Predicting the likelihood of not receiving a pap smear based on individual-level factors and access to healthcare

By: Anna Wuest, Anja Shahu, Ligia Flores

I. Abstract

Although cervical cancer mortality has declined with increased availability and use of the pap smear, there are still many disparities in pap smear screening that need to be addressed to improve cervical cancer outcomes in the United States (U.S.). Using the 2018 Full Year Consolidated Data File from the Medical Expenditure Panel Survey (MEPS) by the U.S. Department of Health and Human Services (HHS), our primary analysis aims to find what individual-level and access to healthcare factors predict the likelihood of not getting a pap smear in the last 5 years among American women aged 21-65. Our final prediction model had a sensitivity of 68.25%, specificity of 68.28%, positive predictive value (PPV) of 86.74%, negative predictive value (NPV) of 41.42 % and overall accuracy of 68.26%. Our results suggested that socio-demographic, health status, smoking, access to healthcare and medical expenditure variables were predictive of not getting a pap smear. In our secondary analysis, we investigated the association between not getting a pap smear and access to healthcare, and we found statistical evidence for this association. For example, among those who were uninsured, the odds of not getting a pap smear was estimated to be, on average, 2.01 times that among the people who had private insurance, adjusting for the other covariates in our final association model (95%: 1.63 to 2.49, p-value = 1.03e-10).

KEYWORDS: cervical cancer, pap test, disparities, women's health

II. Introduction and Review of Literature

Decades ago, cervical cancer was one of the most common causes of cancer death among American women.[5] Since then, use of the pap smear test as a diagnostic tool has increased, contributing to the decrease in cervical cancer mortality by helping to diagnose cases early, before progression to terminal stages.[5] However, cervical cancer disparities still exist. Overall, those from low socio-economic backgrounds,[9,11] from racial/ethnic minority groups, and from the LBGTQIA+ community[6] have a higher risk of developing cervical cancer.

Continued disparities in cervical cancer outcomes can be partially attributed to the fact that although pap smear uptake has increased, pap smears are still not widely and equally accessible; people are often limited by socio-demographic factors and access to care.[1] Akers et al. noted that factors, such as race/ethnicity, age, immigration status, health literacy, education, socio-economic position, geography, provider characteristics and health system deficiencies, were consistently reported as factors influencing pap smear uptake in numerous studies.[1] The American Cancer Society found that women without health insurance were less likely to get screened for cervical cancer.[12] Looking among uninsured women, Akinlotan et al. found barriers to testing, including language barriers, fear of finding cancer, having male physicians and lack of knowledge about cervical cancer, and they found that age, marital status, and previous screening status influenced the barriers that people identified.[2] Other studies found that risk factors for cervical cancer, such as not getting an HPV vaccination,[7] having multiple sexual partners, smoking and having HIV,[3] were associated with lower pap smear uptake.

Lastly, according to the CDC and the U.S. Preventive Services Task Force (USPSTF), there is an optimal age window where pap smears are most effective when done regularly. The CDC recommends that only women over 21 should receive pap smears.[4] The USPSTF gives a more conservative recommendation, saying that only women aged 21-65 should receive a pap

smear based on clinical trials and cohort studies.[13] For our analysis, we follow the USPSTF recommendation.

In this paper, our primary analysis aims to find what individual-level and access to healthcare factors predict the likelihood of not getting a pap smear in the last 5 years among American women aged 21-65. For our secondary analysis, we investigated the association between not getting a pap smear and access to healthcare.

III. Research and Analysis Methods

A. Data

We used the 2018 Full Year Consolidated Data File from the Medical Expenditure Panel Survey (MEPS), which was collected by the HHS.[10] MEPS is an extensive survey of U.S. citizens. We restricted our analysis to the 6636 women aged 21-65 to follow USPSTF recommendations.

B. Outcome and Predictor Variables

We first identified the predictor variables that were mentioned in our literature review that were also found in MEPS. Then we narrowed down the list of variables further to exclude any that had over 800 missing values, as these would have decreased our sample size too much. In our analysis, we ultimately considered these variables: race/ethnicity, age, marital status, education, self-reported general health status, region, smoking frequency, limitation in work/housework/school, ability to afford care, individual income, family income, total medical expenditures, out of pocket medical expenditures, having a usual source of care (USC) and insurance coverage. For more explanation on the variables, look at Table 1 in the Extra Tables & Figures section.

For our analysis, we used the binary variable of if someone had received a pap smear in the last five years (0 - pap smear; 1 - no pap smear) as our outcome variable. As a result, we used multivariable logistic regression for our models.

