Intra-observer Variability: Should we worry?

Abstract

Many papers have identified concerns about intra-observer variability of radiotherapy outlining. Levels of variability in this are much higher than in diagnostic CT image evaluation tasks. This is likely to arise from the need for an assessment to be made on every image slice with a substantially more complex outcome. Further challenges are due to the lack of a "gold standard" or "ground truth" where an imaging finding can be directly confirmed by biopsy or clinical examination. This paper provides an epistemological approach to the realistic expectations for intra-observer variability in this scenario and the extent to which this is an issue.

In most aspects of medical practice it is commonly understood that provided clinicians have had sufficient training then their judgement is to be trusted. A true constructivist approach would accept intra-observer variability as an inevitable consequence of individual clinical decision making. Certainly there are few other aspects of medical practice where a clinician's "expert opinion" is subjected to such rigorous investigation. Training, guidelines and provision of high quality imaging data can improve observer agreement but beyond a threshold level there exists a range of acceptable outlines that are all valid clinically.

A constructivist approach to variability may empower clinicians to accept this variability as an inherent aspect of their practice. Research efforts should perhaps be focussed on maximising impact of training and guidelines as well as the development of a target minimum agreed measure of intra-observer variability that educational interventions should seek to facilitate.

Introduction

Many papers have identified concerns about intra-observer variability of repeat outlining by the same clinician. These variations in individual performance in turn make it challenging to determine values for inter-observer variability since these depend largely on the assumption that each observer's outline is accurate. Aside from the concerns about inaccuracy, variability is a potential component of the PTV margin and thus minimisation of this has the potential to reduce normal tissue dose and morbidity. One accepted measure of intra-observer agreement since 1960 [1] has been the Kappa (k) correlation coefficient which varies from 0 (agreement by chance) to 1 (full agreement). The accepted subdivisions of kappa [2] are "excellent" (0.81-1.00), "good" (0.61-0.80), "moderate" (0.41-0.60), "fair" (0.21-0.40) and "poor" (0-0.20). It is clear from the evidence base that kappa is common to many aspects of medical practice. Despite the kappa assumptions concerning observer independence [3], it has been used extensively to report both intra- and inter-observer variability in the interpretation of CT imaging data. Table 1 summarises the results of these studies from the last 10 years.

Table 1: Best reported kappa for intra-observer variability in CT-based studies

Paper	Region or Pathology	Task	Kappa (best case)
Meirelles 2006	Pleural plaques	Diagnosis	1
Branstetter 2006	Middle ear	Diagnosis	0.99
Tan 2007	Spinal allograft fusion	Classification	0.95
Lee 2009	Ear otosclerosis	Classification	0.94
Brunner 2009	Proximal humerus fractures	Diagnosis	0.91
Panou 2015	Lower limb torsional profile	Evaluation	0.88
Hopyan 2010	Stroke	Diagnosis	0.88
Wattjes 2009	Brain	Classification	0.88
Arduini 2015	Hip muscle	Classification	0.872
Chang 2010	Cervical spine	Evaluation	0.86
Lee 2010	Lung cavitary mass	Evaluation	0.854

Brinjikji 2010	Haemorrhage	Classification	0.8
Ridge 2015	CT pulmonary node	Evaluation	0.792
Hoomweg 2008	Abdominal Aortic Aneurysm rupture	Diagnosis	0.78
Abul-kasim 2009	Scoliosis screw placement	Evaluation	0.76
Renou 2010	Brain haemorrhage	Classification	0.75
Roll 2011	Calcaneal fractures	Evaluation	0.75
Ozgen 2008	Temporal bone	Evaluation	0.682
De Souza 2012	Neck metastases	Diagnosis	0.66
Bogot 2005	Pulmonary nodule	Evaluation	0.659
Arealis 2014	Bone fractures	Diagnosis	0.65
Bishop 2013	Glenoid bone	Evaluation	0.64
Burkes 2014	Bone fractures	Diagnosis	0.6
Aukland 2006	Chest	Diagnosis	0.54
Carreon 2007	Spine posterolateral fusion	Evaluation	0.48
Van de Velde 2014	Brachial plexus	Outlining	0.45
Stroet 2011	Tibial fractures	Classification	0.45

Although the papers in Table 1 all relate to clinician CT interpretation skills there are clearly aspects that make some tasks more prone to variability than others. A diagnosis or classification task generally requires a clinician to use the imaging data as a whole to arrive at a single simple answer; a definitive diagnosis or rating. The mean best case kappa values in the diagnosis and classification studies are 0.78 and 0.80 respectively. Evaluation tasks usually require additional clinical expertise and decision making across the range of images which can potentially lead to wider variability; the mean best case kappa in the published studies was 0.74.

