Logistic regression model-building strategies for predicting regular mammography screening adherence among uninsured immigrant women

Jesse A. Canchola, University of California, San Francisco & California State University, Hayward
Regina Otero-Sabogal, University of California, San Francisco
Desiree Owens, University of California, San Francisco
Farzaneh Tabnak, California State Department of Health Services

ABSTRACT

While one-time mammography screening rates have improved, repeat-screening rates have not followed this trend. Little is known about those factors that help sustain the regular use of mammography over time, especially among low-income, uninsured immigrant women. We utilized model-building strategies using the SAS LOGISTIC procedure to find a set of candidate logistic regression models that included a candidate set of predictors for repeat-screening mammography among 562 uninsured immigrant women who participated in the California No-Cost Screening Program (CANCSP). Rationale for the strategies and procedures used is given and limitations are discussed.

INTRODUCTION

For the uninitiated analyst, statistical model-building in the logistic regression context can be a challenging experience on the one hand or yield poor or inappropriate models at worst. In investigating mammography repeated screening for low-income immigrant women, we examine a robust model-building methodology that presents the investigator with rich models from which to choose that might best explain the phenomena at hand. By repeat mammography screening, we mean regular mammography screening according to set governmental guidelines.

In this section we give a brief overview of these model-building strategies. In the next section, we proceed in step using the CANCSP data, while giving some background, to find candidate models from which the investigator can choose. Finally, we exhibit results and limitations of the methodology.

MODEL-BUILDING STRATEGIES/TECHNIQUES

A general model-building strategy for the analyst might be portrayed as follows:

Steps

1. Choose the substantive/important variables and interactions;
2. Choose additional variables thought to be important (e.g., demographic, other auxiliary); Choose possible interactions.
3. Examine chosen variables via univariate and bi-variate analyses (i.e., frequencies, cross-tabulating categorical variables with the outcome as well as compare continuous predictors with the outcome and obtain significance tests; e.g., chi-square, t-test/nonparametric Wilcoxon rank-sum).
4. Screen/edit chosen variables based on distributions in *Step 3*. Modify variables as necessary;
5. Fit the logistic regression model with initial variables chosen in *Step 4*. Examine general model and specific variable characteristics;
6. Use a backward elimination method in your logistic regression algorithm with variables from *Steps 1-3*. Note candidate models;
7. Use a forward addition method in your logistic regression algorithm with variables from *Steps 1-3*. Note candidate models;
8. Use a stepwise approach in your logistic regression algorithm with variables from *Steps 1-3*. Note candidate models;
9. If all predictors are continuous, use a best subsets scoring method approach in your logistic regression algorithm with variables from *Steps 1-3*. Note candidate models;
10. Examine and array "best" models within *Steps 5-9*, above, using, say, the Hosmer-Lemeshow Goodness-Of-Fit (HL-GOF) as one criterion of fit and the Akaike Information Criterion (AIC). Check continuous variables for linearity in the logit (especially if you get poor fit);
11. Compare all models in *Step 10* on various fit indices and present to investigator to choose one that best explains the data.

Step 1. Typically, when studies are done, the investigator chooses important variables and interactions which she posits will answer the primary (and secondary) aims of the study then collects data on those variables. For example, in the present study, the investigator is interested in finding predictors of whether women will return for a doctor-recommended mammography screening within a certain period of time (10-18 months post initial screen). In addition, some interactions with ethnicity are posited. The variables chosen as the primary substantive variables and interactions are given in Appendix A, under the headings, "Primary Vars for Main Hypotheses" and "Posited Interactions".

Step 2. Additional variables thought to be related to the outcome can then be chosen by the investigator for input into *Step 3* and, if viable, into the modeling process.

Step 3. The chosen variables from *Steps 1-2* are then examined in a series of bi-variate tests with the outcome. All variables in *Step 1* are included in the logistic regression model. Eventually, only variables at the 0.25 level of significance should be included from *Step 2 and Step 4*. This inclusion level is chosen as the most conservative value in practice and allows each variable,

though not significant at the usual 0.05 significance level bi-variately, to "become" significant in the modeling process—jointly so, with other variables in the model.

