Noname manuscript No.
(will be inserted by the editor)

Machine Learning for Health: Algorithm Auditing & Quality Control

Luis Oala · Andrew G. Murchison · Pradeep Balachandran · Shruti Choudhary · Jana Fehr · Alixandro Werneck Leite · Peter G. Goldschmidt · Christian Johner · Elora D. M. Schörverth · Rose Nakasi · Martin Meyer · Federico Cabitza · Pat Baird · Carolin Prabhu · Eva Weicken · Xiaoxuan Liu · Markus Wenzel · Steffen Vogler · Darlington Akogo · Shada Alsalamah · Emre Kazim · Adriano Koshiyama · Sven Piechottka · Sheena Macpherson · Ian Shadforth · Regina Geierhofer · Christian Matek · Joachim Krois · Bruno Sanguinetti · Matthew Arentz · Pavol Bielik · Saul Calderon-Ramirez · Auss Abbood · Nicolas Langer · Stefan Haufe · Ferath Kherif · Sameer Pujari · Wojciech Samek · Thomas Wiegand

Received: date / Accepted: date

L. Oala, E. D. M. Schörverth, E. Weicken, M. Wenzel, W. Samek and T. Wiegand
Fraunhofer HHI, Germany, {luis.oala, elora-dana.schoerverth, eva.weicken, markus.wenzel, wojciech.samek, thomas.wiegand}@hhi.fraunhofer.de

A.G. Murchison
Oxford University Hospitals NHS Foundation Trust, United Kingdom, agmurchison@gmail.com

P. Balachandran
Technical Consultant (Digital Health), India, pbn.tvm@gmail.com

S. Choudhary
University of Oxford, United Kingdom, shruti.choudhary@kellogg.ox.ac.uk

J. Fehr
Hasso-Plattner-Institute of Digital Engineering, Germany, jana.fehr@hpi.de

Alixandro Werneck Leite
Machine Learning Laboratory in Finance and Organizations, Universidade de Brasília, Brazil
alixandrowerneck@outlook.com

P. G. Goldschmidt
World Development Group Inc, United States, pgg@worlddg.com

C. Johner
Johner Institute, Germany, christian.johner@johner-institut.de

R. Nakasi
Makerere University, Uganda, g.nakasirose@gmail.com

M. Mayer
Siemens Healthineers, Germany, martin.mm.meyer@siemens-healthineers.com

F. Cabitza
University of Milano-Bicocca, Italy, federico.cabitza@unimib.it

P. Baird
Philips, United States, pat.baird@philips.com

C. Prabhu
Office of the Auditor General of Norway, Norway, cap@riksrevisjonen.no

X. Liu
University Hospitals Birmingham NHS Foundation Trust & Academic Unit of Ophthalmology, Institute of Inflammation and Ageing, College of Medical and Dental Sciences, University of Birmingham, United Kingdom, x.liu.8@bham.ac.uk

S. Vogler
Bayer AG, Germany, steffen.vogler@bayer.com

D. Akogo
minoHealth AI Labs, Ghana, darlington@gudra-studio.com

S. Alsalamah
Information Systems Department, College of Computer and Information Sciences, King Saud University (Saudi Arabia) & Digital Health and Innovation Department, Science Division, World Health Organization (Switzerland), alsalamahs@who.int

E. Kazim and A. Koshiyama
University College London, United Kingdom
{e.kazim,adriano.koshiyama.15}@ucl.ac.uk

S. Piechottka
Open Regulatory, Germany, sven@openregulatory.com

S. Macpherson and Ian Shadforth
MIOTIFY LTD, United Kingdom, {sheena.macpherson,ian.shadforth}@miotify.co.uk

R. Geierhofer
IEC TC62 and Siemens Healthineers, Germany, geierhofer@cocir.org

C. Matek
Helmholtz Zentrum München, Germany, christian.matek@helmholtz-muenchen.de

J. Krois
Oral Diagnostics Digital Health Health Services Research, Charité - Universitätsmedizin, Germany, joachim.krois@charite.de

B. Sanguinetti
Dotphoton AG, Switzerland, bruno.sanguinetti@dotphoton.com

M. Arentz
Department of Global Health, University of Washington, United States, marentz@uw.edu

P. Bielik
LatticeFlow & ETH Zurich, Switzerland, pavol.bielik@inf.ethz.ch

S. Calderon-Ramirez
De Montfort University & Instituto Tecnologico de Costa Rica, Costa Rica, sacalderon@itcr.ac.cr

A. Abbood
Robert Koch Institut, Germany, abbooda@rki.de

N. Langer
Department of Psychology, University of Zurich, Switzerland, n.langer@psychologie.uzh

S. Haufe
Technische Universität Berlin, Germany, haufe@tu-berlin.de

F. Kherif
Laboratory for Research in Neuroimaging, Department of Clinical Neuroscience, Lausanne University Hospital and University of Lausanne, Switzerland, ferath.kherif@chuv.ch

S. Pujari
Digital Health and Innovation Department, Science Division, World Health Organization, Switzerland, pujaris@who.int

Abstract Developers proposing new machine learning for health (ML4H) tools often pledge to match or even surpass the performance of existing tools, yet the reality is usually more complicated. Reliable deployment of ML4H to the real world is challenging as examples from diabetic retinopathy or Covid-19 screening show. We envision an integrated framework of algorithm auditing and quality control that provides a path towards the effective and reliable application of ML systems in healthcare. In this editorial, we give a summary of ongoing work towards that vision and announce a call for participation to the special collection *Machine Learning for Health: Algorithm Auditing & Quality Control* in this journal to advance the practice of ML4H auditing.

