Population-Based Disease Risk Prediction Modeling using Survey, Clinical, and Registry Data: Application to Risk Prediction for Oropharyngeal Cancer in the US

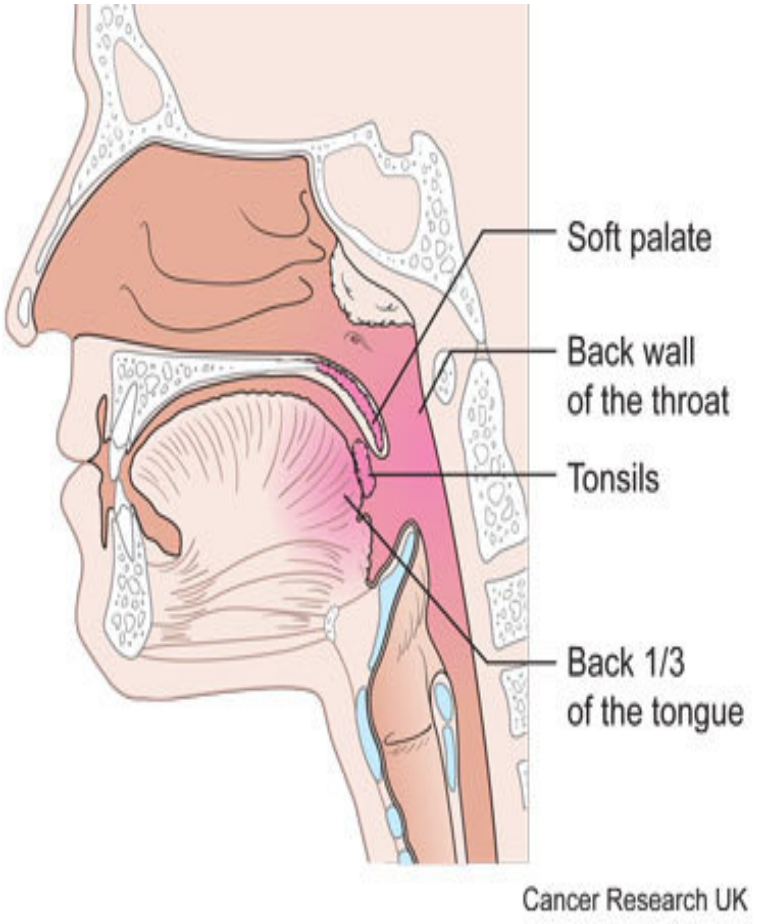
Barry I Graubard, PhD
Senior Investigator
Biostatistics Branch
DCEG, NCI

NIH NATIONAL CANCER INSTITUTE

Background

Oropharyngeal Cancer (OPC)

- Subset of head and neck cancers
- OPC risk factors
  - Tobacco and alcohol
  - Sexual behavior
  - **Oral** human papillomavirus (HPV)
  - Oncologic HPV subtypes 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58 & 59
- U.S. burden
  - <10 OPC cases per 100,000 in US
  - Increasing in US since 1980's
  - >70% of OPCs HPV+ (~7,500/year)
  - >90% subtype HPV16+

