EClinicalMedicine 9 (2019) 52–59

Contents lists available at ScienceDirect

EClinicalMedicine

journal homepage: https://www.journals.elsevier.com/eclinicalmedicine

EClinicalMedicine
Published by THE LANCET

Research Paper

Diagnostic Efficacy and Therapeutic Decision-making Capacity of an Artificial Intelligence Platform for Childhood Cataracts in Eye Clinics: A Multicentre Randomized Controlled Trial

Haotian Lin [a,*,1], Ruiyang Li [a,1], Zhenzhen Liu [a,1], Jingjing Chen [a,1], Yahan Yang [a,1], Hui Chen [a,1], Zhuoling Lin [a], Weiyi Lai [a], Erping Long [a], Xiaohang Wu [a], Duoru Lin [a], Yi Zhu [a,b], Chuan Chen [a,b], Dongxuan Wu [c], Tongyong Yu [c], Qianzhong Cao [a], Xiaoyan Li [a], Jing Li [a], Wangting Li [a], Jinghui Wang [a], Mingmin Yang [d], Huiling Hu [d], Li Zhang [e], Yang Yu [f], Xuelan Chen [f], Jianmin Hu [f], Ke Zhu [g], Shuhong Jiang [h], Yalin Huang [i], Gang Tan [j], Jialing Huang [k], Xiaoming Lin [a], Xinyu Zhang [a], Lixia Luo [a], Yuhua Liu [a], Xialin Liu [a], Bing Cheng [a], Danying Zheng [a], Mingxing Wu [a], Weirong Chen [a], Yizhi Liu [a,*]

[a] State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, Guangdong 510060, China
[b] Department of Molecular and Cellular Pharmacology, University of Miami Miller School of Medicine, Miami, FL 33136, USA
[c] Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, Guangdong 510060, China
[d] Shenzhen Eye Hospital, Shenzhen Key Ophthalmic Laboratory, The Second Affiliated Hospital of Jinan University, Shenzhen, Guangdong 518040, China
[e] Department of Ophthalmology, The Central Hospital of Wuhan, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 430014, China
[f] Department of Ophthalmology, The Second Affiliated Hospital of Fujian Medical University, Quanzhou, Fujian 362000, China
[g] Kaifeng Eye Hospital, Kaifeng, Henan 475000, China
[h] Inner Mongolia People's Hospital, Hohhot, Inner Mongolia 010017, China
[i] Henan Eye Institute, Henan Eye Hospital, Henan Provincial People's Hospital and People's Hospital of Zhengzhou University, Zhengzhou, Henan 450003, China
[j] The First Affiliated Hospital of the University of South China, Hengyang, Hunan 421001, China
[k] School of Public Health, Sun Yat-sen University, Guangzhou, Guangdong 510080, China

ARTICLE INFO

Article history:
Received 13 September 2018
Received in revised form 12 February 2019
Accepted 3 March 2019
Available online 17 March 2019

Keywords:
Artificial intelligence
Childhood cataracts
Multicentre randomized controlled trial
Ophthalmology

ABSTRACT

Background: CC-Cruiser is an artificial intelligence (AI) platform developed for diagnosing childhood cataracts and providing risk stratification and treatment recommendations. The high accuracy of CC-Cruiser was previously validated using specific datasets. The objective of this study was to compare the diagnostic efficacy and treatment decision-making capacity between CC-Cruiser and ophthalmologists in real-world clinical settings.
Methods: This multicentre randomized controlled trial was performed in five ophthalmic clinics in different areas across China. Pediatric patients (aged ≤ 14 years) without a definitive diagnosis of cataracts or history of previous eye surgery were randomized (1:1) to receive a diagnosis and treatment recommendation from either CC-Cruiser or senior consultants (with over 5 years of clinical experience in pediatric ophthalmology). The experts who provided a gold standard diagnosis, and the investigators who performed slit-lamp photography and data analysis were blinded to the group assignments. The primary outcome was the diagnostic performance for childhood cataracts with reference to cataract experts' standards. The secondary outcomes included the evaluation of disease severity and treatment determination, the time required for the diagnosis, and patient satisfaction, which was determined by the mean rating. This trial is registered with ClinicalTrials.gov (NCT03240848).
Findings: Between August 9, 2017 and May 25, 2018, 350 participants (700 eyes) were randomly assigned for diagnosis by CC-Cruiser (350 eyes) or senior consultants (350 eyes). The accuracies of cataract diagnosis and treatment determination were 87.4% and 70.8%, respectively, for CC-Cruiser, which were significantly lower than 99.1% and 96.7%, respectively, for senior consultants ($p < 0.001$, OR = 0.06 [95% CI 0.02 to 0.19]; and $p < 0.001$, OR = 0.08 [95% CI 0.03 to 0.25], respectively). The mean time for receiving a diagnosis from CC-Cruiser was 2.79 min, which was significantly less than 8.53 min for senior consultants ($p < 0.001$, mean difference 5.74 [95% CI 5.43 to 6.05]). The patients were satisfied with the overall medical service quality provided by CC-Cruiser, typically with its time-saving feature in cataract diagnosis.
Interpretation: CC-Cruiser exhibited less accurate performance comparing to senior consultants in diagnosing childhood cataracts and making treatment decisions. However, the medical service provided by CC-Cruiser

* Corresponding authors at: Zhongshan Ophthalmic Center, Sun Yat-Sen University, Xian Lie South Road 54#, Guangzhou 510060, China.
E-mail addresses: haot.lin@hotmail.com (H. Lin), yizhi_liu@aliyun.com (Y. Liu).
[1] These authors contributed equally to the work.

https://doi.org/10.1016/j.eclinm.2019.03.001

was less time-consuming and achieved a high level of patient satisfaction. CC-Cruiser has the capacity to assist human doctors in clinical practice in its current state.
Funding: National Key R&D Program of China (2018YFC0116500) and the Key Research Plan for the National Natural Science Foundation of China in Cultivation Project (91846109).

