Aakanksha Khandwaha
April 18, 2024

Benchmark Fetishism: Flaws in Medical AI Model Evaluation

Benchmarking is a core mechanism for measuring progress in artificial intelligence (AI). Most commonly, a 'benchmark' is defined as a combination of a dataset and metric(s) that is used to evaluate the performance of a model on a particular task (Raji et al. 2021). Achieving state-of-the-art performance on benchmarks has been synonymous with meaningful advancement in model capability and performance.

At the same time, it has been said that artificial intelligence has the potential to transform medicine and clinical care. From improving diagnostic accuracy to optimizing clinical workflows and unlocking personalized treatments, AI benchmarked at human-level or "superhuman performance" promises to revolutionize healthcare and patient outcomes.

However, reality has yet to match this potential. Despite the proliferation of medical AI models that demonstrate remarkable results on benchmarks, comparatively few have seen significant clinical adoption or real-world impact. Models that appear to equal or exceed human practitioners in silico often fail to translate those gains into improvements in actual clinical environments and use cases.

I argue that that this disconnect stems largely from fundamental flaws in the benchmarks themselves and the ways they are used to evaluate AI models, using examples from general and medical applications. I will first outline how we are over-relying on benchmarks, then explain two problems with this approach – what I'll refer to as the "external" and "internal" flaws with benchmarking. "External" flaws relate to the lack of scientific rigor and standardization in the benchmarking process, whereas "internal" flaws are inherent to the design and composition of the benchmarks themselves. I will also argue that, most importantly, benchmark performance does not represent how well a model will perform in the real world.

Overreliance on Benchmarks

While it is important to determine ways to measure progress in the field of artificial intelligence, there has been an increasing reliance of benchmarks and metrics as the be-all end-all of evaluation. This results in chasing incremental increases in performance to achieve "state-of-the-art" rather than hypothesis-based scientific inquiry (Raji et al. 2021). According to Saitta, benchmark chasing allows researchers to "publish dull papers that proposed small variations of existing […] algorithms" (Raji et al. 2021). This is known as *diminishing returns,* when "increasingly large efforts achieve smaller and smaller performance gains" (Varoquaux and Cheplygina 2022). For example, in 8 public medical imaging challenges (for either disease diagnosis or image segmentation), the "best" algorithm in 6 of the challenges had performance gains that were smaller than the expected variance compared to algorithms in the top 10% performance-wise (Varoquaux and Cheplygina 2022).

This emphasis on performance on benchmarks also influences the direction of AI development, potentially shifting research in entire fields. For example, due to the popularity of chess in the 1960s, deep-tree searching and the minimax algorithms dominated the field because of their effectiveness in improving game performance (Raji et al. 2021). In natural language processing (NLP), a focus on metrics overtook the rest of the field since it was a key condition for funding from DARPA, resulting in a lack of funding for projects that did not have immediate performance gains (Gururaja et al. 2023).

Such tunnel vision in AI development may be due to misaligned incentives in academia. In a survey, many NLP researchers found that the immediate goal of a research paper was to "get another 2% [increase in performance] and get the boldface black entry on the table." (Gururaja et al. 2023). Publishing especially incentivizes research into "state-of-the-art" models where reviewers seemed to find improvements on benchmarks to be the only justification in the validity of a paper (Gururaja et al. 2023). This results in researchers exploiting tricks to achieve "state-of-the-art" results on benchmarks rather than exploring the "deeper mechanisms by which models function", write unnecessarily "math-y" papers, or use suggestive language (e.g., 'human-level performance') when describing their AI model's efficacy (Varoquaux and Cheplygina 2022; Gururaja et al. 2023). In a systematic review of more than 150 studies on prediction models, most articles contained at least 7 examples of 'spin' - language exaggerating the benefits while downplaying the costs, risks, and limitations (Morley 2023). These academic incentives

undermine the needs of stakeholders such as clinicians and patients (Varoquaux and Cheplygina 2022).

This is not to say that benchmarks are completely irrelevant, but that they not enough – the goals of AI cannot be fully evaluated through "data-defined benchmarks" (Raji et al. 2021). A survey of NLP researchers found that benchmarks are necessary, but becoming increasingly insufficient (Gururaja et al. 2023). There is a similar tendency in medical AI research to focus solely on technical evaluations rather than performance in real-world, clinical settings (Morley 2023).

