SUPPLEMENTAL MATERIAL for "How Well Do the Standard Body-Mass Index or Variations With A Different Exponent Predict Human Lifespan?"

AUTHORS: Dean Foster, Howard Karloff, Kenneth E. Shirley.

Dean Foster: Amazon, Inc., New York, NY, and Department of Statistics, University of Pennsylvania, Philadelphia, PA.

Howard Karloff: None.

Kenneth E. Shirley: None.

CONTACT INFO: Howard Karloff, howard@cc.gatech.edu

Supplemental Material

Models

We fit a series of Cox proportional hazards regression models to the right-censored times until death for men and women separately, using age as the underlying time variable (see [1] and [2] for a discussion of how to choose the time variable when fitting Cox proportional hazard models). Table 1 summarizes the seven models that we fit. We begin the discussion of the models by describing model M_2, and then we will describe the other models as they compare to M_2.

Model M_2 models the hazard function for participant i as $h_i(t) = h_0(t)\exp(\mathbf{X}_i\boldsymbol{\beta})$, where $h_0(t)$ is the baseline hazard function, t is the subject's age in years, $\mathbf{X}_i$ is the vector of predictor variables for participant i corresponding to the first 11 input variables listed in the "Data Set" subsection of the main paper and $\boldsymbol{\beta}$ is the vector of unknown coefficients that we wish to estimate.[1] Additionally, we used a cubic function of the input variable BMI_2, rather than just a linear function, to allow the model to capture the well-known "J-shaped" relationship between BMI_2 and mortality. Thus, there are 3 model degrees of freedom associated with BMI_2 in model M_2, and counting $J-1$ model degrees of freedom for each categorical input variable with J levels, and one degree of freedom for each continuous or binary input variable in the model, model M_2 contains 74 model degrees

[1] We will use the convention that the categorical *input* variable "Smoking," for example, which has 32 levels, results in 31 *predictor* (as opposed to *input*) variables being entered into the model, where 31 is also the number of model degrees of freedom associated with the input variable "Smoking." For another example, "Education" is an input variable with 8 levels, and $1\{\text{Education} = \text{College Degree}\}$ is one of 7 predictor variables associated with the input variable "Education."

of freedom. The maximum likelihood estimates of $\boldsymbol{\beta}$ for model M_2, for both men and women, are available here in the "Coefficients of model M_2" section.

To put model M_2 into context, we also fit model M_0, the null model with no predictors, and model M_1, a model with a cubic function of BMI_2 and the same additional input variables as were used in [3] (race, education, smoking status, alcohol consumption, and physical activity frequency) to the data. Note that model M_2 fits substantially better than model M_1 using only 13 additional model degrees of freedom.

Next, we investigated whether interaction effects between the input variables would improve the fit of the model. Model M_3 includes all of the predictor variables in model M_2, as well as the two-way interaction effects between all pairs of input variables, and quadratic terms for all variables except for diabetes (which is binary). To fit such a model using the existing categorical variables, however, would have resulted in a huge number of parameters in the model ($(7-1) \times (32-1) = 186$ variables, for example, for the interaction between the 7-level input variable "physical activity frequency" and the 32-level input variable "smoking"). To alleviate this problem, we "tied together" the parameters for the interactions to reduce the number of resulting model degrees of freedom. See the section on how we tied the parameters together for details. In short, instead of estimating 186 interaction effects for "smoking" $\times$ "physical activity frequency," for example, we employed a two-stage procedure:

- First, we computed the estimated linear effects of "smoking" and "physical activity frequency" for each participant from model M_2, effectively creating a single, continuous variable describing the effect of each of these categorical variables, denoted $X_i^{\text{total-smoking}}$ and $X_i^{\text{total-physical}}$ for each participant. These "tied together" variables are measured on the log-hazard scale, such that larger values are associated with a higher risk of death.

- Second, we estimated, in model M_3, a single interaction effect between these "tied together" variables.

