Modernizing Healthcare

How Electronic Health Records Predict Health Outcomes

Abstract: In this paper, we measure how technological innovation in the healthcare sector predicts intensity and quality of care in 2015. Utilizing individual-level health survey data from the Behavioral Risk Factor Surveillance System and statewide electronic health record (EHR) implementation rates from the National Electronic Health Record Survey (NEHRS), we find that some measures of EHR implementation modestly predict better self-described health, while some predict worse health. To come to this conclusion, we implemented a stepwise subset selection procedure on our trimmed dataset. We chose the model that had the best BIC criterion due to its performance on a number of other test metrics. Our final model has 46 predictor terms, and no influential outliers were detected. In measuring how electronic health record use statewide predicts health care efficacy, we aim to contribute to the ongoing debate about the use and implementation of EHR.

1. Background and Significance

Healthcare, especially in the time of COVID-19, has become an extremely relevant and pressing issue; however, there continues to be a lack of attention and discussion on the ways that technology is able to improve the inefficiencies in our healthcare system. Our paper contributes to this subject by studying the usage of electronic health records (EHRs). EHRs are collections of patient and population health information stored in a digital format. Potential advantages of using EHRs include increased interoperability between health care providers (HCPs) and improved accuracy of pharmacy script referrals. While some HCPs in the US have transitioned to EHR systems, many still use primarily paper-based systems to manage their patients' care.

In our project, we are interested in how EHR usage in the healthcare sector predicts health outcomes. More specifically, our research question is: How does electronic health record adoption on a state level predict self-reported health? Based on previous research, we expect that higher EHR uptake rates would predict individuals having a higher probability of being in good health.

Figure 1

Name	Description	Type
Demographic Information		
NUMADULT	Number of adults	numeric
PVTRESD1	Private residence (yes/no)	binary
SEX	Sex (male/female)	binary
MARITAL	Marital status	categorical
EDUCA	Education	ordinal
RENTHOM1	Own or rent home	categorical
VETERAN3	Veteran status (yes/no)	binary
EMPLOY1	Employment status	categorical
CHILDREN	Number of children	numeric
INCOME2	Income level	ordinal
X_BMI5	Body Mass Index	numeric
PREGNANT	Pregnancy status (yes/no)	binary
SCNTWORK1	Hours of work per week	numeric
X_HISPANC	Hispanic origin (yes/no)	binary
HLTHPLN1	Is Insured (yes/no)	binary
INTERNET	Uses Internet (yes/no)	binary
EXERANY2	Exercises (yes/no)	binary
X_SMOKER3	Smoking status	categorical
EHR Usage		
pct_phys_any_ehr	Percent Physicians (PP) Use Any EHR	numeric
pct_phys_basic_ehr	PP Use Basic EHR	numeric
pct_phys_cert_ehr	PP Use Certified EHR	numeric
pct_primary_care_phys_cert_ehr	Percent Primary Care Physicians Use Certified EHR	numeric
pct_surg_med_spec_phys_cert_ehr	Percent Medical Specialists Use Certified EHR	numeric
pct_small_practice_phys_cert_ehr	Percent Small Practice Physicians Use Certified EHR	numeric
pct_phys_send_receive_any_patient_info	PP Electronically Send or Receive Any Patient Info	numeric
pct_phys_e_share_patients	PP Electronically Share Patient Info	numeric
pct_phys_patient_secure_message	PP Electronically Send Secure Messages to Patients	numeric

2. Data

2.1 Data Description

Our data came from two surveys: the Behavioral Risk Factor Surveillance System (BRFSS) and the National Electronic Health Records Survey (NEHRS).

The BRFSS is an annual nationwide landline and cellphone survey conducted by the CDC. It observes population level health trends in the United States. We implemented our model using data from the 2015 annual survey which has 441,456 observations.

The NEHRS, conducted by the CDC in conjunction with the Office of the National Coordinator for Health Information Technology (ONC), is an annual survey of non-federal office-based physicians practicing in the United States. This dataset reports yearly average statistics related to electronic health record implementation at the state-level (50 observations).

2.2 Data Cleaning

We first selected demographic and health outcome variables from the BRFSS that could be relevant to our research question (figure 1).

