Predicting Unplanned Medical Visits Among Patients with Diabetes: Translation From Machine Learning to Clinical Implementation

Arielle Selya (arielle.selya@sanfordhealth.org)
University of North Dakota School of Medicine & Health Sciences https://orcid.org/0000-0001-7026-6988

Drake Anshutz
Advanced Analytics, St. Luke's Health System

Emily Griese
Sanford Health Plan

Tess L Weber
Sanford Research https://orcid.org/0000-0002-1257-3469

Benson Hsu
Sanford Health

Cheryl Ward
EDCO Health Information Systems

Research article

Keywords: Diabetes, Unplanned medical visits, Machine learning, Predictive model

Posted Date: September 21st, 2020

DOI: https://doi.org/10.21203/rs.3.rs-72164/v1

Version of Record: A version of this preprint was published at BMC Medical Informatics and Decision Making on March 31st, 2021. See the published version at https://doi.org/10.1186/s12911-021-01474-1.

Abstract

Background: Diabetes is common and an economic burden in the United States. In this study, a machine learning predictive model was developed to predict unplanned medical visits among patients with diabetes.

Methods: Data were drawn from electronic medical records (EMRs) from a large healthcare organization in the Northern Plans region of the US, from adult (≥18 years old) patients with type 1 or type 2 diabetes who received care at least once during the 3 year period. A variety of machine-learning classification models were run using standard EMR variables as predictors (age, body mass index (BMI), Systolic blood pressure (BP), Diastolic BP, low-density lipoprotein (LDL), high-density lipoprotein (HDL), glycohemoglobin (A1C), smoking status, number of diagnoses and number of prescriptions). The best-performing model after cross-validation testing was analyzed to identify strongest predictors.

Results: The best-performing model was a radial-basis support vector machine, which achieved a prediction accuracy (average of sensitivity and specificity) of 66.2%. This outperformed a conventional logistic regression by 1.5 percentage points. High BP and low HDL were identified as the strongest predictors, such that eliminating these from the model decreased its overall prediction accuracy by 1.9 and 1.8 percentage points, respectively.

Conclusion: Our machine-learning predictive model more accurately predicted unplanned medical visits among patients with diabetes, relative to conventional models. Post-hoc analysis of the model was used for hypothesis generation, namely that HDL and BP are the strongest contributors to unplanned medical visits among patients with diabetes. In this way, this predictive model can be used in moving from prediction to implementation and improved diabetes care management in clinical settings.

Introduction

There are approximately 1.5 million new diabetes diagnoses among people 18 years and over every year, and in 2018, approximately 34.2 million persons (10.5%) in the US had diabetes.(1) In 2017, 83,564 deaths were attributed to diabetes in the United States, and diabetes is the 7th leading cause of death in the United States (25.7 deaths per 100,000 population).(1)

Diabetes imposes significant healthcare utilization and costs.(2) Americans with diabetes in 2017 spent approximately $16,700 annually in health care costs, 2.3 times higher than those without diabetes.(3) Total costs of diabetes in 2017 were $327 billion annually, of which $237 billion were in direct medical costs.(3) In addition, there is a positive relationship between lack of health insurance and prevalence of diagnosed diabetes, exacerbating the risks for uninsured Americans.(4) By 2034, the population with diabetes is expected to increase by 100% and the cost is expected to increase by 53%.(5)

Patients with diabetes generally have increased healthcare utilization, including clinic visits, outpatient departments, and emergency departments, compared to those without diabetes.(3, 6) The 2011 National

Health Interview Survey Diabetes revealed that 30% of diabetic patients had at least one emergency department visit within the last year, compared to only 20% of the general population.(6) The majority of emergency department visits among patients with diabetes are likely related to acute glycemic complications (hyperglycemia and hypoglycemia);(6) however, most adults with diabetes have at least one comorbid chronic condition(7) which could contribute to these visits as well.

