Using Artificial Intelligence Reading Label System in Diabetic Retinopathy Grading Training of Junior Ophthalmology Residents and Medical Students

Ruoan Han
Peking Union Medical College Hospital, Chinese Academy of Medical Sciences
Weihong Yu
Peking Union Medical College Hospital, Chinese Academy of Medical Sciences
Huan Chen
Peking Union Medical College Hospital, Chinese Academy of Medical Sciences
Youxin Chen (chenyx@pumch.cn)
Peking Union Medical College Hospital, Chinese Academy of Medical Sciences

Research Article

Keywords: diabetic retinopathy, artificial intelligence, grading training

Posted Date: July 22nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-671168/v1

Version of Record: A version of this preprint was published at BMC Medical Education on April 9th, 2022.
See the published version at https://doi.org/10.1186/s12909-022-03272-3.

Abstract

Purpose

Evaluate the efficiency of using artificial intelligence reading label system in diabetic retinopathy grading training of junior ophthalmology resident doctors and medical students.

Methods

Loading 520 diabetic retinopathy patients' color fundus images in the artificial intelligence reading label system. 13 participants (including 6 junior ophthalmology residents and 7 medical students) read the images randomly for 8 rounds. They evaluated the grading of images and labeled the typical lesions. The sensitivity, specificity and kappa score were determined by comparison with the participants' results and expert golden standards.

Results

Through 8 round reading, average kappa score was elevated from 0.67 to 0.81. Average kappa score of round 1 to 4 was 0.77, and average kappa score of round 5 to 8 was 0.81. The participant was divided into two groups. Participants in group 1 were junior ophthalmology resident students and participants in group 2 were medical doctors. Average kappa score of group 1 was elevated from 0.71 to 0.76. Average kappa score of group 2 was elevated from 0.63 to 0.84.

Conclusion

The artificial intelligence reading label system was a useful tool in training resident doctors and medical students in doing diabetic retinopathy grading.

Summary Statement

This article evaluate the efficiency of using artificial intelligence reading label system in diabetic retinopathy grading training of junior ophthalmology resident doctors and medical students. Through the reading training, the kappa score of DR grading elevated. It showed out that the artificial intelligence reading label system was a useful tool in training resident doctors and medical students in doing diabetic retinopathy grading.

Introduction

Diabetic retinopathy (DR) is the most common microvascular complication of diabetes and the leading cause of irreversible vision loss in adults of working age [1]. Early diagnosis and treatment of DR can

cause timely medical intervention thus preventing progression of the disease and avoid the occurrence of severe visual impairment. [2, 3] Therefor, it is important to accurately screen and grade the disease. In China's existing medical system, qualified specialists of fundus diseases are in a serious shortage. Long-term medical education and fundus disease's professional training are needed to meet the requirement of DR grading. This makes it difficult for primary hospitals to manage the screening of DR. Most patients went to medical institutions only if they had obvious symptoms, resulting in a more serious condition when diagnosed.

Recent years, because of the rapid development of artificial intelligence (AI) techniques, the AI technique based on machine leaning plays an important role in DR screening, which acquired high sensitivity and specificity through the learning of large number of fundus photo training data sets. But the fundus photo training data sets needed manual annotation by qualified specialist and the AI reading results also needed confirmed by retina experts. Thus, in order to train the junior ophthalmologists of DR reading and AI data set annotation, DR reading training is a very important work in ophthalmology residency training. The purpose of this study was to evaluate the efficiency of using artificial intelligence reading labeling system in diabetic retinopathy grading training of junior ophthalmology resident doctors and medical students.

Methods

Reading methods

A total of 520 fundus photographs centered on the macular region were included in this study. Photographs were randomly divided into 8 groups, 65 images for each group. The severity of diabetic retinopathy was graded based on the international clinical diabetic retinopathy severity scale.[4] Photographs of no DR, mild non-proliferative DR (NPDR), moderate NPDR, severe NPDR and proliferative DR (PDR) were included of each group. Three senior consultants made the diagnosis golden standard of each image. 13 junior ophthalmology residents and medical students participated in the training. Six of them were the first year residents of ophthalmology residency training programme in Peking Union Medical College Hospital (PUMCH). Seven of them were medical students of Peking Union Medical College (PUMC). 13 participants did DR reading using AI reading labeling system, made DR grading and labeled classic lesions of each image. Reading training was performed for 8 rounds with 65 images per round. After each round's labeling, the participants were gathered to study the diagnosis golden standard. Each round lasted for one week and the whole process lasted for 8 weeks. The sensitivity, specificity and mean (SD) values of kappa score according to the diagnosis golden standard were summarized after each round.

