EBioMedicine 51 (2020) 102621

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.elsevier.com/locate/ebiom

Research paper

A practical model for the identification of congenital cataracts using machine learning

Duoru Lin[a], Jingjing Chen[a], Zhuoling Lin[a], Xiaoyan Li[a], Kai Zhang[a,b], Xiaohang Wu[a], Zhenzhen Liu[a], Jialing Huang[c], Jing Li[a], Yi Zhu[a,d], Chuan Chen[a,d], Lanqin Zhao[a], Yifan Xiang[a], Chong Guo[a], Liming Wang[b], Yizhi Liu[a], Weirong Chen[a,**], Haotian Lin[a,*]

[a] State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Jinsui Road #7, Guangzhou, Guangdong 510060, People's Republic of China
[b] School of Computer Science and Technology, Xidian University, Xi'an, Shanxi 710071, People's Republic of China
[c] School of Public Health, Sun Yat-sen University, Guangzhou, Guangdong 510060, People's Republic of China
[d] Department of Molecular and Cellular Pharmacology, University of Miami Miller School of Medicine, Miami, FL 33136, USA

ARTICLE INFO

Article History:
Received 8 August 2019
Revised 18 December 2019
Accepted 19 December 2019
Available online xxx

Keywords:
Congenital anomaly
Congenital cataract
Identification model
Machine learning

ABSTRACT

Background: Approximately 1 in 33 newborns is affected by congenital anomalies worldwide. We aimed to develop a practical model for identifying infants with a high risk of congenital cataracts (CCs), which is the leading cause of avoidable childhood blindness.

Methods: This case-control study was performed in the Zhongshan Ophthalmic Center and involved 2005 subjects, including 1274 children with CCs and 731 healthy controls. The CC identification models were established based on birth conditions, family medical history, and family environmental factors using the random forest (RF) and adaptive boosting methods (trained by 1129 CC cases and 609 healthy controls), which were tested by internal 4-fold cross-validation and external validation (145 CC cases and 122 healthy controls). The models were also tested using 4 datasets with gradually reduced proportions of CC patients (bilateral cases) to validate their performance in an approximate simulation of a clinical environment with a relatively low disease prevalence.

Findings: The CC identification models showed high discrimination in both the 4-fold cross validation (area under the curve (AUC)=0.91 [95% confidence interval: 0.88–0.94] in bilateral cases; 0.82 [0.77–0.89] in unilateral cases) and external validation (AUC=0.93±0.05 in bilateral cases; 0.86±0.01 in unilateral cases), and achieved stable performance in the clinical tests (AUC=0.94–0.96 in the four subgroups by RF). Furthermore, family history of CC, low parental education level, and comorbidity were identified as the top three most relevant factors to both bilateral and unilateral CC diagnosis.

Interpretation: Our CC identification models can accurately discriminate CC patients from healthy children and have the potential to serve as a complementary screening procedure, especially in undeveloped and remote areas.

1. Introduction

According to the World Health Organization, approximately 1 in 33 newborns is affected by congenital anomalies worldwide [1]. This global health issue has become one of the main causes of long-term illness, disability and even death in infants, resulting in economic and emotional burdens on individuals, families, healthcare systems and society. Congenital/infantile cataracts (CCs), with a global prevalence ranging from 2.2 to 13.6 per 10,000 children [2], are a typical congenital anomaly that occurs before or during the critical stage of visual development and has become one of the leading causes of avoidable childhood blindness worldwide [3].

Due to the difficulties associated with treatment and the poor prognosis, as well as the time limitation imposed by visual development among patients with CCs, prevention and early detection are the best disease management strategies [4,5]. In underdeveloped regions where medical resources are in short supply, CCs, an anomaly with a low prevalence but long-term effects, may not be included in routine congenital disease screening programmes or may be missed due to poor screening coverage. Although most infants in developed

* Corresponding author.
** Corresponding author.
E-mail addresses: chenwr_q@aliyun.com (W. Chen), gddlht@aliyun.com (H. Lin).

https://doi.org/10.1016/j.ebiom.2019.102621

Research in context

Evidence before this study

Approximately 1 in 33 newborns are affected by congenital anomalies worldwide according to the World Health Organization. Congenital cataracts (CCs) are a typical congenital anomaly that occurs before or during the critical stage of visual development and has become one of the leading causes of avoidable childhood blindness worldwide. We searched PubMed, Web of Science, and Wanfang Database for published articles with the keywords "congenital anomaly", "congenital disease", "congenital cataract", "peadiatric cataract", "prediction model", "screening", and "machine learning" (published between Jan 1, 2001, and Sept 30, 2019) with no language restrictions, but identified no known studies established the practical identification model for timely screening infants with high risk of developing CC based on nonimaging data, which is of great clinical significance. While studies have identified a few of risk factors for CC, but they were mostly studied independently in a relatively small number of patients and have not established the CC prediction models.

Added value of this study

This national study compared eleven potential risk factors of CC between CC patients and healthy controls, who exhibited distinct characteristics. Additionally, to our knowledge, we established a practical identification model, with high discrimination, for identifying infants with a high risk of CCs based on 11 easily obtainable predictive factors. This study assessed the most comprehensive collection of nonimaging-based risk/relevant factors and their predictive value in the early detection of CCs using a novel AI model based on the largest number of nationally representative subjects to date.

Implications of all the available evidence

The identification model has the potential to serve as a complementary screening procedure for the early detection or prediction of CC development, which could be especially useful in underdeveloped and remote areas. More broadly, our study may provide a reference for the development of AI-based preventive strategies for other congenital diseases.

regions undergo neonatal screening shortly after birth, many with late-onset CC [6] are not diagnosed at the time of screening. The delayed presentation of CC patients at hospitals remains very common, especially among those living in remote and undeveloped areas with poor medical resources [7,8]. A practical identification model that can screen infants at high risk for future CCs would be of great clinical significance.

