JMIR Publications
The Leading eHealth Publisher

Consolidated reporting guideline recommendations for prognostic and diagnostic machine learning models: CREMLS

Khaled El Emam, Tiffany I Leung, Bradley Malin, William Klement, Gunther Eysenbach

Submitted to: Journal of Medical Internet Research
on: April 04, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Consolidated reporting guideline recommendations for prognostic and diagnostic machine learning models: CREMLS

Khaled El Emam[1, 2] BEng, PhD; Tiffany I Leung[3, 4] MD, MPH; Bradley Malin[5] BA, MSc, PhD; William Klement[2]; Gunther Eysenbach[3, 6] MD, MPH

[1]School of Epidemiology and Public Health, University of Ottawa Ottawa CA
[2]Children’s Hospital of Eastern Ontario Research Institute Ottawa CA
[3]JMIR Publications, Inc Toronto CA
[4]Department of Internal Medicine (adjunct), Southern Illinois University School of Medicine Springfield US
[5]Department of Biomedical Informatics, Vanderbilt University Nashville US
[6]University of Victoria Victoria CA

Corresponding Author:
Khaled El Emam BEng, PhD
School of Epidemiology and Public Health, University of Ottawa
401 Smyth Road
Ottawa
CA

Abstract

The number of papers presenting machine learning (ML) models that are being submitted, and published, in the Journal of Medical Internet Research and other JMIR Publications journals has steadily increased. Editors and peer reviewers involved in the review process for such manuscripts often go through multiple review cycles to enhance the quality and completeness of reporting. The use of a reporting guideline, or checklist, can help ensure consistency in the quality of submitted (and published) scientific manuscripts and, for instance, avoid instances of missing information. In this Editorial, JMIR Publications journal editors discuss the general JMIR Publications policy with regards to authors’ application of reporting guidelines, then focus specifically on the reporting of machine learning studies in JMIR Publications journals.

(JMIR Preprints 04/04/2024:52508)
DOI: https://doi.org/10.2196/preprints.52508

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

Please make my preprint PDF available to anyone at any time (recommended).
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
✓ **No, I do not wish to publish my submitted manuscript as a preprint.**

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Consolidated reporting guideline recommendations for prognostic and diagnostic machine learning models: CREMLS

Introduction

The number of papers presenting machine learning (ML) models that are being submitted, and published, in the *Journal of Medical Internet Research* and other JMIR Publications journals has steadily increased over time. The cross-journal JMIR Publications e-collection “Machine Learning” includes nearly 1300 articles, as of April 1, 2024 [1] and there are additional sections in other journals that collate articles related to the field (e.g. “Machine Learning from Dermatological Images” [2] in *JMIR Dermatology*). From 2015 to 2022, the number of published articles with "artificial intelligence" or "machine learning" in the title and abstract in JMIR Publications journals increased from 22 to 298 (13.5x growth), and already there are 312 in 2023 (14x growth). For *JMIR Medical Informatics*, the number of articles increased from 10 to 160 (16x growth) until 2022. This is consistent with the growth in the research and application of medical artificial intelligence (AI) in general where a similar Pubmed search (+ “medicine”) showed 22x growth (from 640 to 14147) between 2015 and 2022, and already there are 11272 matching articles in 2023.

Many ML papers in medicine utilize a large clinical dataset to make diagnostic or prognostic predictions [3–6]. However, the use of data from electronic health records (EHRs) and other resources is often not without pitfalls, as these data are typically collected for other purposes (e.g. medical billing) [7].

Editors and peer reviewers involved in the review process for such manuscripts often go through multiple review cycles to enhance the quality and completeness of reporting [8]. The use of a reporting guideline, or checklist, can help ensure consistency in the quality of submitted (and published) scientific manuscripts and, for instance, avoid instances of missing information. For example, in the experiences of the *JMIR AI* Editors-in-Chief, missing information is especially notable because for manuscripts reporting on ML models that are submitted to *JMIR AI* it can delay the overall review interval by adding more revision cycles.

