Journal of Medical Imaging and Radiation Sciences 47 (2016) 217-220

Journal of Medical Imaging and Radiation Sciences

Journal de l'imagerie médicale et des sciences de la radiation

www.elsevier.com/locate/jmir

Brief Communication

Intraobserver Variability: Should We Worry?

Pete Bridge MSc[a*], Andrew Fielding PhD[a], Pamela Rowntree FIR[a], and Andrew Pullar MBBS[b]

[a] *School of Chemistry, Physics and Mechanical Engineering, Queensland University of Technology, Brisbane, Queensland, Australia*
[b] *Department of Radiation Therapy, Radiation Oncology Mater Centre, Brisbane, Queensland, Australia*

Keywords: Radiotherapy; outlining; variability

Introduction

Many papers have identified concerns about intraobserver variability of repeat outlining by the same clinician. These variations in individual performance in turn make it challenging to determine values for interobserver 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 planning target volume margin and thus minimization of this has the potential to reduce normal tissue dose and morbidity. One accepted measure of intraobserver 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 intraobserver and interobserver variability in the interpretation of CT imaging data. Table 1 summarizes the results of these studies from the last 10 years.

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, center 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 intraobserver variability in radiotherapy outlining is a problem [5, 6] and the requirement to assess multiple slices independently makes it extremely difficult to exclude intraobserver variability from the process.

Most of the "nonoutlining" studies are also characterized 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 et al [6] for example, acknowledges the lack of clinical target volume 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 anatomic 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 et al's [6] educational intervention that included use of established guidelines and group feedback.

Conflicts of Interest: There are no conflicts of interest.

* Corresponding author: Pete Bridge, MSc, Faculty of Science and Engineering, Queensland University of Technology, Brisbane, Queensland 4001, Australia.

E-mail address: peter.bridge@connect.qut.edu.au (P. Bridge).

http://dx.doi.org/10.1016/j.jmir.2016.06.004

Table 1
Best Reported Kappa for Intraobserver 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	Hemorrhage	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 hemorrhage	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

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 utilized more widely but also admitted a lack of ''ground truth''.

While this certainly suggests that guidelines from cooperative groups such as Radiation Therapy Oncology Group [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 judgment. A recent paper attempting to evaluate guidelines for RTOG brachial plexus outlining [9] interpreted continuing intraobserver 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.

Intraobserver 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 two 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 substandard 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 utilized.

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 summarize the realistic expectations for intraobserver variability in this scenario and discuss the extent to which this is an issue. It adopts an epistemologic approach to the issue to postulate a new variability paradigm and aims to highlight the deeper philosophical issue underlying intraobserver variability to determine whether intraobserver variability can actually be eliminated and, more fundamentally, whether it actually matters.

Discussion

From an epistemologic 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 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 et al [11] recently underscored the importance of involving the clinician in the outlining process to utilize 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 intraobserver 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