Received: 18 February 2024 | Accepted: 19 April 2024

DOI: 10.1002/jeo2.12039

REVIEW PAPER

Journal of Experimental Orthopaedics WILEY

A practical guide to the implementation of AI in orthopaedic research, Part 6: How to evaluate the performance of AI research?

Felix C. Oettl[1,2] | **Ayoosh Pareek**[3] | **Philipp W. Winkler**[4,5,6] | **Bálint Zsidai**[5,6] | **James A. Pruneski**[7] | **Eric Hamrin Senorski**[6,8] | **Sebastian Kopf**[9] | **Christophe Ley**[10] | **Elmar Herbst**[11] | **Jacob F. Oeding**[5,12] | **Alberto Grassi**[13] | **Michael T. Hirschmann**[14,15] | **Volker Musahl**[16] | **Kristian Samuelsson**[5,6,17] | **Thomas Tischer**[18,19] | **Robert Feldt**[20] | **ESSKA Artificial Intelligence Working Group**

[1]Hospital for Special Surgery, New York, New York, USA

[2]Schulthess Klinik, Zurich, Switzerland

[3]Sports Medicine and Shoulder Institute, Hospital for Special Surgery, New York, New York, USA

[4]Department for Orthopaedics and Traumatology, Kepler University Hospital GmbH, Johannes Kepler University Linz, Linz, Austria

[5]Department of Orthopaedics, Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

[6]Sahlgrenska Sports Medicine Center, Göteborg, Sweden

[7]Department of Orthopaedic Surgery, Tripler Army Medical Center, Honolulu, Hawaii, USA

[8]Department of Health and Rehabilitation, Institute of Neuroscience and Physiology, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

[9]Center of Orthopaedics and Traumatology, University Hospital Brandenburg an der Havel, Brandenburg Medical School Theodor Fontane, Germany

[10]Department of Mathematics, University of Luxembourg, Esch-sur-Alzette, Luxembourg

[11]Department of Trauma, Hand and Reconstructive Surgery, University Hospital Muenster, Muenster, Germany

[12]Mayo Clinic Alix School of Medicine, Mayo Clinic, Rochester, Minnesota, USA

[13]IIa Clinica Ortopedica e Traumatologica, IRCCS Istituto Ortopedico Rizzoli, Bologna, Italy

[14]Department of Orthopaedic Surgery and Traumatology, Kantonsspital Baselland, Bruderholz, Switzerland

[15]University of Basel, Basel, Switzerland

[16]Department of Orthopaedic Surgery, UPMC Freddie Fu Sports Medicine Center, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

[17]Department of Orthopaedics, Sahlgrenska University Hospital, Mölndal, Sweden

[18]Department of Orthopaedic Surgery, Universitymedicine Rostock, Rostock, Germany

[19]Department of Orthopaedic and Trauma Surgery, Malteser Waldkrankenhaus Erlangen, Erlangen, Germany

[20]Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg, Sweden

Abbreviations: AI, artificial intelligence; AUC-ROC, area under the curve-receiver operator curve; BLEU, bilingual evaluation understudy; DVT, deep venous thrombosis; FN, false negative; FP, false positive; LLM, large language models; MAE, mean absolute error; MAP, mean average precision; MAPE, mean absolute percentage error; MCC, Matthews Correlation Coefficient; ML, machine learning; NDCG, normalised discounted cumulative gain; RMSE, root-mean squared error; TN, true negative; TP, true positive; TRIPOD, transparent reporting of a multivariable prediction model for individual prognosis or diagnosis.

Correspondence
Kristian Samuelsson, University of Gothenburg, Göteborgsvägen 31, 431 80 Mölndal, Sweden.
Email: kristian.samuelsson@gu.se

Funding information
None

Abstract
Artificial intelligence's (AI) accelerating progress demands rigorous evaluation standards to ensure safe, effective integration into healthcare's high-stakes decisions. As AI increasingly enables prediction, analysis and judgement capabilities relevant to medicine, proper evaluation and interpretation are indispensable. Erroneous AI could endanger patients; thus, developing, validating and deploying medical AI demands adhering to strict, transparent standards centred on safety, ethics and responsible oversight. Core considerations include assessing performance on diverse real-world data, collaborating with domain experts, confirming model reliability and limitations, and advancing interpretability. Thoughtful selection of evaluation metrics suited to the clinical context along with testing on diverse data sets representing different populations improves generalisability. Partnering software engineers, data scientists and medical practitioners ground assessment in real needs. Journals must uphold reporting standards matching AI's societal impacts. With rigorous, holistic evaluation frameworks, AI can progress towards expanding healthcare access and quality.

Level of Evidence: Level V.

KEYWORDS
AI, digitalization, healthcare, ML, performance metrics

INTRODUCTION

All models are wrong, but some are useful.
George Box [4]

The rapid development of artificial intelligence (AI) has led to systems capable of predicting outcomes, analysing and reporting on data produced by imaging procedures, as well as generating creative works such as generated music, graphics or even art. However, as these powerful technologies advance, it is essential that their outputs are properly interpreted and evaluated, in particular, for applications in medicine. While high predictive capacity is desirable, values nearing or reaching 100% should be examined closely to rule data leakage or other methodological concerns. Very small sample sizes can make models prone to overfitting. Respective studies warrant careful scrutiny regarding their evaluation methodology and the appropriateness of any claims.