1. Introduction

The implementation of artificial intelligence (AI), including robotic surgery, medical imaging, and automated diagnosis, has become increasingly popular in modern medical industry [1–4]. For example, IBM-Watson, a question-answering computer system, can provide diagnosis and treatment suggestions for breast cancer [5]. An AI generated through deep convolutional neural network (CNN) algorithms can effectively screen skin disease and classify skin cancer as dermatologists [6]. Medical AI has significant advantages in continuous lifelong learning from human experts, convenient open-source sharing, and efficient decision-making [4,7]. The traditional medical care service modality has limited capacity for providing high-quality healthcare to large populations, as experienced clinicians require extensive training [8–10]. In contrast, medical AI can imitate and replace the primary work of human doctors through deep learning, and provide medical guidance to multiple hospitals simultaneously, especially in those less-developed and remote areas [11,12]. Therefore, advances in medical AI are expected to provide high-quality medical services and alleviate the uneven distribution of medical resources [13,14].

Research in context

Evidence before this study

Advances in medical AI are expected to provide high-quality medical services and alleviate the shortage of medical resources. We searched PubMed for clinical trials published in all language between Jan 1, 2000, and Dec 20, 2018, using the search terms "artificial intelligence", "diagnose/diagnosis", and "treatment". We also searched the reference lists of the retrieved articles. Previous studies of medical AI, such as diagnosis of skin diseases, breast cancer, retinopathy, cataract, glaucoma and cardiovascular diseases, have mainly focused on machine learning by specific screened datasets or observational study to assess the AI performance. However, our scientific literature review found that all the available evidence had not evaluated the efficacy of diagnostic medical AI using large-scale unscreened clinical data in a real-world-comparative trial.

Added value of this study

To the best of our knowledge, this study is the first multicentre randomized controlled trial to compare the diagnostic accuracy and efficiency of medical AI to that of senior consultants in real-world clinical settings. The results of this study suggest that the accuracy of diagnosis and treatment-decision making of medical AI is lower than that of senior consultants. However, our medical AI requires less time for diagnosis and still achieves a high level of patient satisfaction in eye clinics.

Implications of all the available evidence

The study demonstrates that medical AI has the capacity to assist human doctors in clinical practice. However, the real-world diagnostic performance of all medical AI must be evaluated in clinical controlled trials before regular clinical application.

Previous studies on the application of medical AI, such as detecting diabetic retinopathy, macular degeneration, glaucoma, and cardiovascular diseases, mainly focused on machine learning by screening images collected from specific datasets [15–19]. However, the efficacy of medical AI in disease diagnosis and therapeutic decision-making has not been evaluated using large-scale unfiltered clinical data in a real-world-comparative trial. At present, inaccurate diagnoses and inappropriate treatment decisions are common, especially among patients with rare diseases, mainly due to insufficient medical resources in non-specialized hospitals [20–22]. Childhood cataract is a rare disease that can cause irreversible vision loss without urgent early diagnosis and treatment [23,24]. Moreover, the diagnosis and treatment of cataract is mainly based on morphology and AI has showed great advantages in image recognition [1]. Thus, childhood cataract is a suitable test case for the exploration of a medical AI.

CC-Cruiser is an ophthalmic AI platform developed by Zhongshan Ophthalmic Center (ZOC) for diagnosing childhood cataracts and providing risk stratification and treatment guidance [25]. This collaborative cloud platform enables patient data sharing between individual hospitals for data integration and patient screening. CC-Cruiser was trained from a dataset including 410 ocular images of childhood cataracts and 476 images of normal eyes from the Childhood Cataract Program of the Chinese Ministry of Health (CCPMOH), a specialized care centre for rare diseases in China. In addition, the high accuracy of CC-Cruiser was previously validated in an in silico test, a website-based study using eye images from websites, a 'finding a needle in a haystack' test (a test using a dataset with a normal lens: cataract ratio of 100:1), and a small clinical trial [25]. Here, we performed a multicentre diagnostic randomized controlled trial [26], which is the final frontier to evaluate the clinical difference between the AI diagnostic procedures using CC-Cruiser and traditional eye clinics. We also investigated patients' feedback regarding the medical services provided by CC-Cruiser and senior consultants.

2. Methods

2.1. Study Design and Participants

This is a large, multicentre, parallel-group, randomized controlled trial performed in five Chinese ophthalmic clinics. The Consolidated Standards for Reporting Trials (CONSORT) guidelines have been followed in our study [27]. The leading study centre of this trial is the ZOC, located in Guangzhou in southern China. The other four eye clinics are affiliated with Shenzhen Eye Hospital, the Central Hospital of Wuhan, the Second Affiliated Hospital of Fujian Medical University, and Kaifeng Eye Hospital. We selected these collaborating hospitals from different areas to represent the diversity of healthcare settings across China.

Participants were recruited by the investigators according to standard inclusion criteria of the ophthalmic clinics in these hospitals. Participants were eligible for the study if they were less than 14 years old, with or without eye symptoms, and had no history of previous eye surgery. All participants were required to undergo slit-lamp photography, and sedatives such as chloral hydrate were used when necessary. Patients who already had a definitive diagnosis of cataract, other ocular abnormalities or ocular trauma were excluded.

Written informed consent was obtained at enrollment from at least one guardian of each participating child, and the principles outlined in the Declaration of Helsinki were followed throughout the study. The study protocol was approved by the ethics committee of ZOC and the institutional review boards at all collaborating centres, including Shenzhen Eye Hospital, the Central Hospital of Wuhan, the Second Affiliated Hospital of Fujian Medical University, and Kaifeng Eye Hospital. This trial is registered with ClinicalTrials.gov (NCT03240848).