We "tied together" the effects of each of the seven categorical variables in model M_2: race, education, smoking status, physical activity frequency, alcohol consumption, self-reported health

Supplemental Table 1: Summary of Models

	Description	df	LL (men)	LL (Women)
M0	Null	1	-6993.7	-4432.2
M1	Adams 2006 Variables	61	-1933.4	-1177.4
M2	All Variables	74	-241.6	-138.7
M3	All Interactions($\alpha = 2.0$)	145	-0.8	-0.2
M4	All Interactions($\alpha = 2.0$) except BMI x Height	144	-17.7	-4.4
M5	All Interactions($\alpha =$ optimal) except BMI x Height	144	-0.1	-0.0
M6	All Interactions($\alpha =$ optimal)	145	0.0	0.0

Here, the column "df" reports the *model* degrees of freedom, i.e., the number of predictors in the model, and "LL" is the log likelihood of the parameters of the model given the data, with M_6 used as the comparison model. See the main text for descriptions of the models.

status, and marital status, and estimated two-way interaction effects between all pairs.

Model M_4 is identical to model M_3 except that it omits the interaction between BMI_2 and height.

Results

The differences in likelihood ratios in Table 1 are all statistically significant, for men and women, except between models M_3, M_5, and M_6. This suggests several interesting findings:

1. The inclusion of interaction terms in model M_3, compared to using only main effects in model M_2, suggests that one's optimal BMI_2 depends on the values of his or her other covariates. Specifically, the estimated coefficients of the interaction effects between BMI_2 and other variables in model M_3 are:

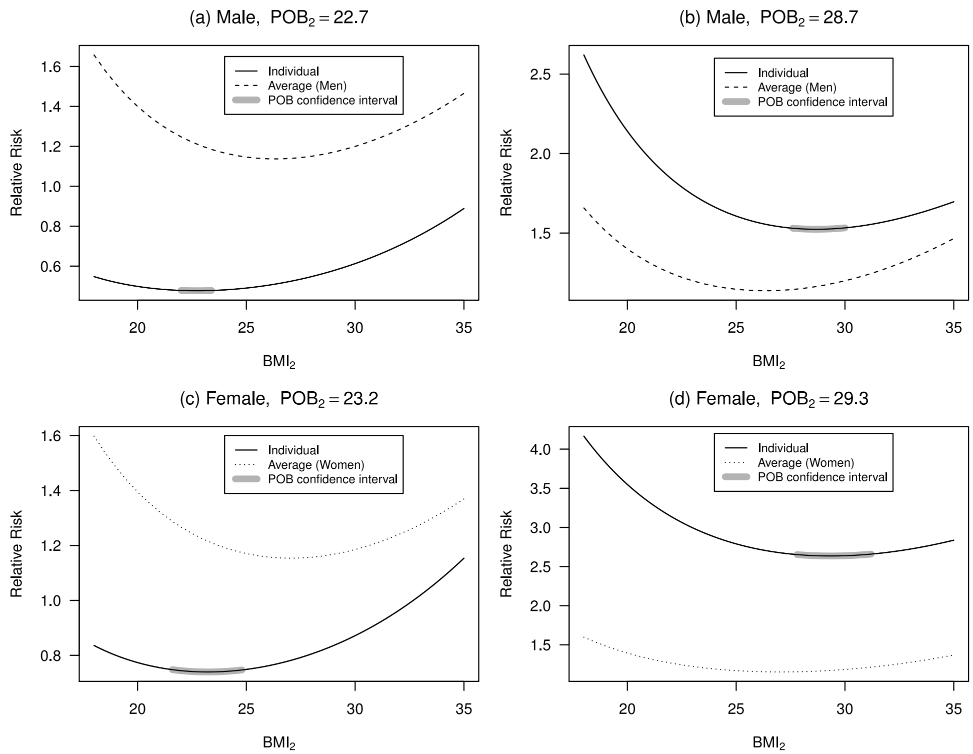
```
Men:                                      Women:
                        z Pr(>|z|)                                z Pr(>|z|)
bmi:age            -1.800    0.072       bmi:age              1.151    0.250
bmi:diabetes       -0.864    0.388       bmi:diabetes        -1.011    0.312
bmi:race           -2.013    0.044       bmi:race            -0.114    0.909
bmi:edu             1.162    0.245       bmi:edu              0.841    0.400
```