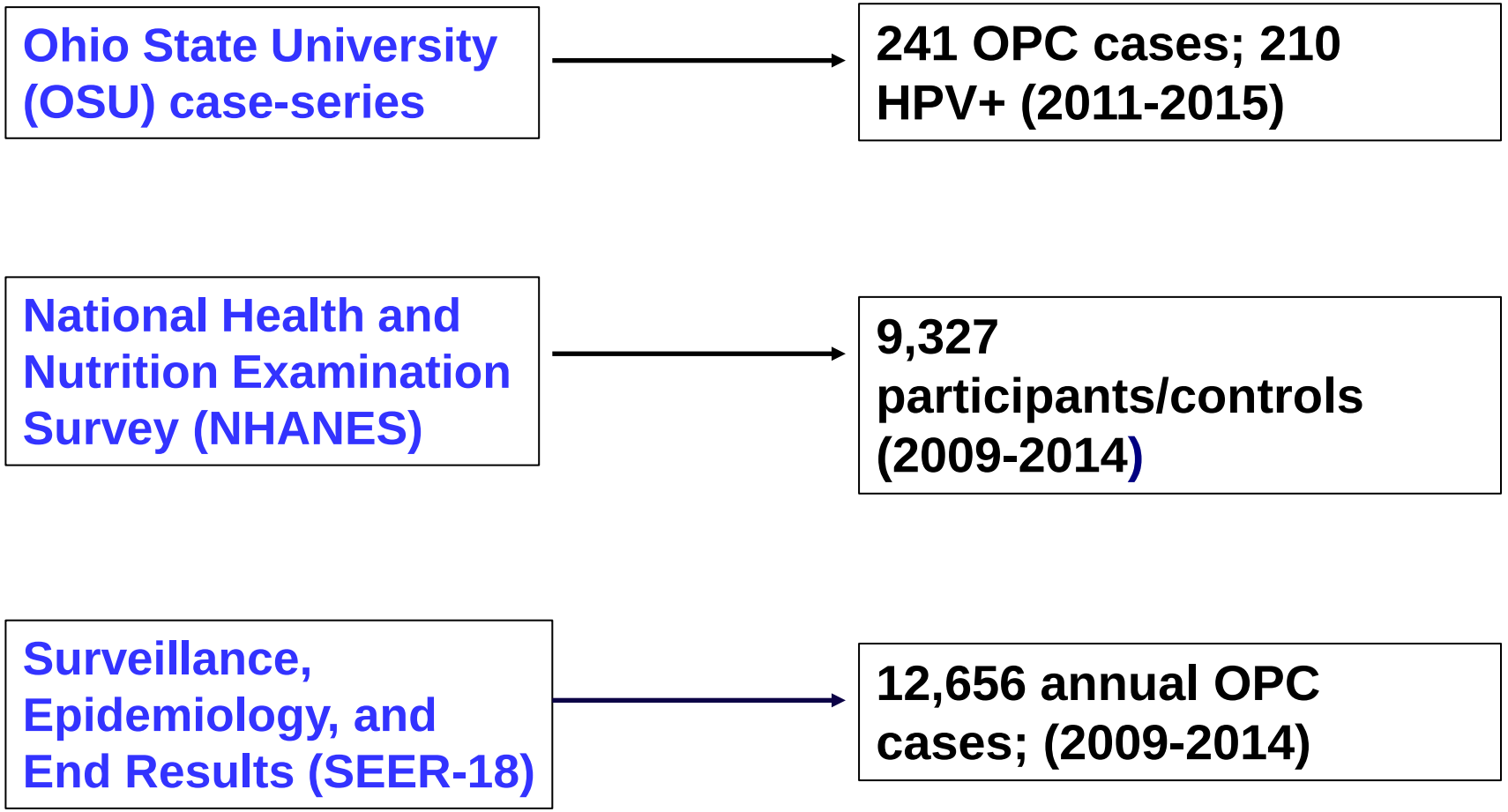


Objectives

- Develop risk model using oncologic HPV subtypes as a risk factor that predict one-year individualized risk of OPC among 30-69 yr. in U.S. population by creating a “synthetic” population-based case-control study.
- Validate the OPC risk model
- Characterize the extent risk stratification
- Alternative risk models and their validation and calibration

Sources of Data

3 data sources for OPC risk model



Comparison of demographic and behavioral information from subjects in the Ohio State University study and NHANES

	Information available/used from cases (OSU Study)	Information available/used from controls (NHANES)	Coding applied
Age	Age in years (continuous)	Age in years (continuous)	Continuous (age)
Gender	Male or female (binary)	Male or female (binary)	Binary (male, female)
Race	White, Black or African American, Other or Multi-Racial (categorical)	Non-Hispanic White, Non-Hispanic Black, Mexican American, Other Hispanic, Other or Multi-Racial (categorical)	Categorical (white, black, other)
Smoking	Smoking status: -Never smoker or <100 cigarettes lifetime -Former smoker -Current smoker (categorical) Intensity: -Average number of cigarettes smoked per day (during smoking period(s) among former smokers and currently among current smokers; continuous) Duration: -Age started smoking (continuous) -Age quit smoking (continuous)	Smoking status: -Never smoker or <100 cigarettes lifetime -Former smoker -Current smoker (categorical) Intensity: -Average number of cigarettes smoked per day (at the time of cessation among former smokers and in the past 30 days among current smokers; continuous) Duration: -Age started smoking (continuous) -Age quit smoking (continuous)	Continuous (log pack-years) Note: Pack-years calculation = duration of smoking (years) x intensity (cigarette packs/day) 20 cigarettes = 1 pack
Alcohol	Current drinks/week: -0, >0 and ≤14, >14 (categorical)	Average number of drinks per week in past 12 months (continuous)	Binary (0-14, >14 drinks/week)
Lifetime Sex Partners	Total lifetime number of sexual partners (vaginal, anal, and oral sex), both genders (continuous)	Total lifetime number of sexual partners (vaginal, anal, and oral sex), both genders (continuous)	Categorical (0-1, 2-5, 6-10, >10 lifetime sex partners)

Distribution of age, gender and race among oropharynx cancer cases in Ohio State University study and SEER 18 Registries

	OSU study cases n=241	Annual oropharynx cancer cases in the US n=12,656	Beta coefficient	*P value*	Weight
Age				*0.99*	
30-39	1.2%	1.9%	Ref		71.00
40-49	14.5%	13.9%	0.44		50.34
50-59	43.6%	42.1%	0.43		50.49
60-69	40.7%	42.1%	0.36		54.89
Gender				*0.82*	
Male	85.9%	82.3%	Ref		50.44
Female	14.1%	17.7%	-0.25		65.11
Race				*0.87*	
White	87.6%	85.8%	Ref		51.39
Black	7.1%	9.9%	-0.35		71.69
Other	5.4%	4.4%	-0.22		45.63

Abbreviations: OSU, Ohio State University; SEER, Surveillance, Epidemiology, and End Results.

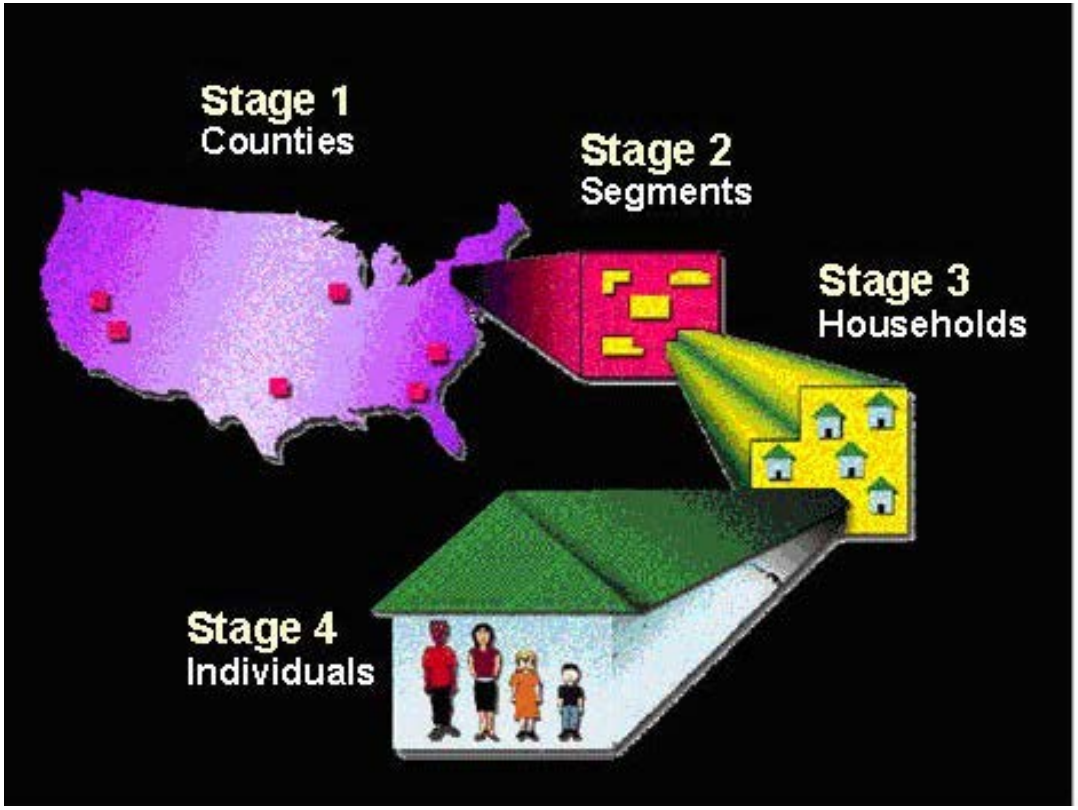
Weighting OSU cases to be representative of US OPC cases

