European Review for Medical and Pharmacological Sciences 2022; 26: 7344-7348

A study to create a risk factor models to predict the development of keratoconus among the Saudi Arabian population

S.M. ALJOHANI[1,2]

[1]College of Applied Medical Sciences, Qassim University, Qassim, Saudi Arabia
[2]College of Health Sciences, Salus University, PA, USA

Abstract. – OBJECTIVE: An interaction between hereditary and environmental variables is thought to be the cause of keratoconus, a progressive ectatic corneal condition. The identification of risk factors is necessary since they are currently the subject of intense debate and are crucial to the management and prevention of the disease. The objective of this study is to gain a better understanding of the risk factors associated with the onset and progression of keratoconus. It would be valuable for both eye care professionals and patients in Saudi Arabia.

PATIENTS AND METHODS: Patients seeking eye care at Qassim University eye clinic were included in this study. Participants were divided into: cases (with keratoconus) and control (without keratoconus but with other ocular problems). Keratoconus diagnoses of the participants were made by the attending optometrists or ophthalmologists. Multivariate logistic analyses were performed to identify the risk factors for keratoconus. Moreover, by performing logistic regression and CART analysis, supervised learning algorithms were developed to predict the likelihood of keratoconus based on the risk factors.

RESULTS: There were 75 keratoconus patients and 75 control. The CART model to predict the chances of keratoconus occurrence has an accuracy of 73%. Our prediction model can be a baseline model for future risk factor analysis studies that will be done in the Middle Eastern region. The models can be better trained by refining the risk factor quality and also by increasing the keratoconus population in the study. Including clinical parameters in the prediction models would result in complex as well as models with better prediction accuracy.

CONCLUSIONS: Clinical ocular parameters including the corneal topographic variables have to be obtained to better correlate the risk factors with specific changes or the subtypes of the keratoconus. Complex diseases like keratoconus require machine learning models apart from statistical analysis for association and causation. Machine learning models would not only predict the disease but also provide insight into how the risk factors interact with each other.

Key Words:
Keratoconus, Risk factors, Qassim, Saudi Arabia.

Introduction

The estimated keratoconus varies in different parts of the world, with a relatively higher prevalence in Middle Eastern countries[1]. The Saudi population has typical characteristics of kinship and larger family size that might significantly contribute to the prevalence of keratoconus in Saudi Arabia[2]. Literature[3-5] supports that first cousin marriages were more likely to produce offspring with keratoconus. The onset of keratoconus has been linked to parental consanguinity. Thus, the role of genetics in the prevalence of keratoconus and progression is important[6]. A Middle Eastern study[6] showed that children born from consanguineously married (first and second cousins) parents have a 4-fold risk of developing keratoconus. A European study[7] concluded that the familial association was present in 19 out of 101 (19%) families and 5 out of 58 (9%) families, with a higher proportion associated with patients from greater family sizes. The Saudi Arabian culture has more consanguineous marriages and larger family sizes than other cultures contributing significantly to our research questions[2].

Machine learning deals with computational learning by utilizing pattern recognition as its principle. It extracts knowledge or information from the data, which is then used as input[8]. Machine learning acknowledges the concepts related to the study and formulation of algorithms that can learn from and make predictions on the given data load[8,9]. Dealing

Corresponding Author: Saad Aljohani, Ph.D; e-mail: sa.aljuhani@qu.edu.sa

with complex diseases before their expression is a better approach than combating the complications, and for this purpose, machine learning and artificial intelligence have proved to be provident[10-12]. They have been demonstrated[12] to be a valuable resource for the healthcare industry by enabling it to achieve much higher accuracy in predicting the occurrence of a disease. Supervised machine learning is the most efficient and widely used type of machine learning[13]. It is used when there is a need to predict a specific outcome from a given input and real source input/output pairs are present[10,13]. These source data were used to train the algorithm to make accurate predictions for new unknown data. In supervised learning, the clinician's effort is required to build the dataset with as many variables as possible but afterward automates and often speeds up an otherwise laborious or infeasible task, even for a complex disease condition[11,13]. This study is designed to develop and compare supervised learning algorithms in predicting keratoconus based on the risk factors. We chose to perform a logistic model, a commonly used prediction model to describe the relationship between an outcome and the risk factors, and a classification modeling with better prediction accuracy.