"External" Flaws in Benchmarking

The evaluation of AI through benchmarks is not scientifically rigorous and lacks standardization. This includes data leakage, where many popular large language model (LLM) benchmarks have been found to leak into the training data of the latest models thus falsely increasing performance (McIntosh et al. 2024). Synonymously in healthcare, the 'fingerprinting phenomenon' results in different samples of the same patient ending up in both training and validation datasets where the algorithm learns to recognize that patient rather than markers of the disease in question (Morley 2023; Varoquaux and Cheplygina 2022). For example, some studies diagnosing ADHD based on brain imaging performed circular analysis, where they "perform[ed] feature selection on the full dataset, before cross-validation" (Varoquaux and Cheplygina 2022).

Similarly, there is a prevalence of 'p-hacking' in AI research, which occurs when models overfit to the training data. For example, in a study developing an AI model to detect polyps, they removed the videos on which the model performed poorly and ones with more than one polyp (Hicks et al. 2022). Using a subset of the actual dataset can give potentially false impressions of a model's actual performance. There are also inconsistencies amongst research papers in reporting metrics and benchmarks. For example, Area Under Curve (AUC) metrics are usually referenced as 'the AUC curve' even though there are different meanings depending on if it's the plotting precision and recall against each other (PR-AUC), or recall and the false-positive rate (ROC-AUC) (Blagec et al. 2021). Both result in an obfuscation of an AI model's performance, making it difficult for the public to interpret findings. Proper scientific communication is especially important in interdisciplinary domains such as medical AI, where

healthcare practitioners need to understand what certain benchmarks represent in terms of clinical care.

Furthermore, only a small subset of benchmarks and metrics tend to be reported. In a comparative study, the performance of more than three-quarters of benchmark datasets was represented using only a single metric (Blagec et al. 2021). Specifically, classification metrics (e.g., accuracy) were the most presented in benchmark papers (Blagec et al. 2021). It is much easier to present and interpret one metric, but it only provides one perspective on the model's performance and devalues others. Calculating a range of metrics, especially for classification tasks, does not require extra measurements or design considerations because most are based on the true positive, false positive, true negative, and false negative rates. Thus, a researcher would exclude them only if they lacked space or knowledge, or were deliberately obfuscating actual performance (Hicks et al. 2022).

At the same time, there are differences in the interpretation of the same metrics amongst different domains/datasets and are not comparable even though they are all on the scale of 0-100%. In botany, for example, 80% accuracy may be sufficient for classifying Irises but would be dangerous if trying to classify between a poisonous or edible mushroom (Raji et al. 2021). For unbalanced datasets, which are common in medical applications[1], accuracy is a worse indicator of performance than precision (Blagec et al. 2021). This is due to the 'accuracy paradox' where if class A exists in the dataset 90% of the time and a classifier predicts A for all data points, it will have an accuracy of 90%. Sensitivity[2] and specificity[3] are metrics that are especially critical in healthcare because they help evaluate the accuracy of a test in correctly identifying patients with and without a particular condition or disease. However, they are not provided as often by AI researchers, even when they are working on medical applications (Morley 2023).

It is also extremely difficult to replicate or reproduce findings. In LLMs, for example, simple changes in text input - switching from *A)* to *1)*, or inserting an extra space between the option and answer – results in up to 5% variance in text output (McIntosh et al. 2024). These prompts are integral to evaluating LLMs on popular benchmarks such as MMLU and produce such significant differences in output that there is a whole field dedicated to this, known as

[1]Malignant, or positive, diagnoses are less common than benign in medical datasets
[2] Sensitivity, or the True Positive Rate, measures the proportion of actual positives that are correctly identified
[3] Specificity, or False Positive Rate, measures the proportion of actual negatives that are correctly identified

'prompt engineering'. Furthermore, AI models are increasingly privatized, developed with proprietary (thus not publicly available) model architecture and datasets which prevents external researchers from validating model performance. This is especially true in healthcare where model evaluation omits key details such as training data due to privacy risks, resulting in the replication rate in clinical decision support system research being 3 in 1000 studies (Morley 2023).

