for the whole structure. Green means the model is confident in its prediction, whereas for red, it is less confident. |
| C2: Binary traffic light (slice-level) |   | Outputs one binary value per slice for the structure. Green means the model is confident in its prediction whereas, for red, it is less confident. |
| C3: Continuous indicator (structure-level) |   | Outputs one continuous value (illustrated by a percentage thermometer) for the whole structure. |
| C4: Continuous indicator (slice-level) |   | Outputs one continuous value (illustrated by a percentage thermometer) per slice for the structure. |
| C5: Binary contour (voxel-level) |   | The contour segmentation is divided into subcomponents displaying part of the contour where the model is confident (i.e. green) and where the model is not confident (i.e. red). This is done for each slice. |
| C6: Continuous contour (voxel-level) |   | The contour segmentation is coloured according to the uncertainty of the model, illustrated with a continuous percentage value. This is done for each slice. |

(continued on next page)

**Table 2** (*continued*)

| Uncertainty visualisation option | Illustration | Description |
|---|---|---|
| C7: Binary envelope (voxel-level) |   | Uncertainty map displaying in white the region setting a threshold on the uncertainty (right). This is the binary version of Case 8. This is done for each slice. |
| C8: Continuous envelope (voxel-level) |   | Uncertainty map displaying the uncertainty percentage of the prediction as a continuous value (right). This is done for each slice. |
| C9: Binary traffic lights (slice-level) with dosimetric information |   | Uncertainty is displayed in the same way as Case 2: one binary value per slice for the structure. There is additional dosimetric information displayed with isodose lines (levels arbitrarily chosen here). |
| C10: Continuous envelope (voxel-level) with dosimetric information |   | Uncertainty is displayed in the same way as Case 8: per slice percentage of the prediction as a continuous value. There is additional dosimetric information displayed with isodose lines (levels arbitrarily chosen here). |

– Binary vs continuous information:

Whether the information displayed should be binary like in the flagging system of Cases 1, 2, and 9, the binary contour Case 5, or binary envelope Case 7, or their counterparts with continuous information. Many clinicians argue that they have a "binary mind" and therefore prefer the simplicity of binary information, whereas continuous information would leave room for interpretation. If the percentage methods are used, then they would want a consensus on the threshold above which the models can be trusted (which comes back to a binary output). Otherwise, using the raw percentage without a consensual threshold might lead to large inter- and intra-observer variability. Moreover, percentage values will require more training to understand the meaning behind the numbers. The colour code will also have an impact on the visualisation and should be selected carefully.

– Structure-level vs slice-level vs voxel-level information:

Time gain is expected to be higher with a green flag for the whole structure (Case 1) rather than green flags for each slice (Case 3) as the clinicians will still need to go through every slice of the structure. On the other hand, if the structure gets a red flag, then it could be useful to get voxel-level uncertainty information such as in Case 5. This led the sessions in multiple institutions to conclude that they prefer subsequent information: first a structure-level flagging, and if it is red, then display a voxel-level uncertainty information.

– Dosimetric information:

Opinions on adding dosimetric information varied. Some clinicians viewed it negatively, citing the excessive information it presented, while others viewed it positively, noting its potential to help eliminate dosimetrically irrelevant errors (e.g. very low dose regions). The latter is especially true for an online adaptive workflow where time is crucial and dosimetric information can help prioritise the edits to perform. Some participants suggested masking the areas with low dosimetric impact instead of displaying isodoses. Another suggestion involved incorporating the dosimetric relevance in the binary flag. For example, even though the network is unsure about the contour of an OAR but editing will have no dosimetric impact, then the flag may still be green.