Grading Methods

Fundus photographs were divided to 5 levels according to the DR severe degrees. No DR, mild NPDR, moderate NPDR, severe NPDR and PDR were labeled as degree 0, 1, 2, 3 and 4 separately. Degree 0 defined as “without DR” and degree 1 to 4 defined as “with DR”. Degree 0 and 1 defined as “Non referral DR”, while degree 2 to 4 defined as “referral DR”. Degree 0 to 2 defined as “Non severe DR”, while degree 3 and 4 defined as “severe DR”.

Statistical Methods

Diagnosis results were collected according to the diagnosis golden standard and analyzed statistically using SPSS 25 (IBM, NY, USA). The sensitivity and specificity of different grading were calculated. The kappa score was used to determine the agreement between the diagnosis and participants’ results. Kappa score 0.61 to 0.80 was determined to be significant consistency, while kappa score above 0.80 was determined to be highly consistency. The discrepancy of kappa score before and after training was compared to evaluate the effect of DR reading training.

Results

Training result of all the participates

In the DR reading training, the average of harmonic mean and Kappa score of each diagnosis group were shown in Table 1. Through the eight round of reading, the average of kappa score was elevated from 0.67 to 0.81. The average kappa score of first 4 rounds was 0.77, which means significant consistency. The average kappa score of latter 4 rounds was elevated to 0.81, which means highly consistency. There was an escalating trend of diagnosis accuracy. The growth curve of reading training was shown as Fig. 1.

The harmonic mean of “with or without DR” was elevated from 0.55 to 0.73, and the harmonic mean of “non referral or referral DR” was elevated from 0.76 to 0.81. “Non severe or severe DR” group got a harmonic mean of 0.75 to 0.85.

Table 1
The average values of harmonic mean and Kappa score.

	1	2	3	4	5	6	7	8
With or without DR	0.55	0.64	0.71	0.65	0.79	0.74	0.82	0.73
Non referral or referral DR	0.76	0.79	0.84	0.75	0.85	0.86	0.85	0.81
Non severe or severe DR	0.75	0.76	0.88	0.86	0.79	0.89	0.84	0.85
Kappa Score	0.67	0.72	0.80	0.76	0.78	0.83	0.83	0.81

Training Result Of Each Group

13 participants were divided into two groups. Group 1 consisted of junior ophthalmology residents who got basic knowledge of ophthalmology. Group 2 consisted of medical students who didn't have ophthalmology knowledge basis. The average kappa score of each group was calculated separately. As shown in Table 2. After eight round of reading, the average kappa score of group 1 was elevated from 0.71 to 0.76. The average kappa score of group 2 was elevated from 0.63 to 0.84. Figure 2 and Fig. 3 showed the growth curves according to the kappa score of the two groups.

Table 2
Average kappa scores of two groups

	1	2	3	4	5	6	7	8
Group1	0.71	0.72	0.86	0.77	0.83	0.81	0.82	0.76
Group2	0.63	0.71	0.74	0.76	0.75	0.84	0.84	0.83

Discussion

In recent years, artificial intelligence technology based on classic machine learning (ML) or deep learning (DL) has been widely used in a variety of fundus disease screening including DR. Gulshan et al. used the deep learning algorithm for the screening of DR and obtained extremely high sensitivity and specificity. [5] Takahashi et al. used a modified deep learning algorithm model for the screening and grading of DR, which can obtain grading results similar to those of ophthalmologists. [6] However, even if the application of artificial intelligence technology in DR screening and grading has achieved very high accuracy, the final results can only be used as a diagnostic reference. Training junior ophthalmologists to grow rapidly and perform DR reading accurately is still an important part of ophthalmologist training. If junior ophthalmologists can master the DR reading method through centralized training quickly, it is not only conducive to the growth of ophthalmologists, but also reserves the strength of physicians for labeling AI training set. Therefore, it is of great significance to find an efficient DR reading training method. There is no previous discussion on the standard way of DR reading training, and there is no literature exploring the use of AI reading labeling system for reading training and learning. In this study, the AI reading labeling system was used for DR reading training of junior ophthalmology residents and medical students. Compared with the traditional reading training requiring at least 1,500 to 2,000 pictures, This training using AI labeling system require only 500 pictures to obtain a high diagnostic accuracy. This training showed very obvious advantages in terms of time and number of pictures compared with the traditional reading training.

