Title

Usage of Electronic Health Record Phenotyping in American Adult Patients with Schizophrenia to Improve Detection of Type II Diabetes Mellitus

Author

Haynesworth, Austin

Publication Date

2020

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Usage of Electronic Health Record Phenotyping in American Adult Patients with Schizophrenia to Improve Detection of Type II Diabetes Mellitus

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Applied Statistics

by

Austin Haynesworth

2020

ABSTRACT OF THE THESIS

Usage of Electronic Health Record Phenotyping
in American Adult Patients with Schizophrenia
to Improve Detection of Type II Diabetes Mellitus

by

Austin Haynesworth
Master of Applied Statistics
University of California, Los Angeles, 2020
Professor Frederic Paik Schoenberg, Chair

Objective: Prevalence of type II diabetes is higher in diagnosed patients with schizophrenia when compared with the general adult population. With the propagation of electronic health records (EHR), we assess whether electronic health record phenotyping can improve diabetes mellitus type 2 (T2DM) screening when compared to conventional diabetic screening metrics in patients with schizophrenia.

Method: EHR data from 1267 patients with schizophrenia was used to develop a pre-screening tool for current T2DM using logistic regression and random forest models. Three types of models were created: the first used conventional diabetic screening criteria (BMI, age, gender, etc.) and their interactions. The second utilized conventional factors plus prescribed medications. The third utilized conventional factors, prescriptions, and diagnoses as determined by ICD-9 codes.

Results: EHR phenotyping models utilizing conventional factors, prescriptions, and diagnosis information (EHR-DX) had improved ability to detect T2DM, when compared to models containing conventional (EHR-conventional) or conventional & prescription metrics (EHR-RX). Conversely, there was not a statistically significant improvement in T2DM detection between models containing conventional metrics and models containing conventional & prescription metrics (likelihood ratio test, $p < 0.001$). The AUC for the full EHR DX, EHR RX, and conventional models using logistic regression were 85.8%, 74.8%, and 72.7% respectively.

Several ICD codes were shown to have a significant association with diabetes diagnosis, including several from established predictors or comorbidities, such as prediabetes and heart disease. Associations between antipsychotic drugs and diabetes diagnosis were generally insignificant and lacked a discernible positive or negative association trend.

The thesis of Austin Haynesworth is approved.

Ariana Anderson

Mahtash Esfandiari

Frederic Paik Schoenberg, Committee Chair

University of California, Los Angeles

2020

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge everyone who generously helped me during this project. Foremost, I would like to express my sincere gratitude to Dr. Ariana Anderson, for her guidance, patience, and support throughout this thesis.

I would also like to thank the rest of my committee members, Dr. Mahtash Esfandiari and Dr. Frederic Paik Schoenberg for their generous encouragement and insightful comments.

CHAPTER 1

INTRODUCTION

The association between schizophrenia and type II diabetes has been recognized for more than a century. While diabetes has a prevalence of 13% within the general population, the prevalence of diabetes increases 2-3x in patients with schizophrenia (*Centers for Disease Control, 2020; Cohn, 2012*). This relationship is specific to type II diabetes mellitus (T2DM), as type 1 diabetes mellitus (T1DM) is less common in patients with schizophrenia (*Cohn, 2012*). It is thought that between 20% and 30% of patients with schizophrenia will develop diabetes or prediabetes during the course of psychiatric treatment (*Cohn, 2012*).

Many factors have been proposed to explain this comorbidity, including side effects of antipsychotic medication, poorer overall physical health, unhealthy lifestyle choices, and poorer health care options and accessibility (*Dixon, 2000*). A commonly purported explanation for the increased prevalence of diabetes in schizophrenics is that weight gain, a consequence of lifestyle factors as well as antipsychotics that promote obesity, leads to progressive insulin resistance (*Cohn, 2012*; *Ng-Mak 2019*). For example, many clinicians consider clozapine and olanzapine to be effective in treating schizophrenia. These drugs, however, also have high metabolic liability, which has been suggested to lead to antipsychotic-induced weight gain and development of metabolic syndrome and obesity (*Cohn, 2012; Dixon, 2000*). Quetiapine has also been linked to similar symptoms (*Dixon, 2000*). These metabolic changes and resulting obesity have been associated with diabetes diagnoses. Thus, recognizing the role certain antipsychotics play in diabetes diagnosis can help steer physicians in both schizophrenia and diabetes management.

There have also, however, been suggestions of a genetic basis for the connection between T2DM and schizophrenia. Genetic linkage analyses have identified several loci associated with

schizophrenia that have also been identified in linkage studies in T2DM (*Bellivier, 2005)*. Meanwhile, proteomic studies have shown perturbances in expression of genes involved in glucose metabolism in brain tissue and elevated insulins in patients with schizophrenia when compared to controls. This suggests that the metabolic effects of antipsychotic usage and lifestyle factors may at all least be partially mediated by genetic disposition in patients with schizophrenia (*Hackinger, 2018.)*

Untreated diabetes has serious long-term health consequences including blindness, amputations, and even potential early death from heart attacks (*Cohn, 2012*). As such, early and aggressive intervention for diabetic patients with schizophrenia or patients at high risk of diagnosis for diabetes is important for lowering prevalence, lowering mortality, and improving prognosis.

Unfortunately, currently roughly 25% of individuals with T2DM are undiagnosed (*Anderson, 2015*.) The cost of not treating diabetes is detrimental, both in relation to health and finances. According to the American Diabetes association, the total estimated costs of diabetes in 2012 was $245 billion dollars, with roughly 2/3 of this cost being direct medical costs, and the remainder being costs associated with reduced productivity due to absenteeism and early mortality (*American Diabetes Association, 2013*). These costs underscore the importance for early screening and detection, such that complications can be avoided and progression can be slowed,

Historically, diabetes screening risk scores combine basic demographic, lifestyle, and historical information with laboratory testing to predict the likelihood of developing diabetes (*Anderson, 2015*.) Beyond these factors, however, EHRs have demonstrated potential for detecting and monitoring diabetes. It has been shown that usage of the full electronic medical record -- beyond simple conventional metrics -- to extend screening models can be a useful predictive tool for the development of diabetes (*Anderson, 2015*.) These EHR-based phenotypes

can identify individuals who may benefit from interventions and thereby improve patient treatment and prognosis. While significant phenotyping studies exist for general adults with type 1 and type 2 diabetes, EHR phenotyping for diagnosed schizophrenics to predict diabetes diagnosis has not been extensively studied (*Ng-Mak, 2019*).