`bmi:smoking`	-6.155	0.000	`bmi:smoking`	-3.199	0.001
`bmi:physical`	-2.461	0.014	`bmi:physical`	-2.222	0.026
`bmi:alcohol`	-2.352	0.019	`bmi:alcohol`	-1.103	0.270
`bmi:health`	-5.823	0.000	`bmi:health`	-5.695	0.000
`bmi:marriage`	0.281	0.779	`bmi:marriage`	-0.490	0.624
`bmi:height`	5.814	0.000	`bmi:height`	2.927	0.003

Note that for men and women, every statistically significant interaction between BMI_2 and a "tied-together" categorical variable is negative. (Recall that the sign of a z statistic matches that of the corresponding coefficient.) This means that a person with a large value of BMI_2 and a large linear effect of smoking on mortality, for example, would have a negative interaction effect (on mortality)—his or her hazard rate is less than the product of the individual multipliers of the hazard rate from these two variables. Equivalently, there is a "bonus" effect of a lower hazard rate if a person has a low BMI_2 and also has a low linear effect of smoking, for example. The full sets of maximum likelihood estimates for model M_3 for men and women are in a separate subsection.

2. The fact that models M_3 and M_5 are statistically significantly better than model M_4 means that the association between BMI_2 and mortality is not independent of height. (This is also evident from the statistically significant effects of $\mathrm{BMI}_2 \times$ Height shown above.) This is illustrated in Figure 5 of the main paper, where we show that the relative risk as a function of BMI_2 under model M_4 is different for women of different heights (and is also different for men of different heights). In other words, the association between one's weight and his or her mortality is not fully accounted for by his or her BMI_2.

For illustrative purposes, Supplemental Figure 1 shows the estimated relative risk curve for four specific participants in our study under model M_3, compared to the the average relative risk curve across the population (stratified by sex). The curves are J-shaped, as has been found in previous studies, but their minima have different locations along the x-axis, depending on the demographic variables of the participant. These results motivate our computation and discussion



Supplemental Figure 1: The relative risk curves of four specific participants are plotted (in solid lines) compared to the average relative risk curves (in dashed or dotted lines, stratified by sex), as a function of BMI_2 for values between 18 and 35 under model M_3. The full set of demographic variables for each of these participants appears in a separate subsection. The main factors that push the minimum point of the relative risk curves in plots (b) and (d) to the right are relatively poor self-reported health and little physical activity.

of "Personalized Optimal BMI's" in the main body of the paper.

Details of Data Cleaning

The data contained 566,398 survey responses in its original form. We removed respondents according to the following steps:

1. We removed 16,987 respondents with extreme values for height or weight, defined as height less than 1.4 meters or greater than 2.1 meters, and weight less than 31.8 kg or more than 181.8 kg.

2. We removed 3,180 respondents with extreme values of caloric consumption, defined as fewer than 200 calories or more than 6000 calories consumed per day.

3. We removed 4,099 respondents who reported heavy alcohol consumption, defined as more than 200 grams per day.
4. We removed 1,455 respondents with extreme values of BMI_2, defined as BMI_2 less than 15 or greater than 50.
5. We removed 133,178 respondents who were chronically ill, defined as having one of either cancer, heart disease, renal disease, emphysema, or stroke.

The remaining data set of "participants" consisted of 407,499 respondents, 235,546 men and 171,953 women.

Coefficients of Model M_2

The estimated coefficients from model M_2 are listed below. The variables BMI (which is BMI_2), age, and height, were standardized to have a mean of zero and a standard deviation of 1. The variable alcohol was treated as categorical with 9 levels, where level 0 implied zero drinks/day, level j indicates the number of drinks/day was in $(j-1, j]$ for $j = 1, 2, ..., 7$, and level 8 indicates that the person consumed 7 or more drinks per day.

