Predicting Depression among Patients with Diabetes Using Longitudinal Data*

A Multilevel Regression Model

H. Jin; S. Wu; I. Vidyanti; P. Di Capua; B. Wu

1Epstein Department of Industrial and Systems Engineering, University of Southern California, Los Angeles, California, USA;
2School of Social Work and Epstein Department of Industrial and Systems Engineering, University of Southern California, Los Angeles, California, USA;
3Division of Chronic Disease and Injury Prevention, Los Angeles County Department of Public Health, Schaeffer Center of Health Policy and Economics, University of Southern California, Los Angeles, USA;
4Baptist Health South Florida, Miami, Florida, USA;
5Keck School of Medicine, University of Southern California, Los Angeles, California, USA

Keywords
Depression, diabetes mellitus, comorbidity, machine learning, multilevel regression

Summary
Introduction: This article is part of the Focus Theme of Methods of Information in Medicine on “Big Data and Analytics in Healthcare”.

Background: Depression is a common and often undiagnosed condition for patients with diabetes. It is also a condition that significantly impacts healthcare outcomes, use, and cost as well as elevating suicide risk. Therefore, a model to predict depression among diabetes patients is a promising and valuable tool for providers to proactively assess depressive symptoms and identify those with depression.

Objectives: This study seeks to develop a generalized multilevel regression model, using a longitudinal data set from a recent large-scale clinical trial, to predict depression severity and presence of major depression among patients with diabetes.

Methods: Severity of depression was measured by the Patient Health Questionnaire PHQ-9 score. Predictors were selected from 29 candidate factors to develop a 2-level Poisson regression model that can make population-average predictions for all patients and subject-specific predictions for individual patients with historical records. Newly obtained patient records can be incorporated with historical records to update the prediction model. Root-mean-square errors (RMSE) were used to evaluate predictive accuracy of PHQ-9 scores. The study also evaluated the classification ability of using the predicted PHQ-9 scores to classify patients as having major depression.

Results: Two time-invariant and 10 time-varying predictors were selected for the model. Incorporating historical records and using them to update the model may improve both predictive accuracy of PHQ-9 scores and classification ability of the predicted scores. Subject-specific predictions (for individual patients with historical records) achieved RMSE about 4 and areas under the receiver operating characteristic (ROC) curve about 0.9 and are better than population-average predictions.

Conclusions: The study developed a generalized multilevel regression model to predict depression and demonstrated that using generalized multilevel regression based on longitudinal patient records can achieve high predictive ability.

1. Introduction

Individuals with diabetes are twice as likely as the general population to experience clinically significant depressive symptoms [1, 2]; however, 45% of those with depression are not assessed for depressive symptoms, and their depression goes undiagnosed and untreated [3]. Since depression has significant impact on healthcare outcomes, utilizations and costs as well as elevating suicide risk [4–7], a model to predict depression among diabetes patients is a promising and valuable tool for providers to proactively assess depressive symptoms.
and identify those with depression. Depression is also an often lifelong and highly recurrent condition, and its symptoms may change and be correlated over time [8–11]. Thus, a longitudinal model that takes into account patient historical records may achieve better predictive accuracy than a model that uses only information assessed at one time point.

There are only a few studies in the literature related to developing depression prediction models. The predictD study [12, 13] developed regression models to predict the onset of major depression, using predictors selected from 39 known depression risk factors; those models were trained on a dataset from European countries and validated on an independent dataset from Chile. Huang, LePendu [14] developed regression-based models to predict depression, its severity, and its response to treatment, using electronic health record data that included structured diagnosis and medication codes as well as free-text clinical reports. Wang, Sareen [15] developed a regression model to predict first onset of major depression in the general US population, using data from a national survey.

Longitudinal data are desirable for use in a prediction model because they contain information regarding developments and progressions in patient condition and health status over time. In a longitudinal data set, the set of observations on one subject are not statistically independent over time (16). Since this correlation must be taken into account to make valid predictions, many widely used prediction models such as the ordinary regression methods are inappropriate because those methods lack the ability to model the correlations.