Keywords Machine learning · Artificial intelligence · Algorithm · Health · Auditing · Quality control

1 Introduction

Machine learning (ML) technology promises to automate, speed up or improve medical processes. A large number of institutions and companies are ambitiously working on fulfilling this promise spanning tasks such as medical image classification [23], segmentation [80] or reconstruction [3], protein structure prediction [65] and electrocardiography interpretation [73], among others[1]. However, the deployment of machine learning for health (ML4H) tools into real-world applications has been slow because existing approval processes [77] may not account for the particular failure modes and risks that accompany ML technology [56,21,54,10,5]. Certain changes to image data that may not change the decision of a human expert can completely alter the output of an image classification [68] or regression [41,51] model. Model performance estimates are often not valid for the types of varying input distribution that can occur during real world deployment [26,69,75]. The decision heuristics a model learns can differ from the heuristics we may expect a human to use [46,39,23, 47], and model predictions may come with ill-calibrated statements of confidence [22, 6,44] or no estimate of uncertainty altogether [37]. Developers proposing new ML4H technologies sometimes promise to match or even surpass the performance of existing methods [58] yet the reality is often more complicated. Classical ML performance evaluation does not automatically translate to clinical utility as examples from large diabetic retinopathy projects [25] or Covid-19 diagnosis illustrate [49]. The reliable and integrated management of these risks remains an open scientific and practical hurdle.

In order to overcome this hurdle, we envision a framework of algorithm auditing and quality control that provides a path towards the effective and reliable application of ML systems in healthcare. In this editorial we give a brief summary of ongoing work towards that vision from our open collective of collaborators. Many of the considerations presented here originate from a consensus finding effort by the International Telecommunication Union (ITU) and World Health Organization (WHO) which started in 2018 as the Focus Group on Artificial Intelligence for Health (FG-AI4H) [74].

[1] The larger machine learning community maintains a good overview of tasks, benchmarks and state-of-the-art methods at `https://paperswithcode.com/`.

We are convinced that success on this path heavily depends on practical feedback. Auditing processes that are developed on paper have to be put to the test to ensure that they translate to utility in the actual auditing practice [50]. That is why we are introducing the special collection *Machine Learning for Health: Algorithm Auditing & Quality Control* in this journal (see the Call for Participation for more details[2]). The special collection will provide a platform for the submission, discussion and publication of audit methods and reports. The resulting compendium is intended to be a useful resource for users, developers, vendors and auditors of ML4H systems to manage and mitigate their particular risks.

2 ML4H Algorithm Auditing & Quality Control

From a bird's eye view, many ML tools share a set of core components comprising data, an ML-model and its outputs, as visualized in Figure 1 Ⓐ. The typical ML product life cycle goes through stages of planning, development, validation and, potentially, deployment under appropriate monitoring (see Figure 1 Ⓑ). Feedback loops between stages, for example from product validation back to development, are commonplace[3].

An audit entails a detailed assessment of an ML4H tool at one or more of the ML life cycle steps. It can be carried out to anticipate, monitor, or retrospectively review operations of the tool [66,38]. The audit output should consist of a comprehensive standardized report that can be used by different stakeholders to efficiently communicate the tool's strengths and limitations [50]. We envision a process by which an independent body, for example appointed by a government, carries out the audit using the methods and tools outlined below. Further, they can also be used by manufacturers and researchers themselves to carry out internal quality control [59]. In either scenario, the assessment is carried out with respect to a dynamic set of technical, clinical and regulatory considerations (see Figure 1 Ⓒ) that depend on the concrete ML technology and the intended use of the tool. Audit teams should thus comprise expertise in all these dimensions and have to be able to synthesize related requirements across disciplines. In the following, we list a selection of considerations for all three of these auditing dimensions, tools that can be used to aid the auditing process as well as the role so called trial audits can play in advancing ML4H quality control.

2.1 Auditing dimensions

The **technical validation** of an ML4H tool comprises the application of data and ML model quality assessment methods to detect possible failure modes in the model's behavior. These include model-oriented metrics, such as predictive performance, robustness [8,27], interpretability [61,23], disparity [60] or uncertainty [37,52,41] but also data-oriented metrics related to sample size determination [2], sparseness [43],

[2] In the supplement and at this address `https://aiaudit.org/joms/`

[3] Both representations Ⓐ and Ⓑ in Figure 1 are high level abstractions. A granular taxonomy of ML tools or their life cycles is beyond the scope of this editorial. We refer the interested reader to [24] and our documentation [34] for an in-depth treatment.

