Training Data Tell Us a Lot About Whom Health AI Tools Are Likely to Benefit

COMMENTARY

P. Alison Paprica, PhD PMP
Professor (Adjunct) and Senior Fellow
Institute for Health Policy, Management and Evaluation
University of Toronto
Toronto, ON

ABSTRACT

Appropriate training data are a prerequisite for health AI tools. Policy makers, clinicians and patients can assess the datasets used to train AI models as a practical step in determining whom health AI tools are likely to benefit. Analyses of training datasets can help prioritize which health AI tools to validate and help identify where changes are needed to improve the equity of health AI.

Introduction

In their article in this issue, Kueper and Pandit (2025) identify how data have contributed to advances and challenges for AI in healthcare in Canada. For example, the authors note how the availability of large datasets led to advances in the data-centric AI methods used today and include joint guidance from the Food and Drug Administration, Health Canada and the UK Regulatory Authority that focuses on addressing issues related to data in their *Table 2. International guidelines for AI development and use relevant to Canadian healthcare*.

I argue here that even more emphasis on training datasets is warranted because (1) appropriate training data are a prerequisite for AI; (2) health AI tools cannot be expected to work well for subgroups and individuals who are grossly under-represented in training data; (3) transparency about the data used to train health AI models can help policy makers, clinicians and individual patients understand whom AI tools are likely to work for; and (4)

analyzing the appropriateness of training data is a practical step that can help prioritize future validation work and improve the equity of health AI tools in the long term.

AI Requires Alignment Between the Data, the Objective/Task and the Model/Method

When I started work as the inaugural vice president of Health Strategy and Partnerships at the Vector Institute for Artificial Intelligence, I was fortunate to learn a simple way of understanding what is required for machine learning from Vector faculty member Graham Taylor. Graham explained to me that machine learning AI requires alignment between (1) the data; (2) the objective or task; and (3) the AI model or method (Taylor, Personal communication, 2017). Graham's triad of alignment for AI resonates with me, and with others who are accustomed to making decisions based on data, probably because the need for alignment between data, objectives and methods applies to both AI and traditional statistical methods.

Since the usual starting point for health AI is a large dataset that has the potential to be reused for public benefit, fulfilling the triad of alignment for a health AI tool generally entails choosing a model/method and objective/task that align with the data. For example, if the objective/task is to have a human-understandable prediction, alignment could be achieved by choosing an interpretable AI model (Vokinger et al. 2021). In contrast, it may not be possible to achieve alignment, no matter what AI model/method is selected, if the objective/task is to have a health AI tool that works for all subgroups of a population, but the training dataset includes biased, incomplete or incorrect information about some subgroups (Gupta and Treviranus 2022).

An easy-to-understand example of this kind of misalignment comes from research focused on AI tools for dermatology. Though there are efforts to improve the representation of people with dark skin types in dermatology image datasets (ISIC 2024), most current datasets do not include descriptions of patient ethnicity or race or any information about skin tone (Daneshjou et al. 2021; Wen et al. 2022), and dermatological AI models have been reported to exhibit worse performance for people with dark skin types (Daneshjou et al. 2022; Fliorent et al. 2024). An editorial by Tschandl (2021: 1271) sums up the problem succinctly, "We cannot expect AI to work well for rare disease variants or expressions of diseases in different contexts if these cases are underrepresented, or not included at all, in the training set."

Statistical Discrimination and Thresholds Below Which AI Tools May Not Work for Some Subgroups and People

This same sentiment is apparent in the Accessibility Standards Canada's (2024) *Accessible and Equitable Artificial Intelligence Systems – Technical Guide*. The Technical Guide includes the concept of "statistical discrimination" and notes that even when people with disabilities are represented proportionately in datasets, AI predictions are often wrong or exclude them. This happens because high error rates for minority subgroups can be obfuscated when AI models are focused on good performance for the majority (Gupta and Treviranus 2022).