– Alternative segmentation proposed by other models:

Most clinicians are in favour of alternatives, but it should be for the whole structure, not per-slice. Ideally, the system should automatically select the winner instead of letting the clinician choose, which would be

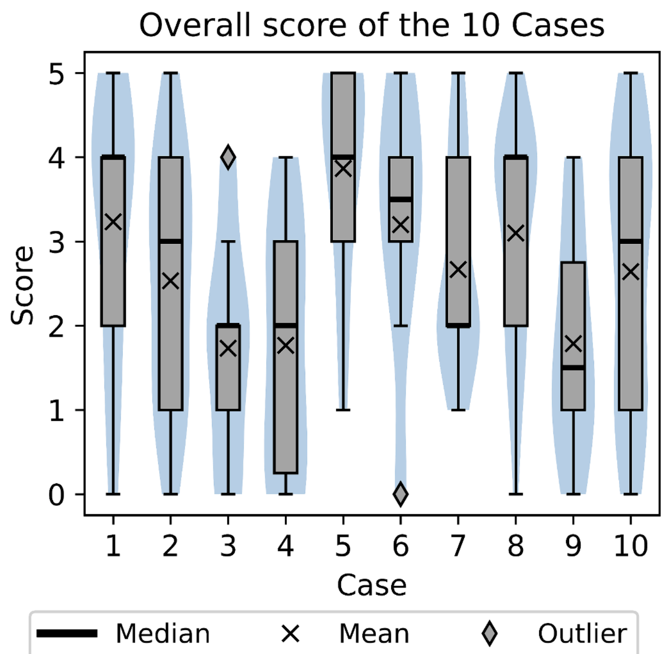



**Fig. 1.** Boxplot of the overall scores given by all participants. Violin added in the background to show the distribution of data per Case.

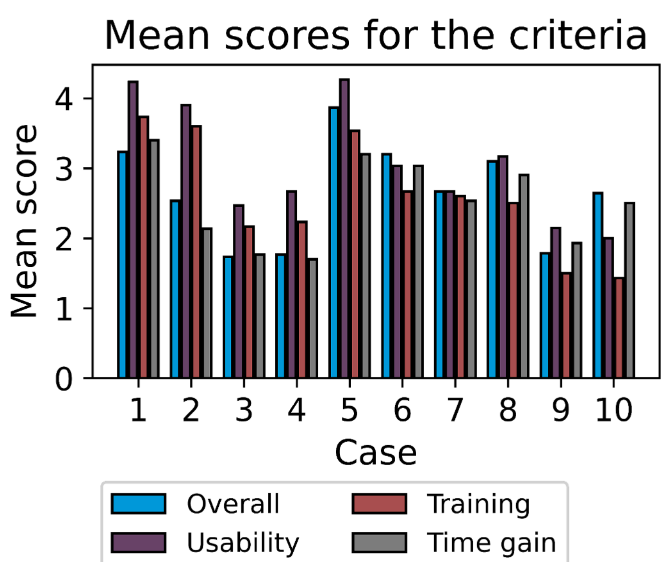



**Fig. 2.** Bar chart visualising the mean of the scores given by all participants.

a waste of time.

**Discussion**

Through the initial questionnaire, we confirmed that clinicians in radiation therapy are interested in tools that estimate DL segmentation uncertainty. However, the method for presenting this uncertainty information was not straightforward and warranted further investigation. We devised a range of visualisation examples that were presented to the teams of clinicians in a second questionnaire session and evaluated their feedback. Interestingly, the scores provided by RTTs and ROs followed the same distribution according to our statistical analysis. In general, representations with binary values were preferred, since a green or red flag is easier to interpret and leaves less room for interpretation compared with a continuous, provided that the binary thresholds are set following a clinical consensus. One of the centres did prefer the continuous visualisation, but also argued that a consensus threshold value is required before using such tools. While binary information is easier for ROs and RTTs to understand and reduces interobserver variability, continuous information on voxel-based uncertainty could be helpful in specific clinical cases where higher uncertainty is reported such as identifying patients that are out-of-distribution for the segmentation model at hand and select an alternative.

The discussions in the different institutions converged toward a combination of representation levels. First, a structure-level representation with binary flag to rapidly sort the structures requiring attention which could be followed by more detailed voxel-level information. Including dosimetric information was generally considered to be too complex but might be of value in an online adaptive workflow. A feature suggested by participants to reduce the overload of information was to mask uncertainty regions. For instance, clinicians might not need to visualise segmentation uncertainty in areas with low dosimetric impact. The way to display this mask was not investigated in this study but a solution could be to incorporate the dosimetric relevance in the binary flag. For example, even though the network is unsure about the contour of an OAR but editing will have no dosimetric impact, then the flag may still be green. This masking feature, along with the idea of combining different levels of information seem to be promising for future implementations of such systems.