- The 241 OPC OSU cases are a nonprobability sample.
- Fit a logistic regression with age, sex, and race to estimate propensities $p(x_i)$for being a OSU case vs. SEER-18 case among 30-69 yr. in US (2009-2014).

$$ln[p(x_i)] = \beta_0 + \beta_{Age1} I_{[30,39]} + \cdots + \beta_{Age4} I_{[60,69]} + \beta_{sex} I_{[male]} + \beta_{race1} I_{[black]} + \beta_{race2} I_{[other]}$$

- Case weights are inverse of propensities, $w_i = [p(x_i)]^{-1}$, *i=1, …, 240*
- As SEER18 registries cover ~28% of the US population the propensities rescaled to reflect the total US OPC cases, $w_i/0.28$.

Stratified Multistage Cluster Sample of the NHANES 2009-2014



- Counties are the Primary Sample Units (PSUs)
- PSUs partitioned into L strata
- Random sampling at each stage starting with PSU's
- Differential sampling rates at each stage → sample weights w_i for each sample person $i = 1, \ldots, 9{,}327$
- Re-postratified 9,327 controls, to match age/gender/race distribution of US

Risk Modeling

OPC risk prediction model development

- Weighted (propensity weights (w_{pi}) for cases and NHANES sample weights (w_{sj}) for controls) multivariable binary logistic regression used to estimate the 1-year risk.

  - Main effects: age, gender, race (white, black, other races), smoking (pack-years with a cubic spline), alcohol consumption (>14 drinks per week), lifetime no. sexual partners (any type of sex), and at least one prevalent oral oncogenic HPV subtype infection

  - Interactions: gender x HPV and smoking x HPV

  - Weighted estimating equations for β :

$$\sum_{i \epsilon cases} w_{pi} \pi_i \, x_{ik} + \sum_{j \epsilon controls} w_{sj} \pi_j \, x_{jk} = \sum_{i \epsilon cases} w_{pi} \, x_{ik}, \quad k = 0, 1, \ldots, p-1$$

$$\text{where } \pi_i = \Pr(OPC|x_i),\ logit\ (\pi_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{ip-1} + interactions$$

Logistic regression model for the prediction of 1-year risk of oropharynx cancers in the US population

	Beta coefficient	Odds ratio (95% CI)	*P value*
Intercept	-12.40		
Age (per year)	0.09	1.09 (1.07-1.11)	*<0.001*
Gender			*<0.001*
Male	Reference		
Female	-1.09	*See interaction below*	
Race			*<0.001*
White	Reference	1.0	
Black	-0.25	0.78 (0.45-1.36)	
Other	-1.30	0.27 (0.15-0.51)	
Pack-years of smoking	0.16	*See interaction below*	*0.04*
Alcohol use			*0.03*
≤14 drinks/week	Reference	1.0	
>14 drinks/week	0.60	1.82 (1.05-3.15)	
Lifetime sex partners			*0.12*
0-1	Reference	1.0	
2-5	0.16	1.17 (0.96-1.44)	
6-10	0.32	1.38 (0.92-2.06)	
>10	0.48	1.62 (0.88-2.95)	
High-risk HPV status			*<0.001*
Negative	Reference		
Positive	3.80	*See interaction below*	
Gender x HPV interaction	1.38		*0.004*
Male x HPV-		1.0	
Male x HPV+		44.8 (24.0-83.7)	
Female x HPV-		0.33 (0.19-0.60)	
Female x HPV+		59.7 (24.1-147)	
Smoking x HPV interaction	-0.33		*0.003*
Effect of HPV (x smoking)			
HPV+ in never smokers		44.8 (24.0-83.7)	
HPV+ in smokers		37.8 (21.8-65.7)	
Effect of smoking (x HPV)			
Smoking in HPV-		1.18 (1.01-1.37)	
Smoking in HPV+		0.84 (0.71-1.00)	