This prompted the present study which attempted to evaluate the usefulness of expanded diabetic screens that include antipsychotic prescriptions and full EHR ICD-9 code data.

CHAPTER 2

MATERIALS AND METHODS

2.1 Overview

A data set including patients with schizophrenia was mined from University of California Health Systems EHR data. Data included transcripts from 1997 patients seen between 1991-2017. On average, patients had 2988 days (median = 2876 days) between their first and last encounter. On average, patients developed diabetes after 2524 days (median = 2661 days) of being followed. Further demographic information is provided in Table 1.1.

Patients were determined to have T2DM using a surveillance algorithm developed by Klompas et al. This algorithm utilizes laboratory diagnostic criteria, suggestive medication prescriptions, and ICD-9 codes. Specifically, Klompas utilized the following T2DM diagnostic metrics:

- Hemoglobin A1C > 6.5%
- Fasting glucose > 126 mg/dL
- Prescription for insulin outside of pregnancy (Cases where the prescription of insulin happened during or after the pregnancy were filtered out)
- ICD-9 code 250 on two or more occasion.
- Prescription for one or more of the following antidiabetic medications: glyburide, gliclazide, glipizide, glimepiride, pioglitazone, rosiglitazone, repaglinide, nateglinide, meglitinide, sitagliptin, exenatide, pramlintide

The Klompas algorithm was also used to distinguished between type 1 and type 2 diabetes. As the connection between schizophrenia and diabetes only exists in T2DM, only non-diabetics and

those with type II diabetes were included in this study. Patients with type 1 diabetes (n=11) were removed. Incomplete cases were also removed (n=730).

We then assessed whether T2DM risk scores could be improved with prescription profiles and full EHR phenotypes, created using the additional medical and diagnostic information contained in the EHR. Our methodology roughly mimicked that which was used in Anderson et al. We predicted current T2DM status using a multivariate logistic regression model in R comparing three separate models:

- "Conventional model" mimicking conventional risk scores;
- ''EHR RX" model which contained conventional information along with prescription information;
- A full EHR model ("EHR DX"), based upon the EHR phenotype, containing conventional information, diagnostic information based on ICD codes, and prescription information.

The EHR RX model is specifically important for patients with schizophrenia, as prescribing metabolic syndrome-inducing antipsychotics is believed to lead to diabetes diagnosis (Cohn).

2.2 Introduction to data & data processing

Once missing data was removed, our dataset consisted of 1267 complete cases, from an original 1997 cases. Our response variable, diabetes diagnosis contained 133 subjects being classified as having diabetes and 1134 classified as without diabetes (Table 1.1).

	Non-Diabetic	Type 2 Diabetes
Number of patients	1134	133
Male %	53.10%	52.60%
Age (years)	47 (18)	56 (15)
BMI	27(6)	29(7)
Systolic BP mmHG	123(17)	127 (20)
Diastolic BP mmHG	76(11)	76(12)
Total Diabetes Risk Factors	1.7(1.2)	2.3(1.5)
Hypertension DX (%)	31.30%	64.70%
High Cholesterol (%)	14.10%	17.30%
Smoking (%)	31.50%	27.10%

Table 1.1: Introduction to the 1267 complete cases that were used in the model. Values indicate mean and standard deviation in parenthesis for a given value (unless otherwise stated as a percentage).

2.3 Selecting Predictor Variables and Sample Size Consideration

Initially, our dataset consisted of 1267 complete cases with 193 predictor variables. Due to the relatively small number of observations and the fact that our response variable was unbalanced, a reduction of our independent variables was necessary (*Peduzzi et al. ,1996*). Reduction choice was based on:

1. To reduce bias, we removed as predictors established treatments and complications of T2DM. This included primary and secondary diabetes-related diagnosis (ICD-9 250.x2, 249.x), diabetic retinopathy (ICD-9 366.41), and medications used to treat diabetes such as metformin and insulin (Anderson, et al). (4)
2. Antipsychotic medications that lacked a single prescription amongst the 1267 patients were removed. (7)
3. Medications that are no longer prescribed due to being withdrawn by the FDA were removed. (1)

4. Predictors with high degrees of multicollinearity or aliasing, as determined by the variance inflation factor quotient (VIF), were removed. (6)
5. Initially 145 most common ICD-9 codes were included in the dataset. Because this large number of predictors violates our sample size restriction (and led to convergence issues for the logistic models), the 50 most common ICD-9 codes were retained (*Fig. 2.1*) while the rest were removed. (95)