"Internal" Flaws in Benchmarking

Each individual benchmark is value-laden and cannot be treated as neutral – it embodies a limited, subjective perspective. More recently, benchmarks are being used to define 'general' cognitive abilities such as visual or language understanding (Raji et al. 2021). Many of these benchmarks were originally developed to measure performance on a tightly scoped, finite, and domain-specific task such as automatic speech recognition (ASR) or machine translation (MT). ImageNet[4], for example, was described by its creators as "the most comprehensive and diverse coverage of the image world" and an "attempt to map the entire world of objects" (Raji et al. 2021). However, as we'll see later, this is an inaccurate representation of the benchmark. Inversely, exams and scientific texts have been hastily converted into benchmarks and used to boast about the model's performance. Most notably, OpenAI's GPT-4 was evaluated on "exams that were originally designed for humans" including the bar exam (OpenAI 2023). In healthcare, the United States Medical Licensing Examinations (USMLE) exam is most commonly used to compare the ability of LLMs to healthcare practitioners. However, it has been shown that such exams "fail to fully assess the skills required for modern medical practice" (Mbakwe et al. 2023). This is because the ability to "regurgitate mechanistic models of health and disease" may not be as important as critical thinking skills and respect for patients in a clinician (Mbakwe et al. 2023).

Image datasets are a great example of this. It was found that images in the same category but different datasets were distinguishable from one another – each image encodes a certain perspective when captured (Raji et al. 2021). Specifically in ImageNet, most objects tend to be centred in pictures which is not usually how these objects would "naturally appear" (Raji et al. 2021). There is also a significant lack of geo-diversity, with 45% of the images sourced from the

[4] ImageNet is a large database of images labelled according to the WordNet hierarchy (Raji et al. 2021)

United States and over 60% from the Global North (Raji et al. 2021). On the other hand, only 1% and 1.2% of images are from China and India, respectively, even though those countries are the most populous countries on the planet (Raji et al. 2021). Furthermore, marginalized people are underrepresented and often tagged with racial or ethnic slurs in them. These perspectives are similarly encoded in medical imaging datasets, which tend to comprise of patients who received clinical care at one of the elite few institutions in North America and Europe (Varoquaux and Cheplygina 2022). Furthermore, most dermatology datasets contain images of light-skinned patients (Daneshjou et al. 2022). Thus, benchmarking on datasets that are skewed towards WEIRD[5] people hides unequal performance on underrepresented groups, pushing them to the margins while also justifying the deployment of biased models[6].

There is also little to no intentional data curation during the development of most benchmarks. ImageNet adapted the English-language WordNet hierarchy despite the fact that it was developed in a different field for a different purpose (Raji et al. 2021). Furthermore, the images were labelled arbitrarily (e.g., from specific dog breeds to a generic 'New Zealand beach') and used a range of derogatory and offensive categories (Raji et al. 2021). Similarly, the GLUE benchmark was curated based on what a select few NLP researchers thought to be interesting at the time (Raji et al. 2021). As a result, the final benchmark neither systematically mapped out a range of specific linguistic skills nor present a truly varied range of ways to deploy linguistic knowledge in comprehension.

There have also been cases of AI "inbreeding" during the development of benchmarks. A survey found that 9 out of 23 LLM benchmark datasets were generated by LLMs themselves (McIntosh et al. 2024). Similarly, the ChestX-ray14 dataset produced labels using NLP from radiology reports (Gichoya et al. 2018). Both datasets now encode the biases and inaccuracies of the models used to generate the data, making any evaluations based on these benchmarks unreliable.

[5] WEIRD is an acronym for Western, Educated, Industrialized, Rich and Democratic

[6] Most medical datasets do not include data about the patient's age, sex, or gender which makes it even more excruciatingly difficult to analyze performance amongst different subgroups

Benchmark performance does not represent real-world performance

The UCI Machine Learning repository is a set of benchmarks for a variety of individual subtasks, including Iris, Adult, Wine, and Breast Cancer classification datasets (Raji et al. 2021). The popularity of these benchmarks has led to hordes of AI researchers working on developing the best Iris classifier, but do botanists truly need this type of model? Are they publishing about the topic in journals? NLP researchers similarly state that their "field that, like fundamentally, is about something about people, knows remarkably little about people" (Gururaja et al. 2023). Many benchmarks have been created perpendicular to needs in the real world.

