Evaluating the influence of data source and labelling variability on a Deep Learning (DL) based OCT B-scan abnormality classification

Ganesh Babu T C, MD[1], Sandipan Chakroborty, PhD[1], Krunalkumar Ramanbhai Patel, ME[1], Niranchana Manivannan, PhD[2], Mary Durbin, PhD[2] and Alexander Freytag, PhD[3]

[1]Center for Applications and Research in India, Carl Zeiss India (Bangalore) Pvt. Ltd.; [2]Carl Zeiss Meditec, Inc., Dublin, CA, USA; [3]Corporate Research and Technology, Jena, Carl Zeiss AG, Jena, Thuringia, Germany

Poster # 3531779

PURPOSE

- The accuracy of a DL based algorithm for detecting abnormalities in OCT B-scans is **dependent on quality of training data and its labels**.
- The sources of data could be either from a clinical study or some eye hospitals with high patient load.
- In contrast to clinical studies, factors such as image quality, disease prevalence, age of subjects, etc. are **not easily controllable** when collecting data from eye hospitals.
- The involvement of **multiple labelers may further introduce inconsistencies while creating labels**, because each expert has their own clinical judgment and differs depending on how and where (Primary, Secondary or Tertiary clinics) they practice.
- We **evaluate the effects** of above two aspects on OCT abnormality classification performance.

METHODS

- To assess the effect of the data collection scope, we gathered CIRRUS™ 4000/5000 and PRIMUS 200 macular OCT cube data during i) clinical studies and ii) from eye hospitals.
- To measure the effect of labelling variability, the data was labelled at B-scan level by five retina specialists. Among five labelers, two of them (labelers X & Y in Table 2) had similar expertise levels practicing in the same hospital, while the rest (labelers A, B & C) were from three different eye clinics. Each cube has been seen by single pair of eyes.
- Data from each labeler was split into training and test sets.
- Inception_V1 model was developed on the training set for each labeler, and the performance was evaluated on all test sets. Table 1 and 2 show the number of samples used for trainings and evaluations.

Data Source	# Train Sample	# Test Sample
Controlled	28,544	7,040
Non-controlled	72,512	18,432

Table 1. Number of OCT B-scans in train and test split of dataset

Data Source	# Train Sample	# Test Sample
Labeler A	14,199	3,537
Labeler B	31,293	7,440
Labeler C	42,132	10,219
Labeler X	61,121	15,360
Labeler Y	61,088	15,349

Table 2. Dataset wise number of B-scans in train and test split

Email: ganesh.babu@zeiss.com; sandipan.chakroborty@zeiss.com; Krunal.patel@zeiss.com

Disclosures: GB(E), SC(E), KP(E), NM(E), MD(E), Carl Zeiss Meditech, Inc.

RESULTS

- Fig 1 shows that evaluating **a model of data from clinical trials does not always indicate good generalizability** and may over-estimate the model accuracy.
- Training on 'uncontrolled' data sources leads overall to improved performance in a typical clinical setting, even if such a model underperforms in the clinical trial setting.

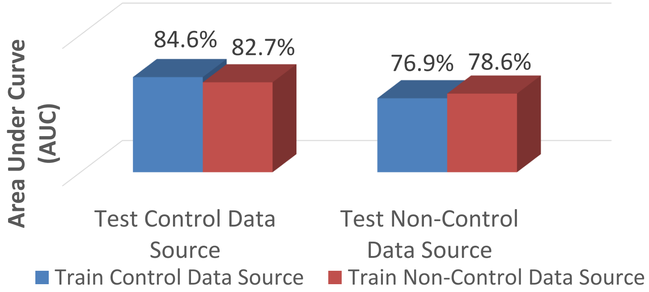

Fig 1. Data source wise AUC performance evaluation

- From Fig 2, we observe that models perform well on data labelled by experts with similar background. It shows **strong differences in accuracy for abnormality prediction with labelers from different backgrounds**, showing significant AUC drop.

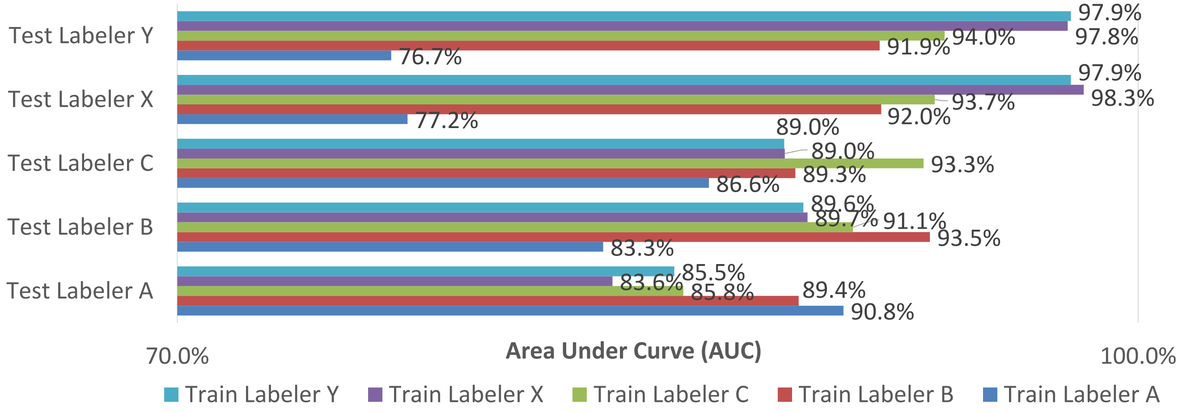

Fig 2. Data source wise AUC performance evaluation

CONCLUSIONS

We conclude that prediction models are not easily transferable across labelers from different backgrounds. Furthermore, model accuracy from clinical trial model may not be transferrable to busy clinical environments.