An audit of outlining agreement in the male pelvis

Dr. Keith Langmack

Head of Radiotherapy Physics

Nottingham University Hospitals

Overview of Talk

- Background
- Method
- Analysis
- Results
- Interpretation

Background

- Aim to reduce amount of Oncologist and Radiographer time devoted to outlining
- Literature indicates reduced outlining time by use of auto-segmentation tools
- As part of introducing these in a pilot site (prostate) audited consistency of outlines (prostate, rectum, bladder, femoral heads, seminal vesicles, body)

Method

- Selected 8 previously treated prostate patients as test group
- Their original outlines were taken as the gold standard
- Each patient was re-outlined by a different person (inter-observer agreement)
- Each patient was also outlined with ABAS and edited by the same person who performed the re-outlining (assisted intra-observer agreement)

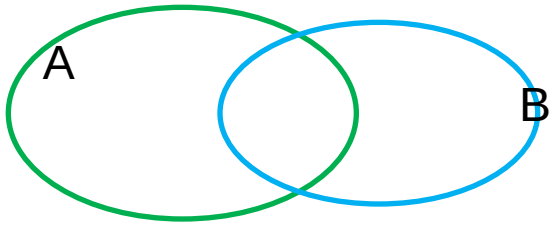
Quantifying Agreement

- There have been a number of recent reviews of techniques for analysing contouring agreement.
- These all conclude that no one metric can be used to summarise agreement between contours and recommend quoting several metrics
- However there is little guidance on what values are “acceptable”

Jameson et al J. Med. Imag. and Rad. Oncol. 2010;54:401–10.
Hanna et al Clin. Oncol. 2010;22:515-25.
Fotina et al Strahlenther Onkol 2012;188:160–7

Possible Metrics (1)

- Simple centre of mass
- Volume comparisons
- Overlap measures (e.g. conformity index, Dice's similarity coefficient)
- CI = $(A \cap B)/(A \cup B)$, DSC = $2(A \cap B)/(A+B)$



Possible Metrics (2)

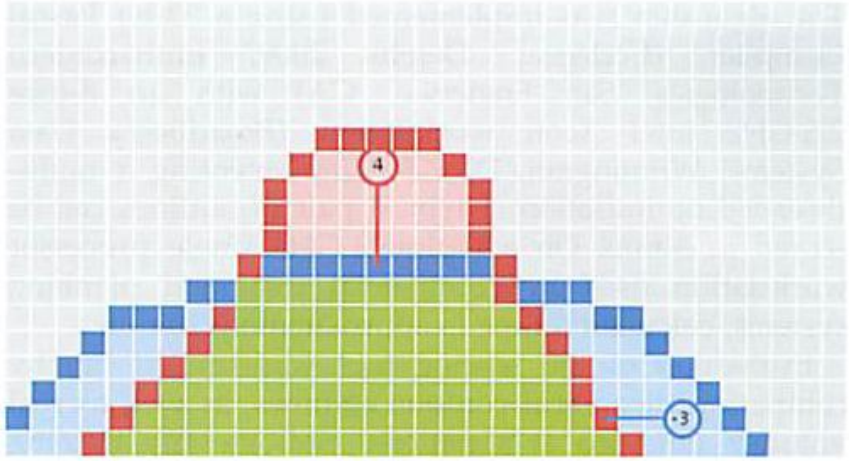


Figure 2. Diagram representation of the cellular automaton network. A series of nodes from a single axis of the network is shown. Two outlines have been entered into the node. The blue contour is the reference contour and the red contour the evaluation contour. The green voxels have been determined to lie within both outlines. The light blue voxels are within the reference outline but not within the evaluation outline and represent under-contouring errors. The light red voxels are within the evaluation outline but not the reference outline and represent over-contouring errors. For two of the nodes indicated by the red and blue circles, the distance to the nearest contour has been displayed. Note that over-contouring errors are given positive values and under-contouring errors are given negative values.

The British Journal of Radiology, 83 (2010), 44–51

A novel algorithm for the morphometric assessment of radiotherapy treatment planning volumes

[1]R JENA, MD, MRCP, FRCR, [2]N F KIRKBY, PhD, [1]K E BURTON, Msc, [1]A C F HOOLE, PhD, [1]L T TAN, MD, MRCP, FRCR and [3]N G BURNET, MD, FRCS, FRCR

Mean Distance to Conformity (MDC)

Mean distance that an outlying point in a volume must be moved to achieve perfect conformity with the reference volume.