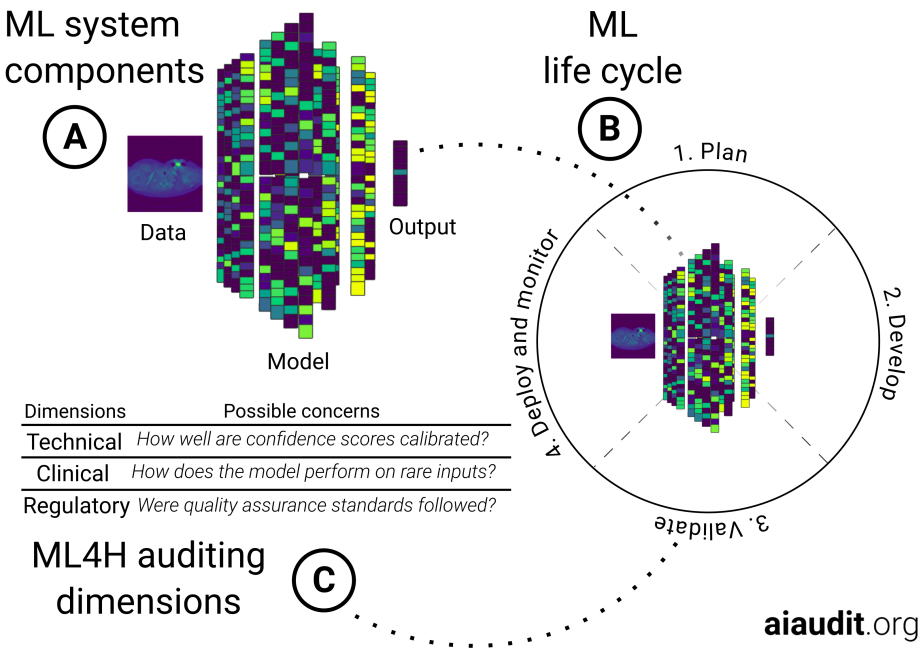


Fig. 1 Process overview. Ⓐ: Most ML tools share a set of core components comprising data, a ML-model and its outputs Ⓑ: The typical ML life cycle goes through stages of planning, development, validation and, potentially, deployment under appropriate monitoring Ⓒ: An ML4H audit is carried out with respect to a dynamic set of technical, clinical and regulatory considerations that depend on the concrete ML technology and the intended use of the tool.

bias [48] distribution mismatch [55,42] and label quality [5]. Rigorous statistical analysis of the model metrics is a common pitfall in both research and industry, and thus plays an important role during technical validation [53]. FG-AI4H has formulated a standardized quality assessment framework based on existing good practices [34,70, 17] and provides practical guidance and examples for performing technical validation audits on three ML4H tools [50].

Clinical evaluation comprises an "ongoing procedure to collect, appraise and analyse clinical data pertaining to a medical device and to analyse whether there is sufficient clinical evidence to confirm compliance with relevant essential requirements for safety and performance when using the device according to the manufacturer's instructions for use" [13]. The EQUATOR-network, including STARD-AI [67], CONSORT-AI [40] and SPIRIT-AI [57], as well as different scientific journals and associations [4,64,29,63], have developed guidelines for the design, implementation, reporting and evaluation of AI interventions in various study designs. Key concerns are whether the ML4H tool delivers utility in clinical pathways, how cost-effective the clinician-tool interaction is [62] and whether it provides the desired benefits for the intended users [19]. To demonstrate reliable performance, it is important to look beyond common machine learning performance statistics such as accuracy and to evaluate in addition whether the ML4H tool is suited to the clinical setting in which it will be used; for example, whether the training and test data represent patient

populations that are similar to the intended use population [5,35] and whether the output translates to medically meaningful parameters [45].

Regulatory assessment comprises the systematic evaluation of ML4H tools with respect to the applicable regulatory requirements found in laws (MDR [12], IVDR [11], 21 CFR [15], among others), to international standards (such as IEC 62304 [30], IEC 62366-1 [31] and ISO 14971 [33]), to guidelines by regulatory bodies (for example FDA [16], IMDRF [32]) or to guidelines and drafts by other organizations (for example AAMI [1] or European Commission [14]). Such guidance is of practical concern for stakeholders in the ML4H ecosystem including manufacturers (e.g. product managers, developers, developers and data scientists, quality and regulatory affairs managers) and for regulatory bodies (authorities, notified bodies). The FG-AI4H has identified and critically reviewed general yet fundamental regulatory considerations related to ML4H. This overview of regulatory considerations assessment have been converted into specific and verifiable requirements and subsequently published as a comprehensive assessment checklist entitled "Good practices for health applications of machine learning: Considerations for manufacturers and regulators" [34] which covers the entire life cycle outlined in Figure 1 Ⓑ at a higher resolution. It includes checklist items which should be given high priority in the presence of limited time - an important practical constraint for real-world audits. Examples and comments give further guidance to users. New regulatory developments, such as predetermined change control plans [71], imply faster software update cycles and potentially more frequent audits. Hence, good tooling can become an important means to make effective as well as efficient audits possible.