2.2. Randomization and Masking

The participants were randomized (1:1) to receive a diagnosis from either CC-Cruiser or senior consultants, where one participant (two eyes) was randomized to the same group. Centralized randomization was done via a random number generating program with no stratification factors to avoid selection bias. Investigators in each study centre assessed the eligibility of each patient. If the patient met the inclusion criteria, the investigator sent the patient's information to a study coordinator, and the coordinator notified the investigators about the allocated group. Slit-lamp photography and patient recruitment were performed in each participating clinic by trained clinical staffs. The clinical staffs, investigators involved in data management and analysis, and experts providing the golden standard diagnosis by consensus in each clinic were blinded to the group assignments to help prevent ascertainment bias. The study participants, senior consultants, the study coordinator and study personnel responsible for randomization were not masked.

2.3. Procedures

The CC-Cruiser platform at the Children's Cataract Center of the ZOC was connected with all collaborating clinics through internet. A CC-Cruiser website (https://www.cc-cruiser.com/version1) has been established with a demonstration video of guidelines and instructions. Registered users can upload new cases to CC-Cruiser, and the output will include: diagnosis (normal lens versus cataract), comprehensive evaluation (opacity area, density and location), and treatment recommendation (surgery versus follow-up). Senior consultants with at least 5 years of clinical experience in pediatric ophthalmology provided initial diagnoses in each centre. The investigators created a profile for every eligible and consenting participant and documented their demographic information and baseline clinical characteristics, including sex, date of birth, family history of cataract, and eye symptoms. The participating investigators and clinical staffs at each centre received standardized training for the study procedures before the trial. All eligible participants underwent slit-lamp photography with pupil dilation and unified standard of diffused light, appropriate illumination intensity of slit-lamp, and normal eye position before group assignment. The clinical staffs attempted no more than three times for each eye. The investigators used sedatives (chloral hydrate) for 43 very young patients who would otherwise not cooperate with this examination.

The participants in the AI group were assigned to the AI clinic after slit-lamp photography. The investigators sent images of the ocular anterior segment to CC-Cruiser and received the initial diagnoses (normal lens versus cataract) with comprehensive evaluations of disease severity (lens opacity and the opacity area, density, and location) and treatment suggestions (surgery versus follow-up). The investigators calculated the time required for visiting CC-Cruiser and receiving initial diagnoses.

The participants in the senior consultants group were assigned to the regular ophthalmic clinic. The senior consultants provided patients with initial diagnostic reports including the disease severity and treatment decision. The investigators also calculated the time required for the diagnostic process.

After receiving an initial diagnosis, all the participants with identification numbers masked received a gold standard diagnosis from an expert panel including three cataract experts with more than 10 years of clinical experience in ophthalmology. The expert panel performed the slit-lamp examination and reached a consensus to make a final definitive diagnosis and treatment-decision for every patient. After the initial diagnostic report and standard diagnosis, the participants and their guardians were asked to complete a questionnaire regarding their satisfaction with diagnostic accuracy and efficiency.

2.4. Outcomes

The primary outcome was the accuracy of the diagnosis normal lens versus cataract. Because there is no available international classification system for pediatric cataracts, the reference standard for the evaluation of pediatric cataracts is the diagnosis from the cataract experts. The investigators compared the diagnostic accuracy of CC-Cruiser to that of the senior consultants using the gold standard diagnoses from the cataract experts.

The secondary outcomes included the evaluation of the disease severity, the time required for making the diagnosis, and patient satisfaction. The disease severity was comprehensively evaluated with the opacity area (extensive versus limited), density (dense versus non-dense), location (central versus peripheral), and treatment recommendations (surgery versus follow-up). The opacity area was defined as extensive when the opacity covered more than 50% of the pupil; otherwise, it was defined as limited. The opacity density was defined as dense when the opacity fully disrupted fundus imaging; otherwise, it was defined as non-dense. The opacity location was defined as central when the opacity fully covered the visual axis area; otherwise, it was defined as peripheral. Because the diagnosis was based on the slit-lamp image of the ocular anterior segment, the time required for diagnosis was calculated from the beginning of image acquisition to the completion of initial diagnostic reports and treatment recommendations by CC-Cruiser or senior consultants. The level of patient satisfaction was evaluated and analyzed via a seven-item questionnaire. A score of 1 indicated disagree; 2 indicated neutral; 3 indicated agree; and 4 indicated strongly agree. Both the number and percentage of participants who responded to each item were documented, and the mean rating for each item was calculated.

2.5. Statistical Analysis

Using the data from a comparative test with CC-Cruiser [25], we calculated that a sample size of at least 700 eyes (assuming a 1:1 allocation ratio, 350 eyes in each group) was required to compare diagnostic accuracy between CC-Cruiser and senior consultants based on the expected accuracy of 90% in the AI arm and 95% in the senior consultants arm, an 80% statistical power, and a 5% statistical significance level [27–30].

The study analyses followed a comprehensive, prespecified statistical analysis plan. Demographic and clinical data were recorded at baseline. Baseline demographics and diseases characteristics were statistically analyzed to confirm that all 350 participants (700 eyes) were well randomized into two study groups. The intention to treat population is same with the population of per protocol in this trial since no patients discontinued or withdrew after recruitment. Then, our primary analysis included all patients as originally allocated after randomization. The analysis of diagnostic accuracy was at eye level, and bilateral eyes in the same person were separately analyzed in the same group. We calculated the sensitivity, specificity, accuracy, positive predictive value (PPV), and negative predictive value (NPV) of CC-Cruiser and the senior consultants with reference to the gold standard (the cataract experts). The correct diagnosis of cataract was further analyzed with comprehensive evaluations of disease severity and

treatment recommendations. The generalized estimating equations (GEE) method, an extension of the quasi-likelihood approach, is being increasingly used to analyze longitudinal and other correlated data, especially when they are binary or in the form of counts [31]. We applied two eyes data from one person, which belonged to correlated data, therefore we performed the GEE to identify significant differences in the accuracy, true positive fraction (TPF), and false positive fraction (FPF) between CC-Cruiser and the senior consultants. The TPF is equivalent to sensitivity, and the FPF is equivalent to 1-specificity. The time required by CC-Cruiser and the senior consultants was assessed by the Mann–Whitney U test. Patient satisfaction with the medical service was also calculated as the mean rating with standard deviation. The Mann–Whitney U test was performed to identify significant differences in the responses to each question between the two groups. The criterion for significance was set at $\alpha = 0.05$. For all models, the results are expressed as an estimate of the effect size with odd ratio (OR), 95% CIs and *p*-values. All statistical analyses were performed with SPSS (version 20; SPSS, Inc., Chicago, IL, USA).

