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

Anja Shahu, Ligia Flores, Anna Wuest

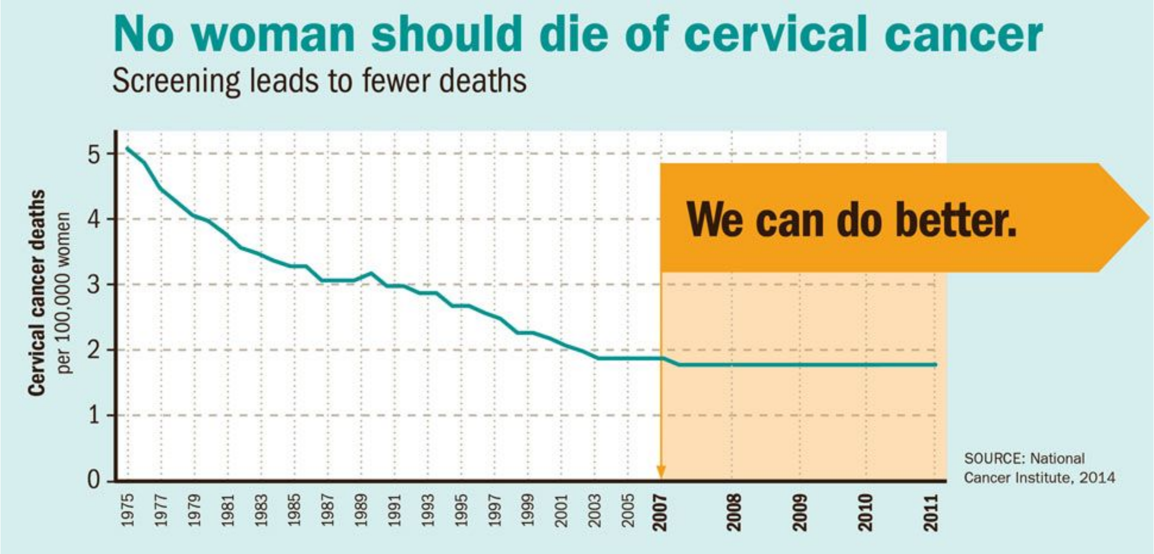
Introduction

Primary Question:

- How well do individual-level factors and accessibility to healthcare predict the likelihood of not getting a pap smear in the last 5 years among American women aged 21-65?

Secondary Question:

- What is the effect of access to healthcare variables on the probability of not getting a pap smear?



Source: CDC | Cervical Cancer is Preventable infographic

Variables and Type of Modeling

Data set

- 2018 Full Year Consolidated Data File from the Medical Expenditure Panel Survey (MEPS) by the U.S. Department of Health and Human Services (HHS)
- 5863 observations of women aged 21-65 after removing 773 missing observations

Outcome variable

- If someone has received a pap smear in the last five years (0 - pap smear; 1 - no pap smear)
- Used complete case multivariable logistic regression analysis

Predictor 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)
- Insurance coverage

Primary Question

Goal

- Predict the likelihood of not getting a pap smear in the last 5 years among American women aged 21-65 using individual-level and access to healthcare factors

Methodology

- 70% train set (4105 observations) and 30% test set (1758 observations)
- Use cross validation to build model on train set and test on test set
- Selected model that maximized AUC on the test set.

Model	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%

Primary Question

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

Table 1: Predicted vs observed pap smear values for p-cutoff of 0.25 based on final predictive model for the test set

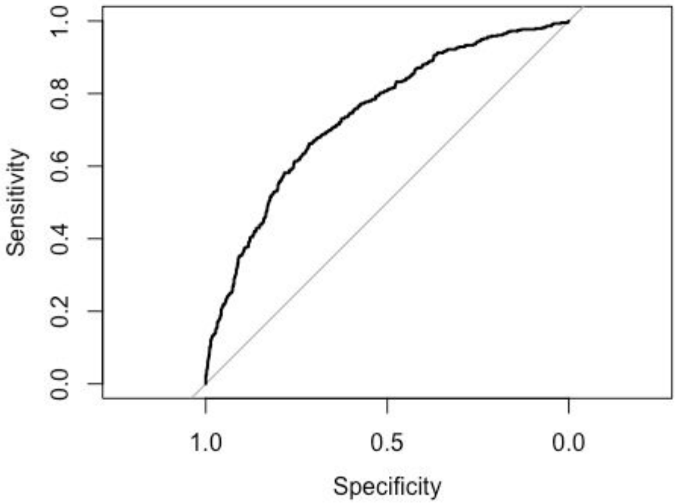


Figure 1: ROC of final prediction model with AUC of 74.09% on the test set

- Accuracy: 68.26%
- Sensitivity: 68.25%
- Specificity: 68.28%
- PPV: 86.74%
- NPV: 41.42%
- Positive class = “yes pap smear”

Secondary Question

How does access to healthcare impact pap smear use?

- To assess this, we focused on health-care related variables:
  - Ability to afford care
  - Usual source of care
  - Insurance coverage

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**

Secondary Question Con’t

	exp(estimate)	exp(95% CI)	Std. Error	Z-value	P-value
Ability to Afford Care	0.7468	(0.5915, 0.9375)	0.1174	-2.487	0.012873
Usual Source of Care	0.5192	(0.4507, 0.5983)	0.07228	-9.068	< 2e-16
Public Insurance	1.1430	(0.9596, 1.3604)	0.08902	1.501	0.133343
No Insurance	2.0136	(1.6281, 2.4896)	0.1083	6.462	1.03e-10

Takeaway

Conclusion

- Socio-demographic, health status, smoking, access to healthcare and medical expenditure variables were predictive of not getting a pap smear.
- There is an association between access to health care and not getting a pap smear in the U.S.
- The highest accuracy we were able to get that balanced sensitivity and specificity was 68.26%

Limitations

- Results cannot be generalized to women in other countries
- Our assumption on the type of missing data could be inaccurate leading to bias
- Limited to variables in the dataset

Future Scope

- Use a more expansive dataset that includes variables not included in the MEPS dataset
- Look at machine learning methods, such as Random Forest
- Look into who are at higher risk of cervical cancer instead