Digital Society (2023) 2:46
https://doi.org/10.1007/s44206-023-00076-w

ORIGINAL PAPER

Human-Curated Validation of Machine Learning Algorithms for Health Data

Magnus Boman[1,2,3]

Received: 2 March 2023 / Accepted: 18 September 2023 / Published online: 19 October 2023

Abstract

Validation of machine learning algorithms that take health data as input is analysed, leveraging on an example from radiology. A 2-year study of AI use in a university hospital and a connected medical university indicated what was often forgotten by human decision makers in the clinic and by medical researchers. A nine-item laundry list that does not require machine learning expertise to use resulted. The list items guide stakeholders toward complete validation processes and clinical routines for bias-aware, sound, energy-aware and efficient data-driven reasoning for health. The list can also prove useful to machine learning developers, as a list of minimal requirements for successful implementation in the clinic.

Keywords Gold standard · Ground truth · Health data · Machine learning · Deep learning · Artificial intelligence · Validation · Bias

1 Introduction

Machine learning (ML) algorithms require validation, externally as well as internally. Many professional ML developers lack experience in the relatively strict standards of external validation of medical research, possibly contributing to the relatively slow uptake of, in particular, deep learning in medicine (Topol, 2019). The practice of registering protocols and analysis plans before a study commences are two examples of external validation not employed in most other domains (Chambers et al., 2015; Klau et al., 2021; Ioannidis, 2022). By contrast, professional ML

✉ Magnus Boman
mab@kth.se; magnus.boman@ki.se

1 Software and Computer Sciences (EECS/SCS), KTH Royal Institute of Technology, Electrum 229, Kista SE-16440, Sweden

2 Division of Clinical Epidemiology, Department of Medicine Solna and LIME, Karolinska Institutet, Stockholm SE-17177, Sweden

3 MedTechLabs, BioClinicum, Karolinska University Hospital, Solna, Stockholm SE-17176, Sweden

developers are self-proclaimed experts in matters of internal validation, with a 5- or 10-fold cross-validation and advanced variable importance determination procedures as typical examples (Wei et al., 2015). For internal validation too, however, there are cultural differences to consider. For example, medical researchers often keep a holdout sample out of their training, in order to validate on completely unseen data. In ML, the 10 or 20 per cent kept out of training and being reserved for testing only over the cross-validation folds is usually considered enough (Eloranta & Boman, 2022). This may sound as if the validation data are kept out of the training data in both instances, but it is important to note that some ML models may benefit from recalling training data points at the time of testing, making the model overfit to training data. The worst examples include developers choosing the best fold when deciding how to report quantitative training results, but there are more subtle forms of cherry-picking too, like choosing cut-offs for test data runs (so-called hyperparameter tuning) informed by earlier test data runs; the methodologically sound thing to do is instead to only employ the training runs for this purpose. Further differences emerge when the training or testing data is stratified, to allow for the test population to meet the target population (Zendel et al., 2017).

Humans curate validation processes for both internal and external validation of ML algorithms. Given quantitative measurements of model performance, a comparison is sometimes made with quantitative measurements of human performance on the same task. The latter is typically a prediction or classification task, which could improve practice, by supporting less experienced or non-specialised clinicians in assessing risks (Kadir & Gleeson, 2018). Many reviews of the feasibility of ML models for problems in medicine compare human to machine performance (Nagendran et al., 2020). The machine can then be seen as emulating a human *ground truth*, consisting of labels that type data instances, for example indicating urgency using a high/medium/low scale. Such a set of labels is often an important part of a clinical *gold standard*. The folklore now has it that ML algorithms, the learning of which is supervised by human labels, can outdo humans on some problems, in particular within medical imaging (Liu et al., 2019a; Nagendran et al., 2020). This seems natural, as anybody would agree that computers are better than humans on purely computational tasks, and many imaging problems are so-called compute problems. This could be Fourier transforms to turn raw data into comprehensible images for the purpose of studying pathologies or voxel-by-voxel comparisons of two scans of the same region in the same body to study change. Machines excel in detecting change patterns, the *Delta*, so problem and technical solution seem well matched. But to outdo the very humans that the algorithm is supposed to emulate would mean that any step away from an entirely human-produced ground truth would also risk diminishing the quantitative performance of the algorithm, a first indication that the game might be rigged against AI.

The gold standard, when it exists, is non-trivial to relate even to simple regression-like ML prediction and classification tasks in the health domain. Moreover, the latter is to an increasing extent associated with such tasks, since the awareness of and the number of applications of artificial intelligence (AI) steadily grow. Some such associations are problematic enough to have wide-ranging ethical implications. Reasoning by example, mostly from radiology, but representing larger and other

categories, methodological issues can be overcome for this entire set of problems. This can in turn make the best use of ML in the clinic a possibility by devising and employing sound processes for validation, cf. (Bera et al., 2022). The purpose is therefore to identify and illustrate key problems, and a few pitfalls, as well as how to address these.