Our results are mostly in line with previous work and build thereupon.

Gillman et al. [45] used three criteria to score uncertainty visualisations: (1) whether the underlying image is shown; (2) whether the segmentation uncertainty is displayed and (3) whether the cognitive workload is kept minimal. Criteria 1 and 2 were fulfilled for all our Cases. Our results reflected that the more complex the visualisation (such as adding isodoses to the representation of Cases 9 and 10), the lower the usability and interpretability values.

Maruccio et al. [17] presented cases very similar to our Case 8 to two clinicians and found a correlation between the level of uncertainty perceived by the clinicians and two uncertainty metrics. Our work builds on that as they discussed a study with more observers and also assessing the potential time saved would be needed.

Balagopal et al. [16] used the method we presented as Case 7 to present the uncertainty of post-operative prostate CTVs. They could show their network predicted contours in better agreement with physician contours than the contours of the residents. They concluded such confidence bands could streamline the contouring workflow.

Van Rooij et al. [18] also used uncertainty thresholds and displayed *iso*-probability dashed segmentation lines. They suggested that checking all OARs using this visualisation tool might be too time-consuming. Some options to overcome this were suggested: the clinician selects which OAR he/she would like to see, or only particular areas are highlighted based on the distance between the *iso*-probability lines. Also, using a standard threshold in which OARs would be shown could be an option. A similar approach could be foreseen for Case 5, which was favoured in our study.

Our study also has some limitations. The ten Cases presented to the participants focused on the type of uncertainty visualisation, like its level (structure, slice, or voxel) or its format (binary versus continuous, dosimetric versus geometric only). However, as suggested by the participants, a lot of other factors in the display will have an impact on the interpretation of uncertainty, such as the choice of thresholds for binary metrics, the colour codes, opacity and line thickness or styles.

When analysing the data, we noticed that some participants, sometimes clustered within institutes, seemed to have very strong opinions, grading only with extreme scores of 0 or 5, while others were more moderate and nuanced in their answers. Whether this truly reflects relevant differences between institutions is impossible to say, as they could also reflect individual differences. It is noteworthy that while we obtained individual evaluation grids, some participants were actually discussing the cases with each other, which may have influenced the results. We designed a variation mitigation strategy to set up the study and standardise as many aspects as possible. The study setup was designed by a multidisciplinary team of clinicians and researchers, including an expert in Cognitive Psychology, Human Factors and Artificial Intelligence. It included frequent meetings with all researchers involved to ensure standardised execution of the study, ways to present the cases to the participants, response collection and analysis of the data. To avoid that different experimenters explained the cases in different ways to the participants, descriptions of the visualisation cases were standardised and presented in written form in all institutions. Moreover, data processing and data analysis was done after data from all institutions were collected in one data file.

Future work may include focusing on the preferred Cases from this study and further refinement of the implementation preferences, as well as exploring practical use cases (i.e. with real patients) that can give us quantitative results about the time gain for each option, especially in an adaptive workflow. In particular, it is worth investigating the possible combinations of Cases that might be of additional value.

## Conclusion

Uncertainty quantification tools for DL segmentation models are becoming more and more popular in the research community, but their translation into clinical practice and within commercial treatment planning systems is still very prospective. Our exploratory study confirmed the interest of both radiation oncologists and radiation therapists in having these tools in their clinical routine to ease and speed up the delineation process, and provided valuable feedback on the format of such tools. After evaluating uncertainty visualisation cases at different levels (structure, slice or voxel) with binary or continuous values, participants expressed their preference for the binary voxel-level uncertainty options in the first place, followed by binary structure-level uncertainty. Expressing uncertainty as a continuous value was disliked by most participants. Adding dosimetric information was perceived differently according to each participant but was considered helpful, particularly in online adaptive workflows.