C. Missing Values

In our final data set, there were 815 missing values. Of those missing values, 595 were from our outcome variable and the remaining 220 were from our predictors. Much of the missingness, especially in the outcome variable, resulted from the random skip pattern methodology that MEPS uses to survey people, so it is likely that some of the variables are missing completely at random (MCAR). However, we recognize that as a whole we likely do not have MCAR, as there are certain variables that may have patterns to their missingness that are related to the covariates (e.g. 118 people refused to answer a question about whether they had a provider). For the purposes of our project, we will be assuming that these variables are missing at random (MAR) so that we can perform a complete case analysis. By removing our missing values, we removed 773 observations, decreasing our sample size from 6636 to 5863.

D. Primary vs. Secondary Analysis Approach

During the model building stage for both analyses, we accounted for potential confounding, effect modification and nonlinearity of our continuous variables. Since we did not have a primary predictor of interest in either of our analyses, there were many potential confounding relationships to consider; therefore, rather than doing many tests to identify statistical confounding, we used the subject matter knowledge sourced from our literature review to decide which potential confounders to adjust for in the MEPS dataset. Our research also suggested that the variables of marital status and education could be potential effect modifiers of the relationship between barriers to receiving a pap smear and failure of pap smear uptake, so we decided to focus on these two variables while assessing effect modification during model

building. For both analyses, we assessed model fit for every model that we built using the Hosmer-Lemeshow test for calibration and the Receiver Operating Characteristic (ROC) curve for discrimination; we threw out any models that had a p-value < 0.05 for the Hosmer-Lemeshow test, as this suggested poor goodness of fit. Although our initial exploration of the correlation between our different variables found low correlation between most of the variables and did not suggest we should expect major problems with multicollinearity (see Table 2 in the Extra Tables & Figures section), we still tested for potential multicollinearity using the variance inflation factor (VIF) during model building as well, focusing on making sure VIF values for variables were < 2.

Our primary and secondary analyses had different modeling aims, which meant we had to take a different approach for the final model selection of each. For the primary analysis, our goal was to create a prediction model to predict the likelihood of not getting a pap smear in the last 5 years among American women aged 21-65, so we split our data up into a 70% train set (4105 observations) and 30% test set (1758 observations). We built all our models using the train set and used cross validation to assess the fit of the model on the test set. We selected the prediction model that best maximized the area under the ROC (AUC) on the test set. For the secondary analysis, our goal was to create an associational model that explored the effect of access to healthcare on not getting a pap smear; as a result, we built our models on the entire 5863 observation data set, used backward/forward selection and hypothesis testing (all of which were done at the $\alpha = 0.05$ level) to drive our model building and selected the model that best balanced AIC and interpretability for the final model.

IV. Findings and Analysis

A. Primary Analysis

The aim of our primary analysis was to predict the likelihood of not getting a pap smear in the last 5 years among American women aged 21-65. We built numerous models on the train set and then compared them using AUC on the test set. We first built a full model that adjusted for all 15 predictor variables. Then we used backward and forward selection on the full model to create a pared down model; both forward and backward selection produced the same model. Since the full model had a higher AUC than the backward/forward selection model (AUC of 0.7214 and 0.7205, respectively), we decided to build off the full model rather than the backward/forward selection model as we assessed if adding interaction and nonlinear terms improved our prediction.

Inspired by the results of our secondary associational analysis, we first started by adding in nonlinear terms for age, trying models with quadratic age, a cubic age spline with 3 knots and a GAM with 4 df for age (AUC of 72.66%, 72.65% and 72.14%, respectively). Since it had the highest AUC, we built off the full model with quadratic age, and again inspired by the results of our secondary analysis, we added in an interaction term between marital status and family income, finding that AUC increased (AUC of 73.04%) from the full model with just quadratic age. We then added an interaction term between education and total medical expenditures and found that AUC increased (AUC of 73.17%) from the previous interaction model, so we kept it in the model. Then we tested for nonlinearity of the remaining four continuous predictors, trying out quadratic terms, cubic splines with 3 knots and GAMs with 4 df, keeping terms that improved the AUC and continuously building to make more complex models, as can be seen in Table 1. In the end, we had two models tied with the highest AUC of 74.09%, but we chose the simpler model as our final model because it was more parsimonious.

Prediction models	**AUC**
Full model	72.14%

Backward/forward selection model	72.05%
Full model + quad. age	72.66%
Full model + quad. age + marital status * family income	73.04%
Full model + quad. age + marital status * family income + education * total exp.	73.17%
Full model + quad. age + marital status * family income + education * total exp. + quad individual income	73.14%
Full model + quad. age + marital status * family income + education * total exp. + quad family income	73.14%
Full model + quad. age + marital status * family income + education * total exp. + total exp. cubic spline w/ 3 knots	**74.09%**
Full model + quad. age + marital status * family income + education * total exp. + total exp. cubic spline w/ 3 knots + out of pocket exp. quad.	74.09%

Table 1: Summary of some of the prediction models that were built. The bolded model denotes the final model that was chosen.