Nongenetic factors account for approximately 2/3 of CC cases [9]. Disease screening based mainly on nongenetic factors is clinically practical and affordable. Sporadic nongenetic risk factors for CCs have been previously reported. An association between CCs and toxoplasmosis, rubella, cytomegalovirus, and herpes simplex (TORCH) has been reported based on the presence of specific IgM antibodies in developing countries [10]. The British Congenital Cataract Interest Group (BCCIG) reported that children with CC were overrepresented among those who were born preterm or had low birth weight [11]. Furthermore, intrauterine infection [12], histories of disease and medication use during pregnancy, toxin exposure, X-ray exposure [13], and poverty [1] are also reportedly associated with CCs. However, these factors have mostly been studied independently in relatively small samples of patients.

Recently, artificial intelligence (AI) has been applied in the diagnosis of ocular diseases. Ting et al. [14] developed a deep learning system to screen for diabetic retinopathy and related ocular diseases using retinal images. A review by Reid et al. [15] reported that machine learning had been applied to the classification of pediatric cataracts and the prediction of postoperative complications. We also evaluated the diagnostic efficacy and therapeutic decision-making capacity of an AI platform for childhood cataracts (CC—Cruser) in eye clinics. [16] However, most of these AI models were trained using image data collected with professional ophthalmological equipment, meaning that they might be of limited use for disease screening in the areas where there is inadequate access to medical resource.

Here, we assessed the most comprehensive collection of nonimaging-based risk/relevant factors and their predictive value in the early detection of CCs using a novel AI model based on the largest number of nationally representative subjects to date.

2. Methods

2.1. Patient enrollment and ethics statement

The primary dataset for this case-control study included 1129 patients with CCs and 609 healthy controls examined between February 2012 and January 2017. All patients with CC were recruited from the national center for CC prevention and treatment, the Childhood Cataract Programme of the Chinese Ministry of Health (CCPMOH) [17], which is located in the Zhongshan Ophthalmic Center (ZOC), Guangzhou, and receives transferred patients with CCs from 21 provinces or regions in China (61.72%, 21/34) (Supplementary file 1, Table S1). Due to the similar clinical management of both congenital and infantile cataracts and the interchangeable use of these terms in practice [18], a clinical definition of CCs was adopted for this study. CC patients aged $\leq$18 years had their diagnoses confirmed by two experienced pediatric ophthalmologists (WRC and HTL) based on slit-lamp (BX900; HAAG-STREIT AG, Bern, Switzerland) and Pentacam (Pentacam HR; Oculus, Inc., Wetzlar, Germany) examinations. Infants with newly or recently diagnosed ($<$ 2 weeks) vision-threatening congenital or infantile cataracts within the first year of life were included. Children in whom cataracts were diagnosed after the age of 1 year were eligible for inclusion only if the cataracts were confirmed to be due to a congenital cause or had specific ophthalmic features indicative of early onset, such as cataract morphology, presence of nystagmus, or associated congenital ocular anomaly [19]. Eligibility was reconfirmed based on the patients' ocular examinations, diagnosis, medical history, progress notes and other detailed medical records. Drug-induced cataracts, metabolic cataracts, secondary cataracts, traumatic cataracts, and developmental cataracts were excluded. Patients who were unable to actively cooperate were sedated with 10% chloral hydrate (0.8 ml/kg, oral or rectal administration) [20]. Patients who came from welfare homes and lacked clear information regarding heredity and gestation history were also excluded. Healthy children without cataracts were randomly recruited from kindergartens and communities located in different regions and served as controls. The healthy children were selected so that their ages and geographical distribution would be comparable to those of the children with CCs. The healthy controls were examined by the same ophthalmologists to verify the absence of CCs. Similarly, healthy children with unclear family conditions or pregnancy-labor history were excluded.

A total of 145 CC cases and 122 healthy controls collected from March 2017 to January 2018 were included in the external validation dataset. CC cases were obtained from the CCPMOH, which were referred from 12 provinces of China (Supplementary file 1, Table S2). Patients were eligible only if they had been diagnosed in other health care institutes and transferred to the ZOC for further treatment. Patients directly diagnosed in the ZOC were excluded from the external validation dataset. The healthy controls were recruited from

kindergartens and communities that differed from those of the primary dataset.

This study was registered with ClinicalTrials.gov (NCT03215186) and approved by the institutional review board of the ZOC at Sun Yat-sen University (IRB-ZOC-SYSU). All procedures followed the tenets of the Declaration of Helsinki. Written informed consent was obtained from at least one parent of each participant.

2.2. Collection of data on potential predictive factors

Questionnaire investigations were performed by experienced interviewers (ZLL and JL) in private conversation rooms at the ZOC (CC group) or kindergartens and neighborhood committees (control group). According to previous reports and clinical experience regarding the possible predictive factors of CCs, the following three classes of information were collected for each participant and confirmed independently by two other researchers (XYL and JJC): (1) demographic variables: age at recruitment, sex, and laterality; (2) birth conditions: illness during pregnancy, birth parity, preterm or term, eutocia or caesarian, history of supplemental oxygen inspiration/infant incubator use, and comorbidity; (3) family medical history and environmental factors: family history of CC (information on family history was included because this factor is readily obtainable and is important to the establishment of disease prediction model), radiation/pollution exposure, parental smoking, parental education level, and annual household income. The age and sex of the respondents were acquired by their identity cards or residence booklets. Comprehensive information regarding the family and pregnancy-labor histories, comorbidity, and living environment of all children was obtained based on medical records, physical examinations, abnormal appearances, and parent statements. All data were collected through a structured questionnaire designed and revised by the members of the CCPMOH (Supplementary file 2). Returned questionnaires were considered invalid and excluded from further analysis if they met the following criteria: more than one-third of the items were unclear, or left blank, or all items were regularly marked. Each item related to the relevant factors was carefully rechecked, sorted, and presented. The researchers who conducted the analysis were blinded to the data collection. All collected data was kept strictly confidential and deidentified before the analysis.

2.3. Data processing and construction of CC identification model

Fig. 1 shows an overview of the experimental strategy of the research performed in this study. First, we analyzed the potential predictive factors of CC and compared these factors between patients with CCs and healthy controls. Next, we established CC identification models, and the importance ranking of the factors in each identification model was assessed.