Variance estimation OPC risk prediction model

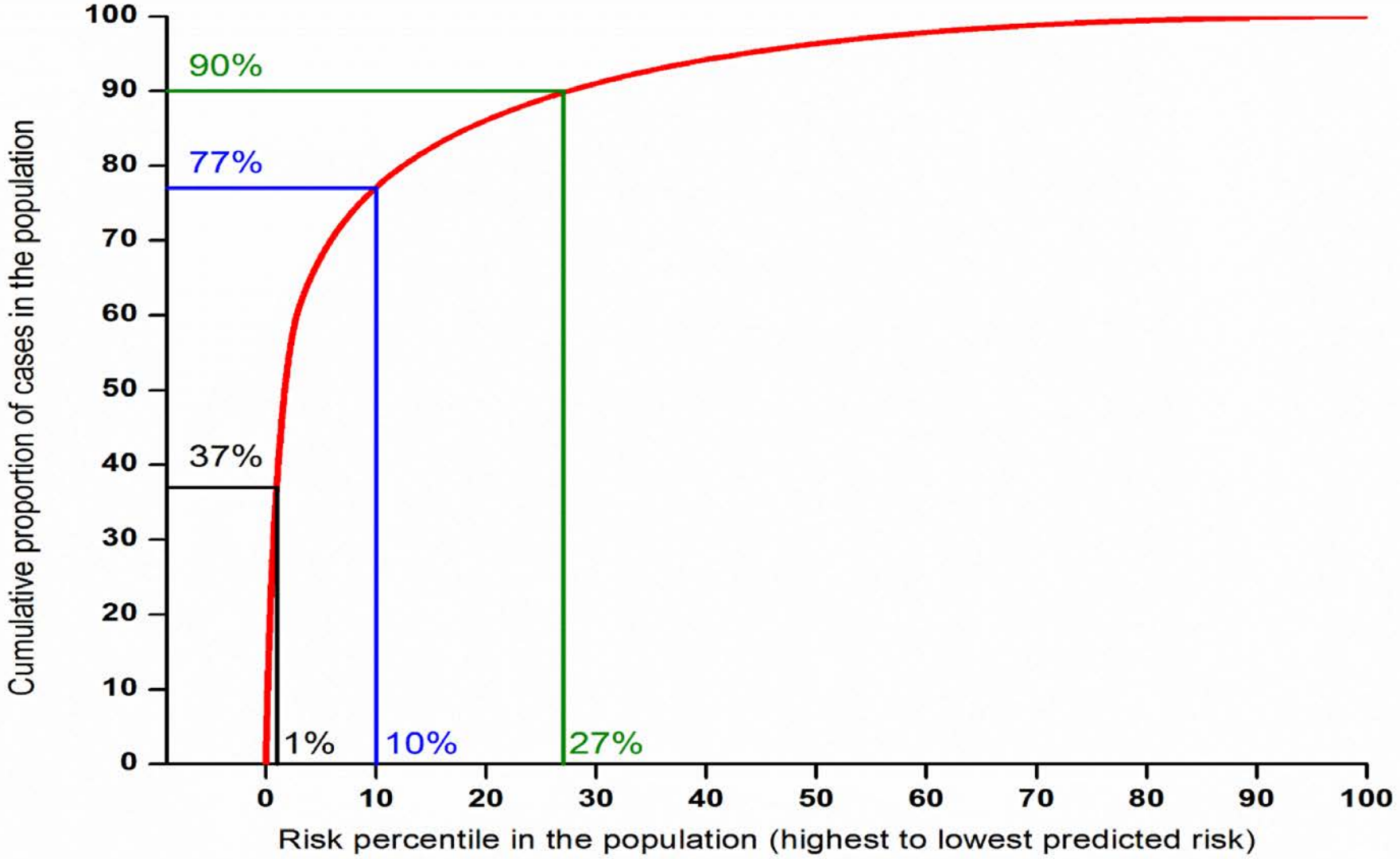
- Leaving-one-out Jackknife variances to account for estimation of OPC case weights & NHANES sample design:

1. Combine the cases and NHANES controls
2. Individual cases are PSU's from a stratum and the controls with PSU's from the L NHANES strata.
3. Leave-one PSU out at time recomputing the propensity weights for the cases left out and readjusting the sample weights for the control PSU's left out.
4. Re-estimate risk model with $\hat{\beta}_{(li)}$ leaving out PSU i in stratum l.
5. Repeat 1-4 and estimate variance:

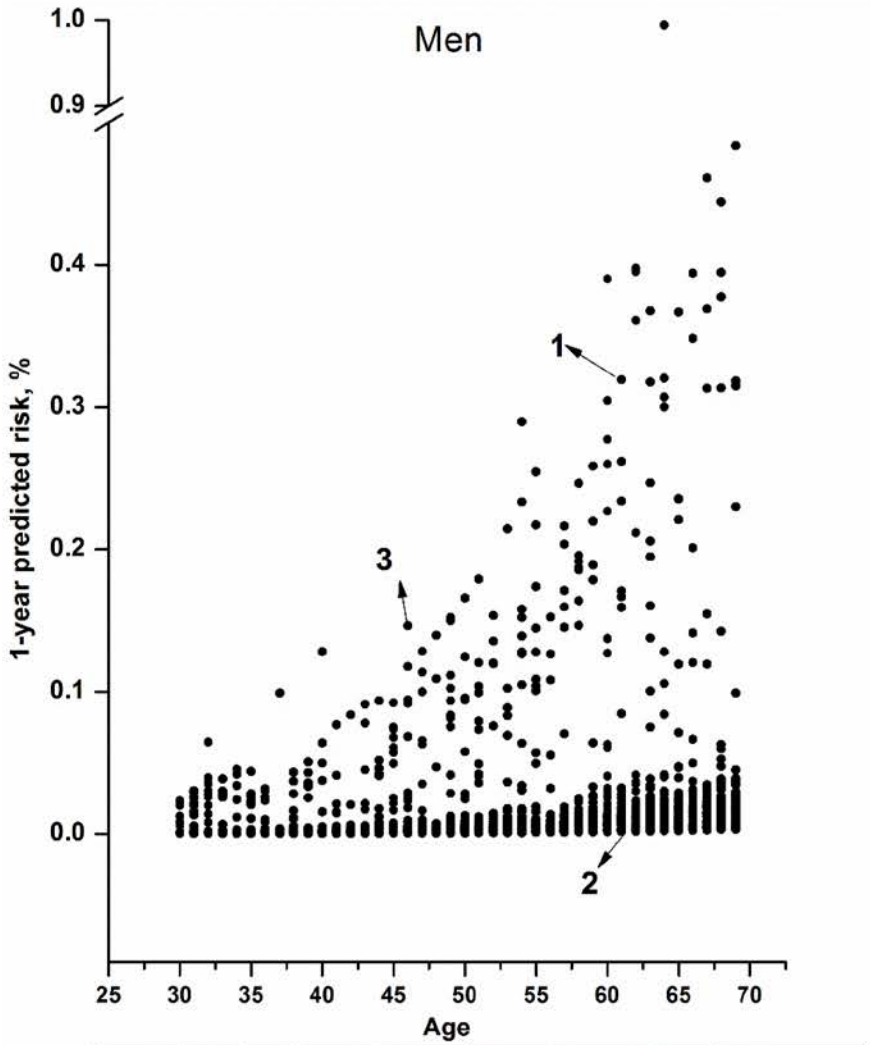
$$\sum_{l=1}^{L+1} \frac{k_l - 1}{k_l} \sum_{i=1}^{k_l} \left(\hat{\beta}_{(li)} - \hat{\beta}\right)^2$$

