Real-World Data Generalization for Glaucoma Prediction

Homa Rashidisabet [1,2], R.V. Paul Chan [1,2], Thasarat S. Vajaranant [1,2], Darvin Yi [1,2]

[1] Illinois Eye and Ear Infirmary, Department of Ophthalmology and Visual Sciences, University of Illinois at Chicago

[2] Artificial Intelligence in Ophthalmology (Ai-O) Center, University of Illinois at Chicago

INTRODUCTION

Despite the rapid growth of AI in ophthalmology research, translations to clinical settings remain challenging. This study aims at showing AI models trained on standardized publicly available datasets perform well while they are not generalizable to Real-World (RW) clinical data to predict glaucoma. In this study, we answer three main questions below.

1. To what extent publicly trained DL models can generalize to RW data?
2. Whether more heterogeneous training data improves generalization?
3. How well DL models can learn from heterogeneous data in a small data regime?

METHODS

DATA

We used Illinois Eye and Ear Infirmary (IEEI) data [1] as an example for RW data. A sample of these images is shown in Figure 1. We also used three public datasets including RIME-ONE-DL, Drishti-GS, and REFUGE. Given the limited available number of glaucoma images (n=362) in the public domain, we randomly sampled the same number of non-glaucoma images to create a balanced dataset. Consistently, we sampled 362 images per class from the IEEI dataset.

FIGURE 1

(a) A sample of heterogeneous IEEI RW data. (b) A sample of standardized public data.

CROPPING FUNDUS IMAGES

Images in RIME-ONE-DL are cropped. We cropped Drishti-GS, REFUGE, and RW images to mimic RIME-ONE-DL images. We used the procedure shown in Figure 2 to crop images.

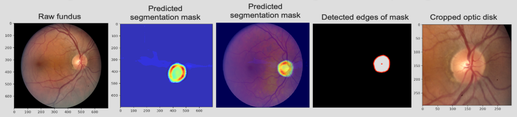


FIGURE 2

The cropping procedure of Optic Disc (OD) from an original fundus image.

GLAUCOMA CLASSIFICATION

We used cropped images as input to a classification network which makes the model effectively learn the low-level features of the optic disc. Cropped images are passed through a series of convolutions, ReLU activation functions, pooling, and fully connected layers. We classified public and RW images as glaucomatous or non-glaucomatous. We used ResNet-50 architecture [2] and Cross Entropy loss with ADAM optimizer. We performed model selection through HP tuning.

EXPERIMENTAL SETUP

We trained two models per train set (e.g., public, RW) and tested each model on both datasets (e.g., public, RW), as shown in Figure 3. Referring to each model as a train-test pair, experimented models include RW-RW, Public-Public, RW-Public, and Public-RW models.

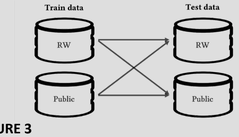


FIGURE 3

Schematic view of our experimental setup for the glaucoma classification task.