Radiotherapy outlining, however, requires an assessment to be made on every image slice and results in a substantially more complex outcome; the only reported kappa in an outlining study was 0.45. Most radiotherapy outlining studies do not report kappa but instead use a range of measures [4] including volume-ratios, volume overlap indices, centre of volume comparison or coefficients of variation to quantify the range of different volumes created; this absence of an agreed measure makes comparison challenging. It is clear, however, that intra-observer variability in radiotherapy outlining is a problem [5,6] and the requirement to assess multiple slices independently makes it extremely difficult to exclude intra-observer variability from the process.

Most of the “non-outlining” studies are also characterised by a “gold standard” or “ground truth” where an imaging finding can be directly confirmed by biopsy or clinical examination. The lack of this gold standard in radiotherapy outlining is a constant theme in published data; Khoo [6] for example, acknowledges the lack of CTV gold standard data as a limitation of his study. Unlike many other aspects of medicine, accuracy of radiotherapy outlining can only be confirmed using another expert opinion with no alternative validation method. An outline is an expression of clinical opinion concerning apparent anatomical configuration and not a predictor of a potentially measurable outcome. Combined with the major impact that this outline will have on the planned and delivered intervention, this makes variability in radiotherapy outlining a constant topic of research.

Several initiatives including educational interventions [6] and adherence to guidelines [7] have been published that have purported to help reduce variability. A good example of this was Khoo’s educational intervention [6] that included use of established guidelines and group feedback. This resulted in a 9% improvement in variability for CT outlining, although one of the participants experienced increased variability after the intervention. The authors concluded that education should be utilised more widely but also admitted a lack of “ground truth”.

While this certainly suggests that guidelines from cooperative groups such as RTOG [8] combined with training can be of value, these measures have failed to eliminate variability altogether or even attain the low levels of variability seen in diagnostic studies. This implies that there are still outstanding issues relating to either

clinician interpretation of medical imaging data or variation in clinical judgement. A recent paper attempting to evaluate guidelines for RTOG brachial plexus outlining [9] interpreted continuing intra-observer variability as evidence that the guidelines were inaccurate or insufficient. An alternative hypothesis could be that there is an underlying variability associated with some complex clinical tasks that guidelines and training cannot completely eliminate.

Intra-observer variability is of course not detectable in a single outline and every outline performed by a clinician represents the end product of a process that they are satisfied with. Provided sufficient training has been undertaken; to suggest that variability is an issue implies that clinician-approved outlines are not appropriate. There are two potential reasons why an appropriately trained and experienced clinician supported by guidelines would outline a structure differently on 2 separate occasions. Either on one occasion the clinician is unhappy with it or on both occasions they are satisfied that the outline is clinically acceptable. It must be assumed that the first reason is invalid and that clinicians would never be satisfied with sub-standard work. This leaves the conclusion that although the outlines are different, on both occasions the individual is satisfied with the output; thus they are both clinically acceptable. The clinical decision-making skills on each occasion have created a level of variability. This paper maintains that this variability is NOT a problem as each provided that training and guidelines have been utilised.

The challenge for the profession is to manage the possibility that several different outlines can be acceptable when this contradicts the desire for a single "ground truth". This paper aims to summarise the realistic expectations for intra-observer variability in this scenario and discuss the extent to which this is an issue. It adopts an epistemological approach to the issue in order to postulate a new variability paradigm and aims to highlight the deeper philosophical issue underlying intra-observer variability to determine whether intra-observer variability can actually be eliminated and more fundamentally whether it actually matters.