This phenomenon can be described by the "measure-target confusion", coined by Müller (2020). We observe progress in AI development and then try to see how this can be measured – via benchmarks. But then we turn the measure into the target: We ask for research that improves on benchmarks. The example Müller provides is progress of individual scientists: we need quantifiable metrics to evaluate scientific success so publication numbers, h-index, funding, etc. are used to make future decisions. However, this results in researchers targeting the metrics instead of the original target - scientific progress. Note that this is also an "internal" flaw of benchmarks and one so significant that it deserves a more thorough examination[7].

This confusion can easily be observed in medical AI benchmarks. A recent survey showed that very few benchmarks of direct clinical relevance exist, and the ones that do fail to cover many tasks that clinicians most want to see addressed (Blagec et al. 2023). In medical imaging specifically, there is enormous research showing state-of-the-art performance on benchmarks with no "practical improvement" for the clinical problem (Varoquaux and Cheplygina 2022). Similarly, 62 studies in AI for COVID were reviewed, and none of them had potential for clinical use (Varoquaux and Cheplygina 2022). There may be a few reasons this disjointedness exists.

In some cases, the data distribution in a benchmark may not match the target population where the model was intended to be deployed. A review found that more than half of the studies (54%) provided no clarity on whether the population included in the study matched, or at least aligned to, the population of the area (Morley 2023). The same phenomenon has been demonstrated in medical imaging (chest X-rays, retinal imaging, brain imaging, histopathology and dermatology) due to privacy/security concerns as well as disparities in who has access to a

[7] At least, I think it does!

healthcare system (Varoquaux and Cheplygina 2022). Evaluation on the same data pool from one of the few benchmark datasets does not assess clinical relevance since there could be key differences in the benchmark and the actual population the model will perform on.

The benchmark may also include proxies that would not be present in real-time clinical data. For example, on chest X-ray datasets, images for the "pneumothorax" condition sometimes show a chest drain, which is a treatment for this condition, and which would not yet be present before diagnosis (Varoquaux and Cheplygina 2022). Similarly, a dataset of different skin conditions may include markings placed by the dermatologist which would usually not be present before consultation.

Even if the above two biases are accounted for, real data is inherently messier than clean, pre-processed benchmark datasets. Benchmark creators may remove imperfect data samples, or outliers. For example, a system trained only on high-quality images might provide incorrect diagnosis when classifying images of low quality, with different lighting, or other differences in real-world clinical settings. Furthermore, data for a single patient may be stored in "a multitude of medical imaging archival systems, pathology systems, EHRs, electronic prescribing tools and insurance databases" (Kelly et al. 2019). This data is not prepared to be inputted into an AI model.

Benchmark datasets also use historically labelled data, with most evaluations of medical AI models done retrospectively (Kelly et al. 2019). Patient populations and clinical practices change over time, but AI models do not (and tend to be brittle!). Evaluating solely on historical data does not guarantee that the model will perform well on future patients. Currently, no benchmarks exist to evaluate the model's ability to respond to such data drift.

Outside of the data, this disconnect may also be because AI researcher's lack proper understanding of the medical field and clinical care. An example of this in radiology AI is the CheXNet model which claimed to diagnose pneumonia, without specifying whether it is clinical or radiological pneumonia[8] (Gichoya et al. 2018). Thus, the benchmarks used to evaluate CheXNet performance clearly do not have any clinical significance. Diagnosis is rarely a binary classification problem as current benchmarks suggest, but instead require complex decision making.

[8] In comparison, understanding the difference between the two types of pneumonia is said to be "common sense" to clinicians

Moving Beyond Benchmarks

For the reasons outlined above, it is imperative that AI research moves beyond (perhaps away from) benchmarks and instead refocus on the original goals of AI development. Gururaja et al. state that we should move on from benchmarks not when they are saturated but when "it wouldn't really improve the world to improve this performance anymore" (2023). By over-relying on and misusing flawed benchmarks, medical AI research risks sidelining model development and giving false impressions of capability. Shifting to a more nuanced, clinically grounded approach to model evaluation is crucial for closing the gap between medical AI's imagined potential and real-world utility.

In healthcare, this means improving clinical care by integrating technology within the current workflow. We should also aim to improve upon the "external" flaws of benchmarking by emphasizing the importance of scientific communication amongst AI researchers and ensuring they fairly interpret the implications of good benchmark performance. However, ultimately what matters is whether the model resulted in a beneficial change in patient care. To do this, AI researchers must collaborate with healthcare practitioners to understand what parts of their job is the most burdensome, such as routine documentation and patient data administration (Blagec et al. 2023). There should also be a greater focus on evaluating AI models during the deployment stage, with consistent iterations to ensure it is integrated properly – "even if the system's knowledge is indeed accurate, complete, and consistent, it will be of little help if it's clinical interface is faulty." (Morley 2023). Lastly, but most importantly, stakeholders including patients and healthcare practitioners should be deeply involved in the development of medical AI models to ensure that to measurements of progress translate computational achievement into meaningful clinical impact.