Implemented in IMSimQA

Overlap Metrics and Statistics

- Zijdenbos et al indicates that DSC is a special case of the kappa statistic
- Analysis show “excellent agreement” for DSC>0.7, CI>0.54
- Zou et al recommend using logit transform of DSC for statistical analysis

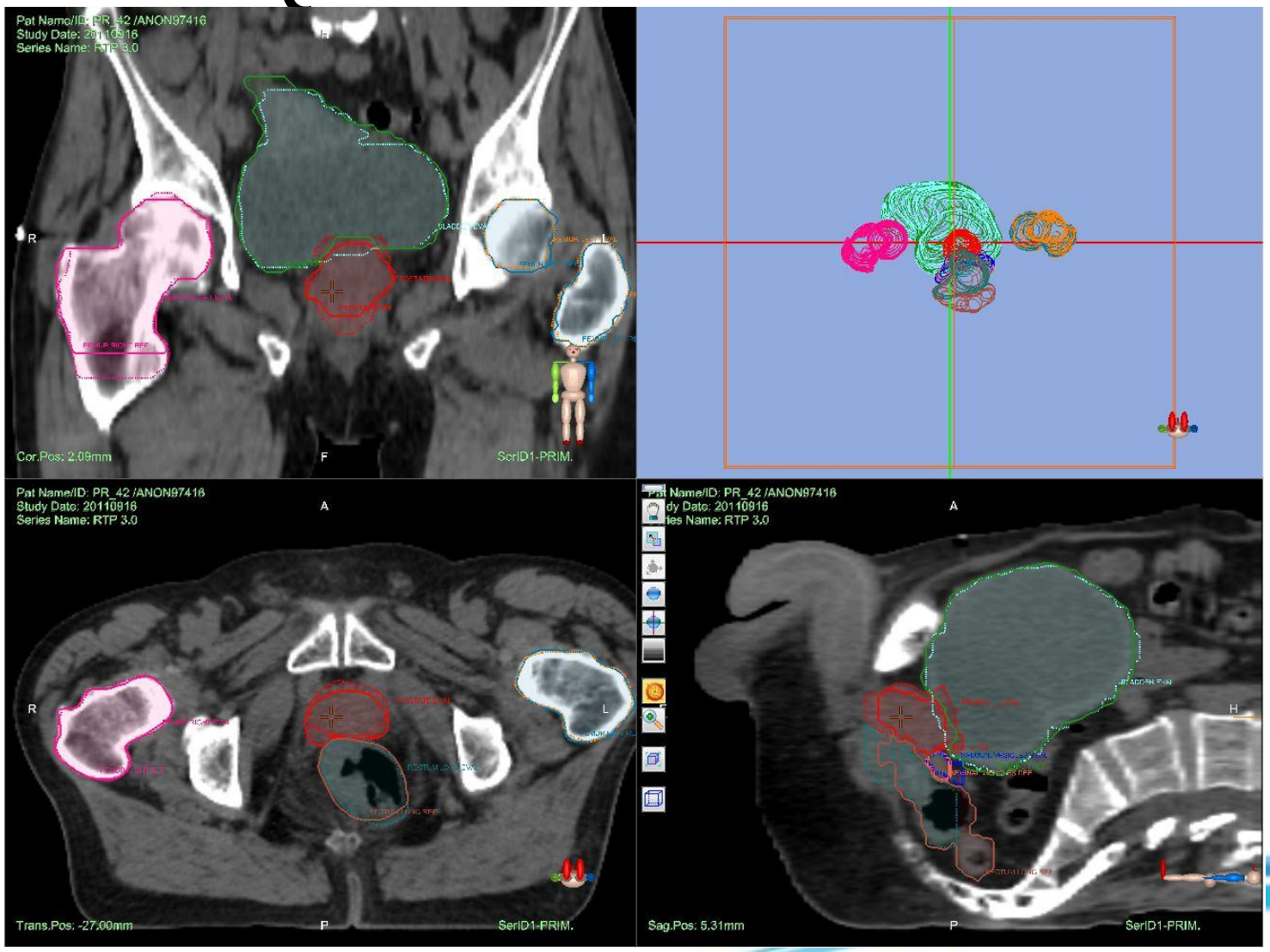
Kappa	Agreement
<0	Less than chance
0.01 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 0.99	Almost perfect