3. Results

Between August 9, 2017, and May 25, 2018, 353 patients were screened for eligibility (Fig. 1). After screening, three very young children were excluded because they could not take chloral hydrate and undergo slit-lamp photography. The remaining 350 participants (700 eyes) were randomly assigned to either the AI group (350 eyes) or a senior consultant's group (350 eyes). No participant withdrew from the study after randomization. Three hundred and fifty participants (700 eyes) were included in the analysis. The baseline demographics and disease characteristics, including sex, age, family history, eye symptoms, patients with cataracts, eyes with cataracts and severity of cataract were comparable between the two groups (Table 1).

With reference to the cataract experts' standards, the sensitivity, specificity, accuracy, PPV, and NPV of the diagnoses (normal lens versus cataract) were 89.7%, 86.4%, 87.4%, 74.4%, and 95.0%, respectively, for CC-Cruiser, compared to 98.4%, 99.6%, 99.1%, 99.2%, and 99.1%, respectively, for the senior consultants (Table 2). The diagnostic accuracy and TPF for childhood cataracts for CC-Cruiser were significantly lower ($p < 0.001$, OR = 0.06 [95% CI 0.02 to 0.19]; and $p = 0.012$, OR = 0.14 [95% CI 0.03 to 0.65], respectively) and the FPF for CC-Cruiser was significantly higher than those for the senior consultants ($p < 0.001$, OR = 43.05 [95% CI 5.42 to 341.70]) (Table 2). CC-Cruiser was significantly less accurate in diagnosing cataracts than senior consultants. The percentages of correct comprehensive evaluations of lens opacity including the opacity area, density, and location were 90.6%, 80.2%, and 77.1%, respectively, in the AI group, compared to 93.3%, 85.0%, and 87.5%, respectively, in the senior consultants group (Table 3). Compared to senior consultants, CC-Cruiser exhibited no statistical difference when evaluating the opacity area, density, and opacity location ($p = 0.463$, 0.286, and 0.130, respectively) (Table 3). The treatment recommendations (surgery versus follow-up) provided by CC-Cruiser were significantly less accurate than those provided by the senior consultants (70.8% vs. 96.7%, $p < 0.001$, OR = 0.08 [95% CI 0.03 to 0.25], Table 3).

The time required for CC-Cruiser to make a diagnosis and treatment recommendation was less than that required for the senior consultants (2.79 min vs. 8.53 min, $p < 0.001$, mean difference 5.74 [95% CI 5.43 to 6.05], Table 4).

At the end of the study, 345 participants completed the evaluation questionnaire (172 in the CC-Cruiser group and 173 in the senior consultant group). Five participants' guardians did not complete the questionnaires for personal reasons. The responses to each statement are summarized in Table 5. The response rates for the completion of questionnaire were 98.3% for the AI group and 98.9% for the senior consultant group. The patients had high

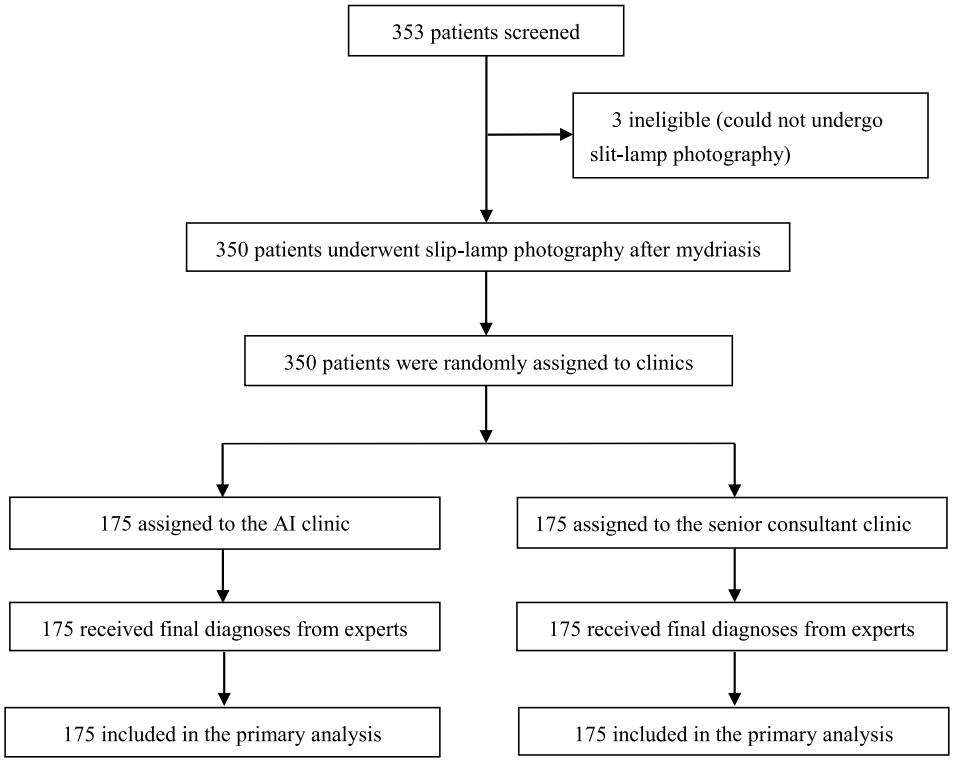

Fig. 1. Trial profile. AI = artificial intelligence.

Table 1
Baseline demographics and disease characteristics.