Machine learning (ML) models have shown promising predictive abilities for select tasks in medicine, such as screening skin lesions for cancer risk or predicting protein folding structures [16, 43]. However, ML models' ability to make reliable clinical judgements across all domains of medicine remains limited at this time. While ML algorithms can find patterns and correlations in data, correlations alone are not sufficient to justify clinical actions for specific patient care. To make credible clinical judgements, or even to support the decisions of medical staff, AI needs access to extensive medical data and an understanding of diagnostic methods, treatment options and surgical techniques grounded in clinical expertise. For example, an ML algorithm may discover gene mutations associated with increased disease risk, but qualified medical professionals must validate and interpret these findings to determine appropriate next steps for each patient and put these findings into context with what is already known about the pathophysiology and nature of the disease. More research is needed to develop AI systems with the reasoning capabilities necessary for sound clinical decision-making [8, 20, 21, 25, 44, 51–53, 55, 56].

To enable proper evaluation, researchers should report on the ML model's training process, data sources, modelling capabilities and performance [9, 31]. Comparisons to ground truth benchmarks, measured error rates and comparisons with human experts can contextualise the AI's performance. Additionally, AI outputs should be critically examined through peer review, replication studies and real-world, clinical testing before being integrated into standard practices. This process is known as external validation, where models are tested on data sets from different patient populations, geographic locations or hospitals to ensure that no bias was introduced during the internal training process. With thoughtful evaluation methods, AI can augment human performance and ameliorate patient outcomes. Standards for reporting and interpretation of AI and ML model performance can help in this, but it is important to acknowledge that evaluations may need to be reassessed if a model is used over time. The underlying population of patients, and thus the

data, might drift, potentially rendering the models obsolete.

ML MODELS AND METRICS

While generative AI (ChatGPT [OpenAI Inc.] probably being the most well-known in 2024) is becoming increasingly important in all fields of science, the three main applications utilising AI in healthcare remain classification, regression and clustering [11, 35, 47]. Other ML algorithms are often permutations and combinations of these underlying models. Commonly proposed algorithms in medicine that define themselves 'Ranking' or 'Forecasting' are permutations and/or combinations of Classification, Regression and Clustering models.

Classification

A common use of AI is to classify (e.g., X-rays) into discrete categories or labels [36]. For example, an image classifier may categorise X-rays as normal or abnormal [34]. Classification models output a predicted class, possibly along with a probability score reflecting the model's confidence. Key evaluation metrics are confusion matrices from which accuracy, precision, recall, specificity and F1-scores, as well as area under the curve–receiver operator curve (AUC-ROC) can be calculated [36]. Arguments have been made that if only a single metric is to be used, then the Matthews Correlation Coefficient (MCC) has many benefits since it summarises all the other basic rates (sensitivity, specificity, precision and negative predictive value) while AUC-ROC does not [6]. However, high scores on internal validation (commonly referred to as test set) do not necessarily mean that AI will generalise to real-world use [15, 54]. When assessing classification models, it is vital to critically evaluate how accurately the distribution of ground truth labels represents the true underlying prevalence across classes. Real-world data sets often exemplify class imbalance, where the positive disease cases comprise a disproportionately smaller fraction compared to the negative cohort. For example, in a prediction model for deep venous thrombosis (DVT) after surgery, only a fraction of the patient population will present with DVT, this is called an imbalanced data set.

Regression

Regression models predict continuous numeric values instead of discrete classes, such as patient length of stay based on clinical data [36]. Evaluation focuses on deviation from true values, using metrics like percentiles of errors, mean absolute error (MAE), mean absolute percentage error (MAPE) and root-mean-squared error (RMSE) [24, 36, 38]. However, solely chasing better numeric scores can overlook whether outputs are clinically meaningful [29].

Clustering

Evaluating the performance of unsupervised clustering algorithms requires metrics that quantify how well the clusters separate dissimilar observations and group similar ones [36]. A cluster refers to a set of observations that is more related to each other than to data points in other clusters. For example, a clustering algorithm may separate patients into distinct clusters based on symptoms and test results, with each cluster representing a potential undiscovered disease subtype. Two popular performance metrics are the Silhouette Coefficient and Dunn's Index [14, 42]. The Silhouette Coefficient measures how close each observation is to others in its cluster versus the next nearest cluster. It ranges from −1 (poor clustering) to +1 (dense, well-separated clusters) [42]. The Dunn's Index computes the ratio of the minimal intercluster distance to the maximal cluster diameter. Here, distance refers to the chosen similarity metric used to compare data points during clustering. For medical data, this could be the Euclidean distance between feature vectors. Cluster diameter measures dispersion within a cluster by the greatest distance between any two members [14].

Larger intercluster separation gaps and more compact cluster sizes (lower diameter) produce higher Dunn's Index values, indicating better delineation of distinct groups. However, an inherent assumption is that the natural clusters in the data are dense and well-separated [14].