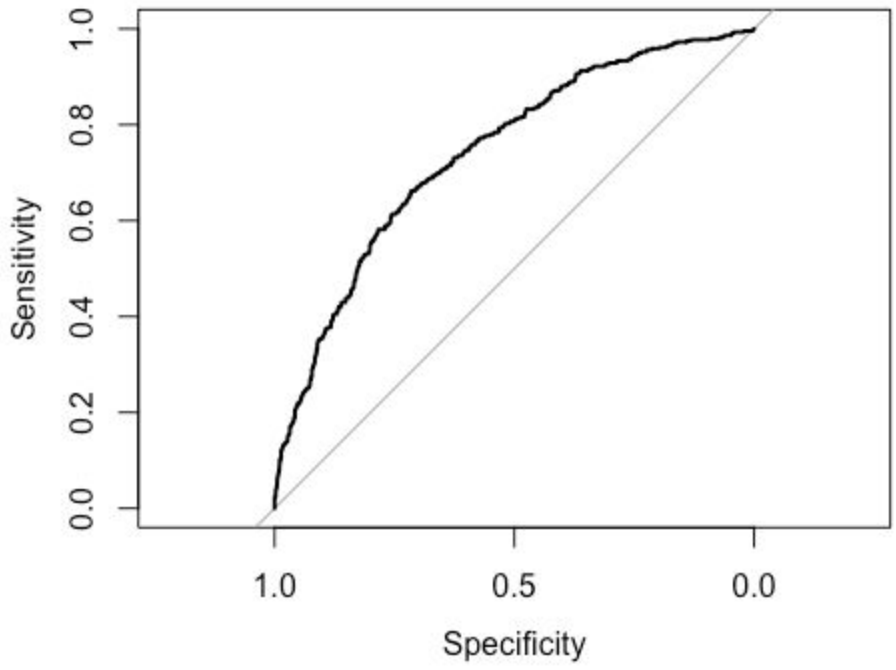

Figure 1: ROC of final prediction model with AUC of 74.09% on the test set

Our final model adjusted for all 15 of our variables as well as quadratic age, interaction terms between marital status and family income and between education and total medical expenditures, and a cubic spline with 3 knots for total expenditures. To see the summary output for the model, refer to Table 3 in the Extra Tables & Figures section.

We looked for the p-cutoff that balanced sensitivity and specificity in our final model best, and we found that occurred at p = 0.25. With a p-cutoff of 0.25, we have an overall accuracy of 68.26%, sensitivity of 68.25% , specificity of 68.28%, PPV of 86.74% and NPV of 41.42%.

	Yes pap smear (observed)	No pap smear (observed)
Yes pap smear (predicted)	903	138
No pap smear (predicted)	420	297

Table 2: 2x2 table comparing predicted vs observed pap smear values for p-cutoff of 0.25 based on final predictive model for the test set

B. Secondary Analysis

The goal of our secondary analysis was to create an associational model that explored the effect of access to healthcare on not getting a pap smear. Starting with a full model of all 15 variables from our data set, we used backward and forward selection to refine the model. Both selection methods gave us the same model. Building off the backward/forward selection model, we added a quadratic term for age and found that it was statistically significant at the $\alpha = 0.05$ level (p-value = 1.32e-11), giving us statistical evidence for a nonlinear effect of age on not getting a pap smear. We also tried a cubic spline term with 3 knots for age and found from the likelihood ratio test (LRT) that at least one of the spline terms was statistically significant (p-value = < 2.2e-16). Since the backward/forward selection model adjusted for quadratic age is nested within the backward/forward selection model adjusted for the cubic spline of age with 3

knots, we were able to perform a LRT; we rejected the null hypothesis that linear and quadratic age were sufficient to model the effect of age on not getting a pap smear, so we went forward with the spline model instead of the quadratic one. Next we tested if marital status was an effect modifier of the effect of family income on not getting a pap smear using LRT. We rejected the null hypothesis that the simpler backward/forward selection model adjusting for a cubic spline of age with 3 knots was sufficient for modeling (p-value = 2.351e-05), concluding that there is statistical evidence of interaction. Building off of this interaction model, we added another interaction term; this time between education and total medical expenditures. We rejected the null hypothesis that the earlier model with just one interaction term was sufficient (p-value = 0.0122), concluding that there is statistical evidence that education modifies the effect of total medical expenditures on not getting a pap smear. In order to ensure interpretability of the access to healthcare coefficients in our model, we did not run any other more complex models.

