Deploying clinical machine learning? Consider the following...

Charles Lu,[1] **Ken Chang,**[2,3] **Praveer Singh,**[2] **Stuart Pomerantz,**[1] **Sean Doyle,**[1] **Sujay Kakarmath,**[1]
Christopher Bridge,[1,2] **Jayashree Kalpathy-Cramer**[1,2]

[1] Massachusetts General Hospital, Boston, MA
[2] Harvard Medical School, Boston, MA
[3] Massachusetts Institute of Technology, Cambridge, MA

Abstract

Despite the intense attention and considerable investment into clinical machine learning research, relatively few applications have been deployed at a large-scale in a real-world clinical environment. While research is important in advancing the state-of-the-art, translation is equally important in bringing these techniques and technologies into a position to ultimately impact healthcare. We believe a lack of appreciation for several considerations are a major cause for this discrepancy between expectation and reality. To better characterize a holistic perspective among researchers and practitioners, we survey several practitioners with commercial experience in developing CML for clinical deployment. Using these insights, we identify several main categories of challenges in order to better design and develop clinical machine learning applications.

Introduction

In the wake of the COVID-19 pandemic, hundreds of AI papers were published but few were clinically practical or useful (Born et al. 2021; Roberts et al. 2021). Even large technology companies have recently encountered unexpected difficulties in trying to productionize state-of-the-art research into real clinical products for healthcare applications (Vincent 2021; Strickland 2019). We believe there is a disconnect between the clinical machine learning (CML) research community and delivered clinical impact from the translation of useful CML tools (Wolff et al. 2021). While there have been extensive claims of AI performance being equivalent or superior to human doctors within the research literature (Liu et al. 2019), emerging clinical studies reveal contradictions in those claims of purported superhuman performances (Freeman et al. 2021; Zech et al. 2018; Voter et al. 2021).

Machine learning (ML) in the healthcare domain faces numerous challenges, such as immense difficulty in acquiring and annotating large amounts of medical data, recruiting the necessary clinical expertise in validating models, and integration into existing clinical workflows and infrastructure (Kelly et al. 2019; Stead 2018). As such, developing CML software is considered especially challenging compared to traditional software systems (Yang et al. 2020;

Role	Years of experience	Considerations
Machine learning scientist	5	2.1, 3.2, 4.1, 4.2
Machine learning scientist	10	1.2, 2.2, 3.1, 3.3
Software engineer	30	2.4, 3.2, 3.3, 4.1
Clinical project manager	5	1.1, 1.3, 2.1, 4.2
Radiologist	25	1.3, 2.3, 4.3
Neurologist	10	1.3, 2.3
Clinical researcher	5	2.1, 4.2

Table 1: Role and (approximate) experience level of surveyed participants and their (general) contributions to specific considerations.

Lu et al. 2020; He et al. 2019). Furthermore, sub-fields of healthcare such as oncology and dentistry have even more nuanced challenges to adopting AI to their respective specialties (Bi et al. 2019; Schwendicke, Samek, and Krois 2020). Others have attempted to operationalize these challenges into a formal specification; Oala et al. (2020) apply the ITU/WHO FG-AI4H framework to audit several case studies of AI applications for diagnostic retinopathy, Alzheimer's diagnosis, and cytomorphologic classification for leukemia.

To gain a more holistic understanding of the challenges in CML translation, we survey researchers and clinicians and distill their insights into better practices. The contributions of our survey complement other works such as the perspective by Wiens et al. (2019) by providing empirical insights grounded in actual interviews from working CML practitioners. We hope that by enumerating these considerations explicitly, we can promote and drive more targeted CML research that is better situated into clinical context and facilitate the translation of research into real clinical benefit.