To mitigate the effect of statistical discrimination, Figure 1 in the *Accessible and Equitable Artificial Intelligence Systems – Technical Guide* presents thresholds for when and how AI tools should be used based on the percentage of people, groups or communities with a relevant attribute in the training data, that is:

- If the relevant attribute is present in more than 50% of the training data, the AI tool can be used with standard precautions and monitoring.
- If the relevant attribute is present in less than 50% but more than 20% of the training data, the AI tool should only be used with human monitoring and oversight to address inaccurate determinations.
- If the relevant attribute is present in less than 20% of the training data, the AI tool should not be used.

Acknowledging that the numerical values of 20% and 50% would likely change based on the context, risks and potential harms associated with an AI tool, the key point of the thresholds is that some AI tools may not work for minority subgroups, especially people who are outliers in datasets because they have multiple intersecting relevant attributes that diverge from the statistical mean (Gupta and Treviranus 2022). It follows that detailed information about training data needs to be available to help people decide if and how to use a health AI tool.

... AI tools may not work for minority subgroups, especially people who are outliers in datasets ...

Transparency About the Datasets Used to Train and Customize AI Models

Gupta and Treviranus (2022) identify two broad approaches to transparency and auditing of AI; one focuses on providing information to domain and technical experts, and the other focuses on providing information to the people to whom the AI tools would be applied. For the former, there are a growing number of resources to support experts in assessing AI, including health research reporting guidelines (e.g., see EQUATOR Network 2024) and tools to communicate information about data and AI models (e.g., see Data Nutrition Project n.d.; Nsoesie and Ghassemi 2024; Partnership on AI n.d.). However, if information about training data is going to be used to help policy makers, clinicians and patients understand whom an AI tool is likely to benefit, simple tools to communicate information about training datasets to non-expert audiences will also be required.

Fortunately, there is excellent work on "Dataset Nutrition Labels" to build upon (Data Nutrition Project n.d.), including an example draft nutrition label for a large health-related dataset (Data Nutrition Project 2023). While the primary audience for the Dataset Nutrition Label is the data science and developer community who build AI models (Data Nutrition Project n.d.), it is easy to imagine how learnings from dashboards, decision aids and other health decision-making tools could bring information about AI training data to policy makers, clinicians and patients.

For clarity, I am not suggesting that analyzing and communicating information about the representation of subgroups in training data would be sufficient to assess whether a health AI tool should be used, and for whom. For one thing, even if a training dataset appears to adequately represent minority subgroups, the labels or categories established when the data were collected for their original purpose may be misleading or insufficient for the intended AI objective/task (Gupta and Treviranus 2022). Second, it may be possible to address under-representation in training datasets through oversampling and other techniques (e.g., see Daneshjou et al. 2022; Vokinger et al. 2021). Third, as noted by Kueper and Pandit (2025), many health AI risks and harms are not related to training data, so "a simple checklist approach to responsible AI will not be sufficient." Still,

analyzing training data can be a practical step that tells us a lot about health AI tools.

Analyses of Datasets Should Help Prioritize Future Work on Health AI Tools

For example, analyses of training data can help with two key issues identified by Kueper and Pandit (2025): (1) much work is needed to make health AI tools generalizable and (2) far more AI predictive models have been developed than are externally validated. Comparing an AI model's training data with the data at a potential new implementation site can be a first step in determining the likelihood that a health AI tool will be generalizable to the new setting. In addition, analyzing training data can be a practical way to shed light on which AI models are likely to be generalizable, and therefore priority candidates for validation. Training data can also clarify for which groups and subpopulations it is possible to validate a health AI tool.

In the short term, analyses of training data may result in scoped-down objectives/tasks to ensure that health AI tools are only applied to groups who are adequately represented in training data. In the longer term, such scoping down would make it easier to identify equity gaps in terms of those whom health AI tools are benefiting. Information about gaps can, in turn, help prioritize work to improve training data and AI models so that health AI becomes more equitable.

Conclusions

Just as we assess the appropriateness of data when interpreting studies and analytics performed with traditional statistical methods, we can learn a lot about whom health AI tools are likely to benefit by analyzing the datasets used to train health AI models. Analyses of training data are not sufficient to ensure responsible AI. However, presenting information about training data in simple and understandable ways can help policy makers, clinicians and patients understand what health AI tools can do for whom and inform planning for investments in validation and improvements in health AI equity.