Based on our hypothesis testing, it would seem that the backward/forward selection model adjusted for a cubic spline of age with 3 knots and interaction terms between marital status and family income and between education and total medical expenditures modeled the data best. Furthermore, it also had the lowest AIC value from the six models, so we decided to choose it as our final association model. Results from the Hosmer-Lemeshow test and the ROC indicated that the model had good calibration (p-value = 0.5089) and discrimination (Figure 2, AUC of 74.00%), respecitively. A summary of the output of the final association model can be seen in Table 4 of the Extra Tables & Figures section.

Association models	df	AIC
Full model	30	6030.07
Backward/forward selection model	22	6025.70
Backward/forward selection model + quad. age	23	5982.08

Backward/forward selection model + cubic spline for age with 3 knots	27	5913.56
Backward/forward selection model + cubic spline for age with 3 knots + marital status * family income	31	5894.92
Backward/forward selection model + cubic spline for age with 3 knots + marital status * family income + education * total medical exp.	**33**	**5890.11**

Table 3: Summary of the association models we built and compared. Bolded model is the one chosen as the final model.

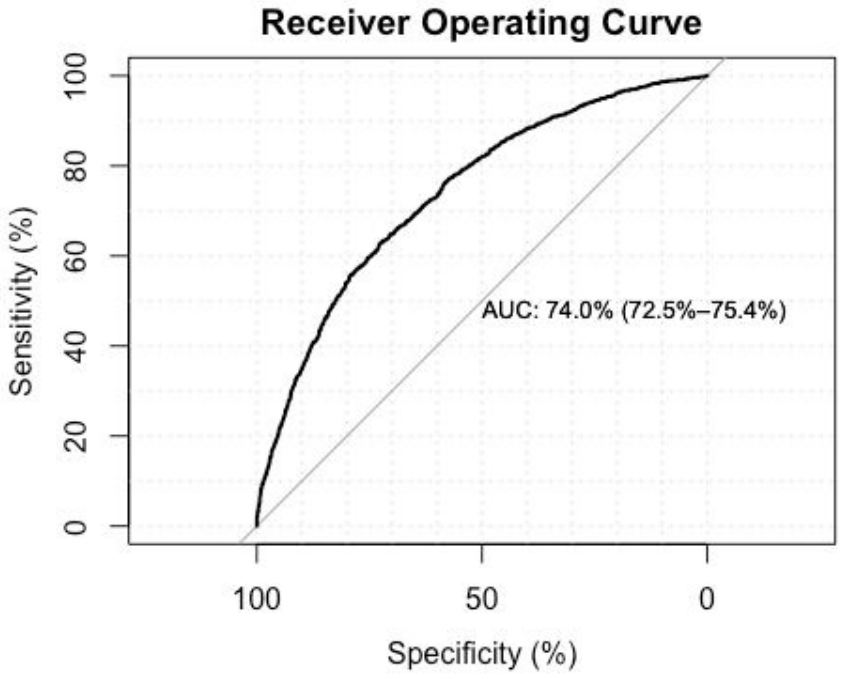

Figure 2: ROC of final association model

For the interpretation of the final model, we decided to focus on three of the access-related coefficients—specifically, ability to afford care (0 - no; 1 - yes), having a USC (0 - no; 1 - yes) and insurance coverage (0 - any private insurance; 1 - public only; 2 - uninsured) which were all fit as indicator variables in our model. The results for these coefficients can be seen in Table 4.

	exp(estimate)	exp(95% CI)	Std. Error	Z-value	P-value
afford_care_f	0.7468	(0.5915,	0.1174	-2.487	0.012873

yes		0.9375)			
have_usc_fyes	0.5192	(0.4507, 0.5983)	0.07228	-9.068	< 2e-16
inscov_gen_2018_fpubliconly	1.1430	(0.9596, 1.3604)	0.08902	1.501	0.133343
inscov_gen_2018_funinsured	2.0136	(1.6281, 2.4896)	0.1083	6.462	1.03e-10

Table 4: Results of three access-related coefficients from the final association model (see Extra Figures & Tables section for entire model). Note that the estimates and 95% confidence intervals (CIs) presented have been exponentiated. The baseline used for the afford care indicator is not being able to afford care. The baseline used for the USC indicator is not having a USC. The baseline used for the insurance indicators is having private insurance.

V. Discussion

In our primary analysis, we found that all 15 variables that we considered were predictive of not getting a pap smear, that nonlinear terms for age and for total medical expenditures helped improve prediction, and that interaction terms between marital status and family income and between education and total medical expenditures also helped improve prediction. Among those who did get a pap smear in the test set, 68.25% were correctly predicted as getting a pap smear by the model. Among those who did not get a pap smear in the test set, 68.28% were correctly predicted as not getting a pap smear. Among those who were predicted to get a pap smear by the model, 86.74% of them did in fact get a pap smear in the test set. Among those who were predicted to not get a pap smear by the model, 41.42% of them did in fact not get a pap smear in the test set. The model is incorrectly predicting a large number of people as not getting a pap smear only for it to turn out that they did get a pap smear. The stark difference in PPV and NPV can be explained by the relatively high prevalence of people who get a pap smear in the test set (75.26%).