## CRediT authorship contribution statement

**M. Huet-Dastarac:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation. **N.M. C. van Acht:** Writing – original draft, Methodology, Investigation, Formal analysis, Data curation. **F.C. Maruccio:** Writing – original draft, Methodology, Investigation, Formal analysis, Data curation. **J.E. van Aalst:** Writing – review & editing, Investigation. **J.C.J. van Oorschodt:** Writing – review & editing, Investigation. **F. Cnossen:** Writing – review & editing, Methodology. **T.M. Janssen:** Writing – review & editing, Methodology, Investigation, Conceptualization. **C.L. Brouwer:** Writing – review & editing, Supervision, Methodology, Investigation, Conceptualization. **A. Barragan Montero:** Writing – review & editing, Supervision, Methodology, Investigation, Conceptualization. **C.W. Hurkmans:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Formal analysis, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

We acknowledge all the valuable discussions of all the participants in the sessions.

Margerie Huet-Dastarac is funded by the Walloon region in Belgium (PROTHERWAL/CHARP, grant 7289). Ana Barragan is funded by Fonds spécial de recherche (FSR) from UCLouvain. Tomas Janssen and Federica Maruccio reported institutional funding by Elekta AB. All remaining authors have declared no conflicts of interest.

## Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.radonc.2024.110545.

## References

[1] Heilemann G, Buschmann M, Lechner W, Dick V, Eckert F, Heilmann M, et al. Clinical implementation and evaluation of auto-segmentation tools for multi-site contouring in radiotherapy. Phys Imaging Radiat Oncol 2023;28:100515. https://doi.org/10.1016/j.phro.2023.100515.

[2] Rong Y, Chen Q, Fu Y, Yang X, Al-Hallaq HA, Wu QJ, et al. NRG oncology assessment of artificial intelligence deep learning-based auto-segmentation for radiation therapy: current developments, clinical considerations, and future directions. Int J Radiation Oncol, Bio, Phy 2024;119:261–80. https://doi.org/10.1016/j.ijrobp.2023.10.033.

[3] Ginn JS, Gay HA, Hilliard J, Shah J, Mistry N, Möhler C, et al. A clinical and time savings evaluation of a deep learning automatic contouring algorithm. Med Dosim 2023;48:55–60. https://doi.org/10.1016/j.meddos.2022.11.001.

[4] van der Veen J, Willems S, Bollen H, Maes F, Nuyts S. Deep learning for elective neck delineation: more consistent and time efficient. Radiother Oncol 2020;153: 180–8. https://doi.org/10.1016/j.radonc.2020.10.007.

[5] van der Veen J, Willems S, Deschuymer S, Robben D, Crijns W, Maes F, et al. Benefits of deep learning for delineation of organs at risk in head and neck cancer. Radiother Oncol 2019;138:68–74. https://doi.org/10.1016/j.radonc.2019.05.010.

[6] Bakx N, Rijkaart D, van der Sangen M, Theuws J, van der Toorn P-P, Verrijssen A-S, et al. Clinical evaluation of a deep learning segmentation model including manual adjustments afterwards for locally advanced breast cancer. Tech Innov Patient Support Radiat Oncol 2023;26:100211. https://doi.org/10.1016/j.tipsro.2023.100211.

[7] Kendall A, Gal Y. What uncertainties do we need in bayesian deep learning for computer vision? Adv Neural Inf Process Syst 2017;30:Available. https://proceedings.neurips.cc/paper/2017/hash/2650d6089a6d640c5e85b2b88265dc2b-Abstract.html.

[8] Lambert B, Forbes F, Doyle S, Dehaene H, Dojat M. Trustworthy clinical AI solutions: a unified review of uncertainty quantification in Deep Learning models for medical image analysis. Artif Intell Med 2024;150:102830. https://doi.org/10.1016/j.artmed.2024.102830.

[9] Barragán-Montero A, Bibal A, Dastarac MH, Draguet C, Valdés G, Nguyen D, et al. Towards a safe and efficient clinical implementation of machine learning in radiation oncology by exploring model interpretability, explainability and data-model dependency. Phys Med Biol 2022;67. https://doi.org/10.1088/1361-6560/ac678a.

