OPEN ACCESS

Predicting the 10-year risk of cataract surgery using machine learning techniques on questionnaire data: findings from the 45 and Up Study

Wei Wang ,[1] Xiaotong Han ,[1] Jiaqing Zhang,[1] Xianwen Shang,[2,3] Jason Ha,[3] Zhenzhen Liu ,[1] Lei Zhang,[3,4,5] Lixia Luo ,[1] Mingguang He [1,3,6]

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bjophthalmol-2020-318609).

For numbered affiliations see end of article.

Correspondence to
Professor Lixia Luo, Zhongshan University Affiliated Eye Hospital, Guangzhou, China; luolixia@gzzoc.com

LL and MH contributed equally.

LL and MH are joint senior authors.

Received 8 December 2020
Revised 22 April 2021
Accepted 5 May 2021

ABSTRACT

Background/aims To investigate the feasibility and accuracy of using machine learning (ML) techniques on self-reported questionnaire data to predict the 10-year risk of cataract surgery, and to identify meaningful predictors of cataract surgery in middle-aged and older Australians.

Methods Baseline information regarding demographic, socioeconomic, medical history and family history, lifestyle, dietary and self-rated health status were collected as risk factors. Cataract surgery events were confirmed by the Medicare Benefits Schedule Claims dataset. Three ML algorithms (random forests [RF], gradient boosting machine and deep learning) and one traditional regression algorithm (logistic model) were compared on the accuracy of their predictions for the risk of cataract surgery. The performance was assessed using 10-fold cross-validation. The main outcome measures were areas under the receiver operating characteristic curves (AUCs).

Results In total, 207 573 participants, aged 45 years and above without a history of cataract surgery at baseline, were recruited from the 45 and Up Study. The performance of gradient boosting machine (AUC 0.790, 95% CI 0.785 to 0.795), RF (AUC 0.785, 95% CI 0.780 to 0.790) and deep learning (AUC 0.781, 95% CI 0.775 to 61 0.786) were robust and outperformed the traditional logistic regression method (AUC 0.767, 95% CI 0.762 to 0.773, all p<0.05). Age, self-rated eye vision and health insurance were consistently identified as important predictors in all models.

Conclusions The study demonstrated that ML modelling was able to reasonably accurately predict the 10-year risk of cataract surgery based on questionnaire data alone and was marginally superior to the conventional logistic model.

To cite: Wang W, Han X, Zhang J, *et al*. *Br J Ophthalmol* Epub ahead of print: [*please include* Day Month Year]. doi:10.1136/bjophthalmol-2020-318609

INTRODUCTION

The combination of machine learning (ML) technology and big data has attracted considerable interests and has been adopted with great success in various research fields, including medicine, genomics, epidemiology and economics.[1 2] Predictive medicine provides clinicians with a familiar concept, through which they can look for related characteristics to identify high-risk subjects and corroborate with their diagnosis or prognosis. It is well believed that predictive medicine will mature with the development of big data and data modelling techniques, to the point that it will assist ophthalmologists in making clinical decisions and making itself a new tool for disease prediction.[3]

Cataract is the leading cause of visual impairment and blindness globally. The only effective treatment for cataract is surgery, which is the most commonly performed clinical procedure worldwide.[4] With ageing of the population, the global burden of cataract is projected to escalate substantially, and it is increasingly important to identify modifiable risk factors for cataract that requires surgery as targets for preventative measures. Older age, diabetes mellitus, ultraviolet radiation and steroid use have been widely reported as risk factors for cataract.[5 6] However, the impacts of other modifiable factors for cataract and cataract surgery, including physical activity, alcohol intake, obesity, diet and female reproductive factors, are inconclusive.[7–9] Additionally, the generalisability of previous studies has been mostly limited by population-based data with limited sample size.

Conventional regression techniques are commonly used for constructing predictive models, which requires to select variables with priori assumptions based on data distribution during the development process, leading to potential loss of information.[10] Although ML stands out as an effective and efficient technique for predictive medicine, few ML models exist for the prediction of cataract development, and to the best of our knowledge, there have been no mature prediction models for the risk of cataract surgery built on prospective cohort data collected over an extended period.[11]

Our study aims therefore to evaluate the performance of different ML algorithms in comparison to the conventional logistic model for the prediction of cataract surgery based on the prospective population-based study with 10 years of follow-up data in Australia.