Patients and Methods

Sample Size

The detectable effect size was calculated based on the contingent Chi-square test, comparing the prevalence of a family history of keratoconus between cases and controls with alpha = 0.05 and power = 0.8. The proportion of controls having a family history of keratoconus was estimated at 1.7%, which was the prevalence of having a family history of keratoconus[5]. The study population in Millodot et al[5] was from Middle East, as the study population in this dissertation. The sample was set at 75 cases and 75 controls, a feasible sample size for this study. With these parameters, the detectable odds ratio is 11.0. Although this is a large effect, the Millodot et al[5] study found an odds ratio of 17.1, larger than the detectable. Therefore, the sample size of 75 cases and 75 controls is sufficient to detect a realistic effect size for a family history of keratoconus.

The study was approved by the Ethics Committe of Qassim University and Salus University institutional review board. It was conducted according to the ethical principles of the Declaration of Helsinki and by the current legislation on clinical research rules developed by the National Committee of Bioethics in Saudi Arabia. All participants signed an informed consent form before initiating the study.

Data Collection

Patients seeking eye care at Qassim University eye clinic were included in the study. Participants were contacted through emails and by posting flyers on the Qassim University campuses in Buraidah. Interested people were contacted and scheduled for a screening to determine their eligibility for the study based on the inclusion and exclusion criteria. If eligible, they were given a full explanation of their involvement and role in the study, and written consent was obtained. Because the study focused on assessing risk factors of the Saudi population who had lived and experienced the Saudi lifestyle since birth, it only targeted Saudi nationals. The age range for the participants was 10-30 years.

Statistical Analysis

We modeled the risk factors using R Studio (V 1.2.5001, RStudio, Boston, MA, USA). The first model was the logistic regression, and the second model was the Classification and Regression Trees (CART). Both the models were compared for prediction accuracy. The logistics model was suitable to perform when the outcome variable was discrete and, in our study, it was dichotomous, that is presence or absence of keratoconus.

The strengths of CART are that it applies to complex decision-making in the clinical setting, especially in the decision-making process for predicting and diagnosing a clinical condition, such as keratoconus. In this study, we performed a CART analysis to predict keratoconus by classifying the risk factors. The other reason we chose CART modeling is because of its usefulness when predicting outcomes in the presence of non-linear interaction of the risk factors. The receiver operator curve (ROC) was plotted between sensitivity and specificity for both models. An appropriate cut-off value was chosen to calculate the Area Under the Curve (AUC), which was then used to calculate the prediction accuracy.

Results

The Model I: Logistic Regression

The feeling of eye dryness increases from 1 to 5, where 1 is no eye dryness, and 5 is a very severe feeling of eye dryness. Eye dryness levels 2, 3, and 4 are included in the model. Also, the

male gender and the positive family history are retained in the logistic model. The confusion matrix and the prediction accuracy for the training and testing data are given below:

Confusion Matrix [Training Data]

		Actual 0	Actual 1
Predicted	0	53	12
	1	7	48

Prediction accuracy = 0.84 [95% CI = 0.76-0.90].

Confusion Matrix [Test Data]

		Actual 0	Actual 1
Predicted	0	13	6
	1	2	9

Prediction accuracy = 0.73 [95% = CI 0.54-0.88].

The Model II: Classification and Regression Trees (CART)

0 = Controls, 1 = Keratoconus. Positive Family History = 0 → No family history of Keratoconus, Gender = 0 → Female.

Eye dryness = 1→No feeling of eye dryness.

Confusion Matrix [Training Data]

		Actual 0	Actual 1
Predicted	0	13	6
	1	2	9

Prediction accuracy = 0.84 [95% CI = 0.76-0.90]

Confusion Matrix [Test Data]

		Actual 0	Actual 1
Predicted	0	53	12
	1	7	48

Prediction accuracy = 0.73 [95% = CI 0.54-0.88]

Discussion

The factors of gender, positive family history, and eye dryness symptoms were retained in the logistic regression model (Figure 1). Males have a 2.5 times higher risk of developing keratoconus than females. A positive family history of keratoconus is associated with 2.6 times higher risk; as the severity of eye dryness increases from level 2 to level 4, the risk of developing keratoconus increases from 3.7 to 4.7 times. The other risk factors like smoking, eye rubbing, and age were excluded from the equation. The training and testing data are split at 80:20, and the prediction accuracy for training is 0.86, which is higher than the test data prediction accuracy, which is 0.63. Since there is a significant difference in the model prediction accuracy between the training and test sets, this logistic regression model is likely overfitting. Overfitting is a modeling error that leads to bias in the model because it is too closely related to the data set. This makes the model only more relevant to its data set and less relevant to other data sets. Accordingly, selecting this logistic regression model would result in poor accuracy when applying a new data set. Hence, the CART classification model was performed (Figure 2).

