Human versus Machine: Comparing a Deep Learning Algorithm to Human Gradings for Detecting Glaucoma on Fundus Photographs

Jammal AA (1) , Thompson AC (2) , Mariottoni EB (2) , Berchuck SI (3) , Urata CN (2) , Estrela T (2) , Wakil SM (2) , Costa VP (4) , Medeiros FA (5)

1 Vision, Imaging and Performance Laboratory (VIP) , Duke Eye Center and Department of Ophthalmology, Duke University, Durham, NC; Department of Ophthalmology, State University of Campinas, Campinas, Brazil.
2 Vision, Imaging and Performance Laboratory (VIP) , Duke Eye Center and Department of Ophthalmology, Duke University, Durham, NC.
3 Vision, Imaging and Performance Laboratory (VIP) , Duke Eye Center and Department of Ophthalmology, Duke University, Durham, NC; Department of Statistical Science and Forge, Duke University, Durham, NC.
4 Department of Ophthalmology, State University of Campinas, Campinas, Brazil.
5 Vision, Imaging and Performance Laboratory (VIP) , Duke Eye Center and Department of Ophthalmology, Duke University, Durham, NC. Electronic address: felipe.medeiros@duke.edu.

PURPOSE: To compare the diagnostic performance of human gradings versus predictions provided by a machine-to-machine (M2M) deep learning (DL) algorithm trained to quantify retinal nerve fiber layer (RNFL) damage on fundus photographs.

DESIGN: Evaluation of a machine learning algorithm.

METHODS: A M2M DL algorithm trained with RNFL thickness parameters from spectral-domain optical coherence tomography was applied to a subset of 490 fundus photos of 490 eyes of 370 subjects graded by two glaucoma specialists for the probability of glaucomatous optical neuropathy (GON) , and estimates of cup-to-disc (C/D) ratios. Spearman correlations with standard automated perimetry (SAP) global indices were compared between the human gradings versus the M2M DL-predicted RNFL thickness values. The area under the receiver operating characteristic curves (AUC) and partial AUC for the region of clinically meaningful specificity (85-100%) were used to compare the ability of each output to discriminate eyes with repeatable glaucomatous SAP defects versus eyes with normal fields.

RESULTS: The M2M DL-predicted RNFL thickness had a significantly stronger absolute correlation with SAP mean deviation (rho=0.54) than the probability of GON given by human graders (rho=0.48; PCONCLUSION: A M2M DL algorithm performed as well as, if not better than, human graders at detecting eyes with repeatable glaucomatous visual field loss. This DL algorithm could potentially replace human graders in population screening efforts for glaucoma.

Am J Ophthalmol. 2019 Nov 12. pii: S0002-9394(19) 30543-4. doi: 10.1016/j.ajo.2019.11.006.

http://www.ncbi.nlm.nih.gov/pubmed/31730838