The abovementioned information regarding the CC patients and healthy controls, collected between February 2012 and January 2017, was used as the training dataset for the CC identification models. Missing data (fewer than one-third of questions answered, other than the annual household income) were imputed using the missForest algorithm [21]. The CC identification models were established in bilateral and unilateral cases using two common AI analysis algorithms: random forest (RF) [22] and adaptive boosting (Ada) [23]. Four-fold cross validation was performed: the training data were randomly and equally divided into four sub-samples, three of which were used to train the prediction models, and the remaining one served as the validation dataset. This procedure was repeated until each sub-sample had been used as the validation set. Four-fold cross validation may reduce the risk of overfitting and bias (Supplementary file 1, Fig. S1). In addition to this internal cross validation, the CC identification models were also externally validated using data from patients who were diagnosed in hospitals other than the ZOC. Assessments of accuracy, sensitivity, specificity, false positive rate, false negative rate, receiver operating characteristic (ROC) curve, and area under the curve (AUC) were performed to evaluate the discriminatory ability of the identification models. ROC and AUC, two important evaluation indexes for classifier performance in machine learning, were calculated according to previous reports [24,25] using R and Python. To test the identification models in a real-world clinical environment with a relatively low prevalence of CC, we applied the models to four clinical datasets with gradually decreasing proportions of CC patients. In addition, the importance of the variables among all investigated factors was scored by RF in both bilateral and unilateral cases.

2.4. Statistical analysis

The statistical analysis was performed using SPSS (version 19.0, IBM SPSS Inc., Chicago, Illinois, USA), R (version 3.4.2), and Python (version 3.5.2). The sample size was determined by PASS (version 15.05) with the following parameters setting: confidence interval (CI) formula, Score (Wilson); Interval type, Two-Sided; Confidence level (1-alpha), 0.95; CI width (two-sided), 0.1; Proportion, 0.8. The Shapiro–Wilk test and Levene's test were used to test the normal distribution and equality of variance of continuous variables. The

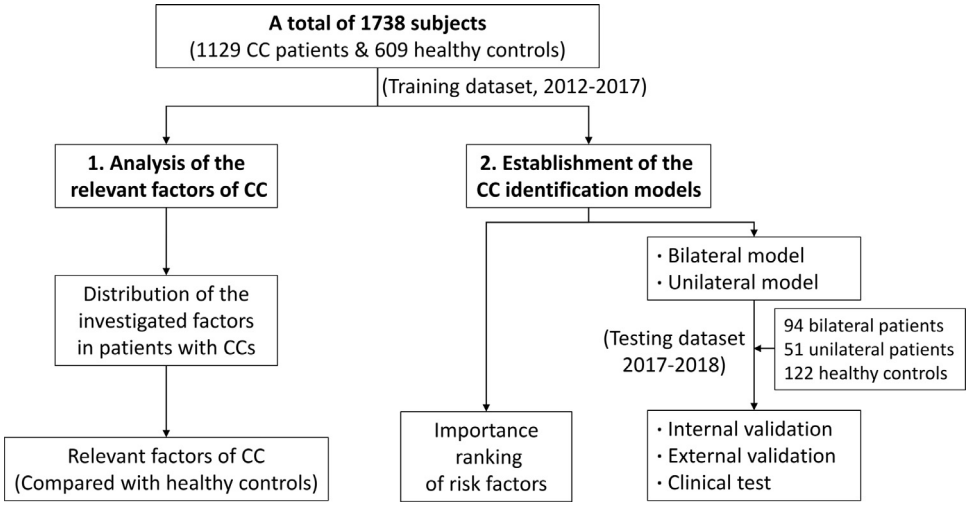

Fig. 1. Flowchart of the research performed in this study. CC: congenital/infantile cataracts.

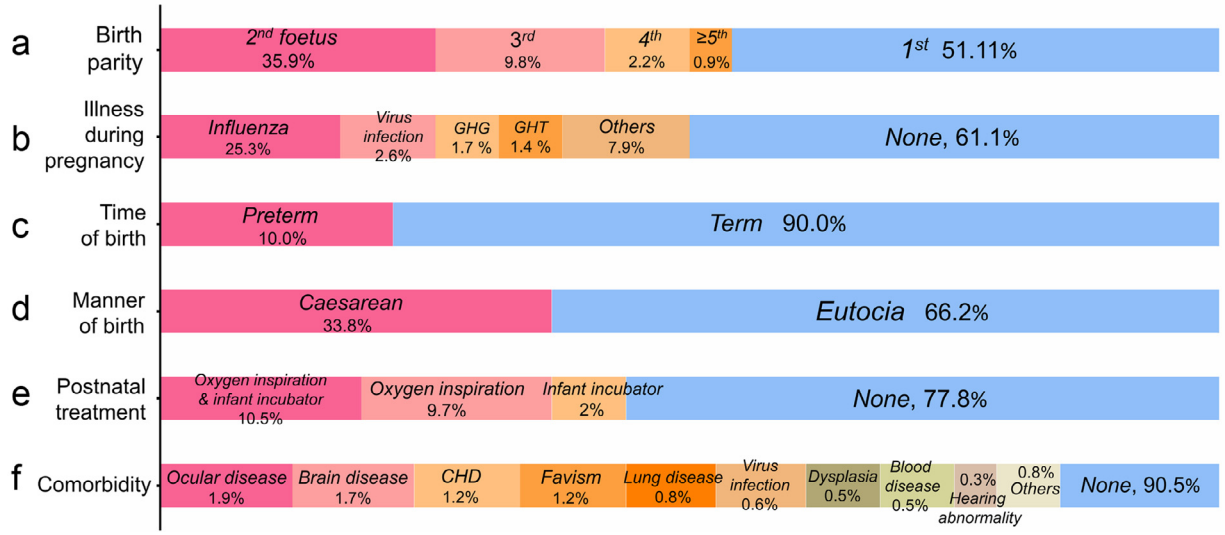

Fig. 2. Birth information of children with CCs. Notes: Only the patients with relevant data were included in the distribution analysis. GHG: gestational hyperglycaemia; GHT: gestational hypertension; CHD: congenital heart disease.

distributions of the potential predictive factors were compared between CC patients and healthy controls using Pearson's chi-squared test for categorical factors or Student's *t*-test for continuous factors. All tests were two tailed, and statistical significance was defined as $P < 0.05$.