Considerations

First, we detail the survey methodology. We conducted several semi-structured interviews with clinical and technical practitioners from the Center of Clinical Data Science at the Massachusetts General Hospital and Brigham and Women's Hospital. Participants were chosen based on their prior experience in developing and deploying CML products in collaboration with industry partners (such as GE Healthcare and Nvidia) for applications in stroke, emergency medicine, and radiology. In each session, we asked several open-ended

questions to identify reoccurring themes and asked probing followup questions tailored to the background and expertise of each participant. Examples questions included:

> What are the major challenges you encountered when deploying clinical machine learning?

and

> If you had to start this project over, what would you have done differently?

We analyzed the responses, made broad thematic categories, and synthesized specific considerations that would be generally applicable to most CML contexts. Table 1 summarizes the roles, experience, and contributions from the interviewed participants. While the considerations collected should not be taken as an exhaustive list, we believe they can still serve as a basic and approachable checklist of real-world considerations for most ML engineers and researchers working towards clinical application. We group these considerations into four broad areas: clinical context, clinical validation, deployment, and monitoring.

1) Clinical context

The clinical context is often overlooked in CML research in favor of demonstrating high performance or technical novelty on benchmark datasets (Yan et al. 2018; Codella et al. 2019; Abràmoff et al. 2016). Commercial CML systems designed for clinical practice also meet unexpected challenges and friction in deployment and integration (Strickland 2019; Wang et al. 2021). For example, Wang et al. (2021) report the tensions that arose between the AI clinical decision support system and the reality of clinical practice in rural China.

1.1 Identify pain points We contend that CML research that primarily focuses on automated prediction for critical diseases may often miss lower-hanging fruit with greater potential utility to the clinical workflow. Specifically, tools that are most likely to be appreciated are tools that improve the quality of life for the clinician (e.g. by lowering administrative burden in routine clinical practice). In radiology, auto-generated report templates and speech dictation are conveniences that have been widely adopted by radiologists because they decrease friction and interpretation time in the overall workflow.

Some of the best product ideas come from clinicians and vendors of clinical software. Clinicians are well-positioned to estimate the potential value generated from CML tools that directly address experienced pain points. Industry vendors will have already assessed market value and feasibility of software integration costs into existing clinical infrastructure. Focus should be on developing tools for a particular stage or aspect of the clinical workflow (such as triage, screening, or diagnosis), the associated downstream risks, and how they change or affect the existing care pathway. As such, close collaboration between researchers, clinicians, and industry vendors is often a wise route in CML research and development.

1.2 Assess feasibility Many projects often underestimate the difficulty of realistic validation and deployment of CML models. CML applications are uniquely challenging due to the need of clinical expertise to define cohort selection, annotate data, and evaluate model performance. Lack of easily accessible and available data resources is a major issue in CML. In addition to maintaining patient privacy, annotation of large amounts of sensitive medical data can be extremely difficult and more complicated than other kinds of data. Inter-annotator and intra-annotator variability is a known issue and disagreements often arise due to differences in interpretation, training, and expertise of the annotator. Consensus labeling requires multiple annotators, which places additional overhead and cost constraints.

Clinical evaluation can also be subjective due to lack of established evaluation standards. In such scenarios, performance metrics would need to be defined specifically for the clinical application. Furthermore, this assessment criteria may evolve over time in accordance to changing clinical guidelines or protocols. The exclusion criteria may need to be continuously refined as outliers are uncovered. We stress that CML projects are an extremely iterative development process that require constant feedback between clinical and technical teams.

1.3 Promote clinical champions All CML projects should have a dedicated clinical champion, preferably a practicing specialist in the area of the specific application. The success of a project (even progressing past the initial development to the deployment testing) often depends on finding "clinical use-case fit". A clinical champion will be able to continuously reaffirm the clinical utility and guide early development and provide a path for later deployment. Ideally, a clinical champion will be an insider that is able to promote the project within complex social hierarchies within a hospital that would be impenetrable to outsiders. Having their dedicated buy-in often allows for the prioritization of the project within a myriad of competing interests and help in sourcing users to test CML prototypes. Especially for critical decision-making, trust is a major barrier to adoption (Jacobs et al. 2021), and a clinical champion can help in defining what trust means to their colleagues, as well as assist in educating and training users of CML tools.