2 Related Research

For validation in the domain at hand, it is not meaningful to discuss state-of-the-art, as this would require commensurability between solutions. Even for relatively simple problems, such as correctly identifying symptoms of well-known conditions, quantitative comparisons between different ML models are often dubious. Statements such as "Random forest is the best model for patient data, unless we are interested in development over time" are often heard at seminars, but are utterly meaningless. Regrettably, the literature is not rich on such methodological quandaries; only a few recent exposés exist. Varoquaux and Cheplygina point to problems with internal cross-validation in ML when it comes to clinical relevance: the test data should not be a random subset of the training data (Varoquaux & Cheplygina, 2022). The authors also discuss in exemplary detail dataset bias and poorly chosen baselines and poor benchmarking for quantitative results. For all of those problems, strict stratification to match outcomes of studies is necessary and may, in some cases, even reveal hidden bias (Carbonell et al., 2021), although it hurts in the short term to validate your ML models on the hardest challenges possible, which is what correct stratification does.

Adamson and Welch bring up two sides of what they call *the gold-standard problem* (Adamson & Welch, 2019). Firstly, the clinical problem of what constitutes cancer is a dynamic problem, while the pathology is based on static observations. Second, pathologists disagree on histopathological diagnoses. The latter have changed, the authors note, from tumours that could be felt with the human hand to microscopic cellular abnormalities. The authors also point to the ubiquity of ML as a source of overdiagnosis, the automation of which could turn into a nightmare for both patients and clinicians. Mitigating that risk is best done via panels of human and artificial decision makers who may focus on the information disagreed upon on the pathology side. Strand et al. have adopted such an approach for breast cancer studies (Cossío et al., 2023; Dembrower et al., 2020b), for triaging by ML algorithms, combining human experts with commercial AI cancer detection algorithms. Their purpose was to move radiologist focus from clearly negative mammograms to more closely investigating women at risk of having a false-negative screening. Such approaches have been found to improve inter-physician consistency (Freer & Ulissey, 2001). Strand's group has also produced reference data sets (Dembrower et al., 2020a), in turn leading to a call for controlled validation data sets (Strand et al., 2021). Current research in general is described in a recent review (Anderson et al., 2022), but the preference of radiologists on how to best improve the current best practice remains unknown (Hendrix et al., 2022).

Finally, strict protocols for research, such as TRIPOD (Moons et al., 2015), recommend that folding to be done over time (variables) only and that model adequacy be measured with ROC-AUC. As this is sometimes not convenient or even sound for some ML models, a special protocol TRIPOD-ML is under development, as are ML-adapted versions of several other protocols (Faes et al., 2020). The adherence to such protocols is still low in many disciplines, however (Belue et al., 2022), arguably due to lack of trust (Ferrario et al., 2020).

3 Example of Human Validation: Radiology

Machine learning is in healthcare often applied to diagnosis or triaging tasks. Even relatively simple diagnosis tasks, like the ordering by urgency of a set of scans to be assessed by radiologists, can have great clinical value (Sachs et al., 2020). It has proved more useful to have humans correct automated triaging than to have machines give a second opinion on human triaging (Kooli & Al Muftah, 2022). The concept of AI-enhanced human decision support systems has to a large extent been replaced by AI systems that generate and apply their own rules for decision support. Depersonalisation is sometimes listed as a risk of such approaches, but at least today, AI in healthcare is self-learning, one or two steps away from human validation by ocular inspection. An example here would be using self-configuring so-called *U-net* approaches to learning segmentation. A neural network U-net (Isensee et al., 2021) can cleverly segment interesting areas, e.g. for computerised tomography scans, and the Dice coefficient then measures the success rate with which the learning system can emulate human segmentation.

In a recently concluded pilot study of such a system for finding metastases in the adrenal glands using CT data (paper in preparation), a Dice coefficient median score of 0.89 was achieved when comparing a self-supervised ML approach to radiologists, where 1.00 indicates a perfect match, down to the last voxel. This is therefore an example in which one may choose to go beyond successful pilot into the clinic. The learning approach is also generalisable to other organs, making further validation such as through procedures of CE marking, Medical Device Regulation classification and randomised controlled trials a future possibility. The distribution of Dice scores was interesting, in that several of these were zero, meaning that no voxel was in the intersection between the ML segmentation and the ground truth. A radiologist (different from the radiologists who had done the segmentation for the data) scrutinised such cases manually and found various reasons for the ML algorithm doing so poorly, some of which were interesting in their own right. But the crucial finding was some CT scans for which the ML algorithm actually outperformed the ground truth, as discovered by the radiologist. This points to a problem of validation mentioned in the introduction: what is measured is the extent to which the algorithm emulates the human-generated ground truth, not how well it is doing its job. The main reason for zero Dice score was that in some cases, the open-source software tool used by the radiologists for the segmentation mixed up the left and right adrenal gland, as a result of a bug in how scans were named and labelled. This bug was reported by other users of the software already in 2020, when the developers put it

on their ToDo list. The bug is still marked as unresolved, however, with a promise to fix it before the year 2030.