	AI group (N = 175 P/350 E)	SC group (N = 175 P/350 E)	*p*-Value
Sex			
Male	77 (44.0%)	82 (46.9%)	$p = 0.591$[a]
Female	98 (56.0%)	93 (53.1%)	
Age (years)	6.58 (0.45)	5.89 (0.45)	$p = 0.124$[b]
Family history of cataracts			
No	165 (94.3%)	163 (93.1%)	$p = 0.660$[a]
Yes	10 (5.7%)	12 (6.9%)	
Eye symptoms			
No	96 (54.9%)	93 (53.1%)	$p = 0.748$[a]
Yes	79 (45.1%)	82 (46.9%)	
Patients with cataracts			
Normal	106 (60.6%)	100 (57.1%)	
Monocular cataracts	31 (17.7%)	28 (16.0%)	$p = 0.527$[a]
Bilateral cataract	38 (21.7%)	47 (26.9%)	
Eyes with cataracts			
No	243 (69.4%)	228 (65.1%)	$p = 0.342$[c]
Yes	107 (30.6%)	122 (34.9%)	
Opacity area			
Extensive	71 (66.4%)	85 (69.7%)	$p = 0.764$[c]
Limited	36 (33.6%)	37 (30.3%)	
Density			
Dense	69 (64.5%)	77 (63.1%)	$p = 0.696$[c]
Non-dense	38 (35.5%)	45 (36.9%)	
Location			
Central	77 (72.0%)	83 (68.0%)	$p = 0.776$[c]
Peripheral	30 (28.0%)	39 (32.0%)	
Treatment recommendations			
Surgery	63 (58.9%)	76 (62.3%)	$p = 0.575$[c]
Follow-up	44 (41.1%)	46 (37.7%)	

Data are presented as the number *n* (%) or mean (standard deviations). Percentages do not add up to 100% in some cases because of rounding. The χ^2 test was performed to compare the characteristics of sex, family history of cataracts, eye symptoms and patients with cataracts between the AI group and the senior consultant group. An independent samples *t*-test was performed to compare age between the two groups. The generalized estimating equation was performed to compare the eyes with cataracts, disease severity, and treatment recommendations. None of the baseline characteristics differed significantly at the 0.05 level between groups. AI = artificial intelligence. SC = senior consultant. P = participants. E = eyes.
[a] χ^2 test.
[b] *t*-Test.
[c] Generalized estimating equation.

satisfaction levels regarding the medical services provided by CC-Cruiser, especially for the time required for diagnosis. The mean rating for overall satisfaction with CC-Cruiser was 3.47 ± 0.501, which was higher than that of the senior consultants (3.38 ± 0.554, $p = 0.007$, Table 5), indicating that patients preferred medical AI than real doctors when receiving medical services.

4. Discussion

In this study, we showed that CC-Cruiser was less accurate in diagnosing childhood cataracts and making treatment decisions in clinical practice than the senior consultants. However, compared to senior consultants, CC-Cruiser required less time for diagnosis, and achieved a high level of patients' satisfaction. These results highlighted the clinical importance of diagnostic randomized controlled trials for evaluating real-world performance of CC-Cruiser before regular use in outpatient settings.

The real-world diagnostic accuracy of CC-Cruiser is lower than that reported in our previous study conducted with screening datasets [25]. Although CC-Cruiser was highly accurate in evaluating 306 standard images of the ocular anterior segment, cataracts were misdiagnosed and evaluated inaccurately more often by CC-Cruiser than senior consultants for 43 images of poor quality in this clinical trial, which could be attributed to several reasons. First, some pediatric patients could not cooperate sufficiently and fix their eyes on the cameras due to photophobia or lack of attention. Therefore, the slit-lamp could not be focused properly on the lens. Second, the eyelids and eyelashes could obscure the lens, compromising the quality of the captured images. Third, if the reflective point was focused near the visual axis, the features on the reflective point on the lens could not be accurately extracted, leading to a misdiagnosis of cataract and a higher false positive fraction for CC-Cruiser. Fourth, the strong illumination intensity of slit-lamp may result in artefactual lens opacities, which was another reason for a higher false positive fraction for CC-Cruiser. However, these problems could usually be identified by senior consultants, as they could adjust the focus point manually and evaluate the opacity from different sites or angles of the lenses. The higher false positives may increase the burden and cost of medical resources and may result in physical or mental injury to the patients. In addition, although diagnosis by CC-Cruiser at the current stage may still need inputs from clinicians (including using sedative drugs) to ensure the quality of image capture, we believe that further improvement in autofocus technology of medical AI will achieve more diagnostic accuracy with less requirement of human input. For example, an improvement in the recognition of reflective point on the lens can significantly reduce false positive rate.

Previous studies indicated that AI-facilitated diagnosis can alleviate doctors' workload and contribute to high-quality medical care provision to patients in need [3,12]. Here, we showed that in clinical application, the medical AI platform exhibited superiority to real human doctors in terms of shortened diagnostic time. Consistently, the participants in the CC-Cruiser group felt that they received a faster diagnosis, and the waiting time required for the outpatient visit was significantly reduced. It is reported that AI have the potential to reduce costs in health care

Table 2
Diagnostic performance regarding childhood cataract.