According to the EQUATOR Network (Enhancing the QUAlity and Transparency Of health Research), a reporting guideline is “a simple, structured tool for health researchers to use while writing manuscripts. A guideline provides a minimum list of information needed to ensure a manuscript can be, for example: understood by a reader, replicated by a researcher, used by a doctor to make a clinical decision, and included in a systematic review” [9]. These can be presented in the form of a checklist, flow diagram, or structured text.

In this Editorial, we discuss the general JMIR Publications policy with regards to authors’ application of reporting guidelines. We then focus specifically on the reporting of ML studies in JMIR Publications journals.

JMIR Publications Policy on Use of Reporting Guidelines

Accumulating evidence suggests that when authors apply reporting guidelines and reporting checklists in health research, they can be beneficial for authors, readers, and the discipline overall by enabling the replication or reproduction of studies. Recent evidence suggests that asking reviewers to use reporting checklists, instead of authors, offer no added benefits on reporting quality [10]. However, one other study reported a positive association between reviewer ratings of adherence to reporting guidelines and favorable editorial decisions [11], while another reported a significant positive correlation between adherence to reporting guidelines and citations, and between adherence to reporting guidelines and publication in higher impact factor journals [12].

At JMIR Publications, editorial policy recommends authors to adhere to applicable study design and reporting guidelines for the study when preparing manuscripts for submission [13]. Authors should note that most reporting guidelines are strongly recommended, particularly because they can improve the quality, completeness, and organization of the presented work. At this time, JMIR Publications *requires* reporting checklists to be completed and supplied as supplementary appendices for randomized trials without [14–16] or with eHealth or mHealth components [17] and systematic and scoping literature reviews across the portfolio; and implementation reports for this article type in *JMIR Medical Informatics* [18]. Although some medical journals have made the use of certain reporting guidelines and checklists mandatory, JMIR Publications recognizes that authors may have concerns about the additional burden that the formalized use of checklists may bring to the submission process. As such, JMIR Publications has chosen to begin with recommending the use of ML reporting guidelines and will evaluate their benefits and gather feedback on implementation costs before considering stronger requirements.

Reporting on Machine Learning Models

With respect to the reporting of prognostic and diagnostic machine learning studies, multiple checklists have been developed that are directly relevant. The manuscript by Klement and El Emam [19] consolidates these guidelines and checklists into a single set that we refer to as the Consolidated REporting of Machine Learning Studies (CREMLS) checklist. CREMLS serves as a reporting checklist for journals publishing this type of research, including all journals from JMIR Publications, which has officially adopted these guidelines. CREMLS was developed by identifying existing relevant reporting guidelines and checklists. The initial item list were identified through a structured literature review and expert curation, then the quality of the methods used for their development was assessed to narrow them down to a high quality subset. This high quality item subset was further filtered to those that meet specific inclusion and exclusion criteria. The resultant items were converted to guidelines and a checklist which went through a review with members of the *JMIR AI* editorial board, and was followed by a preliminary application to assess articles published in *JMIR AI*. The final checklist offers a present-day best practice for high-quality reporting of machine learning studies.

Examples of the application of the CREMLS item checklist are presented in Table 1. For that we identified 7 articles published in JMIR Publications journals that exemplify each checklist. Note that not all of the items are relevant to each article, and some articles are particularly good examples of how to operationalize a checklist item. These are illustrated in that table.

Table 1: Illustration of how various articles published in JMIR journals implement each of the REML guideline checklist items.