Step 4. One often overlooked data quality procedure is data variable screening or editing. Data screening/editing, in this context, refers to the process of examining variables *univariately* for (1) out of range values, (2) logically inconsistent values, (3) data sparseness, (4) outliers and with that (5) skewness (kurtosis), and (6) *bivariately* to detect potential distributional problems in its relationship with the outcome as well as (7) high correlations (.70 or above) between predictor variables that might identify potential multicollinearity on the predictor side.

Although it may seem that this step can be placed as a first step in our methodology, we are assuming most of this data screening/editing actually has been done as a general part of the data quality process at the end of data collection and is not specific to any analysis, as we use it in the present context. In the current framework, data screening/editing becomes important to understanding the "behavior" of *bivariate* representations of those same variables with the outcome of study. For example, if we find that the frequency distribution of ethnicity (White/Caucasian, African-American, Hispanic, Asian, Other) is sparse on the "Other" category, we will see this in the data screening/editing stage. This will allow the investigator to make a decision as to what to do with this category (e.g., drop it if not too many subjects or fold it into another category if that category is not different than the "Other" category with the outcome.) This will prevent estimation problems "down the road" when model-building begins (and having to re-do many time-consuming analyses if caught later!)

Data screening/editing may be accomplished using a combination of one-way frequency tables and measures of central tendency and variability for continuous variables (e.g., means, medians, standard deviations, skewness, and kurtosis), as well as various graphical methods such as histograms and box-and-whisker plots. Additionally, it is often helpful to screen data on a bivariate basis prior to performing statistical modeling. For categorical outcomes and predictors, bivariate data screening/editing can often identify zero cell problems that can lead to improper solutions in logistic regression models. One commonly used practice of outlier remediation is categorization of variables containing outliers.

Step 5. Next, fit the logistic regression model with the initial variables chosen in the previous steps. This step is crucial since the analyst will examine the general model results as well as the results for the specific parameter estimates. The analyst should note the (1) overall model fit indices and (2) individual variable parameter estimates and their contribution to the general model. This is to note if the overall model is significant (omnibus test) and a simultaneous determination of the initial statistical importance of variables in the model-fitting process.

Steps 6-9. The model-building process begins with backward elimination, forward addition, and stepwise approaches for *Steps 6* through *8*, respectively. If all predictors are continuous, the analyst can use, in addition, a best subsets approach using the score method (*Step 9*). The analyst is cautioned, however, not to be complacent and allow these automatic processes to work entirely on their own. The dangers of letting the computer algorithms decide what is important or significant are well known (Flack and Chang, 1987; Freedman, 1983; Hauck, 1991). It is the analyst's responsibility to carefully follow the results of each step of the variable selection process to determine if a better model may be indicated from the intermediates steps (as might frequently be the case from the first author's experiences).

Step 10. Array the "best" models within *Steps 6-9* above in a table such as that found in Table 1, below. Note that in Table 1, only one model appears for each model due to space limitation reasons. In practice, one might have two or more models for each model-building process. Include fit indices such as the Hosmer-Lemeshow Goodness-Of-Fit (HL-GOF) as one (of many) criterion of fit. Make sure to check continuous variables for linearity in the logit (especially important if you get poor fit; see Hosmer and Lemeshow, 1989, page 89). If a parsimonious (i.e., simple) model is desired, the analyst may re-run the model-building process *Steps 5-9* using the variables in Table 1 where: (1) all variables appear in every final model (e.g., in Table 1: Var3 and Var5), plus (2) variables that appear in a majority or higher of the models (e.g., in Table 1, if the variable appears in, say, two out of the three models: Var1, and Var2.)

Table 1. Results of model-building process for outcome

Model 1: Backward	Model 2: Forward	Model 3: Stepwise
HL-GOF[1]=.73	HL-GOF[1]=.65	HL-GOF[1]=.23
Var1	Var1	
Var2		Var2
Var3	Var3	Var3
	Var4	
Var5	Var5	Var5

1 HL-GOF=Hosmer-Lemeshow Goodness-of-Fit test

Step 11. Compare all models in *Step 10* on various fit indices (both numerical and graphical) and present to investigator as candidate models to choose from. For simplicity and space limitations, we use the Hosmer-Lemeshow Goodness-of-Fit test (HL-GOF; Hosmer and Lemeshow, 1989). In practice, this test indicates a good model fit if its p-value $\geq$ .20 (closer to 1.0 is better.) The HL-GOF has two forms and has been shown to have sufficient power with sample sizes of about 400 or more.