In some medical contexts, underlying conditions may better manifest as overlapping, sparse or elongated clusters. For example, co-morbidities could link symptoms of two diseases, preventing clean separation. In such cases, poor Dunn Index scores do not necessarily indicate ineffective clustering, but a mismatch between analysis assumptions and real-world ambiguity. The clustering itself may still provide clinical utility. However, interpretation should account for complexity in the disease patterns defying assumptions. In these situations, different similarity metrics or clustering approaches optimised for interconnected data may be warranted [28].

Recommendation and ranking

Recommendation and ranking systems suggest items likely of interest to a certain user or patient, such as research papers relevant to a surgeon's specialty, or a

clinical study to a patient, based on their previous behaviour and known metrics.

Recommendation systems can be classified into two main categories: collaborative filtering and content-based filtering [41]. Collaborative filtering algorithms recommend items based on the ratings or preferences of other users. For example, if you have rated a publication highly, a collaborative filtering algorithm might recommend other publications that have been rated highly by people who have similar interests as you. Content-based filtering algorithms recommend items based on the features of the items themselves. For example, if you have read mainly arthroscopic literature, a content-based filtering algorithm might recommend other arthroscopy-related papers. Both approaches can be combined. In addition to collaborative filtering and content-based filtering, recommendation systems can also use classification and regression algorithms. Classification algorithms are used to predict a categorical value, such as whether a user will like or dislike an item. Regression algorithms are used to predict a continuous value, such as the number of citations of an item.

In healthcare, such algorithms can be utilised to recommend patients clinical trials that they are more likely to participate in, online health resources that they are more likely to adhere to, and personalised lifestyle advice [46].

The performance of recommendation systems is commonly evaluated using metrics, such as precision, recall, mean average precision [3], normalised discounted cumulative gain (NDCG) and AUC-ROC. The choice of performance metrics will depend on the specific application and the goals of the system. A key determinant is often how many recommended items are likely to be manually checked before a final selection is done.

Forecasting and time series

AI forecasting leverages historical time series data to predict future values through ML regression techniques tailored to capture temporal patterns and trends [18, 27]. Common evaluation metrics for forecasting systems include MAE, RMSE and MAPE to quantify deviation from ground truth over the prediction horizon. While similar metrics are utilised for general regression tasks, the distinction lies in the explicit modelling of time-based effects [45]. Judicious selection of appropriate skill metrics, testing on ample time series data encompassing variability, and reporting detailed performance across near-term and longer-range forecasts facilitates rigorous assessment of model accuracy and generalisability. Adoption of robust evaluation protocols tailored to the nature of forecasting problems enables standardised benchmarking and continued advancement of predictive technologies.

Anomaly detection

Anomaly detection is an analytical task applied across various domains, notably in healthcare, cybersecurity and finance, to identify data instances that deviate from established norms [13, 49]. While its evaluation metrics exhibit some resemblances with classification, subtle distinctions emerge due to the unique data characteristics and primary objectives of anomaly detection. In the realm of anomaly detection, the foremost evaluation metrics include sensitivity, precision, recall, F1-score and the AUC-ROC. These metrics offer essential insights into the performance of anomaly detection algorithms. Anomaly detection frequently contends with imbalanced data sets wherein anomalies constitute a minority. Consequently, precision and recall assume heightened significance as anomalies necessitate meticulous scrutiny to minimise false alarms. This requires careful threshold definitions as this choice exerts a profound influence on the intricate precision-recall trade-offs inherent to anomaly detection.

In summation, anomaly detection is a specialised form of classification, notably within imbalanced data contexts, which necessitates a deliberate contemplation of precision and recall.

Text generation

With the rise of ChatGPT (OpenAI Inc.), Claude 2 (Anthropic), Bard (Google) and other generative AI, text generation has already entered the healthcare system (e.g., chat bots) [10].

These large language models generate coherent text based on an initial text input that provides context and guides the direction of the output, known as a ‘prompt’. Outputs should be evaluated both automatically (grammar, coherence) and manually (factual correctness, creativity). An example of the former is the bilingual evaluation understudy (BLEU) which quantifies the degree to which a generated answer corresponds to a pre-existing, expected one [37]. Kaarre et al. used expert orthopaedic surgeons as expert evaluators to judge responses from the GPT-4 generative model [26]. Due to the nature of their output, these systems are notoriously hard to compare and contrast objectively. However, automatically calculated scores like BLEU have known limitations [5].

Image/video generation

Generative deep-learning models can create realistic images, audio and video (e.g., Dall-E, Midjourney) [2]. The quality and fidelity of generated media can be measured via human evaluation, and similarity metrics

such as Fréchet Inception Distance which uses deep neural network representations to quantify the statistical similarity of synthetically generated images compared to real images [22]. However, manipulation risks necessitate cautious deployment, and all output should undergo human review before being deployed.

In summary, AI outputs take a variety of numeric, textual and visual forms. Rigorously evaluating results for a given application requires selecting meaningful performance metrics, testing models on appropriate real-world data, and considering changes in data over time. Interdisciplinary collaboration between technical and subject matter experts can help determine if AI outputs are reasonable, useful and generalisable.

INTERPRETING AND EVALUATING OUTPUTS

Properly interpreting and evaluating AI outputs is crucial before applying models to real-world tasks. Here, we further discuss aforementioned metrics and detail their interpretation.