2) Clinical validation

Rigorous evaluation of many CML papers according to clinical standards is often lacking (Nagendran et al. 2020). Historically, the ML research community has often focused on achieving state-of-the-art performance while the medical community focused on demonstrating utility in broad clinical contexts and trials. Some of these differences can be resolved by using multi-site evaluation, choosing more clinically meaning performance metrics, and performing user studies on end-to-end systems that simulate real-world clinical environments.

2.1 Multi-site evaluation Generalizability across institutions is one of the biggest challenges to developing robust CML. A model might go from having excellent performance at one site to having considerably worse performance at another seemingly similar institution (Chen et al. 2021; Chang et al. 2020). This generalization gap can arise from factors

such as differences in patient population, disease characteristics, or data acquisition settings (Gupta et al. 2021). This is why multi-site testing is crucial in order to truly assess and validate CML models. However, a study by Wu et al. (2021) analyzed 130 FDA approvals of AI medical devices and found that 93 (72%) did not publicly report evaluation on a multi-site evaluation set.

A metastudy by Freeman et al. (2021) finds that the evaluation procedures in many papers on CML for breast cancer screening are vulnerable to numerous biases in patient selection, ground truth reference standard, and human comparator evaluation – ultimately, that inflating claims about CML being "better than radiologists" in detecting breast cancer. In fact, 34 out of 36 (94%) CML systems were determined to be less accurate than a single radiologist, while all 36 were less accurate than the consensus of two or more radiologists. We advocate testing CML on heterogeneous, multi-site datasets. Ideally, the gold standard evaluation should be a prospective study used in clinical trials such as a multi-case multi-reader design with statistically meaningful endpoints (Dendumrongsup et al. 2014).

2.2 Choose meaningful metrics While many performance metrics overlap between ML and CML, traditional ML do not fully take into account potential clinical benefit or demonstrated impact when deployed into real-world clinical workflows. A CML system to triage stroke patients might be evaluated for high sensitivity (true positive rate) as well as time advantage over standard-of-care, while a CML application to screen breast cancer risk might consider specificity (true negative rate) as well as callback rate.

Meaningful evaluation metrics should align with the intended clinical application use-case. For example, the goal of a CML application to detect sepsis is not merely absolute rate of sepsis detection, but to identify those cases of sepsis in which the clinician missed it; additionally, the CML must be able to detect sepsis early enough for there to be time for some clinically actionable intervention to impact care (Goh et al. 2021).

Also, despite the theoretical and pedagogical emphasis on binary classification, most tasks in medicine are not simple binary decisions but nuanced gradations that take into account the clinical context and patient history. These grey-zones of ambiguity and ambivalence should be recognized and deliberately acknowledged when choosing to operationalize the problem into a more amendable form for CML.

2.3 Perform user studies CML papers often ignore human factors and other social components in their evaluation. If the goal is to develop more useful CML systems for clinicians, more attention should be placed on the interaction between human and CML applications (Amershi et al. 2019). User studies with appropriate domain experts should participate in the evaluation of human-centered tools that empower clinicians (Cai et al. 2019; Kiani et al. 2020; Gaube et al. 2021).

User studies can also identify and alleviate points of friction that hamper successful adoption of CML applications. The more rigorous the user study, the more assurance can be had that the CML system will be used as intended (or at least that off-label uses can be discovered early). Performing realistic user studies can also uncover small changes that result in substantial improvements to the user experience. For example, instead of outputting out uncalibrated probabilities from a black box model, semantically meaningful outputs can often be designed that better map onto user intuition (e.g. displaying "mild", "moderate", or "severe" instead of 73.2% to reduce number fatigue and cognitive overhead). Clear visualizations and thoughtful interface design can bring greater transparency that assist medical decision-making and facilitate trust in a algorithmic system (Xie et al. 2020).