In Table 1, above, it can be seen that the best-fitting model is Model 1 with an HL-GOF of 0.73 and not far behind is Model 2 with an HL-GOF=0.65. However, though Model 3 has an HL-GOF of 0.23, it should not be discounted as a viable candidate model since its p-value is above the 0.20 standard cutoff. If Model 3 had a p-value of, say, less than 0.10, we might discount this model because of the poor fit or investigate its constituent variable properties to achieve a better fit.

METHODS & RESULTS

Background of the CANCSP Data

Data for the CANCSP were collected in 1996 and 1997 from 102 final clinic sites in California and from women 50

years of age or older receiving services at those sites (n=1,050 women.) To reduce survey costs, the data were first stratified into high and low density strata and then sites randomly chosen (as clusters) within those strata. Probability of selection weights were constructed to reflect the stratified-clustered sampling design.

Motivation

Periodic mammography screening has a significant role in reducing mortality rates if utilized in the earlier stages. Even though one-time mammography use has increased over the past decade, repeat-screening mammography rates have not followed this trend. This is particularly true among low-income immigrant populations who have a disproportionate number of breast cancer related deaths. While repeat-screening mammography in vulnerable populations has been indicated as a national priority, there is limited information available on factors associated with regular mammography screening adherence among uninsured, low income, immigrant populations. In fact, scant research has examined the relationship between the needs of uninsured immigrant women and health care system factors that have the potential to improve repeat-screening mammography.

Cost and lack of insurance have repeatedly been cited as primary deterrents to regular breast cancer screening and as a reason for tumor stage difference at detection. However, simply having access to free services is not enough to eliminate barriers to regular mammography screening. This study explores individual, attitudinal, clinician, and health care system factors that may further explain repeat screening mammography adherence among uninsured immigrant women served by a state-wide no-cost-to-the-patient mammography screening program. For a more detailed discussion of barriers to repeat-screening mammography in the literature, see, for example, (Sabogal et al., 2003.)

We next apply the methodology discussed in the *Introduction*, above. Variables in *Steps 1-3* are exhibited in Appendix A.

Step 1. The substantive variables chosen by the investigator are measures of (1) acculturation [ACULT4C; =less acculturation, 1=more acculturation], (2) access to healthcare [ACCHLTH; Range: 0-8; higher values mean more access], (3) decisional balance [DECBAL; -6 to 11; higher values mean more pros than cons], (4) intensity of the intervention [INTNSITY; Range: 3-15; higher values means more intense intervention], (5) doctor patient communication [MDPATCOM; Range: 3-7; higher values indicate more doctor/patient communication], and measures of whether the patient (6) got a mammogram because of a history of breast problems [R_H4B; 1=yes, 2=no], (7) imputed version of patient got mammogram because the doctor or nurse recommended it [R_H4FIMP: 1=yes, 2=no], (8) gets a Pap test following the recommended guidelines [R_M9PAPC; 1=yes, 0=no], and whether the clinic site has (9) an information or advice telephone line [A22G; 1=yes, 0=no] or (10) has in-reach services [A25; 1=yes, 0=no].

Step 2. Relevant demographic variables are included in this step (see Appendix A under the "Demographic vars" heading). Also secondary variables thought to be important in predicting regular mammography screening are whether the patient (11) took hormone pills [R_M4; 1=yes, 0=no], (12) got their 1996 Pap test at the same time as their clinical breast exam (CBE) [R_M10R; 1=yes, 0=no], and the clinic site (13) percentage of White/Caucasian patients [A9B; Range: 0-84], (14) language fluency [A12_B; 1=≥50%, 0=<50%], (15) percent re-imbursement from Medi-Cal (all patients) [A8R; 1=≥50%, 0=<50%], and whether the patients, (16) got a mammogram at another location [MAMOLOC; 1=yes, 0=no].