In our secondary analysis, we found that among the people who self-reported as being able to afford medical care, the odds of not getting a pap smear was estimated to be, on average, 0.75 times that among the people who self-reported as not being able to afford care, adjusting for the other covariates in the model (95%: 0.59 to 0.94, p-value = 0.013). Among the people who had a USC, the odds of not getting a pap smear was estimated to be, on average, 0.52 times that among the people who did not have a USC, adjusting for the other covariates in the model (95%: 0.45 to 0.60, p-value = < 2e-16). Among the people who only had public insurance, the odds of not getting a pap smear was estimated to be, on average, 1.14 times that among the people who had private insurance, adjusting for the other covariates in the model (95%: 0.96 to 1.36, p-value = 0.13). Among the people who were uninsured, the odds of not getting a pap smear was estimated to be, on average, 2.01 times that among the people who had private insurance, adjusting for the other covariates in the model (95%: 1.63 to 2.49, p-value = 1.03e-10). These results suggest that there is an association between access to health care and not getting a pap smear in the U.S.

VI. Limitations

The MEPS data set is U.S.-specific. Since the U.S. has unique issues with income inequality and access to healthcare and has a unique populace relative to other nations, we cannot generalize our results to women in other countries. In both of our analyses, we assumed that we had MAR data and used a complete case analysis. However, it is possible that we could have missing not at random (MNAR) data and as a result there could be bias in the results. We also were limited to the variables that were available in MEPS, so we could be potentially missing important predictive variables, confounders and effect modifiers, leading to potential bias in our models.

VII. Future Scope

By using these 15 predictors from the MEPS dataset, the highest accuracy we were able to get that balanced sensitivity and specificity was 68.26%. While not terrible, this accuracy is not optimal for the prediction model to be used confidently. Future research could consider using a more expansive dataset that includes variables that we did not have in the MEPS dataset, such as HPV vaccination status and sexual history, as the inclusion of these variables could potentially improve the prediction model. Future research might also consider looking at machine learning methods, such as Random Forest, as those might be more successful at accurately predicting the likelihood of not getting a pap smear than the multivariable logistic regression that was used for this paper.

Lastly, our paper focuses on identifying people who are at risk of not getting a pap smear. However, failure to get a pap smear is likely more detrimental for certain populations than others. A more holistic approach could be to not only determine who is failing to receive pap smears, but also who needs them the most. As a result, future predictive modeling might want to focus on identifying those who do not get a pap smear and who are at higher risk of cervical cancer and thus will benefit most from getting regular pap smears.

VIII. Extra Tables & Figures

Variable (as seen in model outputs)	Description
race_f	Race/ethnicity (0 - white; 1- hispanic; 2 - black; 3 - asian; 4 - other or multiple races)
age	Age (continuous; 21-65 year old only)
marital_stat_f	Marital status (0 - never married, 1 - married, 2 - widowed; 3 - divorced; 4 - separated)

educ_f	Education (0 - any college; 1 - any high school but no college; 2 - none or any elementary but no high school or college)
genhlth_avg_f	Self-reported general health status (0 - poor; 1 - fair; 2 - good; 3 - very good; 4 - excellent)
region_f	Region (0 - northeast; 1 - midwest; 2 - south; 3 - west)
smoke_f	Smoking frequency (0 - never; 1 - some days; 2 - every day)
limitation_f	Limitation in work/housework/school (0 - no; 1 - yes)
afford_care_f	Ability to afford care (0 - no; 1 - yes)
income_indiv	Individual income (continuous)
income_fam	Family income (continuous)
totexp	Total medical expenditures (continuous)
outofpocket_exp	Out of pocket medical expenditures (continuous)
have_usc_f	Having a usual source of care (USC) (0 - no; 1 - yes)
inscov_gen_2018_f	Insurance coverage (0 - any private, 1 - public only, 2 - uninsured)