Discussion

From an epistemological perspective, a phenomenon can be considered using a positivist or a constructivist paradigm [10]. The positivist approach assumes that there is an absolute truth that can be measured and that exists irrespective of observer experience. This paradigm has traditionally underpinned mainstream medical research and is supported by quantitative research methods. The constructivist, on the other hand, arises from the assumption that truth arises from how an observer experiences a phenomenon. The constructivist paradigm collects and analyses qualitative data in order to develop a theory relating to a phenomenon.

The positivist approach: the elusive gold standard

In the case of structure outlining it can be seen that most of the current research adopts a positivist approach with the fundamental assumption that there is a single truth; in this case a "gold standard" of an outline. An excellent review by Whitfield [11] recently underscored the importance of involving the clinician in the outlining process in order to utilise clinical expertise and visual processing skills. There is still an underlying assumption, however, that a "gold standard" can be provided by an expert opinion. Research relating to intra-observer variation is therefore aimed at helping eliminate variation from this truth completely. Guidelines and training, along with clinical experience can certainly help with this but even the most comprehensive support has this far failed to achieve a zero level of variability. Several studies have concluded that reliance on an expert opinion as a gold standard is unreliable [12-14] while Khoo [6]and Van de Velde [9] both express dismay at the lack of a gold standard for absolute comparison. Many of the current problems arising from the use of automatic or semi-automatic model-based outlining systems stem from the simple fact that a single gold standard is not appropriate for all situations or all clinicians.

Towards a constructivist approach: the consensus standard

Acknowledging the inherent variability in outlining is a clear conclusion to draw from the previous discussion. Recent research has built on this and led to the development of software such as STAPLE [15] that is capable of creating a "consensus gold standard" from an amalgamation of several different "acceptable" outlines. This complex and well-received approach adopts a combination of probabilistic and confidence-based algorithms. While this is a valuable step towards acknowledging the validity of multiple different acceptable outlines there is still a perceived need for a "gold standard" and a series of ratings of each user against which to compare future attempts. From an epistemological perspective, this approach suggests the value of a constructivist approach but still converts the findings to a positivist-based output of a single "ground" truth.

A constructivist approach: embracing variability

Adopting a purely constructivist approach to this problem assumes that there will be different truths (structure outlines) due to different experiences. These differences can be due to different inherent biases, experiences or judgements not only between individuals but also within the same individual on different occasions. In this case there is not a single "truth" gold standard but rather a range of acceptable truths that are each valid. The inherent danger in relying solely on this approach and assuming that all outlines are correct is demonstrated by the improvement in variability seen by participants following educational interventions and application of protocols.

This, of course, is true of all medical professional training and in most other aspects of medical practice. Yet it is commonly understood that provided clinicians have had sufficient training in medical procedures then their judgement is to be trusted. Sufficient access to training resources, clear guidelines and high quality images [11] to eliminate misinterpretation errors is essential, as in other aspects of medical practice. Provided these aspects are in place and clinicians are working without misplaced confidence then perhaps a true constructivist approach should be adopted that accepts intra-observer variability as an inevitable consequence of individual clinical decision making. Certainly there are few other aspects of medical practice where a clinician's "expert opinion" is subjected to such rigorous investigation.

It is hard to distinguish between the issues of error and variability. Variability due to error is expected and dangerous in the absence of guidelines, training and good images. The constructivist assumes, however, that once these errors have been eliminated there will still be intra-observer variability arising from experiential input. As previously highlighted, for complex tasks such as 3D outlining, this variability can be more substantial than other medical situations where the outcome is a diagnosis or rating. This paper proposes that variability is an inherent function of the unique radiotherapy outlining clinical paradigm and does not constitute a problem. Thus beyond the threshold level imposed by training, guidelines and imaging data there are multiple truths. Identification of clinically acceptable variability levels along with agreement on the most appropriate measure to be reported would be of considerable value in identifying this threshold.