Additionally, social and behavioral factors are associated with unplanned medical visits among the population of patients with diabetes. Lower socioeconomic status, longer disease duration, disease severity, and co-morbid depression are all significant determinants of unplanned medical visits and hospitalizations.(8) More precisely, patients with diabetes with very high current depressive symptoms were two times more likely to have an unplanned emergency department visit, and patients who were diagnosed more than 10 years ago were 1.3 times more likely of an unplanned emergency visit.(8) Additionally, cigarette smoking is associated with a greater likelihood of unplanned medical visits.(9) However, unplanned visits remain a high-impact problem for patients and healthcare systems alike, highlighting the need for improved prediction models that can be implemented clinically.

Because of the increased risks and associated costs for patients with diabetes, there is a significant need to improve prediction capabilities aimed at reducing unplanned medical visits for this group of patients. A majority of medical risk prediction models have been developed using stepwise logistic regression, while machine learning classification methods have been largely unexplored.(10) Machine learning methods offer the additional possibility to improve prediction based on pattern detection of many variables simultaneously, as has been shown in applications on predicting(11) and compliance with dietary recommendations,(12) predicting metabolic syndrome from physical characteristics and lab results,(13) identifying binge drinkers from parenting variables(14) and drinking motives,(15) and predicting high blood pressure using body measures.(16) The current study utilizes electronic medical record (EMR) data from a large healthcare system, and develops a machine learning based predictive model to predict any vs. no unplanned medical visits over a 3-year period among adult patients with diabetes.

Methods

Sample

Data were obtained from electronic medical records (EMRs) in Epic from Sanford Health, a not-for-profit rural healthcare system that primarily serves South Dakota, North Dakota, Northern & Southwest Minnesota, Northwest Iowa, and parts of Nebraska. Sanford Health includes roughly 44 hospitals 1,382 physicians and 9,703 nurses delivering care in more than 80 specialty areas. All data were de-identified according to the Health Insurance Portability and Accountability Act HIPAA de-identification method Safe Harbor § 164.514(b)(2). The dataset included records from all patients who visited a Sanford healthcare facility between January 1, 2014 and December 30, 2016 (N=1,143,028). Only adult patients (age≥18; N=875,168) with a diagnosis of diabetes (ICD-10 codes E10.xx and E11.xx; N=67,575) were included in the current study. Further, only patients who reported a residential zip code in Minnesota (MN), North

Dakota (ND), or South Dakota (SD) were included in the current study (*N*=63,781), due to low sample sizes in other states. Finally, patients who had missing data on the outcome variable of unplanned medical visits or any of the predictor variables were excluded, for a final sample size of *N*=43,831.

Measures

The outcome was any vs. no unplanned medical visits during the 3-year period over which EMR data were collected. This was derived from four separate variables: emergency department visits, hospitalizations, hospital observations, and urgent care visits. All four types of visits were summed and dichotomized at ≥1 vs. 0 unplanned medical visits.

Predictor variables included all numeric variables that were common and readily available in Sanford's EMR's. Ten variables were selected and are described in detail below.

Age was measured in years at time of initial analyses (12/1/2016).

Body mass index (BMI) was obtained from EMR's as kg/m^2. Extreme values (<15 or >60) were assumed to be errors and were set to missing. Values from the most recent visit in the 3-year period were used.

Blood pressure (BP) was obtained in mm/Hg. Values from the most recent visit in the 3-year period were used. Systolic BP and diastolic BP were included as two separate variables.

Serum cholesterol was obtained as both low-density lipoprotein (LDL) and high-density lipoprotein (HDL) in mg/dL. Extreme values in HDL (<10 or >100) or LDL (<20 or >200) were assumed to be errors and were set to missing. Values from the most recent laboratory result were used. LDL and HDL were analyzed as two separate variables.

Glycohemoglobin (A1C) was measured from the most recent laboratory result. A1C values below 4 or above 15 were assumed to be errors and were set to missing.

Ranked smoking status was obtained by patient self-report as a vital sign on their most recent visit. A ranked variable was created as follows from the several possible response categories, with higher values indicating more smoke exposure: never smoker (0), passive smoker (1), former smoker (2), current some day smoker (3), current every day smoker, light tobacco smoker, or heavy tobacco smoker (4).