$$\text{Logit}(DSC) = \ln\left(\frac{DSC}{1-DSC}\right)$$

$$[0,1] \rightarrow [-\infty, +\infty]$$

Zou et al. Statistical validation of image segmentation quality based on a spatial overlap index: scientific reports. Acad Radiol 2004;11:178–89.
Zijdenbos et al Morphometric analysis of white matter lesions in MR images: method and validation. IEEE Trans Med Imaging 1994; 13:716–724.

Qualitative Results



Inter-Observer Consistency Results

Structure	CI	DSC	MDC(mm)	PVD(%)
Bladder	0.78 (0.07)	0.88 (0.05)	2.2 (0.3)	9.2 (6.7)
Body	0.97 (0.02)	0.99 (0.01)	2.6 (2.8)	1.9 (1.1)
LFH	0.81 (0.05)	0.89 (0.03)	3.0 (1.0)	7.4 (6.9)
RFH	0.83 (0.04)	0.91 (0.02)	2.6 (0.8)	5.9 (4.3)
Prostate	0.54 (0.10)	0.70 (0.08)	3.1 (0.7)	28.3 (27.1)
Rectum	0.63 (0.14)	0.76 (0.11)	4.8 (2.4)	18.3 (18.6)
SV	0.42 (0,10)	0.58 (0.10)	3.4 (0.9)	27.7 (16.1)

Original outlines v new outlines

Inter-Observer Consistency Results

Structure	CI	DSC	MDC(mm)	PVD(%)
Bladder	0.85 (0.11)	0.92 (0.06)	2.1 (0.6)	3.6 (3.7)
Body	0.99 (0.02)	0.99 (0.01)	1.4 (0.4)	0.9 (1.2)
LFH	0.80 (0.08)	0.88 (0.05)	4.9 (3.4)	10 (3.6)
RFH	0.82 (0.11)	0.90 (0.07)	4.5 (3.3)	7.2 (5.4)
Prostate	0.72 (0.12)	0.83 (0.09)	2.1 (0.5)	20.9 (14.0)
Rectum	0.66 (0.13)	0.79 (0.10)	5.2 (2.0)	14.1 (11.7)
SV	0.41 (0.12)	0.57 (0.13)	3.5 (0.8)	23.7 (15.8)

Original outlines v Edited ABAS outlines

Intra-Observer Consistency Results

Structure	CI	DSC	MDC(mm)	PVD(%)
Bladder	0.78 (0.09)	0.88 (0.05)	2.0 (0.4)	10.8 (6.0)
Body	0.97 (0.02)	0.98 (0.01)	2.6 (2.6)	2.0 (1.1)
LFH	0.78 (0.11)	0.87 (0.07)	4.3 (3.3)	7.3 (6.8)
RFH	0.81 (0.12	0.89 (0.08)	3.8 (3.2)	6.0 (6.5)
Prostate	0.59 (0.11)	0.74 (0.09)	3.1 (0.9)	12.7 (7.7)
Rectum	0.67 (0.06)	0.80 (0.04)	3.4 (0.8)	12.6 (12.6)
SV	0.49 (0.12)	0.65 (0.11)	2.8 (0.7)	45.7 (34.3)

New outlines v edited ABAS outlines

And this means?

Structure	Metric	p values		
		T1 v T2	**T1 v T3**	**T2 v T3**
Bladder	PVD	0.109	0.329	**0.020**
Body	logit CI	**0.004**	0.298	**0.001**
Prostate	CI	**0.013**	0.143	**0.045**
	logit CI	**0.014**	0.123	**0.047**
	DSC	**0.014**	0.185	**0.042**
	logit DSC	**0.014**	0.123	**0.047**
	MDC	**0.008**	0.897	**0.018**