MATERIALS AND METHODS

Participants

The Sax Institute's 45 and Up Study is a large-scale prospective cohort study undertaken in New South Wales (NSW), Australia. Participants aged 45 and older were randomly sampled from the general population using the Services Australia (formerly the Australian Government Department of Human Services) Medicare enrolment database. In total, 267 153 participants were recruited from 2006 to 2009, corresponding to 11% of the entire NSW population of this age group. Each participant

completed a mailed self-administered questionnaire at the baseline, capturing information on a broad range of demographic, socioeconomic, medical and lifestyle factors. These participants were also linked to a range of Australian health databases, including the Medicare Benefits Schedule (MBS) that tracks claims records for diagnostic tests and procedures and the Pharmaceutical Benefits Scheme (PBS) that provides information on medications dispensed to the participants in the community, facilitated by the Sax Institute using a unique identifier provided by Services Australia. The method used to link records in the MBS and PBS is deterministic matching. The MBS data were available from 24 January 2001 to 31 December 2016, and the PBS data were available from 1 June 2004 to 31 December 2016. Therefore, all procedures and medications received by the participants could be tracked from baseline to the end of 2016. The study methods have been extensively described in earlier studies.[12 13]

Participants without history of cataract surgery at baseline based on their responses to the baseline questionnaire and the MBS database records were included in the study. To ensure that only those with visually impaired, age-related cataract were captured in this study, we excluded cataract surgery performed for other causes such as congenital cataract, secondary cataract, and traumatic cataract and other conditions via MBS codes for juvenile cataract extraction, corneal surgery, scleral surgery, glaucoma procedures, vitreoretinal procedures and traumatic surgeries during the follow-up period (for further information, please refer to the online supplemental material 1).

Written informed consent was obtained from all participants for linkage of their information to routine health databases.

Predicting variables

In contrast with the previous epidemiological studies using biological data and ocular parameters that required ongoing clinical investigations and examinations, ML algorithms used in this study were derived completely from the information obtained from the baseline questionnaire completed by the participants. All variables of this study were self-reported, which could be easily obtained from the participants, without examinations. The variables were classified into five categories: demographic characteristics, chronic diseases/family history, lifestyle/dietary indicators, self-rated health status and social support/psychological distress. The definition and the classification of each variable are detailed in the online supplemental material 1.

Outcome definition

The primary outcome was the first occurrence of cataract surgery for age-related cataract, which was defined as the corresponding MBS codes detailed in the supplemental material. Participants without MBS claims records of cataract surgery during the study period (up to 2016) were defined as persons who did not undergo cataract surgery.

Algorithms used for prediction

One traditional regression model and three state-of-the-art ML models (all available online) were constructed to predict the risk of cataract surgery, and their relative performance was compared. The frameworks of the models' construction are illustrated in figure 1. In brief, the dataset was split into the training (60%) and the validation (40%) cohorts, with all models tuned using a 10-fold cross-validation. As a benchmark, a logistic regression model was employed to characterise the association between the predictor variables and the subsequent incidence of cataract surgery. Three ML models, including the gradient boosting machine (GBM), random forest (RF) and multilayer feedforward deep learning (DL) models, were used for predicting cataract surgery. Apart from the basic framework, the sensitivity analyses were performed by using the participants in cities as training set for models building, and the participants in regional/rural as validation set, considering the huge difference in lifestyle, socioeconomic status, diet, etc, between the populations (online supplemental figure 1).