Estimated OPC Risks

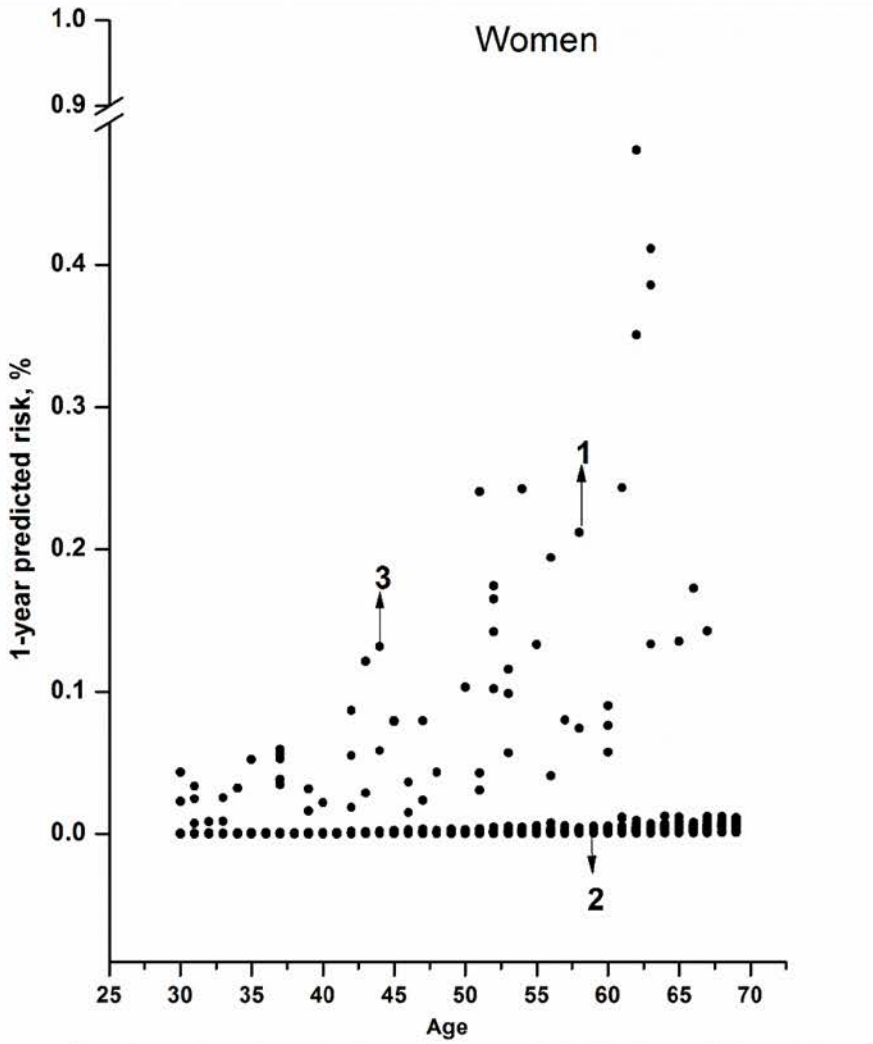
Concentration (Lorenz) curve of risk model for oropharynx cancer for 30-69 yr.