2.2 Auditing Tools

The auditing process can be supported by appropriate tools to make it more targeted and time-efficient. This can include process and requirements descriptions, as mentioned above [17,34,19], which help to manage dynamic workflows that may vary by use case and ML technology. It also includes reporting templates to present the audit results in a standardized way for the communication between different stakeholders. [72,50]. In addition, the nature of ML4H tools, as primarily software that interacts with data, lends itself to the application of test automation and simulations for the purpose of auditing. This requires software tools which can handle custom evaluation scripts, the flexible processing of different ML4H model formats and data modalities as well as security protocols that protect intellectual property and sensitive patient information [20]. We are working with open source frameworks such as EvalAI [78] and MLflow [9] to develop solutions for automated auditing[4], federated auditing in remote teams[5] and automated report creation. Our first demo platform is available via `http://health.aiaudit.org/`[6] and hosted on ITU provisioned infrastructure. While quantitative performance measures can already be provided, it is essential to also offer qualitative measures. This is realized by requiring the users to fill out a standardized questionnaire [18]. Quantitative and qualitative performance results are then provided to the users as a comprehensive and standardized report card [72].

[4] `https://github.com/aiaudit-org/health-aiaudit-public`

[5] `https://github.com/aiaudit-org/amazon-sagemaker-mlflow-fargate`

[6] You are welcome to reach out to any of the contributors `https://aiaudit.org/contributors/` for information on how to join the efforts.

2.3 Trial Audits

We are convinced that success on the path towards a framework for algorithm auditing and quality control depends heavily on practical feedback. The development and refinement of auditing processes should routinely be accompanied by trial audits. In trial audits, draft processes and standards are applied to ML4H tools. The purpose of such an exercise is to ensure that auditing processes developed on paper translate to utility in actual auditing practice [50]. In order to facilitate the implementation of trial audits, we are introducing the special collection *Machine Learning for Health: Algorithm Auditing & Quality Control* in this journal. We welcome contributions pertaining to methods, tools, reports or open challenges in ML4H auditing.

3 Outlook

The materials summarized above bear testimony to the initial progress that has been made towards the creation of frameworks for ML4H algorithm auditing and quality control. Nevertheless, new challenges emerge as we collectively pull at the complex fabric that ML4H systems are.

From the perspective of technical validation, the identification of factors which bias or deteriorate algorithmic performance is often constrained by the absence of relevant metadata. For example, the measurement device types (and related acquisition parameters) used to produce the validation inputs should be available in order to validate if the model performance is robust under device type changes. This problem can be alleviated by identifying and routinely recording this information during data acquisition.

For clinical evaluation, future considerations include extending and refining the specific requirements related to how the clinical effectiveness of a tool should be monitored after implementation of the algorithm and with ongoing monitoring [12]. This also requires agreement over the clear and clinically useful procedures to obtain ground truth annotations. It might be necessary to refine the ML algorithm to the target population, if demographics or clinical character are different from training settings or if medical guidelines for diagnostics or treatment have changed [36]. Therefore, in order for these insights to be effective it is imperative that auditors exhibit a solid understanding of the training data, ML algorithm, independent test data and evaluation metrics specific to the intended use.

A challenge for regulatory assessment is that standardization organizations, notified bodies and manufacturers need to efficiently formulate and parse applicable regulatory requirements for each individual ML4H tool. Comprehensive assessment checklists [34, 4] can help with that task. However, more support is needed in terms of workflow management and assisting tools if we consider the limited time and budgets which professional auditors have at their disposal. Future regulatory checklists should allow for interactive selection of use-case specific sub-checklists, an automated audit report creation, a collection of standard minimum test cases as well as accompanying glossaries and education materials for auditors. We also have to ensure that protocols are in place which translate the audit insights to actual improvements in the ML4H tool. Managing the risks presented by the exciting advances of AI in healthcare is a formidable undertaking, but with collaborative pooling of expertise and resources we believe we can rise to the task.

