Title
Predictive Models for Diabetic Retinopathy from Non-Image Teleretinal Screening Data.

Authors
Ogunyemi, Omolola I
Gandhi, Meghal
Tayek, Chandler

Publication Date
2019

Peer reviewed

Predictive Models for Diabetic Retinopathy from Non-Image Teleretinal Screening Data

Omolola I. Ogunyemi, PhD[1], Meghal Gandhi, MS[1], Chandler Tayek, BS[1,2]

[1]Center for Biomedical Informatics, Charles R. Drew University of Medicine and Science, Los Angeles, CA; [2]Dept of Computer Science, University of Massachusetts, Amherst, Amherst, MA

Abstract

Introduction: *Timely diabetic retinopathy detection remains a problem in medically underserved settings in the US; diabetic patients in these locales have limited access to eye specialists. Teleretinal screening programs have been introduced to address this problem.*
Methods: *Using data on ethnicity, gender, age, hemoglobin A1C, insulin dependence, time since last eye examination, subjective diabetes control, and years with diabetes from 27,116 diabetic patients participating in a Los Angeles County teleretinal screening program, we compared different machine learning methods for predicting retinopathy. The dataset exhibited a class imbalance.*
Results: *Six classifiers learned on the data were predictive of retinopathy. The best model had an AUC of 0.754, sensitivity of 58% and specificity of 80%.*
Discussion: *Successfully detecting retinopathy from diabetic patients' routinely collected clinical data could help clinicians in medically underserved areas identify unscreened diabetic patients who are at risk of developing retinopathy. This work is a step towards that goal.*

Introduction

Diabetic retinopathy is a sight-threatening complication of diabetes, and the leading cause of blindness in working age adults within the US.[1, 2] It is particularly problematic in medically underserved areas of the US, where annual screening rates for diabetic retinopathy are often much lower than the US national average.[3-5] Teleretinal screening services are increasingly being used in these settings to extend specialty eye care services to patients who would otherwise have limited access to them. A complicating factor in addressing diabetic retinopathy is that many diabetic patients with retinopathy experience no symptoms, even when they are at advanced stages of the condition.[6] Risk factors for diabetic retinopathy include the length of time a person has had diabetes,[6-9] high blood glucose/poor blood sugar control,[6-10] high blood pressure,[7-10] dyslipidemia/[7] high cholesterol,[8] pregnancy,[7] nephropathy,[9] obesity,[7] inflammation,[7] ethnicity,[7] and insulin treatment for Type II diabetes.[9] High blood glucose, duration of diabetes and high blood pressure are considered to be the strongest predictors of retinopathy.[7]

Our ultimate goal is to develop a decision support system that helps primary care physicians who practice in medically underserved settings to identify unscreened diabetic patients who are at high risk of retinopathy, using the diabetic patients' routinely collected clinical data stored within Electronic Health Record Systems. This study represents a step towards achieving that goal. We are particularly interested in clinical data that correspond to known risk factors for retinopathy.

In previous work, we examined data from six federally qualified health centers (FQHCs) in South Los Angeles, obtaining data for machine learning from a class-imbalanced[11-13] subset of 513 diabetic patients. Using standard classifiers on this dataset, the best classification result we achieved was with a Bayesian network that had a sensitivity of 26.2%, a specificity of 94.5% and an Area Under the ROC Curve (AUC) of 0.71.[14] Using a majority class undersampling technique in combination with an ensemble of weak decision tree learners, we achieved an AUC of 0.72, a sensitivity of 69.2% and a specificity of 55.9%.[15] Table 1 below shows the 24 risk factors that were available to us when we developed the machine learning models for this prior work.

