Editorial

Recommendations on how to establish evidence from auto-segmentation software in radiotherapy

Vincenzo Valentini, Luca Boldrini, Andrea Damiani, Ludvig P. Muren

Along with the improved treatment conformity achieved with the recently implemented radiotherapy (RT) planning and delivery approaches [1], there is growing awareness of the uncertainties connected to the definition and delineation of the RT targets as well as the organs at risk (ORs). To assure accurate reproducibility of the planned treatment in order to avoid ‘geographical miss’ of the target, it is mandatory to correctly identify the volumes to be irradiated and to appropriately manage these uncertainties in the definition of the gross target volume (GTV) as well as of standardized, disease-specific, nodal clinical target volumes (CTVs).

Accurate segmentation of primarily the GTVs and CTVs, as well as the ORs, therefore represents the foundation for successful RT. International or institutional guidelines, contouring atlases, case libraries and numerous recommendations have thus recently been developed [2–12]. Atlases and guidelines are widely recognized and contribute to reducing inter-observer variability, but they are static documents that also lack interactivity.

To address this challenge, several commercial auto-contouring software solutions have recently been released, representing an opportunity for individualizing the existing atlases, automatically propagating clinically reliable contours to patients’ specific anatomies [13]. They also have potential for lowering the segmentation time, increasing the adherence to existing guidelines and, not the least, to reduce the inter-observer variability that still may be a major source of uncertainty in RT. The current issue presents several studies related to the use of such software in RT [14–16].

A basic requirement of the application of auto-segmentation software is the existence of a widely recognized and reliable definition of the volumes, according to the anatomical region, the histology as well as the stage of the disease. This can be described as software ontology [17], covering clinical, anatomical, pathological and imaging information. This ontology should be linked to benchmark performance values obtained through comparisons with the agreement between multiple observers/operators, or between manual delineations and auto-contouring. There are several metrics for the agreement between the contours, and establishing ranges indicating clinically acceptable agreement on the scales of these metrics plays an essential role in translating the obtained observations to everyday practice.

In this paper we would like to share comments and criticisms on how evidence can be derived from the use of auto-segmentation software, addressing the key aspects that should be considered in future studies in this field: ontology definition, benchmark evaluation methods and performance evaluation tools. We also discuss the potential benefits that can be achieved with these tools.

Ontology definition

The term ontology refers to a form of dictionary where information is specified and organized in a well-defined semantic data collection model; a set of concepts within a domain, and also the relationships between those concepts. In the auto-contouring setting the ontology is therefore represented by the prior reference knowledge regarding GTV and CTV definition, in practise related to the existence of contouring guidelines and the adherence to these.

The segmentation of the GTV relies on interpretation of multimodal imaging data (e.g. CT, MRI, PET), often involving the use of specific image registration and segmentation tools [18]. GTV segmentation is frequently performed in close consultation with other medical disciplines (e.g. radiologists, surgeons). Delineation of the nodal CTV reflects the lymphatic drainage of the disease site and its relapse risk, which are generally identified through retrospective pathological and/or surgical observations performed for the specific stage of disease. Prospective randomized studies of the effect of lymph node irradiation, such as in the RTOG 9413 study of prostate cancer with or without pelvic irradiation [19], still remains uncommon. Areas of nodal sub-volumes of different risks (both for lymphatic spread and/or relapse) are generally described by contouring atlases (with related text descriptions), defining their boundaries and anatomical relations. Together with the interpretation of the TNM stage of the disease [20] the evaluation of nodal involvement will dictate which sub-volumes should be included in the nodal CTV [2,3,5]. Reliable definition and segmentation of the normal tissues has also considerable dose/volume and clinical consequences, in particular for dose-limiting ORs [21].
For all main RT tumour sites there have been several ontologies (i.e. delineation guidelines) suggested, yet there are seldom any criteria to define which represents the most accurate and clinically reliable standard for each considered anatomical district. A first attempt of a hierarchical organization of these proposals was the endorsement of the proposed contours by national and international scientific societies but, having prostate as an example, at least eight atlases with the endorsement of four different organisations can be counted today, demonstrating a substantial lack of harmony [22–29].

Dealing with auto-contouring software, it is necessary to understand how the auto-segmentation system propagates the chosen ontology on the single case, e.g. by the adoption of a commonly agreed atlas delineation by the institution’s organ oriented expert team, or through the creation of an atlas through selection of cases uploaded into a library system available in an atlas-based contouring system [17,30].

Performance evaluation tools

The performance of auto-segmentation software is usually evaluated with respect to contour similarity indices as well as potential time/workload savings. So far a number of approaches have been used to quantify the amount of similarity between the proposed contours and others. Analysis and comparison among the five main methods (area of intersection, Dice similarity index, Jaccard index, Conformation number, Hausdorff distance) found that they are not effective in distinguishing random from systematic errors, or to separate false positives from false negatives [17,31–34]. They also operate “out of context”, without any consideration to the potential effects of a contouring error. What the measures do quantify is how close a given contour is to an ‘ideal’ contour. This is suitable for a first level analysis and is useful as a comparison tool, especially if a combination of a dimensionless index and a measure such as the Hausdorff distance is used; this is even more appropriate if the attempted contour is sufficiently close to its reference.