	Sensitivity	Specificity	Accuracy	Positive predictive value	Negative predictive value	Accuracy difference (*p*-value, OR [95% CI])	TPF difference (*p*-value, OR [95% CI])	FPF difference (*p*-value, OR [95% CI])
CC-Cruiser	89.7%	86.4%	87.4%	74.4%	95.0%			
						−11.7 ($p < 0.001$, OR = 0.06 [95% CI 0.02 to 0.19])	−8.7 ($p = 0.012$, OR = 0.14 [95% CI 0.03 to 0.65])	13.2 ($p < 0.001$, OR = 43.05 [95% CI 5.42 to 341.70])
Senior consultants	98.4%	99.6%	99.1%	99.2%	99.1%			

Eyes were the units of analysis (N = 700). There were 350 eyes in the CC-Cruiser group and 350 eyes in the senior consultant group. OR = odd ratio. CI = confidence interval. TPF = true positive fraction. TPF is equivalent to sensitivity. FPF = false positive fraction. FPF is equivalent to 1-specificity. We performed a diagnostic accuracy analysis with reference to the cataract specialists' standards. The TPF and FPF of diagnosis (normal lens versus cataract) were 89.7%, and 13.6%, respectively, for CC-Cruiser and 98.4%, and 0.4%, respectively, for the senior consultants. The generalized estimating equation (GEE) was performed to identify significant differences in accuracy, TPF, and FPF between CC-Cruiser and the senior consultants. GEE results (adjusted results) and logistic regression results (unadjusted results) for cataract diagnosis in the supplementary table were presented to show the impact of the cluster at the level of participants (Supplementary table 1 for adjusted and unadjusted results).

Table 3
Comprehensive evaluations of childhood cataract and treatment recommendations.

	Sensitivity	Specificity	Accuracy	Accuracy difference (p-value, OR [95% CI])	TPF difference (p-value, OR [95% CI])	FPF difference (p-value, OR [95% CI])
Opacity area						
CC-Cruiser	91.3%	88.9%	90.6%	−2.7 ($p = 0.460$, OR = 0.66 [95% CI 0.22 to 1.98])	−2.8 ($p = 0.564$, OR = 0.68 [95% CI 0.18 to 2.56])	2.5 ($p = 0.439$, OR = 2.11 [95% CI 0.32 to 14.05])
Senior consultants	94.1%	91.4%	93.3%			
Density						
CC-Cruiser	85.3%	67.9%	80.2%	−4.8 ($p = 0.286$, OR = 0.64 [95% CI 0.28 to 1.45])	3.5 ($p = 0.867$, OR = 1.09 [95% CI 0.40 to 2.97])	22.8 ($p = 0.042$, OR = 4.24 [95% CI 1.05 to 17.13])
Senior consultants	81.8%	90.7%	85.0%			
Location						
CC-Cruiser	84.2%	50%	77.1%	−10.4 ($p = 0.130$, OR = 0.52 [95% CI 0.22 to 1.21])	−7.4 ($p = 0.351$, OR = 0.59 [95% CI 0.20 to 1.78])	28.4 ($p = 0.134$, OR = 2.91 [95% CI 0.72 to 11.71])
Senior consultants	91.6%	78.4%	87.5%			
Treatment						
CC-Cruiser	86.7%	44.4%	70.8%	−25.9 ($p < 0.001$, OR = 0.08 [95% CI 0.03 to 0.25])	−8.0 ($p = 0.247$, OR = 0.44 [95% CI 0.11 to 1.77])	55.6
Senior consultants	94.7%	100.0%	96.7%			

Eyes were the units of analysis. A total of 216 eyes (correctly diagnosed as cataracts in both groups, 96 eyes in the CC-Cruiser group and 120 eyes in the senior consultant group) were further analyzed by comprehensive evaluation of lens opacity, including the opacity area (extensive versus limited), density (dense versus non-dense), and location (central versus peripheral), and the recommended treatment (surgery versus follow-up) with reference to the cataract specialists' standards. OR = odd ratio. CI = confidence interval. TPF = true positive fraction. TPF is equivalent to sensitivity. FPF = false positive fraction. FPF is equivalent to 1-specificity. The generalized estimating equation was performed to identify significant differences in the accuracy, TPF, and FPF of the opacity area, density, and location and the treatment recommendations between CC-Cruiser and the senior consultants. The p-value and OR of the difference in FPF of treatment between two groups couldn't be calculated because of the 100% specificity for senior consultants. GEE results (adjusted results) and logistic regression results (unadjusted results) for evaluation of cataract and treatment in the supplementary table were presented to show the impact of the cluster at the level of participants (Supplementary table 1 for adjusted and unadjusted results).

economies [32]. With the widespread application of AI technology in health care, economic cost will also be lower than human doctors since only the cost of development and operation of machine will be assumed. Therefore, AI technology is a promising modality for providing high-quality health services to large populations in time- and cost-effectiveness [13].

Patients' satisfaction with medical AI has not been fully studied. Laure et al. assessed patient satisfaction with rheumatoid arthritis (RA) care using Sanoia, an e-health website [33]. The authors showed discordance between patient satisfaction and access to the AI platform, primarily because RA is a chronic disease, and patients may lose interest in using Sanoia and become less dedicated to regular disease self-management when the disease is in remission [33]. However, childhood cataract can be vision-threatening if early diagnosis and appropriate management are not provided [24]. Therefore, parents of pediatric patients are eager to access a medical service for diagnosis and treatment-decisions with high efficiency. Our study showed that the overall patient satisfaction with CC-Cruiser was slightly higher than that with the senior consultants, indicating that patients had good experience in the AI medical service. The satisfaction of patients may due to their curiosity or interest to medical AI, or the fact that patients need a balance between the diagnostic accuracy and diagnostic time and are more willing to receive medical service that is less time-consuming and with acceptable diagnostic accuracy. Our results support that CC-Cruiser at least achieves a comparable satisfaction metrics as human doctors do. Therefore, CC-Cruiser, at its current stage, has shown potential to assist human doctors in clinical applications. In future studies, we will dedicate in the improvement of accuracy of CC-Cruiser to increase patient satisfaction.