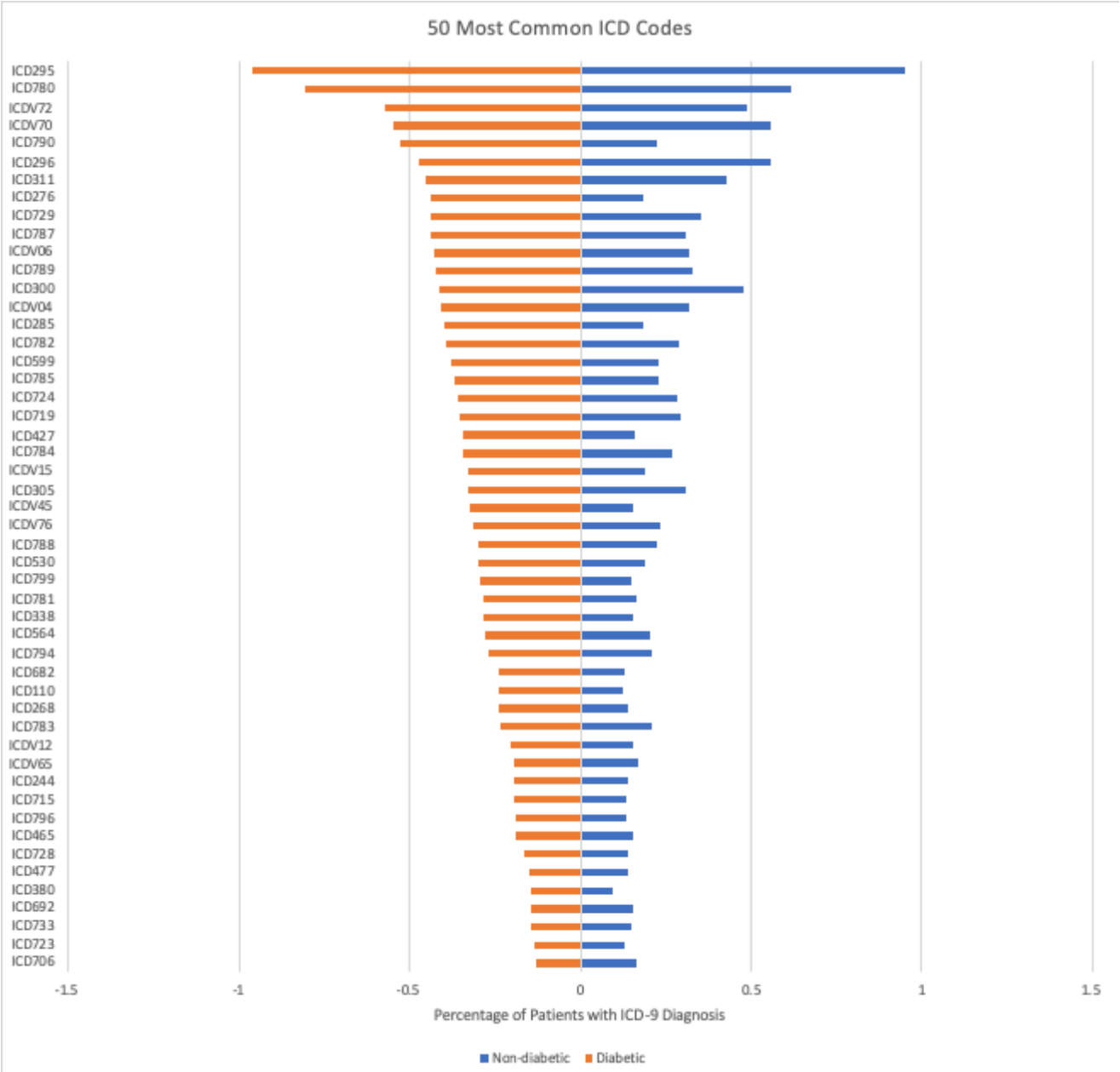

Fig 2.1: *The 50 most common ICD-9 diagnoses across 1267 subjects. Proportion of diabetic and non-diabetic cohort with the diagnosis is reported for each ICD-9 code.*

2.4 Description of models

The first logistic regression model (conventional model) mimics conventional risk scores by including only the limited subset of covariates (smoking status, sex, age, BMI, cholesterol, and hypertensive status) that have been used in current diabetes risk models. Systolic & diastolic blood pressure were also included. Lastly a backwards stepwise regression was performed to determine relevant interaction effects to be included in the model.

The second logistic regression model (EHR RX) included conventional information but also included prescriptions. As described in Cohn et al, antipsychotic use and the related weight gain and lifestyle side effects, are suggested factors for diabetes prevalence in patients with schizophrenia. Medications up until the date of Klompas diagnosis were included, as it is assumed a clinician's prescribing behavior could be influenced by patient's diabetes status (*Anderson et al, 2015*.)

For the full EHR logistic regression model (EHR DX), 50 most commonly diagnosed ICD-9 codes were included, in addition to basic conventional screening information and prescriptions. An additional variable, total number of diabetes risk factors (as described in Anderson et al 2015), was also included. These risk factors are common comorbidities of diabetes. Patient's full ICD-9 code record was included, sans codes associated with diabetes diagnosis and complications as described above. Diabetes is often not diagnosed until long after symptoms begin to occur (*Katulanda, 2016.)* As the Klompas classification algorithm identifies both confirmed diagnosed and undiagnosed T2DM, usage of the full ICD-9 record allows for more complete screening.

To gauge performance, we compared the performance of these models using a likelihood ratio test. Briefly, a likelihood ratio test is used to compare two nested models and takes the form:

(1)
$$LRT = -2log\frac{L_s(\hat{\theta})}{L_g(\hat{\theta})}$$
$$= deviance_s - deviance_g$$

where L_s is the simpler model and L_g is the more complex model. This test statistic is approximately χ^2.

Next, we computed performance statistics such as receiver operating characteristic (ROC) curves and area under the curve (AUC) to determine the ability to detect T2DM.

Within each model, the significance of covariates was evaluated based on Wald chi-squared statistic, and the most significant covariates were highlighted.

Finally, we validated the findings from our logistic models externally using a random forest prediction model.

Data was split into test and train sets at a 70:30 ratio. All models were trained and tested on the same train and test sets.