The CART classification analysis has not been documented in the literature for the risk factor analysis specific to keratoconus. When CART was applied in our data set at the top of the node, also called the root node, the percentage of the case

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -2.794454   1.524307  -1.833 0.066764 .
Age                     -0.032179   0.043965  -0.732 0.464211
Gender1                  2.518575   0.790190   3.187 0.001436 **
Positive_Family_History1 2.606647   0.718581   3.627 0.000286 ***
Eye_Rubbing2             1.611833   1.099147   1.466 0.142529
Eye_Rubbing3            -0.223279   0.882051  -0.253 0.800163
Eye_Rubbing4             0.003535   0.905795   0.004 0.996886
Eye_Rubbing5             0.211812   1.017682   0.208 0.835126
Drynes2                  3.721663   0.964458   3.859 0.000114 ***
Drynes3                  4.268804   0.919226   4.644 3.42e-06 ***
Drynes4                  4.748603   1.649125   2.879 0.003983 **
Drynes5                  1.886347   1.196521   1.577 0.114905
Smoking1                -0.943539   0.579405  -1.628 0.103427
Ocular_Allergy2         -0.512334   0.896611  -0.571 0.567720
Ocular_Allergy3          0.210570   0.978686   0.215 0.829646
Ocular_Allergy4         -1.317805   0.977462  -1.348 0.177597
Ocular_Allergy5         -0.281331   0.850410  -0.331 0.740782
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 1. Logistic regression.

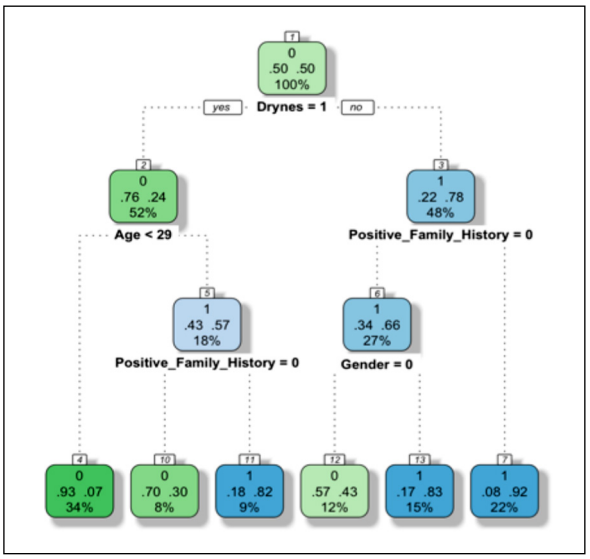

Figure 2. CART classification and regression trees.

and controls were 50-50% (Figure 2). According to this model, node one was divided into nodes 2 and 3 based on the eye dryness symptom, which was the most influential risk factor for keratoconus. As discussed in the second aim, literature supports the idea that eye dryness is one of the risk factors associated with keratoconus in both Middle Eastern as well as in other countries' populations. Nodes 2 and 3 were the group with an age less than 29 years and the group with a positive family history. The age < 29 years node was then tested for positive family history in node 5.

Meanwhile, positive family history was node three and was tested for gender being male in node 6. In this CART classification analysis, the difference between the prediction accuracy of the train data to the test data was not wider compared to the logistic regression. Therefore, it has been ideal to choose CART over the logistic regression model for the prediction of keratoconus. However, additional studies are required in both the Saudi population and other country populations to generate models and machine learning algorithms to improve the fitting of test data to the trained model before generalization can be achieved for the model.

Conclusions

We developed models that would predict the risk of keratoconus by fitting risk factors. Logistic regression and CART analysis were performed. For both models, we found that the eye dryness symptom is positively associated with keratoconus and is considered the major predictor variable. Also, family history was found to be significantly associated in both models. Age was a predictor only in the CART model, and gender was included in both the CART and the logistic regression. In the CART model, the prediction pathway is visually represented in a flow chart and is easy to apply.

The CART model will become complex when more risk factors are included in the model. The difference between the prediction accuracy of training and test data is lower in the CART model compared to the logistic regression model. So, it is ideal to consider the CART over the logistic regression model. Considering the CART model, if a person feels eye dryness and has a positive family history of keratoconus and is a male, he/she will have a 92% chance of developing keratoconus.

Conflict of Interest

The Author declares that he has no conflict of interests.

Acknowledgements

The author would like to thank the scientific research Deanship at Qassim University for supporting this work.

Ethics Approval

The study was approved by the Ethics Committee of Qassim University and Salus University institutional review board. It was conducted according to the ethical principles of the Declaration of Helsinki and by the current legislation on clinical research rules developed by the National Committee of Bioethics in Saudi Arabia.

