Clinicians must lead the development, standardised evaluation and adoption of AI in clinical practice: a case study from regional anaesthesia

Authors

James S Bowness[1,2], Xiaoxuan Liu[3,4], Pearse A Keane[5,6]

Author Institutions

1. Nuffield Department of Clinical Anaesthesia, University of Oxford, Oxford, UK
2. Department of Anaesthesia, Aneurin Bevan University Health Board, Newport, UK
3. University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK
4. College of Medical and Dental Sciences, University of Birmingham, UK
5. Institute of Ophthalmology, Faculty of Brain Sciences, University College London, London, UK
6. NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust, London, UK

Corresponding Author

James S Bowness

OxSTaR Centre, Nuffield Division of Anaesthetics, Nuffield Department of Clinical Neurosciences,

John Radcliffe, Hospital, Oxford, OX3 9DU

Tel: 0044 1865 231 510. E-mail: james.bowness@jesus.ox.ac.uk

Author contributions

Manuscript preparation & editing: all authors

ORCID ID

James Bowness https://orcid.org/0000-0002-8665-1984

Xiaoxuan Liu https://orcid.org/0000-0002-1286-0038

Pearse A Keane http://orcid.org/0000-0002-9239-745X

Twitter

James S Bowness @bowness_james, Xiaoxuan Liu @DrXiaoLiu, Pearse A Keane @pearsekeane.

Acknowledgements & Competing Interests (including Funding, Employment and Financial Interests)

JSB is supported by a Maurice Freeman Barema & Association of Anaesthetists Grant (HMR03690/HM00.01) and a University of Oxford stipend (Intelligent Ultrasound Limited; IUL); Cardiff, UK). JSB is a Senior Clinical Advisor for IUL, receiving research funding and honoraria. Outside of this work, XL is a Health Studies Scientist at Apple. She has previously acted as a consultant for Hardian Health, and has received research funding from NIHR, the NHS AI Lab, the Health Foundation, and the MHRA. PAK has acted as a consultant for DeepMind, Roche, Novartis, Apellis and BitFount, and is an equity owner in Big Picture Medical. He has received speaker fees from Heidelberg Engineering, Topcon, Allergan and Bayer. PAK is supported by a Moorfields Eye Charity Career Development Award (R190028A) and a UK Research & Innovation Future Leaders Fellowship (MR/T019050/1).

Running Header

Clinicians must lead the development, evaluation and adoption of medical AI

Keywords

Artificial intelligence; medical devices; standardisation; evaluation; regulation; ultrasound; regional anaesthesia

Summary

Suissa et al explore the clinical relevance of a medical image segmentation metric (Dice metric) commonly used in the field of artificial intelligence (AI). The investigators showed that pixel-wise agreement for physician identification of structures on ultrasound images is variable, and a relatively low Dice metric (0.34) correlated to a substantial agreement on subjective clinical assessment. This manuscript highlights the need to bring structure and clinical perspective to the evaluation of medical AI, which clinicians are best placed to direct.