Since we only wanted to consider completed survey observations recorded in 2015, we removed 10,915 observations from the BRFSS dataset that were taken in 2016 and 138,817 observations whose survey was terminated early. We then restricted our analysis to a subset of the BRFSS and all of the NEHRS variables. We also filtered out observations from Guam and Puerto Rico, as the NEHRS does not have observations for those places. After filtering, we had 274,896 observations.

We identified 17 variables with missingness (bolded in fig. 1). For predictor variables missing less than 15% of their observations, we imputed them with

pct_phys_vdt	PP Let Patients View, Download, or Transmit Info	numeric
pct_phys_vd_and_t	PP Let Patients View, Download, and Transmit Info	numeric
pct_phys_find_clin_info	PP Can Search External Patient Health Info	numeric
pct_phys_send_any_clin_info	PP Can Send Patient Health Info	numeric
pct_phys_send_summary_care_card	PP Can Send Summary Care Records	numeric
pct_phys_receive_any_clin_info	PP Receive External Patient Clinical Info	numeric

multivariate imputation by chained equations. The variable imputation specifications are as follows: normal imputation ignoring model error for numeric variables, logistic regression for binary variables, proportional odds logistic regression for categorical variables.

3. Methods and Results

3.1 Multicollinearity

We examined multicollinearity of numeric variables using variance inflation factor (VIF) as a criterion. We set our VIF threshold for high multicollinearity at 10, and eliminated pct_phys_cert_ehr, pct_phys_e_share_patients, pct_phys_send_receive_any_patient_info from the dataset.

3.2 Model Fitting and Evaluation

We used a stepwise procedure to find the best first-order logistic regression model considering both the AIC and BIC criteria. After converting our predicted probabilities into binary outcomes using a decision threshold of 0.5, we compared results using the likelihood ratio test and also computed accuracy, sensitivity, specificity, precision, deviance, and AUC using 5-fold cross-validation (figure 2).

Figure 2

	Subset Selection Criteria	
	AIC	BIC
# Preds	55	**46**
AIC	**159724**	159782
BIC	160311	**160267**
Accuracy	0.8488	**0.8491**
Sensitivity	0.9665	**0.9667**
Specificity	0.298	**0.2987**
Precision	0.8657	**0.8658**
AUC	**0.8147**	0.8144
Deviance	**199543**	199622

In our evaluation, we prioritized parsimony, since we would like to have an understandable model, as well as having high accuracy, sensitivity, and AUC. We are particularly interested in models that have high accuracy, as we would like to correctly predict someone being in good health.

Based on our evaluation criteria, we chose the BIC logit model, since it had fewer predictors and slightly higher accuracy and sensitivity, even though the AIC model had a lower AUC score. Additionally, the difference in AUC, accuracy, and sensitivity were all very close between the AIC and BIC models, so with the BIC model being more parsimonious, we chose it.

3.4 Model Diagnostics

To examine if there were any influential outliers, we measured the delta deviance of each of the observations within our dataset. We did not observe any distinct peaks in a graph against the observation row number, and so did not eliminate any observation for being influential.

4. Conclusion and Other Considerations

Our final model includes the predictor variables shown in Figure 3 for a total of 46 predictors. Note that almost all predictors, barring a few levels of categorical variables, are significant at α = 0.05 level, while most predictors have p-values that are virtually zero.

4.1 Discussion

Our final model included a diverse group of predictor variables with both demographic and EHR-usage predictors left in the model. To address our research question, we surprisingly do not find a clear relationship between the variables for EHR uptake and our outcome variable. While five predictors have a positive relationship with the outcome, such as pct_primary_care_ phys_cert_ehr or pct_phys_ receive_any_clin_info, the remaining three have a negative relationship with it. At first we suggested this was due to multicollinearity, but further investigation into the correlation between these predictors alone and the outcome variable revealed the same sign as their coefficients in the logit model. Another interesting observation about our final model is that almost all the EHR-usage predictors relate to specific specialties of types of EHRS (such as pct_surg_med _spec_phys_cert_ehr which is about medical or surgical specialists), while most EHR-usage variables that cover broader categories (such as pct_phys_