Evaluation metrics

Classification models output predicted labels and confidence scores. Metrics like accuracy, precision, recall, F1 and AUC-ROC provide quantification, but have limitations, as no single metric fully captures performance [23, 36]. Nevertheless, in order to compare model performance, quantifiable performance metrics serve as an important tool. All these values rely on the underlying confusion matrix, summarising prediction of the model versus real-world data. Four values are presented in a confusion matrix, similar in concept to methodology for assessing a new clinical diagnostic test:

	Known positive	Known negative
Predicted positive	TP	FN
Predicted negative	FP	TN

- True positive (TP): Positive outcome correctly classified as positive.
- True negative (TN): Negative outcome correctly classified as negative.
- False positive (FP): Negative outcome incorrectly classified as positive.
- False negative (FN): Positive outcome incorrectly classified as negative.

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{FP} + \text{FN} + \text{TN})},$$

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})},$$

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})},$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

$$\text{F1} - \text{score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}.$$

Regression models' performance metrics derive from the value between the forecasted and actual variable of the test data [34]. Let A_t and F_t denote the actual and forecasted values of the test data point t, respectively. Then the MAPE is given by:

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|,$$

while the RMSE corresponds to:

$$\text{RMSE} = \sqrt{\frac{\sum_t (A_t - F_t)^2}{2}}.$$

The F1 − score, also known as the harmonic mean of precision and recall, is most useful for problems with imbalanced data set classes, as accuracy alone can be misleading if there is a majority negative class that is simple to predict [36, 39]. F1 handles class imbalance better as it incorporates precision and recall, considering true positives, false positives and false negatives. It ranges from 0 to 1, with 1 being a perfect prediction [39]. A drawback of the F1-score is that it does not consider true negatives; however, it remains useful for healthcare, where minimising false positives and false negatives is critical [48].

The AUC-ROC metric refers to the area under the receiver operating characteristic curve, which plots the true positive rate (recall) against the false positive rate at different classification thresholds and measures the entire area under this curve, from (0, 0) to (1, 1) (Figure 1) and is sometimes referred to as a model's predictive capacity [17, 36, 39, 48]. A higher AUC indicates the model is better at distinguishing between positive and negative classes across thresholds and does not require setting a single classification threshold-like accuracy. Accuracy—amongst others—relies on setting a specific classification threshold, such

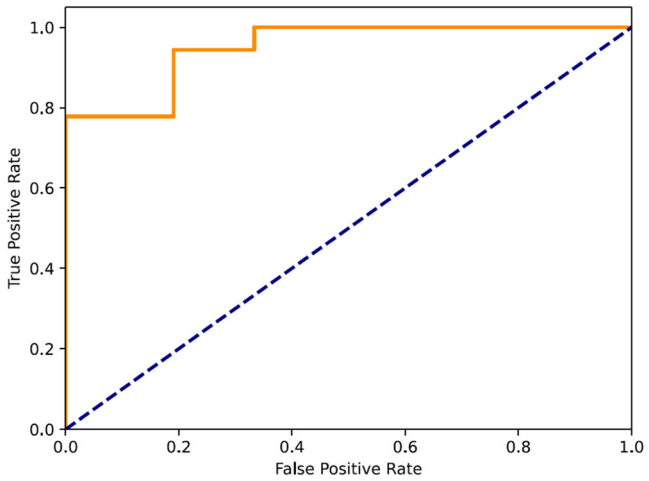

FIGURE 1 Area under the curve (AUC)–receiver operator curve (ROC) graph, the orange line displays the models true positive rate and false positive rate at various thresholds, the dashed blue line represents an AUC of 0.5, no better than chance.

as 0.5, to map prediction scores to discrete classes. Values above that set threshold are classified as positive, while values below are negative. The AUC value ranges from 0 to 1, with 1 being perfect classification. A drawback is that a high AUC can sometimes overstate model performance if there is a large false positive rate [17, 39]. Still, AUC-ROC is commonly used in medicine, biometrics and other applications where understanding the trade-off between true positives and false positives is important [48]. The MCC is another evaluation metric that summarises the confusion matrix into a value between −1 and +1, with higher scores indicating better classification performance. An MCC of +1 represents perfect prediction, 0 is equivalent to random guessing and −1 indicates total disagreement between predictions and true labels.

Compared to metrics like accuracy or AUC-ROC, MCC provides a more balanced assessment when evaluating imbalanced data sets. Given the relative benefits of the MCC metric, it is recommended to complement the AUC-ROC score with MCC calculation since it can give a more balanced point of comparison across different model types and data sets [6].

Human evaluation

Human evaluation is required for generative outputs like text and images. Automatic metrics have limitations, so expert judges should evaluate quality, coherence and correctness, considering the limitations of quantitative evaluation methods. However, human evaluation can be subjective and inconsistent between judges [30]. It is prudent to report interrater agreement measures when using multiple human evaluators of performance [19, 26].