Number of diagnoses on "problem list" was derived from the most recently available list over the 3-year period.

Number of prescriptions were aggregated over the 3-year period and was used as a numeric variable.

Analyses

Machine learning.

All analyses predicted the unplanned medical visit status of each patient (i.e. which patients had at least one vs. no unplanned medical visits in the 3-year period), and this classification task was based on the 10 EMR variables above (age, BMI, BP, HDL and LDL cholesterol, A1C, ranked smoking status, number of diagnoses on the patient's "problem list," and the number of prescriptions in the 3-year period). Three types of machine learning were utilized: discriminant analysis (linear and quadratic), support vector machines (SVM; linear basis and radial basis), and artificial neural nets (NN's; single and double hidden layer). R software (17) was used for all analyses, including the packages MASS for discriminant analysis, (18) e1071 for SVM's,(19) nnet for single-layer NN's,(18) and deepnet four double-layer NN's.(20) A logistic regression was run for purposes of comparing machine learning results with conventional prediction approaches.

Cross-validation testing.

Since classifiers are susceptible to overtraining (i.e. when the classifier can predict the training dataset with high accuracy, but fits noise and thus has not learned patterns that generalize to other datasets), cross-validation testing is important to identify models that have identified patterns that are truly important in the prediction task. Cross-validation testing is performed by partitioning all available data points into a training set and a testing set; the classifier is trained on the data from the training set, and the generalization of the prediction task learned by the classifier is tested using the data from the testing set. In particular for this study, repeated subsampling cross-validation was used by withholding a randomly selected 10% of trials as the testing set, and iterating (with different random selections of the testing set) 1000 times.

Both training accuracy (prediction accuracy on the testing dataset) and generalization accuracy (prediction accuracy on the training set) were assessed using confusion matrices. SVM's and NN's were optimized by running several iterations over different parameter values: for SVM, the cost parameter was varied from 0.1-10, and for radial SVM the gamma parameter was varied from 0.001-0.5; for single-layer NN's, the size of the hidden layer was varied from 1-20, the number of training iterations was varied from 100-200, and the decay parameter was varied from 0-0.9; and for double-layer NN's, the size of each hidden layer was varied from 0-20, the learning rate was varied from 0-1, the momentum of the learning rate was varied from 0-1, and the number of training iterations is varied from 10-20. For each classifier, the model with the highest generalization accuracy is reported. Since there are two possible categories, chance performance is 50%. Accuracy vs. chance was measured using a binomial test of the success rate out of the 1000 cross-validation iterations.

Sensitivity testing.

In order to derive clinical implications from the predictive model, it is valuable to know which variables are most strongly predictive of unplanned medical visits. Although being important for prediction does not necessarily indicate causality, many of the modifiable predictors (A1C, BMI, BP, cholesterol, smoking) do have plausible causal effects on diabetes and its complications. If risky values of these modifiable predictors are important for prediction (through a causal mechanism or an association), then removing

risky values should disrupt prediction accuracy. Thus, in order to determine the modifiable variables that are most strongly indicative of unplanned medical visits, a variant of sensitivity testing was performed: for one variable at a time, the dataset was restricted to observations within the normal or healthy range, and the disruption in the model's generalization accuracy was assessed when predicting on this restricted dataset. Restrictions to the normal/healthy range were based on current guidelines, namely BMI < 30,(21) BP < 120/80,(22) ranked smoking status < 2 (indicating never smoker or former smoker), LDL < 130, HDL > 50,(23) and A1C < 6.5.(24) Larger disruptions to the generalization accuracy as a result of restricting a variable to a healthy range indicates a greater importance of that variable to the prediction task, and potentially as a clinical target for intervention.

