AbstractID: 11492 Title: DTA-based metrics for the evaluation of autosegmentation algorithms in clinical radiotherapy workflow

Purpose: To establish quantitative metrics to assess the performance of autosegmentation programs in the context of clinical radiotherapy workflow. **Method and Materials:** One of the more time intensive and predominantly un-automated tasks in radiation therapy treatment planning is the segmentation of the anatomical structures which are relevant for a particular intended radiotherapy treatment. We present performance metrics based on the distance-to-agreement (DTA) between manually-contoured reference sets and the results of a generic autosegmentation platform. **Results:** To correlate to the time expended to correct erroneous auto-contouring results we posit the slice-by-slice DTA = $\min\{ | \mathbf{x}_{r,i} - \mathbf{x}_{t,i} | \} \; \forall \; \{\mathbf{x}_{t,i}\}$, where$\{\mathbf{x}_{r,i}\}$ and $\{\mathbf{x}_{t,i}\}$ are the vertices on the reference and test sets, respectively, on the i^{th} axial CT slice. The histogram of DTAs (or simple pass/fail DTA thresholds) will assess the amount of manual correction required to bring an auto-contoured structure into acceptable agreement with the reference set. Beyond this application, full automation of the contouring process (i.e. no manual correction) would require stricter, 3-dimensional, and site-specific DTA metrics, for which we propose generalized quality factors. As a test bed for these metrics, we have compared manually-generated clinical contours to those created by CMS' Atlas-Based Autosegmentation system and shown that, for some sites, the autosegmentation is nearly indistinguishable from the variability in our clinic's manual contouring. **Conclusion:** Depending on the desired use of an autosegmentation package, either slice-by-slice or 3D DTA metrics facilitate the assessment of clinical performance. Regardless of intended application, the DTA-based metrics may be used as (a) a quantitative benchmark of an autosegmentation algorithm against the inherent variability in manual contouring and (b) a potential correlate to efficiency gains should the software be brought into full clinical use.