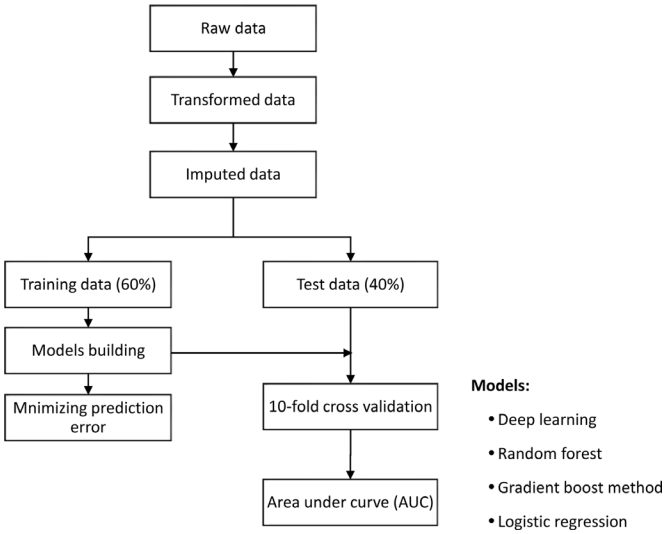

Figure 1 A general framework of the machine learning algorithm.

Statistical analysis

The ML modelling was performed using R V.3.4.1 (R Programming) with toolbox h2o V.3.16.0.2, while other statistical analyses were performed using SAS V.9.4 (SAS Institute). Descriptive statistics, including frequencies and proportions, were used to characterise the study population. χ^2 tests were performed for categorical variables to compare baseline characteristics of participants with and without claim records of cataract surgery. Poisson regression models with robust variance were used to evaluate associations of potential predictors with the risk of cataract surgery. Potential predictor variables were identified for each respective model and automatically ranked by information gain. The area under the receiver operating characteristic curve (AUC) was used to compare the performance of different models. A p value <0.05 was considered to be of statistical significance.

RESULTS

Participant selection and baseline characteristics

Among 267 153 participants recruited at baseline (2006–2009), 59 580 were excluded due to cataract surgery performed for reasons other than age-related cataract, leaving 207 573 participants eligible for the final analysis (online supplemental figure 2). During the median 9-year follow-up period (range of 7.0–11.5 years), 23 573 (11.4%) eligible participants had linked MBS claims for cataract surgery, and the remaining 184 000 (88.6%) participants had not claimed for cataract surgery. The eligible participants' baseline characteristics were summarised in online supplemental table 1. There were 107 427 participants in cities and 100 146 in regional or rural regions, with significant difference in most variables at baseline (online supplemental table 2).

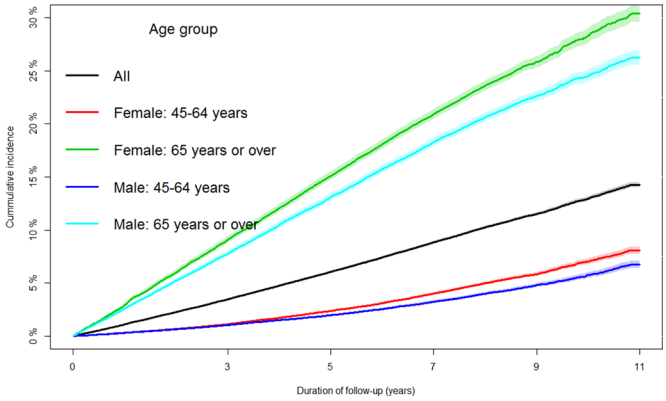

Figure 2 Cumulative incidences of cataract surgery stratified by age and gender during follow-up.

Cumulative incident cataract surgery by gender and age groups

Overall, the cumulative incidence of cataract surgery was 11.4% during the follow-up period (figure 2). Cataract surgery incidence increased steadily with age, with 5.31% of the participants aged 45–64 (younger age group) and 23.84% of participants aged ≥65 at baseline (older age group) requiring cataract surgery during the subsequent follow-up. This higher cumulative incidence of cataract surgery with age was observed irrespective of gender; however, female participants had a higher cumulative incidence in both age groups during follow-up.

Prediction of cataract surgery risk with ML techniques

All predictive models demonstrated reasonably high predictive accuracy for cataract surgery (figure 3). In particular, the GBM achieved an AUC of 0.790 (95% CI 0.785 to 0.795), the highest among the four models, followed by RF (AUC 0.785, 95% CI 0.780 to 0.790), deep learning (AUC 0.781, 95% CI 0.775 to 0.786) and logistic regression (AUC 0.767, 95% CI 0.762 to 0.773). The superior performance of the three ML models compared with the conventional logistic regression was statistically significant (all $p<0.05$).