Results

Table 1 shows the characteristics of the sample, summarized by patients who did vs. did not have unplanned medical visits during the 3-year period. Patients with at least one unplanned visit tended to be slightly older (66 vs. 65 years old), rank higher on the smoking scale (2 vs. 1), have more diagnoses on the problem list (4 vs. 3), and have lower HDL values (42 vs. 44), and have been prescribed considerably more medications over the 3-year period (205 vs. 88) (all $p < .05$). The two groups had similar mean levels of diastolic blood pressure, but those with at least one unplanned visit had a wider interquartile range (IQR: 64–80 vs. 66–80). Similarly, the two groups had similar mean levels of A1C, but those with at least one unplanned visit had a wider IQR (6.3–7.8 vs. 6.3–7.9). No significant difference was observed for BMI, systolic blood pressure or LDL cholesterol (p-value > 0.05).

Table 1
Characteristics of patients with diabetes by unplanned visit status.

Predictor Variable	No Unplanned Visits (*N* = 18,771)	≥ 1 Unplanned Visits (*N* = 25,060)	*p*-value
Age	**65 (55–74)**	**66 (55–76)**	**< .0001**
BMI	32.3 (28.3–37.0)	32.2 (28.0–37.3)	= .2454
Systolic BP	126.0 (118.0–134.0)	126.0 (116.0–136)	= .0089
Diastolic BP	**72.0 (66.0–80.0)**	**72.0 (64.0–80.0)**	**< .0001**
LDL cholesterol	85.0 (67.0–106.0)	84.0 (65.0–106.0)	= .0053
HDL cholesterol	**44.0 (37.0–53.0)**	**42.0 (35.0–52.0)**	**< .0001**
A1C	**6.9 (6.3–7.8)**	**6.9 (6.3–7.9)**	**= .0001**
Ranked smoking status	**1.0 (0.0–2.0)**	**2.0 (0.0–2.0)**	**< .0001**
Number of diagnoses on problem list	**3.0 (2.0–4.0)**	**4.0 (3.0–6.0)**	**< .0001**
Number of prescriptions	**88.0 (40.0–179.0)**	**205.0 (96.0–408.0)**	**< .0001**

Note. Variables are summarized as median (interquartile range). A1C: glycohemoglobin. BMI: body mass index. BP: blood pressure. HDL: high-density lipoprotein. LDL: low-density lipoprotein. *p*-values are based on t-tests. Bold: $p < .05$.

Table 2 shows the best-case instance of each classifier in terms of generalization accuracy. Logistic regression (bottom row) is intended as a comparison, as it only models main effects of each predictor and does not contain any interaction terms. Logistic regression performed reasonably well, with a sensitivity (true positive rate) of 70.2% and a specificity (true negative rate) of 60.4%. Linear discriminant analysis found the highest specificity of all models (75.4%), but the sensitivity (51.0%) is barely above chance performance; thus, this considered a low performing model, especially relative to logistic regression, due to its inability to distinguish between classes. For similar reasons, quadratic discriminant analysis also resulted in a low performing model, although in this case the sensitivity was high (82.5%) and the specificity was below chance (44.1%). Linear SVM and radial SVM were the only two models that outperformed the logistic regression. Radial SVM found the most accurate overall prediction (average of sensitivity and specificity = 66.9%) with the sensitivity (67.8%) and specificity (65.9%) both being significantly above chance. Single and double hidden layer neural networks were found to underperform in comparison to logistic regression prediction accuracy.

Table 2
Generalization accuracy of best-case classifiers, presented as confusion matrices and average prediction accuracy.

Classifier	Parameters		Predicted: No Unplanned Visits	Predicted: ≥1 Unplanned Visit
Linear discriminant analysis	N/A	Actual: No Unplanned Visits	75.4%	24.6%
		Actual: ≥1 Unplanned Visit	49.0%	51.0%
		Average	63.2%	
Quadratic discriminant analysis	N/A	Actual: No Unplanned Visits	44.1%	55.9%
		Actual: ≥1 Unplanned Visit	17.5%	82.5%
		Average	63.3%	
Linear SVM	Cost = 1	Actual: No Unplanned Visits	60.3%	39.7%
		Actual: ≥1 Unplanned Visit	29.1%	70.9%
		Average	65.6%	
Radial SVM	Cost = 1; Gamma = 0.007	Actual: No Unplanned Visits	65.9%	34.1%
		Actual: ≥1 Unplanned Visit	32.2%	67.8%
		Average	66.9%	
Single hidden layer NN	Hidden layer = 10 nodes; Iterations = 200; Decay = 0.2	Actual: No Unplanned Visits	65.7%	34.3%
		Actual: ≥1 Unplanned Visit	38.9%	61.1%
		Average	63.4%	
Double hidden layer NN	Hidden layers = 10 nodes; Learning = 0.8; Momentum = 0.5;	Actual: No Unplanned Visits	62.4%	37.6%