One might think of this as a cautionary tale on the uncertainty attached to open source code quality, but proprietary software suffers from a general problem even more serious. In order to keep the customer satisfied, many (though far from all) commercial pieces of ML software lend themselves to producing the optimal quantitative results by means of methods that do not hold up to scientific scrutiny. A recent example involves software that always trains on the validation set, leading to higher quantitative scores on the validation set than on the test set, a case of spectrum bias in which the target population is inadequately represented (Park & Han, 2018). Any published result for which this is true should alert the reader to a possibly circular analysis problem, producing extreme overfitting of the data and less scalability (Pulini et al., 2019), and sadly, there are many such publications to be found. That the software is proprietary makes direct line-by-line inspection of the code impossible and could lead to sponsorship bias. It also forces anybody interested in reproducing results to buy software licenses, which may come at prohibitive cost. Only in careful reproduction—known as *docking* models, by aligning them computationally (Axtell et al., 1996)—are such tricks revealed, and then only by reverse engineering of the unavailable code.

Another problem lies in the interdisciplinarity of AI employment in healthcare. Many ML programmers want to contribute to solving health problems and feel it is natural to compete internally so that the best technical solutions can be offered to solve such problems. For simplicity and perhaps in part by ignorance, many such competitions are addressing problems that seem relevant but in fact might be red herrings. There are, for instance, many competitions in which ML programmers compete on performance of binary classifiers of scans (cancer or no cancer?). The clinical question is rarely one of cancer or not, since the final diagnosis can only in exceptional cases be made from a CT scan only, but instead of choosing the appropriate investigative algorithm for a given patient in a given clinical situation (referral or not, biopsy or not, etc.) (Sachs et al., 2020). As a result, there might be a gap between such competitions and clinical needs (Kadir & Gleeson, 2018). This is a general problem of such competitions (Masnick, 2012), but in healthcare, it might also contribute to frustration with inadequacy of AI solutions in the clinic, diminishing human trust in such algorithms (Roberts et al., 2021). In fact, in radiology for solid tumours, the reasoning is often counterfactual: if there is no growth within a suspected pathological structure, some clinical actions can be ruled out. In ML research, uptake of counterfactual reasoning mechanisms has been slow (Verma et al., 2021), often under the guise of interpretability and explainability of AI systems.

Finally, in radiology, human assessments have vastly improved over the years, in part thanks to improved methodology and technology. Hence, not only algorithms move toward perfection, humans do too. This produces a problem that has only recently been highlighted within ML research, namely that training on old (and less informed than now) human labels may not have the expected positive results on the output of the algorithm. With a general shortage of training data, it is tempting to use older data if it is available. There are also many distinctions with respect to data

quality between purposefully gathered data and found data. To just throw more data and compute at a problem is not always a good strategy.

4 Methods

As part of a strategic AI project at Karolinska Institutet (KI) at Stockholm, principal investigators and key researchers in more than 30 AI projects in the KI ecosystem were interviewed, see (Boman, 2022) for examples and quotes. Something that clinicians found crucial was scope: What does AI and ML denote, and how do they relate? A full scientometric analysis of AI and ML in that ecosystem was therefore completed, which was useful for the present article too (Boman et al., 2022). Rather than an auto-ethnographic study of problems encountered with key stakeholders, findings from the set of semi-structured interviews, with the interviewees scoped using the scientometry, provided the foundation. The findings also helped identify the radiology example used above. Closing the gap between what is there today for efficient human validation of medical and care data and what is needed for such procedures to merge with automated ones, the findings are collected as a laundry list (framed below, and detailed in Table 1).

In connection with references for that table, the concept of automation includes self-learning models, models that write code, and so-called *foundation models*: huge pre-trained structures that require little to no domain adaptation to work for health data. Adaptive treatment platforms and learning machines were also scrutinised, so as to include systems that change over time with repeated use, even if left to their own devices. The list is a place to start rather than a complete specification of all research that is required to close the gap; such indications can instead be found inside the reports referred to.