2.4 Embed situated workflows Integration of CML can also affect the flow of information. For example, should the results of radiology CML be made available to the referring clinician? What if there are multiple algorithm outputs, each of which may not be wholly consistent with the report? Which parts of the CML output belong in the patient's medical record? There should be a way for the clinician to curate algorithmic results into clinically useful outputs.

3) Deployment

Deployment in CML to often has to follow stringent integration requirements with existing clinical infrastructure, which often supports other functions beyond that specific AI tool. Practical implementation challenges of deploying into hospitals often necessitate considerable coordination between internal and external teams to setup orchestration and inference platforms (usually with interaction between third-party vendors of clinical software).

3.1 Development To Deploy Decisions made during model development can have significant knock-on effects that ultimately affect the ease with which the model can be deployed. Thus, many pain points in the model deployment process can be mitigated by developing the model with deployment in mind from the outset. Relying on complex or slow data processing routines can significantly complicate deployment efforts. Furthermore, choosing a modular code-base structure more easily allows for re-use of code between training and deployment stages.

3.2 Choose an integration layer Difficulty of integration greatly depends on where in the workflow the CML application will be injected. For example, an application to detect motion artifacts might be deployed into the MRI scanner; while another application to triage stroke patients may be added directly into the clinical work-list software; while a critical findings notification application for stroke may be integrated into a mobile device environment. Additionally, different inference contexts will have different requirements for the application. An application deployed in the emergency room may require a faster inference time than that of an application that runs in the background on routine cases.

3.3 Embrace healthcare standards The lack of support for standards often increases integration costs. Hospitals are often unaware of the importance of standards and fail to specify newer standards in their Request for Proposals (RFPs) when purchasing products. Embracing existing

healthcare standards can help avoid proprietary APIs, which can lock users and hospitals into workflows which may be sub-optimal. The Integrating the Healthcare Enterprise Initiative (IHE) has guidelines for integrating imaging models into clinical PACS (IHE Radiology Technical Committee 2020) through the use of existing standards such as DICOM Structured Reporting (Clunie 2000), DICOM Segmentation, and HL7. Unfortunately, support for these standards is not yet widespread in existing enterprise software. We contend that support for standards should be a key criterion used by hospitals in their RFP process when purchasing systems in order to be as prepared as possible for future CML deployments.

4) Monitoring

Even after being deployed into clinical practice, CML models must be continuously monitored to record and audit predictions for safety (especially the side-effect of software changes and model updates). CML applications must take into account trace-ability and consider how the system will be equipped to handle changing data streams while keeping track of predictions in a controlled and documented procedure.

4.1 Design for traceability Who bears liability for when CML (inevitably) makes an error and causes harm? Rather than conceal failures and shirk from reporting, we believe that the long-term interest for all parties lies in responsible disclosures (already widely practiced by computer security researchers) and review (similar to mortality conferences practiced by surgeons) (Kravet, Howell, and Wright 2006). Gradual accumulation of incident reports should alert manufacturers to failure modes, which might be more prone in certain risky subgroups or edge cases (Masic, Miokovic, and Muhamedagić 2008). Tracking CML predictions will also help to verify the intended-use (as opposed to off-label use) and inform the development of proper mitigation strategies, such as user training or re-calibration of prediction thresholds.

Patterns of traceability can be very subtle. If an algorithm has been part of workflow for a long period of time and has known failure modes – when a new version of the algorithm is installed, it might be prudent to run both versions for a while and alert users if predictions on the new algorithm differ substantially from the old algorithm. If the CML is from a different vendor, this task may be made more difficult if the CML output is not in a standardized format.

4.2 Beware of bias Any ML model is susceptible to encoding and exacerbating existing disparities and biases from the data (during collection, selection, annotation, etc.) to algorithm development (design choices, training parameters, testing evaluation) to the framing of the application itself (purpose, reification, intended vs. actual usage) (Barocas, Hardt, and Narayanan 2019; Green and Hu 2018; Jacobs and Wallach 2019; Jr. et al. 2020). As one recent study demonstrated on multiple medical datasets, CML models can discriminate patient race even in extremely down-sampled chest radiographs as small as 4x4 pixels, that would be unrecognizable to any human (Banerjee et al. 2021). Other

studies have shown that algorithms can have differential performance across subgroups, raising concerns surrounding equity and fairness (Larrazabal et al. 2020; Pierson et al. 2021).