Related work by others include a study utilizing public health records from the Korea National Health and Nutrition Examination Surveys (KNHANES) for retinopathy assessment.[16] Data from 327 diabetic patients were used to create predictive models, followed by an internal validation on 163 patients from the KNHANES V-1 dataset. External validation was performed using data from 562 diabetic patients in the KNHANES V-2 dataset. The best results were achieved with support vector machines:[17] an AUC of 0.83 and sensitivity of 71% on the internal validation set, and an AUC of 0.81 and sensitivity of 75.7% on the external validation set. In another study, authors developed predictive models for diabetic retinopathy using data on 266 individuals from the 2005-2008 versions of the US National Health

and Nutrition Examination Surveys (NHANES). The study reported a best AUC of 0.74, a low precision of 22% and a high negative predictive value of 99%, but did not list sensitivities or specificities.[18]

For this study, we performed machine learning using a limited set of risk factors currently available to us from data collected through EyePACS,[19] a web-based teleretinal screening software platform. The data is from diabetic patients utilizing the Los Angeles County Department of Health Services/LACDHS' Teleretinal Diabetic Retinopathy Screening Program and Reading Center.

Table 1. Clinical variables obtained from six FQHCs for machine learning on diabetic retinopathy

Clinical variables collected that might impact diabetic retinopathy risk	
Age	Gender
Ethnicity/race	Marital Status
Education	Household income
Insulin dependence	Insurance
Number of years patient has had diabetes	Body mass index
Hemoglobin A1C value	Primary language
Co-morbid conditions	
Peripheral vascular disease	Cerebrovascular accident/Stroke
Hypertension	Other heart-related diagnosis
Nephropathy	Neuropathy
Depression	Erectile dysfunction
Dyslipidemia	Obesity
Other (hypothyroidism, etc.)	Previous diagnosis & treatment of retinopathy

Methods

We obtained institutional review board approval to use clinical data for the study from the Charles R. Drew University of Medicine and Science under IRB#: 16-10-2491-01.

Data Source and Description:
Clinical data for the study were obtained from a retrospective review of LACDHS EyePACS records. Records obtained included data on eight predictor variables: age, ethnicity/race, gender, hemoglobin A1C, insulin dependence, time since last eye examination, years with diabetes and a patient's subjective assessment of how well-controlled his/her diabetes was, and one outcome variable: diabetic retinopathy (dichotomized for this study as "yes" or "no"). EyePACS records for LACDHS included data from 27,116 patients with Type I or Type II diabetes seen by the County's Teleretinal Diabetic Retinopathy Screening Program and Reading Center between January 1, 2015 and December 31, 2016. Of the 27,116 records obtained, 9,233 records or 34.1% of the sample were cases involving diabetic retinopathy (i.e., the diabetic retinopathy outcome variable value was "yes") and 17,883 records or 65.9% of the sample were cases that did not involve diabetic retinopathy (i.e., the diabetic retinopathy outcome variable value was "no"). This represents a *class imbalance*, since the majority of the data available to us corresponded to cases that did not involve diabetic retinopathy.

Classification Methods:
Missing data were handled using k-nearest neighbor imputation techniques with a k of 9. Numeric variables (age and hemoglobin A1C) were normalized prior to use in machine learning. We learned three standard classifiers on the data: penalized logistic regression, support vector machines (SVMs) with radial basis function kernels, and artificial neural networks (ANNs). Since many machine learning approaches work best when the dataset used for learning has a relatively balanced distribution of class instances (e.g., an equal number of "yes" and "no" instances for diabetic retinopathy), we also learned three classifiers incorporating data pre-processing methods such as minority class oversampling and majority class undersampling, which can help to improve models learned from class-imbalanced data. These included: penalized logistic regression with the Synthetic Minority Oversampling TEchnique/SMOTE,[20] SVM with underbagging, a majority class undersampling technique, and ANNs with SMOTE.

We reserved a random selection of 34% of the dataset, a total of 9,039 cases, as a hold-out test set/internal validation set. We then performed 10-fold cross validation with parameter tuning on the remaining 66% of the dataset (18,077 cases) using the three standard classifiers described above as well as the three classification approaches that incorporated data pre-processing methods. The best models developed from the cross-validation process were assessed on the internal validation set. Analyses were performed in R,[21] using the caret[22] package and the VIM[23] package.