any_ehr) do not remain in our model. For further investigation, we also repeated the process above with the same initial predictors for another outcome variable we created that measured the number of self-reported annual checkups, where 1 denoted individuals who received annual checkups equal to or more than the doctor-recommended threshold, and 0 denoted otherwise. We again chose the BIC logit model as it was more parsimonious (44 vs. 52 predictors) and had similar accuracy (0.775 for both), sensitivity (0.978 for both) and AUC (0.6664 for BIC vs. 0.6665 for AIC). Compared to our BIC model for general health, the final model for checkups included more EHR-related predictors, notably many that described broader categories (e.g. pct_phys_any_ehr and pct_phys_basic_ehr), as well as those related to sending clinic information or secure messages to patients (e.g. pct_phys_vdt and pct_phys_send_any_clin_info). This made sense, given that being able to send information means physicians can send digital checkup reminders.

Figure 3

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.361e+00	1.513e-01	8.996	< 2e-16	***
NUMADULT	-4.307e-02	6.269e-03	-6.870	6.42e-12	***
MARITAL2	-3.865e-02	1.827e-02	-2.115	0.03441	*
MARITAL3	5.177e-02	1.953e-02	2.651	0.00803	**
MARITAL4	-1.871e-01	3.522e-02	-5.311	1.09e-07	***
MARITAL5	2.484e-01	1.957e-02	12.689	< 2e-16	***
MARITAL6	5.354e-02	3.385e-02	1.582	0.11375	
EDUCA2	1.528e-01	1.176e-01	1.299	0.19407	
EDUCA3	4.691e-01	1.164e-01	4.030	5.58e-05	***
EDUCA4	7.195e-01	1.154e-01	6.234	4.55e-10	***
EDUCA5	7.527e-01	1.157e-01	6.505	7.77e-11	***
EDUCA6	9.312e-01	1.161e-01	8.022	1.04e-15	***
RENTHOM12	-1.423e-01	1.476e-02	-9.644	< 2e-16	***
RENTHOM13	-2.199e-01	2.783e-02	-7.901	2.77e-15	***
VETERAN31	-1.820e-01	1.678e-02	-10.844	< 2e-16	***
EMPLOY12	1.288e-02	2.569e-02	0.501	0.61606	
EMPLOY13	-4.095e-01	5.953e-02	-6.880	6.01e-12	***
EMPLOY14	-1.261e-01	6.140e-02	-2.054	0.03994	*
EMPLOY15	-1.584e-01	5.677e-02	-2.791	0.00526	**
EMPLOY16	5.252e-01	6.692e-02	7.849	4.19e-15	***
EMPLOY17	-3.747e-01	5.295e-02	-7.076	1.48e-12	***
EMPLOY18	-1.775e+00	5.411e-02	-32.798	< 2e-16	***
CHILDREN	9.731e-02	6.544e-03	14.871	< 2e-16	***
INCOME22	1.285e-02	2.805e-02	0.458	0.64680	
INCOME23	1.268e-01	2.671e-02	4.747	2.06e-06	***
INCOME24	2.234e-01	2.661e-02	8.396	< 2e-16	***
INCOME25	3.415e-01	2.689e-02	12.701	< 2e-16	***
INCOME26	5.791e-01	2.749e-02	21.065	< 2e-16	***
INCOME27	7.311e-01	2.896e-02	25.246	< 2e-16	***
INCOME28	1.038e+00	2.967e-02	34.992	< 2e-16	***
X_BMI5	-3.851e-04	7.604e-06	-50.643	< 2e-16	***
PREGNANT2	-4.295e-01	2.299e-02	-18.677	< 2e-16	***
SCNTWRK1	8.643e-03	1.208e-03	7.156	8.31e-13	***
X_HISPANC2	3.247e-01	2.156e-02	15.061	< 2e-16	***
INTERNET2	-3.667e-01	1.461e-02	-25.099	< 2e-16	***
EXERANY22	-7.311e-01	1.206e-02	-60.624	< 2e-16	***
X_SMOKER32	2.812e-02	2.806e-02	1.002	0.31626	
X_SMOKER33	1.434e-01	1.875e-02	7.651	1.99e-14	***
X_SMOKER34	4.666e-01	1.786e-02	26.122	< 2e-16	***
pct_primary_care_phys_cert_ehr	4.554e-01	9.083e-02	5.013	5.35e-07	***
pct_surg_med_spec_phys_cert_ehr	8.705e-01	1.133e-01	7.682	1.56e-14	***
pct_small_practice_phys_cert_ehr	-7.099e-01	1.233e-01	-5.757	8.58e-09	***
pct_phys_send_summary_care_record	1.104e+00	1.659e-01	6.653	2.87e-11	***
pct_phys_receive_any_clin_info	6.075e-01	8.111e-02	7.490	6.86e-14	***
pct_phys_receive_summary_care_record	-1.990e+00	1.790e-01	-11.116	< 2e-16	***
pct_phys_integrate_any_clin_info	-7.355e-01	1.225e-01	-6.002	1.95e-09	***
pct_phys_integrate_summary_care_record	1.254e+00	1.829e-01	6.854	7.18e-12	***