One clinician expressed the need for a platform with the ability to audit CML model for algorithmic bias across patient demographics, such as age and race, as well as acquisition metadata, such as time and version of model prediction (Lu et al. 2021a,b). Dashboards that interactively display different cohorts would allow clinicians more insight into how CML models perform in real-time CML monitoring should flag any trends among sensitive subgroups, which, depending on the context, may be clinically warranted or not.

4.3 Expect distribution shift We found consensus with other researchers, for the need to consider non-stationarity in medical data (Ghassemi et al. 2018). Populations, technology, and human behavior change over time. Software will be modified, equipment will be replaced, and clinical guidance will be updated (Ovadia et al. 2019; Finlayson et al. 2020). Lazer et al. (2014) studied concept drift in Google searches over a period that systematically overestimated flu incidence based on user's previous search behavior; predictive performance was observed to have deteriorated over time as their ML model become stale and outdated.

Methods of detecting and characterizing distribution shift in clinical contexts should be a crucial area of continued research for CML deployment (Rabanser, Günnemann, and Lipton 2019; Zhang et al. 2021). Interesting dynamics will emerge as CML systems begin to affect clinical decision-making, which in turn affect patient outcomes, whcih in turn feedback into population data that will be used in future CML systems (Pianykh et al. 2020).

4.4 Missing outputs Once CML has become an established part of a clinical workflow, then a clinician may begin to rely on its presence and expect some default behaviors. If the algorithm fails for whatever reason (e.g. network issue, data issue, algorithm issue), then a clinician may wait for the results believing that the CML output is in the queue. Currently, there is no standard mechanism to report and handle the failure in which the prediction result will be unavailable, incomplete, or missing.

Conclusion

We detail a number of important considerations for the deployment of CML systems. We hope these consideration promotes and drives more targeted research to addresses to better situate CML into clinical situations and facilitates more translation of promising CML research into clinical practice to effect higher standards of healthcare.

References

Abràmoff, M. D.; Lou, Y.; Erginay, A.; Clarida, W.; Amelon, R.; Folk, J. C.; and Niemeijer, M. 2016. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Investigative ophthalmology & visual science*, 57(13): 5200–5206.

Amershi, S.; Inkpen, K.; Teevan, J.; Kikin-Gil, R.; Horvitz, E.; Weld, D.; Vorvoreanu, M.; Fourney, A.; Nushi, B.; Collisson, P.; Suh, J.; Iqbal, S.; and Bennett, P. 2019. Guidelines for Human-AI Interaction. 1–13. ISBN 978-1-4503-5970-2.

Banerjee, I.; Bhimireddy, A. R.; Burns, J. L.; Celi, L. A.; Chen, L.; Correa, R.; Dullerud, N.; Ghassemi, M.; Huang, S.; Kuo, P.; Lungren, M. P.; Palmer, L. J.; Price, B. J.; Purkayastha, S.; Pyrros, A.; Oakden-Rayner, L.; Okechukwu, C.; Seyyed-Kalantari, L.; Trivedi, H.; Wang, R.; Zaiman, Z.; Zhang, H.; and Gichoya, J. W. 2021. Reading Race: AI Recognises Patient's Racial Identity In Medical Images. *CoRR*, abs/2107.10356.

Barocas, S.; Hardt, M.; and Narayanan, A. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

Bi, W. L.; Hosny, A.; Schabath, M.; Giger, M.; Birkbak, N.; Mehrtash, A.; Allison, T.; Arnaout, O.; Abbosh, C.; Dunn, I.; Mak, R.; Tamimi, R.; Tempany, C.; Swanton, C.; Hoffmann, U.; Schwartz, L.; Gillies, R.; Huang, R.; and Aerts, H. 2019. Artificial intelligence in cancer imaging: Clinical challenges and applications. *CA: A Cancer Journal for Clinicians*, 69.