One way ANOVA used to find significant differences, then pairwise t-test

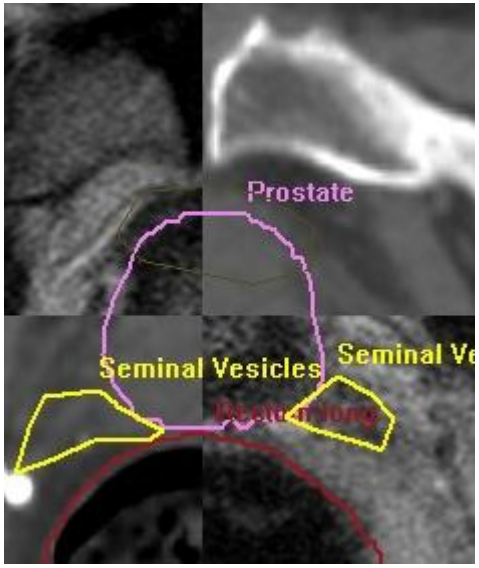
Hint that ABAS may increase prostate outlining consistency between observers?

T1 = original outlines v new outlines

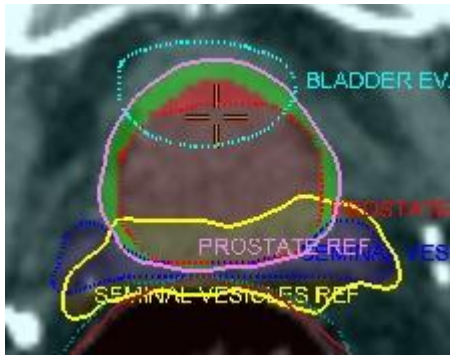
T2 = original outlines v edited ABAS outlines

T3 = new outlines v edited ABAS outlines

An Example

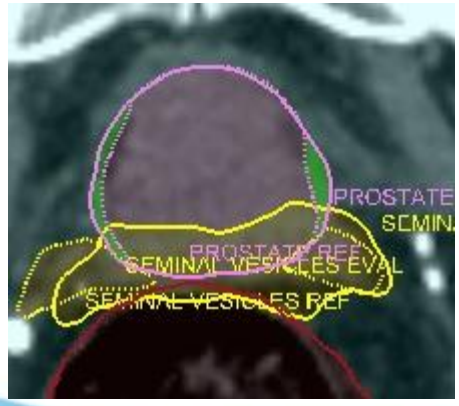


Original v Re-outlined

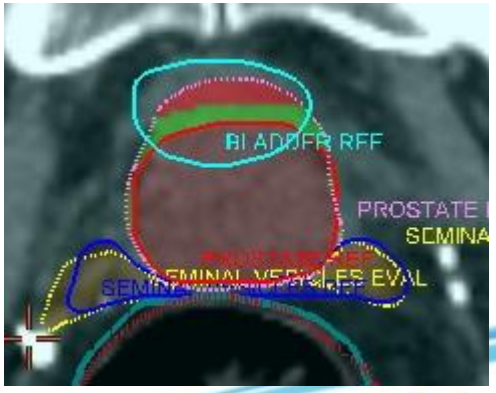


PROSTATE	
C.I.	0.72
MDC Mean(mm)	1.38
MDC Under(mm)	2.76
MDC Over(mm)	0.00

Original v ABAS edited



Re-outlined v ABAS edited



PROSTATE	
C.I.	0.77
MDC Mean(mm)	2.02
MDC Under(mm)	1.00
MDC Over(mm)	3.05

Potential Clinical Impact

- Volume accuracy is important for IMRT planning in particular
- If a dose constraint is say no more than 50% of the rectum is to get 40 Gy, then if the volume is found to be 100cc, then 50% is 50cc, but if it is outlined as 110cc then 50% is 55cc. This might affect toxicity

Summary

- This sort of work can be quite time consuming and requires a large clinical commitment
- There is no agreement as to which metric or combination of metric to use, and very limited data on what level of agreement is good enough
- However qualitative analysis indicates that structures where there isn't an obvious boundary are particularly prone to disagreements

Thanks to...

- Dr. D. Saunders
- Dr. J. Mills
- Dr. S. Sundar
- Dr. C. Perry
- Dr. P. Dickinson
- Dr. H. Reddy
- Clare Sinstead
- Carol Kennedy
- The rest of the pre-treatment team