In this DR reading training, after 8 round of reading, the mean Kappa score value of 13 participants increased from 0.67 in the first reading to 0.81 in the eighth reading, the mean Kappa score value of the first 4 rounds was 0.77, indicating significant agreement, and the mean Kappa score value of the last 4 rounds was 0.81, indicating that after training, the overall reading accuracy of participants was significantly improved. The Kappa score value is not linearly increased each time, which may be due to

the fact that the difficulty level cannot be completely consistent with the picture loaded in each time, resulting in the bias of the results.

At the same time, the trainees were also divided into two groups for statistics. The first group was junior ophthalmology residents with certain basic knowledge of ophthalmology, who also attended ophthalmology course and participated in the clinical work of ophthalmology. The second group was medical students who had not learn ophthalmology basic knowledge before the start of reading training, and had not participated in the course and clinical work of ophthalmology after the start of reading. The initial Kappa score of the two groups reflected the difference in the knowledge base of the two groups of readers, with an initial Kappa score of 0.71 in group 1 and 0.62 in group 2, reflecting that the accuracy of the basic reading was higher in group 1 than in group 2. As the training progressed, the difference between the two groups gradually narrowed, and the Kappa scores increased to 0.76 in group 1 and 0.84 in group 2 for the eighth reading, with a more significant increase in the medical student group. The mean Kappa score of the first four rounds was 0.77, the mean Kappa score of the last four rounds was 0.81 in group 1, 0.71 in the first four rounds and 0.82 in the last four rounds in group 2, which also reflected that the gap in reading accuracy between the two groups was reduced, and after reading training, even medical students without an ophthalmological knowledge base could be familiar with the law of DR reading and achieve a certain diagnostic accuracy.

The results of harmonic mean value of presence or absence of DR, referral of DR and severe DR showed that the harmonic mean value of determination of presence or absence of DR was the lowest, and the harmonic mean value of referral of DR and severe DR was relatively higher, which may be because it may not be very accurate for the presence or absence of microhemangioma based on fundus color photography alone. The small microhemangioma in the picture may be confused with poor quality artifacts at the time of photography, leading to incorrect conclusions. This also suggests that for the reading training, we should be cautious in selecting the fundus photographs used for the training, try to select the pictures with good quality, and eliminate the possible confounding factors caused by the poor quality of picture shooting.

This study also has some limitations. Since the original application of the reader labeling system used in the training is to train the AI deep learning model, which is not used for the reading training of physicians, the system cannot immediately give the correct grading answer after labeling, and needs to uniformly conduct the retrospective learning of picture grading after each labeling, which has an effect on the reading learning efficiency. In addition, the number of people included in the training was small, and there may be some error in the statistical mean. In order to make this system more conducive to reading training, the AI reading result prompt function can be added, and the gold standard is given after each round of picture labeling for comparison, which can make the training efficiency and strengthen the effect of reading training.

In conclusion, the use of artificial intelligence DR reading labeling system can effectively improve the DR reading level of junior ophthalmologists, and can achieve a certain reading accuracy in a short time and

using less reading volume, which is a feasible reading training method.

Declarations

Ethics approval and consent to participate: Our manuscripts was conducted in accordance with the Declaration of Helsinki and approved by ethics committee board of Peking Union Medical College Hospital.

Consent for publication: All participants provided written informed consent.

Availability of data and materials: Data are available on reasonable request.

Competing interests: The authors declare that they have no competing interests.

Funding: This work was funded by Non-profit Central Research Institute Fund of Chinese Academy of Medical Sciences (2018PT32029), CAMS Initiative for Innovative Medicine (CAMS-I2M, 2018-I2M-AI-001), Pharmaceutical collaborative innovation research project of Beijing Science and Technology Commission (Z191100007719002), National Key Research and Development Project (SQ2018YFC200148), Beijing Natural Science Foundation Haidian original innovation joint fund (19L2062).

The funding organizations had no role in design or conduct of this research.

Authors' contributions: YC contributed to the conception of the study. RH organized the training, wrote the main manuscript text and performed the data analyses. WY supervised the training and performed the analysis with constructive discussions. HC performed the data analyses and contributed to the manuscript preparation. All authors reviewed the manuscript.