4.2 Further Considerations

One area for further exploration would be to implement a regression model that allowed for interaction terms. Given the amount of variables in our model and dataset, selecting interaction terms with subset selection is a computationally intense problem. The amount of variables we considered coupled with the limited computational power lead us to not try to implement this model for this paper.

An extension of our work could implement an ordered logistic regression using the original categorical variable for self-described health as an outcome. However, since this method is outside of the scope of our class, we did not implement it here.

In the BRFSS dataset, we acknowledge that there are alternate ways to interpret a respondent refusing to answer a question. In order to have the most complete data that we could, we imputed values for both individuals who refused to answer a question, and those who were not asked that question to begin with.

Lastly, our data presents an ecological inference problem: mainly, the data for the electronic health record uptake is at the state level while our health outcome data is at the individual level. Since we are using population means to estimate individual outcomes, we could be committing an ecological fallacy. However, we do not have access to more granular data, nor the background in ecological inference to address this problem here.

Appendix

Predictor Name	VIF (**when removed**)
NUMADULT	1.047103
CHILDREN	1.062768
X_BMI5	1.023213
SCNTWRK1	1.046985
pct_phys_any_ehr	7.119985
pct_phys_basic_ehr	3.242681
pct_primary_care_phys_cert_ehr	3.900294
pct_phys_cert_ehr	**133.13889**
pct_surg_med_spec_phys_cert_ehr	6.475013
pct_small_practice_phys_cert_ehr	4.49054
pct_phys_send_receive_any_patient_info	**30.71440**
pct_phys_e_share_patients	**21.17662**
pct_phys_patient_secure_message	7.407546
pct_phys_vdt	6.405377
pct_phys_vd_and_t	2.954226
pct_phys_find_clin_info	1.906412
pct_phys_send_any_clin_info	3.675686
pct_phys_send_summary_care_record	6.42179
pct_phys_receive_any_clin_info	4.362476
pct_phys_receive_summary_care_record	7.575905
pct_phys_integrate_any_clin_info	2.40616
pct_phys_integrate_summary_care_record	3.979136

Figure A.1
Multicollinearity Analysis: In bold: variance inflation factor of variables removed at the time they were removed. Otherwise, variance inflation factor of variables after removing variables with VIF scores greater than or equal to 10.

AIC and BIC Stepwise Procedure:

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
	NA	NA	274833	199538.3	199664.3
- pct_phys_find_clin_info	1	0.0287335	274834	199538.3	199662.3
- HLTHPLN1	1	0.0387570	274835	199538.4	199660.4
- pct_phys_send_any_clin_info	1	0.7456649	274836	199539.1	199659.1
- pct_phys_basic_ehr	1	0.7753187	274837	199539.9	199657.9
- pct_phys_vd_and_t	1	0.8718572	274838	199540.8	199656.8
- pct_phys_vdt	1	0.4775031	274839	199541.2	199655.2
- PVTRESD1	1	1.9830296	274840	199543.2	199655.2

Figure A.2
Variables removed during stepwise procedure using AIC as the criterion.