[10] Zou K, Chen Z, Yuan X, Shen X, Wang M, Fu H. A review of uncertainty estimation and its application in medical imaging. Meta-Radiology 2023;1:100003. https://doi.org/10.1016/j.metrad.2023.100003.

[11] Roy AG, Conjeti S, Navab N, Wachinger C. Inherent Brain Segmentation Quality Control from Fully ConvNet Monte Carlo Sampling. Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. Springer International Publishing; 2018. pp. 664–672. doi:10.1007/978-3-030-00928-1_75.

[12] Camarasa R, Bos D, Hendrikse J, Nederkoorn P, Kooi E, van der Lugt A, et al. In: Quantitative Comparison of Monte-Carlo Dropout Uncertainty Measures for Multi-class Segmentation. Springer International Publishing; 2020. p. 32–41. https://doi.org/10.1007/978-3-030-60365-6_4.

[13] Jungo A, Balsiger F, Reyes M. Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. Front Neurosci 2020. https://doi.org/10.3389/fnins.2020.00282.

[14] Wang G, Li W, Ourselin S, Vercauteren T. Automatic brain tumor segmentation based on cascaded convolutional neural networks with uncertainty estimation. Front Comput Neurosci 2019;13:56. https://doi.org/10.3389/fncom.2019.00056.

[15] Ng M, Guo F, Biswas L, Petersen SE, Piechnik SK, Neubauer S, et al. Estimating uncertainty in neural networks for cardiac MRI segmentation: a benchmark study. IEEE Trans Biomed Eng 2023;70:1955–66. https://doi.org/10.1109/TBME.2022.3232730.

[16] Balagopal A, Nguyen D, Morgan H, Weng Y, Dohopolski M, Lin M-H, et al. A deep learning-based framework for segmenting invisible clinical target volumes with estimated uncertainties for post-operative prostate cancer radiotherapy. Med Image Anal 2021;72:102101. https://doi.org/10.1016/j.media.2021.102101.

[17] Maruccio FC, Eppinga W, Laves M-H, Navarro RF, Salvi M, Molinari F, et al. Clinical assessment of deep learning-based uncertainty maps in lung cancer segmentation. Phys Med Biol 2024:69. https://doi.org/10.1088/1361-6560/ad1a26.

[18] van Rooij W, Verbakel WF, Slotman BJ, Dahele M. Using spatial probability maps to highlight potential inaccuracies in deep learning-based contours: facilitating online adaptive radiation therapy. Adv Radiat Oncol 2021;6:100658. https://doi.org/10.1016/j.adro.2021.100658.