3. Results

Fourteen CC cases and 8 healthy controls were excluded in the questionnaire quality control process. The primary dataset represented 1738 subjects, including 1129 patients with CC and 609 healthy controls. The mean age of these subjects was 37.95 ± 29.81 months (95% CI 36.42−39.23), and the ratio of males to females was 1.36 (1001:737). A total of 11 factors were summed up from the questionnaires for the distribution analysis and the establishment of identification models: family history of CCs, birth parity, virus infection during gestation, preterm delivery, eutocia, supplemental oxygen inspiration/infant incubator use, comorbidity, radiation/pollution, parental smoking, parental education level, and annual household income.

3.1. Analysis of relevant factors

Among the patients with CCs, bilateral cataract involvement represented most of the study population (71.48%, 807/1129), and the remaining patients had unilateral cataracts (28.52%, 322/1129).The birth conditions, family medical history and family environmental conditions of CC patients are shown in Figs. 2 and 3.

To analyze the possible predictive factors, we compared the eleven chosen factors between the CC patients and healthy controls (Table 1). Other than radiation/pollution exposure, the proportions of CC patients with all investigated factors were significantly higher than those of the healthy controls.

3.2. Models to identify children at high risk for CCs

The numbers of bilateral and unilateral patient, and healthy controls were 807:322:609 in the training dataset and 94:51:122 in the external validation dataset. Information on the training and external validation datasets is presented in Table S3 (Supplementary file 1). A missing-data mechanism analysis was performed. The most serious missing data were found in annual household

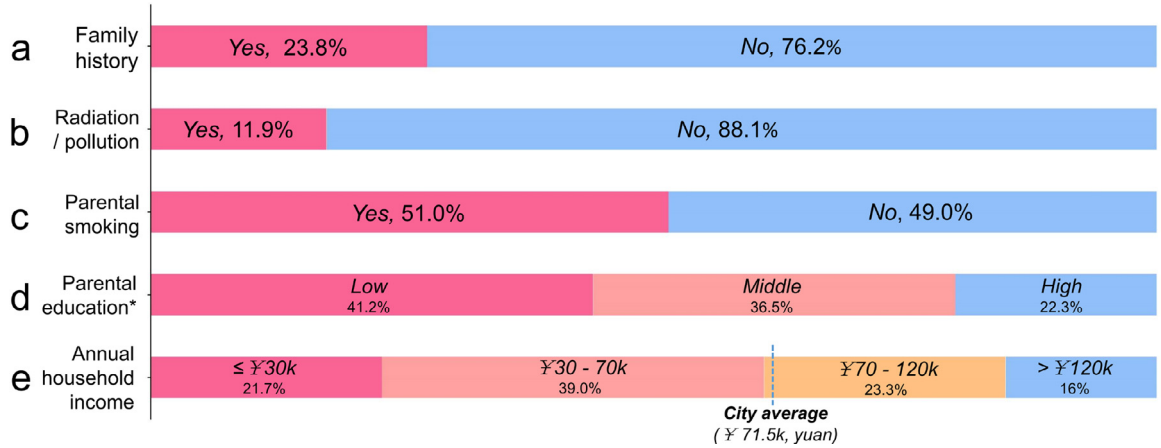

Fig. 3. Histories of family heredity and family conditions of children with CC. Notes: Only the patients with relevant data were included in the distribution analysis. *: Because no significant difference was found in education level between the fathers and mothers, the mothers were used to represent the parental education level in this study. ¥: Chinese Yuan.

Table 1
Comparisons of the pregnancy-labor history, living environment and family variables between the children with CCs and the healthy controls.

	Children with CCs	Healthy controls	χ^2/t	*P*
Number	1129 (Bil=807; Unil=322)	609	–	–
Age (months)	31.28±33.23	39.09±12.80	6.99	<0.001#
Male	59.9%	53.7%	6.194	0.015*
Family history	23.83% (269/1129)	0% (0/609)	171.67	<0.001*
≥2nd foetus	48.89% (419/857)	23.06% (140/607)	142.74	<0.001*
Pregnant virus infection	27.83% (310/1114)	20.39% (124/608)	11.53	0.001*
Preterm delivery	9.97% (112/1123)	3.78% (23/608)	21.02	<0.001*
Eutocia	66.19% (742/1121)	56.58% (344/608)	15.59	<0.001*
Oxygen inspiration/ infant incubator	22.17% (237/1069)	6.74% (41/608)	66.70	<0.001*
Comorbidity	11.78% (133/1129)	1.65% (10/606)	53.51	<0.001*
Radiation/pollution	11.86% (114/961)	9.20% (55/598)	2.71	0.111
Parental smoking	51.05% (537/1052)	34.44% (208/604)	42.77	<0.001*
Low/medium parental education level†	77.69% (846/1089)	33.06% (200/605)	327.95	<0.001*
Low household income¶	60% (391/644)	22.89% (111/485)	253.01	<0.001*

Notes: †: junior, primary and below; ¶: an average family income less than 71.5 K (Chinese yuan) was defined as a low household income; results are marked if statistically significant according to Pearson's chi-squared test (*) or an independent-sample *t*-test (#) ($P<0.05$); Bil: bilateral patients; Unil: unilateral patients.

Table 2
Performance of four-fold cross validation and external validation of the CC identification models.