Main article

Background

Ultrasound image guidance for regional anaesthesia was first described in 1989.[1] It has been described as the "*most important advance in regional anaesthesia practice of the new millennium*"[2] and is now the predominant technique used to direct the targeted blockade of peripheral nerves.[3] The accurate identification of key sono-anatomical structures is central to ultrasound-guided regional anaesthesia (UGRA)[4] and artificial intelligence (AI) has been proposed as a technology to support this.[5, 6] Several commercially-available systems have recently emerged, which segment (divide the images into multiple parts or regions of a certain class) ultrasound using a colour overlay.[7-10] The intent is to draw attentional gaze to key structures which, in turn, will support the acquisition of an appropriate ultrasound view and improve identification of those structures on that view. As with any innovation in healthcare, it is important to robustly interrogate the system(s) before applying them to patients. To make an informed decision on clinical use and potential patient benefit, one must understand the accuracy of structure identification, whether this provides benefit to the **operator**, and whether this benefit translates to improved patient care. However, evaluations of this type of AI model are typically limited and heterogenous, and there is often little publicly available external validation data for commercial systems. Academic publications from members of the technology fields in medical AI are often early stage, proof of concept studies, which are limited in scope.[11] They typically **use metrics which** assess pixel-wise agreement of the model/system output and that of a ground truth (interpretation that is assumed to be true, based on assessment of the images by experts in the field) for individual image frames.[11] This is not representative of the overall performance of system as ultrasound machines have a refresh rate of 25-30 frames/second and statistical significance for pixel agreement does not necessarily take into context clinical significance. Evaluation from a clinical perspective typically involves experts visualising image frames or videos;[11] this provides clinical significance and assesses performance over the course of a video of ultrasound footage, but can lack the **detail provided by** metrics typically used in technology fields.

The Current Study & its Potential Impact

Suissa et al attempt to address the disparity in these approaches through a study for which they recruited 10 anaesthetists, with varying levels of expertise, to each segment 10 images from seven selected nerve blocks. Most images were sourced from 217 YouTube (San Bruno, USA) videos, whilst the dataset for one region (femoral nerve block) was taken from an open-source repository on GitHub (San Francisco, USA). The investigators used the Dice metric to compare pixel-wise agreement of segmentation for all permutations of label pairs (**Figure 1**). Subsequently, segmentations for one image were overlayed on the relevant image and shown to participants, who were asked "*regarding the <nerve region> would you guide the needle to the highlighted area to block this region?*" The yes/no answers were converted to 1 (yes) and 0 (no) scores, then used to calculate correlation between the pixel-wise Dice score and the clinical assessment. **Dice score of 0.34 was found to be a threshold for agreement between pixel-wise score and clinical assessment**. Though they do not evaluate an AI model itself, the authors conclude that objective, pixel-wise metrics are best placed to determine AI model performance. **However, they also state their belief that this emphasises** "*high [metric] scores without critically translating model performance to real-world clinical significance*" and relatively low Dice metric scores correlated to clinical agreement.

Research in this area, linking technological progress to clinical practice, is highly important during an era where technology will redefine our lives.[6] Often, studies evaluating pixel-wise model performance have been published but made little clinical impact due to the avenue through which they have been published and the lack of clinician involvement/clinical context.[11] **Suissa et al's** study is a novel and useful approach to bring clinical meaning to objective assessment. However, we feel the study's impact is constrained by several limitations. Firstly, the rationale for choice of nerve block regions and sono-anatomical structures assessed for each region is unclear. The images selected were done so by a single (non-clinician) study investigator with the intent to "*feature differences in anatomical*

depositions so as not to include similar-looking images". However, standardisation has become a prevalent theme in UGRA clinical practice and research,[12-14] and international consensus work has recently been published on the identification of sono-anatomical structures for specific blocks.[15, 16] Many of these works are endorsed by international specialist societies in regional anaesthesia and consistency in the regions/structures considered would benefit **Suissa et al's** study and future investigations. A second limitation is the source of ultrasound images/videos used for the study. There is little information provided regarding the provenance of ultrasound scans and images used (though the femoral nerve block region images are available on an open-access repository for those with the inclination to investigate). The lack of information and standardisation regarding ultrasound machine/probe and demographic data for subjects of the scans limits validation and generalisability of the study findings. Thirdly, the authors of this editorial believe that choice of accuracy metric is not optimal for all structures in the study. The Dice metric is useful for considering enclosed areas (e.g., when annotating the outer border of a blood vessel, nerve or muscle). However, **Suissa et al** include the transversus abdominis plane, PECS I [interpectoral], and rectus sheath blocks in their work. These are fascial plane blocks, which aim to identify a plane of tissue between muscles and/or fascial boundaries, therefore comparing the accuracy of identifying the relevant planes would be better performed using a metric which assesses the identification of lines (e.g., Hausdorff metric).[17]

Based on the summary above, we believe that the study in question addresses an important question and the authors deserve credit for the elegant link between pixel-wise assessment and clinical relevance. However, the impact of these findings is stifled by the study **methodology**, which we believe **includes issues which are representative of research** in the healthcare AI field.