Note. Cross-validation matrices show the generalization accuracy with respect to the actual class (rows) against the predicted class (columns). NN: neural nets. SVM: support vector machines.

Classifier	Parameters		Predicted: No Unplanned Visits	Predicted: ≥1 Unplanned Visit
	Iterations = 10			
		Actual: ≥1 Unplanned Visit	38.0%	62.0%
		Average	62.2%	
Logistic Regression	N/A	Actual: No Unplanned Visits	60.4%	39.6%
		Actual: ≥1 Unplanned Visit	29.8%	70.2%
		Average	65.3%	

Note. Cross-validation matrices show the generalization accuracy with respect to the actual class (rows) against the predicted class (columns). NN: neural nets. SVM: support vector machines.

Table 3 shows the sensitivity analysis of the best-case radial SVM classifier presented in Table 2. Both blood pressure and HDL cholesterol were found to contribute most significantly to the prediction task: removing blood pressure from the model decreases the model's accuracy by 1.9 percentage points, and removing HDL cholesterol results in a decrease of 1.8 percentage points. The lowest change in accuracy comes from A1C which when removed from the model only resulted in a drop of 0.7 percentage points. These analyses show that BP and HDL seem to be the most important indicators of unplanned medical visits among patients with diabetes, among the potentially modifiable variables.

Table 3
Sensitivity analysis showing the disruption of generalization accuracy when restricting each variable to a healthy range.

Restricted variable range	New prediction accuracy	Change in accuracy
A1C < 6.5 (*N* = 13,857)	66.2%	-0.7%
BMI > 30 (*N* = 15,885)	65.9%	-1.0%
BP < 120/80 (*N* = 11,996)	65.0%	-1.9%
HDL > 50 (*N* = 30,058)	65.1%	-1.8%
LDL < 130 (*N* = 39,384)	65.8%	-1.1%
No current smoking (*N* = 38,370)	65.8%	-1.1%

Note. Prediction accuracy is the average of the hit and correct rejection accuracies. Change in prediction accuracy is relative to the best-case results using the full data sample in Table 2. A1C: glycohemoglobin. BMI: body mass index. BP: blood pressure. HDL: high-density lipoprotein. LDL: low-density lipoprotein.

Discussion

This study utilized machine learning to predict unplanned medical visits among patients with diabetes over a 3-year period, using readily available variables from EMRs as prediction variables. Machine learning methods (radial-basis SVM in particular) were able to achieve more accurate prediction relative to conventional logistic regression, with average accuracy ranging from 65.1–66.2%. Further, post-hoc analysis of the trained prediction model revealed that HDL and BP are possibly the most important modifiable variables that predict unplanned medical visits among patients with diabetes.

HDL and BP may be driving unplanned medical visits among patients with diabetes due to their individual risks for unplanned medical. HDL is generally known as being the “good cholesterol” because of its atherogenesis inhibitory properties. In addition, HDL is normally anti-inflammatory; however, HDL often has a loss of function in patients with diabetes, and thus the anti-inflammatory properties are inhibited.(25, 26) The disease mechanisms in both diabetes and hypertension are similar, and have commonalities in etiology including obesity, inflammation, oxidative stress, and insulin resistance.(27) Similarly, high BP in diabetes patients is associated with increased risk of death and diabetes-related complications, which explains the finding that high BP is especially predictive of unplanned medical visits.(28)