References

1. Yau JW, Rogers SL, Kawasaki R, et al. Global prevalence and major risk factors of diabetic retinopathy. Diabetes Care 2012; 35: 556–564.
2. Cheung N, Mitchell P, Wong TY. Diabetic retinopathy. Lancet 2010; 376: 124–136.
3. Wong TY, Cheung CM, Larsen M, et al. Diabetic retinopathy. Nat Rev Dis Primers 2016; 2: 16012.
4. Wilkinson CP, Ferris FL 3rd, Klein RE, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology 2003; 110: 1677.
5. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016; 316: 2402–2410.
6. Takahashi H, Tampo H, Arai Y, et al. Applying artificial intelligence to disease staging: Deep learning for improved staging of diabetic retinopathy, PLoS One.

Figures

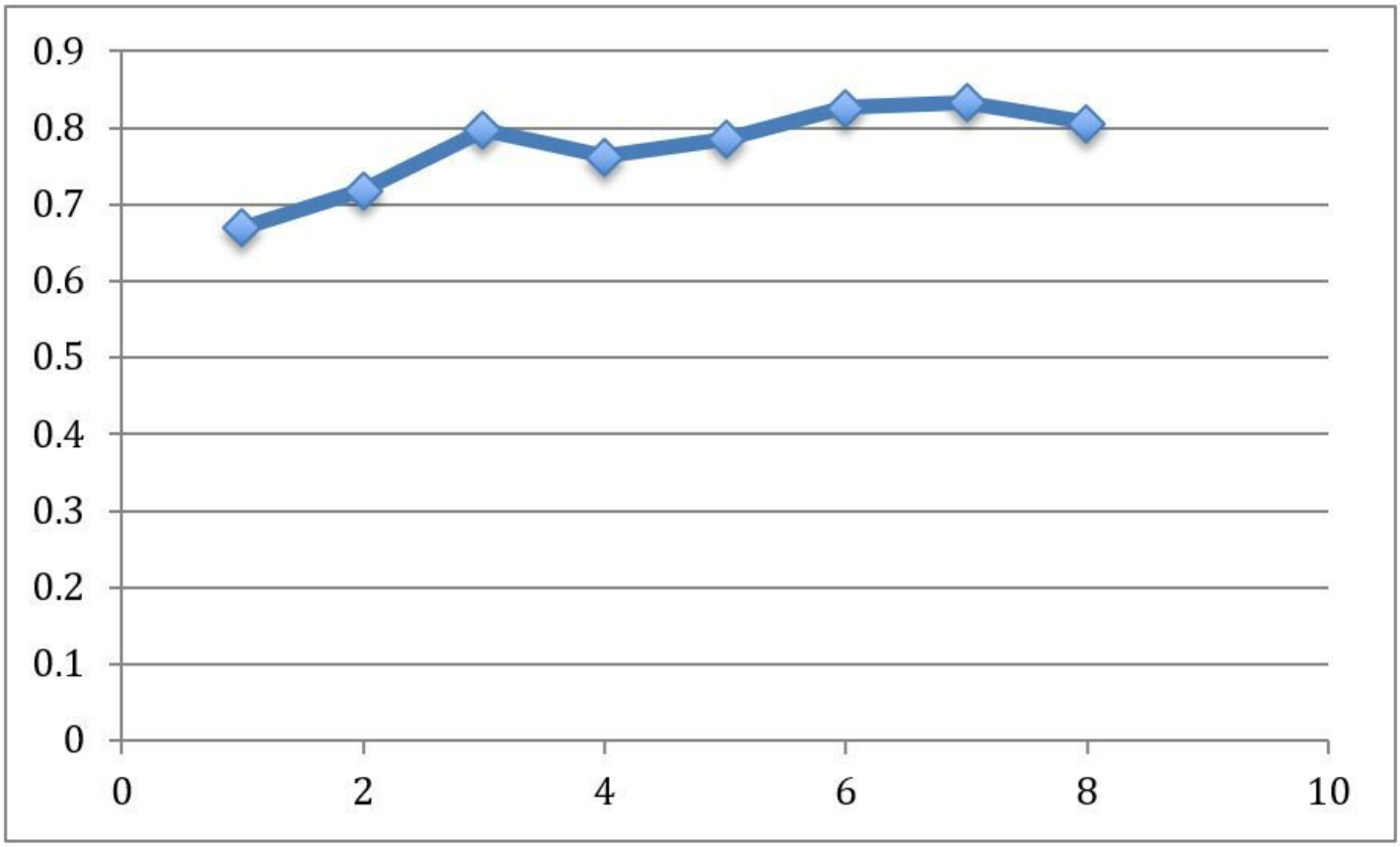

Figure 1

Growth curve of the mean number of kappa scores for each reading by the 13 readers training participants. The abscissa is the number of training sessions and the ordinate is the Kappa score.

