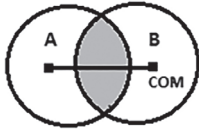
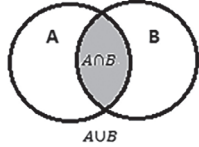
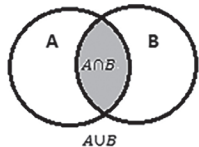
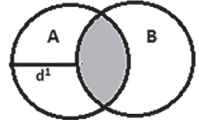


Table S1 Summary of metrics used to assess interobserver variation. The accuracy of the metrics depends on their ability to assess variations in volume, shape, location, margins required to account for interobserver variations and ultimately, treatment outcomes. Some metrics are easily exported from the treatment planning software and are more widely used (12,16,19,90-94)

Metric type	Method	Perfect value	Advantages	Limitations
Simple volume measurements	Compares delineated volume with a reference contour; e.g., volume A, volume B	1	Easily exported from planning software; correlates well with NTCP; provides information on over or under outlining	Contours can have the same volume but different shape and location; cannot be used to calculate margins
Centre of mass (COM)	Calculates the difference in the centre coordinates (x,y,z) of different contours	0	Easily exported from planning software	Contours can have the same COM but different shape and volume; cannot be used to calculate margins
 <p>Centre of mass (COM)</p>				
Overlap metrics	The overlap between observer contour (A) and reference contour (B) can be calculated using; $Jaccard = \frac{A \cap B}{A \cup B}$ or $Dice = \frac{2(A \cap B)}{(A \cup B)}$ The general conformity index (CI _{gen}) can be used to calculate the overlap between many pairs of observers. No reference contour required. $CI_{gen} = \frac{\sum^n pairs (A \cap B)}{\sum^n pairs (A \cup B)}$	1	Easily exported from planning software; they are widely used in the literature	Provides no information on shape and volume variations; overestimates variations for small contours; cannot be used to calculate margins
 <p>Overlap metrics</p> <p>A ∩ B: Intersection A ∪ B: Union</p>				
Over or under outlining	General miss index (GMI): calculates the amount of under outlining. $GMI = \frac{B - (A \cap B)}{B}$. Discordance index (DI) calculates the amount of over outlining $DI = \frac{1 - (A \cap B)}{A}$	0	Easily exported from planning software; provides information about over and under outlining	Provides no information on shape and volume variations; site and case dependent; cannot be used to calculate margins
 <p>Over or under outlining</p> <p>A ∩ B: Intersection A ∪ B: Union</p>				
Shape surface metrics	Local SD measures the perpendicular distance (d^{\perp}) reference contour (B) to the observers' contours (A). The standard deviation in the distance between all observers is calculated at each point, and then the average is calculated using the root mean square. Other similar algorithms include Mean distance to agreement, ComGrad distance and bidirectional local distance(96). These vary in the method used to measure the distance from reference contour	0	Widely used in the literature; provides information about shape and location; can be used to estimate margins	Requires specialised software; no information about volume; overall score influenced by outliers; accuracy for irregularly shaped contours depends on the algorithm
 <p>Shape surface metrics</p>				
3D shape surface method	Distribution of local SD over tumour surface area plotted as a histogram or 3D surface map	0	Provides information about the percentage tumour surface area affected by large interobserver variation	Requires specialised software; no information about volume; accuracy for irregularly shaped contours depends on the algorithm
Dosimetric assessment	Involves applying the dose distribution from a reference expert plan to each of the observers' contours with or without plan optimisation according to the observers' contours. Or plan on each observer and evaluate coverage of reference contours. Dosimetric deviations from the reference plan are evaluated.	Within acceptable margins and PTV tolerances according to clinical guidelines	Provides a correlation with clinical outcomes	Time consuming; not widely used in interobserver studies; depends on the ability of the planner to optimise the plan
Semi-quantitative analysis	An algorithm creates a percentage score by identifying the voxels falling outside or missing from the reference contour. A penalty can be applied by the teacher based on the distance of the voxels from the reference contour and severity of error (94,95)	100%	Provides a correlation with clinical outcomes	Limited research on the penalties that should be applied; specialised software required
Expert qualitative visual analysis	Expert/s visually classify contours as acceptable or unacceptable based on the clinical impact of the error (18)there is little research documenting its impact in the setting of stereotactic body radiation therapy (SBRT	Accept	Provides a correlation with clinical outcomes; no specialised software required; facilitates the identification of factors leading to interobserver variation	Subjective; time consuming; no quantitative measurement provided

Local SD, local standard deviation; TCP, tumour control probability; NTCP, normal tissue complication probability.

References

90. Jameson MG, Holloway LC, Vial PJ, et al. A review of methods of analysis in contouring studies for radiation oncology. *J Med Imaging Radiat Oncol* 2010;54:401-10.
91. ProKnow. Radiation Oncology Residency Programs 2018. Available online: <https://proknowsystems.com/benefits/educators-rorp?content=proknow> (accessed 9 January 2019).
92. Van Herk M, Duppen J, Massoptier L, et al. EP-1801: A novel web-based delineation and scoring system for teaching target volume delineation. *Radiother Oncol* 2014;111:S290.
93. Nelms BE, Tomé WA, Robinson G, et al. Variations in the contouring of organs at risk: test case from a patient with oropharyngeal cancer. *Int J Radiat Oncol Biol Phys* 2012;82:368-78.
94. Kim HS, Park SB, Lo SS, et al. Bidirectional local distance measure for comparing segmentations. *Med Phys* 2012;39:6779-90.