The impact of contouring protocols on machine learning input data quality

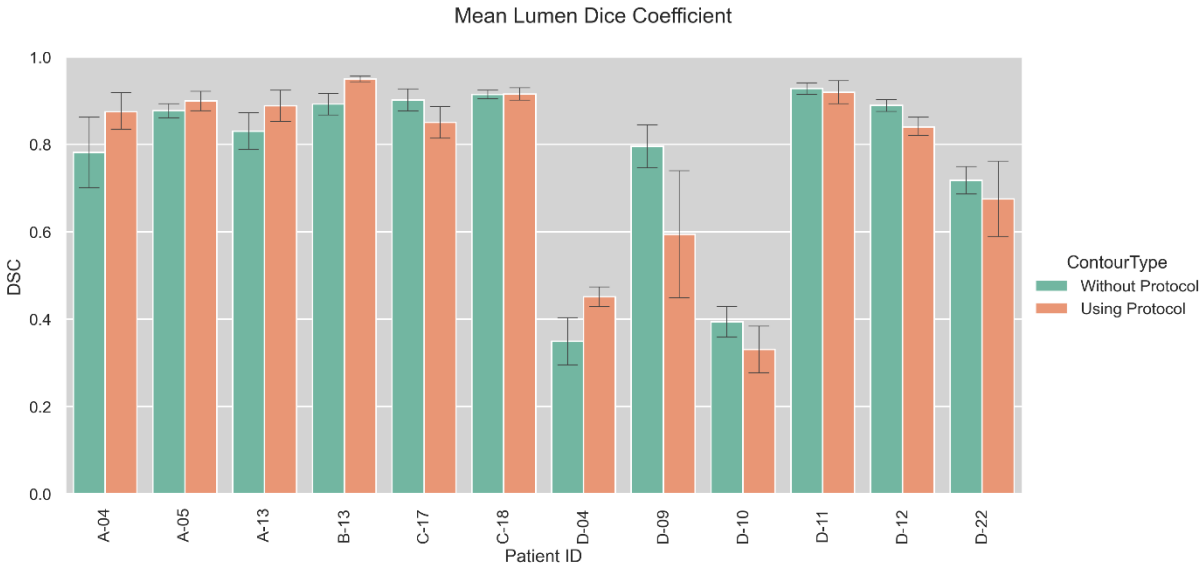
C. Boyd[1,2], T. J. Kleinig[3,4], J. Dawson[5,6], S. Patel[7], W. Mayer[8], E. Bezak[2, 9]
[1] Medical Physics and Radiation Safety, South Australia Medical Imaging, Adelaide, Australia
[2] Allied Health & Human Performance, University of South Australia, Adelaide, Australia
[3] Department of Neurology, Royal Adelaide Hospital, Adelaide, Australia;
[4] Adelaide Medical School, The University of Adelaide, Adelaide, Australia;
[5] Discipline of Surgery, University of Adelaide, Adelaide, Australia;
[6] Department of Vascular & Endovascular Surgery, Royal Adelaide Hospital, Adelaide, Australia;
[7] Department of Medical Imaging, Royal Adelaide Hospital, Adelaide, Australia
[8] Industrial AI Research Centre, UniSA STEM, University of South Australia, Australia;
[9] Department of Physics, University of Adelaide, Adelaide, Australia

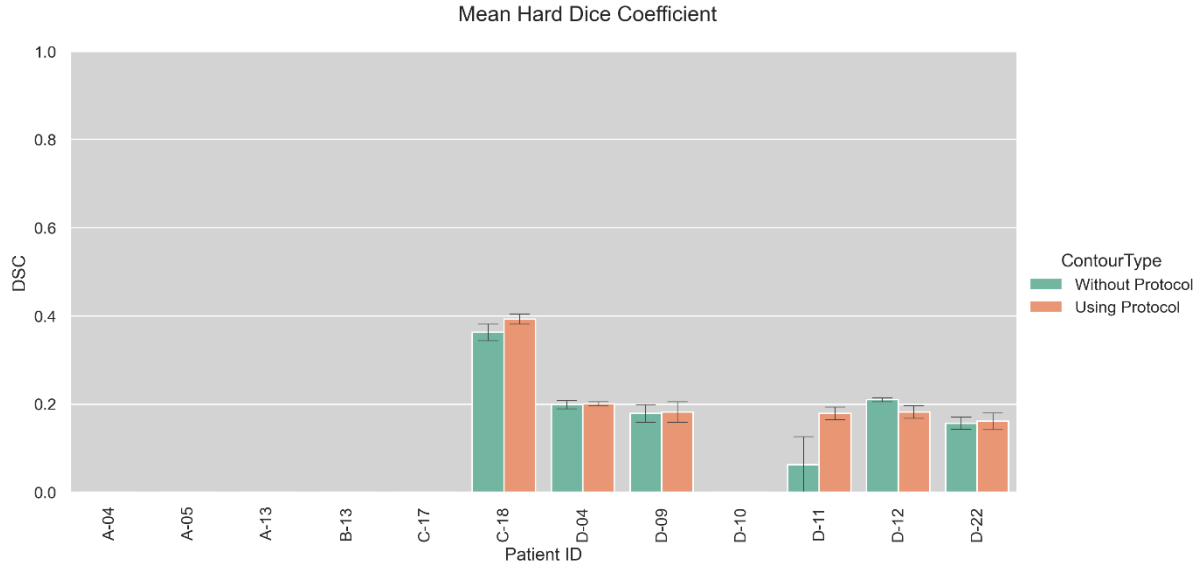
Aim
Though contour analysis is a mature field in radiation oncology fields, contouring is increasingly being performed by other groups seeking to develop training datasets for the training of machine learning. This presents new challenges in producing contours that are accurate, precise, and reproducible. This work sought to quantify the variability in contours of carotid atherosclerosis CTA, from five professionals from differing clinical backgrounds and experience.