Step 3. We use the FREQ procedure to examine the univariate distributions of all chosen variables, including the outcome. Due to space limitations, only variables that are "flagged" for further inspection are shown. The rest are given in the on-line supplementary material. Next, we use the FREQ and MEANS/NPAR1WAY procedures in SAS to look at crosstabulations of the chosen categorical variables with the outcome and compare means of continuous variables, respectively. The output is summarized in the on-line supplementary material.

Step 4. Note that in Table 1, the "Other" ethnicity category is sparse and cannot be combined with other categories. A decision is made by the principal investigator to exclude this category to reduce the likelihood of invalid parameter estimation results in the logistic regression analyses. This brings the sample size down to 581. In general, the analyst should seriously consider combining categories before resorting to excluding cases. In this case, the investigator chose to look at only the known ethnic categories. Additionally, it was decided to exclude 19 cases on the outcome with missing data. The final working sample size is then 562. The analyst should also consider multiple imputation methods as a tool to analyze the full data to compare with the results found with missing data excluded. This is to make sure no biases have been introduced by excluding any missing data at the outset.

Correlations between predictors were found to be less than 0.70. All other variables have been screened previously to their current states.

Table 1. Abbreviated CANCSP initial ethnic distribution

The FREQ Procedure Ethnicity

ETHNICTD	Frequency	Percent	Cumulative Frequency
Latina	418	69.44	418
Chinese	85	14.12	503
Filipina	78	12.96	581
Other	21	3.49	602

Frequency Missing = 3

Step 5. As a preliminary analysis of the omnibus model, the full model is run using the SAS LOGISTIC procedure, without regard to automatic procedures. Results are given in the supplementary material on-line. For all logistic regression runs, the probability modeled is regular adherence (versus non-regular adherence). This means that final interpretation of odds ratios will be in terms of regular adherence. For categorical predictors, the values used as the reference are given in Appendix A with an "(r)" next to the variable level.

Step 6. We used the "`/include=23 selection=backward slstay=0.05 stopres lackfit`" model options in the LOGISTIC procedure.

The "selection=backward" utilizes a backward elimination approach to model-building while the "include=23" keeps the first 23 terms in the model statement fixed while testing the posited interactions with ethnicity. In a subsequent model, we use the "`include=3`" option to keep the one interaction found and its constituent main effects. The "slstay=0.05" option specifies the significance level of the Wald chi-square for an effect to stay in the model in a backward elimination step. The "stopres" option specifies that the removal or entry of effects be based on the value of the residual chi-square. Finally, the "lackfit" option asks that the Hosmer-Lemeshow Goodness-of-Fit index to be output. After analyzing the backward elimination procedure, the final model is shown in Table 2 under the "Model 1" heading. Note that the interaction between ethnicity and acculturation is statistically significant at the 0.05 significance level (p=0.048). The HL-GOF also indicates good fit at p=0.38.

Table 2. Results of model-building process for regular screening mammography [1=regular vs. 0=non-regular]; probability modeled: adhere=regular.

Model 1: Backward	**Model 2: Forward**	**Model 3: Stepwise**
Fit Indices	**Fit Indices**	**Fit Indices**
HL-GOF=0.38[1]	HL-GOF=0.38[1]	HL-GOF=0.40[1]
AIC=551.4[2]	AIC=551.4[2]	AIC=556.4[2]
Variables	**Variables**	**Variables**
ETHNICTD [E]	ETHNICTD [E]	ETHNICTD [E]
ACULT4C [A]	ACULT4C [A]	ACULT4C [A]
ExA	ExA	ExA
YRSINUSD	YRSINUSD	YRSINUSD
DECBAL	DECBAL	DECBAL
ACCHLTH	ACCHLTH	ACCHLTH
R_M9PAPC	R_M9PAPC	R_M9PAPC
		A22G
A25	A25	A25

1 p-value for Hosmer-Lemeshow Goodness-of-Fit test. Closer to 1 is better. Good fitting models have p-values $\geq$ 0.20.