Future Directions

Artificial intelligence has profoundly impacted our daily lives and has the potential to revolutionise healthcare, yet its impact on clinical practice has so far been limited.[18] This is forecast to change;[6] as

it does, it is essential that novel AI devices undergo thorough and transparent scrutiny to fully understand their strengths and limitations, appropriate deployment in clinical practice, and ensure patient safety. The paper by **Suissa et al** serves to highlight several important issues in the field of translational AI in healthcare, which currently hamper the realisation of this potential. As UGRA is an early field within anaesthesia where AI is beginning to emerge, we feel appropriate structure, rigour and transparency can be established here as an exemplar for other areas.

Framework for Evaluation

A key limiting factor, in our opinion, is the lack of a standardised framework for the evaluation of such devices. The authors have personal experience of the regulatory approval process in the UK, Europe and the USA, where the same data considered sufficient for approval in one region has been deemed insufficient in another. This presents a patient safety concern as commercially available devices have gone through different regulatory approval processes, requiring different data and scrutiny, in different territories. In addition, this hinders clinician understanding of the performance and utility of available devices, which in turn negatively impacts their appropriate adoption in clinical practice. It is impossible to fully understand the comparative performance of two devices that have been tested on different structures and body regions, using different metrics, with scans from different machines and different patients. We propose that a standardised evaluation framework be developed for AI devices and suggest essential elements of this in the context of devices for ultrasound scanning for regional anaesthesia (**Figure 2 & Table 1**).

Standardised Open-access Database for Transparent Evaluation

The framework should include creation of an open-access set of ultrasound scans for peripheral nerve blocks and sono-anatomical structures for those blocks, to provide a standardised dataset for accuracy assessment. This should cover at least a set of agreed high value procedures, commonly taught in anaesthesia training programmes and not techniques solely employed by domain experts. On such

example is the seven "Plan A Blocks" proposed by **Turbitt et al, 2020**.[14] There must be agreement on the structures to be included in views selected for each scan, such as those proposed in a recent international consensus of opinion on this subject endorsed by the UK, European and American specialist societies in regional anaesthesia.[16] Scans must include subjects with a diversity of factors affecting ultrasound appearance (e.g., anatomical variation, body habitus and age) to be able to fully investigate performance of AI systems across the full spectrum of the population. This metadata should be provided in the database, to allow subgroup analyses in specific populations.

Accuracy Assessment

Once an appropriate database of ultrasound images is established, agreed pixel-wise and clinical approaches to assessing accuracy should be determined. This could include use of the Dice metric for pixel-wise assessment, but others (e.g., intersection over union) should be considered. Input from experts in the technical fields of AI, such as computer science and engineering, will be essential in this regard. Equally, appropriate methods of providing clinical evaluation should also be agreed. This could include existing methodologies such as rates of true/false positive/negative. In a previous study, **Bowness et al** used these data to report the true/false positive/negative rates (TPr, TNr, FPr, FNr), accuracy (TPr + TNr), and error (FPr + FNr).[19] We recommend that reporting also include sensitivity, specificity, positive predictive value and negative predictive value. However, it is highly desirable to avoid adding unnecessary complexity to these analyses, to ensure the information is accessible and understandable to all. Importantly, using both pixel-wise and clinical approaches for the same ultrasound scan dataset would facilitate the synchronous presentation of objective pixel-wise assessment with subjective clinical context, which will support a detailed, complete, and clinically-grounded evaluation.