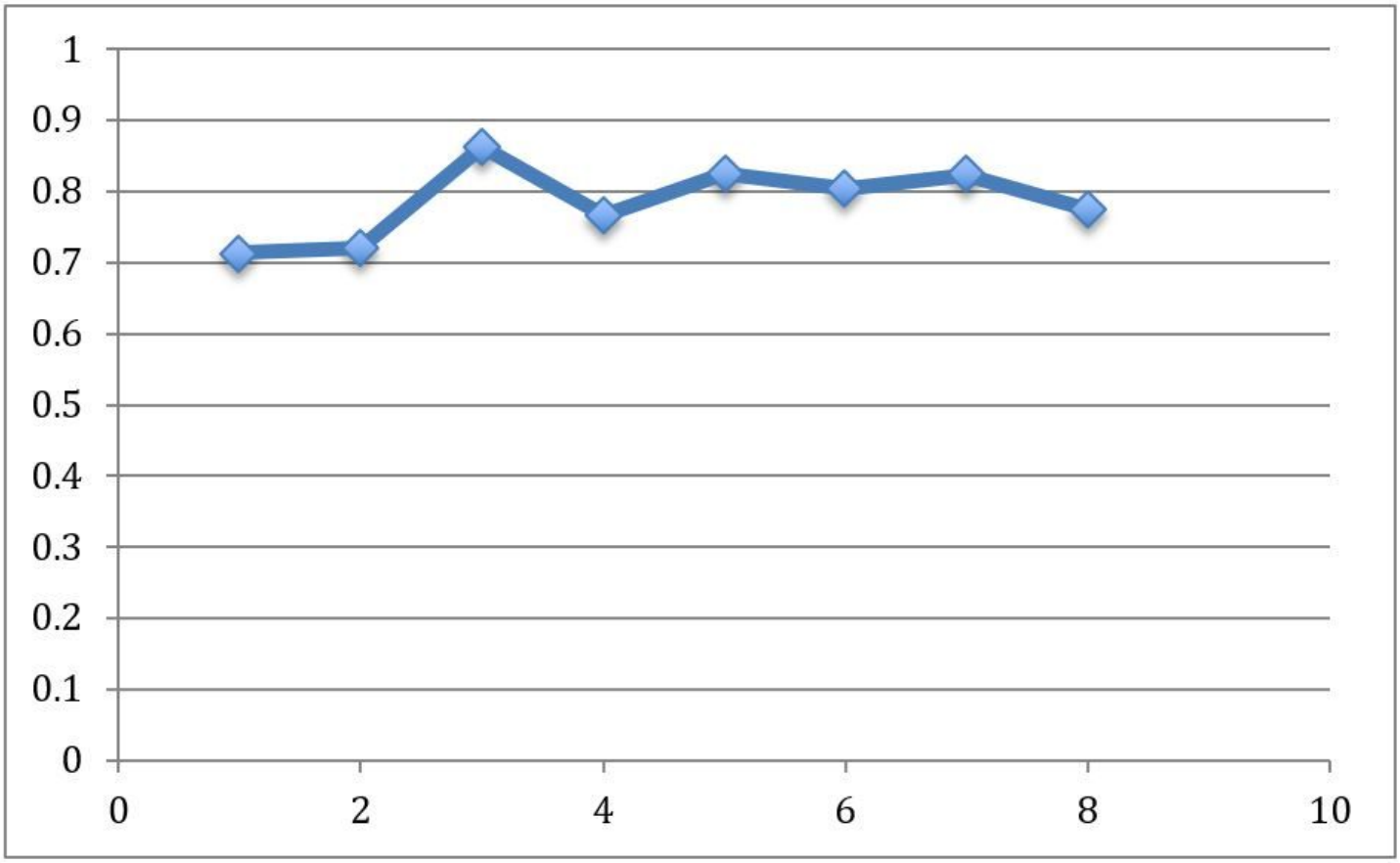

Figure 2

Growth curve of the mean number of Kappa scores for group 1 (junior ophthalmology residents). The abscissa is the number of training sessions and the ordinate is the Kappa score.