1-year risk of oropharynx cancer in the US population

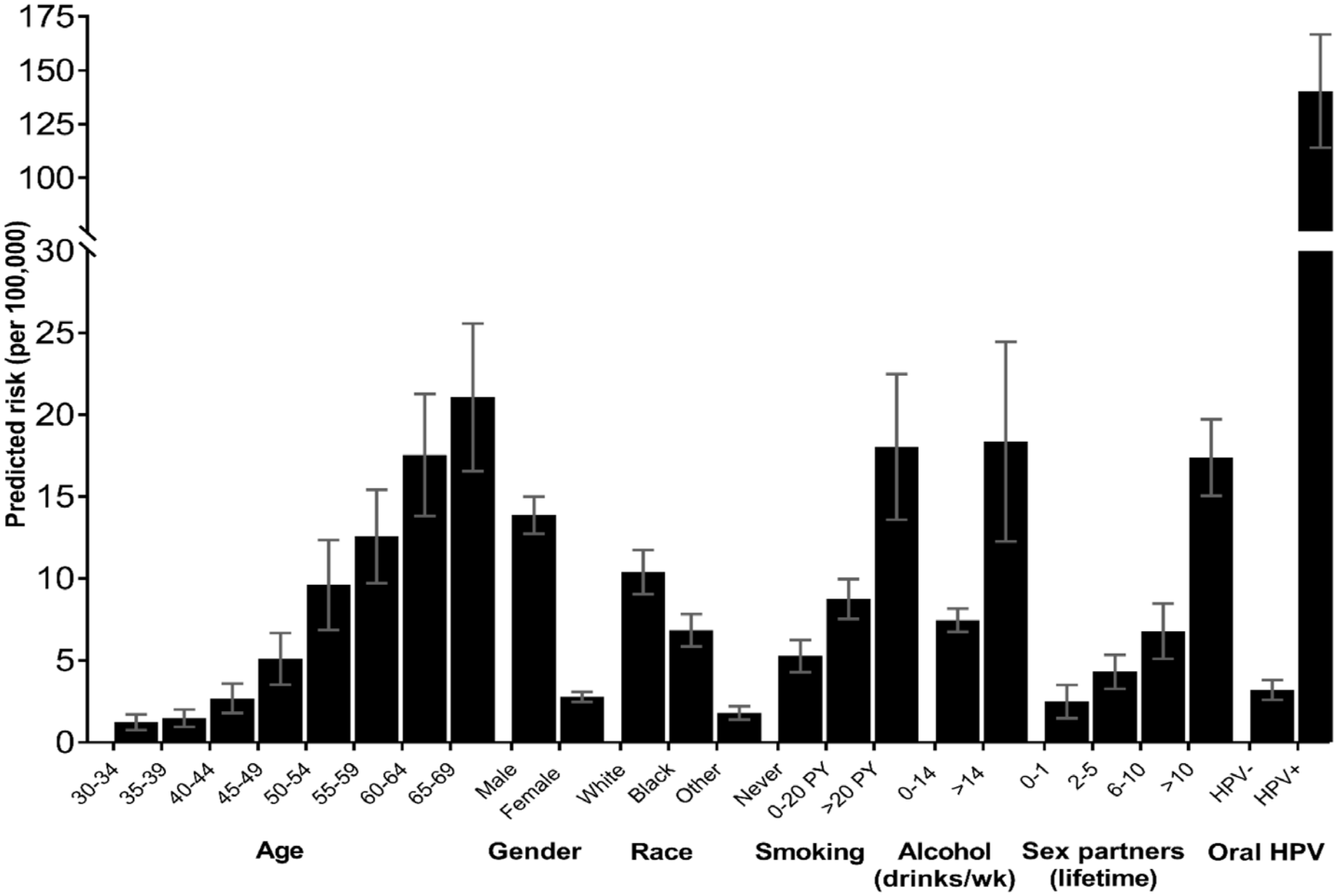


Obs	Age	Sex	Race	Packyrs	Alcohol	Partners	HPV	Predicted risk
1	60	M	Black	0	Low	20	+	0.3%
2	60	M	Other	0	Low	7	-	0.002%
3	46	M	White	8.4	Heavy	12	+	0.15%



Obs	Age	Sex	Race	Packyrs	Alcohol	Partners	HPV	Predicted risk
1	58	F	Black	5.6	Low	10	+	0.21%
2	58	F	White	4.8	Low	3	-	0.002%
3	44	F	White	0	Low	15	+	0.13%

Mean 1-year risk of oropharynx cancer across demographic and risk-factor subgroups in the US population



Risk Model Validation Calibration

Validation of the risk model

Internal Validation

- Training set included first 2/3 of OSU oropharynx cancer cases recruited into study + 2/3 of NHANES controls from 2009-2012
  - Estimated the risk model with the training set

- Test set included last 1/3 of OSU cases + NHANES 2013-2014

"External" Validation

- 116 oropharynx cancer cases (recruited 2000 to 2005 from John Hopkins University)

- Propensity-weighted JHU cases to SEER-18 for aged 30-69 yrs in 2009-2014.

- NHANES 2013-2014 as controls (closest to cases in calendar time)

Internal and external validation and calibration of oropharynx cancer risk prediction model

Measure	Internal validation	External validation
[1]AUC (95%CI)	0.94 (0.92-0.97)	0.87 (0.84-0.90)
Overall O/E ratio (95%CI)	1.01 (0.70-1.32)	1.08 (0.77-1.39)
O/E ratio in the highest decile of predicted risk (95%CI)	1.05 (0.67-1.44)	0.91 (0.57-1.25)
Percentage of oropharynx cancers in the highest decile of predicted risk	76.5%	61.7%

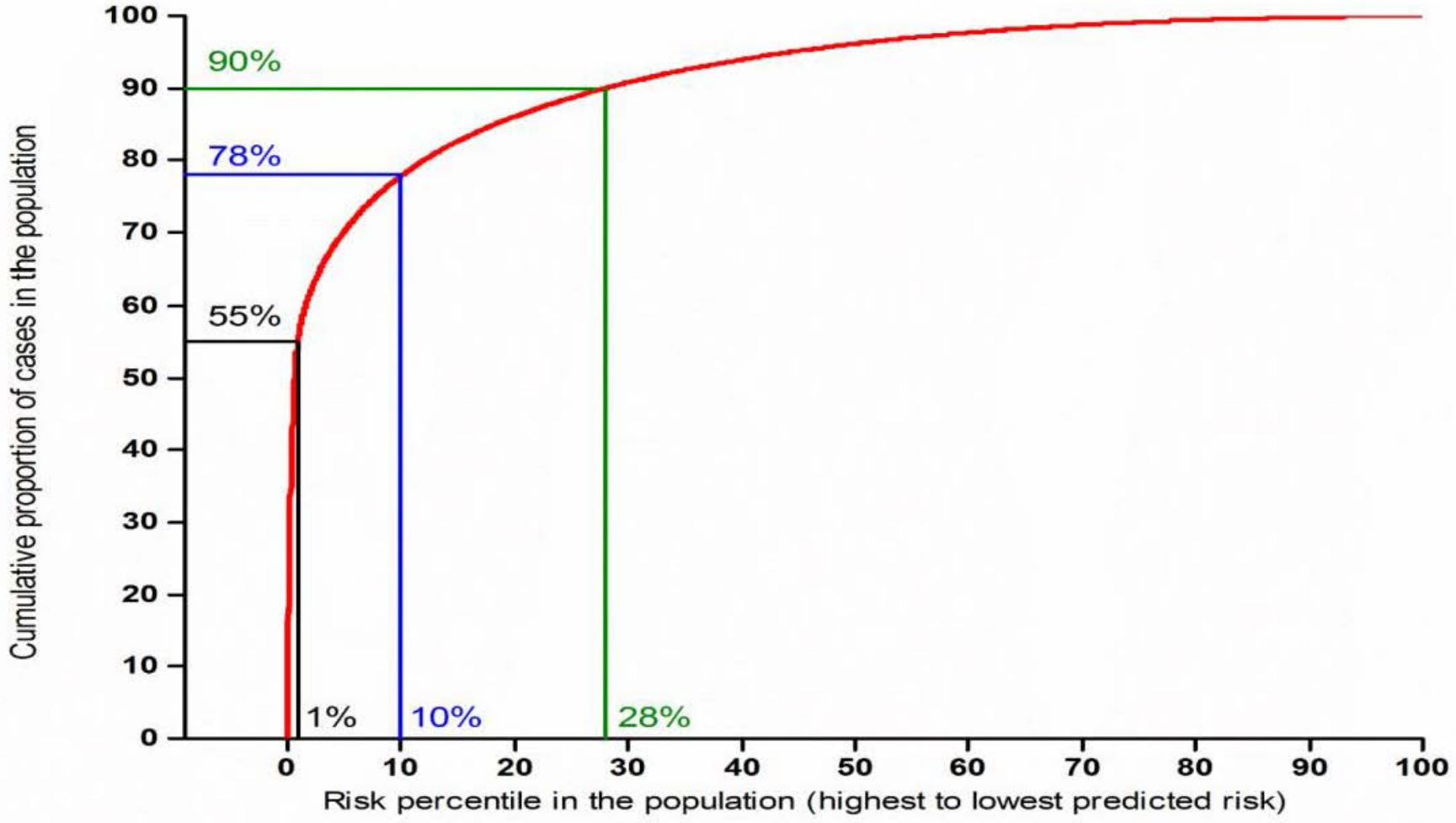
[1]AUC for survey data see Yao, Li, Graubard Stat. Med. 2015 with joint sample wts. approximated by product of individual sample wts.