An agreed ground truth is essential to provide alongside the scans – a uniform set of scans is not meaningful without a uniform benchmark to compare models against. This is applicable to both the

pixel-wise assessments, the ground truth for which could be provided in the open-access database, and the level of experience/expertise required in assessors for clinically orientated evaluations (e.g., define level of qualification/experience required of an assessor). Finally, as there is no agreed consensus on what level of performance is considered acceptable for an AI device,[20] this should be established – at least in some form of recommendation if not absolute.

Utility Assessments

After establishing accuracy, researchers and companies should follow recommended approaches to evaluating benefits. This should involve identifying benefits to patients (e.g., increased block efficacy), which practitioners will benefit most (e.g., novice vs expert)[21] and in what manner (e.g., improved block efficacy vs improved patient access).[22] Demonstrating these benefits in the simulation setting first, followed by real-world clinical data, may help mitigate any risk to patients posed by untested technology. Furthermore, outcome data should include endpoints which address the patient safety profile of new devices. Institutional benefits and health economic evaluations are equally important, but potentially lie outside the scope of this framework.

Reporting

As well as standardised evaluations, findings must be consistently reported according to an agreed standard. Guidelines exist for different study formats,[23] many of which have AI extensions, and it may be most appropriate to select an appropriate format for each different study or develop a specific reporting guideline for the use case of segmentation of ultrasound images in regional anaesthesia. However, ensuring this is consistent with a single overarching standard for reporting all evaluations, such as the *Develop*mental *and Exploratory Clinical Investigations of Decision Support Systems driven by Artificial Intelligence* (DECIDE-AI),[24] would again support the consistent generation and transfer of information.

The Role of Industry & Regulators

Most AI systems used in clinical practice will be developed, evaluated and brought to market by commercial organisations. The role of the technology industry in this field is inadvertently highlighted in the study by **Suissa et al**, who use resources provided by technology companies – YouTube, owned by Google (Mountain View, USA) and GitHub, owned by Microsoft (Redmond, USA). The resources available to such companies ensure that the commercial sector will play a major role in the evolution of AI in healthcare. It is essential that clinicians engage with industry to direct effort and resources to appropriate clinical endpoints and adopt the same standards of evaluation and reporting.

In our experience, there can be a lack of consistency between regulatory requirements and/or their implementation across jurisdictions. In addition, there is a no requirement to make regulatory approval data publicly available, and collaboration with this sector should seek to encourage standardisation and visibility of data. Furthermore, respected journals should be encouraged to require investigators to conduct and report their work in line with the above framework, to establish it as standard practice for academia and industry.

Led by Clinicians with Oversight by a Responsible Body

In developing such a framework, contribution from technology experts, academia, industry, regulatory authorities, and medical bodies overseeing clinical practice is essential. Importantly, clinicians must lead this process, to ensure the development, evaluation and deployment of AI technology addresses pain points relevant to clinical practice. This is not to say that clinicians should become experts in the respective fields contributing to medical AI, rather that guiding development in academia/industry, then conducting and reporting evaluations with a clinical perspective is an essential element in the toolbox of translational science for AI implementation. This approach will facilitate understanding in the clinical community, which will in turn support appropriate adoption.

With such wide-ranging contributions and requirements, these efforts should be undertaken by a working party under the remit of a medical body with responsibility for standards of practice, in a similar manner to which the Association of Anaesthetists make recommendations for the standards of monitoring for anaesthesia and recovery.[25] This would benefit from collaboration with specialist societies and learned bodies, and organisations which inform medical practice in respect to new treatments and technologies in the respective territory – such as the UK's National Institute for Health and Care Excellent, which already has guidance on appraising medical technology (though not specifically for AI).[26]

Conclusions

Anaesthesia is a data-rich specialty, from risk stratifications systems, continuous or near-continuous monitoring of multiple physiological parameters, infusion algorithms, imaging, and frequent calculations/predictions of resource requirements (e.g., emergency theatre use, duration of cases, requirement for critical care). Therefore, AI has huge potential to impact our practice. The study by **Suissa et al** offers a potential advance in approaches to evaluation of AI for ultrasound image interpretation in regional anaesthesia, but also highlights limitations prevalent in the field at present. This framework outlined above seeks to address these. Whilst it is relevant to UGRA, elements may be adopted elsewhere, and similar frameworks should be established for different anaesthesia subspecialties. Clinicians are central to shaping the new era of AI-enabled healthcare; it is time for us not to just seek education and partake, but to lead the process and enable this technology to realise the potential benefits it holds.