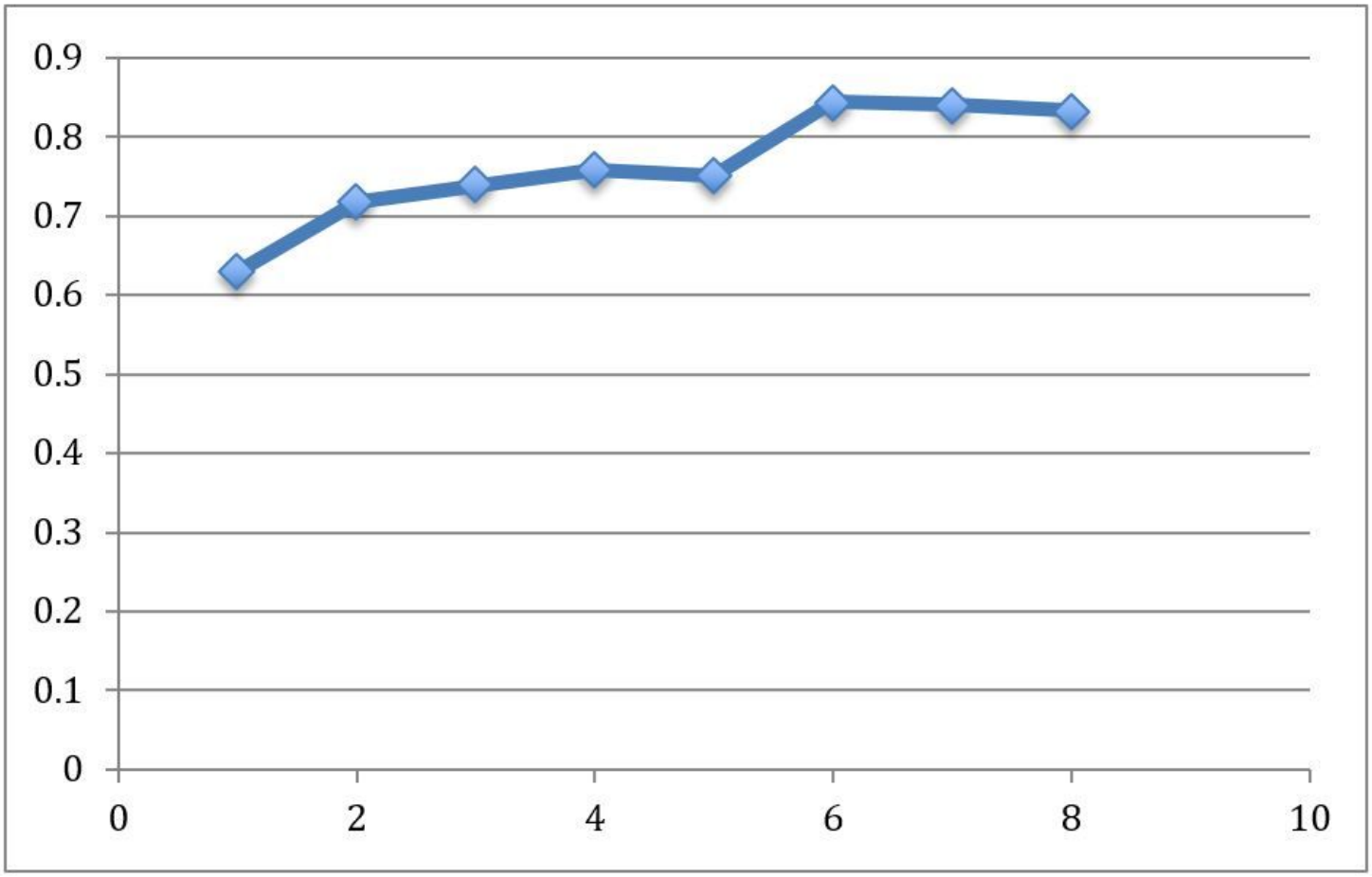

Figure 3

Growth curve of the mean number of Kappa scores for group 2 (medical students). The abscissa is the number of training sessions and the ordinate is the Kappa score.