#	Item	Example Illustrating the Item
Study Details		
1.1	*The medical/clinical task of interest*	examines chronic disease management — a clinical problem with 4 example solutions using ML models [20]
1.2	*The research question*	proposes a framework to transfers old knowledge to a new environment to mange drifts [21]
1.3	*Current medical/clinical practice*	provides a review of current practice and issues associated with chronic disease management [20]
1.4	*The known predictors and confounders to what is being predicted / diagnosed*	described variables defined as part of a well-established health test available to the public [20]
1.5	*The overall study design*	presents experimental design with data flow and data partitions used at various steps of the experiment (Figure 1) [22]
1.6	*The medical institutional setting(s)*	describes the institution as an academic (teaching) community

		hospital where the data was collected [23]
1.7	*The target patient population*	clear partitioning of target patient populations and comparator group [20]
1.8	*The intended use of the ML model*	describes how the prediction model fits in the clinical practice of scheduling operating room procedures [5]
1.9	*Existing model performance benchmarks for this task*	reviews exiting research and presents achieved performance (AUC) [20]
1.10	*Ethical and other regulatory approvals obtained*	ethics approvals [5]
The Data		
2.1	*Inclusion / exclusion criteria for the patient cohort*	defined in Figure 1 [5]
2.2	*Methods of data collection*	describes source and methods of data collection, what type of data was used, and potential implied bias in interpretation [23]
2.3	*Bias introduced due to the method of data collection used*	discusses potential bias in data collection and in outcome definition [23]
2.4	*Data characteristics*	uses descriptive statistics to show data characteristics for different types of data (demographics and clinical measurements) [23]
2.5	*Methods of data transformations and preprocessing applied*	imputation is discussed [5]
2.6	*Known quality issues with the data*	missingness and outlier detection were discussed [5]
2.7	*Sample size calculation*	brief section dedicated to power analysis [5]
2.8	*Data Availability*	explains how to obtain a copy of the data [24]
Methodology		
3.1	*Strategies for handling missing data*	describes how missing values were replaced [20]
3.2	*Strategies for addressing class imbalance*	describes the approach of using SMOTE to adjust class ratios to address imbalance [23]
3.3	*Strategies for reducing dimensionality of data*	describes the vectorization of a dimension of 100 into a 2-dimensional space using an

		established algorithm [22]
3.4	*Strategies for handling outliers*	the authors stated the threshold values used to detect outliers [5]
3.5	*Strategies for data augmentation*	showed how variable similarity is achieved between synthetic and real data in the context of augmentation [24]
3.6	*Strategies for model pre-training*	describes and illustrates (Figure 1) how models from other data sets were trained and used in the new model [23]
3.7	*The rationale for selecting the machine learning algorithm*	discusses properties of the selected algorithm relevant to the problem at hand as motivation [20]
3.8	*The method of evaluating model performance during training*	presents separate discussion of evolution in cross-validations settings and external evaluation while also describing hyperparameter tuning [23]
3.9	*The method used for hyperparameter tuning*	comprehensive description of tuning within nested cross-validation (this is a tutorial but illustrates how to describe the process) [25]
3.10	*Model's output adjustments*	describes the final model, how it was calibrated and discusses the impact of embedding on patient data for interpretation [22]
Evaluation		
4.1	*Performance metrics used to evaluate the model*	comprehensive and detailed discussion of evaluation and quality metrics [24]
4.2	*The cost or consequence of errors*	comprehensive error analysis [25]
4.3	*The results of internal validation*	detailed validation discussion (internally and externally) [25]
4.4	*The final model hyperparameters*	presents details of the final model and the winning parameters [5]
4.5	*Model evaluation on an external dataset*	detailed and comprehensive external validation that is separate from model testing [5]
4.6	*Characteristics relevant for detecting data shift and drift*	implements performance monitoring, addresses data shifts over time and illustrates them in detail [21]
Explainability and Transparency		
5.1	*The most important features and how they relate to*	presents variable importance (SHAP

	the outcome(s)	values) in the context of interpretation and compares it to existing literature [5]
5.2	*Plausibility of model outputs*	shows sample output (Figure 4) [5]
5.3	*Interpretation of model's results by an end-user*	good discussion about interpretability and use of final model [5]