Model M2: Men

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
BMI	0.02	1.02	0.01	3.26	0.00
BMI^2	0.09	1.09	0.01	14.87	0.00
BMI^3	-0.01	0.99	0.00	-7.54	0.00
race2=black	0.05	1.06	0.03	1.88	0.06
race3=hispanic	-0.22	0.81	0.04	-5.31	0.00
race4=asian	-0.18	0.84	0.05	-3.53	0.00
race5=pacificislander	-0.12	0.88	0.16	-0.76	0.45
race5=unknown	0.03	1.03	0.05	0.65	0.52
race6=nativeamerican	0.07	1.07	0.10	0.67	0.50
edu2=8-11years	0.02	1.02	0.05	0.37	0.71

edu3=highschool	-0.04	0.96	0.05	-0.77	0.44
edu4=vocation/techschool	-0.06	0.94	0.05	-1.23	0.22
edu5=somecollege	-0.05	0.95	0.05	-1.09	0.28
edu6=collegegrad	-0.12	0.88	0.05	-2.60	0.01
edu7=postgrad	-0.19	0.83	0.05	-3.97	0.00
edu9=unknown	0.05	1.05	0.05	0.94	0.34
smoking02=quit10+dose1-10	0.06	1.06	0.02	2.79	0.01
smoking03=quit10+dose11-20	0.15	1.17	0.02	8.35	0.00
smoking04=quit10+dose21-30	0.21	1.23	0.02	9.79	0.00
smoking05=quit10+dose31-40	0.31	1.36	0.02	12.77	0.00
smoking06=quit10+dose41-60	0.36	1.44	0.03	13.41	0.00
smoking07=quit10+dose60+	0.43	1.53	0.04	9.64	0.00
smoking08=quit5-9dose1-10	0.32	1.37	0.06	5.61	0.00
smoking09=quit5-9dose11-20	0.49	1.64	0.04	13.30	0.00
smoking10=quit5-9dose21-30	0.59	1.81	0.04	15.34	0.00
smoking11=quit5-9dose31-40	0.68	1.97	0.04	16.36	0.00
smoking12=quit5-9dose41-60	0.55	1.73	0.05	10.40	0.00
smoking13=quit5-9dose60+	0.68	1.97	0.09	7.35	0.00
smoking14=quit1-4dose1-10	0.49	1.63	0.07	6.65	0.00
smoking15=quit1-4dose11-20	0.55	1.73	0.05	11.48	0.00
smoking16=quit1-4dose21-30	0.74	2.11	0.05	15.25	0.00
smoking17=quit1-4dose31-40	0.78	2.18	0.06	13.59	0.00
smoking18=quit1-4dose41-60	0.87	2.39	0.07	12.16	0.00
smoking19=quit1-4dose60+	0.92	2.51	0.13	7.04	0.00
smoking20=quit<1dose1-10	0.71	2.03	0.08	9.03	0.00
smoking21=quit<1dose11-20	0.80	2.21	0.06	13.70	0.00
smoking22=quit<1dose21-30	0.92	2.51	0.07	12.56	0.00
smoking23=quit<1dose31-40	1.22	3.39	0.09	13.53	0.00

smoking24=quit<1dose41-60	1.00	2.73	0.14	7.28	0.00
smoking25=quit<1dose60+	0.71	2.04	0.41	1.75	0.08
smoking26=currentdose1-10	0.77	2.16	0.03	24.84	0.00
smoking27=currentdose11-20	0.99	2.68	0.02	43.10	0.00
smoking28=currentdose21-30	1.13	3.11	0.03	43.09	0.00
smoking29=currentdose31-40	1.29	3.63	0.03	40.41	0.00
smoking30=currentdose41-60	1.31	3.70	0.05	24.03	0.00
smoking31=currentdose60+	1.25	3.49	0.14	9.13	0.00
smoking32=unknown/missing	0.36	1.43	0.03	13.13	0.00
physical2=rarely	-0.11	0.89	0.03	-4.18	0.00
physical3=1-3permonth	-0.23	0.80	0.03	-8.45	0.00
physical4=1-2perweek	-0.25	0.78	0.03	-9.65	0.00
physical5=3-4perweek	-0.31	0.74	0.03	-11.91	0.00
physical6=5+perweek	-0.31	0.74	0.03	-11.60	0.00
physical7=unknown/missing	-0.17	0.84	0.06	-2.96	0.00
alcohol1	-0.19	0.83	0.01	-14.55	0.00
alcohol2	-0.22	0.80	0.02	-11.94	0.00
alcohol3	-0.18	0.84	0.03	-6.80	0.00
alcohol4	-0.06	0.94	0.03	-2.20	0.03
alcohol5	-0.03	0.97	0.04	-0.88	0.38
alcohol6	0.01	1.01	0.05	0.21	0.84
alcohol7	-0.05	0.95	0.04	-1.29	0.20
alcohol8	0.08	1.09	0.03	2.38	0.02
health2=verygood	0.17	1.18	0.02	9.98	0.00
health3=good	0.36	1.44	0.02	21.17	0.00
health4=fair	0.70	2.01	0.02	31.25	0.00
health5=poor	1.25	3.50	0.05	27.24	0.00
health6=unknown	0.52	1.68	0.04	12.31	0.00