For each classification approach utilized, we measured sensitivity or the true positive rate (the total number of cases classified as having diabetic retinopathy divided by the total number of cases actually involving retinopathy), specificity or the true negative rate (the total number of cases classified as not having diabetic retinopathy divided by the total number of cases that did not involve retinopathy), and the AUC, which represents the trade-off between the true positive rate/sensitivity and the false positive rate or 1 – specificity. Since general model accuracy (the total number of correct predictions divided by the total number of records) is often misleading in the context of class-imbalanced data, we chose to focus on the three metrics listed above (sensitivity, specificity and AUC), since they are better suited to evaluating models learned from class-imbalanced data.

Results

The set of predictor variables/features utilized in developing each of the models used is shown in Table 2. Table 3 shows the results of 10-fold cross-validation using standard classifiers on the training set. Table 4 gives the results of applying the best standard classifiers to the internal validation set. Table 5 shows the results of 10-fold cross-validation using classification methods adapted to handle class-imbalances on the training set. Table 6 gives the results of applying the best classification methods adapted to handle class-imbalances to the internal validation set.

Table 2. Variables obtained from EyePACS that could impact development of diabetic retinopathy

Predictor Variables Obtained from LACDHS EyePACS	
Age	Gender
Ethnicity/race	Hemoglobin A1C
Insulin dependence	Time since last eye examination
Years patient has had diabetes	Patient's subjective assessment of diabetes control

Table 3. Ten-fold cross-validation results using standard classifiers

	Penalized logistic regression	**Support Vector Machines (with RBF kernel)**	**Artificial Neural Networks**
Sensitivity	45.9%	39.8%	45.4%
Specificity	87.2%	89.3%	87.6%
AUC	0.758	0.728	0.762

Table 4. Internal validation set results using standard classifiers

	Penalized logistic regression	**Support Vector Machines (RBF kernel)**	**Artificial Neural Networks**
Sensitivity	44.8%	39.2%	44.8%
Specificity	87.2%	89.4%	87.9%
AUC	0.753	0.729	0.756

Table 5. Ten-fold cross-validation results using classifiers that take into account class-imbalances

	Penalized logistic regression with SMOTE	**Underbagging SVM (RBF kernel)**	**Artificial Neural Networks with SMOTE**
Sensitivity	57.6%	65.7%	57.2%
Specificity	79.4%	72.0%	79.4%
AUC	0.755	0.754	0.751

Table 6. Internal validation set results using classifiers that take into account class-imbalances

	Penalized logistic regression with SMOTE	Underbagging SVM (RBF kernel)	Artificial Neural Networks with SMOTE
Sensitivity	57.5%	65.6%	58.0%
Specificity	79.4%	71.3%	80.0%
AUC	0.752	0.745	0.754

Discussion

A great deal of work has been done in recent years on automated identification of diabetic retinopathy from the digital retinal images of patients who utilize teleretinal screening services.[24-28] However, these advances do not address the needs of the substantial proportion of diabetic patients in medically underserved areas who are not receiving annual eye screenings and who, as a consequence, do not have any digital retinal images that can be assessed. Given the sizable number of records used for the present study, which utilized data from 27,116 diabetic patients, the results presented here are, to our knowledge, the largest thus far focusing on diabetic retinopathy prediction from non-image data.

In general, the results show classifiers that are moderately predictive of retinopathy. Classification methods that were adapted to handle class-imbalances yielded the best results, with the combination of ANNs and SMOTE having a sensitivity of 58%, a specificity of 80% and the highest area under the ROC curve (0.754) on the validation set. As observed in previous studies, for classifiers that incorporate methods to address class-imbalances, improvements in sensitivity came with the trade-off of decreases in specificity. Underbagging SVMs with radial basis function kernels produced the highest sensitivity on the internal validation set (65.6%), which from the standpoint of reaching the largest number of diabetic patients at risk for retinopathy would be very important. Underbagging SVMs maintained a specificity of 71.3% on the internal validation set, which, while worse than the specificity results for ANNs and SMOTE, is still reasonable. They achieved an overall AUC of 0.745. Penalized logistic regression with SMOTE on the internal validation set produced results very similar to ANNs with SMOTE (a sensitivity of 57.5%, specificity of 79.4% and AUC of 0.752).