References

1 Ting PL, Sivagnanaratnam V. Ultrasonographic study of the spread of local anaesthetic during axillary brachial plexus block. *Br J Anaesth* 1989; **63**: 326-9

2 Abdallah FW, Macfarlane AJ, Brull R. The Requisites of Needle-to-Nerve Proximity for Ultrasound-Guided Regional Anesthesia: A Scoping Review of the Evidence. *Reg Anesth Pain Med* 2016; **41**: 221-8

3 Neal JM, Brull R, Horn JL, et al. The Second American Society of Regional Anesthesia and Pain Medicine Evidence-Based Medicine Assessment of Ultrasound-Guided Regional Anesthesia: Executive Summary. *Reg Anesth Pain Med* 2016; **41**: 181-94

4 Sites BD, Chan VW, Neal JM, et al. The American Society of Regional Anesthesia and Pain Medicine and the European Society Of Regional Anaesthesia and Pain Therapy Joint Committee recommendations for education and training in ultrasound-guided regional anesthesia. *Reg Anesth Pain Med* 2009; **34**: 40-6

5 Bowness J, El-Boghdadly K, Burckett-St Laurent D. Artificial intelligence for image interpretation in ultrasound-guided regional anaesthesia. *Anaesthesia* 2021; **76**: 602-7

6 McKendrick M, Yang S, McLeod GA. The use of artificial intelligence and robotics in regional anaesthesia. *Anaesthesia* 2021; **76 Suppl 1**: 171-81

7 https://www.gehealthcare.com/products/ultrasound/venue-family/regional-anesthesia. Accessed 17.072023 - 21.09.2023.

8 https://www.intelligentultrasound.com/scannav-anatomy-pnb/. Accessed 17.07.2023 - 21.09.2023.

9 https://www.mindray.com/uk/products/ultrasound/point-of-care/tex20-series. Accessed 17.07.2023 - 21.09.2023.

10 https://www.nerveblox.com/. Accessed 17.07.2023 - 21.09.2023.

11 James SB, David M, Kariem E-B, Neal T, Noble JA, Helen H. Artificial Intelligence for Anatomical Structure Identification on Ultrasound in Regional Anaesthesia: A Scoping Review Protocol. *medRxiv* 2023: 2023.07.04.23291560

12 Ahmed HM, Atterton BP, Crowe GG, et al. Recommendations for effective documentation in regional anesthesia: an expert panel Delphi consensus project. *Reg Anesth Pain Med* 2022; **47**: 301-8

13 El-Boghdadly K, Wolmarans M, Stengel AD, et al. Standardizing nomenclature in regional anesthesia: an ASRA-ESRA Delphi consensus study of abdominal wall, paraspinal, and chest wall blocks. *Reg Anesth Pain Med* 2021; **46**: 571-80

14 Turbitt LR, Mariano ER, El-Boghdadly K. Future directions in regional anaesthesia: not just for the cognoscenti. *Anaesthesia* 2020; **75**: 293-7

15 Ashken T, Bowness J, Macfarlane AJR, et al. Recommendations for anatomical structures to identify on ultrasound for the performance of intermediate and advanced blocks in ultrasound-guided regional anesthesia. *Reg Anesth Pain Med* 2022; **47**: 762-72

