Interpretation and Reporting of Predictive or Diagnostic Machine Learning Research in Trauma & Orthopaedics

Authors: L.Farrow[1,2] – Clinical Research Fellow *MBChB, Bsc(Intercalated), MRCS*, M.Zhong[1] – Clinical Lecturer in Machine Learning *PhD*, L.Anderson[1] – Chair in Health Data Science *PhD MPHe BSc(Hons) PGCHET FHEA*, G.P.Ashcroft[1,2] – Senior Clinical Lecturer *Bsc, MBChB, FRCSEd (Tr&Orth)*, R.M.D Meek[3] – Consultant Orthopaedic Surgeon *MBChB, BSc (Hons), MD, FRCS (Tr & Orth)*

1. University of Aberdeen, Aberdeen, United Kingdom
2. Aberdeen Royal Infirmary, Aberdeen, United Kingdom
3. Queen Elizabeth University Hospital, Glasgow, United Kingdom

Conflict of Interest: Luke Farrow, Mingjun Zhong, Lesley Anderson, George Patrick Ashcroft declare that they have no conflict of interest. R. M. D. Meek reports board membership by The Bone & Joint Journal, and payment for lectures (including service on speakers' bureaus) from Palacademy and Johnson & Johnson, all of which are unrelated to this article.

Funding: None

Corresponding author

Luke Farrow

Institute of Applied Health Sciences

University of Aberdeen

Foresterhill, Aberdeen, AB25 2ZD

Scotland, United Kingdom

Tel: +44 (0) 1224 437841

ORCID: 0000-0002-1443-5908

Luke.farrow@abdn.ac.uk

Interpretation and Reporting of Predictive or Diagnostic Machine Learning Research in Trauma & Orthopaedics

Abstract

There is increasing popularity in the use of Artificial Intelligence and Machine Learning techniques to provide diagnostic and prognostic models for various aspects of Trauma and Orthopaedic surgery. Correct interpretation of these models is however difficult to those without specific knowledge of computing or health data science methodology. Lack of current reporting standards leads to potential for significant heterogeneity in the design and quality of published studies. We provide an overview of Machine Learning techniques for the “lay” individual, including key terminology and best practice reporting guidelines.

Introduction

Trauma and Orthopaedic research using Artificial Intelligence (AI – performance of computer tasks that usually require human intelligence), and more specifically a subset of AI called Machine Learning (ML – computer algorithms that automatically improve performance through experience), has become increasingly popular over the last decade, with a rapid rise seen in the number of related publications (Figure 1). This is likely related to the widespread use of patient registries and other sources of electronic health information that provide significant quantities of readily accessible data to which ML techniques can be applied. The ability to manage complex healthcare data is a key feature of ML techniques and hence why it is often seen as having significant potential in the field of Trauma & Orthopaedics,[1,2] where clinical decision making often relies on integrating multiple sources of healthcare information.

ML applications to Trauma & Orthopaedics mainly relate to predictive or diagnostic classification of patients that seek to aid timely diagnosis of adverse outcomes or provide important prognostic information about patient outcomes that can support medical decision-making.

With the increasing number of studies relating to ML published within mainstream orthopaedic journals, “lay” readers will inevitably find difficultly in interpreting results, where knowledge and training in computing & health data science methodology are required to conduct and interpret research in this field. A current lack of reporting standards for non-interventional AI & ML research means there is potential for significant heterogeneity in the design and quality of published manuscripts.

In this article we therefore provide an overview of current commonly utilised ML techniques within the field of Trauma & Orthopaedics, including best practice recommendations for manuscript reporting for the ML researcher and a guide to interpretation for the "lay" clinician or scientist. Table 1 provides a glossary of important terminology to assist with key concepts and definitions.

Research Reporting

It is important to recognise that development of clinical prediction or diagnostic models such as those used in Trauma & Orthopaedics is a longitudinal process that requires significant oversight and input for clinically useful information to be produced. This initially starts with identification of the clinical problem to be addressed and runs through data extraction, cleaning and coding, to model development, testing and validation[3]. It is important to reflect that data pre-processing is perhaps the most critical step in this "pipeline" and encompasses over ¾ of the required work in producing a valid ML output.