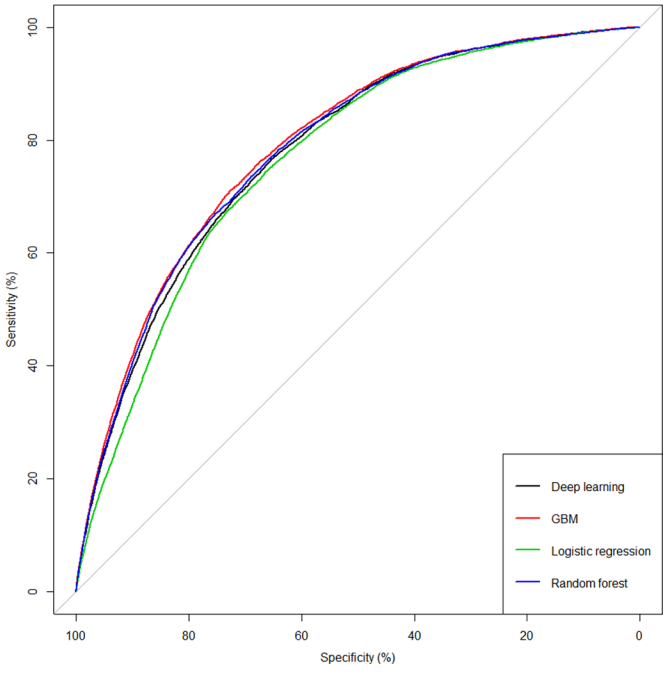

Figure 3 Receiver operating characteristic curves for each cataract surgery prediction model. GBM, gradient boosting machine.

The relative importance of the predictive variables in the logistic regression and ML models is summarised in online supplemental figure 3. As expected, all models consistently demonstrated age as the most important predictor for cataract surgery, contributing 30% (deep learning) to 95% (logistic regression) of the variance of their respective models alone. Health insurance was ranked as the second most important predictor in three models (deep learning, RF and logistic regression), explaining over 10% of the prediction variance in each model.

Sensitivity analysis (external validation)

When the prediction models were constructed based on participants in city, the external validation of prediction models achieved similar results with the primary analysis, with AUC of 0.768 for logistic model, 0.786 for RF model, 0.790 for GBM model and 0.782 for DL model. Both GBM and RF models had significantly better performance than logistic model (all $p<0.05$).

DISCUSSION

To date, systematic identification and ranking of predictors of cataract surgery have been unavailable for the middle-aged and elderly population in developed countries. Our study presents the first attempt to use data from a prospective population-based cohort to evaluate the performance of different ML models on risk prediction for cataract surgery and identify key predictive variables without reliance on ocular biometric data. Our results show that the 10-year risk cataract surgery could be adequately predicted through self-reported variables only. All models demonstrated reasonable performance, with AUCs ranging from 0.767 to 0.790. The GBM and RF approaches outperformed the conventional logistic regression model and were well suited for accurate risk prediction of cataract surgery.

Accurate prediction of cataract surgery remains a major challenge in ophthalmic public health. Given the paucity of population-wide ocular biometric data, this study demonstrates that employing ML models on self-reported questionnaire responses alone is sufficient to undertake this cataract surgery prediction. This mirrors the success of these predictive ML models in other areas of oncological, cardiovascular, and ophthalmic disease research, including studies that have accurately predicted risks of high myopia, and improvement in prediction of visual acuity post-anti-vascular endothelial growth factor (VEGF) treatment.

Our study proved that ML methods, which took into account all available questionnaire-based risk factors for automatic evaluation and modelling, were more suitable for both individual risk prediction and population surveillance. The ML algorithms differ from conventional prediction techniques with its independence of prior assumptions, which avoids the possibility to overlook the unexpected but significant variables, or to identify essential risks in patients with several marginal risk factors (or no risk factors at all).[1 14] Another nature of ML algorithm lies in the minimum input in the developing model stage, which can seamlessly update and optimise with new data, thus leading to higher performance of the model over time. The risk of cataract surgery was also predicted by DL, which is the most advanced ML technique. DL plays an indispensable role in certain fields, such as image recognition, self-driving cars, Google DeepMind AlphaGo/AlphaZero and machine vision software in cameras.[15–17]

However, the DL model did not exert higher performance than GBM model in predicting cataract surgery in this study.