			Accuracy	Sensitivity	Specificity	False negative rate	False positive rate
4-fold cross validation	Bilateral	RF	0.81±0.01	0.79±0.02	0.82±0.04	0.21±0.02	0.18±0.04
		Ada	0.79±0.02	0.78±0.03	0.81±0.03	0.22±0.03	0.19±0.03
	Unilateral	RF	0.79±0.01	0.56±0.05	0.92±0.03	0.44±0.05	0.08±0.03
		Ada	0.75±0.01	0.70±0.08	0.78±0.05	0.30±0.08	0.22±0.05
External validation	Bilateral	RF	0.86	0.80	0.91	0.20	0.09
		Ada	0.85	0.77	0.90	0.23	0.10
	Unilateral	RF	0.86	0.58	0.98	0.42	0.02
		Ada	0.85	0.58	0.97	0.42	0.03

Notes: RF: random forest; Ada: adaptive boosting.

income (35.1%), and the missing data related to the other 10 factors were considered missing at random. The missing data were imputed using the missForest algorithm. CC prediction models were trained in bilateral and unilateral patients separately. The accuracy, sensitivity, specificity, and false negative and positive rates of the CC identification models established by RF and Ada in the 4-fold cross validation and external validation are shown in Table 2. The ROC curves and AUC values of models with different algorithms and lateralities of cataracts (bilateral or unilateral) are compared in Fig. 4. The results show that CC prediction models

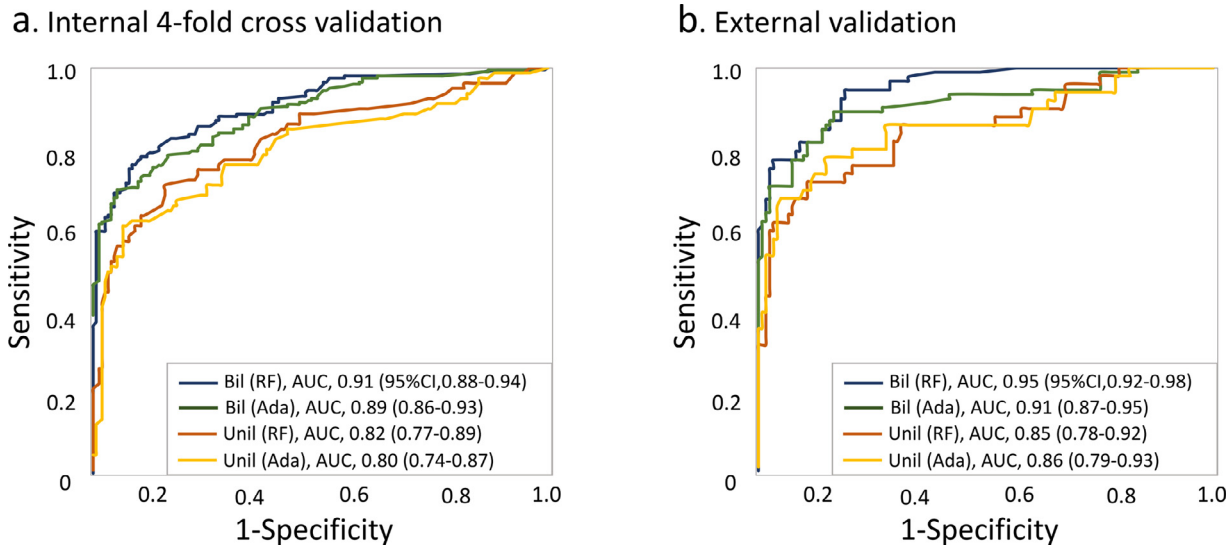

Fig. 4. ROC curves and AUC values of models with different algorithms and type of cataracts (bilateral or unilateral) in internal 4-fold cross validation and external validation. CC prediction models performed better in bilateral patients than in unilateral cases, and RF yielded better performance than Ada. ROC: receiver operating characteristic; AUC: area under the curve; RF: random forest; Ada: adaptive boosting; CI: confidence interval.

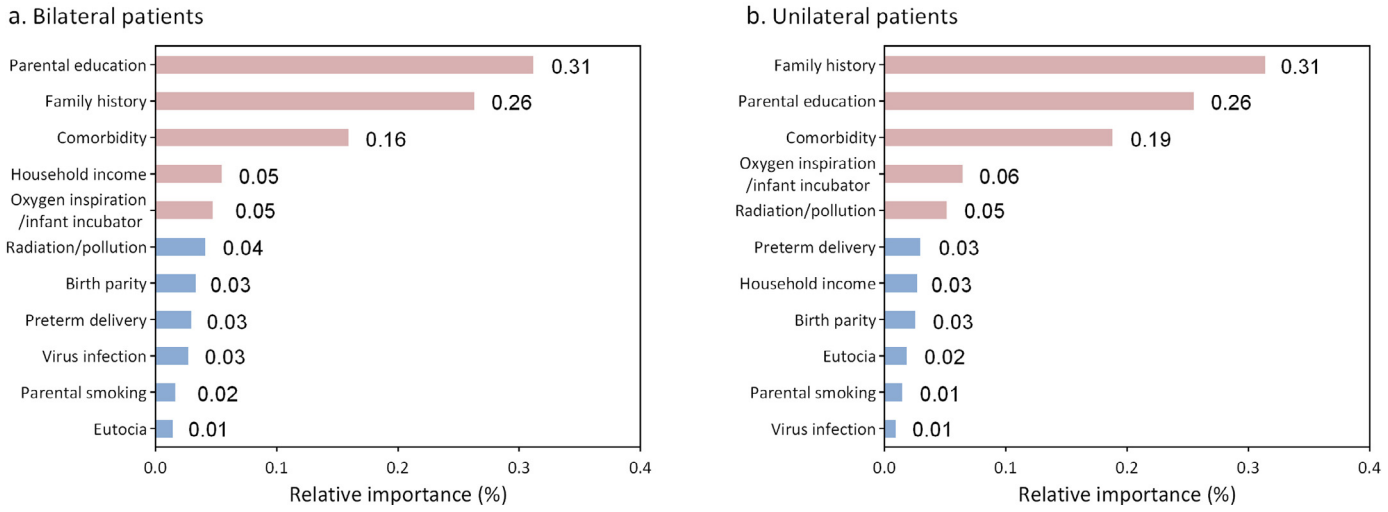

Fig. 5. Relevance ranks of the 11 relevant factors of CCs in bilateral and unilateral patients. Family history of CC, low parental education level, and comorbidity were identified as the top three most relevant factors to both bilateral and unilateral CC diagnosis.

Table 3
Clinical test of the stability of CC identification models.