2 Akaike Information Criterion value. Smaller is better.

Step 7. Here, we used the "`/include=23 selection=forward slentry=0.05 stopres lackfit`" model options in the LOGISTIC procedure. The "`selection=forward`" option utilizes a forward addition approach to model-building. The "slentry=0.05" option specifies the significance level of the score chi-square for entering an effect into the model in the "selection=forward" or "selection=stepwise" methods. In a subsequent model, we use the "`include=3`" option to keep the one interaction found and its constituent main effects. Other options are as indicated in *Step 6*. The final model is shown in Table 2 under the "Model 2" heading. Note that the final model is the same as that found in *Step 6*.

Step 8. Here, we used the "`/include=23 selection=stepwise slentry=0.05 lackfit`" model options in the LOGISTIC procedure. The "`selection=stepwise`" option utilizes a stepwise approach to model-building. As in *Step 6 and 7*, in a subsequent model, we use the "`include=3`" option to keep the one interaction found and its constituent main effects. The final model is shown in Table 2 under the "Model 3" heading. The final model contains nine terms that were also found in the models from *Steps 6 and 7*, except for A22G (site has phone/info advice line). This model is slightly better as measured by the HL-GOF (p=0.40). In addition, the Akaike Information Criterion (AIC) is greater than the AIC for the models in Steps 6 and 7 (smaller AIC are better).

Step 9. Unfortunately, since we have a mix of categorical and continuous, we cannot use this step to help us with our model-building process.

Step 10. We array the model results in Table 2 and use selected fit indices to help us choose candidate models.

Step 11. It appears from the arrayed models in Table 2 that both the backward elimination and forward addition models converge to the same model with good fit as determined by the HL-GOF (p=0.38). Compare this to the stepwise procedure which gives a final model that has an additional term but produces a slightly better model (p=0.40). Effectively there are two models to present to the investigator that are close in fit to each other (Models 1 and 2 versus Model 3). Upon presentation and explanation, the investigator chooses the simpler one (Models 1 and 2). The final model characteristics are shown below in Table 3 where odds ratios are exhibited using the LOGISTIC model option, "`/clodds=both`". This option requests two types of confidence intervals for the odds ratios. Only the Wald intervals are shown.

Table 3. Final logistic regression model results (abbreviated).

```
    MODEL 1C: FITTING REDUCED MODEL
          BACKWARD ELIMINATION
        The LOGISTIC Procedure

            Model Information

Data Set                    WORK.ETHNICTD1
Response Variable           ADHERE
Deri. Adherent clients (J2,J3)
Number of Response Levels   2
Number of Observations      469
Weight Variable             PWT2
Sum of Weights              469.24935709
Model                       binary logit
Optimization Technique      Fisher's scoring

               Response Profile

Ordered                 Total       Total
  Value   ADHERE  Frequency      Weight
      1  regular        216   216.15675
      2  non-regular    253   253.09261

Probability modeled is ADHERE='regular'.

NOTE: 93 observations were deleted due to
      missing values for the response or
      explanatory variables.

             Model Fit Statistics
                                 Intercept
                Intercept          and
Criterion         Only         Covariates

AIC               649.607        551.350
SC                653.758        605.308
-2 Log L          647.607        525.350

  Testing Global Null Hypothesis: BETA=0
```

Test	Chi-Square	DF	Pr > ChiSq
LR	122.2569	12	<.0001
Score	108.2208	12	<.0001
Wald	87.5504	12	<.0001

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
ETHNICTD	2	9.7286	0.0077
ACULT4C	1	2.5208	0.1124
E*A	2	6.0724	0.0480
YRSINUSD	3	8.1136	0.0437
DECBAL	1	9.0809	0.0026
ACCHLTH	1	29.7345	<.0001
R_M9PAPC	1	42.2214	<.0001
A25	1	6.4030	0.0114

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
YRSINUSD 2 vs 1	1.239	0.681	2.253
YRSINUSD 3 vs 1	2.366	1.262	4.438
YRSINUSD 4 vs 1	1.555	0.792	3.053
DECBAL	1.153	1.051	1.264
ACCHLTH	1.340	1.206	1.489
R_M9PAPC 1 vs 0	0.238	0.154	0.367
A25 1 vs 0	1.736	1.132	2.662

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
8.5983	8	0.3773

DISCUSSION

Our final model found the following significant predictors for predicting regular screening among low-income, uninsured immigrant women living in California: (1) the number of years living in the U.S., (2) decisional balance, (3) access to healthcare, (4) whether the patient obtained a Pap test according to guidelines and (5) whether the site has done in-reach as a reminder to patients. Also, (6) a two-factor interaction was found between ethnicity and acculturation. This interaction and its constituent main effects are shown in Table 3 under the heading, "`Type III Analysis of Effects`" as E*A and ETHNICTD, ACULT4C, respectively.

