Evaluating Consistency in CT Derived Body Composition Analysis: A Comparative Study of Two Deep Learning Algorithms at L3 Vertebral Level

Adam P. Dachowicz, PhD, Data Science Analyst, Radiology, Mayo Clinic
Jason Klug, PhD; Timothy Kline, PhD, MS; Panos Korfiatis, PhD; Daniel Blezek, PhD; Bill Ryan; Eric Williamson, MD; Steve Langer, PhD, FSIIM; Jeremy Collins, MD; Francis Baffour, MD; Andy Missert, PhD; Gian Marco Conte, MD, PhD

Introduction

The increasing availability of deep learning algorithms for body composition analysis from medical imaging presents opportunities to improve clinical workflows. However, potential variability in algorithm output raises concerns about the consistency of measurements. This study compares the performance of two state-of-the-art algorithms for body composition analysis, evaluating their agreement on key segmentation metrics.

Hypothesis

We hypothesize that despite high segmentation accuracy, differences in training will lead to measurable differences in output between the algorithms.

Methods

The study included 1,132 abdomen-pelvis CT scans from 615 patients. The analysis focused on the L3 vertebral level, a standard reference for body composition studies, using an internally developed model and Comp2Comp, an open-source tool. Both algorithms were applied to segment skeletal muscle (SKM), visceral adipose tissue (VAT), subcutaneous adipose tissue (SAT), and inter-muscular adipose tissue (IMAT). Segmentation was performed on identical CT slices taken at the L3 vertebra level. Segmentation accuracy was evaluated using the Dice Similarity Coefficient (DSC). Additionally, Bland-Altman analysis was used to assess agreement in total segmented volume. We also compare agreement across demographic and scanner variables of interest, including patient sex, age, and slice thickness.

Results

The algorithms demonstrate variable segmentation agreement across tissues, with highest agreement for SKM and lowest for IMAT, yielding DSC of 0.957±0.038 (mean +/- standard deviation) (SKM), 0.900+/-0.12 (VAT), 0.911+/-0.071 (SAT), and 0.843+/-.0.128 (IMAT). We see absolute segmented area differences of 4.32+/-6.55 cm^2 (SKM), 9.01+/-11.52 cm^2 (VAT), 18.22+/-26.34 cm^2 (SAT), and 1.18+/-1.16 cm^2 (IMAT). We observed evidence of a significant difference in the distribution in DSC for patient sex across all segmentation tissues (2-Sample KS p-value < 0.05), for age across all tissues (multi-sample A-D p-value < 0.05 and KW p value < 0.05), and for slice thickness across all tissues (2-Sample KS p-value < 0.05), suggesting segmentation performance is sensitive to these parameters.

Conclusion

The algorithms yield high but variable segmentation agreement across tissues. We observe evidence of varying levels of agreement across several demographic variables of interest, holding the L3 slice and originating CT scan constant. These result highlights the utility of benchmarking algorithms with similar outputs against each other to capture where disagreements occur. Such differences are important to be aware of in clinical practice.

Figure(s)