Offline versus online performance

Offline ML, also called batch learning, describes engineering a model trained on a fixed training set, evaluated by a fixed test data set (internal validation), without changing them during the iteration process. This is the most commonly used type in the medical literature. Concept drift and covariate shifts are not considered after deployment as the model is based on an original data set. Online ML takes into account evolving learning environments and changes in real-world performance are immediate [36]. In the context of the healthcare sector, both approaches can be utilised with offline learning, for example, being suited for image recognition and classification, while online learning is more suitable for data sets with continuous data streams such as prediction of hospital capacity utilisation. However, even online ML has many risks and the model can drift due to invalid or erroneous data, faulty sensors and so on.

Uncertainty and interpretability

All models should provide uncertainty estimates like confidence intervals, as point estimates are insufficient in presenting the complete picture, especially in small data sets [1]. Scoring models can also help in this regard since they are both interpretable, for example, it is clear from their construction why they predict a certain outcome and directly predict patient risk via their score [50].

In summary, while metrics provide quantification, responsible evaluation goes deeper to test model limitations and ensure outputs are sound. Evaluation is an iterative process, not a one-time event. Interdisciplinary collaboration between technical and subject matter experts grounds evaluation in real-world needs. With rigorous and thoughtful evaluation, we can realise AI's benefits while building trust through transparency. Guidelines and reporting standards have already been proposed for the judicial use of AI and ML models in medical applications and should be consulted both in scientific reporting and in clinical evaluation of AI-based solutions such as transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement [9, 12].

CASE STUDIES

Case study 1: Image recognition

Cho et al. analysed 1394 arthroscopic rotator cuff repair images from 580 patients. Images were

categorised as 1138 nonretear and 256 retear based on magnetic resonance imaging follow-up within 2 years postoperatively [7]. The authors implemented standard ML practices for model development and validation. The image data set was split into training (80%) and held-out test (20%) sets. The training set was further divided into three folds for stratified *k*-fold cross-validation to fine-tune model hyperparameters and prevent overfitting. Performance metrics including accuracy, AUC-ROC, sensitivity, specificity and $F1-score$ were reported for both cross-validation and final model evaluation on the unseen test set. Three pretrained convolutional neural network architectures (VGG16, DenseNet121, Xception) were initialised with transferred weights and fine-tuned on the training folds. The models achieved cross-validation accuracy of 80%–99% across folds. On final model testing, DenseNet121 performed best with 91% accuracy, 0.92 AUC, 84% sensitivity and 93% specificity in predicting retearing from the arthroscopic images [7]. Limitations also noted by the authors were the imbalance of the retear to nonretear classes and lack of external validation.

By implementing robust ML methodology including cross-validation and reporting performance on an unseen hold-out test set, the authors have demonstrated that deep-learning analysis of arthroscopic rotator cuff repair images can accurately predict postoperative integrity.

Case study 2: Database analysis and prediction

Martin et al. analysed data from 62,955 patients in the Norwegian and Danish ACL reconstruction registries [33]. The aim was to develop an ML model to predict risk of revision surgery at 1, 2 and 5 years postoperatively. The data set was randomly split into 75% training and 25% test sets. Four algorithms were tested: Cox lasso regression, random survival forest, gradient boosting machines and super learner ensemble [40]. Hyperparameters were optimised via grid search with cross-validation on the training set. Performance was evaluated on the test set using Harrell's concordance index (C-index) for predictive discrimination and Hosmer–Lemeshow calibration plots. Multiple imputation was used to assess potential bias from missing data. The nonparametric ML models (random survival forest, gradient boosting, super learner) demonstrated moderate predictive performance, with indices around 0.67. Despite the large sample size, this was similar to prior models developed using the Norwegian ACL registry alone [32]. A key weakness noted by the authors was that, despite using multiple ML methods, the prediction accuracy for knee revision surgery outcomes showed limited improvement compared to previous simpler models, likely due to substantial missing preoperative data.

By splitting data into training and test sets, tuning hyperparameters via cross-validation and evaluating discrimination and calibration, the authors implemented rigorous ML methodology. However, model accuracy reached a performance ceiling, indicating that enhancing variable capture may be needed to improve predictions.

CONCLUSION

In conclusion, as AI continues advancing at a rapid pace, adopting rigorous evaluation, and reporting standards is imperative. The methodologies outlined here, including thoughtful selection of performance metrics, testing on real-world data distributions, assessing model uncertainties and transparent reporting of details based on existing guidelines, serve as a framework for critical model appraisal. There is still significant work needed to realise AI's potential benefits while mitigating risks. Model interpretability and explainability techniques must continue advancing to enable practitioners to understand how systems reach conclusions. Moving forward, a cross-disciplinary emphasis on rigorous analytical evaluation, clinical collaboration and ethical deployment will help foster continued AI progress. This will require commitment from researchers, clinicians, journal editors and regulatory agencies alike, to uphold AI evaluation and reporting standards that match these powerful technologies' capabilities and societal impacts.

AUTHOR CONTRIBUTIONS

All listed authors have contributed substantially to this work: Felix C. Oettl, Ayoosh Pareek, Bálint Zsidai and Eric Hamrin Senorski performed literature review and primary manuscript preparation. Editing and final manuscript preparation were performed by Philipp W. Winkler, Sebastian Kopf, Christophe Ley, Elmar Herbst, Jacob F. Oeding, Alberto Grassi, Michael T. Hirschmann, Volker Musahl, Kristian Samuelsson, Thomas Tischer and Robert Feldt. All authors read and approved the final manuscript.