(i) Employ contextualised representation learning models. (ii) Maintain control of intellectual property rights of successful algorithms. (iii) Respect and try to anticipate legal changes to fair use of machine learning. (iv) Apply clinical thresholds in the strife for evidence-based self-learning apps. (v) Define and explain a concept of safe learning spaces. (vi) Assess critically the black box solutions employed within health, without asking for full transparency. (vii) Incorporate the training costs of large deep learning models into the overhead cost for their deployment, even if the model was pre-trained by somebody else. (viii) Leverage knowledge in statistical modelling to avoid incommensurability in meta-studies. (ix) Acknowledge and make visible any model bias and do not simply throw more data at a biased learning model.

Table 1 A laundry list of domains within which actions are necessary to perfect human validation of machine learning results for health data

#	Problem	Effect	Domain
1	The ground truth is not the absolute truth	Capacity to emulate humans rather than to solve is stressed	Statistics, methodology, philosophy
2	Sharing optimised ML open source code might have ethical consequences	Regulatory bodies lagging permits exploitation without full validation	Ethics, judicial, regulatory
3	Pre-registering protocols for studies involving ML are hard to publish	Standards are lagging, requiring protocol-external comments and explanations	Standardisation bodies, research community efforts, judicial
4	Apps that learn are not always evidence-based	Physicians do not put much trust into ML-powered apps	App developers, innovation managers, policy making
5	MDR classification and CE marking did not originally consider learning software	No one knows if new validation of software changed since it was certified is needed	Standardisation bodies, research community efforts, judicial, regulatory
6	ML software packages are often black boxes	Low interpretability yields low trust and zero opportunity for full validation	Research community efforts, methodology
7	Large deep models are energy-hungry	Pre-training and validation of large deep models are almost reserved for Big Tech	Regulatory, judicial, policy making
8	The quantitative performance measurements of ML models are usually incommensurable	Folklore of model adequacy is mistaken for scientific fact	Research community efforts, methodology, statistics
9	ML model validation must incorporate bias, which is hard to estimate and control for	Biased ML models are unknowingly used, with erroneous results	Research community efforts, methodology, statistics, philosophy

5 Discussion and Results

In Table 1, a set of domains are listed, within which one has to merge human validation efforts with automated procedures, and in particular take action to merge ML validation methods with those of the highly regulated world of health data. The data is sometimes referred to as primary, for care data, as compared to secondary, for research. In future care, these two categories could in theory be unified, if privacy and security issues can be successfully handled. For each domain, in which action must be taken, the problems associated with it are indicated, as are their effects. For the cases not already commented on in detail, additional explanations on the actions themselves are provided in the present section. The first problem to be detailed is that of ground truth.

Item 1 Not all ML requires labels; there are unsupervised learning algorithms, for instance, as well as reinforcement learning approaches. Labels can also be produced without human intervention, for example by masking. For text, this means training a bi-directional model on large corpora, leaving both past and future tokens masked, and then asking the model to predict the masked words (Devlin et al., 2018). The bi-directionality extends ordinary language models that only work with past tokens, making many new means to scoring model capacity possible (Salazar et al., 2019). This is a clever trick, since human labelling is costly, and much of the ground truth can be found simply by removing the masking (Liu et al., 2019b). Such models will in the near future be used for radiology too, since the contextual representation built is not limited to text, but are quickly transferring to other modalities. For images, this is not limited to segmentation either, but the list of tasks now includes detection, classification, reconstruction, synthesis, registration, clinical report generation, and more, according to a recent review, which also describes limitations of such approaches (Shamshad et al., 2022). Such developments mean that one can not focus only on the hitherto successful approaches but instead one should expect the principles of huge pre-trained deep learning models for text to generalise to many applications within the medical domain. Only by relying on such contextualised representation learning can one move naturally from emulation to self-learning machines. This is not the only innovative approach either. There are analogue computers like reservoirs that are extremely energy-efficient, as they train only on output, and yet they can solve meaningful problems in health (Tanaka et al., 2019). AI-based approaches to quantum sensors are also being rolled out in several fields of application, such as magnetoencephalography (Hari & Salmelin, 2012; Westin et al., 2020). The role of AI then becomes a translator or an interpreter of sensor results, as the output itself rarely lends itself to human validation directly.