2. Objective

The objective of this study is to develop a generalized multilevel regression model, using a longitudinal dataset from a recent large-scale clinical trial, to predict depression severity and the presence of major depression among patients with diabetes.

3. Methods

3.1 Predicted Outcome

The predicted outcome in this study is the Patient Health Questionnaire PHQ-9 score. PHQ-9, a well-validated scale for assessing severity of depressive symptoms [17, 18], consists of nine questions that are the same as the nine criteria used for the diagnosis of depressive disorders as defined by the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV). Each question has four ordinal response categories with assigned scores from 0 to 3; thus, the overall scale has a score range from 0 to 27, where the higher score indicates more severe depressive symptoms. PHQ-9 can also be used to diagnose major depression. Kroenke, Spitzer [18] validated the PHQ-9, a well-validated scale for assessing depressive disorders as defined by the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV). Each question has four ordinal response categories with assigned scores from 0 to 3; thus, the overall scale has a score range from 0 to 27, where the higher score indicates more severe depressive symptoms. PHQ-9 can also be used to diagnose major depression. Kroenke, Spitzer [18] validated the PHQ-9, a well-validated scale for assessing depressive disorders as defined by the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV). Each question has four ordinal response categories with assigned scores from 0 to 3; thus, the overall scale has a score range from 0 to 27, where the higher score indicates more severe depressive symptoms. PHQ-9 can also be used to diagnose major depression. Kroenke, Spitzer [18] validated

3.2 Dataset

The dataset used in this study was from the Diabetes-Depression Care-management Adoption Trial (DCAT), a comparative effectiveness study from 2010 to 2013 with three arms: Usual Care (UC), Supported Care (SC), and Technology Care (TC). Participants in the UC arm received usual care from clinics in Los Angeles County Department of Health Services (LACDHS), the second largest safety-net healthcare system in the United States. Patients in the SC and TC arms enrolled in a collaborative care management program [19] during the first 6 months of the trial, after which they went back to usual care. Additionally, patients in the TC arm received automated telephonic depression screening and monitoring during the first 12 months of the trial. DCAT details have been described elsewhere [20–22]. The prediction model was developed from the 853 patients who completed all four assessments: one to establish study baseline plus three follow-up assessments (at 6, 12, and 18 months).

3.3 Modeling the Change of PHQ-9 Score Over Time

Generalized multilevel regression is an extension to the ordinary generalized linear regression method and is useful for modeling longitudinal data [23]. As PHQ-9 score is nonnegative and only takes on integer values, we used a 2-level Poisson regression model to relate both time-varying and time-invariant predictors to the PHQ-9 score. The level-1 model excluding the residual can be written as:

\[
\ln(\text{PHQ9}_{i,t}) \sim \pi_{0,i} + \pi_{1,i}t + \pi_{t\text{vp}}X_{t\text{vp},i,t} + \cdots + \pi_{n\text{vp}}X_{n\text{vp},i,t}
\]

where \(t\) is the time variable (represented in years), where \(t = 0\) is the time when a patient is assigned for the PHQ-9 screening for the first time and the values of the predictors are recorded. In DCAT, \(t = 0\) indicates the study baseline. PHQ9_{i,t} is subject \(i\)'s PHQ-9 score at time \(t\). \(X_{t\text{vp},i,t}\) to \(X_{n\text{vp},i,t}\) are the values of \(l\) time-varying predictors at time \(t\). The level-1 model links the right-hand side linear predictor to the natural log-transformed PHQ-9 score.

The level-2 model relates subject \(i\)'s time-invariant predictors (such as gender and depression history) to the level-1 intercept \(\pi_{0,i}\) and the coefficient of time variable \(\pi_{1,i}\). The level-2 model can be seen in Figure 1.

Figure 1 The second level of the generalized multilevel regression model.
variant predictors for predicting the level-1 coefficient of time variable \( \pi_{1,i} \). A time-invariant predictor can be used to predict \( \pi_{0,i} \) or \( \pi_{1,i} \) or both. \( \zeta_{0,i} \) and \( \zeta_{1,i} \) represent between-individual variances, which are assumed to be bivariate normally distributed with mean 0; unknown variances, \( \sigma_{0}^{2} \) and \( \sigma_{1}^{2} \), respectively; and unknown covariance, \( \sigma_{01} \).