ACKNOWLEDGEMENTS

The authors have no funding to report.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

ETHICS STATEMENT

The authors have nothing to report.

ORCID

Felix C. Oettl http://orcid.org/0000-0001-9721-685X
Ayoosh Pareek http://orcid.org/0000-0001-8683-1697
Bálint Zsidai http://orcid.org/0000-0002-5697-6577
James A. Pruneski http://orcid.org/0000-0002-8645-9386
Eric Hamrin Senorski http://orcid.org/0000-0002-9340-0147
Elmar Herbst http://orcid.org/0000-0002-5652-0692
Jacob F. Oeding http://orcid.org/0000-0002-4562-4373
Michael T. Hirschmann http://orcid.org/0000-0002-4014-424X
Volker Musahl http://orcid.org/0000-0001-8881-6212
Kristian Samuelsson http://orcid.org/0000-0001-5383-3370
Robert Feldt http://orcid.org/0000-0002-5179-4205

REFERENCES

1. Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M. et al. (2021) A review of uncertainty quantification in deep learning: techniques, applications and challenges. *Information Fusion*, 76, 243–297. Available from: https://doi.org/10.1016/j.inffus.2021.05.008
2. Adams, L.C., Busch, F., Truhn, D., Makowski, M.R., Aerts, H.J.W.L. & Bressem, K.K. (2023) What does DALL-E 2 know about radiology? *Journal of Medical Internet Research*, 25, e43110. Available from: https://doi.org/10.2196/43110
3. Ashraf, S., Wibberley, H., Mapp, P.I., Hill, R., Wilson, D. & Walsh, D.A. (2011) Increased vascular penetration and nerve growth in the meniscus: a potential source of pain in osteoarthritis. *Annals of the Rheumatic Diseases*, 70, 523–529. Available from: https://doi.org/10.1136/ard.2010.137844
4. Box, G.E.P. (1976) Science and statistics. *Journal of the American Statistical Association*, 71, 791–799. Available from: https://doi.org/10.1080/01621459.1976.10480949
5. Chen, A., Stanovsky, G., Singh, S. & Gardner, M. (2019) Evaluating question answering evaluation. *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, 1 January 2019. Hong Kong, China: Association for Computational Linguistics, pp. 119–124. Available from: https://doi.org/10.18653/v1/D19-5817
6. Chicco, D. & Jurman, G. (2023) The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining*, 16, 4. Available from: https://doi.org/10.1186/s13040-023-00322-4
7. Cho, S.H. & Kim, Y.S. (2023) Prediction of retear after arthroscopic rotator cuff repair based on intraoperative arthroscopic images using deep learning. *The American Journal of Sports Medicine*, 51, 2824–2830. Available from: https://doi.org/10.1177/03635465231189201
8. Clancey, W.J. & Hoffman, R.R. (2021) Methods and standards for research on explainable artificial intelligence: lessons from intelligent tutoring systems. *Applied AI Letters*, 2, e53. Available from: https://doi.org/10.1002/ail2.53
9. Collins, G.S., Reitsma, J.B., Altman, D.G. & Moons, K.G.M. (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *European Urology*, 67, 1142–1151. Available from: https://doi.org/10.1016/j.eururo.2014.11.025
10. Dave, T., Athaluri, S.A. & Singh, S. (2023) ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Frontiers in Artificial Intelligence*, 6, 1169595. Available from: https://doi.org/10.3389/frai.2023.1169595
11. Davenport, T. & Kalakota, R. (2019) The potential for artificial intelligence in healthcare. *Future Healthcare Journal*, 6, 94–98. Available from: https://doi.org/10.7861/futurehosp.6-2-94
12. De Hond, A.A.H., Leeuwenberg, A.M., Hooft, L., Kant, I.M.J., Nijman, S.W.J., Van Os, H.J.A. et al. (2022) Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *npj Digital Medicine*, 5, 2. Available from: https://doi.org/10.1038/s41746-021-00549-7
13. Deng, H. & Li, X. (2022) Self-supervised anomaly detection with random-shape pseudo-outliers. *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, *Glasgow, Scotland, United Kingdom*. pp. 4768–4772. Available from: https://doi.org/10.1109/EMBC48229.2022.9871621
14. Dunn, J.C. (2008) Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4, 95–104. Available from: https://doi.org/10.1080/01969727408546059
15. Eche, T., Schwartz, L.H., Mokrane, F.-Z. & Dercle, L. (2021) Toward generalizability in the deployment of artificial intelligence in radiology: role of computation stress testing to overcome underspecification. *Radiology. Artificial intelligence*, 3, 210097. Available from: https://doi.org/10.1148/ryai.2021210097
16. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M. et al. (2017) Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542, 115–118. Available from: https://doi.org/10.1038/nature21056
17. Fawcett, T. (2006) An introduction to ROC analysis. *Pattern Recognition Letters*, 27, 861–874. Available from: https://doi.org/10.1016/j.patrec.2005.10.010
18. Festag, S., Denzler, J. & Spreckelsen, C. (2022) Generative adversarial networks for biomedical time series forecasting and imputation. *Journal of Biomedical Informatics*, 129, 104058. Available from: https://doi.org/10.1016/j.jbi.2022.104058
19. Gisev, N., Bell, J.S. & Chen, T.F. (2013) Interrater agreement and interrater reliability: key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, 9, 330–338. Available from: https://doi.org/10.1016/j.sapharm.2012.04.004
20. Gunning, D., Vorm, E., Wang, J.Y. & Turek, M. (2021) DARPA's explainable AI (XAI) program: a retrospective. *Applied AI Letters*, 2, e61. Available from: https://doi.org/10.1002/ail2.61
21. Hendricks, L.A., Rohrbach, A., Schiele, B., Darrell, T. & Akata, Z. (2021) Generating visual explanations with natural language. *Applied AI Letters*, 2, e55. Available from: https://doi.org/10.1002/ail2.55
22. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. & Hochreiter, S. (2017) GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *arXiv e-prints*. Available from: https://doi.org/10.48550/arXiv.1706.08500
23. Hicks, S.A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M.A., Halvorsen, P. et al. (2022) On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports*, 12, 5979. Available from: https://doi.org/10.1038/s41598-022-09954-8
24. Hoenders, C.S.M., Harmsen, M.C. & van Luyn, M.J.A. (2008) The local inflammatory environment and microorganisms in "aseptic" loosening of hip prostheses. *Journal of Biomedical Materials Research, Part B: Applied Biomaterials*, 86, 291–301. Available from: https://doi.org/10.1002/jbm.b.30992