Alternate models

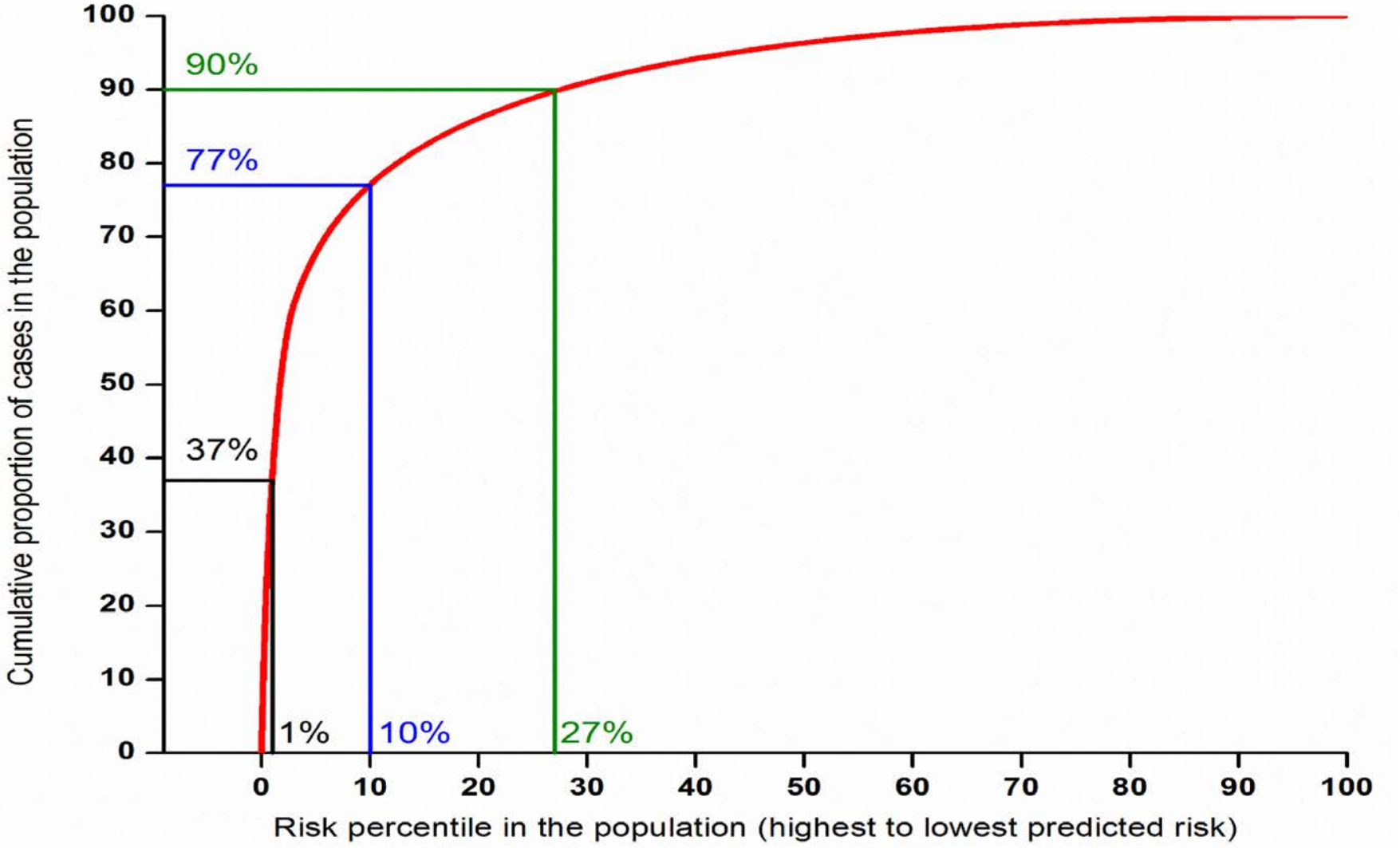
1) Separately considers HPV16 infection as risk factor

2) Model that includes only demographic and behavioral factors (w/o HPV)