[19] Sahlsten J, Jaskari J, Wahid KA, Ahmed S, Glerean E, He R, et al. Application of simultaneous uncertainty quantification for image segmentation with probabilistic deep learning: Performance benchmarking of oropharyngeal cancer target delineation as a use-case. medRxiv 2023. https://doi.org/10.1101/2023.02.20.23286188. p. 2023.02.20.23286188.
[20] Bhat I, Pluim JPW, Viergever MA, Kuijf HJ. Influence of uncertainty estimation techniques on false-positive reduction in liver lesion detection. J Mach Learn Biomed Imaging 2022;1:1–33. https://doi.org/10.59275/j.melba.2022-5937.
[21] Ghoshal B, Tucker A, Sanghera B, Wong WL. Estimating uncertainty in deep learning for reporting confidence to clinicians in medical image segmentation and diseases detection. Comput Intell 2021:701–34. https://doi.org/10.1111/coin.12411.
[22] Mehrtash A, Wells WM, Tempany CM, Abolmaesumi P, Kapur T. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. IEEE Trans Med Imaging 2020;39:3868–78. https://doi.org/10.1109/TMI.2020.3006437.
[23] Roshan S, Tanha J, Zarrin M, Babaei AF, Nikkhah H, Jafari Z. A deep ensemble medical image segmentation with novel sampling method and loss function. Comput Biol Med 2024;172:108305. https://doi.org/10.1016/j.compbiomed.2024.108305.
[24] Zheng H, Zhang Y, Yang L, Liang P, Zhao Z, Wang C, et al. A new ensemble learning framework for 3D biomedical image segmentation. AAAI 2019;33:5909–16. https://doi.org/10.1609/aaai.v33i01.33015909.
[25] Dang T, Nguyen TT, McCall J, Elyan E, Moreno-García CF. Two-layer ensemble of deep learning models for medical image segmentation. Cognit Comput 2024. https://doi.org/10.1007/s12559-024-10257-5.
[26] Georgescu M-I, Ionescu RT, Miron AI. In: Diversity-Promoting Ensemble for Medical Image Segmentation. New York, NY, USA: Association for Computing Machinery; 2023. p. 599–606. https://doi.org/10.1145/3555776.3577682.
[27] Buddenkotte T, Escudero Sanchez L, Crispin-Ortuzar M, Woitek R, McCague C, Brenton JD, et al. Calibrating ensembles for scalable uncertainty quantification in deep learning-based medical image segmentation. Comput Biol Med 2023;163:107096. https://doi.org/10.1016/j.compbiomed.2023.107096.
[28] Huang L, Ruan S, Belief DT, Segmentation F-B-S. IEEE 18th International Symposium on Biomedical Imaging (ISBI). IEEE 2021;2021:160–4. https://doi.org/10.1109/ISBI48211.2021.9433885.
[29] Zou K, Yuan X, Shen X, Wang M, Fu H. TBraTS: Trusted Brain Tumor Segmentation. Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. Springer Nature Switzerland; 2022. pp. 503–513. doi:10.1007/978-3-031-16452-1_48.
[30] Huang L, Ruan S, Decazes P, Denœux T. Evidential segmentation of 3D PET/CT images. Belief Functions: Theory and Applications Springer International Publishing 2021:159–67. https://doi.org/10.1007/978-3-030-88601-1_16.
[31] Li H, Nan Y, Del Ser J, Yang G. Region-based evidential deep learning to quantify uncertainty and improve robustness of brain tumor segmentation. Neural Comput & Applic 2023;35:22071–85. https://doi.org/10.1007/s00521-022-08016-4.
[32] Chen Y, Yang Z, Shen C, Wang Z, Zhang Z, Qin Y, et al. Evidence-based uncertainty-aware semi-supervised medical image segmentation. Comput Biol Med 2024;170:108004. https://doi.org/10.1016/j.compbiomed.2024.108004.
[33] Murugesan B, Liu B, Galdran A, Ayed IB, Dolz J. Calibrating segmentation networks with margin-based label smoothing. Med Image Anal 2023;87:102826. https://doi.org/10.1016/j.media.2023.102826.
[34] Rodríguez Outeiral R, Ferreira Silvério N, González PJ, Schaake EE, Janssen T, van der Heide UA, et al. A network score-based metric to optimize the quality assurance of automatic radiotherapy target segmentations. Phys Imaging Radiat Oncol 2023;28:100500. https://doi.org/10.1016/j.phro.2023.100500.
[35] González C, Gotkowski K, Fuchs M, Bucher A, Dadras A, Fischbach R, et al. Distance-based detection of out-of-distribution silent failures for Covid-19 lung lesion segmentation. Med Image Anal 2022;82:102596. https://doi.org/10.1016/j.media.2022.102596.
[36] Diao Z, Jiang H, Shi T. A unified uncertainty network for tumor segmentation using uncertainty cross entropy loss and prototype similarity. Knowl-Based Syst 2022;246:108739. https://doi.org/10.1016/j.knosys.2022.108739.