Substituting the level-2 model into the level-1 model leads to the so-called mixed-effects form of the model (see Figure 2).

\[
\ln(\text{PHQ-9}_{i,t}) \sim \left( \gamma_{0,0} + \gamma_{0,1} x_{tip_{0,1}} + \cdots + \gamma_{0,m} x_{tip_{0,m}} + \pi_{tvp_{1}} x_{tvp_{1},t} + \cdots + \pi_{tvp_{1}} x_{tvp_{1},t} \right) \\
+ \left( \gamma_{1,0} + \gamma_{1,1} x_{tip_{1,1}} + \cdots + \gamma_{1,n} x_{tip_{1,n}} \right) t + \left( \zeta_{0,i} + \zeta_{1,i} t \right)
\]

(5)

Figure 2 Mixed-effects form of the generalized multilevel regression model

\[
\text{PHQ-9}_{i,t_0}^{\text{pop}} = \exp(\hat{\gamma}_{0,0} + \hat{\gamma}_{0,1} x_{tip_{0,1}} + \cdots + \hat{\gamma}_{0,m} x_{tip_{0,m}} + \hat{\pi}_{tvp_{1}} x_{tvp_{1},t_0} + \cdots + \hat{\pi}_{tvp_{1}} x_{tvp_{1},t_0}) \\
+ \left( \hat{\gamma}_{1,0} + \hat{\gamma}_{1,1} x_{tip_{1,1}} + \cdots + \hat{\gamma}_{1,n} x_{tip_{1,n}} \right) t_0 + \left( \hat{\zeta}_{0,i} + \hat{\zeta}_{1,i} t_0 \right)
\]

(6)

Figure 3 Population-average prediction of the PHQ-9 score

\[
\text{PHQ-9}_{i,t_0}^{\text{sub}} = \exp(\hat{\gamma}_{0,0} + \hat{\gamma}_{0,1} x_{tip_{0,1}} + \cdots + \hat{\gamma}_{0,m} x_{tip_{0,m}} + \hat{\pi}_{tvp_{1}} x_{tvp_{1},t_0} + \cdots + \hat{\pi}_{tvp_{1}} x_{tvp_{1},t_0}) \\
+ \left( \hat{\gamma}_{1,0} + \hat{\gamma}_{1,1} x_{tip_{1,1}} + \cdots + \hat{\gamma}_{1,n} x_{tip_{1,n}} \right) t_0 + \left( \hat{\zeta}_{0,i} + \hat{\zeta}_{1,i} t_0 \right)
\]

(7)

Figure 4 Subject-specific prediction of the PHQ-9 score

3.4 Using the Model for Prediction

We can use the 2-level Poisson regression model described above to make two types of prediction: the population-average prediction, which is based only on fixed effects, and the subject-specific prediction, which is based on both fixed and random effects [24]. Specifically, the population-average prediction of the PHQ-9 score for patient \( i \) at time \( t \geq 0 \), \( \text{PHQ-9}_{i,t}^{\text{pop}} \), can be seen in Figure 3 and the subject-specific prediction of the PHQ-9 score for patient \( i \) at time \( t_0 \geq 0 \), \( \text{PHQ-9}_{i,t_0}^{\text{sub}} \), can be seen in Figure 4.

3.5 Prediction Model Development

Many factors have been shown to be correlated with depression [8]. We selected 29 factors related to diabetes care from the DCAT dataset as the candidate predictors to develop the prediction model: 20 time-varying factors and nine time-invariant factors, measuring demographics, diabetes, health conditions, healthcare utilizations, and intervention programs in which the patients enrolled (patients from SC and TC arms of DCAT enrolled in the collaborative care management program for 6 months and patients from the TC arm received automated telephonic depression screening and monitoring for 12 months). Summary of time-invariant and time-varying predictors are provided in Online Table 1 and 2, respectively.