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
	NA	NA	274833	199538.3	200327.3
- X_PRACE1	6	54.5172903	274839	199592.8	200306.7
- HLTHPLN1	1	0.0915472	274840	199592.9	200294.3
- pct_phys_find_clin_info	1	0.1188048	274841	199593.0	200281.9
- pct_phys_vdt	1	0.4071325	274842	199593.4	200269.7
- pct_phys_vd_and_t	1	0.3185279	274843	199593.8	200257.5
- pct_phys_basic_ehr	1	0.4734505	274844	199594.2	200245.5
- pct_phys_send_any_clin_info	1	1.2165272	274845	199595.5	200234.2
- PVTRESD1	1	1.9760975	274846	199597.4	200223.6
- pct_phys_any_ehr	1	6.2606910	274847	199603.7	200217.4
- SEX	1	8.1195506	274848	199611.8	200213.0
- pct_phys_patient_secure_message	1	12.0584169	274849	199623.9	200212.5

Figure A.3
Variables removed during stepwise procedure using BIC as the criterion.

Predictor Name	AIC Criterion	BIC Criterion
Intercept	1.4752***	1.361***
NUMADULT	-0.0399***	-0.0431***
SEX	-0.0381**	-
MARITAL2	-0.0403*	-0.0386*
MARITAL3	0.0447*	0.0518**
MARITAL4	-0.1823***	-0.1871***
MARITAL5	0.2581***	0.2484***
MARITAL6	0.0563	0.0535
EDUCA2	0.1482	0.1528
EDUCA3	0.4654***	0.4691***
EDUCA4	0.7103***	0.7195***
EDUCA5	0.7409***	0.7527***
EDUCA6	0.9193***	0.9312***
RENTHOM2	-0.1371***	-0.1423***
RENTHOM3	-0.2158***	-0.2199***
VETERAN3	-0.1594***	-0.182***
EMPLOY2	0.0173	0.0129
EMPLOY3	-0.3943***	-0.4095***
EMPLOY4	-0.1063	-0.1261*
EMPLOY5	-0.1626**	-0.1584**
EMPLOY6	0.5393***	0.5252***
EMPLOY7	-0.3649***	-0.3747***
EMPLOY8	-1.7607***	-1.7748***
CHILDREN	0.0992***	0.0973***
INCOME2	0.0084	0.0129
INCOME3	0.1237***	0.1268***
INCOME4	0.2181***	0.2234***
INCOME5	0.3365***	0.3415***

Predictor Name	AIC Criterion	BIC Criterion
INCOME6	0.5734***	0.5791***
INCOME7	0.7247***	0.7311***
INCOME8	1.0326***	1.0384***
X_BMI5	-4e-04***	-4e-04***
PREGNANT	-0.4286***	-0.4295***
SCNTWRK1	0.009***	0.0086***
X_PRACE2	-0.0392*	-
X_PRACE3	-0.2356***	-
X_PRACE4	-0.0495	-
X_PRACE5	-0.2562***	-
X_PRACE6	-0.0856*	-
X_PRACE7	-0.0786	-
X_HISPANC	0.3051***	0.3247***
INTERNET	-0.3594***	-0.3667***
EXERANY2	-0.7316***	-0.7311***
X_SMOKER2	0.0339	0.0281
X_SMOKER3	0.1431***	0.1434***
X_SMOKER4	0.465***	0.4666***
pct_phys_any_ehr	-0.4472*	-
pct_primary_care_phys_cert_ehr	0.51***	0.4554***
pct_surg_med_spec_phys_cert_ehr	0.9246***	0.8705***
pct_small_practice_phys_cert_ehr	-0.6688***	-0.7099***
pct_phys_patient_secure_message	0.3843***	-
pct_phys_send_summary_care_record	0.9802***	1.104***
pct_phys_receive_any_clin_info	0.4059***	0.6075***
pct_phys_receive_summary_care_record	-1.9338***	-1.99***
pct_phys_integrate_any_clin_info	-0.6705***	-0.7355***
pct_phys_integrate_summary_care_record	1.0665***	1.254***

Figure A.4
Model coefficients for the models found through AIC and BIC stepwise selection.