Presently, literature shows more evidence of hospitals employing predictive analytics to help with predicting readmission risk and less regarding preventing and predicting unplanned medical visits. Commonly used are the HOSPITAL and LACE screening tools to help predict readmission risk. The HOSPITAL score uses 7 clinical predictors to help identify patients at high risk of hospital readmissions within 30 days of discharge. This score has been validated and shown to have superior discriminative ability over other prediction tools(29) Similarly, the LACE index uses only four variables to predict death or 30-day readmission after hospital discharge of 66.3% and a correct rejection accuracy of 53.3%.(30) While this tool has also been validated, LACE has been shown to only have moderate discriminative ability.(29)

Therefore, the higher prediction accuracy in the current model demonstrates the utility of machine learning approaches for prediction of medical risks. Though the improvement in accuracy may be considered small (~ 1.5 percentage points), this could have substantial implications at a large scale. For example, back-of-the envelope calculations show that, under the assumption that these visits can be anticipated and prevented with perfect accuracy, an improvement of 1.5% for a patient population of 1 million translates into approximately 3000 people and 10,000 visits that could be avoided.

The higher accuracy is likely attributable to the increased predictive information contained in *patterns* of variables, over and above each variable’s statistically independent association with the outcome.(11, 31) Though this pattern-based information is difficult to extract in “black-box” models (e.g. SVM), we present a form of sensitivity analysis that estimates each variable’s total contribution to the model (accumulated across its statistically independent main effect and all interactions with other predictor variables) and thus can quantify each variable’s “diagnostic information.”

Identifying the most salient predictors is an important step towards moving this predictive algorithm towards concrete implementation in clinical settings. That is, the trained predictive model can be used for hypothesis generation (i.e. that risky HDL and BP values lead to unplanned medical visits). Since the predictive model itself cannot test or establish causality, further longitudinal research in clinical settings is needed to test these hypotheses; nevertheless, this hypothesis generation is an essential step in that it reduces the number of likely hypotheses that must be tested in clinical settings, leading to a more efficient use of resources. Following the hypothesis validation stage, an evidence-based intervention then can be designed and implemented which flags high-risk patients for an appropriate protocol (e.g. more aggressive targeting of BP and HDL through clinical or behavioral measures).

In fact, such an intervention is currently under development at the sponsoring healthcare organization, brought about in part by this research. Thus, this study demonstrates the value of machine learning models not only in improved prediction of costly unplanned medical visits, but also in moving towards clinical implementation.

Limitations

This study has several limitations. First, causality cannot be established using observational data; however, the current procedure of performing a sensitivity analysis on modifiable predictor variables produces a more refined set of causal hypotheses that can be pursued in follow-up research. A related limitation is that, factors that may be relevant for prediction may not be pertinent for treatment (e.g. age which is not modifiable). Limitations of EMR data, such as imprecise measures of smoking status, are likely to negatively impact the prediction accuracy. However, basing our prediction model on standard EMR fields increases its utility within this healthcare system, as well as its potential generalizability to other healthcare systems. Finally, results may not be generalizable to other populations outside the North Dakota, South Dakota, and Minnesota, and further validation is needed in other independent samples.

Strengths:

The use of EMR data from a large healthcare system in the US allows for the capture of large proportion of the population, and a large sample size. This study also utilizes innovative machine learning methods with cross-validation, which leads to improved prediction accuracy and generalizability of results. Finally, the current study demonstrates a relatively novel procedure for moving a machine-learning model from pure prediction towards making clinical improvements to care management.

Conclusions

This study shows improved prediction of unplanned medical visits among patients with diabetes by utilizing machine learning methods, relative to conventional prediction models. A post-hoc sensitivity analysis identified low HDL and high BP as the strongest predictors of unplanned medical visits among this patient population, prompting future research in clinical settings on whether these are causal

relationships Future research is underway based on this predictive model on a behavioral health intervention aimed at improving diabetes management in clinical settings.