The heterogeneity of reporting in ML approaches to prediction modelling has been recognised by others, both in the wider medical literature,[4,5] and also specifically within Trauma & Orthopaedics.[6,7]

Use of existing guidelines such as the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement[8] provide an excellent starting point for study reporting, but do not cover AI techniques (although an updated TRIPOD-AI is currently in development and should provide a definitive methodological overview for reporting prediction modelling using ML).[9]

Guidelines for the reporting of interventional studies involving AI are covered by the Consolidated Standards of Reporting Trials (CONSORT) AI Extension[10] and its complementary Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT) AI Extension.[11] They specifically highlight how the reporting of AI focused studies should differ from traditional trials, with a particular focus on how participant selection, data pre-processing and where AI is utilised in the intervention pathway. It is imperative that journals, authors and readers are aware of their existence and utilise these when considering interventional ML research in Trauma & Orthopaedics.

Our expert opinion recommendations for the reporting of prognostic or diagnostic ML models are included in Table 2.

Whilst this is not an exhaustive list, and one which may change over time as ML techniques develop, we believe it provides a minimum reporting guide for journals, authors and reviewers that will ensure that all studies are of sufficient quality that they provide meaningful additions to the current evidence base.

Interpretation of machine learning output – a guide for the "lay" audience

Most individuals involved in the clinical practice of Trauma and Orthopaedics will have minimal knowledge and experience of interpreting studies that report on ML output. This is perhaps one of the main reasons, alongside a lack of clear reporting, why the majority of published research in this area has yet to influence clinical practice.

In order to maximise the significant potential of ML in the field it is therefore imperative that healthcare professionals and policymakers have at least a basic understanding of how to correctly appraise and interpret ML studies. We therefore highlight key areas for output interpretation and important factors for consideration in study evaluation.

The first key step is to understand the data that is included within the model. ML algorithms follow the simple mantra that "garbage in = garbage out". It is important therefore that the data utilised is from a reliable source with good data quality, and ideally one that has been externally validated regarding its accuracy (e.g. automated electronic capture of healthcare data audited against manual searching of individual patient records).

Secondly it is important to understand the nature of the control group and gold standard reference to ensure they are appropriate if using supervised ML techniques (where the potential outcomes are predefined). For example, when considering the diagnosis of distal radius fractures: if a control group consists of young healthy controls and the case group consists of low-energy fragility fractures then then ML model may pick up on the presence of osteopenia in fracture subjects rather than the actual diagnosis.

It is also important to remember that ML output can only relate to the data that has been included within it. Therefore, a model that was trained on a specific dataset may not apply to other data, even within a similar field. For example, a ML algorithm to predict attainment of the Minimum Clinically Important Difference following knee arthroplasty may only relate to the specific implants utilised in the initial dataset. It therefore may not relate to other implant types (e.g. uncemented or unicondylar designs) if these were not commonly found within the data the model was trained on. This is why external validation of ML models is a vital part of the assessment process and a critical requirement before any claims to clinical applicability can be made.

Interpretation of the results from ML output is also vital to determining the importance of studies. Like many things in life, if the predicted accuracy of a ML model sounds too good to be true, it probably is.

One of the key results often offered in ML literature is the Area Under Curve (AUC) of the Receiver Operating Characteristic (ROC) metric (or Concordance/C-statistic). This is typically classified in a range from 0-1 and relates to the discriminatory capacity (predictive ability) of the model. An AUC of 1 therefore represents a model which is always correct in its prediction, and an AUC of 0 is always incorrect. An AUC of 0.5 represents a model which is no better than random chance at predicting a specific outcome. Though there are no fixed parameters for what constitutes an "acceptable" AUC generally 0.6-0.7 is considered moderate, 0.7-0.8 good, 0.8-9 excellent and 0.9+ outstanding. [12] AUCs should be reported with 95% Confidence intervals.

Understanding the calibration slope and intercept of a model is also integral to appreciating how well a ML model functions,[13] and is often missed from analyses.[14] Calibration is defined as the level of agreement between the predictive and observed values. [15] An example is shown in Figure 2. These assessments are important help determine how well the model fits the sample it is being applied to, including potential population sub-groups. A perfect model is one where the intercept (Alpha) is 0, and the slope (Beta) is 1.[3] Incorrect calibration may have serious consequences in terms of underestimating or overestimating risk for certain populations and is again another vital reason for external validation of ML models before determination of potential clinical applicability. For example, a machine learning algorithm to predict mortality using data from all hip fracture patients may underestimate mortality in those with pathological fracture related to malignancy if appropriate clinical categorisation is not considered during algorithm development. Further details regarding the specifics of calibration, including visual examples, can be found in *Van Calster et al* – Calibration: the Achilles heel of predictive analytics.[14] A Brier Score may also sometimes be presented as a way to assess model calibration, with a score of 0 the best achievable and 1 the worst.