Confusion Matrices

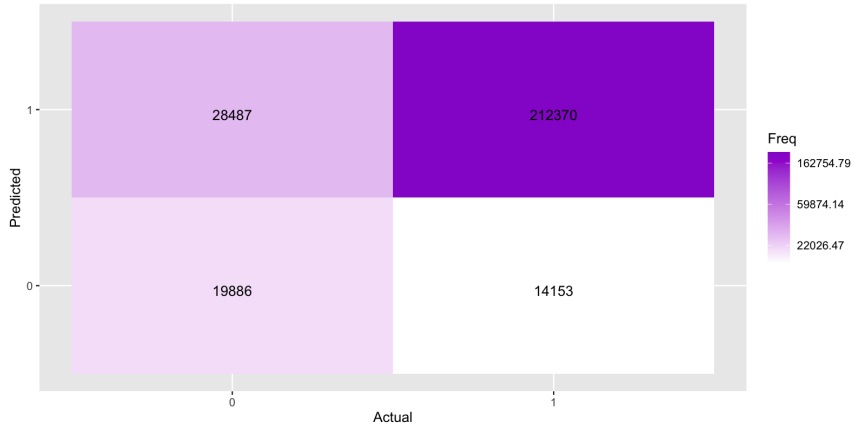

Figure A.5
AIC Stepwise Final Model Confusion Matrix

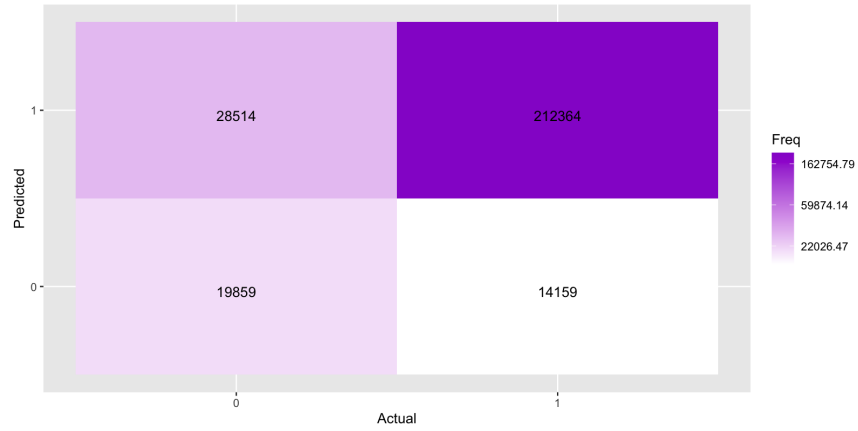

Figure A.6
BIC Stepwise Final Model Confusion Matrix

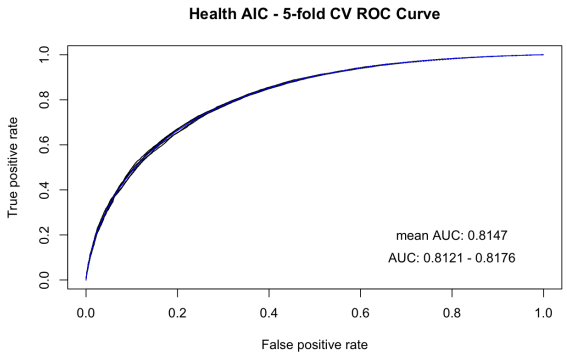

Figure A.7
ROC curves found using 5-fold cross validation, mean AUC, and range of AUC for the AIC stepwise final model.

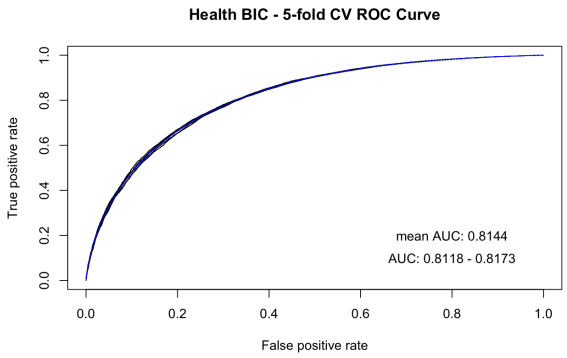

Figure A.8
ROC curves found using 5-fold cross validation, mean AUC, and range of AUC for the BIC stepwise final model.