Table 1: Descriptions of the variables used during modeling

	pap_num	age	income_indiv	income_fam	totexp	outofpocket_exp	genhlth_avg_f	region_f	race_f	marital_stat_f	educ_f	smoke_freq_f	limitation_f	afford_care_f	have_usc_f	inscov_gen_2018_f
pap_num	1.00	-0.01	-0.18	-0.15	-0.07	-0.08	-0.06	0.03	0.13	-0.06	0.17	0.08	0.05	0.01	-0.16	0.20
age	-0.01	1.00	0.14	0.09	0.12	0.15	-0.13	-0.05	-0.06	0.36	0.07	0.04	0.16	0.03	0.19	-0.08
income_indiv	-0.18	0.14	1.00	0.66	0.03	0.14	0.21	-0.02	-0.11	0.08	-0.32	-0.13	-0.16	-0.09	0.09	-0.34
income_fam	-0.15	0.09	0.66	1.00	0.02	0.14	0.26	0.00	-0.11	-0.04	-0.31	-0.18	-0.16	-0.14	0.10	-0.37
totexp	-0.07	0.12	0.03	0.02	1.00	0.34	-0.21	-0.06	-0.05	0.05	-0.02	0.03	0.25	0.01	0.13	-0.06
outofpocket_exp	-0.08	0.15	0.14	0.14	0.34	1.00	-0.04	-0.03	-0.12	0.06	-0.10	-0.02	0.05	0.04	0.12	-0.13
genhlth_avg_f	-0.06	-0.13	0.21	0.26	-0.21	-0.04	1.00	0.02	-0.07	-0.09	-0.21	-0.18	-0.38	-0.21	-0.09	-0.17
region_f	0.03	-0.05	-0.02	0.00	-0.06	-0.03	0.02	1.00	0.11	0.00	0.06	-0.07	-0.03	0.01	-0.07	0.06
race_f	0.13	-0.06	-0.11	-0.11	-0.05	-0.12	-0.07	0.11	1.00	-0.07	0.10	-0.04	0.04	-0.01	-0.08	0.12
marital_stat_f	-0.06	0.36	0.08	-0.04	0.05	0.06	-0.09	0.00	-0.07	1.00	0.05	0.04	0.09	0.08	0.09	0.02
educ_f	0.17	0.07	-0.32	-0.31	-0.02	-0.10	-0.21	0.06	0.10	0.05	1.00	0.12	0.09	0.05	-0.06	0.32
smoke_freq_f	0.08	0.04	-0.13	-0.18	0.03	-0.02	-0.18	-0.07	-0.04	0.04	0.12	1.00	0.15	0.07	0.00	0.12
limitation_f	0.05	0.16	-0.16	-0.16	0.25	0.05	-0.38	-0.03	0.04	0.09	0.09	0.15	1.00	0.08	0.11	0.14
afford_care_f	0.01	0.03	-0.09	-0.14	0.01	0.04	-0.21	0.01	-0.01	0.08	0.05	0.07	0.08	1.00	-0.06	0.15
have_usc_f	-0.16	0.19	0.09	0.10	0.13	0.12	-0.09	-0.07	-0.08	0.09	-0.06	0.00	0.11	-0.06	1.00	-0.18
inscov_gen_2018_f	0.20	-0.08	-0.34	-0.37	-0.06	-0.13	-0.17	0.06	0.12	0.02	0.32	0.12	0.14	0.15	-0.18	1.00

Table 2: Correlation coefficients between variables in our data set

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.5992557	0.6365338	2.5124442	0.0119898
age	-0.1163195	0.0245117	-4.7454689	0.0000021
I(age^2)	0.0015846	0.0002781	5.6974811	0.0000000
income_indiv	-0.0000079	0.0000018	-4.2860707	0.0000182
income_fam	0.0000052	0.0000014	3.7004430	0.0002152
bSpline(totexp, df = 6, degrees = 3)1	-0.1855308	0.1918211	-0.9672079	0.3334401
bSpline(totexp, df = 6, degrees = 3)2	-1.1014068	0.1657245	-6.6460089	0.0000000
bSpline(totexp, df = 6, degrees = 3)3	-1.2363050	0.1542857	-8.0130911	0.0000000
bSpline(totexp, df = 6, degrees = 3)4	-1.9088076	1.1272849	-1.6932788	0.0904024
bSpline(totexp, df = 6, degrees = 3)5	-2.4926742	4.0449373	-0.6162454	0.5377325
bSpline(totexp, df = 6, degrees = 3)6	-7.5130246	8.0930300	-0.9283327	0.3532350
outofpocket_exp	0.0000131	0.0000241	0.5446767	0.5859759
genhlth_avg_ffair	0.5255282	0.3591940	1.4630762	0.1434465
genhlth_avg_fgood	0.3217651	0.3561637	0.9034191	0.3663035
genhlth_avg_fvery good	0.1467213	0.3611996	0.4062055	0.6845916
genhlth_avg_fexcellent	0.1554651	0.3712456	0.4187663	0.6753869
region_fmidwest	-0.0698802	0.1372735	-0.5090583	0.6107113
region_fsouth	0.0944786	0.1226282	0.7704473	0.4410346
region_fwest	-0.0640889	0.1312492	-0.4882993	0.6253378
race_fhispanic	0.0427627	0.1100171	0.3886919	0.6975041
race_fblack	0.1855162	0.1183269	1.5678281	0.1169212
race_fasian	1.0750124	0.1637258	6.5659300	0.0000000
race_fother or multiple races	-0.1474384	0.2270754	-0.6492925	0.5161493
marital_stat_fmarried	-0.1817249	0.1400919	-1.2971830	0.1945682
marital_stat_fwidowed	-0.0542058	0.3066086	-0.1767914	0.8596723
marital_stat_fdivorced	-0.0670234	0.2087529	-0.3210658	0.7481605
marital_stat_fseperated	-0.2977082	0.2928810	-1.0164819	0.3094000
educ_fany high school	0.2762522	0.0958433	2.8823332	0.0039474
educ_fnone or any elementary	0.3363555	0.1955191	1.7203209	0.0853741
smoke_freq_fsome days	0.0722232	0.1854421	0.3894649	0.6969323
smoke_freq_fevery day	0.3833683	0.1243295	3.0834850	0.0020459
limitation_fyes	0.2593218	0.1587643	1.6333758	0.1023900
afford_care_fyes	-0.3023831	0.1418649	-2.1314870	0.0330490
have_usc_fyes	-0.4016763	0.0894287	-4.4915825	0.0000071
inscov_gen_2018_fpublic only	0.0827237	0.1084989	0.7624380	0.4457986
inscov_gen_2018_funinsured	0.3888730	0.1332907	2.9174811	0.0035287
income_fam:marital_stat_fmarried	-0.0000069	0.0000016	-4.2457175	0.0000218
income_fam:marital_stat_fwidowed	-0.0000045	0.0000053	-0.8397525	0.4010471
income_fam:marital_stat_fdivorced	-0.0000052	0.0000036	-1.4335945	0.1516880
income_fam:marital_stat_fseperated	0.0000019	0.0000059	0.3274364	0.7433379
educ_fany high school:totexp	0.0000145	0.0000073	1.9945394	0.0460931
educ_fnone or any elementary:totexp	-0.0000134	0.0000189	-0.7088432	0.4784218