Nos. of patients vs. controls	Algorithm	Accuracy	Sensitivity	Specificity	False negative rate	False positive rate
94 vs. 100	RF	0.86	0.80	0.93	0.20	0.07
(1:1)	Ada	0.88	0.82	0.93	0.18	0.07
50 vs. 100	RF	0.86±0.01	0.72±0.01	0.93±0.01	0.28±0.01	0.07±0.01
(1:2)	Ada	0.85±0.02	0.72±0.02	0.92±0.02	0.28±0.02	0.09±0.02
30 vs. 100	RF	0.88±0.01	0.69±0	0.93±0.01	0.31±0	0.07±0.01
(1:3)	Ada	0.87±0.02	0.72±0.02	0.91±0.03	0.28±024	0.09±0.03
10 vs. 100	RF	0.92±0.01	0.78±0	0.93±0.01	0.22±0	0.07±0.01
(1:10)	Ada	0.89±0.02	0.75±0.05	0.91±0.03	0.25±0.05	0.09±0.03

performed better in bilateral patients than in unilateral cases, and RF outperformed Ada. The model for the identification of children at high risk for CC based on RF is available (www.ccpmohprediction.cn) for the preclinical application testing stage.

To improve the development of prevention and identification strategies for CCs, we assessed and ranked the relative importance of the eleven factors using RF. As shown in Fig. 5, family history of CCs, parental education, and comorbidity exhibited the greatest relevance to CCs in both bilateral and unilateral cases. A close relationship between history of supplemental oxygen inspiration or infant incubator use and CCs was also observed.

3.3. *Clinical test of the CC identification models*

To evaluate the performance of the CC identification model using real CC patient in an approximation of the clinical environment with a relatively low disease prevalence, we tested the model in four clinical datasets with gradually decreasing proportions of CC patients. Ninety four patients with bilateral CCs and 100 randomly selected healthy controls from the external dataset were included in this clinical test. The ratios of CC patients to healthy controls in the 4 datasets were approximately 1:1, 1:2, 1:3, and 1:10. Four analyses using sampling with replacement were performed in each dataset to reduce the sampling error, except in the dataset with a 1:1 patient:control ratio. As shown in Table 3, the CC identification models showed good discriminatory ability between the bilateral CC patients and healthy children with high accuracies and AUC values (Fig. 6 (a) and (b)). Furthermore, the CC identification models achieved stable performance in the four subgroups and maintained acceptable accuracy and AUC values even in the 1:10 dataset.

4. Discussion

This study compared potential predictive factors for CCs between children with CCs and healthy controls, who exhibited distinct characteristics. Additionally, we established, to the best of our knowledge, the first practical identification model that effectively identifies children at high risk for CC based on 11 easily obtainable predictive factors. Benefiting from the advantages of feasibility and nearly zero cost, the models have the potential to compensate for the deficiency in current CC screening, especially in underdeveloped areas.

Most AI algorithms used for disease identification or prediction models have been trained based on different modalities of medical imaging [26,27], and the performance of medical AI techniques using non-imaging-based structured data remains unclear. Liang et al. [28] recently trained an AI model that demonstrated high diagnostic accuracy across common childhood diseases based on 101.6 million data points from 1,362,559 pediatric patients. However, collecting such a large volume of hospital-based medical data from patients with CCs and other rare congenital diseases is very difficult. In this study, we analyzed 11 variables in 2005 CC patients and healthy children and trained CC identification models, which exhibited satisfactory discrimination and stability, indicating that medical AI techniques achieve satisfactory performance in identifying congenital diseases even with limited data. Furthermore, the data used for the model training in the current study were collected noninvasively from both mothers and children.

Given the advantages of technical feasibility, and the ease and noninvasive nature of the data collection, this CC prevention and screening strategy has great potential. We designed a program (available at: www.ccpmohprediction.cn) to enable a wider range of testing after several rounds of validation, including internal cross

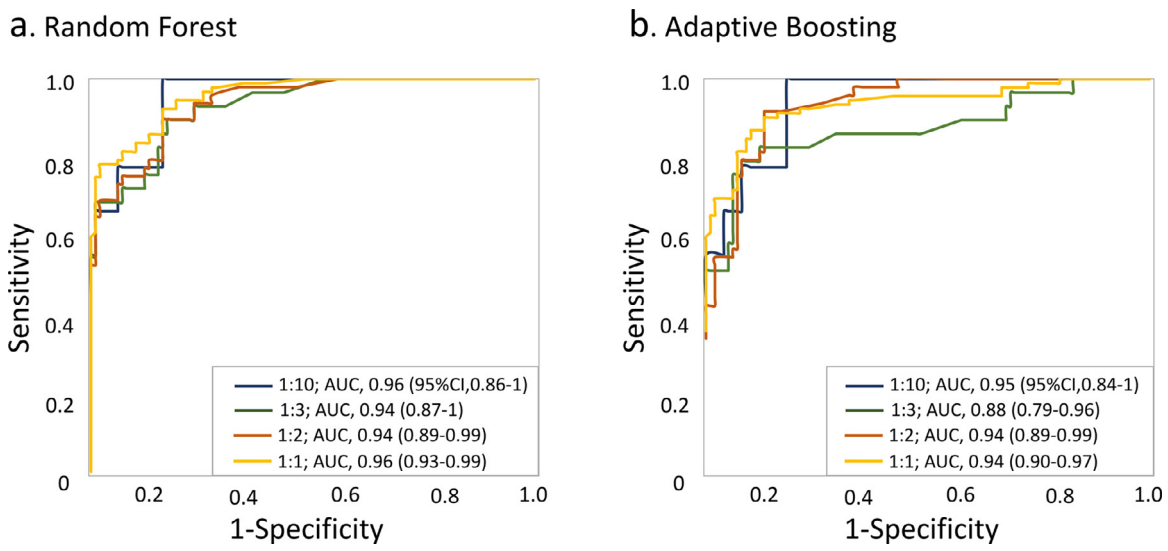

Fig. 6. ROC curves and AUC values of two AI algorithms for bilateral patients in the clinical test. ROC: receiver operating characteristic; AUC: area under the curve; AI: artificial intelligence; CI: confidence interval.