We strongly advise authors who seek to submit their manuscripts on prognostic and diagnostic ML studies to the *Journal of Medical Internet Research*, *JMIR AI*, *JMIR Medical Informatics*, or other JMIR Publications journals to utilize the CREMLS guidelines and checklist to ensure that they have considered and addressed all relevant details for their work prior to initiating the submission and review process. More complete and high-quality reporting benefits the authors by accelerating the review cycle and also reducing the burden on reviewers. Hence the need for reporting guidelines and checklists for papers describing prognostic and diagnostic ML studies. This is expected to assist, for example, in reducing missing documentation on hyperparameters for an ML model and to clarify how data leakage was avoided. We have observed that peer reviewers have been in practice asking authors to improve reporting on the same topics covered in the CREMS checklist. This is not a surprise given that peer reviewers are experts in the field and would note important information that is missing. Nevertheless, we would encourage reviewers to use the checklist on a regular basis to ensure completeness and consistency.

The CREMLS checklist is limited in scope to ML models using structured data that are trained and evaluated in-silico and in shadow mode. This leaves significant opportunity to expand on CREMLS to different data modalities and additional phases of model deployment. Should such extended reporting guidelines and checklists be developed, they may be considered for recommendation in JMIR Publications journals, incorporating lessons learned from the initial checklist for machine learning studies.

Conclusion

There is evidence that the completeness of reporting of research studies is beneficial to the authors, as well as the broader scientific community. For prognostic and diagnostic machine learning studies, many reporting guidelines have been developed, and these have been consolidated into CREMLS, capturing the combined value of the source guidelines and checklists in one place. In this editorial we extend journal policy and make the recommendation to authors to follow these guidelines when submitting articles to journals in the JMIR portfolio. This will improve the reproducibility of research studies using machine learning methods, accelerate review cycles, and improve the quality of published papers overall. Given the rapid growth in this literature, it is important to establish reporting standards early.

Abbreviations

AI: artificial intelligence

CREMLS = Consolidated REporting of Machine Learning Studies

ML: machine learning

Conflicts of Interest

KEE and BM are Co-Editors-in-Chief of *JMIR AI.* KEE is co-founder of Replica Analytics, an Aetion company, and has financial interests in the company. TIL is scientific editorial director at JMIR Publications. GE is the founder, chief executive officer, and executive editor of JMIR Publications, receives a salary and owns equity.

Author Contributions

KEE: Conceptualization; Writing-original draft; Writing-review & editing. BM: Conceptualization; Writing-original draft; Writing-review & editing. TIL: Writing-review & editing. GE: Writing-review & editing.

References

1 Machine Learning. JMIR Publications. 2024. https://medinform.jmir.org/themes/500-machine-learning (accessed 1 April 2024)

2 Machine Learning from Digital Images in Dermatology. JMIR Publications. 2023. https://derma.jmir.org/themes/922-machine-learning-from-digital-images-in-dermatology (accessed 22 September 2023)

3 Lee S, Kang WS, Kim DW, *et al.* An Artificial Intelligence Model for Predicting Trauma Mortality Among Emergency Department Patients in South Korea: Retrospective Cohort Study. *J Med Internet Res*. 2023;25:e49283.

4 Deng Y, Ma Y, Fu J, *et al.* Combinatorial Use of Machine Learning and Logistic Regression for Predicting Carotid Plaque Risk Among 5.4 Million Adults With Fatty Liver Disease Receiving Health Check-Ups: Population-Based Cross-Sectional Study. *JMIR Public Health Surveill*. 2023;9:e47095.

5 Kendale S, Bishara A, Burns M, *et al.* Machine Learning for the Prediction of Procedural Case Durations Developed Using a Large Multicenter Database: Algorithm Development and Validation Study. *JMIR AI*. 2023;2:e44909.