2.5 Mathematical Basis of Logistic Regression and Random Forest Models

The two primary classification techniques employed in this study are logistic regression and random forest. We will briefly speak about the mathematical basis of each.

i) Logistic Regression

Logistic regression is a method of classification. It models log odds, which are defined as:

(2)
$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_o + \beta_1 X_1 + \cdots + \beta_{k-1} X_{k-1}$$

where our left term is our log odds, or logit, and *p* is the probability of our event occurring. Unlike in linear regressions where our outcome variable *y* is numeric, in logistic regression, we model the

probability that *y* belongs to a certain class. In binomial logistic regression, this means that $x \in R$ and $p(x) \in [0,1]$, and we model *y* using the function:

(3)
$$E(Y|x) = \pi(x) = \frac{e^{\beta_o + \beta_1 x}}{1 + e^{\beta_o + \beta_1 x}}$$

For estimation of our parameters, β, logistic regression utilizes maximum likelihood estimation:

(4)
$$L(\beta) = \prod_{i=1}^{n} p(x_i)^{y_i} \ (1 - p(x_i)^{1-y_i}$$

$$l(\beta_o, \beta) = \sum_{i=1}^{n} y_i log p(x_i) + (1 - y_i) log 1 - \ p(x_i)$$

$$\frac{\partial l}{\partial \beta_j} = \ - \sum_{i=1}^{n} \frac{1}{1 + e^{\ \beta_o + x_i \beta}} e^{\beta_o + x_i \beta} \ x_{ij} + \sum_{i=1}^{n} y_i x_{ij}$$

$$= \sum_{i=1}^{n} (y_i - p(x_i; \beta_o, \beta)) x_{ij}$$

We end with the final form of the log likelihood function, which can be optimized via finding the value of β that maximizes this function. Solving this likelihood function can be done with Newton – Raphson approach.

The statistical significance of the betas is determined via the Wald chi-squared statistic. For each estimated $\beta = b$ in logistic regression, the Wald chi-square is defined as:

(5)

$$\chi^2_{wald} = (\frac{b}{s.e.(b)})^2 \text{ , with confidence interval: } b \pm Z\ s.e.(b) \text{ ,}$$

where Z is the Gaussian percentile.

In the context of our study, in each binomial logistic regression model, we will calculate the β for each covariate then determine its significance via its Wald statistic.

ii) Random Forest

Random forest is a technique used for classification. It works similarly to a decision tree, but, unlike a singular tree, random forest uses a collection of decorrelated decision trees to reach a final classification recommendation.

Random forest is based on the concept of bootstrapping, a technique of estimating statistics of a population by sampling the dataset with replacement. Specifically, each decision tree within the random forest makes predictions based on a bootstrapped data set and using a random subset of predictor variables to arrive at its final Y classification. The final predicted Y classification is determined as a vote across all trees.

iii) Methods of Evaluation

Our models will be evaluated via receiver operator curve (ROC) and area under the curve (AUC). The ROC provides a graphical way to summarize the true positive rate (sensitivity) versus the false positive rate (1-specificity.) It is a curve that summarizes all of the true positive versus false positive values that a given threshold in our model will produce, where false positive is defined as an error in data reporting in which the model improperly indicates presence of a condition where it does not exist. The area under this curve is AUC. A higher AUC indicates that

we are maximizing true positives while minimizing false positives, with a maximum possible value of 1. In the context of our study, AUC is used to compare various models to determine how good they are at correctly classifying diabetes.

Additionally, each model will be evaluated using a confusion matrix, that will categorize the model's performance on a test data set. Sensitivity (ability to determine diabetic cases correctly), specificity (ability to determine non-diabetic cases correctly) and accuracy (ability to differentiate diabetic and non-diabetic cases) will all be reported.

CHAPTER 3

RESULTS

3.1 Comparing Conventional versus Expanded Model Predictive Ability

Incorporating EHR prescription and ICD-9 data improved classification accuracy when compared with using conventional diabetes screen covariates for the logistic regression models. Incorporating only drug data, however, saw a modest yet statistically insignificant improvement in classification performance. Specifically, the EHR DX model predicted better than the conventional model, but the EHR RX model did not improve classification accuracy over conventional model to a statistically significant degree (likelihood ratio test, $p < 0.001$; *Table3.1*).

Model Comparison	Likelihood Ratio χ^2	Likelihood Ratio df	p-value
Conventional v. EHR RX	15.38	20.00	0.754
Conventional v. EHR DX	125.11	70.00	5.77×10^{-5}
EHR RX v. EHR DX	109.73	50.00	2.31×10^{-6}

Table 3.1: *Likelihood ratio test for conventional, EHR RX, and EHR DX logistic models. (*) indicated significant at p < .001 level.*

For the EHR DX, EHR RX, and conventional logistic regression models, the AUC was 85.8%, 74.8%, and 72.7% respectively (Figure 3.1).

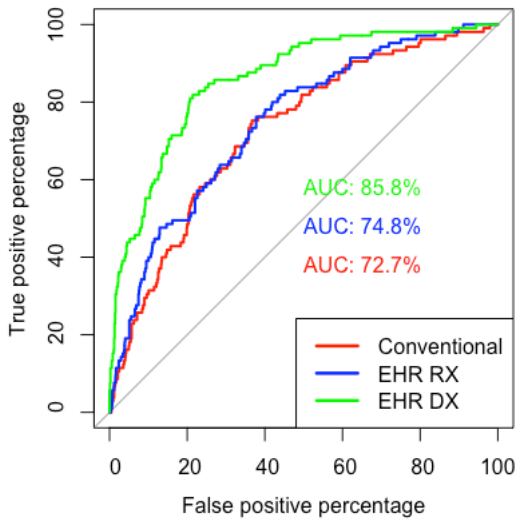

Figure 3.1: *ROC curves with respective AUC percentages for all three logistic models. As more EHR phenotype information was added, AUC increased, indicating that EHR phenotyping improved ability to detect T2DM.*

Model	AUC	AIC	BIC	McFadden R^2
Conventional	0.727	641.19	718.89	0.095
EHR RX	0.748	657.68	832.52	0.130
EHR DX	0.858	629.13	1037.07	0.315

Table 3.2: *Summary statistics for our three logistic regression models*

Performance tests for accuracy, sensitivity, and specificity were performed and confusion matrices were produced (*Figure 3.2*).

