Mitchell et al. BMC Ophthalmology (2021) 21:228
https://doi.org/10.1186/s12886-021-01992-6

BMC Ophthalmology

CORRESPONDENCE Open Access

Generalisability through local validation: overcoming barriers due to data disparity in healthcare

William Greig Mitchell[1,2], Edward Christopher Dee[3] and Leo Anthony Celi[2,3,4,5*]

Abstract

Cho et al. report deep learning model accuracy for tilted myopic disc detection in a South Korean population. Here we explore the importance of generalisability of machine learning (ML) in healthcare, and we emphasise that recurrent underrepresentation of data-poor regions may inadvertently perpetuate global health inequity.
Creating meaningful ML systems is contingent on understanding how, when, and why different ML models work in different settings. While we echo the need for the diversification of ML datasets, such a worthy effort would take time and does not obviate uses of presently available datasets if conclusions are validated and re-calibrated for different groups prior to implementation.
The importance of external ML model validation on diverse populations should be highlighted where possible – especially for models built with single-centre data.

Keywords: Machine learning, Disparity, Healthcare equity, Ophthalmology

We read with great interest the article describing the application of deep learning to recognize optic disc tilt, and the discussion of its importance in considering ophthalmic measurements, by Cho et al. [1] The rapid evolution of artificial intelligence in ophthalmic image recognition has created unprecedented opportunities for efficient, accurate and cost-effective diagnosis with less human input – of particular value in resource-poor settings where specialist input is relatively scarcer [2]. While we are encouraged by the model accuracy and commend the authors for describing strengths and weaknesses of their study, we would like to highlight a limitation and subsequent area of further exploration that would strengthen the utility of their work: the need to evaluate the generalisability of their model outside their single-centre, paediatric South Korean population. We believe validation of the model developed by Cho et al. on other populations, particularly those lacking local imaging repositories, would be of great value.

Sociodemographic disparities in machine learning (ML) are well described; with recurrent underrepresentation of some populations posing substantial risks of unknown ML biases. Indeed, a recent report noted that 172 countries (totalling 3.5 billion people) have no publicly available ophthalmic imaging datasets [3], profoundly illuminating the possibility of sampling-bias and subsequent poor generalisability in global ML studies. Such disparity in data availability, if left unchecked, may inadvertently perpetuate global health inequity.

Generalisability is *itself* not binary, nuancing issues of sampling bias in clinical applications of ML [4]. A proxy for validity, generalisability is challenged when translating findings across different clinical settings – if not by demographic diversity (as may be the case for Cho et al), then by unique patient-level differences or local clinical idiosyncrasies. Creating clinically useful ML systems is

* Correspondence: lceli@mit.edu
[2]Harvard TH Chan School of Public Health, Boston, MA, USA
[3]Harvard Medical School, Boston, MA, USA
Full list of author information is available at the end of the article

therefore contingent not only upon demographic and clinical generalisability, but also on an understanding of how, when, and why different ML models work in different settings. Although we echo the need for increased diversification of ML datasets, such a worthy effort would take time. The need for increased diversity does not obviate the uses of presently available datasets if conclusions are conscientiously validated and re-calibrated for different populations prior to implementation.

For example, a recent study from India outlined the value of validating ML models on diverse populations; the model, built with relatively homogenous sociodemographic data, was demonstrated to be more broadly-applicable to other populations in disease detection [5]. Indeed, there are myriad publicly-available ophthalmic imaging datasets, whose algorithms could be broadly validated to assess the extent of their value to populations in data-poor regions, which may lack the infrastructure to develop their own repositories [6–13]. As shown by Gulshan and others, validation of algorithms based on inevitably imperfect data can identify when and where these models still hold clinical value. While models may not be universally generalisable [14], identifying populations in which they *are* accurate – and to what degree, and in what circumstances – still holds importance, particularly in allowing countries lacking the infrastructure to build local imaging datasets to still benefit from international ML findings. Investing in the infrastructure for local validation and re-calibration will also lay groundwork for eventual contribution of local data to international repositories, which may be required to enhance local validity of models.

While there are no hard and fast rules as regards the amount of data needed to validate and re-calibrate a model trained on population A before deploying to population B, the process of validation and re-calibration requires certain steps and features [15–17]. Variables in the original model from population A must be present in the dataset from population B. Data from population B for model validation has to be as recent as possible. A target acceptable discrimination and calibration should be set by those who will use the model and those who will be affected by the model. Special attention should be made in evaluating the accuracy in marginalized groups. If the model performance is below the set threshold, then re-calibration is necessary. In general, the number of patients required should be an order of magnitude greater than the number of features in the model. For images, a principal component analysis is performed to determine which image features are important. Another crucial factor in determining the minimum cohort size is the prevalence of the diagnosis for a classification algorithm or the event for a prediction algorithm. The less prevalent a diagnosis or event is, the larger the sample size required.

Although medicine stands to benefit immensely from publicly-available anonymised data and its applications in artificial intelligence [18], building equitable sociodemographic representation in data repositories is crucial. In the meantime, conscientious local validation and re-calibration will elucidate how and when current ML findings can be applied to heterogeneous populations; and may help to ameliorate disparities in access to ML-driven tools. The importance of model validation on other diverse populations should be emphasised where possible, especially for models built with single-centre data.