6 Williams DD, Ferro D, Mullaney C, *et al.* An “All-Data-on-Hand” Deep Learning Model to Predict Hospitalization for Diabetic Ketoacidosis in Youth With Type 1 Diabetes: Development and Validation Study. *JMIR Diabetes*. 2023;8:e47592.

7 Maletzky A, Böck C, Tschoellitsch T, *et al.* Lifting Hospital Electronic Health Record Data Treasures: Challenges and Opportunities. *JMIR Med Inform*. 2022;10:e38557.

8 El Emam K, Klement W, Malin B. Reporting and Methodological Observations on Prognostic and Diagnostic Machine Learning Studies. *JMIR AI*. 2023.

9 What is a reporting guideline? EQUATOR Network. https://www.equator-network.org/about-us/what-is-a-reporting-guideline/ (accessed 22 September 2023)

10 Speich B, Mann E, Schönenberger CM, *et al.* Reminding Peer Reviewers of Reporting Guideline Items to Improve Completeness in Published Articles: Primary Results of 2 Randomized Trials. *JAMA Network Open*. 2023;6:e2317651.

11 Botos J. Reported use of reporting guidelines among JNCI: Journal of the National Cancer Institute authors, editorial outcomes, and reviewer ratings related to adherence to guidelines and clarity of presentation. *Res Integr Peer Rev*. 2018;3:7.

12 Stevanovic A, Schmitz S, Rossaint R, *et al.* CONSORT Item Reporting Quality in the Top Ten Ranked Journals of Critical Care Medicine in 2011: A Retrospective Analysis. *PLOS ONE*.

2015;10:e0128061.

13 What reporting guidelines should I follow for my article? JMIR Publications. 2024. https://support.jmir.org/hc/en-us/articles/115001575267-What-reporting-guidelines-should-I-follow-for-my-article (accessed 30 January 2024)

14 Schulz KF, Altman DG, Moher D, *et al.* CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *Int J Surg*. 2011;9:672–7.

15 Schulz KF, Altman DG, Moher D, *et al.* CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. *Ann Intern Med*. 2010;152:726–32.

16 Moher D, Hopewell S, Schulz KF, *et al.* CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ*. 2010;340:c869.

17 Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of Web-based and mobile health interventions. *J Med Internet Res*. 2011;13:e126.

18 Perrin Franck C, Babington-Ashaye A, Dietrich D, *et al.* iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations. *J Med Internet Res*. 2023;25:e46694.

19 Klement W, El Emam K. Consolidated Reporting Guidelines for Prognostic and Diagnostic Machine Learning Modeling Studies: Development and Validation. *J Med Internet Res*. 2023;25:e48763.

20 Lee C, Jo B, Woo H, *et al.* Chronic Disease Prediction Using the Common Data Model: Development Study. *JMIR AI*. 2022;1:e41030.

21 Zhang X, Xue Y, Su X, *et al.* A Transfer Learning Approach to Correct the Temporal Performance Drift of Clinical Prediction Models: Retrospective Cohort Study. *JMIR Medical Informatics*. 2022;10:e38053.

22 Steiger E, Kroll LE. Patient Embeddings From Diagnosis Codes for Health Care Prediction Tasks: Pat2Vec Machine Learning Framework. *JMIR AI*. 2023;2:e40755.

23 Sang S, Sun R, Coquet J, *et al.* Learning From Past Respiratory Infections to Predict COVID-19 Outcomes: Retrospective Study. *Journal of Medical Internet Research*. 2021;23:e23026.

24 Kang HYJ, Batbaatar E, Choi D-W, *et al.* Synthetic Tabular Data Based on Generative Adversarial Networks in Health Care: Generation and Validation Using the Divide-and-Conquer Strategy. *JMIR Medical Informatics*. 2023;11:e47859.

25 Wilimitis D, Walsh CG. Practical Considerations and Applied Examples of Cross-Validation for Model Development and Evaluation in Health Care: Tutorial. *JMIR AI*. 2023;2:e49023.