The dataset was divided into two parts: 80% randomly selected patients (# of patients = 682, # of samples = 2728) in the training set and the rest (# of patients = 171, # of samples = 684) in the validation set. Predictor selection was based on the training set and involved two phases. First, we estimated fixed effect of each candidate predictor in a univariate manner and obtained their p-values. We used univariate versions of Equation 5, that is, Equations 8–10 as seen in Figure 5, to conduct this analysis. Specifically, Equations 8 and 9 were used for each time-invariant candidate predictor to estimate the univariate fixed effect on model intercept (i.e., \( \gamma_{0,1} \) in Equation 8) and on model coefficient of time variable (i.e., \( \gamma_{1,1} \) in Equation 9), respectively. Equation 10 was used to estimate the univariate fixed effect of time variable.
Effect of each time-varying candidate predictor (i.e., \( \pi_{ivp} \) in Equation 10). Univariate analysis results of time-invariant and time-varying candidate predictors are provided in Online Tables 3 and 4, respectively. Then, we included all factors at significance level 0.05 in the univariate analysis and backward eliminated them using Equation 5 until all predictors were significant at \( p < 0.05 \). Selected predictors were used to develop the final prediction model.

We validated predictive ability of the prediction model in a longitudinal context in which newly generated patient records are incorporated with historical records to refit and update the model. First, we fitted a model based on the training set and used the model to predict PHQ-9 scores at study baseline (\( t = 0 \)) for patients in the validation set. Then, we combined the baseline records from the validation set with the training set to refit and update the prediction model and used the updated model to make predictions for PHQ-9 scores at 6-month follow-up in the validation set. We repeated these updating procedures another two times, using the 6- and 12-month follow-up records to update the model; and we made predictions for PHQ-9 scores at 12- and 18-month follow-ups, respectively. As discussed above, only population-average prediction can be made for PHQ-9 scores at study baseline (\( t = 0 \)), and both population-average and subject-specific predictions can be made for PHQ-9 scores at 6-, 12-, and 18-month follow-ups.

We used root-mean-square error (RMSE) to evaluate the accuracy of predicting severity of depression (i.e., the accuracy of the predicted PHQ-9 score). Lower values of RMSE indicate better prediction. We evaluated the classification ability of using predicted PHQ-9 scores to classify patients as having major depression by comparing the area under receiver operating characteristic (ROC) curves for predicted PHQ-9 scores against the real PHQ-9 scores \( \geq 10 \). Statistical analyses in the study were carried out by R, version 3.1.3 [25]. Fixed and random effects for a generalized multilevel model were estimated using quasi-likelihood estimation implemented in R package “glmPQL” function in R package “MASS” [26]. Quasi-likelihood estimation is an approximation method to fit a multilevel model because the exact evaluation of likelihood function is often computationally infeasible since it involves the integral over the random-effect space [27, 28]. P-values of fixed effect coefficients in a model were computed by conditional t-test, which is a variant of Student's t-test to test the marginal significance of each coefficient (29). 95% confidence intervals (CI) of fixed effects were obtained using a normal approximation to the distribution of parameter estimators, implemented by the “intervals” function in R package “nlme” [29].

### 4. Results

The resulting prediction model is a 2-level Poisson regression using two time-invariant predictors and 10 time-varying predictors. Specifically, we selected one of the nine time-invariant predictors, previous diagnosis of major depressive disorder before baseline, to predict the level-1 model intercept. We selected one time-invariant predictor, age at study baseline, to predict the level-1 model coefficient of time variable. We selected ten of the 20 time-varying predictors: 1) diabetes emotional burden, 2) diabetes regimen distress, 3) number of International Classification of Diseases, 9th Revision (ICD-9) diagnoses in past six months \( \geq 10 \), 4) self-rated health \( (1 = \text{poor to } 5 = \text{excellent}) \geq 3 \), 5) unemployed, 6) feeling that my financial situation is getting worse, 7) having difficulty in paying bills, 8) hospitalized overnight, 9) chronic pain, and 10) enrolled in the collaborative care management program.

Table 1 shows the RMSE of population-average and subject-specific predictions for PHQ-9 scores to classify patients as having major depression in the validation set.
tions for PHQ-9 scores of patients in the validation set. Predictive accuracy was improved by incorporating historical records and using them to update the model, as the RMSEs of both population-average and subject-specific predictions were smaller at the three follow-ups than the population-average prediction at study baseline. In addition, subject-specific predictions were shown to be better than population-average predictions for the three follow-ups.