Conventional - Logistic Regression

	Reference	
Prediction	0	1
0	158	5
1	131	23

Conventional - Random Forest

	Reference	
Prediction	0	1
0	212	10
1	77	18

EHR RX - Logistic Regression

	Reference	
Prediction	0	1
0	159	5
1	130	23

EHR RX - Random Forest

	Reference	
Prediction	0	1
0	200	9
1	89	19

EHR DX - Logistic Regression

	Reference	
Prediction	0	1
0	209	6
1	80	22

EHR DX - Random Forest

	Reference	
Prediction	0	1
0	199	7
1	90	21

Figure 3.2: *Confusion matrices of logistic regression and random forest models run on test data set of 317 subjects (289 non-diabetic, 28 diabetic.)*

From these matrices, we can see that the EHR DX model outperformed the conventional and EHR RX model. The EHR RX and conventional model performed nearly identically, with the small difference in performance being statistically insignificant. Random forest showed a similar trend. Models were evaluated on accuracy (the ability of the model to differentiate diabetic and non-diabetic cases correctly), specificity (the ability of the model to correctly determine diabetic cases), and sensitivity (the ability of the model to correctly determine non-diabetic cases.) Full model performance can be seen in *Table 3.3* and graphically in *Figure 3.3*. Train-versus-test data performance can be found in *Appendix Figure A.1*.

Thresholds were determined via Youden's J index. Briefly, the Youden's J index is defined as a combined metric of sensitivity and specificity (Sensitivity + Specificity - 1) and has a value between 0 and 1.

(7) $$J = Max_c(sensitivity_c + specificity_c - 1$$

In diagnostic cases where sensitivity and specificity are diagnostically important, the Youden index will indicate the performance at a given cutoff, and, under these circumstances *J* defines an optimal cutoff, *c* (*Berrar, 2019*). The logic behind choosing this threshold (as opposed to simply optimizing accuracy) was that improved sensitivity, sometimes at the detriment of accuracy, is important in diabetes diagnosis, given that a missed diagnosis (false negative) was more costly than a false positive.

	Accuracy	Sensitivity	Specificity
Conventional - Logistic	0.571	0.820	0.540
EHR RX - Logistic	0.571	0.830	0.550
EHR DX - Logistic	0.729	0.786	0.723
Conventional - RF	0.726	0.643	0.734
EHR RX - RF	0.691	0.679	0.692
EHR DX - RF	0.694	0.750	0.689

Table 3.3: *Accuracy, sensitivity and specificity of logistic and random forest models, run on validation data.*

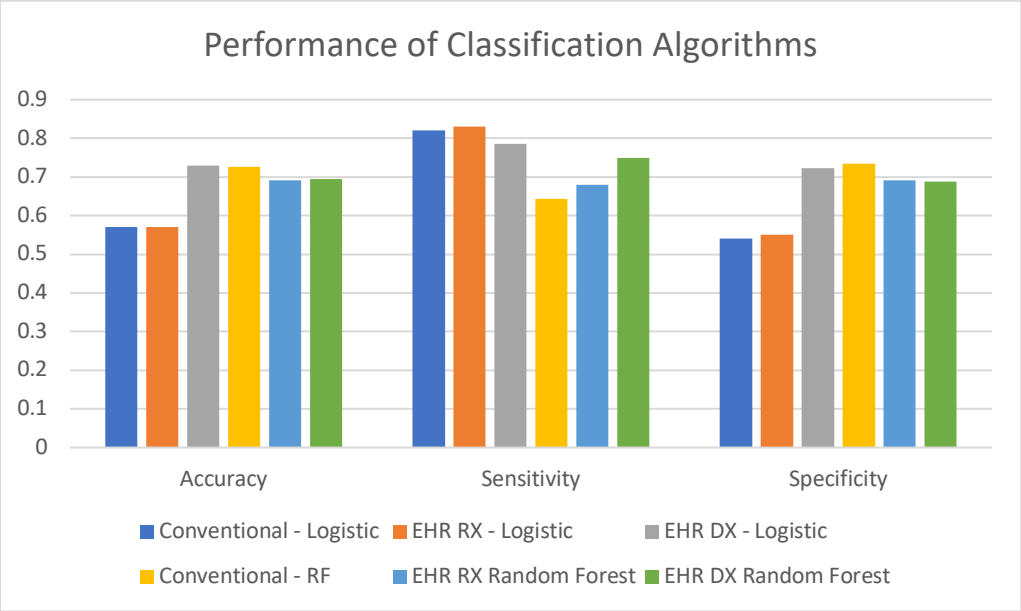

Figure 3.3 *Accuracy, sensitivity and specificity of logistic and random forest models, run on validation data. Result shows increased ability to detect diabetes when incorporating full EHR phenotype data, but only a modest, insignificant increase when incorporating only drug data.*

3.2 Identifying factors significant in diabetes diagnosis.

Beyond model performance, we were also interested in identifying significant factors for diabetes diagnosis. According to conventional logistic regression model, only hypertension status and BMI were significant at the $p < 0.001$ level, though several factors were significant at the $p < 0.10$ level (*see Appendix Table A.1)*.

In the EHR RX model, BMI and hypertension were once again highly significant. Of the 20 drugs included in the model, none were shown to be significant at the $p < 0.05$ level. Antipsychotics like olanzapine and quetiapine, which we hypothesized could be related to antipsychotic-induced weight gain, development of metabolic syndrome, and diabetes, were not found to be significant.

The EHR DX found similar results in relation to conventional factors and medications (*Figure 3.4*). Because of the inclusion of the ICD-9 codes, it also provides information regarding significant comorbidities. A total of 7 ICD codes were found to be significant at the $p < 0.05$ level.