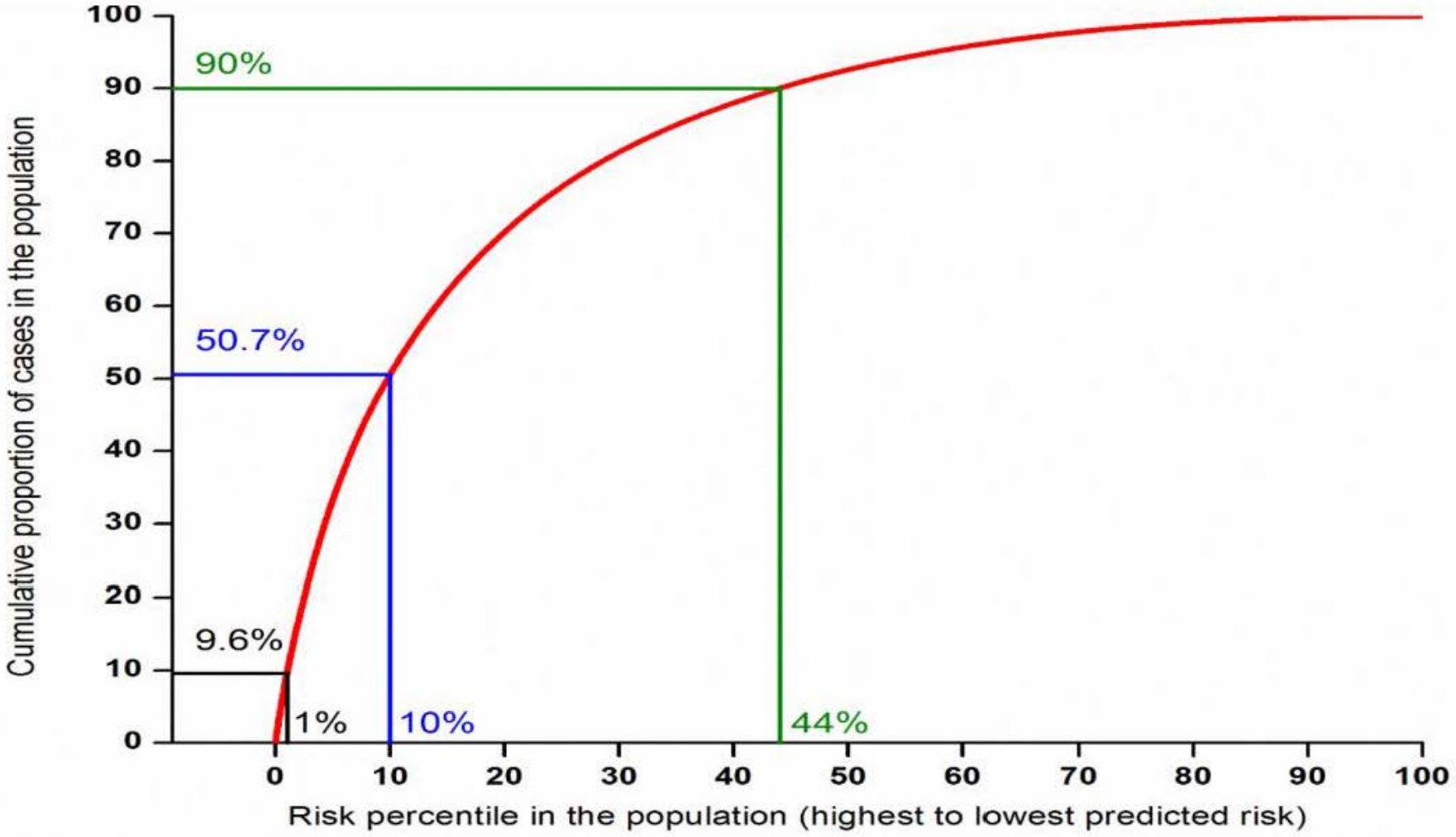
Concentration curve of the cumulative proportion of oropharynx cancers across the cumulative US population based on a model with HPV16 status



Concentration (Lorenz) curve of risk model for oropharynx cancer for 30-69 yr.



Concentration curve of the cumulative proportion of oropharynx cancers across the cumulative US population based on a model that includes demographic and behavioral information



Discrimination of incremental risk models in internal and external validation

Model variables	Internal Validation Dataset	External Validation Dataset
	AUC (95% CI)	AUC (95% CI)
Demographic risk factors		
Age	0.76 (0.72-0.79)	0.74 (0.70-0.78)
Age, gender	0.83 (0.80-0.86)	0.81 (0.77-0.84)
Age, gender, race	0.84 (0.81-0.87)	0.82 (0.79-0.86)
Demographic and behavioral risk factors		
Age, gender, race, smoking	0.84 (0.81-0.88)	0.82 (0.79-0.86)
Age, gender, race, smoking, alcohol	0.84 (0.81-0.88)	0.83 (0.79-0.87)
Age, gender, race, smoking, alcohol, sex partners	0.86 (0.84-0.89)	0.81 (0.77-0.86)
Demographic/behavioral risk factors and oral HPV		
Age, gender, race, smoking, alcohol, sex partners, HPV	0.94 (0.92-0.97)	0.87 (0.84-0.90)
Age, gender, race, smoking, alcohol, sex partners, HPV16/other HPV	0.95 (0.92-0.97)	0.87 (0.84-0.90)

Calibration of across models and percentage of oropharynx cancers in highest decile of predicted risk in the population in internal and external validation

	Internal validation			External validation		
Characteristic	**Primary model**	**Model that separately considers HPV16 status**	**Model with only demographic and behavioral factors**	**Primary model**	**Model that separately considers HPV16 status**	**Model with only demographic and behavioral factors**
Overall O/E (95% CI)	1.01 (0.70-1.32)	0.94 (0.53-1.35)	0.95 (0.68-1.23)	1.08 (0.77-1.39)	1.01 (0.58-1.44)	1.02 (0.74-1.29)
O/E in highest decile of predicted risk (95% CI)	1.05 (0.67-1.44)	0.95 (0.42-1.47)	1.04 (0.63-1.45)	0.91 (0.57-1.25)	0.80 (0.35-1.26)	0.99 (0.60-1.38)
Percentage of cancers in the highest decile of predicted risk	76.5%	77.4%	53.7%	61.7%	63.1%	48.0%

Summary/Future Work

- Combining nonprobability clinical case series of OPC post-stratified (propensity weighted) to SEER-18 and NHANES controls to form a "synthetic" case-control study useful for estimating 1 yr. risk of OPC for U.S.
- Primary OPC risk model can improve efficiency in designing natural history/prevention studies.
  - 77% of cases captured in top risk decile
  - Number Needed to Recruit to get one case may be reduced from 12,195 (random selection) to 1,586 (targeting top 10% of risk) or 323 (targeting top 1%)
- Consider using North American Association of Central Cancer Registries (NAACCR) covering all US states in place of SEER-18 for estimating propensity wts.
- Only used age, race and sex to estimate propensity weights will examine miss-specification of propensity model.
- Study other ways to perform cross-validation for complex samples.

Acknowledgements

JE Tota, NCI

ML Gillison, MD Anderson

HA Katki, NCI

L Kahle, IMS

RK Pickard, OSU

W Xiao, MD Anderson

B Jiang, MD Anderson

AK Chaturvedi, NCI

Thank you