Table 2 shows the areas under ROC curve when predicted PHQ-9 scores are used to classify patients as having major depression. Consistent with the results in Table 1, classification ability was improved by incorporating historical records and using them to update the model, as the areas under ROC of both population-average and subject-specific predictions were larger at the three follow-ups than the population-average prediction at study baseline. Comparisons between the two types of predictions for the three follow-ups show that subject-specific prediction has better classification ability. ROC curves of the population-average prediction for study baseline and subject-specific predictions for the three follow-ups are shown in Figure 6.

Table 3 shows the fixed effects estimated from the whole dataset, which can be used to derive population-average predictions for individuals out of the DCAT dataset using Equation 6. The exponential of the estimated effect of a predictor is the multiplicative term used to calculate the predicted PHQ-9 score when the predictor is increased by one unit. The way to make subject-specific predictions for individuals out of the DCAT dataset is discussed below.

5. Discussion

This study developed a 2-level Poisson regression model from the DCAT clinical trial to predict PHQ-9 scores for patients with diabetes using two time-invariant and 10 time-varying predictors related to demographics, diabetes, health conditions, and healthcare utilizations. The predicted PHQ-9 scores can be used for assessing depression severity and classifying patients as having major depression.

Estimated fixed effects (Table 3) can be used to make population-average predictions for individuals out of the DCAT dataset. To make subject-specific predictions for individuals out of the DCAT dataset, users would need a group of patients with multiple historical records that contain PHQ-9 scores and the 12 predictors to train an initial model. With those data, the model could then be used to make subject-specific predictions for patients in the group. Similar to the validation process described above, newly generated records could be incorporated to update the model.

An important implication of the study results is that incorporating historical records and using them to update the model may improve both the accuracy of predicting PHQ-9 scores and the ability to classify patients with major depression. The subject-specific predictions can achieve very good to excellent classification of major depression, as demonstrated by the 0.88–0.91 areas under ROC curve. These values are significantly better than the 0.72–0.80 areas under ROC curve achieved in prior depression prediction studies [12–15].

Besides multilevel regression, another applicable longitudinal modeling method is generalized estimating equation (GEE). The main difference between using multilevel regression and using GEE for prediction is that GEE can only derive population-average prediction [24]. Application of GEE with the same 12 predictors to the dataset used in this study leads to better predictions than the population-average predictions of generalized multilevel regression, but the predictive ability of GEE for the three follow-ups is worse than the subject-specific predictions of generalized multilevel regression (GEE: RMSE for baseline = 5.26, for 6-month = 4.46, for 12-month = 4.27, for 18-month = 4.54; area under ROC curve for baseline = 0.73, for 6-month = 0.87, for 12-month = 0.84, for 18-month = 0.86). More discussions about GEE and its comparison to multilevel regression can be found elsewhere in the literature [24, 30, 31].
The study has several limitations and opportunities for future research. First, although our results suggest that incorporating historical records can improve predictive ability, the finding is not conclusive because the DCAT dataset includes only four waves of data in 6-month intervals. Future research is indicated to further investigate the influences of using historical records for prediction. Second, some predictors used in the model, such as diabetes emotional burden and diabetes regimen distress [32], are not typically available in current medical practices, which may limit the applicability of the model. Third, the number of candidate predictors in the study is limited, which calls for future research to investigate a broader range of predictors. Fourth, predictor selection in the study is based on p-values obtained from conditional t-tests, which are sometimes not reliable [27, 33]. Future research is needed to address this issue. Finally, study patients are predominantly Hispanics from safety-net clinics in Los Angeles area, which may limit the generalizability of the model and calls for future research on a broader population.

6. Conclusions

The study developed a generalized multilevel regression model to predict depression severity and presence of major depression for patients with diabetes. The study demonstrated that generalized multilevel regression can be used to achieve high predictive accuracy based on longitudinal patient records.

Acknowledgments

This study is sponsored by the US Department of Health and Human Services (1R18AE000054–01) and the Autism Intervention Research Network for Behavior Health.

References