Abbreviation

ML: Machine learning

Acknowledgements

Not applicable.

Authors' contributions

WGM, ECD, and LAC contributed equally to the conception, design, analysis and drafting of the manuscript. All authors have read and approve of the manuscript.

Funding

LAC is funded by the National Institute of Health through NIBIB R01 EB017205.

Availability of data and materials

Not applicable.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

Not applicable.

Author details

[1]Department of Ophthalmology, Massachusetts Eye and Ear Infirmary, Boston, MA, USA. [2]Harvard TH Chan School of Public Health, Boston, MA, USA. [3]Harvard Medical School, Boston, MA, USA. [4]Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, USA. [5]Department of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Centre, Boston, MA, USA.

Received: 4 December 2020 Accepted: 14 May 2021
Published online: 21 May 2021

References

1. Cho BH, Lee DY, Park K-A, et al. Computer-aided recognition of myopic tilted optic disc using deep learning algorithms in fundus photography. BMC Ophthalmol. 2020;20(1). https://doi.org/10.1186/s12886-020-01657-w.
2. He M, Li Z, Liu C, Shi D, Tan Z. Deployment of artificial intelligence in real-world practice: opportunity and challenge. Asia Pac J Ophthalmol. 2020;9(4):299–307. https://doi.org/10.1097/APO.0000000000000301.
3. Khan SM, Liu X, Nath S, Korot E, Faes L, Wagner SK, et al. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. The Lancet Digital Health. 2020;3(1):e51–66. https://doi.org/10.1016/s2589-7500(20)30240-5.

4. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. Lancet Digital Health. 2020;2(9):e489–92. https://doi.org/10.1016/s2589-7500(20)30186-2.
5. Gulshan V, Rajan RP, Widner K, Wu D, Wubbels P, Rhodes T, et al. Performance of a deep-learning algorithm vs manual grading for detecting diabetic retinopathy in India. JAMA Ophthalmol. 2019;137(9):987–93. https://doi.org/10.1001/jamaophthalmol.2019.2004.
6. Decencière E, Cazuguel G, Zhang X, Thibault G, Klein J, Meyer F. Teleophta: machine learning and image processing methods for teleophthalmology. IRBM. 2013;34(2):196–203. https://doi.org/10.1016/j.irbm.2013.01.010.
7. Budai A, Bock R, Maier A, Hornegger J, Michelson G. Robust vessel segmentation in fundus images. Int J Biomed Imaging. 2013;(154860). https://pubmed.ncbi.nlm.nih.gov/24416040/.
8. Almazroa A, Alodhayb S, Osman E, et al. Retinal fundus images for glaucoma analysis: the Riga dataset. Med Imag. 2018;(105790). https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10579/2293584/Retinal-fundus-images-for-glaucoma-analysis-the-RIGAdataset/10.1117/12.2293584.short?SSO=1.
9. Zhuo Z, Shou YF, Jiang L, Kee WW, Meng TN, Hai LB. Origa-light: An online retinal fundus image database for glaucoma analysis and research. Ann Int Conf IEEE Eng Med Biol. 2010; Buenos Aires. https://pubmed.ncbi.nlm.nih.gov/21095735/.
10. Sivaswamy J, Krishnadas S, Joshi GD, Jain M, Tabish AS, Drishti G. Retinal image dataset for optic nerve head (onh) segmentation. IEEE 11th Int Symposium Biomed Imaging (ISBI). 2014; Beijing. https://ieeexplore.ieee.org/document/6867807.
11. Niemeijer M, Xiayu X, Dumitrescu A, Gupta P, Bv G, folk J. Automated measurement of the arteriolar-to-venular width ratio in digital color fundus photographs. IEEE Trans on Med Imaging. 2011;30(11):1941–50. https://doi.org/10.1109/TMI.2011.2159619.
12. Al-Diri B, Hunter A, Steel D, Habib M, Hudaib T, Berry S. Review - a reference data set for retinal vessel profiles. 30th annual international conference of the IEEE engineering in medicine and biology society. Vancouver; 2008.
13. Tong Y, Lu W, YY U, Shen Y. Application of machine learning in ophthalmic imaging modalities. Eye Vision. 2020;7(22):1–15.
14. Beede E, Baylor E, Hersch F, et al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. 2020.
15. Liu Y, Chen P, Krause J. How to read articles that use machine learning users' guides to the medical literature. JAMA. 2019;322(18):1806–16. https://doi.org/10.1001/jama.2019.16489.
16. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med. 2018;169(12):866–72. https://doi.org/10.7326/m18-1990.
17. Balki I, Amirabadi A, Levman J, et al. Sample-size determination methodologies for machine learning in medical imaging research: a systematic review. Can Assoc Radiol J. 2020;70(4):344–53.
18. Kras A, Celi LA, Miller JB. Accelerating ophthalmic artificial intelligence research: the role of an open access data repository. Curr Opin Ophthalmol. 2020;31(5):337–50. https://doi.org/10.1097/icu.0000000000000678.

Publisher's Note