References

Accessibility Standards Canada. 2024, August 7. *Accessible and Equitable Artificial Intelligence Systems – Technical Guide*. Retrieved October 24, 2024. <https://accessible.canada.ca/sites/default/files/2024-09/technicalguide-artificialintelligence-asc.pdf>.

Daneshjou, R., M.P. Smith, M.D. Sun, V. Rotemberg and J. Zou. 2021. Lack of Transparency and Potential Bias in Artificial Intelligence Data Sets and Algorithms: A Scoping Review. *JAMA Dermatology* 157(11): 1362–69. doi:10.1001/jamadermatol.2021.3129.

Daneshjou, R., K. Vodrahalli, R.A. Novoa, M. Jenkins, W. Liang, V. Rotemberg et al. 2022. Disparities in Dermatology AI Performance on a Diverse, Curated Clinical Image Set. *Science Advances* 8(31): eabq6147. doi:10.1126/sciadv.abq6147.

Data Nutrition Project. n.d. The Data Nutrition Project: Empowering Data Scientists and Policymakers With Practical Tools to Improve AI Outcomes. Retrieved October 24, 2024. <https://datanutrition.org/>.

Data Nutrition Project. 2023. California Health Interview Survey (CHIS 2021) Adult Data File. Retrieved October 24, 2024. <https://datanutrition.org/labels/v3/?id=ee3d5314-e1b5-4fc2-a370-ae511a2696ca>.

Enhancing the QUAlity and Transparency Of Health Research (EQUATOR Network). 2024. Search Results for "Machine Learning." Retrieved October 24, 2024. <https://www.equator-network.org/?post_type=eq_guidelines&eq_guidelines_study_design=0&eq_guidelines_clinical_specialty=0&eq_guidelines_report_section=0&s=machine+learning&btn_submit=Search+Reporting+Guidelines>.

Fliorent, R., B. Fardman, A. Podwojniak, K Javaid, I.J. Tan, H. Ghani et al. 2024. Artificial Intelligence in Dermatology: Advancements and Challenges in Skin of Color. *International Journal of Dermatology* 63(4): 455–61. doi:10.1111/ijd.17076.

Gupta, A. and J. Treviranus. 2022. Inclusively Designed Artificial Intelligence. In H. Schaffers, M. Vartiainen and J. Bus, eds., *Digital Innovation and the Future of Work (1st ed.)*. River Publishers.

International Skin Image Collaboration (ISIC). 2024. Gallery. Retrieved October 24, 2024. <http://gallery.isic-archive.com/>.

Kueper, J.K. and J. Pandit. 2025. Artificial Intelligence for Healthcare in Canada: Contrasting Advances and Challenges. *Healthcare Papers* 22(4): 11–30. doi:10.12927/hcpap.2025.27574.

Nsoesie, E.O. and M. Ghassemi. 2024. Using Labels to Limit AI Misuse in Health. *Nature Computational Science* 4(9): 638–40. doi:10.1038/s43588-024-00676-7.

Partnership on AI (PAI). 2024. About ML. Retrieved October 24, 2024. <https://partnershiponai.org/workstream/about-ml/>.

Tschandl, P. 2021. Risk of Bias and Error From Data Sets Used for Dermatologic Artificial Intelligence. *JAMA Dermatology* 157(11): 1271–73. doi:10.1001/jamadermatol.2021.3128.

Vokinger, K.N., S. Feuerriegel and A.S. Kesselheim. 2021. Mitigating Bias in Machine Learning for Medicine. *Communications Medicine* 1(1): 25. doi:10.1038/s43856-021-00028-w.

Wen, D., S.M. Khan, A.J. Xu, H. Ibrahim, L. Smith, J. Caballero et al. 2022. Characteristics of Publicly Available Skin Cancer Image Datasets: A Systematic Review. *The Lancet Digital Health* 4(1): e64–74. doi:10.1016/S2589-7500(21)00252-1.