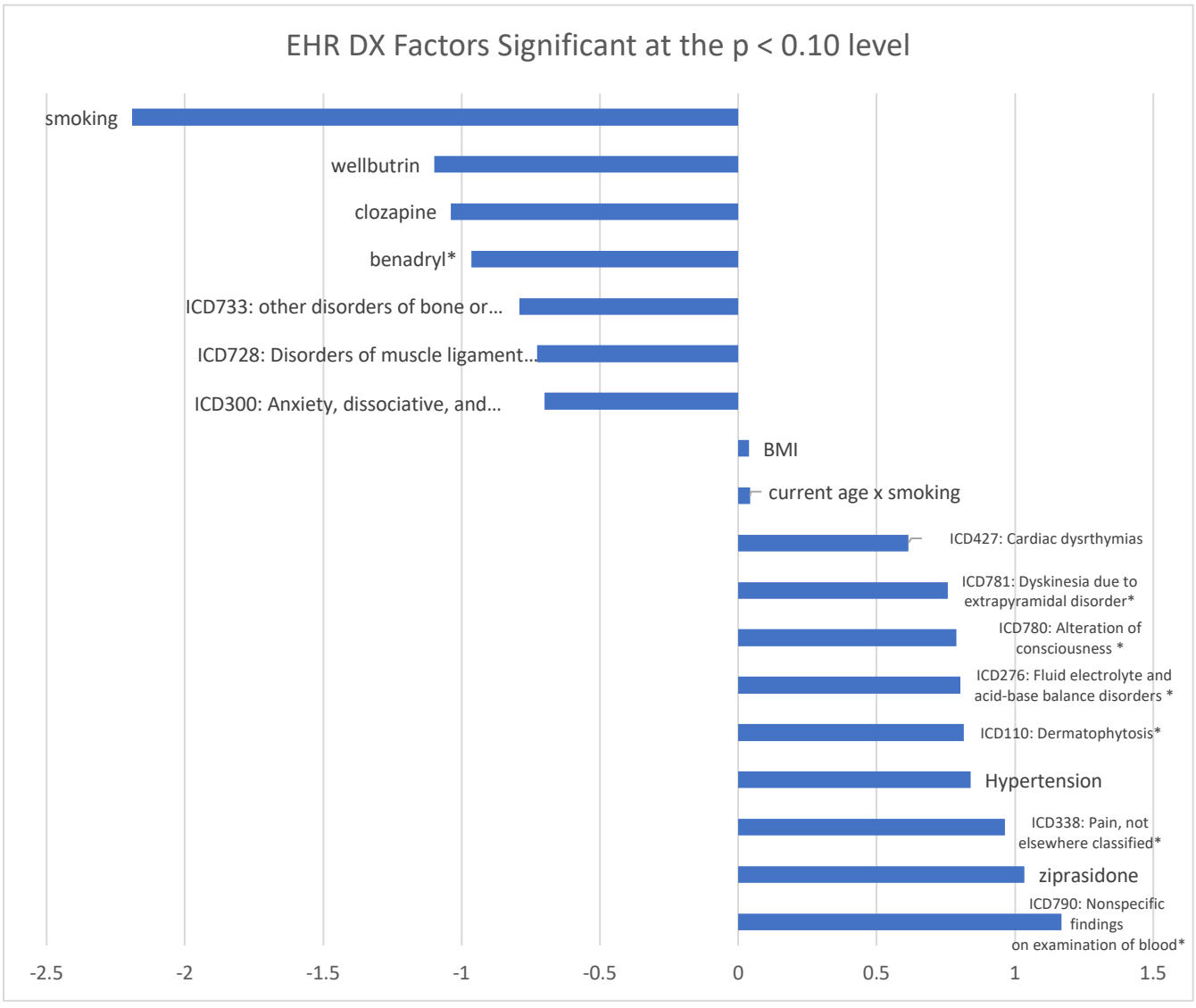

Figure 3.4 *Significant factors for EHR DX.* Factors with (*) were significant at the $p < 0.05$ level.

CHAPTER 4

DISCUSSION

4.1 Usage of EHR Phenotyping in Diabetes Diagnosis for Patients with Schizophrenia

Usage of the full EHR DX model brought significant improvements to diabetes identification when compared to conventional diabetes screening metrics or conventional metrics plus prescription information. Usage of only prescriptions with conventional metrics did not see a statistically significant increase in performance over conventional metrics alone.

4.2 Interpretation of Factors Identified to be Significant

Using the EHR DX model, we were able to identify conventional, drug, and diagnostic factors that were shown to be significant in identifying type II diabetes.

The identified conventional factors were generally unsurprising. Hypertension was consistently identified across all three logistic models as the most significant conventional factor and had a positive association, with the odds of diabetes diagnosis increasing by 2.9, 3.0, and 2.3 times in the conventional, RX, and DX models respectively ($p<0.05$ for all three models). This is consistent with the literature, as hypertension is common amongst diabetic patients and is a strong risk factor for severe cardiac complications, which are the leading cause of morbidity for diabetic patients (*Boer et al, 2017)*. Similarly, BMI had a consistently positive association across all three models, albeit with a smaller effect size compared to hypertension, with the odds of diabetes diagnosis increasing by 4% for all three models ($p < 0.05$.)

The identified drug factors were inconclusive. Across both the EHR RX and EHR DX models, only Benadryl was found significant at a $p < 0.05$ level, with the odds of diabetes diagnosis being 62.9% less likely. This may be because Benadryl is often used to treat extrapyramidal

symptoms such as dystonia from first generation anti-psychotic drugs, which are typically associated with less weight gain and less metabolic side effects.

Associations for various antipsychotic drugs was mixed. Olanzapine and quetiapine, which have been posited to increase diabetes risk due to their impacts on hyperglycemia and metabolic syndrome had a negative, statistically insignificant association. Clozapine, which was significant at the $p < 0.10$ level was significant in the EHR DX model, also had a negative association. Ziprasidone, another antipsychotic, had a positive, statistically significant association, with odds of diabetes diagnosis increasing 2.8 times. This is not entirely inconsistent with established literature, which shows varying (both positive and negative) associations between second-generation antipsychotics and diabetes diagnosis (*Citrome, 2013*). It is also important to consider that these findings may be reflective of prescribing patterns, as doctors may be less likely to prescribe an antipsychotic with known metabolic side effects to an overweight patient. This explanation would be consistent with the findings related to Benadryl.