References

1. AAMI: Techical report (tr) 57 principals for medical device security - risk management. URL https://store.aami.org/s/store#/store/browse/detail/a152E000006j60WQAQ
2. Balki, I., Amirabadi, A., Levman, J., Martel, A.L., Emersic, Z., Meden, B., Garcia-Pedrero, A., Ramirez, S.C., Kong, D., Moody, A.R., et al.: Sample-size determination methodologies for machine learning in medical imaging research: a systematic review. Canadian Association of Radiologists Journal **70**(4), 344–353 (2019)
3. Bubba, T.A., Kutyniok, G., Lassas, M., März, M., Samek, W., Siltanen, S., Srinivasan, V.: Learning the invisible: a hybrid deep learning-shearlet framework for limited angle computed tomography. Inverse Problems **35**(6), 064002 (2019)
4. Cabitza, F., Campagner, A.: The need to separate the wheat from the chaff in medical informatics. International Journal of Medical Informatics p. 104510 (2021). DOI https://doi.org/10.1016/j.ijmedinf.2021.104510
5. Cabitza, F., Campagner, A., Sconfienza, L.M.: As if sand were stone. new concepts and metrics to probe the ground on which to build trustable ai. BMC Medical Informatics and Decision Making **20**(1), 1–21 (2020)
6. Calderon-Ramirez, S., Yang, S., Moemeni, A., Colreavy-Donnelly, S., Elizondo, D.A., Oala, L., Rodríguez-Capitán, J., Jiménez-Navarro, M., López-Rubio, E., Molina-Cabello, M.A.: Improving uncertainty estimation with semi-supervised deep learning for covid-19 detection using chest x-ray images. IEEE Access **9**, 85442–85454 (2021). DOI 10.1109/ACCESS.2021.3085418
7. Candès, E., Fan, Y., Janson, L., Lv, J.: Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **80**(3), 551–577 (2018)
8. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 ieee symposium on security and privacy (sp), pp. 39–57. IEEE (2017)
9. Chen, A., Chow, A., Davidson, A., DCunha, A., Ghodsi, A., Hong, S.A., Konwinski, A., Mewald, C., Murching, S., Nykodym, T., Ogilvie, P., Parkhe, M., Singh, A., Xie, F., Zaharia, M., Zang, R., Zheng, J., Zumar, C.: Developments in mlflow: A system to accelerate the machine learning lifecycle. In: Proceedings of the Fourth International Workshop on Data Management for End-to-End Machine Learning, DEEM'20. Association for Computing Machinery, New York, NY, USA (2020). DOI 10.1145/3399579.3399867. URL https://doi.org/10.1145/3399579.3399867
10. D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M.D., et al.: Underspecification presents challenges for credibility in modern machine learning. arXiv preprint arXiv:2011.03395 (2020)
11. EU: Regulation (eu) 2017/746 of the european parliament and of the council on in vitro diagnostic medical devices (2017). URL https://eur-lex.europa.eu/eli/reg/2017/746/oj
12. EU: Regulation (eu) 2017/746 of the european parliament and of the council on medical devices (2017). URL https://eur-lex.europa.eu/eli/reg/2017/745/oj
13. EUROPEAN-COMMISSION: Meddev 2.7/1 revision 4, clinical evaluation: a guide for manufacturers and notified bodies. https://ec.europa.eu/docsroom/documents/17522/attachments/1/translations/en/renditions/native (2016). (Accessed on 07/01/2021)
14. EUROPEAN-COMMISSION: Eur-lex - 52021pc0206 - en - eur-lex. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206 (2021). (Accessed on 07/01/2021)
15. FDA: Code of federal regulations, title 21 on foods and drugs. URL https://www.ecfr.gov/cgi-bin/text-idx?SID=cc74806513924f0197b7809c8efbefc8&mc=true&tpl=/ecfrbrowse/Title21/21tab_02.tpl
16. FDA: Fda guidance documents. URL https://www.fda.gov/regulatory-information/search-fda-guidance-documents
17. FG-AI4H: Data and artificial intelligence assessment methods (daisam) reference. Reference document DEL 7.3 on FG-AI4H server (2020). URL https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/SitePages/Home.aspx
18. FG-AI4H: Model questionnaire. Reference document J-038 on FG-AI4H server (2020). URL https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/SitePages/Home.aspx
19. FG-AI4H: Clinical evaluation of ai for health. Reference document DEL 7.4 on FG-AI4H server (2021). URL https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/SitePages/Home.aspx