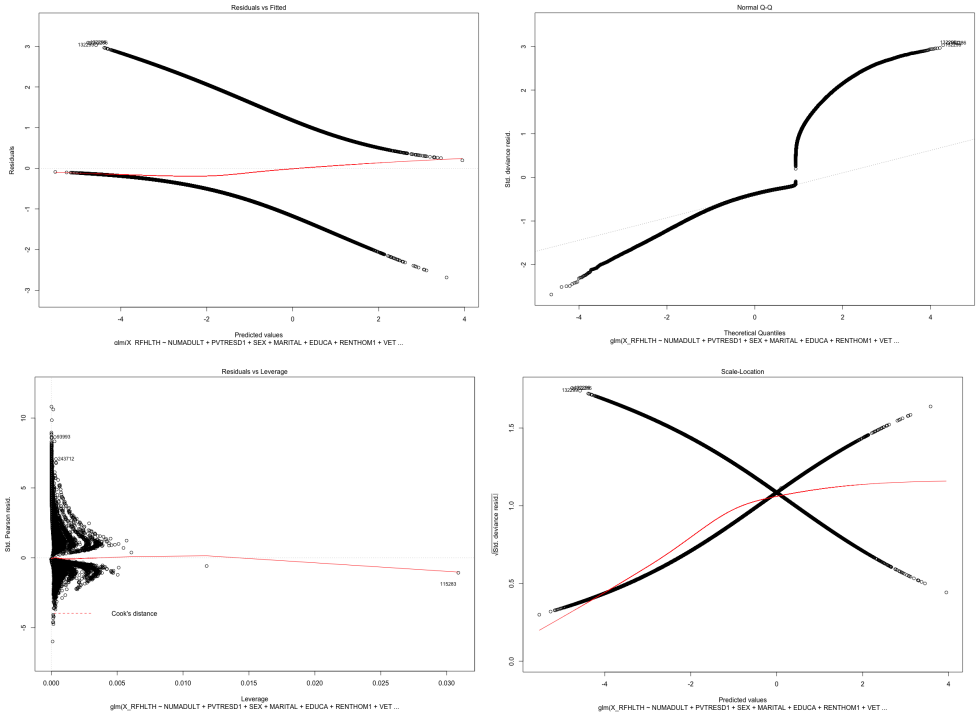

Figure A.9
Error plots for the BIC stepwise selection model.