16 Bowness JS, Pawa A, Turbitt L, et al. International consensus on anatomical structures to identify on ultrasound for the performance of basic blocks in ultrasound-guided regional anesthesia. *Reg Anesth Pain Med* 2022; **47**: 106-12

17 Lloyd J, Morse R, Taylor A, et al. Artificial Intelligence: Innovation to Assist in the Identification of Sono-anatomy for Ultrasound-Guided Regional Anaesthesia. *Adv Exp Med Biol* 2022; **1356**: 117-40

18 Keane PA, Topol EJ. AI-facilitated health care requires education of clinicians. *Lancet* 2021; **397**: 1254

19 Bowness JS, Burckett-St Laurent D, Hernandez N, et al. Assistive artificial intelligence for ultrasound image interpretation in regional anaesthesia: an external validation study. *Br J Anaesth* 2023; **130**: 217-25
20 van der Meijden SL, Arbous MS, Geerts BF. Possibilities and challenges for artificial intelligence and machine learning in perioperative care. *BJA Educ* 2023; **23**: 288-94
21 Bowness JS, El-Boghdadly K, Woodworth G, Noble JA, Higham H, Burckett-St Laurent D. Exploring the utility of assistive artificial intelligence for ultrasound scanning in regional anesthesia. *Reg Anesth Pain Med* 2022; **47**: 375-9
22 Bowness JS, James K, Yarlett L, et al. Assistive artificial intelligence for enhanced patient access to ultrasound-guided regional anaesthesia. *Br J Anaesth* 2023
23 https://www.equator-network.org/. Accessed 12.10.2023.
24 Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nature Medicine* 2022; **28**: 924-33
25 Klein AA, Meek T, Allcock E, et al. Recommendations for standards of monitoring during anaesthesia and recovery 2021: Guideline from the Association of Anaesthetists. *Anaesthesia* 2021; **76**: 1212-23
26 https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/nice-technology-appraisal-guidance. Accessed 11.10.2023.
27 Bowness JS, Morse R, Lewis O, et al. Variability between human experts and artificial intelligence in identification of anatomical structures by ultrasound in regional anaesthesia: a framework for evaluation of assistive artificial intelligence. *British Journal of Anaesthesia* 2023

Table 1. Essential elements of a standardised framework for evaluating AI devices which segment ultrasound scans for regional anaesthesia

Standardised open-source database of ultrasound scans
○ Approved by appropriate body/bodies (e.g., body overseeing clinical practice or specialist society) ○ Include at least the seven "Plan A Blocks"[14] ○ Include the recommended minimum structures to identify on ultrasound[15, 16] ○ Scanned subjects should include a specified range of age and BMI ○ Agreed ground truth for accuracy
Accuracy Assessments
○ Objective pixel-wise assessment on individual images using agreed metrics (e.g., Dice, Haudorff) ○ Subjective clinical assessment on ultrasound video by a panel of experts ○ Defined thresholds for acceptable performance
Utility Assessments

○ Data on which cohort of medical professionals will benefit and how
○ Demonstration of benefit in a simulation setting
  ○ Potential **operator** benefits: accuracy of scanning, speed of scanning, confidence
○ Demonstration of benefits in clinical setting
  ○ **Potential patient benefits: increased access to UGRA, reduced pain scores, reduced opioid usage/side effects**
  ○ **Potential staff/institutional benefits: shortened time in theatre recovery/hospital, ambulatory discharge**
○ Demonstration of safety
  ○ AI-assisted techniques must be at least non-inferior in safety profile (e.g., incidence and/or severity of adverse events)

Reporting
○ Standardised reporting using an agreed guideline (e.g., DECIDE-AI[24]) ○ Academic journals and industry standards to require consistent reporting structure

Figure 1. Illustration of the Dice metric (taken from Bowness et al, 2023[27])

Figure 2. Standardisation, accuracy and utility are the pillars of robust evaluation, which can inform appropriate adoption in clinical practice

(Elements of figure sourced from Bowness et al, 2023[19])