The identified diagnoses create for interesting potential future directions. Some identified ICD-9 codes were quite logical. For example, ICD-9 790, "nonspecific findings on examination of blood", was significant with 3.2 times increased odds ($p < 0.001$). Within ICD-9 790 exists ICD-9 790.2: abnormal glucose, impaired fasting glucose, impaired glucose tolerance test, and other abnormal glucose. This code is used to diagnose pre-diabetes, thus it is unsurprising that it has a significant positive association for diabetes diagnosis. 53% of diabetic patients in our study had an ICD790 diagnosis as opposed to 22.6% of non-diabetic patients. According to the American Diabetes Association, up to 70% of individuals with prediabetes will eventually develop diabetes.

Other common comorbidities or complications of diabetic symptoms were also identified to have positive associations, like ICD 276, which includes ketoacidosis, a complication associated

with secondary diabetes (odds ratio 2.23; p = 0.016). Major cardiac complications (ICD 427) were also found to be significant with a positive association (odds ratio 1.8; p = 0.075).

A few ICD codes identified as significant are more commonly found in schizophrenics than the general population, and could be interesting areas for future discussion. For example, ICD300: Anxiety, dissociative, and somatoform disorders was associated with 51% decreased odds of diabetes diagnosis (p = 0.022). This was surprising, as the literature suggests a bidirectional increase of diabetes in those with anxiety, and increased anxiety in those with diabetes within the general population (*Chien et al, 2016)*. Further investigation of both the behavioral and physiological impacts of anxiety, schizophrenia, and diabetes would be an interesting future direction.

4.3 Comparison of identified factors with non-schizophrenic patients

Within this study, only patients with schizophrenia were analyzed for diabetes risk. This population was selected because although risk is well-studied in the general population it is unknown whether this population may have unique risk factors, given the impact of the disorder. For example, patients with schizophrenia often have disorganized and dissociated thinking, which in turn leads to worse self-care and compliance. This could lead to them achieving a later diagnosis of diabetes, with more associated complications. Comparison of identified factors within this study against similar studies for non-patients with schizophrenia could help pinpoint schizophrenic-specific factors for diabetes diagnosis.

While this study did not analyze non-diabetic patients, the methodology of this study closely mirrors that of Anderson et al 2015. Comparing that results of this study to Anderson's

study, hypertension and BMI had significant, positive associations for diabetes diagnosis across both patients with schizophrenia and those without.

In regards to drug or diagnostic similarities, the two studies observed different drugs, with this study primarily focusing on antipsychotics, thus reducing the comparison value. Similarly, the studies lack substantive ICD-9 code similarities.

4.4 Future directions

Because of the data used, this analysis was limited in its scale. Specifically, all data was mined from the University of California Health System. While this includes a wide range of hospitals, there is inherently a limited geographic range of patients, and consequently a potentially limited scope of environmental factors. Additionally, due to the number of complete case observations available, we were unable to incorporate all 145 ICD codes that existed within the data set. This may have led to missing potentially significant diagnoses. Within this current cohort, it would be interesting to use lesser diagnosed ICD codes as predictors to identify additional significant diagnoses. Alternatively, utilizing a dataset with a higher number of observations would support a larger number of predictor variables. Lastly, another issue is the smaller number of those who have diabetes type II within our dataset compared to those who do not. This imbalance could affect the significance of the findings. As such, for future directions, mining data on a national scale and collecting a larger, more balanced number of observations will allow for a more robust study.

APPENDIX

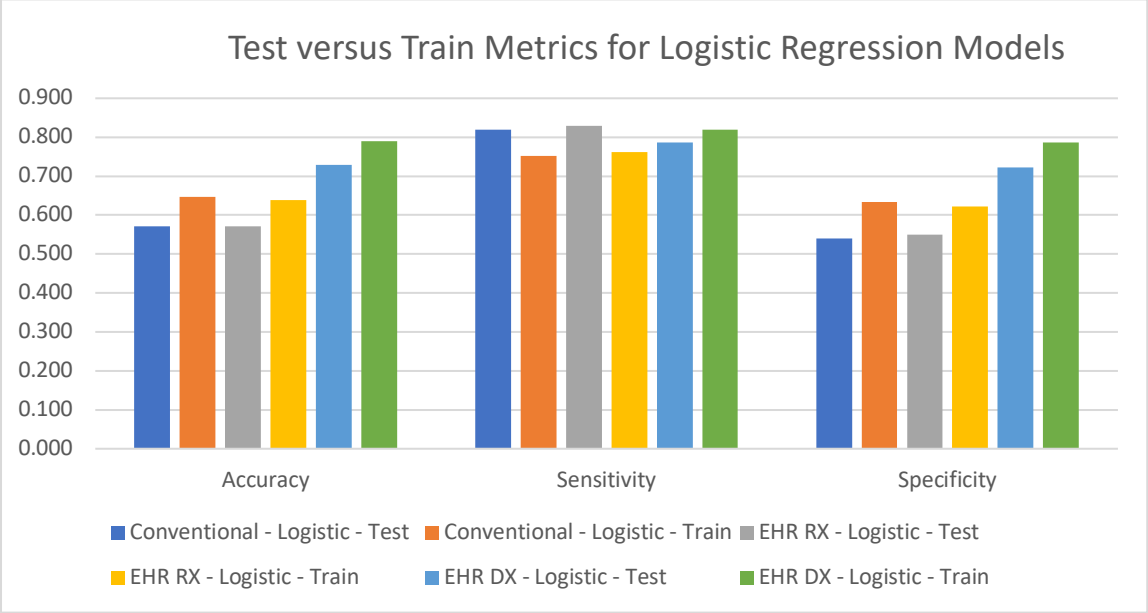

***Figure A.1*:** Train versus test performance of logistic regression models

Factors	Log odds	Std. Error	z value	Pr(>|z|)	
Age	0.015	0.010	1.394	0.163	
Gender (male = 1)	0.334	0.287	1.164	0.244	
BMI	0.044	0.015	2.857	0.004	**
Systolic BP	-0.014	0.009	-1.595	0.111	
Has Hypertension	1.067	0.333	3.208	0.001	**
High Cholesterol	1.953	1.440	1.356	0.175	
Smoking	-2.032	1.110	-1.830	0.067	.
Area	-0.016	0.016	-1.032	0.302	
Age x High Cholesterol	-0.030	0.021	-1.415	0.157	
Age x Smoking	0.030	0.018	1.661	0.097	.
Age x Area depravation index	0.000	0.000	1.213	0.225	
Gender x High Cholesterol	-0.854	0.619	-1.380	0.168	
Gender x Smoking	0.632	0.598	1.057	0.290	
Systolic BP x Area depravation index	0.000	0.000	0.619	0.536	
Has hypertension x Area depravation index	-0.003	0.005	-0.571	0.568	