The strengths of the study include its randomized, controlled design, a large sample size, and data collection from five eye clinics across China. However, our trial has several limitations. First, as patients without symptoms such as blurred vision were less willing to participate in the study, we may have missed some patients with slightly opaque lens. Therefore, assessment of early-stage cataract by CC-Cruiser needs further improvement. Second, CC-Cruiser provided treatment suggestions without considering the patients' general conditions. Therefore, a small proportion (six cases) of treatment recommendations provided by CC-Cruiser were not consistent with those made by the experts, despite that lens opacity had been accurately evaluated. Further improvement of the capacity for treatment determination will require consideration of non-ophthalmic factors, such as age and health status [34]. Third, our AI system was reliance on the computing power and internet accessibility, thus difficulties of widespread application of CC-Cruiser may exist in those developing areas without stable internet. However, those remote locations with internet access can still benefit from medical service provided by CC-Cruiser. Fourth, a cluster randomized controlled trial (cluster at level of the pediatric patients) has been undertaken in this trial because the randomization was on the level of patients and the observation and its analysis was on the level of eyes. However, when the sample size was calculated, the intra-cluster correlation between two eyes from one child was not accounted for, and a randomized controlled trial design was adopted. This would result in a statistical power lower than 0.8 as anticipated since cluster randomized controlled trials require larger sample size than randomized controlled trials to achieve the same statistical power.

In conclusion, this is the first clinical randomized controlled trial to validate the diagnostic accuracy and efficiency of an AI system in eye clinics. This represents the first clinical trial of this kind to robustly evaluate the clinical application of medical AI. CC-Cruiser exhibited less accuracy compared with senior human consultants in diagnosing childhood cataracts and making treatment decisions, but has the capacity to assist human doctors in clinical practice in its current state. Further efforts will be required to perform the clinical controlled trials to

Table 4
Time required for the diagnostic process of CC-Cruiser and senior consultants.

	Mean time (minutes)	Standard deviation	95% CI		Mean difference (p-value, 95% CI)
			Lower	Upper	
CC-Cruiser	2.79	1.11	2.64	2.96	5.74 ($p < 0.001$, 95% CI 5.43 to 6.05)
Senior consultants	8.53	1.75	8.27	8.78	

Three hundred patients were included in the analysis (175 participants in the CC-Cruiser group and 175 participants in the senior consultant group). The Mann–Whitney U test was performed to compare the time required. Significant differences in time required were observed between the CC-Cruiser and senior consultant groups ($p < 0.001$). CI = confidence interval.

Table 5
Questionnaire provided to the participants with their responses to the clinical service.

Question	Response in the AI group (N = 172)				Mean rating (SD)	Response in the SC group (N = 173)				Mean rating (SD)	*p*-Value
	1	2	3	4		1	2	3	4		
The initial diagnosis of the eye clinic was credible.	5.2% (9)	15.9% (27)	32.0% (55)	47.1% (81)	3.21 (0.893)	0	3.5% (6)	58.4% (101)	38.2% (66)	3.35 (0.546)	$p = 0.679$
The initial therapeutic decision of the eye clinic was credible.	4.7% (8)	18.0% (31)	28.5% (49)	48.8% (84)	3.22 (0.902)	0	5.8% (10)	57.0% (98)	37.6% (65)	3.32 (0.578)	$p = 0.972$
The initial diagnosis of the eye clinic was consistent with that of the experts.	3.5% (6)	18.6% (32)	21.5% (37)	56.4% (97)	3.31 (0.896)	0	2.9% (5)	37.6% (65)	59.5% (103)	3.57 (0.552)	$p = 0.053$
The initial therapeutic decision of the eye clinic was consistent with that of the experts.	3.5% (6)	23.3% (40)	17.4% (30)	55.8% (96)	3.26 (0.918)	0	4.0% (7)	37.6% (65)	58.4% (101)	3.54 (0.575)	$p = 0.042$
I was satisfied with the time required to wait for CC-Cruiser/senior consultants in this eye clinic.	0	0.6% (1)	41.9% (72)	57.6% (99)	3.57 (0.508)	0	5.2% (9)	50.9% (88)	43.9% (76)	3.39 (0.586)	$p = 0.005$
I was satisfied with the time required to make the diagnosis and provide treatment recommendations by CC-Cruiser/senior consultants.	0	0	43.6% (75)	56.4% (97)	3.56 (0.497)	0	0.5% (1)	59.5% (103)	39.9% (69)	3.38 (0.554)	$p = 0.002$
Overall, I was satisfied with this medical service provided in this eye clinic.	0	0	52.9% (91)	47.1% (81)	3.47 (0.501)	0	4.0% (7)	61.3% (106)	34.7% (60)	3.31 (0.543)	$p = 0.007$

Data are presented as the number (%) or mean (standard deviation). Percentages do not add up to 100% in some cases because of rounding. Three hundred and forty-five patients were included in the analysis (172 in the CC-Cruiser group and 173 in the senior consultant group). Pediatric participants and at least one of their guardians were asked to complete the questionnaire together. Five participants' guardians were unwilling to complete the questionnaires because of personal reasons. The survey questions used a 4-point scale (1, disagree; 2, neutral; 3, agree; and 4, strongly agree). The Mann–Whitney U test was performed to identify significant differences in responses to each question between the two groups. AI = artificial intelligence. SC = senior consultant. SD = standard deviation.

appropriately evaluate the real-world diagnostic performance of medical AI.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eclinm.2019.03.001.

Contributors

HTL and YZL contributed to the concept of the study and critically reviewed the manuscript. YZL, HTL, RYL, ZZL, HC, YHY, and JJC designed the study and did the literature search. ZLL, WYL, EPL, XHW, DRL, DXW, TYY, QZC, XYL, JL, WTL, JHW, MMY, HLH, LZ, YY, XLC, JMH, KZ, XML, XYZ, LXL, YHL, XLL, BC, DYZ, MXW, and WRC contributed to the data collection. JLH contributed to the design of statistical analysis plan. RYL, YHY, HTL, and JLH did the data analysis and data interpretation. HTL and RYL drafted the manuscript. HTL, RYL, ZZL, HC, YHY, YZ, and CC critically revised the manuscript. HTL provided research funding, coordinated the research and oversaw the project. All the authors reviewed the manuscript for important intellectual content and approved the final manuscript.