25. Hu, R., Andreas, J., Darrell, T. & Saenko, K. (2021) Explainable neural computation via stack neural module networks. *Applied AI Letters*, 2, e39. Available from: https://doi.org/10.1002/ail2.39
26. Kaarre, J., Feldt, R., Keeling, L.E., Dadoo, S., Zsidai, B., Hughes, J.D. et al. (2023) Exploring the potential of ChatGPT as a supplementary tool for providing orthopaedic information. *Knee Surgery, Sports Traumatology, Arthroscopy*, 31, 5190–5198. Available from: https://doi.org/10.1007/s00167-023-07529-2
27. Kaur, J., Parmar, K.S. & Singh, S. (2023) Autoregressive models in environmental forecasting time series: a theoretical and application review. *Environmental Science and Pollution Research*, 30, 19617–19641. Available from: https://doi.org/10.1007/s11356-023-25148-9
28. Khanmohammadi, S., Adibeig, N. & Shanehbandy, S. (2017) An improved overlapping k-means clustering method for medical applications. *Expert Systems With Applications*, 67, 12–18. Available from: https://doi.org/10.1016/j.eswa.2016.09.025
29. Khashei, M., Bakhtiarvand, N. & Etemadi, S. (2021) A novel reliability-based regression model for medical modeling and forecasting. *Diabetes & Metabolic Syndrome*, 15, 102331. Available from: https://doi.org/10.1016/j.dsx.2021.102331
30. Laskar, M.T.R., Bari, M.S., Rahman, M., Bhuiyan, M.A.H., Joty, S. & Huang, J.X. (2023) A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. *arXiv*, 2305, 18486. Available from: https://doi.org/10.48550/arXiv.2305.18486
31. Makarov, V.A., Stouch, T., Allgood, B., Willis, C.D. & Lynch, N. (2021) Best practices for artificial intelligence in life sciences research. *Drug Discovery Today*, 26, 1107–1110. Available from: https://doi.org/10.1016/j.drudis.2021.01.017
32. Martin, R.K., Wastvedt, S., Pareek, A., Persson, A., Visnes, H., Fenstad, A.M. et al. (2023) Ceiling effect of the combined norwegian and danish knee ligament registers limits anterior cruciate ligament reconstruction outcome prediction. *The American Journal of Sports Medicine*, 51, 2324–2332. Available from: https://doi.org/10.1177/03635465231177905
33. Martin, R.K., Wastvedt, S., Pareek, A., Persson, A., Visnes, H., Fenstad, A.M. et al. (2022) Machine learning algorithm to predict anterior cruciate ligament revision demonstrates external validity. *Knee Surgery, Sports Traumatology, Arthroscopy*, 30, 368–375. Available from: https://doi.org/10.1007/s00167-021-06828-w
34. Meena, T. & Roy, S. (2022) Bone fracture detection using deep supervised learning from radiological images: a paradigm shift. *Diagnostics*, 12, 2420. Available from: https://doi.org/10.3390/diagnostics12102420
35. Morris, M.R. (2023) Scientists' perspectives on the potential for generative ai in their fields. *arXiv e-prints*. Available from: https://doi.org/10.48550/arXiv.2304.01420
36. Panesar, A. (2021) *Machine Learning and AI for Healthcare: Big Data for Improved Health Outcomes*, 2nd edition. New York: Apress. Available from: https://doi.org/10.1007/978-1-4842-6537-6
37. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J.B.L.E.U. (2001) *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics—ACL '02*, 1 January 2001. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 311–318. Available from: https://doi.org/10.3115/1073083.1073135
38. Plevris, V., Solorzano, G., Bakas, N. & Ben Seghier, M. (2022) Investigation of performance metrics in regression analysis and machine learning-based prediction models. In: *8th European Congress on Computational Methods in Applied Sciences and Engineering* (*eccomas*). Available from: https://www.scipedia.com/pulic/Plevris_et_al_2022a
39. Powers, D.M.W. (2020) Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv e-prints*. Available from: https://doi.org/10.48550/arXiv.2010.16061
40. Pruneski, J.A., Pareek, A., Kunze, K.N., Martin, R.K., Karlsson, J., Oeding, J.F. et al. (2023) Supervised machine learning and associated algorithms: applications in orthopedic surgery. *Knee Surgery, Sports Traumatology, Arthroscopy*, 31, 1196–1202. Available from: https://doi.org/10.1007/s00167-022-07181-2