Informed Consent

The google form questionnaire included a statement before the start of the survey on informed consent. All participants signed it before initiating the study.

Funding

No funding was declared for this article.

References

1) Gordon-Shaag A, Millodot M, Shneor E, Liu Y. The genetic and environmental factors for keratoconus. Biomed Res Int 2015; 2015: 795738.
2) Monies D, Abouelhoda M, AlSayed M, Alhassnan Z, Alotaibi M, Kayyali H, Al-Owain M, Shah A, Rahbeeni Z, Al-Muhaizea MA, Alzaidan HI, Cupler E, Bohlega S, Faqeih E, Faden M, Al-

younes B, Jaroudi D, Goljan E, Elbardisy H, Akilan A, Albar R, Aldhalaan H, Gulab S, Chedrawi A, Al Saud BK, Kurdi W, Makhseed N, Alqasim T, El Khashab HY, Al-Mousa H, Alhashem A, Kanaan I, Algoufi T, Alsaleem K, Basha TA, Al-Murshedi F, Khan S, Al-Kindy A, Alnemer M, Al-Hajjar S, Alyamani S, Aldhekri H, Al-Mehaidib A, Arnaout R, Dabbagh O, Shagrani M, Broering D, Tulbah M, Alqassmi A, Almugbel M, AlQuaiz M, Alsaman A, Al-Thihli K, Sulaiman RA, Al-Dekhail W, Alsaegh A, Bashiri FA, Qari A, Alhomadi S, Alkuraya H, Alsebayel M, Hamad MH, Szonyi L, Abaalkhail F, Al-Mayouf SM, Almojalli H, Alqadi KS, Elsiesy H, Shuaib TM, Seidahmed MZ, Abosoudah I, Akleh H, AlGhonaium A, Alkharfy TM, Al Mutairi F, Eyaid W, Alshanbary A, Sheikh FR, Alsohaibani FI, Alsonbul A, Al Tala S, Balkhy S, Bassiouni R, Alenizi AS, Hussein MH, Hassan S, Khalil M, Tabarki B, Alshahwan S, Oshi A, Sabr Y, Alsaadoun S, Salih MA, Mohamed S, Sultana H, Tamim A, El-Haj M, Alshahrani S, Bubshait DK, Alfadhel M, Faquih T, El-Kalioby M, Subhani S, Shah Z, Moghrabi N, Meyer BF, Alkuraya FS. The landscape of genetic diseases in Saudi Arabia based on the first 1000 diagnostic panels and exomes. Hum Genet 2017; 136: 921-939.

3) Jamali H, Beigi V, Sadeghi-Sarvestani A. Consanguineous Marriage as a Risk Factor for Developing Keratoconus. Med Hypothesis Discov Innov Ophthalmol 2018; 7: 17-21.

4) Georgiou T, Funnell CL, Cassels-Brown A, O'Conor R. Influence of ethnic origin on the incidence of keratoconus and associated atopic disease in Asians and white patients. Eye 2004; 18: 379-383.

5) Millodot M, Shneor E, Albou S, Atlani E, Gordon-Shaag A. Prevalence and associated factors of keratoconus in Jerusalem: a cross-sectional study. Ophthalmic Epidemiol 2011; 18: 91-97.

6) Gordon-Shaag A, Millodot M, Essa M, Garth J, Ghara M, Shneor E. Is consanguinity a risk factor for keratoconus? Optometry and vision science: official publication of the American Academy of Optometry 2013; 90: 448-454.

7) Ihalainen A. Clinical and epidemiological features of keratoconus genetic and external factors in the pathogenesis of the disease. Acta Ophthalmol Suppl 1986; 178: 1-64.

8) Weiss SM, Kapouleas I. An empirical comparison of pattern recognition, neural nets, and machine learning classification methods. International Joint Conference on Artificial Intelligence1989; 89: 781-787.

9) Schmidhuber J. (2015). Deep learning in neural networks: An overview. Neural networks 2015; 61: 85-117.

10) Mohan S, Thirumalai C, Srivastava G. Effective heart disease prediction using hybrid machine learning techniques. IEEE Access 2019; 7: 81542-81554.

11) Sajda P. Machine learning for detection and diagnosis of disease. Annual review of biomedical engineering 2006; 8: 537-565.

12) Fatima M, Pasha M. Survey of machine learning algorithms for disease diagnostic. Journal of Intelligent Learning Systems and Applications 2017; 9: 1.

13) Sen PC, Hajra M, Ghosh M. Supervised classification algorithms in machine learning: A survey and review. In Emerging technology in modeling and graphics. Springer, Singapore 2020: 99-111.