[37] Wang G, Li W, Aertsen M, Deprest J, Ourselin S, Vercauteren T. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing 2019;335:34–45. https://doi.org/10.1016/j.neucom.2019.01.103.
[38] Ayhan MS, Berens P. Test-time Data Augmentation for Estimation of Heteroscedastic Aleatoric Uncertainty in Deep Neural Networks. Medical Imaging with Deep Learning. 2018. Available: https://openreview.net/forum?id=rJZz-knjz.
[39] Ballestar LM, Vilaplana V. MRI Brain Tumor Segmentation and Uncertainty Estimation Using 3D-UNet Architectures. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer International Publishing; 2021. p. 376–90. https://doi.org/10.1007/978-3-030-72084-1_34.
[40] Wang T, Lu J, Lai Z, Wen J, Kong H. Uncertainty-Guided Pixel Contrastive Learning for Semi-Supervised Medical Image Segmentation. Thirty-First International Joint Conference on Artificial Intelligence. 2022. pp. 1444–1450. doi:10.24963/ijcai.2022/201.
[41] Judge T, Bernard O, Porumb M, Chartsias A, Beqiri A, Jodoin P-M. CRISP - Reliable Uncertainty Estimation for Medical Image Segmentation. Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. Springer Nature Switzerland; 2022. pp. 492–502. doi:10.1007/978-3-031-16452-1_47.
[42] Baumgartner CF, Tezcan KC, Chaitanya K, Hötker AM, Muehlematter UJ, Schawkat K, et al. PHiSeg: Capturing Uncertainty in Medical Image Segmentation. Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Springer International Publishing; 2019. pp. 119–127. doi:10.1007/978-3-030-32245-8_14.
[43] Kohl SAA, Romera-Paredes B, Meyer C, Fauw JD, Ledsam JR, Maier-Hein KH, et al. In: A probabilistic U-net for segmentation of ambiguous images. Red Hook, NY, USA: Curran Associates Inc.; 2018. p. 6965–75.
[44] Li X, Bagher-Ebadian H, Gardner S, Kim J, Elshaikh M, Movsas B, et al. An uncertainty-aware deep learning architecture with outlier mitigation for prostate gland segmentation in radiotherapy treatment planning. Med Phys 2023;50:311–22. https://doi.org/10.1002/mp.15982.
[45] Gillmann C, Saur D, Scheuermann G. How to deal with uncertainty in machine learning for medical imaging? 2021 IEEE Workshop on TRust and EXpertise in Visual Analytics (TREX). IEEE; 2021. doi:10.1109/trex53765.2021.00014.
[46] Al-Taie A, Hahn HK, Linsen L. Uncertainty estimation and visualization in probabilistic segmentation. Comput Graph 2014;39:48–59. https://doi.org/10.1016/j.cag.2013.10.012.
[47] Al-Taie A, Hahn HK, Linsen L. Uncertainty estimation and visualization for multi-modal image segmentation. Eurographics Association 2015. https://doi.org/10.2312/vcbm.20151205.
[48] Gillmann C, Saur D, Wischgoll T, Hoffmann K-T, Hagen H, Maciejewski R, et al. Uncertainty-aware brain lesion visualization. The Eurographics Association 2020. https://doi.org/10.2312/VCBM.20201176.
[49] Nair T, Precup D, Arnold DL, Arbel T. Exploring uncertainty measures in deep networks for Multiple sclerosis lesion detection and segmentation. Med Image Anal 2020;59:101557. https://doi.org/10.1016/j.media.2019.101557.
[50] Kwon Y, Won J-H, Kim BJ, Paik MC. Uncertainty quantification using Bayesian neural networks in classification: application to biomedical image segmentation. Comput Stat Data Anal 2020;142:106816. https://doi.org/10.1016/j.csda.2019.106816.
[51] Saad A, Hamarneh G, Möller T. Exploration and visualization of segmentation uncertainty using shape and appearance prior information. IEEE Trans Vis Comput Graph 2010;16:1366–75. https://doi.org/10.1109/TVCG.2010.152.
[52] Musleh M, Muren LP, Toussaint L, Vestergaard A, Gröller E, Raidou RG. Uncertainty guidance in proton therapy planning visualization. Comput Graph 2023;111:166–79. https://doi.org/10.1016/j.cag.2023.02.002.
[53] Gillmann C, Smit NN, Groller E, Preim B, Vilanova A, Wischgoll T, et al. Ten open challenges in medical visualization. IEEE Comput Graph Appl 2021;41:7–15. https://doi.org/10.1109/MCG.2021.3094858.
[54] D'Agostino R, Pearson ES. Tests for departure from normality. empirical results for the distributions of $b_2$ and $\sqrt{b_1}$. Biometrika 1973;60:613–22. https://doi.org/10.2307/2335012.