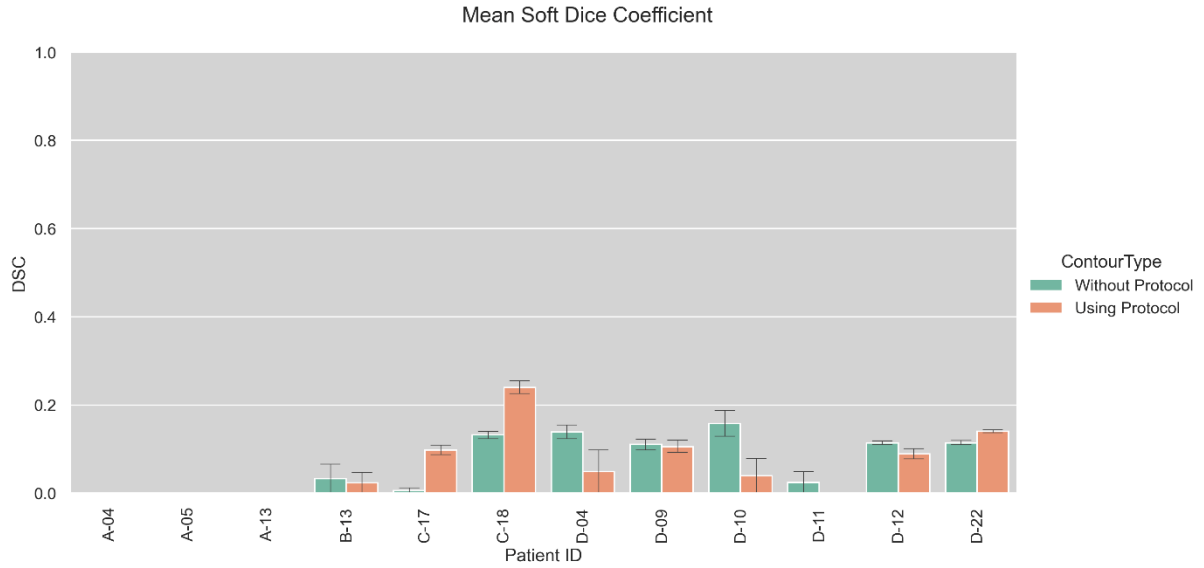
Materials & Methods
We analysed extracranial carotid lesions, with no, mild, moderate, and severe atherosclerotic burden (n=10/group). Lumen, calcific plaque, and soft plaque were manually contoured by three expert experienced clinicians (neuroradiologist, vascular neurologist and vascular surgeon), a medical physicist, and a radiographer. Contouring was repeated following six months for intra-operator variability, and again after development of an agreed protocol to reduce variability. Clinicians also blindly ranked each other's contours for descriptive statistical analysis.

Results
Lumen and calcific plaque contouring by the non-clinicians was similar to clinicians [Dice Similarity Coefficient (DSC): 0.88/0.87 and 0.69/0.71 respectively], with greater soft plaque variation [DSC: 0.37/0.43]. Clinicians generally favoured their own contours, most pronouncedly the neuroradiologist [Standard Deviation: 0.00]. Though DSC of lumen was increased, and Hausdorff Distance of soft plaque was reduced with the implementation of a protocol, this was not found to be statistically significant.

Discussion
CTA carotid pathology contouring inherently has limited agreement between clinicians due to the small structure size and relatively poor contrast. Even datasets from experienced clinicians are prone to potentially terminal variability which requires careful quantification to account for in any future machine learning applications.


Mean Lumen Dice Coefficient
DSC
1.0
0.8
0.6
0.4
0.2
0.0
A-04
A-05
A-13
B-13
C-17
C-18
D-04
D-09
D-10
D-11
D-12
D-22
Patient ID
ContourType
Without Protocol
Using Protocol


Mean Hard Dice Coefficient
DSC
1.0
0.8
0.6
0.4
0.2
0.0
A-04
A-05
A-13
B-13
C-17
C-18
D-04
D-09
D-10
D-11
D-12
D-22
Patient ID
ContourType
Without Protocol
Using Protocol


Mean Soft Dice Coefficient
DSC
1.0
0.8
0.6
0.4
0.2
0.0
A-04
A-05
A-13
B-13
C-17
C-18
D-04
D-09
D-10
D-11
D-12
D-22
Patient ID
ContourType
Without Protocol
Using Protocol