Item 2 In view of the second problem, an example of fixing a bug in software useful for segmentation in CT scans was given, and the deliberation over sharing amended code with the community was considered, especially since other research projects had suffered from that same bug. Sharing all code, however, by extension means sharing it freely with the world; since in theory, anyone could download it. While

near-optimal automated segmentation algorithms have tremendous potential for augmenting current telemedicine solutions, especially in parts of the world where radiologists and imaging are scarce or non-existing resources, there are ethical issues associated with their use. Because such algorithms would shorten the time a human radiologist needs to spend on each patient, there is economical incentive to use them. If the algorithms are available as free open source resources, with simple installation instructions and with no harsh requirements on local hardware or infrastructure, they might for a time become ubiquitous as an unregulated resource. When regulation procedures catch up, the arbitrage disappears, so the ethical problem is a temporary one. But the trade-off between the benefits to patients of regulated and certified use and the risks associated with unregulated and uncertified use needs to be taken into account at this time, even if only a minute share of radiologists would exploit this window of opportunity. Apart from research efforts, there are obvious counter-actions to consider, like keeping the intellectual property reserved for internal use until the ethical problem has been solved. There is also sandboxing: providing trusted research environments in which academia and commercial developers can test algorithm and product safety at lower technology readiness levels.

Item 3 The third problem is likely to be hard, and perhaps standards are lagging because standardisation bodies are trying to future-proof current protocols, like TRIPOD mentioned above. More recommendations are expected from the European Union's long-awaited AI act to be adopted in 2023, and all the while the legal world is dealing with Schrems III, possibly making cloud computing and analytics an impossibility. With so much of health data analytics being cloud-based, making such computing local would hamper development and technology readiness both within health. Generative AI algorithms also write code these days, leading to further possible future tension in protocols. What is fair use and how correct is it that copyrighted code is used for training automated code-generators are among the burning questions here (Caballar, 2022). Again, sandboxes within which developers can test and experiment new innovations under regulatory advice could help solve the problem.

Item 4 While smart AI apps per organ make sense to many a radiologist, imagine a primary care physician who meets a patient that says an app brought them there. In the rare case of full approval for such an app, the physician would have access to the back office of the app software, while the user would see data only through the front office: the app user interface. This diminishes the problem of opacity in how the app processes physiological data, for instance. The physician can then look at measures of things like blood pressure and will most likely abstain from trying to understand why the person was told to seek primary care by the app. This data is then found in the sense that the physician did not order any test or measurement, but can still be used almost as if purposefully gathered to address concerns the physician might have. The far more common case is that the physician has not heard of that particular app before and just asks the person the same type of questions as they would normally ask, disregarding any evidence from the app. If the app is branded as smart or learning over time, the result might be that the physician lumps this

together with the other properties of the app, good or bad. If the physician does find the app interesting, the question arises about what it has actually found out about its user, and possibly how. Is a well-established risk score calculation being made, for example? Are the thresholds used the standard clinical thresholds, or might the app overdiagnose due to lower thresholds? Health apps come with extra restrictions for most infrastructural platforms, having to be manually whitelisted in some cases, and in other cases, the data they save can only be stored on the encrypted part of the device's memory, etc. While those restrictions often make sense, they do not address the fourth problem as such. In particular, there are problems with two-sided (patient and physician) explications of the model, of trust in and trustworthiness of the learning processes in the app, and of patient-physician power asymmetries.

Item 5 Even if they remain a rarity, learning CE-marked platforms that adapt to the person in treatment in accordance with ML algorithms do exist, e.g. for Internet-based cognitive behavioural therapy (Boman et al., 2019). The question naturally arises on whether or not human psychologists using such a platform are nudged toward different actions than they would turn toward without the ML-generated suggestions. Over time, learning could change the platform so much from the one that was once CE-marked that it would need to be re-assessed. On the other hand, most of the adaptive treatment platforms are not likely to change significantly, but the learning takes places within what constitutes a *safe learning space*. A certificate that explains why and how the model adapts, but only in a way that does not change its basic functionality, could save many research and development projects from repeated human validation (Minkkinen et al., 2022).

Item 6 Interpretability and transparency are two terms that would seem to indicate that trust in an apparatus, process or algorithm increases. But the people that most vigorously call for explainable AI, arguably an oxymoron, are from fields where very expensive machinery is operated and where any interference with operations is costly. Hence, trust does not evolve from understanding why deep learning works, or from asking deep learning models to explain their recommendations iteratively, but from situated sensemaking (Boman & Sanches, 2015). Deep learning filters are today a part of many scanners producing medical images, and they do improve image quality, leading to better care, in many cases (Yu et al., 2021), even if there are problems still waiting to be solved (Varoquaux & Cheplygina, 2022). Human validation should thus focus on usefulness and benefits of machines and their associated software and interfaces, never forgetting to maintain a critical perspective.