References

Blagec, Kathrin, Georg Dorffner, Milad Moradi, and Matthias Samwald. 2021. “A Critical Analysis of Metrics Used for Measuring Progress in Artificial Intelligence.” arXiv. http://arxiv.org/abs/2008.02577.

Blagec, Kathrin, Jakob Kraiger, Wolfgang Frühwirt, and Matthias Samwald. 2023. “Benchmark Datasets Driving Artificial Intelligence Development Fail to Capture the Needs of Medical Professionals.” *Journal of Biomedical Informatics* 137 (January): 104274. https://doi.org/10.1016/j.jbi.2022.104274.

Daneshjou, Roxana, Kailas Vodrahalli, Roberto A. Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, et al. 2022. “Disparities in Dermatology AI Performance on a Diverse, Curated Clinical Image Set.” *Science Advances* 8 (32): eabq6147. https://doi.org/10.1126/sciadv.abq6147.

Gichoya, Judy W., Siddhartha Nuthakki, Pallavi G. Maity, and Saptarshi Purkayastha. 2018. “Phronesis of AI in Radiology: Superhuman Meets Natural Stupidity.” arXiv. http://arxiv.org/abs/1803.11244.

Gururaja, Sireesh, Amanda Bertsch, Clara Na, David Gray Widder, and Emma Strubell. 2023. “To Build Our Future, We Must Know Our Past: Contextualizing Paradigm Shifts in Natural Language Processing.” arXiv. http://arxiv.org/abs/2310.07715.

Hicks, Steven A., Inga Strümke, Vajira Thambawita, Malek Hammou, Michael A. Riegler, Pål Halvorsen, and Sravanthi Parasa. 2022. “On Evaluation Metrics for Medical Applications of Artificial Intelligence.” *Scientific Reports* 12 (1): 5979. https://doi.org/10.1038/s41598-022-09954-8.

Kelly, Christopher J., Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. 2019. “Key Challenges for Delivering Clinical Impact with Artificial Intelligence.” *BMC Medicine* 17 (1): 195. https://doi.org/10.1186/s12916-019-1426-2.

Mbakwe, Amarachi B., Ismini Lourentzou, Leo Anthony Celi, Oren J. Mechanic, and Alon Dagan. 2023. “ChatGPT Passing USMLE Shines a Spotlight on the Flaws of Medical Education.” Edited by Harry Hochheiser. *PLOS Digital Health* 2 (2): e0000205. https://doi.org/10.1371/journal.pdig.0000205.

McIntosh, Timothy R., Teo Susnjak, Tong Liu, Paul Watters, and Malka N. Halgamuge. 2024. "Inadequacies of Large Language Model Benchmarks in the Era of Generative Artificial Intelligence." arXiv. http://arxiv.org/abs/2402.09880.

Morley, Jessica. 2023. "Thinking Critically about AI in Healthcare." May 2023. https://drive.google.com/file/d/15PiSt1fEuKX0SzCnnbud-swA2stYUY6v/view?usp=embed_facebook.

Müller, Vincent C. 2020. "Measuring Progress in Robotics: Benchmarking and the 'Measure-Target Confusion.'" In *Metrics of Sensory Motor Coordination and Integration in Robots and Animals*, edited by Fabio Bonsignorio, Elena Messina, Angel P. Del Pobil, and John Hallam, 36:169–79. Cognitive Systems Monographs. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-14126-4_9.

OpenAI. 2023. "GPT-4." March 14, 2023. https://openai.com/research/gpt-4.

Raji, Inioluwa Deborah, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. "AI and the Everything in the Whole Wide World Benchmark." arXiv. http://arxiv.org/abs/2111.15366.

Varoquaux, Gaël, and Veronika Cheplygina. 2022. "Machine Learning for Medical Imaging: Methodological Failures and Recommendations for the Future." *Npj Digital Medicine* 5 (1): 48. https://doi.org/10.1038/s41746-022-00592-y.