A frequently used approach when comparing the ability of human observers and ML models to correctly identify objects (typically radiographs) is development of a "classification model". Here the Kappa index (Cohen's Kappa) is commonly presented.[16] This is a test of inter-observer reliability/agreement that helps determine the comparative performance of this type of model. Values range from 0-1, with 0.8 (80% agreement) typically seen as a minimum acceptable inter-rater agreement in the healthcare setting.[17] It is always important to remember in this setting that there is variability and error in human classification and therefore multiple observers are usually required to ensure accuracy of the "gold standard" comparator. A confusion matrix may also be presented where the predicted and actual values (+/- Kappa) for the two groups are compared. This can also be done where multiple different classes are involved (i.e., determining if radiographic arthritis is mild, moderate or severe).

Finally, it is important to consider common pitfalls for published ML models in Trauma & Orthopaedics, which will be aided by the reporting guidelines outlined above and in Table 2. This may include, but is not limited to:

- Lack of an adequate clinical question or hypothesis to test.
- Use of an inadequate data source e.g., poor data quality, large quantities of missing data, or insufficient data to address the clinical question (including confounding variables).
- Lack of external validation.
- Insufficient reporting of the results, including the discriminatory and classification capabilities of the model.
- Limited explanation or exploration regarding model deployment in clinical practice and how barrier to implementation can be addressed.

Conclusions

ML techniques are often unfamiliar territory for the orthopaedic specialist, with difficulty in determining the relevance of ML studies to clinical practice. ML has previously been described as a "black box", although new techniques that allow us to understand and interpret the decision making of ML algorithms are becoming increasingly prevalent, which will massively increase the potential for healthcare application in the future.[18] This understanding is essential given that ML algorithms may function in an abstract fashion if not correctly applied – for example diagnosis of fracture on X-ray using an abnormality marker applied by the radiographer at time of image acquisition, rather than classification based on the fracture itself.

We highlight important nomenclature to aid in understanding and key elements required in reporting and interpretation of prognostic and diagnostic ML output. The factors identified here are by no means exhaustive but should help support comprehension of the ever-increasing numbers of ML models now widely prevalent in orthopaedic research.

References

1. Jones LD, Golan D, Hanna SA, Ramachandran M. Artificial intelligence, machine learning and the evolution of healthcare: A bright future or cause for concern?. Bone & joint research. 2018 Mar;7(3):223-5.

2. Bayliss L, Jones LD. The role of artificial intelligence and machine learning in predicting orthopaedic outcomes. The bone & joint journal. 2019 Nov;101(12):1476-8.
3. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J. 2014;35(29):1925-31.
4. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. BMJ. 2020;369:m1328.
5. Li Y, Sperrin M, Ashcroft DM, van Staa TP. Consistency of variety of machine learning and statistical models in predicting clinical risks of individual patients: longitudinal cohort study using cardiovascular disease as exemplar. BMJ. 2020;371:m3919.
6. Haeberle HS, Helm JM, Navarro SM, Karnuta JM, Schaffer JL, Callaghan JJ, et al. Artificial Intelligence and Machine Learning in Lower Extremity Arthroplasty: A Review. J Arthroplasty. 2019;34(10):2201-3.
7. Oosterhoff JHF, Doornberg JN, Machine Learning C. Artificial intelligence in orthopaedics: false hope or not? A narrative review along the line of Gartner's hype cycle. EFORT Open Rev. 2020;5(10):593-603.
8. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350:g7594.
9. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet. 2019;393(10181):1577-9.
10. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK, Spirit AI, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. BMJ. 2020;370:m3164.
11. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ, Spirit AI, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Lancet Digit Health. 2020;2(10):e549-e60.
12. Hosmer D, Lemeshow S, Sturdivant R. Applied Logistic Regression. 3rd ed. Statistics WSiPa, editor. New Jersey: Wiley; 2013.
13. Stevens RJ, Poppe KK. Validation of clinical prediction models: what does the "calibration slope" really measure? J Clin Epidemiol. 2020;118:93-9.
14. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Topic Group 'Evaluating diagnostic t, et al. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):230.

15. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128-38.
16. Blackman NJ, Koval JJ. Interval estimation for Cohen's kappa as a measure of agreement. Stat Med. 2000;19(5):723-41.
17. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22(3):276-82.
18. The Lancet Respiratory M. Opening the black box of machine learning. Lancet Respir Med. 2018;6(11):801.