A parallel can be drawn from radiotherapy planning where automation [16] has enabled multi-criteria optimisation IMRT software to develop a range of multiple plans with differing outcomes that are all valid. In this case an element of clinical judgement is required to select the most appropriate "truth" based on the clinician perspective and expertise. A constructivist would argue that there is no single correct plan and that clinician choice will vary whereas a positivist would take umbrage at this suggestion and develop a tool to try and identify the "best" plan. The radiotherapy clinical community may find value from permitting clinicians to create a range of acceptable outlines and stop labelling all contouring variability as an error.

Conclusion

Training, guidelines and provision of high quality imaging data can improve observer agreement but beyond a threshold level there exists a range of acceptable outlines that are all valid clinically. A more constructivist approach to variability may empower clinicians to accept variability as an inherent aspect of their practice.

Research efforts should perhaps be focussed on credentialing methods capable of determining whether or not a clinician has received sufficient training to be confident in their outlining. This could be aided by the development of a target minimum agreed measure of intra-observer variability that educational interventions should seek to facilitate.

References

1. Cohen, J., *A coefficient of agreement for nominal scales.* Educ Psychol Meas 1960; 20: 37-46
2. Altman, D.G. Practical statistics for medical research. 1991; London: Chapman and Hall
3. McHugh, M.L, *Interrater reliability: the kappa statistic.*Biochem Med 2012; 22(3): 276–82
4. Jameson, M.G, Holloway, L.C., Vial, P.J., et al, *A review of methods of analysis in contouring studies for radiation oncology Techniques of contour comparison.* J Med Imag Radiat Oncol 2010; 54(5): 401-10
5. Louie, A.V., Rodrigues, G., Olsthoorn, J. et al, *Inter-observer and intra-observer reliability for lung cancer target volume delineation in the 4D-CT era.* Radiother Oncol 2010; 95(2): 166-71
6. Khoo, E.L.H., Schick, K., Plank, A.W., et al, *Prostate Contouring Variation: Can It Be Fixed?* Int J Radiat Oncol Biol Phys 2012; 82(5): 1923-9
7. Lim, K., Small, J.W., Portelance, L., et al. *Clinical Investigation: Consensus Guidelines for Delineation of Clinical Target Volume for Intensity-Modulated Pelvic Radiotherapy for the Definitive Treatment of Cervix Cancer.* Int J Radiat Oncol Biol Phys 2011; 79: 348-55
8. Radiation Therapy Oncology Group 2016. *Contouring Atlases*: https://www.rtog.org/CoreLab/ContouringAtlases.aspx Accessed 12/05/16
9. Van de Velde, J., Wouters, J., D'Herde, K. et al, *Reliability and accuracy assessment of radiation therapy oncology group-endorsed guidelines for brachial plexus contouring.* Strahlenther Onkol 2014; 190(7): 628-32
10. Ward, K., Hoare, K.J., Gott, M., *Evolving from a positivist to constructionist epistemology while using grounded theory: reflections of a novice researcher.* J Res Nursing 2015; 20(6): 449-62
11. Whitfield, G.A., Price, P., Price, G.J., et al, *Automated delineation of radiotherapy volumes: are we going in the right direction?* B J Radiol 2014; 86(1021): 20110718
12. Stapleford, L.J., Lawson, J.D., Perkins, C., et al, *Evaluation of Automatic Atlas-Based Lymph Node Segmentation for Head-and-Neck Cancer.* Int J Radiat Oncol Biol Phys 2010; 77(3): 959-66.
13. Deeley, M.A., Chen, A., Datteri, R., et al, *Comparison of manual and automatic segmentation methods for brain structures in the presence of space-occupying lesions: a multi-expert study.* Phy Med Biol 2011; 56(14): 4557-77
14. Chao, K.S.C., Bhide, S., Chen, H., et al, *Reduce in Variation and Improve Efficiency of Target Volume Delineation by a Computer-Assisted System Using a Deformable Image Registration Approach.* Int J Radiat Oncol Biol Phys 2007; 68: 1512-21
15. Warfield, S.K., Zou, K.H., Wells, W.M., *Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation.* IEEE T Med Imaging 2004; 23(7): 903-21
16. Khan, F., Craft, D., *Original Report: Three-dimensional conformal planning with low-segment multicriteria intensity modulated radiation therapy optimization.* Pract Radiat Oncol 2015; 5: 103-11