Role of the Funding Source

The source of funding played no role in the design of the study protocol, data collection, data analysis, data interpretation, writing of the report, or the decision to submit the manuscript for publication. The corresponding author had full access to all data in the study and assumes final responsibility for the decision to submit the manuscript for publication. All authors approved the decision to submit.

Declaration of Interests

The authors declare no competing financial interests.

Acknowledgments

We thank all the participants, their families, and the institutions for supporting this study. This study was funded by National Key R&D Program of China (2018YFC0116500) and the Key Research Plan for the National Natural Science Foundation of China in Cultivation Project (91846109). We thank Yimin Chen, Jian Zhang, and Shaodong Hong for their advice in the statistical analysis and interpretation.

References

[1] 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(22):2402–10.
[2] Kim DH, Kim H, Kwak S, et al. The settings, pros and cons of the new surgical robot da Vinci Xi system for transoral robotic surgery (TORS): a comparison with the popular da Vinci Si system. Surg Laparosc Endosc Percutan Tech 2016;26(5):391–6.
[3] Curioni-Fontecedro A. A new era of oncology through artificial intelligence. ESMO Open 2017;2(2):e000198.
[4] Hamet P, Tremblay J. Artificial intelligence in medicine. Metabolism 2017;69S:S36–40.
[5] Somashekhar SP, Sepulveda MJ, Puglielli S, et al. Watson for oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board. Ann Oncol 2018;29(2):418–23.
[6] Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542(7639):115–8.
[7] Printz C. Artificial intelligence platform for oncology could assist in treatment decisions. Cancer 2017;123(6):905.
[8] Wang C, Qi X, Chen X, Yu Q, Xing L. The establishment of China standardized residency training system. Zhonghua Yi Xue Za Zhi 2015;95(14):1041–3.
[9] Jiang Y, Luo L, Congdon N, Wang S, Liu Y. Who will be wielding the lancet for China's patients in the future? The Lancet 2016;388(10054):1952–4.
[10] Wang YX. On the training of young doctors in China. Quant Imaging Med Surg 2015;5(1):182–5.
[11] Azad N, Amos S, Milne K, Power B. Telemedicine in a rural memory disorder clinic-remote management of patients with dementia. Can Geriatr J 2012;15(4):96–100.
[12] Castaneda C, Nalley K, Mannion C, et al. Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. J Clin Bioinforma 2015;5:4.
[13] Patel VL, Shortliffe EH, Stefanelli M, et al. The coming of age of artificial intelligence in medicine. Artif Intell Med 2009;46(1):5–17.
[14] Hashmi S. 'Coming of Age' of artificial intelligence: evolution of survivorship care through information technology. Bone Marrow Transplant 2016;51(1):41–2.

[15] Raju M, Pagidimarri V, Barreto R, Kadam A, Kasivajjala V, Aswath A. Development of a deep learning algorithm for automatic diagnosis of diabetic retinopathy. Stud Health Technol Inform 2017;245:559–63.
[16] Burlina PM, Joshi N, Pekala M, Pacheco KD, Freund DE, Bressler NM. Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. JAMA Ophthalmol 2017;135(11):1170–6.
[17] Ting DSW, Cheung CY, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA 2017;318(22):2211–23.
[18] Poplin R, Varadarajan AV, Blumer K, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nature Biomedical Engineering 2018;2(3):158–64.
[19] Kermany DS, Goldbaum M, Cai W, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 2018;172(5):1122–31 e9.
[20] Schieppati A, Henter J-I, Daina E, Aperia A. Why rare diseases are an important medical and social issue. The Lancet 2008;371(9629):2039–41.
[21] Taruscio D, Vittozzi L, Stefanov R. National plans and strategies on rare diseases in Europe; 2010.
[22] Remuzzi G, Garattini S. Rare diseases: what's next?; 2008.
[23] Lenhart P, Courtright P, Edward Wilson M, et al. Global challenges in the management of congenital cataract: proceedings of the 4th international congenital cataract symposium held on March 7, 2014, New York, New York; 2015.
[24] Medsinge A, Nischal KK. Pediatric cataract: challenges and future directions. Clin Ophthalmol 2015;9:77–90.
[25] Long E, Lin H, Liu Z, et al. An artificial intelligence platform for the multihospital collaborative management of congenital cataracts. Nature Biomedical Engineering 2017;1(2):0024.
[26] Rodger M, Ramsay T, Fergusson D. Diagnostic randomized controlled trials: the final frontier. Trials 2012;13:137.
[27] Chen YL, Yang KH. Consort 2010. Lancet 2010;376(9737):230.
[28] Augestad KM, Berntsen G, Lassen K, et al. Standards for reporting randomized controlled trials in medical informatics: a systematic review of CONSORT adherence in RCTs on clinical decision support. J Am Med Inform Assoc 2012;19(1):13–21.
[29] Abdul Latif L, Daud Amadera JE, Pimentel D, Pimentel T, Fregni F. Sample size calculation in physical medicine and rehabilitation: a systematic review of reporting, characteristics, and results in randomized controlled trials. Arch Phys Med Rehabil 2011;92(2):306–15.
[30] Bacchetti P, Leung JM. Sample size calculations in clinical research. Anesthesiology 2002;97(4):1028–9 [author reply 9–32].
[31] Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73(1):13–22.
[32] Tufail A, Rudisill C, Egan C, et al. Automated diabetic retinopathy image assessment software: diagnostic accuracy and cost-effectiveness compared with human graders. Ophthalmology 2017;124(3):343–51.
[33] Gossec L, Cantagrel A, Soubrier M, et al. An e-health interactive self-assessment website (Sanoia((R))) in rheumatoid arthritis. A 12-month randomized controlled trial in 320 patients. Joint Bone Spine 2018;85(6):709–14.
[34] Wakeman B, MacDonald IM, Ginjaar I, Tarleton J, Babu D. Extraocular muscle hypertrophy in myotonia congenita: mutation identified in the SCN4A gene (V445M). J AAPOS 2009;13(5):526–7.