Title Page

Title: Toward consistent design and reporting of observer studies in imaging

Dr John D. Thompson FHEA PhD MSc BSc (Hons)

University of Salford

Department of Radiography
University of Salford
Frederick Road Campus
M6 6PU
UK

j.d.thompson@salford.ac.uk
+44 (0) 161 295 6509
@radiologyjohn

Word Count: 1035

References: 7

Toward consistent design and reporting of observer studies in imaging

Introduction

In clinical practice, the work of diagnostic radiologists is a visual process based on a successful search of complex images, recognition of abnormalities, and then making a correct decision based on the information presented to them – in essence classifying such abnormalities as benign (or insignificant to health) versus malignant (or potentially harming a patient's health).

Similarly, radiology researchers need to conduct research studies that parallel this diagnostic process. Such studies seek to determine, for example, if a new MRI technique might be better than CT in answering a clinical question. Prior to the expense or difficulty of getting a new technique introduced, there needs to be some certainty that it will offer some advantage over the existing technique. The results of radiology reader studies are presented as a combination of sensitivities and specificities of the radiology readers for the new technique (MRI) and existing technique (CT). The final step involves statistical analysis and construction of the receiver operating characteristic (ROC) curve; the area under the ROC curve has a maximum value of 1 – indicating perfect sensitivity and specificity of the imaging test. Despite seeming straightforward, research studies can be remarkably complex for the radiology researcher to execute.

Objective observer studies using receiver operating characteristic (ROC) methods of this type have maintained their preeminent role in the evaluation of new imaging modalities and techniques. Hundreds of observer studies have been published in Radiology and other medical imaging journals, with variation in image datasets and the number and expertise of observers contributing to differences in statistical power. Despite the availability of software to analyze data from observer studies, such as *Rjafroc*[1] and *VGCAnalyzer*,[2] there is inconsistency in the reporting of the statistics. This can make the comparison of different studies difficult.

Image display and capture of the radiologist's response is another aspect of observer studies that has been inconsistently applied in the literature. In this issue of Radiology, Genske and Jahnke[3] describe a new tool, Human Observer Net, that has potential to allow a more consistent approach to image display and response capture. Because the software is open-source and web-based, it has the potential to be useful to more radiology investigators. If investigators can adequately control viewing conditions in multiple locations, the software may allow radiology readers in different locations to contribute to the same study. This may go some way to overcoming one of the greatest challenges in observer studies – radiologist participation.

The authors have recognized an opportunity to develop a consistent, readily available, and flexible solution to image display and response capture for observer studies. While other platforms are available,[4,5] they are either not open source or platform independent. Another key advantage of a web-based application is version control – there is no worry of researchers using old versions of the application. The new software provided by Genske and Jahnke[3] provides some evidence of validation, with several publications having successfully used the platform for multiple alternative forced choice and location ROC methods. While we await further validation for free-response ROC and visual grading analysis, interested readers can gain familiarity with the software by accessing a trial version of the software.

As part of the validation, Genske and Jahnke present two metrics that summarize the success of the software: system usability scale and reading time. The authors reported a grade A rating for system usability. Having accessed the trial version of the software, I would tend to agree with this for the available examples. In terms of reading time, this is perhaps not as useful as the time to decision in an observer study will be dependent on the task. However, the relatively short time-per-image reported is promising as observer studies can be time consuming. A future area for the authors to investigate would be comparing the time to first fixation, as judged by eye-tracking, to the additional time required to make the localization in the software.

Using the trial version of software with the link provided, it is possible to gain a very good understanding of how it works. Combined with the useful supplemental material it is relatively straightforward to understand what is happening in the background. It was not possible to test some of the functionality with the datasets present. It was not possible to adjust window width and level or scroll through multi-stack images. This is a shame as it could be valuable to know the window width and level at which a lesion had been localized. In addition, a volumetric region of interest covering multiple images in a stack could also be an interesting development. Otherwise, the images are presented well with smooth transition, while the user interface is clean and uncomplicated. The basic principles of functionality explained by Genske and Jahnke are supported well by this trial version of the software.

Currently the software will provide raw data for the researcher, with a single line of data provided for each observer response. With reference to output information provided by the authors it is not clear whether true and false decisions in a location-based study will be separated or whether it is the role of the researcher to determine this based on the intersection over union. It would be useful if a value of intersection over union could be set to determine true and false localizations and reduce the amount of manipulation required before analysis in third party software. Harmonizing data outputs with the required input of popular analysis software would be a welcome development.

Ultimately, the success of any software for observer studies, either image display and response capture or data analysis, will be determined by the end user. Current levels of functionality look promising, and it is hoped that the authors will be receptive to developing the software further to suit methodological choices of other researchers.

For those radiologists interested in designing and running an observer study, Human Observer Net will be a good tool to support the research. I would additionally note that the use of any software must be underpinned by relevant expertise to ensure the design is suitable and the execution is optimal. Below, I have provided references for key readings[6,7] to support your understanding of observer studies and to improve your choice of software for such studies.

References

1. Chakraborty DP. Observer Performance Methods for Diagnostic Imaging: Foundations, Modeling, and Applications with R-Based Examples: CRC Press, 2017.
2. Båth M, Hansson J. VGC Analyzer: A software for statistical analysis of fully crossed multiple-reader multiple-case visual grading characteristic studies. Radiat Prot Dosimetry. 2016 Jun;169(1-4):46-53. doi: 10.1093/rpd/ncv542.
3. Genske U, Jahnke P. Human observer net: a platform tool for human observer studies of image data. Radiology 2022;##(##):###–###.
4. Håkansson M, Svensson S, Zachrisson S, Svalkvist A, Båth M, Månsson LG. VIEWDEX: an efficient and easy-to-use software for observer performance studies. Radiat Prot Dosimetry 2010;139(1-3):42-51. doi: 10.1093/rpd/ncq057
5. Thompson J, Hogg P, Thompson S, Manning D, Szczepura K. ROCView: prototype software for data collection in jackknife alternative free-response receiver operating characteristic analysis. Br J Radiol 2012;85(1017):1320-1326. doi: 10.1259/bjr/99497945
6. Båth, M & Månsson, L. Visual grading characteristics (VGC) analysis: A non-parametric rank-invariant statistical method for image quality evaluation. The British journal of radiology. 2007;80. 169-76. 10.1259/bjr/35012658.
7. Chakraborty, D. A Brief History of Free-Response Receiver Operating Characteristic Paradigm Data Analysis. Academic radiology. 2013;20. 10.1016/j.acra.2013.03.001.

Author Biography

Dr John D. Thompson is a Lecturer in Radiography at the University of Salford and a Senior Radiographer in Nuclear Medicine at the University Hospitals of Morecambe Bay NHS Foundation Trust, UK. His research interests focus on observer performance studies using free-response methodology and visual grading. He has previously obtained grant funding from the Radiological Research Trust and is an Associate Editor for Journal of Medical Imaging and Radiation Science.

@radiologyjohn

j.d.thompson@salford.ac.uk