There are both clinical and methodological implications from the study findings. From a clinical perspective, our study has shown remarkable robustness of ML models to accurately predict of cataract risk from a wide range of metrics from self-reported questionnaire data. Of note, most of metrics are demographic, socioeconomic and lifestyle factors that are amenable for targeted preventative measures. This may facilitate clinical identification of high-risk cataract surgical candidates, estimation of future disease burden, as well as the application of preventative health measures by governments and health organisations to guide targeted public health policies, and manage medical resource allocation.[18]

From a methodological standpoint, this study has proven the feasibility and superiority of ML models that take a comprehensive list of predictive risk factors for automatic evaluation and modelling in predicting cataract surgery risk. The accumulation of electronic medical data in hospitals and the increasing availability of data collection methods (eg, wearable devices, online questionnaires) may pave the way for further future studies on risk prediction using ML methods on a larger dataset with more variables, a larger sample size, a longer follow-up period and more diverse populations.

This study's strengths lie in its population-based prospective cohort design, large sample size, the availability of a comprehensive set of variables and its long-term follow-up. However, several limitations of this study should be noted as well. First, the definition of incident cataract surgery based on MBS in our study may be biased because the diagnosis of cataract is dependent on healthcare-seeking behaviour, coexisting diseases and accessibility. Those who have a great concern about their health would tend to report the positive answer to the questionnaire and also gets the cataract surgery at the early stage of cataract formation. Second, the expected applications of ML models were not evaluated in this study. For example, RNN or survCNN models were not used. Whether an objective ML algorithm without a prior hypothesis outruns the subjective but flexible human mind is a potential topic for future study.

CONCLUSIONS

Using extended 10-year follow-up data from the large-scale population-based 45 and Up Study, we have applied ML methods and established high accuracy risk prediction models for cataract surgery, based on non-clinical self-reported questionnaires. Future applications of ML-based prediction models using large datasets can facilitate powerful disease prediction tools and inform public health change at both individual and governmental levels.

Author affiliations
[1]State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China
[2]Guangdong Eye Institute, Department of Ophthalmology, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou, China
[3]Centre for Eye Research Australia, Royal Victorian Eye and Ear Hospital, Melbourne, Victoria, Australia
[4]Department of Epidemiology and Biostatistics, School of Public Health, Xi'an Jiaotong University Health Science Center, Xi'an, China
[5]School of Public Health and Preventive Medicine, Monash University, Melbourne, Victoria, Australia
[6]Ophthalmology, Department of Surgery, University of Melbourne, Melbourne, Victoria, Australia

Acknowledgements This research was completed using the data collected from the 45 and Up Study (www.saxinstitute.org.au). The 45 and Up Study is managed by the Sax Institute, in collaboration with its major partner Cancer Council New South Wales (NSW) and the following partners: the National Heart Foundation of Australia (NSW Division); NSW Ministry of Health; NSW Government Family & Community Services–Ageing, Careers and the Disability Council NSW; and the Australian Red Cross Blood Service. We acknowledge that Services Australia supplied the Medicare Benefits Schedule (MBS) and Pharmaceutical Benefits Scheme (PBS) data to the Sax Institute. The authors thank the many thousands of people who participated in the 45 and Up Study.

Contributors Writing—Original draft: WW, XH, JZ. Writing—Review and editing: WW, XH, JZ, XS, JH, ZL, LZ, LL, MH. Conceptualisation: LZ, LL, MH. Project administration: LZ, LL, MH. Investigation: WW, XS. Methodology: WW, XS, LZ. Visualisation: LZ, LL, MH. All the authors have been involved in the study conception, data analysis, interpretation of the findings and conclusions.