Interpretation of the parameter estimates in terms of odds ratios is straightforward. These interpretations and implications to mammography rescreening in this population of women are given in a subsequent paper and will be made available with the on-line supplementary material.

CONCLUSIONS

In searching for predictors of mammography rescreening adherence in uninsured low-income immigrant women, we used some readily available model-building techniques in the SAS LOGISTIC procedure. We found logistic regression models that explained the relationships between predictors and outcome reasonably well.

To be more complete, the final analysis in the current example should be followed up with a final run using the GENMOD procedure to account for the complex sampling design and sampling weights. In general, the effect of these adjustments would be to adjust parameter standard errors upward so that borderline statistically significant effects may disappear, but nevertheless, may continue to be important predictors.

ACKNOWLEDGMENTS

We thank Alison J. Canchola at the Northern California Cancer Center, for her helpful comments and masterful editing of the penultimate manuscript. Any remaining errors remain ours.

Funding was provided by the California Department of Health Services, Cancer Detection Section, Contract #0090917.

Users may obtain the supplementary material on-line at http://www.hsru.org/wuss/wuss11.

CONTACT INFORMATION

Jesse A. Canchola, MS
Regina Otero-Sabogal, PhD and Desiree Owens, MS, MSW, LCSW
Institute for Health and Aging
74 New Montgomery Street, Suite 600
San Francisco, California 94105.
415.597.9354; 708.585.3747 (eFax)
e-mail (first author): jcanch@itsa.ucsf.edu

Farzaneh Tabnak, PhD
HIV/AIDS Epidemiology Branch,
California State Department of Health Services
Sacramento, California 95817

REFERENCES

Allison PD (1991). *Logistic Regression using the SAS System: Theory and Applications*. SAS Institute: Cary, NC. 288 pp.

Flack VF and Chang PC (1987). "Frequency of selecting noise variables in subset regression analyses: A simulation study", *The American Statistician.* 41, 84-86.

Freedman, D.A., (1983) "A note on screening regression equations." *The American Statistician.* 37, 152-155

Hauck, W.W. and Mike, R. (1991). "A proposal for examining and reporting stepwise regressions." *Statistics in Medicine*. 10, 711-715.

Hosmer DW, Lemeshow SL (1989). *Applied Logistic Regression*. Wiley: New York. 307 pp.

Menard S (1995). *Applied Logistic Regression Analysis*. Sage: Thousand Oaks. 98 pp.

Sabogal R, Owens D, Tabnak F, Canchola JA (2003). "Mammography adherence among multiethnic, low-income women: Individual, clinician and health care system factors". Submitted to *Journal of Health Care for the Poor and Underserved*.

SAS Institute Inc. (1995). *Logistic Regression Examples using the SAS® System, Version 6*. Cary, NC: SAS Institute Inc. 163 pp.

SAS Institute Inc. (2002). *SAS/STAT® Software: Changes and Enhancements, Release 8.2*, Cary, NC: SAS Institute Inc. 343 pp.

APPENDICES

Appendix A. Variables for CANCSP Study. Note: "(r)" indicates reference category for categorical variables.