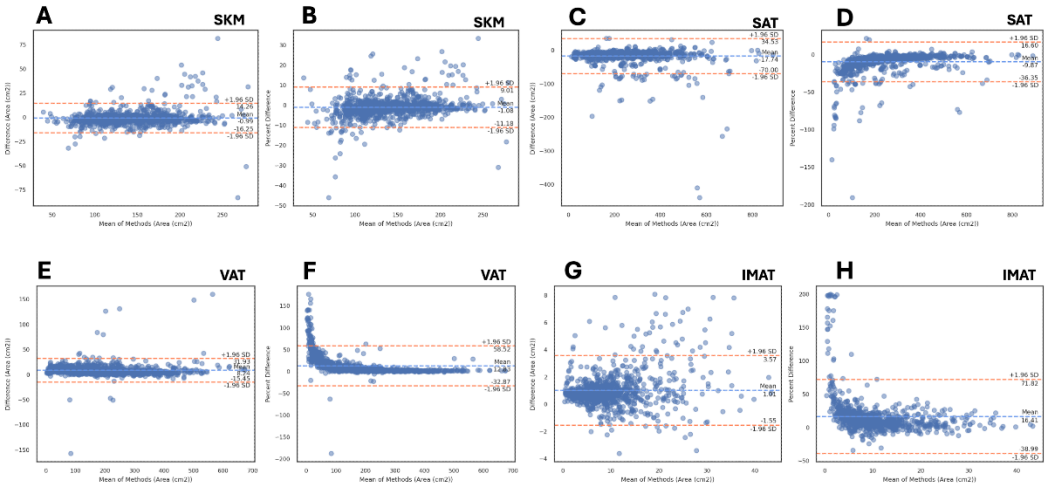

Figure 1. Bland-Altman plots for differences between algorithm segmentation areas across tissues of interest. Bland-Altman plot comparing differences in cm^2 and percent difference for skeletal muscle (SKM; A, B), subcutaneous adipose tissue (SAT; C-D), visceral adipose tissue (VAT; E-F), and inter-muscular adipose tissue (IMAT; G-H).

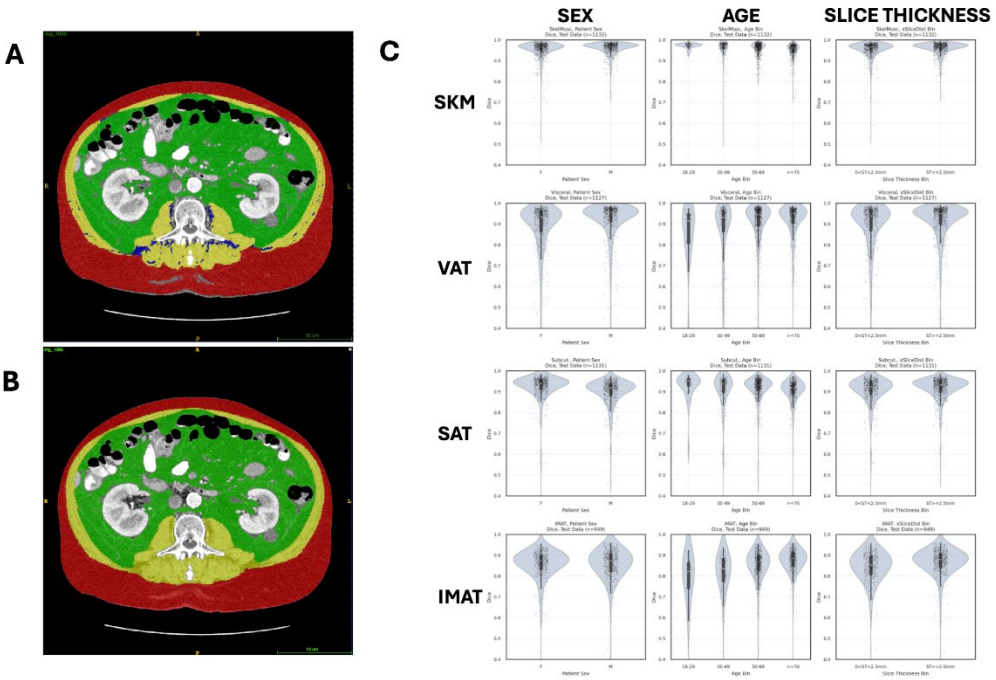

Figure 2. (A) Example segmentation output from the comp2comp algorithm, with categories subcutaneous adipose tissue (SAT, red), skeletal muscle (SKM, yellow), inter-muscular adipose tissue (IMAT, blue), and visceral adipose tissue (VAT, green). (B) Example segmentation from the internal algorithm with categories SAT (red), SKM + IMAT (yellow), and VAT (green). (C) DSC computed between the algorithms outputs across variables of interest (horizontal axis) and tissues of interest (vertical axis).

Keywords

Artificial Intelligence/Machine Learning; Emerging Technologies; Imaging Research