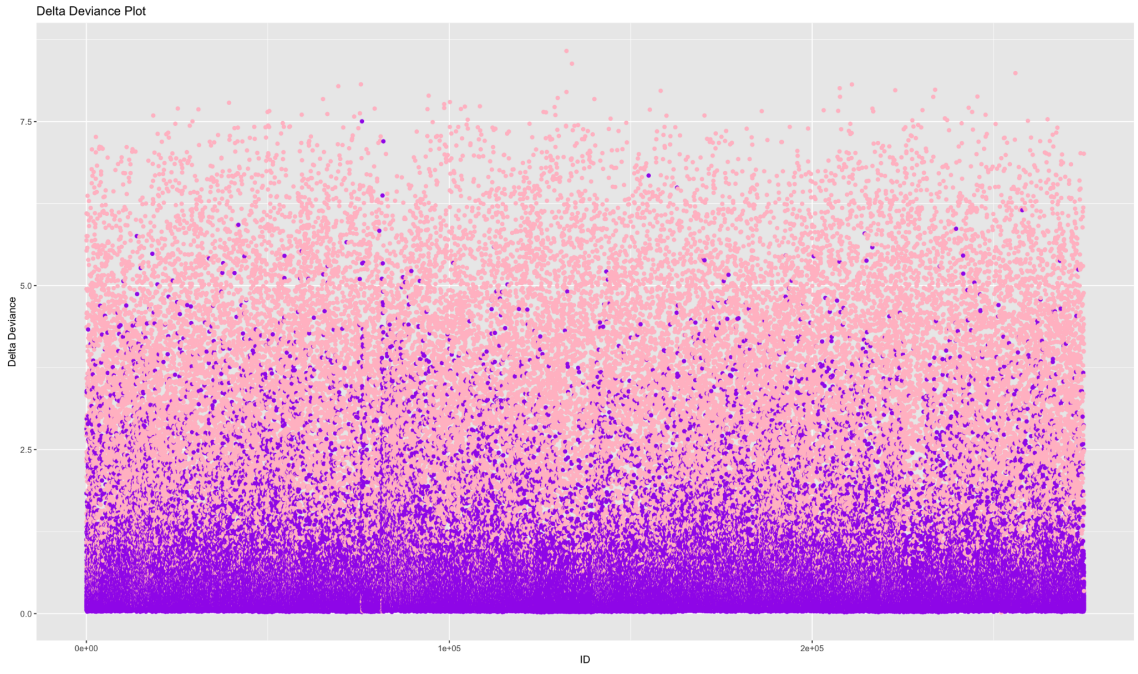

Figure A.10
Delta deviance plot for all observations under the BIC stepwise selection model. Pink observations have ground truth bad health while purple are in good health.

Predictor Name	AIC Criterion	BIC Criterion
Intercept	0.8651***	1.1321***
NUMADULT	–0.0141**	-
PVTRESD1	0.3747***	0.3867***
SEX	0.3929***	0.3923***
MARITAL2	-0.0721***	-0.0635***
MARITAL3	0.1743***	0.1861***
MARITAL4	-0.1288***	-0.1218***
MARITAL5	-0.0465***	-0.0433**
MARITAL6	-0.1828***	-0.1845***
EDUCA2	0.2273	-
EDUCA3	0.3170**	-
EDUCA4	0.3139**	-
EDUCA5	0.2773*	-
EDUCA6	0.2844*	-
VETERAN3	0.3906***	0.3909***
CHILDREN	-0.0678***	-0.0682***
INCOME2	0.0107	0.0114
INCOME3	0.0772**	0.0791**
INCOME4	0.1122***	0.1132***
INCOME5	0.1238***	0.1243***
INCOME6	0.202***	0.2012***
INCOME7	0.2606***	0.2574***
INCOME8	0.3235***	0.3175***
X_BMI5	2e-04***	2e-04***
PREGNANT	-0.1075***	-0.1109***
SCNTWRK1	-0.0105***	-0.0105***
X_PRACE2	0.7176***	0.721***

Predictor Name	AIC Criterion	BIC Criterion
X_PRACE3	0.0545	0.0529
X_PRACE4	0.2385***	0.229***
X_PRACE5	0.256***	0.2466***
X_PRACE6	0.1152***	0.1096**
X_PRACE7	-0.0209	-0.0221
X_HISPANC	-0.1741***	-0.1616***
HLTHPLN1	-1.1143***	-1.1155***
INTERNET	0.0629***	0.0654***
EXERANY2	-0.1579***	-0.1567***
X_SMOKER2	0.2299***	0.2287***
X_SMOKER3	0.4325***	0.4307***
X_SMOKER4	0.3638***	0.3596***
pct_phys_any_ehr	1.0261***	1.0229***
pct_phys_basic_ehr	-1.4598***	-1.4461***
pct_primary_care_phys_cert_ehr	-0.7153***	-0.681***
pct_surg_med_spec_phys_cert_ehr	-0.853***	-0.6793***
pct_small_practice_phys_cert_ehr	0.2082*	-
pct_phys_patient_secure_message	-0.8546***	-0.904***
pct_phys_vdt	1.3463***	1.3933***
pct_phys_vd_and_t	-0.2291	-
pct_phys_find_clin_info	-0.8543***	-0.8563***
pct_phys_send_any_clin_info	-1.1642***	-1.1353***
pct_phys_receive_any_clin_info	2.6878***	2.6736***
pct_phys_receive_summary_care_record	-1.347***	-1.3254***
pct_phys_integrate_any_clin_info	-0.8584***	-0.9189***
pct_phys_integrate_summary_care_record	3.0208***	-3.1869***

Figure A.11
Model coefficients for models with checkup outcome variable found through AIC and BIC stepwise selection.

	# Preds	AIC	BIC	Accuracy	Sensitivity	Specificity	Precision	AUC	Deviance
AIC	52	**223151**	223697	0.775	0.978	0.102	0.783	**0.6665**	**278800**
BIC	**44**	223163	**223627**	0.775	0.978	0.102	0.783	0.6664	278900

Figure A.12
Evaluation of AIC and BIC Stepwise Selection Model with ‘has recommended annual checkups’ as the response variable.