Contrôle qualité des contourages générés par intelligence artificielle : étude rétrospective des variations des différentes versions d'un même logiciel de délinéation automatique.

Maud Suszko*[1], Francesca Di-Franco[2], Jimmy Fontaine†‡[1], Monnet Clemence[1], Emilie Bonnet[1], Bertrand Fleury[1], Sebastien Clippe[1], and Jean-Baptiste Guy[1]

[1]Centre Marie Curie – Centre Marie Curie – France
[2]IRUDIGI SARL – IRUDIGI SARL – France

Résumé

Introduction : L'émergence des logiciels de contourage automatique basés sur l'intelligence artificielle (IA) en radiothérapie a révolutionné la délinéation des structures anatomiques, permettant une réduction significative du temps nécessaire à la délinéation en parallèle d'une validation médicale rigoureuse. Leur intégration en routine clinique nécessite une évaluation approfondie et un processus de validation reste essentiel, notamment lors des mises à jour régulières de ces outils. Ce travail étudie l'impact des mises à jour d'un logiciel de contourage automatique sur la qualité des contours et son intégration dans un processus qualité, tout en confirmant son efficacité clinique.
Matériel et méthodes : Une cohorte de 40 scanners de patients précédemment irradiés et contourés par un médecin (20 thorax, 10 pelvis masculins, et 10 ORL), a été utilisée pour comparer six versions d'un système commercial de contourage automatique (Limbus Contour) (v1.4, v1.5, v1.6, v1.7, v1.8B2, v1.8B3). Le coefficient de similarité de Dice (DSC), la distance de Hausdorff (HD) et la différence de volume relative (RVD) ont été calculés en utilisant un code Python (V3.11.5) pour comparer les contours générés par l'IA aux contours de référence. Un Wilcoxon signed-rank test a été réalisé pour évaluer la significativité statistique.

Résultats : L'indice DSC augmente en moyenne d'une version à l'autre lorsque les structures sont ré-entraînées. Pour les seins, malgré une légère baisse d'environ 1 % du DSC entre v1.5 et v1.8B3, aucune différence statistiquement significative n'a été observée ($p > 0.05$). Les structures n'ayant pas été ré-entrainées entre les versions présentaient des variations de DSC négligeables. En moyenne, l'indice DSC était ≥ 0.75 pour l'ensemble des structures. Les distances HD étaient respectivement $<$ 5mm, $<$ 13mm et $<$ 16mm pour les contours ORL, pelvis et thorax sur l'ensemble des versions. La distance HD moyenne était diminuée de 1.45mm entre les versions 1.8B2 et 1.8B3. La RVD était en moyenne $<$ 15 %, avec une augmentation non statistiquement significative pour la majorité des structures ($p > 0.05$).
Conclusions : Les résultats démontrent une amélioration de la précision du contourage entre les différentes versions du logiciel Limbus Contour jusqu'à présent. Les structures non

*Auteur correspondant: physique@cmc-valence.org
†Intervenant
‡Auteur correspondant: jim_f@hotmail.fr

modifiées d'une version à l'autre ont maintenu des indices DSC, HD et RVD constants. Ces résultats ont motivé l'intégration d'un processus de contrôle qualité rapide et systématique dans notre clinique, pour évaluer quantitativement les contours après chaque mise à jour. Une analyse plus approfondie reste nécessaire pour évaluer d'autres localisations anatomiques et sur des cohortes plus étendues.

Mots-Clés: Assurance Qualite, delineation, contourage automatique IA