20. FG-AI4H: Data sharing practices. Reference document DEL 5.6 on FG-AI4H server (2021). URL https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/SitePages/Home.aspx
21. Gilmer, J., Ford, N., Carlini, N., Cubuk, E.: Adversarial examples are a natural consequence of test error in noise. In: International Conference on Machine Learning, pp. 2280–2289. PMLR (2019)
22. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning, pp. 1321–1330. PMLR (2017)
23. Hägele, M., Seegerer, P., Lapuschkin, S., Bockmayr, M., Samek, W., Klauschen, F., Müller, K.R., Binder, A.: Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. Scientific Reports **10**(1), 1–12 (2020)
24. Hardt, M., Recht, B.: Patterns, predictions, and actions: A story about machine learning. https://mlstory.org (2021)
25. Heaven, W.D.: Google's medical ai was super accurate in a lab. real life was a different story. — mit technology review. https://www.technologyreview.com/2020/04/27/1000658/google-medical-ai-accurate-lab-real-life-clinic-covid-diabetes-retina-disease/. (Accessed on 06/10/2021)
26. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al.: The many faces of robustness: A critical analysis of out-of-distribution generalization. arXiv preprint arXiv:2006.16241 (2020)
27. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261 (2019)
28. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. Proceedings of the International Conference on Learning Representations (2019)
29. Hernandez-Boussard, T., Bozkurt, S., Ioannidis, J.P., Shah, N.H.: Minimar (minimum information for medical ai reporting): developing reporting standards for artificial intelligence in health care. Journal of the American Medical Informatics Association **27**(12), 2011–2015 (2020)
30. IEC: Medical device software — software life cycle processes — amendment 1 (2015). URL https://www.iso.org/standard/64686.html
31. IEC: Medical devices — part 1: Application of usability engineering to medical devices — amendment 1 (2020). URL https://www.iso.org/standard/73007.html
32. IMDRF: Documents by international medical device regulators forum. URL http://www.imdrf.org/documents/documents.asp
33. ISO: Medical devices — application of risk management to medical devices (2019). URL https://www.iso.org/standard/72704.html
34. Johner, C., Balachandran, P., Oala, L., Lee, A.Y., Werneck Leite, A., Murchison, A., Lin, A., Molnar, C., Rumball-Smith, J., Baird, P., Goldschmidt, P.G., Quartarolo, P., Xu, S., Piechottka, S., Hornberger, Z.: Good practices for health applications of machine learning: Considerations for manufacturers and regulators. In: L. Oala (ed.) ITU/WHO Focus Group on Artificial Intelligence for Health (FG-AI4H) - Meeting K, vol. K. ITU (2021). URL https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/SitePages/Home.aspx
35. Kaushal, A., Altman, R., Langlotz, C.: Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms. JAMA **324**(12), 1212–1213 (2020). DOI 10.1001/jama.2020.12067. URL https://doi.org/10.1001/jama.2020.12067
36. Kelly, C.J., Karthikesalingam, A., Suleyman, M., Corrado, G., King, D.: Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine **17**, 195 (2019)
37. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017). URL https://proceedings.neurips.cc/paper/2017/file/2650d6089a6d640c5e85b2b88265dc2b-Paper.pdf
38. Koshiyama, A., Kazim, E., Treleaven, P., Rai, P., Szpruch, L., Pavey, G., Ahamat, G., Leutner, F., Goebel, R., Knight, A., et al.: Towards algorithm auditing: A survey on managing legal, ethical and technological risks of ai, ml and associated algorithms (2021)
39. Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.R.: Unmasking clever hans predictors and assessing what machines really learn. Nature Communications **10**(1), 1–8 (2019)
40. Liu, X., Cruz Rivera, S., Moher, D., Calvert, M., Denniston, A.K., Spirit-ai, T., Group, C.a.W.: CONSORT-AI extension. Nature Medicine **26**(September), 1364–1374 (2020). DOI 10.1038/s41591-020-1034-x

41. Macdonald, J., März, M., Oala, L., Samek, W.: Interval neural networks as instability detectors for image reconstructions. In: C. Palm, T.M. Deserno, H. Handels, A. Maier, K. Maier-Hein, T. Tolxdorff (eds.) Bildverarbeitung für die Medizin 2021, pp. 324–329. Springer Fachmedien Wiesbaden, Wiesbaden (2021)
42. Mårtensson, G., Ferreira, D., Granberg, T., Cavallin, L., Oppedal, K., Padovani, A., Rektorova, I., Bonanni, L., Pardini, M., Kramberger, M.G., et al.: The reliability of a deep learning model in clinical out-of-distribution mri data: a multicohort study. Medical Image Analysis **66**, 101714 (2020)
43. Mendez, M., Calderon-Ramirez, S., Tyrrell, P.N.: Using cluster analysis to assess the impact of dataset heterogeneity on deep convolutional network accuracy: A first glance. In: Latin American High Performance Computing Conference, pp. 307–319. Springer (2019)
44. Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., Lucic, M.: Revisiting the calibration of modern neural networks (2021)
45. Nagendran, M., Chen, Y., Lovejoy, C.A., Gordon, A.C., Komorowski, M., Harvey, H., Topol, E.J., Ionnidis, J.P.A., Collins, G.S., Maruthappu, M.: Artificial intelligence versus clinicians: systematic review of design, reporting standards, and clains of deep learning studies. British Medical Journal **360**, m689 (2020)
46. Nalisnick, E., Matsukawa, A., Teh, Y.W., Gorur, D., Lakshminarayanan, B.: Do deep generative models know what they don't know? arXiv preprint arXiv:1810.09136 (2018)
47. Neves, I., Folgado, D., Santos, S., Barandas, M., Campagner, A., Ronzio, L., Cabitza, F., Gamboa, H.: Interpretable heartbeat classification using local model-agnostic explanations on ecgs. Computers in Biology and Medicine **133**, 104393 (2021)
48. Noseworthy, P.A., Attia, Z.I., Brewer, L.C., Hayes, S.N., Yao, X., Kapa, S., Friedman, P.A., Lopez-Jimenez, F.: Assessing and mitigating bias in medical artificial intelligence: the effects of race and ethnicity on a deep learning model for ecg analysis. Circulation: Arrhythmia and Electrophysiology **13**(3), e007988 (2020)
49. Oakden-Rayner, L.: Ct scanning is just awful for diagnosing covid-19 – luke oakden-rayner. `https://lukeoakdenrayner.wordpress.com/2020/03/23/ct-scanning-is-just-awful-for-diagnosing-covid-19/`. (Accessed on 06/10/2021)
50. Oala, L., Fehr, J., Gilli, L., Balachandran, P., Leite, A.W., Calderon-Ramirez, S., Li, D.X., Nobis, G., Alvarado, E.A.M.n., Jaramillo-Gutierrez, G., Matek, C., Shroff, A., Kherif, F., Sanguinetti, B., Wiegand, T.: Ml4h auditing: From paper to practice. In: Proceedings of the Machine Learning for Health NeurIPS Workshop, vol. 136, pp. 280–317. PMLR (2020)
51. Oala, L., Heiß, C., Macdonald, J., März, M., Kutyniok, G., Samek, W.: Detecting failure modes in image reconstructions with interval neural network uncertainty. International Journal of Computer Assisted Radiology and Surgery pp. 1–9 (2021)
52. Oala, L., Heiß, C., MacDonald, J., März, M., Samek, W., Kutyniok, G.: Interval neural networks: Uncertainty scores. CoRR **abs/2003.11566** (2020). URL `https://arxiv.org/abs/2003.11566`
53. Parmar, C., Barry, J.D., Hosny, A., Quackenbush, J., Aerts, H.J.: Data analysis strategies in medical imaging. Clinical cancer research **24**(15), 3492–3499 (2018)
54. Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., Barnes, P.: Closing the ai accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 33–44 (2020)
55. Ramírez, S.C., Oala, L.: More than meets the eye: Semi-supervised learning under non-iid data. CoRR **abs/2104.10223** (2021). URL `https://arxiv.org/abs/2104.10223`
56. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do imagenet classifiers generalize to imagenet? In: International Conference on Machine Learning, pp. 5389–5400. PMLR (2019)
57. Rivera, S.C., Liu, X., Chan, A.W., Denniston, A.K., Calvert, M.J.: Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension. Bmj **370**, m3210 (2020). DOI 10.1136/bmj.m3210. URL `http://www.bmj.com/lookup/doi/10.1136/bmj.m3210`
58. Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A.I., Etmann, C., McCague, C., Beer, L., et al.: Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans. Nature Machine Intelligence **3**(3), 199–217 (2021)
59. Ryan, J.R.: Software product quality assurance. In: Proceedings of the June 7-10, 1982, National Computer Conference, AFIPS '82, p. 393–398. Association for Computing Machinery, New York, NY, USA (1982). DOI 10.1145/1500774.1500823. URL `https://doi.org/10.1145/1500774.1500823`

60. Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K.T., Ghani, R.: Aequitas: A bias and fairness audit toolkit. arXiv preprint arXiv:1811.05577 (2018)
61. Samek, W., Montavon, G., Lapuschkin, S., Anders, C.J., Müller, K.R.: Explaining deep neural networks and beyond: A review of methods and applications. Proceedings of the IEEE **109**(3), 247–278 (2021)
62. Schwendicke, F., Rossi, J., Göstemeyer, G., Elhennawy, K., Cantu, A., Gaudin, R., Chaurasia, A., Gehrung, S., Krois, J.: Cost-effectiveness of artificial intelligence for proximal caries detection. Journal of Dental Research **100**(4), 369–376 (2021). DOI 10.1177/0022034520972335. URL `https://doi.org/10.1177/0022034520972335`. PMID: 33198554
63. Schwendicke, F., Singh, T., Lee, J.H., Gaudin, R., Chaurasia, A., Wiegand, T., Uribe, S., Krois, J.: Artificial intelligence in dental research: Checklist for authors, reviewers, readers. Journal of Dentistry **107**, 103610 (2021). DOI https://doi.org/10.1016/j.jdent.2021.103610. URL `https://www.sciencedirect.com/science/article/pii/S0300571221000312`
64. Scott, I., Carter, S., Coiera, E.: Clinician checklist for assessing suitability of machine learning applications in healthcare. BMJ Health & Care Informatics **28**(1) (2021)
65. Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A.W., Bridgland, A., et al.: Improved protein structure prediction using potentials from deep learning. Nature **577**(7792), 706–710 (2020)
66. Shneiderman, B.: Opinion: The dangers of faulty, biased, or malicious algorithms requires independent oversight. Proceedings of the National Academy of Sciences **113**(48), 13538–13540 (2016). DOI 10.1073/pnas.1618211113. URL `https://www.pnas.org/content/113/48/13538`
67. Sounderajah, V., Ashrafian, H., Aggarwal, R., De Fauw, J., Denniston, A.K., Greaves, F., Karthikesalingam, A., King, D., Liu, X., Markar, S.R., McInnes, M.D., Panch, T., Pearson-Stuttard, J., Ting, D.S., Golub, R.M., Moher, D., Bossuyt, P.M., Darzi, A.: Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group. Nature Medicine **26**(6), 807–808 (2020). DOI 10.1038/s41591-020-0941-1. URL `http://dx.doi.org/10.1038/s41591-020-0941-1`
68. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
69. Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., Schmidt, L.: Measuring robustness to natural distribution shifts in image classification. arXiv preprint arXiv:2007.00644 (2020)
70. The Supreme Audit Institutions of Finland, Germany, the Netherlands, Norway and the UK: Auditing machine learning algorithms. `https://auditingalgorithms.net/` (2020). (Accessed on 07/02/2021)
71. US-FDA: Aiml_samd_action_plan. `https://www.fda.gov/media/145022/download?utm_medium=email&utm_source=govdelivery` (2021). (Accessed on 07/01/2021)
72. Verks, B., Oala, L.: Daisam audit reporting template. In: ITU/WHO Focus Group on Artificial Intelligence for Health (FG-AI4H) - Meeting J, vol. J. ITU (2020). URL `https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/SitePages/Home.aspx`
73. Wagner, P., Strodthoff, N., Bousseljot, R.D., Kreiseler, D., Lunze, F.I., Samek, W., Schaeffter, T.: Ptb-xl, a large publicly available electrocardiography dataset. Scientific Data **7**(1), 1–15 (2020)
74. Wiegand, T., Krishnamurthy, R., Kuglitsch, M., Lee, N., Pujari, S., Salathé, M., Wenzel, M., Xu, S.: Who and itu establish benchmarking process for artificial intelligence in health. The Lancet **394**(10192), 9–11 (2019)
75. Willis, K., Oala, L.: Post-hoc domain adaptation via guided data homogenization. CoRR **abs/2104.03624** (2021). URL `https://arxiv.org/abs/2104.03624`
76. Wong, A., Otles, E., Donnelly, J.P., Krumm, A., McCullough, J., DeTroyer-Cooley, O., Pestrue, J., Phillips, M., Konye, J., Penoza, C., Ghous, M., Singh, K.: External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. JAMA Internal Medicine **181**(8), 1065–1070 (2021). DOI 10.1001/jamainternmed.2021.2626. URL `https://doi.org/10.1001/jamainternmed.2021.2626`
77. Wu, E., Wu, K., Daneshjou, R., Ouyang, D., Ho, D.E., Zou, J.: How medical ai devices are evaluated: limitations and recommendations from an analysis of fda approvals. Nature Medicine **27**(4), 582–584 (2021)