Born, J.; Beymer, D.; Rajan, D.; Coy, A.; Mukherjee, V. V.; Manica, M.; Prasanna, P.; Ballah, D.; Guindy, M.; Shaham, D.; et al. 2021. On the role of artificial intelligence in medical imaging of covid-19. *Patterns*.

Cai, C. J.; Reif, E.; Hegde, N.; Hipp, J. D.; Kim, B.; Smilkov, D.; Wattenberg, M.; Viégas, F. B.; Corrado, G. S.; Stumpe, M. C.; and Terry, M. 2019. Human-Centered Tools for Coping with Imperfect Algorithms during Medical Decision-Making. *CoRR*, abs/1902.02960.

Chang, K.; Beers, A. L.; Brink, L.; Patel, J. B.; Singh, P.; Arun, N. T.; Hoebel, K. V.; Gaw, N.; Shah, M.; Pisano, E. D.; et al. 2020. Multi-institutional assessment and crowdsourcing evaluation of deep learning for automated classification of breast density. *Journal of the American College of Radiology*, 17(12): 1653–1662.

Chen, J. S.; Coyner, A. S.; Ostmo, S.; Sonmez, K.; Bajimaya, S.; Pradhan, E.; Valikodath, N.; Cole, E. D.; Al-Khaled, T.; Chan, R. P.; et al. 2021. Deep Learning for the Diagnosis of Stage in Retinopathy of Prematurity: Accuracy and Generalizability across Populations and Cameras. *Ophthalmology Retina*.

Clunie, D. A. 2000. *DICOM Structured Reporting*. PixelMed Publishing.

Codella, N.; Rotemberg, V.; Tschandl, P.; Celebi, M. E.; Dusza, S.; Gutman, D.; Helba, B.; Kalloo, A.; Liopyris, K.; Marchetti, M.; et al. 2019. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*.

Dendumrongsup, T.; Plumb, A.; Halligan, S.; Fanshawe, T.; Altman, D.; and Mallett, S. 2014. Multi-Reader Multi-Case Studies Using the Area under the Receiver Operator Characteristic Curve as a Measure of Diagnostic Accuracy: Systematic Review with a Focus on Quality of Data Reporting. *PloS one*, 9: e116018.

Finlayson, S. G.; Subbaswamy, A.; Singh, K.; Bowers, J.; Kupke, A.; Zittrain, J.; Kohane, I. S.; and Saria, S. 2020. The clinician and dataset shift in artificial intelligence. *The New England Journal of Medicine*, 283–286.

Freeman, K.; Geppert, J.; Stinton, C.; Todkill, D.; Johnson, S.; Clarke, A.; and Taylor-Phillips, S. 2021. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. *BMJ*, 374.

Gaube, S.; Suresh, H.; Raue, M.; Merritt, A.; Berkowitz, S. J.; Lermer, E.; Coughlin, J. F.; Guttag, J. V.; Colak, E.; and Ghassemi, M. 2021. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ digital medicine*, 4(1): 1–8.

Ghassemi, M.; Naumann, T.; Schulam, P.; Beam, A. L.; and Ranganath, R. 2018. Opportunities in Machine Learning for Healthcare. *CoRR*, abs/1806.00388.

Goh, K. H.; Wang, L.; Yeow, A. Y. K.; Poh, H.; Li, K.; Yeow, J. J. L.; and Tan, G. Y. H. 2021. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nature communications*, 12(1): 1–10.

Green, B.; and Hu, L. 2018. The Myth in the Methodology: Towards a Recontextualization of Fairness in Machine Learning. In *ICML 2018*.