Variable Name	Description
Primary Vars for Main Hypotheses	
ACULT4C [0(r)=less acculturated, 1=more acculturated]	Recoded acculturation
ACCHLTH [Range: 0-8; higher means more access]	Access to health scale
DECBAL [Range: -6 to 11; higher values mean more pros than cons]	Decisional balance scale
INTNSITY [Range: 3-15; higher values means more intense intervention]	A22a-k,a23a-e. Intensity of intervention scale
MDPATCOM [Range: 3-7; higher values indicate more doctor/patient communication]	Doctor/patient communication scale
R_H4B [1=yes, 2(r)=no]	Got mammogram because of history of breast problems
R_H4FIMP [1=yes, 2(r)=no]	IMP: Got mammogram because MD/RN recommended
R_M9PAPC [1=yes, 0(r)=no]	Pap test following the guidelines
A22G [1=yes, 0(r)=no]	phone info/advice line
A25 [1=yes, 0(r)=no]	site has done in-reach
Demographic Vars	
AGE2C [1=50-64, 2(r)=≥65]	Categorical age (years)
EDUCATN [0(r)=None, 1=1-6, 2=7-12, 3=13+]	Education level (years)
EMPLOY [1(r)=Work F/P time, 2=Unemployed/student, 3=Retired/disabled, 4=Full-time/homemaker]	Employment status
ETHNICTD [1(r)=Latina, 2=Chinese, 3=Filipina]	Recoded ethnicity
MARSTAT [1(r)=Married, 2=Sep/Div/Wid, 3=Never married]	Marital status
RELIGION [0(r)=<2/mo. past year, 1=2+/mo. past year]	Religion
YRSINUSD [1(r)=1-<10, 2=10-<20, 3=20-<30, 4=30+]	No. years living in U.S.
Secondary Vars from Step 4	
R_M4 [1=yes, 0(r)=no]	Take hormone pills
R_M10R [1=yes, 0(r)=no]	Got 1996 Pap test same time as CBE?
A9B [Range: 0-84]	% White/Caucasian Patients
A12_B [1=≥50%, 0(r)=<50%]	Language fluency
A8R [1=≥50%, 0(r)=<50%]	Recoded percent reimbursement from Medi-Cal(all patients)
MAMOLOC [1=yes, 0(r)=no]	Patients got mammogram at other location?
Posited Interactions	
Ethnicity	Interaction with
ETHNICTD	AGE2C
ETHNICTD	YRSINUSD
ETHNICTD	ACULT4C
ETHNICTD	MDPATCOM

Appendix B.

Model 1: Example of SAS LOGISTIC procedure code. See on-line supplementary material for more complete information.

```
/* Results of first model run showed
ETHNICTD*ACULT4C term significant */ ;

title1 "MODEL1B: REDUCED MODEL-TESTING MAIN
EFFECT TERMS NOT INVOLVED IN INTXN" ;
title2 "BACKWARD ELIMINATION" ;
proc format ;
  value adhere 1="regular" 0="non-regular" ;
run ;
proc logistic data = ETHNICTD1 descending ;
weight PWT2 ;
class  YRSINUSD (ref=first) EDUCATN
 (ref=first) AGE2C (ref=last) RELIGION
 (ref=first) A12_B (ref=first) R_M10R
 (ref=first) A22G (ref=first) A25 (ref=first)
 EMPLOY (ref=first) MARSTAT (ref=first) R_M4
 (ref=first) R_M9PAPC (ref=first)
 R_H4B (ref=last) A8R (ref=first) MAMOLOC
 (ref=first) R_H4FIMP (ref=last)
 ETHNICTD (ref=first) ACULT4C (ref=first) ;
model ADHERE = ETHNICTD ACULT4C
 ETHNICTD*ACULT4C YRSINUSD EDUCATN AGE2C
 RELIGION EMPLOY MARSTAT MDPATCOM DECBAL
 ACCHLTH A9B A12_B R_M10R A22G R_M4 R_M9PAPC
 A25 R_H4B A8R MAMOLOC INTNSITY R_H4FIMP
 /include=3 selection=backward slstay=0.05
  stopres lackfit ;
format adhere adhere. ;
run ;

title1 "MODEL 1C: FITTING REDUCED MODEL " ;
title2 "BACKWARD ELIMINATION" ;
proc logistic data = ETHNICTD1 descending ;
weight PWT2 ;
class  YRSINUSD (ref=first) EDUCATN
 (ref=first) A22G (ref=first) A25 (ref=first)
 R_M4 (ref=first) R_M9PAPC (ref=first)
 ETHNICTD (ref=first) ACULT4C (ref=first) ;
model ADHERE = ETHNICTD ACULT4C
 ETHNICTD*ACULT4C
 YRSINUSD DECBAL ACCHLTH R_M9PAPC A25
 /lackfit ;
format adhere adhere. ;
run ;
```