41. Ricci, F., Rokach, L. & Shapira, B. Recommender systems: techniques, applications, and challenges. *Recommender systems handbook*. Springer New York, NY. pp. 1–35. Available from: https://doi.org/10.1007/978-1-0716-2197-4
42. Rousseeuw, P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65. Available from: https://doi.org/10.1016/0377-0427(87)90125-7
43. Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T. et al. (2020) Improved protein structure prediction using potentials from deep learning. *Nature*, 577, 706–710. Available from: https://doi.org/10.1038/s41586-019-1923-7
44. Shao, Y., Cheng, Y., Shah, R.U., Weir, C.R., Bray, B.E. & Zeng-Treitler, Q. (2021) Shedding light on the black box: explaining deep neural network prediction of clinical outcomes. *Journal of Medical Systems*, 45, 5. Available from: https://doi.org/10.1007/s10916-020-01701-8
45. Singh, S., Parmar, K.S., Makkhan, S.J.S., Kaur, J., Peshoria, S. & Kumar, J. (2020) Study of ARIMA and least square support vector machine (LS-SVM) models for the prediction of SARS-CoV-2 confirmed cases in the most affected countries. *Chaos, Solitons, and Fractals*, 139, 110086. Available from: https://doi.org/10.1016/j.chaos.2020.110086
46. Stefanidis, K., Tsatsou, D., Konstantinidis, D., Gymnopoulos, L., Daras, P., Wilson-Barnes, S. et al. (2022) PROTEIN AI advisor: a knowledge-based recommendation framework using expert-validated meals for healthy Diets. *Nutrients*, 14, 4435. Available from: https://doi.org/10.3390/nu14204435
47. Stokel-Walker, C. & Van Noorden, R. (2023) What ChatGPT and generative AI mean for science. *Nature*, 614, 214–216. Available from: https://doi.org/10.1038/d41586-023-00340-6
48. Taha, A.A. & Hanbury, A. (2015) Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging*, 15, 29. Available from: https://doi.org/10.1186/s12880-015-0068-x
49. Teng, H.S., Chen, K. & Lu, S.C. (1990) Adaptive real-time anomaly detection using inductively generated sequential patterns. *Proceedings of the IEEE Computer Society Symposium on Research in Security and Privacy*, *Oakland, CA, USA*. pp. 278–284. Available from: https://doi.org/10.1109/RISP.1990.63857
50. Ustun, B. & Rudin, C. (2016) Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, 102, 349–391. Available from: https://doi.org/10.1007/s10994-015-5528-6
51. Vasu, B., Hu, B., Dong, B., Collins, R. & Hoogs, A. (2021) Explainable, interactive content-based image retrieval. *Applied AI Letters*, 2, e41. Available from: https://doi.org/10.1002/ail2.41
52. Watson, D.S., Krutzinna, J., Bruce, I.N., Griffiths, C.E., McInnes, I.B., Barnes, M.R. et al. (2019) Clinical applications of machine learning algorithms: beyond the black box. *BMJ*, 364, l886. Available from: https://doi.org/10.1136/bmj.l886
53. Witty, S., Lee, J.K., Tosch, E., Atrey, A., Clary, K., Littman, M.L. et al. (2021) Measuring and characterizing generalization in deep reinforcement learning. *Applied AI Letters*, 2, e45. Available from: https://doi.org/10.1002/ail2.45

54. Yang, J., Soltan, A.A.S. & Clifton, D.A. (2022) Machine learning generalizability across healthcare settings: insights from multi-site COVID-19 screening. *npj Digital Medicine*, 5, 69. Available from: https://doi.org/10.1038/s41746-022-00614-9
55. Yang, S.C.-H., Folke, T. & Shafto, P. (2021) Abstraction, validation, and generalization for explainable artificial intelligence. *Applied AI Letters*, 2, e37. Available from: https://doi.org/10.1002/ail2.37
56. Yeh C.-K., Ravikumar P. (2021) Objective criteria for explanations of machine learning models. *Applied AI Letters* 2, e57. https://doi.org/10.1002/ail2.57

How to cite this article: Oettl, F.C., Pareek, A., Winkler, P.W., Zsidai, B., Pruneski, J.A., Senorski, E.H. et al. (2024) A practical guide to the implementation of AI in orthopaedic research, Part 6: how to evaluate the performance of AI research? *Journal of Experimental Orthopaedics*, 11, e12039. https://doi.org/10.1002/jeo2.12039