The Dice index is linear with respect to the intersection, while the other conformation indices are essentially non-linear, showing instead a tendency to amplify the effect of a similar degree of discrepancy when the intersection is larger. This does not pose any problem (the monotonicity is assured), however, the non-uniformity should have an adequate justification from a clinical and/or educational point of view.

A combination of conformation scores, metric elements and clinical risk assessment could lead to a new class of indices, which could prove useful at first in an educational environment evaluating a “test” contour against a reference created by a domain expert.

Closely related to the benchmark contouring definition is the measurement of the contouring time spent, to quantify the time saved with the auto-segmentation approach. It is important to measure the time spent during all components of the process, and to include the time spent in the definition of the atlas cases and in making the right atlas case choice (manual or automatic).

Benchmark evaluation methods

So far only a few studies have described benchmark values in various anatomical regions, often discussing these in relation to the reduction in segmentation time [2–8,17]. These studies usually compare the performance of auto-segmentation tools versus a reference segmentation delineated by one of more human observers. In such settings, the agreement between multiple human segmentations represents the benchmark value the auto-contoured structures should be compared with.

The auto-segmentation studies have reported Dice similarity index values well below unity in all considered anatomical regions, ranging from 0.67 to 0.79 in the anorectal region to 0.86–0.91 in the breast [3]. These values should be evaluated relative to comparisons between contours drawn manually by experts, where the existing variability shows similar values [6,17,32,35].

Struggling to obtain a complete overlap between two manually contoured structures could therefore be deceptive and can lead to dead-end streets in evaluating inter-observer studies, as the existence of a certain degree of variability will remain unavoidable in daily clinical practice.

This translates into the necessity of building up a reliable “gold structure set” which will represent the unique benchmark of the study and will be the referral contour to which all other contours should be compared with. There are several strategies that can be followed when defining these benchmark structures: a commonly agreed delineation by two or more operators with the same expertise; a gold contour coming from the experience of one highly skilled physician or from a team of experts of the chosen anatomical district. Each of these strategies has its own pros and cons: trusting a single operator for the gold contour delineation is quick and easy but lacks the fundamental independent cross check of the structures with other operators, while enrolling a team of experts represents a heavier effort but it could offer a more reliable segmentation.

Discussion

The issues covered in the three previous sections can be summarized into a set of recommendations (Table 1) that should be considered in future studies of auto-segmentation issues.

The benefits expected from the use of auto-contouring software are several, and cover aspects related to clinical impact, the resulting dose distributions as well as education. The correct definition of RT targets and ORs has direct clinical consequences, with respect to both disease control and toxicity. In particular the contouring of nodal CTV sub-volumes is critical, and auto-contouring software can be very useful for this purpose. This is also related to the increasing importance of correlating CTV sub-volumes (and critical choices in this setting) with (imaging) data of local control and survival [36]. Such studies will provide essential information on which structures to include in the auto-contouring atlases and templates. This information is also key input to individualized and adaptive approaches to prediction databases and models [37,38].

Auto-contouring software has clearly a potential to play an important role in future RT, based on sound clinical evidence and linking survival to volume choice and their automatic propagation. However, building up clinical evidence is time consuming (not the least within a prospective setting), and there are considerable future challenges in constructing reliable predictive models [39–41]. Quantifying the potential clinical benefits of auto-contouring in a retrospective way could be a first step in this direction. In any case it should be underlined that present auto-contouring tools are merely a supplementary instrument that should be used together with careful manual editing of the proposed structures.

Besides the simple geometrical comparison, it is useful to explore the relation between auto-segmentation and dose/volume parameters relating to target dose coverage and/or normal tissue irradiation. Voet et al. [42] recently observed that moderate geometrical differences in the PTV may have a large dosimetric impact on the target. Specifically, they observed reductions in D99 up to 11 Gy even for Dice coefficients higher than 0.8 and mean contour distances lower than 1 mm between the autocon-
toured volumes and the manual referral ones. In a paper in this issue, Jameson and colleagues showed how target delineation uncertainties can also be transformed into tumour control probability predictions [14]. Also in this issue, Conson et al. reported on the dose/volume effects of the use of auto-segmentation for normal tissues in the brain [15]. It is therefore necessary to develop this field further to fully describe the existing correlation between the considered similarity indices and dose/volume effects [43]. This must still be considered along with what we refer to as the “benchmark trap” [17]: are we confident that the daily “human” inter-observer variability could show better performances, in terms of dose and volumes, when considering a comparison with the same manual benchmark? Further studies are required in order to characterize the clinical benefit with respect to dose/volume endpoints.

Another component of the benefit offered by the use of auto-contouring software is within the educational dimension, allowing an improvement of the adherence to contouring guidelines for personnel in training, reducing the inter/intra-observer variability and improving the reliability of contouring. Residents/trainees actually learn how to segment using paper atlases, consulting guidelines or referring directly to older and more experienced personnel. The use of auto-contouring software can represent a useful, efficient tool to test oneself with expert drawn structure sets, describing ones learning curve and objectively measuring it through the calculation of the available similarity indices.

In conclusion, auto-segmentation tools represent an attractive opportunity with potentials for both reducing segmentation time and increasing the adherence to existing guidelines. However, they need to be carefully implemented and validated in daily clinical practice. For such studies, we have presented a set of recommendations with regard to the ontology definition, performance evaluation tools and benchmark evaluation methods to contribute in the process of establishing useful clinical evidence from studies in this field.

Declaration of interest

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

References

Evidence from autocontouring software