Funding Professor MH receives support from the University of Melbourne under the Research Accelerator Program and the Centre for Eye Research Australia (CERA) Foundation. CERA receives operational infrastructure support from the Victorian State Government. Professor MH is supported by NHMRC Investigator Grant (GNT1175405) and by the Fundamental Research Funds of the State Key Laboratory in Ophthalmology, National Natural Science Foundation of China (81420108008). Professor LL is supported by the Construction Project of High-Level Hospitals in Guangdong Province (303020102). The sponsor or funding organisation had no role in designing or conducting this research.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval The conduct of the 45 and Up Study was approved by the University of New South Wales Human Research Ethics Committee (HREC). The study protocol was approved by the Royal Victorian Eye and Ear Hospital Human Research Ethics Committee (17/1330HS/20).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data may be obtained from a third party and are not publicly available.

ORCID iDs
Wei Wang http://orcid.org/0000-0002-5273-3332
Xiaotong Han http://orcid.org/0000-0001-6836-3447
Zhenzhen Liu http://orcid.org/0000-0002-4853-2474
Lixia Luo http://orcid.org/0000-0002-4612-2906
Mingguang He http://orcid.org/0000-0002-6912-2810

REFERENCES

1 Deo RC. Machine learning in medicine. *Circulation* 2015;132:1920–30.
2 Ginsberg J, Mohebbi MH, Patel RS, *et al*. Detecting influenza epidemics using search engine query data. *Nature* 2009;457:1012–4.
3 Yan W, He M. Predictive medicine in ophthalmology. *Ophthalmology* 2017;124:420–1.
4 Lam D, Rao SK, Ratra V, *et al*. Cataract. *Nat Rev Dis Primers* 2015;1:15014.
5 West SK, Duncan DD, Muñoz B, *et al*. Sunlight exposure and risk of lens opacities in a population-based study: the Salisbury eye evaluation project. *JAMA* 1998;280:714–8.
6 Floud S, Kuper H, Reeves GK, *et al*. Risk factors for cataracts treated surgically in postmenopausal women. *Ophthalmology* 2016;123:1704–10.
7 Lindblad BE, Håkansson N, Philipson B, *et al*. Hormone replacement therapy in relation to risk of cataract extraction: a prospective study of women. *Ophthalmology* 2010;117:424–30.

8 Tan AG, Kifley A, Tham Y-C, *et al*. Six-Year incidence of and risk factors for cataract surgery in a multi-ethnic Asian population: the Singapore epidemiology of eye diseases study. *Ophthalmology* 2018;125:1844–53.
9 Tian Y, Wu J, Xu G, *et al*. Parity and the risk of cataract: a cross-sectional analysis in the Dongfeng-Tongji cohort study. *Br J Ophthalmol* 2015;99:1650–4.
10 Alanazi HO, Abdullah AH, Qureshi KN. A critical review for developing accurate and dynamic predictive models using machine learning methods in medicine and health care. *J Med Syst* 2017;41:69.
11 Krittanawong C, Zhang H, Wang Z, *et al*. Artificial Intelligence in Precision Cardiovascular Medicine. *J Am Coll Cardiol* 2017;69:2657–64.
12 Yan X, Han X, Wu C, *et al*. Does daily dietary intake affect diabetic retinopathy progression? 10-year results from the 45 and up study. *Br J Ophthalmol* 2020;104:1774–80.
13 Shang X, Peng W, Hill E, *et al*. Incidence of Medication-Treated depression and anxiety associated with long-term cancer, cardiovascular disease, diabetes and osteoarthritis in community-dwelling women and men. *EClinicalMedicine* 2019;15:23–32.
14 Shameer K, Johnson KW, Glicksberg BS, *et al*. Machine learning in cardiovascular medicine: are we there yet? *Heart* 2018;104:1156–64.
15 Silver D, Huang A, Maddison CJ, *et al*. Mastering the game of Go with deep neural networks and tree search. *Nature* 2016;529:484–9.
16 Silver D, Schrittwieser J, Simonyan K, *et al*. Mastering the game of Go without human knowledge. *Nature* 2017;550:354–9.
17 Silver D, Hubert T, Schrittwieser J, *et al*. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science* 2018;362:1140–4.
18 Karczewski KJ, Snyder MP. Integrative omics for health and disease. *Nat Rev Genet* 2018;19:299–310.