validation, external validation, and test in an approximation of the clinical environment. The State Council of China reported that 18.46 million infants were born at hospitals across the country in 2016 [29], one year after the two-child policy was announced. The number of new onset CC cases was estimated to be up to 13,715 per year in China based on the incidence of 7.43/10,000 among Asian populations [2]. Red reflex detection at 6–8 weeks after birth among newborns is recommended for CC screening in the UK [30] and other developed countries [31]. However, due to a shortage of medical resources, CC screening is not included in routine congenital disease screening programmes in China, and patients in undeveloped and remote areas can easily be missed [6]. Furthermore, some late-onset CCs are not manifested immediately after birth but develop gradually until the age of approximately 1 year. In addition, some mild cataracts (such as sutural cataracts) and peripheral cataracts tend to be missed by red reflex testing with undilated pupils, as carried out by non-ophthalmic health care workers. A practical identification model that can screen these infants with a high risk of CC but easily missed would be of great clinical significance, especially in undeveloped and remote areas. It is highlighted that novel digital health solutions could help overcome these clinical barriers by supporting timely diagnosis and referral [32]. Our screening program holds promise for reducing the missed diagnoses by serving as a complementary CC screening procedure. Although the screening model cannot replace screening by a qualified practitioner at present, it would be another way to timely identify those who may potentially benefit from medical intervention, meaning that some could avoid missing the best surgical timing among the key stage of visual development in early postnatal life.

To better understand how the subjects at high risk for CCs were identified and to improve the interpretability of the AI algorithms, we scored and ranked the analyzed risk factors according to their contribution to CC diagnosis by RF. Among these factors, family history of CCs, low parental education level, comorbidity, and history of supplemental oxygen inspiration or infant incubator use were identified as the most relevant factors to both bilateral and unilateral CC diagnosis. These findings are similar to those of previous studies. Nagamoto et al. [33] reported an increased prevalence (31.6%) of concomitant systemic abnormalities among patients with bilateral CCs. SanGiovanni et al. [34] found that preterm infants have 3- to 4-fold higher odds of developing infantile cataracts than those born at term. However, the relative importance of preterm birth was lower than that of supplemental oxygen inspiration or infant incubator use in the current study, indicating that the previously reported increase in infantile cataract risk may have occurred because supplemental oxygen inspiration and infant incubators are disproportionately used among preterm infants. Furthermore, detailed ocular examinations may be warranted among children whose parents have low education levels, although the causal link between this factor and CCs is not completely clear.

This is a preliminary study and there are some issues that should be discussed and thoroughly addressed before clinical application. Regarding study design, the external validation and clinical test in the current study were not so in the true sense. Although relatively stable performance was exhibited by our model under a series of patient proportions, a poorer performance may be shown in real world environment with a disease prevalence of approximately 4.24/10,000 [2] that much lower than those (from 1/10 to 1/1) evaluated in the current study. The model tests should be performed within the setting where the model would be used and in real world environment with a low disease prevalence, which would be a truer clinical validation process. The rarity of CC cases and potential selection bias in the dataset with too little CC cases prevented us from using these ideal methods. Furthermore, all patients in the external validation dataset were diagnosed in other institutes, but the data on potential predictive factors were collected in the ZOC after referral. In addition, the healthy controls were mainly recruited from kindergartens and communities, and the relatively uniform source of the population may create a bias. Regarding model application, although the CC identification models showed relatively high AUC values and stable performance, there is a possibility of missed diagnosis or misdiagnosis. Ocular disease is an important component of the predictive variable of comorbidity. However, except obvious abnormal ocular appearance such as severe strabismus and ptosis, most ocular morbidities cannot be easily detected by the parents, resulting a false negative in the variable of comorbidity. The performance of the model with this incorrect information input may weaken and lead to missed diagnosis. The factors of pregnancy-labor history [35] and living environment [36,37] are also important underlying causes of many other congenital diseases, which may cause misdiagnosis. The combined use of traditional screening methods, such as red reflex and slit-lamp examinations, is still necessary to improve CC identification and reduce the rates of missed diagnosis and misdiagnosis. Furthermore, most subjects in the training and validation of CC identification models were Chinese. Therefore, the results and the generalizability of the CC identification models should be interpreted with

caution due to differences in variables such as ethnicity, medical conditions, and socioeconomic status. Finally, other limitations included poor age matching between CC patients and healthy controls and the possible overlaps between CCs and other types of cataracts in the case recruitment.

In summary, this preliminary study established an accurate and practical identification model for CC screening based solely on AI analyses of eleven easily obtained predictive factors; the most relevant factors for CC development are also uncovered. Our findings are of great clinical significance for the early detection of CCs. The identification model has the potential to serve as a complementary screening procedure for the early detection or prediction of CC development, which could be especially useful in underdeveloped and remote areas. More broadly, our study may provide a reference for the development of AI-based preventive strategies for other congenital diseases.

Data sharing

The dataset and sub-datasets analyzed in the current study are not publicly available because they contain private patient data from the Zhongshan Ophthalmic Center (ZOC). The de-identified parts (information excluding patient identification and demography) are available from the corresponding author for research purposes upon reasonable request after approval by the institutional review board of ZOC at Sun Yat-sen University.

Funding sources

This study was funded by the National Key R&D Programme of China (2018YFC0116500), the China Postdoctoral Science Foundation (2018M640860, 2019T120773), Fundamental Research Funds for the Central Universities (18ykpy33 and 16ykjc28), the Key Research Plan for the National Natural Science Foundation of China in Cultivation Project (91846109), and Fundamental Research Funds of the State Key Laboratory of Ophthalmology (2018–2019). The funders played no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Declaration of Competing Interest

The authors declare that they have no competing interests.

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.ebiom.2019.102621.