Table A.1: *Log odds ratios for conventional logistic regression model.*
*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Factors	Log Odds	Std. Error	z value	Pr(>|z|)	
BMI	0.041	0.016	2.530	0.011	*
Has Hypertension	1.091	0.343	3.185	0.001	**
Smoking	-2.365	1.157	-2.044	0.041	*
wellbutrin	-0.832	0.497	-1.676	0.094	·
Age x Smoking	0.036	0.019	1.921	0.055	·

Table A.2: *Log odds ratios for EHR RX logistic regression model. Only factors identified as significant at the p < 0.1 level were included.*
*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Factor	Log Odds	Std. Error	z value	Pr(>|z|)	
BMI	0.038	0.020	1.952	0.051	·
Has Hypertension	0.840	0.403	2.086	0.037	*
Smoking	-2.190	1.330	-1.647	0.100	·
clozapine	-1.038	0.593	-1.752	0.080	·
ziprasidone	1.034	0.537	1.927	0.054	·
wellbutrin	-1.098	0.577	-1.902	0.057	·
benadryl	-0.965	0.437	-2.208	0.027	*
ICD733	-0.791	0.432	-1.829	0.067	·
ICD728	-0.728	0.410	-1.776	0.076	·
ICD110	0.816	0.352	2.319	0.020	*
ICD338	0.964	0.383	2.513	0.012	*
ICD781	0.757	0.347	2.181	0.029	*
ICD427	0.615	0.345	1.783	0.075	·
ICD300	-0.700	0.305	-2.298	0.022	*
ICD276	0.802	0.332	2.419	0.016	*
ICD790	1.169	0.301	3.883	0.000	***
ICD780	0.788	0.336	2.341	0.019	*
Age x Smoking	0.043	0.023	1.895	0.058	·

Table A.3: *Log odds ratios for EHR DX logistic regression model. Only factors identified as significant at the p < 0.1 level were included.*
*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

REFERENCES

1. Anderson A, Kerr W, et al. *Electronic health record phenotyping improves detection and screening of type 2 diabetes in the general United States population*: *A cross-sectional, unselected, retrospective study*. Journal of Biomedical Informatics 60 (2016) 162–168.
2. Klompas M, Lazarus R, et al. *Automated Detection and Classification of Type 1 Versus Type 2 Diabetes Using Electronic Health Record Data*. Diabetes Care, 2013; 36:914–921
3. Cohn T. *Schizophrenia and diabetes: vigilant metabolic monitoring informs treatment decisions*. 2012. Current Psychiatry Vol 11 No.10.
4. Dixon L, Weiden P, et al. *Prevalence and Correlates of Diabetes in National Schizophrenia Samples*. Schizophrenia Bulletin, (2000) Vol. 26, No.4.
5. Ng-Mak D& Reutsch C. *Association Between Meaningful Use of Electronic Health Records and Patient Health Outcomes in Schizophrenia: A Retrospective Database Analysis*. Am J Manag Care. 2019;25:S159-S165
6. Centers for Disease Control. U.S. Department of Health and Human Services. "*National Diabetes Statistics Report 2020: Estimates of Diabetes and Its Burden in the United States*.
7. American Diabetes Association. *Economic Costs of Diabetes in the U.S. in 2012*. Diabetes Care 2013 Mar; DC_122625
8. Bellivier F. *Schizophrenia, antipsychotics and diabetes: Genetic aspects*. Eur Psychiatry. 2005 Dec;20 Suppl 4:S335-9.
9. Hackinger S, Prins B. *Evidence for genetic contribution to the increased risk of type 2 diabetes in schizophrenia*. Translational Psychiatry, 8, Article number: 252 (2018)
10. Peduzzi P, Concato J, et al. *A simulation study of the number of events per variable in logistic regression analysis*. Journal of Clinical Epidemiology. 1996 Dec;49(12):1373-9.
11. Cosslett, S & Manski, C & McFadden, D. (1981). *Efficient Estimation of Discrete Choice Models*.
12. Boer I, Bangalore S, et al. Diabetes and Hypertension: *A Position Statement by the American Diabetes Association*. *Diabetes Care* 2017 Sep; 40(9): 1273-1284.
13. Citrome L, Collins JM, et. *Incidence of cardiovascular outcomes and diabetes mellitus among users of second-generation antipsychotics*. The Journal of Clinical Psychiatry, 30 Nov 2013, 74(12):1199-1206

14. Ojo O, Wang X, et al. *The Effects of Substance Abuse on Blood Glucose Parameters in Patients with Diabetes: A Systematic Review and Meta-Analysis*. Int J Environ Res Public Health. 2018 Dec; 15(12): 2691.

15. Berrar, D. *Performance measures for Binary Classification*. Encyclopedia of Bioinformatics and Computational Biology, 2019.

16. Chien I, Ching L. *Increased risk of diabetes in patients with anxiety disorders: A population-based study*. *Journal of Psychosomatic Research, Volume 86*. July 2016, Pages 47-52.

17. Katulanda P, Hill NR, et al. *Development and validation of a Diabetes Risk Score for screening undiagnosed diabetes in Sri Lanka (SLDRISK)*. *BMC Endocrine Disorders* volume 16, Article number: 42 (2016)