Item 7 The very first item on the list recommended pre-trained large language models and the like, in spite of these billion-parameter models being extremely costly to pre-train. The commercial development and the academic research community have both sought to address this by making pre-trained structures available at low cost. Easy to access portals collecting such structures are today in wide use, especially if compared to only 5 years back. Even if the natural language processing

(NLP) community has always been good at benchmarking software for continuous validation, the community members have found themselves near obsessed with huge models recently, especially since their use have now left computational linguistics to move into more general use. Very few research and development bodies have the data and the associated infrastructure to do the pre-training, and they have lately revealed their energy costs. This has in turn led to movements like *Green NLP* and increased concern for what transfer learning is possible when reusing a structure originally trained for another purpose (Maronikolakis & Schütze, 2021): a situation comparable to that of the found vs. purposefully gathered data distinction. Today, energy use is part of the overhead cost of using deep learning—or at least it should be—making that factor part of the equation of overall usefulness to health goals.

Item 8 Two models are incommensurable if they share no common measure. There are published reviews of which kinds of ML model have performed best in health. These reviews sport tables of quantitative performance data, as reported in other studies, even conveniently (sic) averaging the results from those studies in a table column. Such reviews have no value and might be misleading by promoting non-optimal families of models for the clinical problem at hand. There is no replacing of philosophy of science truths about what constitutes good science, and any validation at the meta-level too must adhere to those truths. Providing references to such poor reviews here would be contributing to the problem of their propagation, which largely hinges on the numbers of citations, so it is merely concluded here that knowledge on statistical modelling—a notoriously difficult task—is always needed (Breiman, 2001).

Item 9 Last but not least, bias problems must be considered. Popular depictions of what bias in ML can lead to abound. The ubiquity of such reports notwithstanding, one has to deal with found data for many health applications, and increasingly, this found data is used to pre-train deep models, with biased results. The efforts required include to make bias visible (Boman et al., 2020), be aware that it affects also the largest language models (Katsarou et al., 2022), and finally note that some bias problems occur because of a common pitfall: researchers only threw more data at the problem.

6 Conclusion

In Table 1, findings on how human validation of artificial intelligence and machine learning efforts play out within the health domain were listed. Which of the problems identified remain open and thus merit further research was also explained, detailing within which domain. These findings hence provide a starting point for studies, rather than conclusions about efforts already made. This helps

by making solvable the wicked problem of how to merge and validate humans and artificial intelligence for health in the best possible way.

Acknowledgements I thank the main stakeholder of the AI@KI project, the present as well as the former President of Karolinska Institutet, for the time afforded me. Jakob Mökander and Emmanuel Zavalis both provided extensive comments on the near-finished manuscript. Gabriel Westman co-created the concept of Safe Learning Space. Carl Johan Sundberg, Sabine Koch and Peter Sjögårde contributed relevant details on health informatics. Vitali Grozman contributed many insights and also provided important domain knowledge in radiology. Sandra Eloranta and Fredrik Strand provided useful comments, as did the reviewers of the journal. Last but not least, I thank Fehmi Ben Abdesslem for important technical assistance and advice.

Funding Open access funding provided by Royal Institute of Technology. This study was funded by the President of Karolinska Institutet.

Data Availability The generated during and/or analysed during the current study are available in the AI@KI project repository, https://ki.se/en/lime/final-report.

Declarations

Informed Consent Not applicable.

Materials and Method Method is described as Section 4 in the manuscript. No LLMs were employed in the writing.

Competing Interests The author declares no competing interests.

References

Adamson, A. S., & Welch, H. G. (2019). Machine learning and the cancer-diagnosis problem-no gold standard. *The New England Journal of Medicine, 381*(24), 2285–2287.

Anderson, A. W., Marinovich, M. L., Houssami, N., Lowry, K. P., Elmore, J. G., Buist, D. S., Hofvind, S., ... & Lee, C. I. (2022). Independent external validation of artificial intelligence algorithms for automated interpretation of screening mammography: A systematic review. *Journal of the American College of Radiology*.

Axtell, R., Axelrod, R., Epstein, J. M., & Cohen, M. D. (1996). Aligning simulation models: A case study and results. *Computational & Mathematical Organization Theory,1*, 123–141.

Belue, M. J., Harmon, S. A., Lay, N. S., Daryanani, A., Phelps, T. E., Choyke, P. L., & Turkbey, B. (2022). The low rate of adherence to checklist for artificial intelligence in medical imaging criteria among published prostate MRI artificial intelligence algorithms. *Journal of the American College of Radiology*,.

Bera, K., Braman, N., Gupta, A., Velcheti, V., & Madabhushi, A. (2022). Predicting cancer outcomes with radiomics and artificial intelligence in radiology. *Nature Reviews Clinical Oncology, 19*, 132–146.

Boman, M. (2022). *AI@KI: Final report*. Published on January 27, 2022, from https://ki.se/en/lime/final-report

Boman, M., Ben Abdesslem, F., Forsell, E., Gillblad, D., Görnerup, O., Isacsson, N., Sahlgren, M., & Kaldo, V. (2019). Learning machines in internet-delivered psychological treatment. *Progress in Artificial Intelligence, 8*, 475–485.