Table 3: Summary output of final logistic regression model for prediction

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.3507453	0.2009087	1.7457943	0.0808467
bSpline(age, df = 6, degree = 3)1	-1.6356743	0.3396448	-4.8158378	0.0000015
bSpline(age, df = 6, degree = 3)2	-1.5490867	0.2361593	-6.5594989	0.0000000
bSpline(age, df = 6, degree = 3)3	-1.4020432	0.2828291	-4.9572090	0.0000007
bSpline(age, df = 6, degree = 3)4	-0.1769737	0.2767043	-0.6395771	0.5224476
bSpline(age, df = 6, degree = 3)5	-1.2060016	0.2925262	-4.1227126	0.0000374
bSpline(age, df = 6, degree = 3)6	-0.1832088	0.2519446	-0.7271790	0.4671163
income_indiv	-0.0000077	0.0000015	-4.9744563	0.0000007
income_fam	0.0000041	0.0000011	3.6164508	0.0002987
totexp	-0.0000203	0.0000057	-3.5740258	0.0003515
race_fhispanic	0.2399707	0.0871203	2.7544765	0.0058786
race_fblack	0.2355481	0.0963680	2.4442574	0.0145151
race_fasian	1.1813189	0.1330808	8.8767037	0.0000000
race_fother or multiple races	0.1866396	0.1788719	1.0434259	0.2967511
marital_stat_fmarried	-0.2107004	0.1173288	-1.7958108	0.0725246
marital_stat_fwidowed	-0.2538686	0.2620945	-0.9686146	0.3327375
marital_stat_fdivorced	-0.1851787	0.1757359	-1.0537329	0.2920052
marital_stat_fseperated	-0.5084346	0.2398230	-2.1200414	0.0340026
educ_fany high school	0.3616504	0.0792671	4.5624280	0.0000051
educ_fnone or any elementary	0.3021977	0.1633230	1.8503065	0.0642694
smoke_freq_fsome days	0.3325559	0.1491047	2.2303512	0.0257241
smoke_freq_fevery day	0.3941092	0.1028287	3.8326777	0.0001268
limitation_fyes	0.2574488	0.1245457	2.0671038	0.0387244
afford_care_fyes	-0.2920098	0.1174020	-2.4872641	0.0128730
have_usc_fyes	-0.6554217	0.0722767	-9.0682238	0.0000000
inscov_gen_2018_fpublic only	0.1336261	0.0890217	1.5010503	0.1333425
inscov_gen_2018_funinsured	0.6999469	0.1083155	6.4621121	0.0000000
income_fam:marital_stat_fmarried	-0.0000062	0.0000013	-4.6289919	0.0000037
income_fam:marital_stat_fwidowed	-0.0000018	0.0000045	-0.3932157	0.6941602
income_fam:marital_stat_fdivorced	-0.0000058	0.0000032	-1.8059737	0.0709225
income_fam:marital_stat_fseperated	0.0000040	0.0000045	0.8887114	0.3741582
totexp:educ_fany high school	0.0000175	0.0000067	2.6282367	0.0085829
totexp:educ_fnone or any elementary	-0.0000035	0.0000148	-0.2349077	0.8142804

Table 4: Summary output of final logistic regression model for association

IX. References

1. Akers, Aletha Y., Sara J. Newmann, and Jennifer S. Smith. 2007. "Factors Underlying Disparities in Cervical Cancer Incidence, Screening, and Treatment in the United States." *Current Problems in Cancer* 31 (3): 157–81. https://doi.org/10.1016/j.currproblcancer.2007.01.001.