Gupta, S.; Singh, P.; Chang, K.; Qu, L.; Aggarwal, M.; Arun, N.; Vaswani, A.; Raghavan, S.; Agarwal, V.; Gidwani, M.; et al. 2021. Addressing catastrophic forgetting for medical domain expansion. *arXiv preprint arXiv:2103.13511*.

He, J.; Baxter, S.; Xu, J.; Xu, J.; Zhou, X.; and Zhang, K. 2019. The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine*, 25.

IHE Radiology Technical Committee. 2020. IHE Radiology Technical Framework Supplement AI Results (AIR). https://www.ihe.net/uploadedFiles/Documents/Radiology/IHE_RAD_Suppl_AIR.pdf.

Jacobs, A. Z.; and Wallach, H. M. 2019. Measurement and Fairness. *CoRR*, abs/1912.05511.

Jacobs, M.; He, J.; Pradier, M. F.; Lam, B.; Ahn, A. C.; McCoy, T. H.; Perlis, R. H.; Doshi-Velez, F.; and Gajos, K. Z. 2021. Designing AI for Trust and Collaboration in Time-Constrained Medical Decisions: A Sociotechnical Lens. *CoRR*, abs/2102.00593.

Jr., D. M.; Prabhakaran, V.; Kuhlberg, J.; Smart, A.; and Isaac, W. S. 2020. Extending the Machine Learning Abstraction Boundary: A Complex Systems Approach to Incorporate Societal Context. *CoRR*, abs/2006.09663.

Kelly, C.; Karthikesalingam, A.; Suleyman, M.; Corrado, G.; and King, D. 2019. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17.

Kiani, A.; Uyumazturk, B.; Rajpurkar, P.; Wang, A.; Gao, R.; Jones, E.; Yu, Y.; Langlotz, C. P.; Ball, R. L.; Montine, T. J.; et al. 2020. Impact of a deep learning assistant on the histopathologic classification of liver cancer. *NPJ digital medicine*, 3(1): 1–8.

Kravet, S.; Howell, E.; and Wright, S. 2006. Morbidity and Mortality Conference, Grand Rounds, and the ACGME's

Core Competencies. *Journal of general internal medicine*, 21: 1192–4.

Larrazabal, A. J.; Nieto, N.; Peterson, V.; Milone, D. H.; and Ferrante, E. 2020. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23): 12592–12594.

Lazer, D.; Kennedy, R.; King, G.; and Vespignani, A. 2014. The Parable of Google Flu: Traps in Big Data Analysis. *Science (New York, N.Y.)*, 343: 1203–5.

Liu, X.; Faes, L.; Kale, A. U.; Wagner, S. K.; Fu, D. J.; Bruynseels, A.; Mahendiran, T.; Moraes, G.; Shamdas, M.; Kern, C.; et al. 2019. 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.

Lu, C.; Lemay, A.; Chang, K.; Hoebel, K.; and Kalpathy-Cramer, J. 2021a. Fair Conformal Predictors for Applications in Medical Imaging. arXiv:2109.04392.

Lu, C.; Lemay, A.; Hoebel, K.; and Kalpathy-Cramer, J. 2021b. Evaluating subgroup disparity using epistemic uncertainty in mammography. arXiv:2107.02716.

Lu, C.; Strout, J.; Gauriau, R.; Wright, B.; Marcruz, F. B. D. C.; Buch, V.; and Andriole, K. P. 2020. An Overview and Case Study of the Clinical AI Model Development Life Cycle for Healthcare Systems. *CoRR*, abs/2003.07678.

Masic, I.; Miokovic, M.; and Muhamedagić, B. 2008. Evidence Based Medicine – New Approaches and Challenges. *Acta informatica medica : AIM : journal of the Society for Medical Informatics of Bosnia & Herzegovina : casopis Drustva za medicinsku informatiku BiH*, 16: 219–225.

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

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.; and Wiegand, T. 2020. ML4H Auditing: From Paper to Practice. In Alsentzer, E.; McDermott, M. B. A.; Falck, F.; Sarkar, S. K.; Roy, S.; and Hyland, S. L., eds., *Proceedings of the Machine Learning for Health NeurIPS Workshop*, volume 136 of *Proceedings of Machine Learning Research*, 280–317. PMLR.

Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J. V.; Lakshminarayanan, B.; and Snoek, J. 2019. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. arXiv:1906.02530.

Pianykh, O. S.; Langs, G.; Dewey, M.; Enzmann, D. R.; Herold, C. J.; Schoenberg, S. O.; and Brink, J. A. 2020. Continuous Learning AI in Radiology: Implementation Principles and Early Applications. *Radiology*, 297(1): 6–14. PMID: 32840473.

Pierson, E.; Cutler, D. M.; Leskovec, J.; Mullainathan, S.; and Obermeyer, Z. 2021. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nature Medicine*, 27(1): 136–140.

Rabanser, S.; Günnemann, S.; and Lipton, Z. C. 2019. Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift. arXiv:1810.11953.

Roberts, M.; Driggs, D.; Thorpe, M.; Gilbey, J.; Yeung, M.; Ursprung, S.; Aviles-Rivero, A. I.; Etmann, C.; McCague, C.; Beer, L.; et al. 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.

Schwendicke, F.; Samek, W.; and Krois, J. 2020. Artificial Intelligence in Dentistry: Chances and Challenges. *Journal of Dental Research*, 99: 002203452091571.

Stead, W. W. 2018. Clinical Implications and Challenges of Artificial Intelligence and Deep Learning. *JAMA*, 320(11): 1107–1108.

Strickland, E. 2019. IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. *IEEE Spectrum*, 56: 24–31.

Vincent, J. 2021. Google dismantles Health unit in favor of tried-and-tested 'throw everything at the wall' strategy.

Voter, A.; Larson, M.; Garrett, J.; and Yu, J.-P. 2021. Diagnostic Accuracy and Failure Mode Analysis of a Deep Learning Algorithm for the Detection of Cervical Spine Fractures. *American Journal of Neuroradiology*.

Wang, D.; Wang, L.; Zhang, Z.; Wang, D.; Zhu, H.; Gao, Y.; Fan, X.; and Tian, F. 2021. "Brilliant AI Doctor" in Rural China: Tensions and Challenges in AI-Powered CDSS Deployment. *CoRR*, abs/2101.01524.

Wiens, J.; Saria, S.; Sendak, M. P.; Ghassemi, M.; Liu, V. X.; Doshi-Velez, F.; Jung, K.; Heller, K. A.; Kale, D. C.; Saeed, M.; Ossorio, P. N.; Thadaney-Israni, S.; and Goldenberg, A. 2019. Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine*, 1–4.

Wolff, J.; Pauling, J.; Keck, A.; and Baumbach, J. 2021. Success Factors of Artificial Intelligence Implementation in Healthcare. *Frontiers in Digital Health*, 3: 51.

Wu, E.; Wu, K.; Daneshjou, R.; Ouyang, D.; Ho, D.; and Zou, J. 2021. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nature Medicine*, 27: 1–3.

Xie, Y.; Chen, M.; Kao, D.; Gao, G.; and Chen, X. A. 2020. CheXplain: Enabling Physicians to Explore and UnderstandData-Driven, AI-Enabled Medical Imaging Analysis. *CoRR*, abs/2001.05149.

Yan, K.; Wang, X.; Lu, L.; and Summers, R. M. 2018. DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of medical imaging*, 5(3): 036501.

Yang, Q.; Steinfeld, A.; Rosé, C.; and Zimmerman, J. 2020. Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design.

Zech, J. R.; Badgeley, M. A.; Liu, M.; Costa, A. B.; Titano, J. J.; and Oermann, E. K. 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11): e1002683.

Zhang, H.; Dullerud, N.; Seyyed-Kalantari, L.; Morris, Q.; Joshi, S.; and Ghassemi, M. 2021. An Empirical Framework for Domain Generalization in Clinical Settings. *CoRR*, abs/2103.11163.