The key scientific contributions of this work are to demonstrate that:

(1) even the minimal amount of clinical data that accompanies digital retinal images uploaded for teleretinal screening purposes can produce moderately accurate predictions of patients that are at high risk of developing diabetic retinopathy, and,

(2) the use of data pre-processing methods designed to address class imbalances is well suited to the problem of determining diabetic retinopathy risk from data in which the majority of cases do not involve diabetic retinopathy.

The current study shows that at a large scale, predictive models that utilize methods for handling class imbalance can help to detect diabetic patients who have likely developed diabetic retinopathy. Our previous work in this area used data on 24 diabetic retinopathy risk factors collected from 513 diabetic patients. With such a small dataset, there is a danger of the dataset not being adequately representative of the condition we wish to model. Using a larger dataset that contains data from 27,116 diabetic patients gives more confidence in the results observed, as the larger dataset has more than 52 times the number of examples as the previous work, even though a smaller number of features (8 versus 24) was available. Tables 1 and 2 show the differences in the datasets available for machine learning for our prior study and the current work.

Our work also shows that the limited amount of clinical data (independent of digital retinal images) available to readers performing teleretinal screening has some predictive value and could potentially be used to augment deep learning models that utilize fundus images to automatically stage and grade diabetic retinopathy.

Limitations of the current study include the fact that we did not have access to data on known risk factors, such as high blood pressure, that are considered important predictors of retinopathy. However, two key risk factors for predicting diabetic retinopathy, hemoglobin A1C levels and the number of years a patient has had diabetes were available to us in both studies. Future work will involve obtaining access to the complete EHR records of the 27,116 diabetic patients through ORCHID, the LACDHS Cerner implementation, so that we can link existing EHR data on

known retinopathy risk factors to the EyePACs data already obtained for this study. We envision that this will produce models that are even more predictive of retinopathy.

Conclusion

Given the substantial personal, health, and financial costs associated with vision loss from diabetic retinopathy, it is important to develop methods that can assist clinicians in targeting diabetic patients in medically underserved settings who: (1) are not in compliance with American Diabetes Association guidelines on annual eye examinations, and, (2) may be unaware that they have latent retinopathy because they are not yet experiencing any symptoms. We have presented machine learning methods which demonstrate that it is possible to identify high risk patients using clinical data collected in the course of their care. Refinement of these methods to improve their sensitivity and specificity will be an important next step.

Acknowledgments: This work was funded by the National Library of Medicine under grant 1 R01 LM012309. The authors would like to thank Dr. Lauren Patty Daskivich, Director of Ophthalmology and Eye Health programs at LACDHS for providing EyePACS data from the LACDHS Teleretinal Diabetic Retinopathy Screening Program and Reading Center that was used for the current study.