Abbreviations

EMR: Electronic medical records; BMI: body mass index; BP: Blood pressure; LDL: low-density lipoprotein; HDL: high-density lipoprotein; A1C: glycohemoglobin; MN: Minnesota; ND: North Dakota; SD: South Dakota; SVM: support vector machine; NN: neural network.

Declarations

Ethics approval: This study was approved by Sanford Research's Internal Review Board.

Consent for Publication: Not applicable

Availability of data: The datasets used during the current study are available from the corresponding author on reasonable request.

Competing Interests: The authors declare that they have no competing interests.

Funding: This work was supported by an award from the Sanford Data Collaborative, and by the National Institute for General Medical Sciences (NIGMS), grant number 1P20GM121341–01. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' contributions: All authors contributed to the production of this study and the manuscript. AS and EG designed the study. AS and DA performed data analyses. TW and CW drafted the manuscript. BS revised the document for intellectual content. All authors have read and approved the final manuscript.

Author's information: Not applicable

References

1. Heron M. Deaths: Leading Causes for 2017. :77.
2. Rui P, Kang K, Ashman J. National Hospital Ambulatory Medical Care Survey: 2016 Emergency Department Summary Tables. 2016 p. 38.
3. American Diabetes Association. Economic Costs of Diabetes in the U.S. in 2017. Diabetes Care. 2018 May;41(5).
4. Raghupathi W, Raghupathi V. An Empirical Study of Chronic Diseases in the United States: A Visual Analytics Approach to Public Health. Int J Environ Res Public Health [Internet]. 2018 Mar [cited 2019 May 6];15(3). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5876976/
5. Bodenheimer T, Chen E, Bennett HD. Confronting the growing burden of chronic disease: can the U.S. health care workforce do the job? Health Aff Proj Hope. 2009 Feb;28(1):64–74.

6. McEwen L N, Herman WH. Health care utilization and costs of diabetes. In: Diabetes in America [Internet]. 3rd ed. Bethesda, MD: National Institutes of Health; 2018. p. 40-1-40–78. Available from: file:///C:/Users/703398871/Downloads/DIA_Ch40%20(6).pdf
7. Druss BG, Marcus SC, Olfson M, Tanielian T, Elinson L, Pincus HA. Comparing the national economic burden of five chronic conditions. Health Aff Proj Hope. 2001 Dec;20(6):233–41.
8. Begum N, Donald M, Ozolins IZ, Dower J. Hospital admissions, emergency department utilisation and patient activation for self-management among people with diabetes. Diabetes Res Clin Pract. 2011 Aug 1;93(2):260–7.
9. Selya A, Johnson EL, Weber TL, Russo J, Stansbury C, Anshutz D, et al. Smoking is associated with a higher risk of unplanned medical visits among adult patients with diabetes, using retrospective electronic medical record data from 2014 to 2016. BMC Health Serv Res [Internet]. 2020 [cited 2020 May 7];20(1). Available from: https://link.springer.com/epdf/10.1186/s12913-020-05277-4
10. Deo RC. Machine Learning in Medicine. Circulation. 2015 Nov 17;132(20):1920–30.
11. Selya AS, Anshutz D. Machine learning for predicting health outcomes: An example of predicting obesity from dietary and physical activity patterns. In: Advanced Data Analytics in Healthcare. Switzerland: Springer Nature; 2018. p. 77–97.
12. Giabbanelli PJ, Adams J. Identifying small groups of foods that can predict achievement of key dietary recommendations: data mining of the UK National Diet and Nutrition Survey, 2008–12. Public Health Nutr. 2016 Jun;19(9):1543–51.
13. Karimi-Alavijeh F, Jalili S, Sadeghi M. Predicting metabolic syndrome using decision tree and support vector machine methods. ARYA Atheroscler. 2016 May;12(3):146–52.
14. Crutzen R, Giabbanelli PJ, Jander A, Mercken L, de Vries H. Identifying binge drinkers based on parenting dimensions and alcohol-specific parenting practices: building classifiers on adolescent-parent paired data. BMC Public Health. 2015 Aug 5;15(1):747.
15. Crutzen R, Giabbanelli P. Using Classifiers to Identify Binge Drinkers Based on Drinking Motives. Subst Use Misuse. 2014 Jan 1;49(1–2):110–5.
16. Golino HF, Amaral LS de B, Duarte SFP, Gomes CMA, Soares T de J, Reis LA dos, et al. Predicting Increased Blood Pressure Using Machine Learning [Internet]. Vol. 2014, Journal of Obesity. Hindawi; 2014 [cited 2020 Sep 3]. p. e637635. Available from: https://www.hindawi.com/journals/jobe/2014/637635/
17. R Core Team. R: A language and environment for statistical computing. [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2018. Available from: https://www.R-project.org
18. Venables WN, Ripley BD. Modern Applied Statistics with S [Internet]. 4th ed. New York: Springer-Verlag; 2002 [cited 2020 Aug 25]. (Statistics and Computing, Statistics,Computing Venables,W.N.:Statistics w.S-PLUS). Available from: https://www.springer.com/gp/book/9780387954578
19. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F, C++-code) C-CC (libsvm, et al. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien

[Internet]. 2019 [cited 2020 Aug 25]. Available from: https://CRAN.R-project.org/package=e1071

20. Rong X. deepnet: deep learning toolkit in R [Internet]. 2014 [cited 2020 Aug 25]. Available from: https://CRAN.R-project.org/package=deepnet
21. Adults (US) NOEIEP on the I Evaluation, and Treatment of Obesity in. Clinical Guidelines on the Identification, Evaluation, and Treatment of Overweight and Obesity in Adults. National Heart, Lung, and Blood Institute; 1998.
22. Carey RM, Whelton PK, 2017 ACC/AHA Hypertension Guideline Writing Committee. Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: Synopsis of the 2017 American College of Cardiology/American Heart Association Hypertension Guideline. Ann Intern Med. 2018 06;168(5):351–8.
23. Grundy SM, Stone NJ, Bailey AL, Beam C, Birtcher KK, Blumenthal RS, et al. 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA Guideline on the Management of Blood Cholesterol: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. J Am Coll Cardiol. 2019 Jun 25;73(24):e285–350.
24. ACP Guidance Statement on HbA | Annals of Internal Medicine | American College of Physicians [Internet]. [cited 2020 May 7]. Available from: https://annals.org/aim/fullarticle/2674121/hemoglobin-1c-targets-glycemic-control-pharmacologic-therapy-nonpregnant-adults-type
25. Femlak M, Gluba-Brzózka A, Ciałkowska-Rysz A, Rysz J. The role and function of HDL in patients with diabetes mellitus and the related cardiovascular risk. Lipids Health Dis [Internet]. 2017 Oct 30 [cited 2020 Apr 7];16. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5663054/
26. Farbstein D, Levy AP. HDL dysfunction in diabetes: causes and possible treatments. Expert Rev Cardiovasc Ther. 2012 Mar;10(3):353–61.
27. Cheung BMY, Li C. Diabetes and Hypertension: Is There a Common Metabolic Pathway? Curr Atheroscler Rep. 2012 Apr;14(2):160–6.
28. Group BMJP. Tight blood pressure control and risk of macrovascular and microvascular complications in type 2 diabetes: UKPDS 38. BMJ. 1998 Sep 12;317(7160):703–13.
29. Robinson R, Hudali T. The HOSPITAL score and LACE index as predictors of 30 day readmission in a retrospective study at a university-affiliated community hospital. PeerJ [Internet]. 2017 Mar 29 [cited 2019 Aug 28];5. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5374974/
30. Low LL, Lee KH, Hock Ong ME, Wang S, Tan SY, Thumboo J, et al. Predicting 30-Day Readmissions: Performance of the LACE Index Compared with a Regression Model among General Medicine Patients in Singapore. BioMed Res Int. 2015;2015:169870.
31. Hanson SJ, Schmidt A. High-resolution imaging of the fusiform face area (FFA) using multivariate non-linear classifiers shows diagnosticity for non-face categories. NeuroImage. 2011 Jan 15;54(2):1715–34.