RESULTS

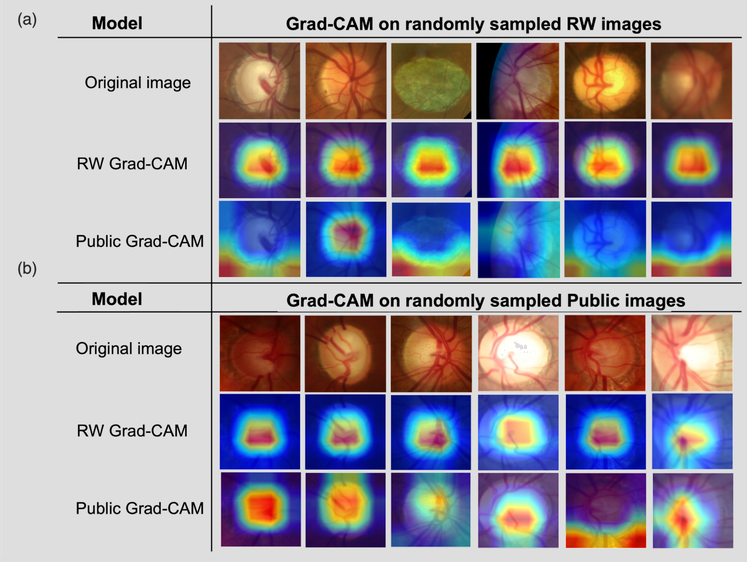
GLAUCOMA CLASSIFICATION RESULTS

Dataset		Evaluation metrics on test set				
Train	Test	Accuracy	Sensitivity	Precision	F_1 score	AUC
Public	Public	**0.88**	**0.81**	**0.81**	**0.81**	**0.95**
RW	RW	0.80	0.75	0.75	0.75	0.86
Public	RW	0.72	0.50	0.72	0.59	0.77
RW	Public	0.80	0.74	0.68	0.71	0.86

TABLE 1

Comparison results of public versus RW trained models for the glaucoma classification task.

Table 1 shows Public-Public model has the highest accuracy for classifying glaucomatous versus non-glaucomatous fundus images. However, this accuracy drops by 16% when the model was tested on RW data. Similarly, other metrics of sensitivity, precision, F_1 score, and AUC dropped by 31%, 9%, 22%, and 18%. However, RW trained model has a stable performance with 80% accuracy on both public and RW test sets for the glaucoma classification task.



In Figure 3, we visualized Grad-CAM [3] predictions. We showed that RW trained model on both public and RW test sets classifies glaucoma based on pixel positions inside the OD region while the public trained model, for the majority of both public and RW test images, classifies glaucoma based on pixel positions outside of OD region. Therefore, we found that similar to glaucoma experts, the RW trained model with a higher generalizability power, consistently predicts glaucoma based on features inside the OD region while the public model, with majorly lower generalizability power, predicts glaucoma based on features outside of the OD region that might only exist in the public data, but not in RW data which might cause overfitting.

FIGURE 4

Top rows in parts (a) and (b) show original input images, the middle and the last rows respectively show Grad-CAM predictions for the last convolutional layer in the last residual block of the RW model and public model.

CONCLUSION

The use of artificial intelligence (AI) techniques is transforming both the clinical and research fields of Ophthalmology. However, one key bottleneck in the deployment of AI-based decision-making tools in medicine is the failure of these models to generalize when deployed across variations in image acquisition protocols, devices, patient populations, and hospitals. We answered the three research questions of this work as follows.

1. We found that most results reported in the past literature for glaucoma prediction on public data are not generalizable to RW data. We verified on three public datasets that DL models trained on public data result in a 16% accuracy drop when tested on RW data.
2. We found that heterogeneous training data can substantially improve the performance of a DL model. We showed that models trained on heterogenous RW data, not only have stable performance across datasets but also their prediction accuracy increases when tested on public data.
3. We achieved comparable results to the past literature (accuracy = 80%) for glaucoma classification while we used heterogenous RW data rather than standardized public data.

REFERENCES

[1] Mojab N, Noroozi V, Aleem A, Nallabothula MP, Baker J, Azar DT, Rosenblatt M, Chan RV, Yi D, Yu PS, Hallak JA. **I-ODA, Real-World Multi-modal Longitudinal Data for Ophthalmic Applications**, arXiv preprint arXiv:2104.02609. 2021.

[2] He, K., Zhang, X., Ren, S. and Sun, J. **Deep residual learning for image recognition**, IEEE conference on computer vision and pattern recognition, 2016.

[3] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra. **Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization**, Int. J. Comput. Vis, 2020.

FINANCIAL DISCLOSURES

Homa Rashidisabet, R.V. Paul Chan, Thasarat Vajaranant, and Darvin Yi report no financial disclosure.

SUPPORT

All work done for this project was supported by the Research to Prevent Blindness (RPB) Foundation.

Supported By Research to Prevent Blindness