Boman, M., Downs, J., Karali, A., & Pawlby, S. (2020). Toward learning machines at a mother and baby unit. *Frontiers in Psychology, 11*, 567310.

Boman, M., Koch, S., & Sjögårde, P. (2022). Scientometric search terms. *Appendix 2 to AI$@$KI: Final report*. Published on January 27, 2022, from https://ki.se/en/lime/final-report

Boman, M., & Sanches, P. (2015). Sensemaking in intelligent health data analytics. *KI-Künstliche Intelligenz, 29*, 143–152.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science, 16*(3), 199–231.

Caballar, R. D. (2022). Ownership of AI-generated code hotly disputed. *IEEE Spectrum*. Retrieved from https://spectrum.ieee.org/ai-code-generationownership

Carbonell, M. F., Boman, M., & Laukka, P. (2021). Comparing supervised and unsupervised approaches to multimodal emotion recognition. *PeerJ Computer Science,7*, e804.

Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P., & Willmes, K. (2015). Registered reports: Realigning incentives in scientific publishing. *Cortex, 66*, A1–A2.

Cossío, F., Schurz, H., Engström, M., Barck-Holst, C., Tsirikoglou, A., Lundström, C., Gustafsson, H., Smith, K., Zackrisson, S., & Strand, F. (2023). VAI-B: A multicenter platform for the external validation of artificial intelligence algorithms in breast imaging. *Journal of Medical Imaging,10*, 061404. Retrieved from https://doi.org/10.1117/1.JMI.10.6.061404

Dembrower, K., Lindholm, P., & Strand, F. (2020). A multi-million mammography image dataset and population-based screening cohort for the training and evaluation of deep neural networks-The cohort of screen-aged women (CSAW). *Journal of Digital Imaging,33*(2), 408–413.

Dembrower, K., Wåhlin, E., Liu, Y., Salim, M., Smith, K., Lindholm, P., Eklund, M., & Strand, F. (2020). Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: A retrospective simulation study. *The Lancet Digital Health,2*(9), e468–e474.

Devlin, J., Chang, M. -W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Eloranta, S., & Boman, M. (2022). Predictive models for clinical decision making: Deep dives in practical machine learning. *Journal of Internal Medicine,262*(2), 278–295.

Faes, L., Liu, X., Wagner, S. K., Fu, D. J., Balaskas, K., Sim, D. A., Bachmann, L. M., Keane, P. A., & Denniston, A. K. (2020). A clinician's guide to artificial intelligence: How to critically appraise machine learning studies. *Translational Vision Science & Technology,9*(2), 7–7.

Ferrario, A., Loi, M., & Viganò, E. (2020). In AI we trust incrementally: A multi-layer model of trust to analyze human-artificial intelligence interactions. *Philosophy & Technology, 33*, 523–539.

Freer, T. W., & Ulissey, M. J. (2001). Screening mammography with computer-aided detection: Prospective study of 12,860 patients in a community breast center. *Radiology,220*(3), 781–786.

Hari, R., & Salmelin, R. (2012). Magnetoencephalography: From squids to neuroscience: Neuroimage 20th anniversary special edition. *Neuroimage*, *61*(2), 386–396.

Hendrix, N., Lowry, K. P., Elmore, J. G., Lotter, W., Sorensen, G., Hsu, W., Liao, G. J., Parsian, S., Kolb, S., Naeim, A., & Lee, C. I. (2022). Radiologist preferences for artificial intelligence-based decision support during screening mammography interpretation. *Journal of the American College of Radiology,19*(10), 1098–1110.

Ioannidis, J. P. (2022). Pre-registration of mathematical models. *Mathematical Biosciences, 345*, 108782. Elsevier.

Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature methods,18*(2), 203–211.

Kadir, T., & Gleeson, F. (2018). Lung cancer prediction using machine learning and advanced imaging techniques. *Translational Lung Cancer Research,7*(3), 304.

Katsarou, S., Rodríguez-Gálvez, B., & Shanahan, J. (2022). Measuring gender bias in contextualized embeddings. In *Computer Sciences and Mathematics Forum* (vol. 3, p. 3). MDPI.

Klau, S., Hoffmann, S., Patel, C. J., Ioannidis, J. P., & Boulesteix, A. L. (2021). Examining the robustness of observational associations to model, measurement and sampling uncertainty with the vibration of effects framework. *International Journal of Epidemiology, 50*(1), 266–278. Oxford University Press.

Kooli, C., & Al Muftah, H. (2022). Artificial intelligence in healthcare: A comprehensive review of its ethical concerns. *Technological Sustainability*.

Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdas, M., Kern, C., Ledsam, J. R., Schmid, M. K., Balaskas, K., Topol, E. J., Bachmann, L. M., Keane, P. A., & Denniston, A. K. (2019a). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. *The Lancet Digital Health,1*(6), e271–e297.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019b). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Maronikolakis, A., & Schütze, H. (2021). Multidomain pretrained language models for green NLP. In *Proceedings of the Second Workshop on Domain Adaptation for NLP* (pp. 1–8).

Masnick, M. (2012). Why Netflix never implemented the algorithm that won the Netflix \$1 million challenge. *TechDirt*. Retrieved from https://www.techdirt.com/2012/04/13/why-netflix-never-implemented-algorithm-that-won-netflix-1-million-challenge/

Minkkinen, M., Laine, J., & Mäntymäki, M. (2022). Continuous auditing of artificial intelligence: A conceptualization and assessment of tools and frameworks. *Digital Society,1*(3), 21.

Moons, K. G., Altman, D. G., Reitsma, J. B., Ioannidis, J. P., Macaskill, P., Steyerberg, E. W., Vickers, A. J., Ransohoff, D. F., & Collins, G. S. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Annals of Internal Medicine,162*(1), W1–W73.

Nagendran, M., Chen, Y., Lovejoy, C. A., Gordon, A. C., Komorowski, M., Harvey, H., Topol, E. J., Ioannidis, J. P., Collins, G. S., & Maruthappu, M. (2020). Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies. *British Medical Journal*, *368*.

Park, S. H., & Han, K. (2018). Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology,286*(3), 800–809.

Pulini, A. A., Kerr, W. T., Loo, S. K., & Lenartowicz, A. (2019). Classification accuracy of neuroimaging biomarkers in attention-deficit/hyperactivity disorder: Effects of sample size and circular analysis. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging,4*(2), 108–120.

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A. I., Etmann, C., McCague, C., Beer, L., Weir-McCall, J. R., Teng, E., & Zhongzhao and Gkrania-Klotsas, AIX-COVNET and Rudd, J. H. F., Sala, E., & Carola-Bibiane, S. (2021). 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.

Sachs, M. C., Sjölander, A., & Gabriel, E. E. (2020). Aim for clinical utility, not just predictive accuracy. *Epidemiology (Cambridge, Mass.)*, *31*(3), 359.

Salazar, J., Liang, D., Nguyen, T. Q., & Kirchhoff, K. (2019). Masked language model scoring. arXiv preprint arXiv:1910.14659.

Shamshad, F., Khan, S., Zamir, S. W., Khan, M. H., Hayat, M., Khan, F. S., & Fu, H. (2022). Transformers in medical imaging: A survey. arXiv preprint arXiv:2201.09873.

Strand, F., Patel, B. K., & Allen, B. (2021). A call for controlled validation data sets: Promoting the safe introduction of artificial intelligence in breast imaging. *Journal of the American College of Radiology,18*(11), 1564–1565.

Tanaka, G., Yamane, T., Héroux, J. B., Nakane, R., Kanazawa, N., Takeda, S., Numata, H., Nakano, D., & Hirose, A. (2019). Recent advances in physical reservoir computing: A review. *Neural Networks, 115*, 100–123.

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine,25*(1), 44–56.

Varoquaux, G., & Cheplygina, V. (2022). Machine learning for medical imaging: Methodological failures and recommendations for the future. *NPJ Digital Medicine,5*(1), 48.

Verma, S., Dickerson, J., & Hines, K. (2021). Counterfactual explanations for machine learning: Challenges revisited. arXiv preprint arXiv:2106.07756.

Wei, P., Lu, Z., & Song, J. (2015). Variable importance analysis: A comprehensive review. *Reliability Engineering & System Safety, 142*, 399–432.

Westin, K., Pfeiffer, C., Andersen, L. M., Ruffieux, S., Cooray, G., Kalaboukhov, A., Winkler, D., Ingvar, M., Schneiderman, J., & Lundqvist, D. (2020). Detection of interictal epileptiform discharges: A comparison of on-scalp MEG and conventional meg measurements. *Clinical Neurophysiology,131*(8), 1711–1720.

Yu, H., Yang, L. T., Zhang, Q., Armstrong, D., & Deen, M. J. (2021). Convolutional neural networks for medical image analysis: State-of-the-art, comparisons, improvement and perspectives. *Neurocomputing, 444*, 92–110.

Zendel, O., Murschitz, M., Humenberger, M., & Herzner, W. (2017). How good is my test data? Introducing safety analysis for computer vision. *International Journal of Computer Vision, 125*, 95–109.