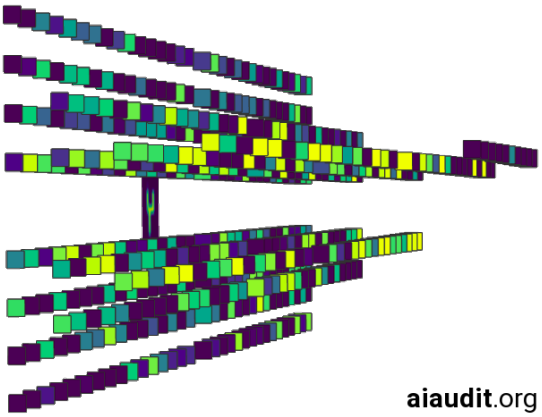
78. Yadav, D., Jain, R., Agrawal, H., Chattopadhyay, P., Singh, T., Jain, A., Singh, S., Lee, S., Batra, D.: Evalai: Towards better evaluation systems for AI agents. CoRR **abs/1902.03570** (2019). URL `http://arxiv.org/abs/1902.03570`
79. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning (still) requires rethinking generalization. Communications of the ACM **64**(3), 107–115 (2021)
80. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep learning in medical image analysis and multimodal learning for clinical decision support, pp. 3–11. Springer (2018)

Appendices

Call for Papers

The Journal of Medical Systems (JOMS) and the aiaudit.org collective invite submissions to the special issue on

Machine Learning for Health: Algorithm Auditing & Quality Control



We invite researchers and practitioners in the fields of machine learning, medicine, regulation and quality management to submit their work. The scope comprises (trial) **audit reports** on ML4H applications as well as work related to **methods**, **tools** or **open challenges** in ML4H auditing.

The resulting compendium is intended as a useful resource of examples and guidance for users, developers, vendors and auditors of ML4H systems to better understand, manage and mitigate their particular risks. Manuscripts which treat topics from an integrated perspective across disciplines are encouraged. Below you can find a list of examples for each category as possible guidance:

Audit reports

- External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients [76]
- ML4H Auditing: From Paper to Practice [50]

Methods

- Benchmarking Neural Network Robustness to Common Corruptions and Perturbations [28]
- Panning for gold:'model-X'knockoffs for high dimensional controlled variable selection [7]

Tools

- EvalAI: Towards Better Evaluation Systems for AI Agents [78]
- Developments in MLflow: A System to Accelerate the Machine Learning Lifecycle [9]

Open challenges

- Understanding deep learning (still) requires rethinking generalization [79]
- Underspecification presents challenges for credibility in modern machine learning [10]

Manuscripts should be submitted as Original Research Articles (up to 3,000 words) or Brief Technical Reports (up to 1,500 words) through the JOMS Editorial Manager[7] and will, upon successful peer-review and acceptance, be published in JOMS. You can find more information on the overall submission guidelines of the journal at the address linked below[8]. Please use the document template provided by JOMS[9,10].

In case of questions you can reach the editorial team via

Ednalyn Reyes
ednalyn.reyes@springernature.com.

We are looking forward to receiving your contributions.

On behalf of the editorial team
Jesse Ehrenfeld, Ednalyn Reyes, Luis Oala
Cyberspace, October 18, 2021

[7] https://www.editorialmanager.com/joms/default.aspx

[8] https://www.springer.com/journal/10916/submission-guidelines

[9] LaTeXhttps://static.springer.com/sgw/documents/468198/application/zip/LaTeX_DL_468198_01072021.zip

[10] MS Word https://static.springer.com/sgw/documents/431298/application/zip/sv-journ.zip