2. Akinlotan, Marvellous, Jane N. Bolin, Janet Helduser, Chinedum Ojinnaka, Anna Lichorad, and David McClellan. 2017. "Cervical Cancer Screening Barriers and Risk Factor Knowledge Among Uninsured Women." *Journal of Community Health* 42 (4): 770–78. https://doi.org/10.1007/s10900-017-0316-9.

3. Bharel, Monica, Carolyn Casey, and Eve Wittenberg. 2009. "Disparities in Cancer Screening: Acceptance of Pap Smears among Homeless Women." *Journal of Women's Health* 18 (12): 2011–16. https://doi.org/10.1089/jwh.2008.1111.

4. CDC Cancer. 2020. "Cervical Cancer Awareness." Centers for Disease Control and Prevention. April 30, 2020. https://www.cdc.gov/cancer/dcpc/resources/features/cervicalcancer/index.htm.

5. "Cervical Cancer Statistics | Key Facts About Cervical Cancer." American Cancer Society. n.d. Accessed December 10, 2020. https://www.cancer.org/cancer/cervical-cancer/about/key-statistics.html.

6. Charlton, Brittany M., Heather L. Corliss, Stacey A. Missmer, A. Lindsay Frazier, Margaret Rosario, Jessica A. Kahn, and S. Bryn Austin. 2011. "Reproductive Health Screening Disparities and Sexual Orientation in a Cohort Study of U.S. Adolescent and Young Adult Females." *The Journal of Adolescent Health : Official Publication of the Society for Adolescent Medicine* 49 (5): 505–10. https://doi.org/10.1016/j.jadohealth.2011.03.013.

7. Guo, Fangjian, Jacqueline M. Hirth, and Abbey B. Berenson. 2017. "Human Papillomavirus Vaccination and Pap Smear Uptake Among Young Women in the United States: Role of Provider and Patient." *Journal of Women's Health* 26 (10): 1114–22. https://doi.org/10.1089/jwh.2017.6424.

8. "H209. MEPS HC-209 2018 Full Year Consolidated Data File." n.d. Accessed December 10, 2020. https://www.meps.ahrq.gov/data_stats/download_data/pufs/h209/h209doc.pdf.

9. Maj, Chiara, Lorraine Poncet, Henri Panjo, Arnaud Gautier, Pierre Chauvin, Gwenn Menvielle, Emmanuelle Cadot, Virginie Ringa, and Laurent Rigal. 2019. "General Practitioners Who Never Perform Pap Smear: The Medical Offer and the Socio-Economic Context around Their Office Could Limit Their Involvement in Cervical Cancer Screening." *BMC Family Practice* 20 (August). https://doi.org/10.1186/s12875-019-1004-x.

10. "Medical Expenditure Panel Survey Public Use File Details." n.d. Agency for Healthcare Research and Quality. Accessed December 10, 2020. https://www.meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-209.

11. Melnikow, Joy, Jillian T. Henderson, Brittany U. Burda, Caitlyn A. Senger, Shauna Durbin, and Meghan A. Soulsby. 2018. *Screening for Cervical Cancer With High-Risk Human Papillomavirus Testing: A Systematic Evidence Review for the U.S. Preventive Services Task Force*. U.S. Preventive Services Task Force Evidence Syntheses, Formerly Systematic Evidence Reviews. Rockville (MD): Agency for Healthcare Research and Quality (US). http://www.ncbi.nlm.nih.gov/books/NBK526306/.

12. "The American Cancer Society Guidelines for the Prevention and Early Detection of Cervical Cancer." n.d. Accessed December 10, 2020a. https://www.cancer.org/cancer/cervical-cancer/detection-diagnosis-staging/cervical-cancer-screening-guidelines.html.

13. US Preventive Services Task Force, Susan J. Curry, Alex H. Krist, Douglas K. Owens, Michael J. Barry, Aaron B. Caughey, Karina W. Davidson, et al. 2018. "Screening for Cervical Cancer: US Preventive Services Task Force Recommendation Statement." *JAMA* 320 (7): 674–86. https://doi.org/10.1001/jama.2018.10897.

X. Appendix

See attached document for appendix of our code.