References

1. Klein R, Klein BE. Vision Disorders in Diabetes. Diabetes in America. 2nd ed: National Diabetes Data Group, National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases; 1995.
2. Zhang X, Saaddine JB, Chou CF, Cotch MF, Cheng YJ, Geiss LS, et al. Prevalence of diabetic retinopathy in the United States, 2005-2008. JAMA. 2010 Aug 11;304(6):649-56.
3. Deeb LC, Pettijohn FP, Shirah JK, Freeman G. Interventions among primary-care practitioners to improve care for preventable complications of diabetes. Diabetes Care. 1988 Mar;11(3):275-80.
4. Payne TH, Gabella BA, Michael SL, Young WF, Pickard J, Hofeldt FD, et al. Preventive care in diabetes mellitus. Current practice in urban health-care system. Diabetes Care. 1989 Nov-Dec;12(10):745-7.
5. Wylie-Rosett J, Basch C, Walker EA, Zybert P, Shamoon H, Engel S, et al. Ophthalmic referral rates for patients with diabetes in primary-care clinics located in disadvantaged urban communities. J Diabetes Complications. 1995 Jan-Mar;9(1):49-54.
6. Garg S, Davis RM. Diabetic Retinopathy Screening Update. Clinical Diabetes. 2009;27(4):140 - 5.
7. Ding J, Wong TY. Current epidemiology of diabetic retinopathy and diabetic macular edema. Current diabetes reports. [Review]. 2012 Aug;12(4):346-54.
8. Yau JW, Rogers SL, Kawasaki R, Lamoureux EL, Kowalski JW, Bek T, et al. Global prevalence and major risk factors of diabetic retinopathy. Diabetes Care. [Research Support, N.I.H., Extramural
Research Support, Non-U.S. Gov't
Review]. 2012 Mar;35(3):556-64.
9. Pedro RA, Ramon SA, Marc BB, Juan FB, Isabel MM. Prevalence and relationship between diabetic retinopathy and nephropathy, and its risk factors in the North-East of Spain, a population-based study. Ophthalmic epidemiology. 2010 Aug;17(4):251-65.
10. Klein R, Klein BE, Moss SE. Epidemiology of proliferative diabetic retinopathy. Diabetes Care. [Research Support, U.S. Gov't, P.H.S.
Review]. 1992 Dec;15(12):1875-91.
11. Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews. 2012;42(4):463 - 84.
12. He H, Ma Y. Imbalanced Learning: Foundations, Algorithms, and Applications. Hoboken, New Jersey: Wiley-IEEE Press; 2013.
13. Feng W, Huang W, Ren J. Class Imbalance Ensemble Learning Based on the
Margin Theory. Applied Sciences. 2018;8(5):1-28.
14. Ogunyemi O, Teklehaimanot S, Patty L, Moran E, George S. Evaluating predictive modeling's potential to improve teleretinal screening participation in urban safety net clinics. Studies in health technology and informatics. [Research Support, N.I.H., Extramural]. 2013;192:162-5.
15. Ogunyemi O, Kermah D. Machine Learning Approaches for Detecting Diabetic Retinopathy from Clinical and Public Health Records. AMIA Annu Symp Proc. 2015:983-90.
16. Oh E, Yoo TK, Park EC. Diabetic retinopathy risk prediction for fundus examination using sparse learning: a cross-sectional study. BMC Med Inform Decis Mak. 2013 Sep 13;13:106.

17. Cortes C, Vapnik V. Support-Vector Networks. Machine Learning. 1995 September 01;20(3):273-97.
18. Cichosz SL, Johansen MD, Knudsen ST, Hansen TK, Hejlesen O. A classification model for predicting eye disease in newly diagnosed people with type 2 diabetes. Diabetes research and clinical practice. 2015 May;108(2):210-5.
19. Cuadros J, Bresnick G. EyePACS: an adaptable telemedicine system for diabetic retinopathy screening. J Diabetes Sci Technol. 2009;3(3):509-16.
20. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. 2002;16:321-57.
21. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2018.
22. Kuhn M, Contributions from Jed Wing SW, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt. caret: Classification and Regression Training. R package version 60-80; 2018.
23. Kowarik A, Templ M. Imputation with the R Package VIM. Journal of Statistical Software. 2016;74(7):1-16.
24. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA. 2016 Dec 13;316(22):2402-10.
25. Ramachandran N, Hong SC, Sime MJ, Wilson GA. Diabetic retinopathy screening using deep neural network. Clin Exp Ophthalmol. 2017 Sep 7.
26. Gargeya R, Leng T. Automated Identification of Diabetic Retinopathy Using Deep Learning. Ophthalmology. 2017 Jul;124(7):962-9.
27. Takahashi H, Tampo H, Arai Y, Inoue Y, Kawashima H. Applying artificial intelligence to disease staging: Deep learning for improved staging of diabetic retinopathy. PLoS One. 2017;12(6):e0179790.
28. Ting DSW, Cheung CY, Lim G, Tan GSW, Quang ND, Gan A, 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 Dec 12;318(22):2211-23.