References

[1] WHO fact sheet on congenital anomalies, updated September 2016, available at: www.who.int/mediacentre/factsheets/fs370/en/, accessed at 5th October 2018.
[2] Wu X, Long E, Lin H, Liu Y. Prevalence and epidemiological characteristics of congenital cataract: a systematic review and meta-analysis. Sci Rep 2016;6:28564.
[3] Stevens GA, White RA, Flaxman SR, et al. Global prevalence of vision impairment and blindness: magnitude and temporal trends, 1990–2010. Ophthalmology 2013;120(12):2377–84.
[4] Taruscio D, Arriola L, Baldi F, et al. European recommendations for primary prevention of congenital anomalies: a joined effort of Eurocat and Europlan projects to facilitate inclusion of this topic in the national rare disease plans. Public Health Genom 2014;17(2):115–23.
[5] Taruscio D, Mantovani A, Carbone P, et al. Primary prevention of congenital anomalies: recommendable, feasible and achievable. Public Health Genom 2015;18(3):184–91.
[6] Yang G, Zhong S, Zhang X, et al. Molecular genetic analysis of autosomal dominant late-onset cataract in a Chinese family. J Huazhong Univ Sci Technol Med Sci 2010;30(6):792–7.
[7] You C, Wu X, Zhang Y, Dai Y, Huang Y, Xie L. Visual impairment and delay in presentation for surgery in chinese pediatric patients with cataract. Ophthalmology 2011;118(1):17–23.
[8] Sheeladevi S, Lawrenson JG, Fielder A, et al. Delay in presentation to hospital for childhood cataract surgery in India. Eye 2018.
[9] Naz S, Sharif S, Badar H, Rashid F, Kaleem A, Iqtedar M. Incidence of environmental and genetic factors causing congenital cataract in children of Lahore. JPMA J Pak Med Assoc 2016;66(7):819–22.
[10] Mahalakshmi B, Therese KL, Devipriya U, Pushpalatha V, Margarita S, Madhavan HN. Infectious aetiology of congenital cataract based on torches screening in a tertiary eye hospital in Chennai, Tamil Nadu, India. Indian J Med Res 2010;131(131):559.
[11] Rahi JS, Dezateux C. Congenital and infantile cataract in the United Kingdom: underlying or associated factors. British congenital cataract interest group. Investig Ophthalmol Vis Sci 2000;41(8):2108–14.
[12] El Fkih L, Hmaied W, El Hif S, et al. Congenital cataract etiology. Tunis Med 2007;85(12):1025–9.
[13] Huo LA, Yang J, Zhang C. Regional difference of genetic factors for congenitals cataract. The results of congenital cataract screening under normal pupil conditions for infants in Tianjin city. Eur Rev Med Pharmacol Sci 2014;18(3):426.
[14] 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.
[15] Reid JE, Eaton E. Artificial intelligence for pediatric ophthalmology. Curr Opin Ophthalmol 2019;30(5):337–46.
[16] Lin H, Li R, Liu Z, et al. Diagnostic efficacy and therapeutic decision-making capacity of an artificial intelligence platform for childhood cataracts in eye clinics: a multicentre randomized controlled trial. EClinicalMedicine 2019;9:52–9.
[17] Lin D, Chen J, Lin Z, et al. 10-Year Overview of the hospital-based prevalence and treatment of congenital cataracts: the ccpmoh experience. PLoS ONE 2015;10(11):e0142298.
[18] Lambert SR, Drack AV. Infantile cataracts. Surv Ophthalmol 1996;40(6):427–58.
[19] Rahi JS, Dezateaux C. Measuring and interpreting the incidence of congenital ocular anomalies: lessons from a national study of congenital cataract in the UK. Investig Ophthalmol Vis Sci 2001;42(7):1444–8.
[20] Chen J, Lin Z, Lin H. Progress of application of sedation technique in pediatric ocular examination. Eye Sci 2014;29:186–92.
[21] Stekhoven DJ, Buhlmann P. MissForest–non-parametric missing value imputation for mixed-type data. Bioinformatics 2012;28(1):112–8.
[22] Lin H, Long E, Ding X, et al. Prediction of myopia development among Chinese school-aged children using refraction data from electronic medical records: a retrospective, multicentre machine learning study. PLoS Med 2018;15(11):e1002674.
[23] Friedman J, Hastie T, Tibshirani R. Special invited paper. Additive logistic regression: a statistical view of boosting. Ann Stat 2000;28(2):337–74.
[24] Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett 2006;27(8):861–74.
[25] Zhang K, Li X, He L, et al. A human-in-the-loop deep learning paradigm for synergic visual evaluation in children. Neural Netw 2019.
[26] Chilamkurthy S, Ghosh R, Tanamala S, et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. Lancet 2018;392(10162):2388–96.
[27] 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.
[28] Liang H, Tsui BY. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. Nat Med 2019;25(3):433–8.
[29] Xinhua. China's two-child policy results in largest number of newborns since 2000. 2017; Available at: http://www.xinhuanet.com//english/2017-03/11/c_136120854.htm. Accessed at 4th October 2018.
[30] Russell HC, McDougall V, Dutton GN. Congenital cataract. BMJ 2011;342:d3075.
[31] Haargaard B, Nystrom A, Rosensvard A, Tornqvist K, Magnusson G. The pediatric cataract register (PECARE): analysis of age at detection of congenital cataract. Acta Ophthalmol (Copenh) 2015;93(1):24–6.
[32] Advances in treatment and diagnosis of vision impairment. EBioMedicine 2019;48:1–2.
[33] Nagamoto T, Oshika T, Fujikado T, et al. Clinical characteristics of congenital and developmental cataract undergoing surgical treatment. Jpn J Ophthalmol 2015;59(3):148–56.
[34] SanGiovanni JP, Chew EY, Reed GF, et al. Infantile cataract in the collaborative perinatal project: prevalence and risk factors. Arch Ophthalmol 2002;120(11):1559–65.
[35] Bagri DR, Yadav KS, Sharma R, Gulati S. Congenital B-cell acute lymphoblastic leukemia with congenital rubella infection. Indian Pediatr 2019;56(1):67–8.
[36] Wang X, Li P, Chen S, et al. Influence of genes and the environment in familial congenital heart defects. Mol Med Rep 2014;9(2):695–700.
[37] Triche EW, Hossain N. Environmental factors implicated in the causation of adverse pregnancy outcome. Semin Perinatol 2007;31(4):240–2.