marriage2=widowed	0.22	1.25	0.02	8.91	0.00
marriage3=divorced	0.22	1.24	0.02	11.44	0.00
marriage4=separated	0.19	1.21	0.05	3.84	0.00
marriage5=nevermarried	0.31	1.37	0.02	12.58	0.00
marriage6=unknown	0.05	1.05	0.06	0.73	0.47
diabetes=yes	0.47	1.60	0.02	31.02	0.00
age	-0.03	0.97	0.01	-3.08	0.00
height	0.08	1.08	0.01	10.03	0.00

Model M2: Women

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
BMI	0.01	1.01	0.01	0.77	0.44
BMI^2	0.07	1.07	0.01	11.86	0.00
BMI^3	-0.01	0.99	0.00	-5.00	0.00
race2=black	-0.02	0.98	0.03	-0.75	0.45
race3=hispanic	-0.19	0.83	0.06	-3.33	0.00
race4=asian	-0.28	0.76	0.08	-3.41	0.00
race5=pacificislander	-0.24	0.79	0.23	-1.04	0.30
race5=unknown	0.14	1.16	0.05	2.73	0.01
race6=nativeamerican	0.12	1.13	0.12	1.04	0.30
edu2=8-11years	0.05	1.05	0.08	0.60	0.55
edu3=highschool	0.04	1.04	0.08	0.46	0.65
edu4=vocation/techschool	0.02	1.02	0.08	0.23	0.82
edu5=somecollege	0.03	1.03	0.08	0.39	0.70
edu6=collegegrad	-0.01	0.99	0.08	-0.09	0.93
edu7=postgrad	-0.05	0.95	0.08	-0.63	0.53
edu9=unknown	0.09	1.09	0.09	1.03	0.30
smoking02=quit10+dose1-10	0.04	1.04	0.03	1.62	0.10
smoking03=quit10+dose11-20	0.21	1.24	0.03	7.16	0.00

smoking04=quit10+dose21-30	0.35	1.42	0.04	8.90	0.00
smoking05=quit10+dose31-40	0.42	1.52	0.05	8.52	0.00
smoking06=quit10+dose41-60	0.34	1.41	0.07	5.15	0.00
smoking07=quit10+dose60+	0.55	1.73	0.12	4.57	0.00
smoking08=quit5-9dose1-10	0.28	1.32	0.06	4.80	0.00
smoking09=quit5-9dose11-20	0.44	1.55	0.05	9.18	0.00
smoking10=quit5-9dose21-30	0.52	1.68	0.06	8.69	0.00
smoking11=quit5-9dose31-40	0.77	2.15	0.07	11.26	0.00
smoking12=quit5-9dose41-60	0.58	1.78	0.10	5.87	0.00
smoking13=quit5-9dose60+	0.99	2.69	0.17	5.74	0.00
smoking14=quit1-4dose1-10	0.34	1.40	0.07	4.68	0.00
smoking15=quit1-4dose11-20	0.58	1.79	0.05	10.62	0.00
smoking16=quit1-4dose21-30	0.71	2.04	0.07	10.19	0.00
smoking17=quit1-4dose31-40	0.81	2.25	0.09	8.84	0.00
smoking18=quit1-4dose41-60	0.54	1.71	0.15	3.52	0.00
smoking19=quit1-4dose60+	0.64	1.91	0.30	2.13	0.03
smoking20=quit<1dose1-10	0.56	1.75	0.08	6.78	0.00
smoking21=quit<1dose11-20	0.89	2.44	0.07	13.06	0.00
smoking22=quit<1dose21-30	1.03	2.79	0.11	9.72	0.00
smoking23=quit<1dose31-40	1.07	2.91	0.16	6.73	0.00
smoking24=quit<1dose41-60	1.02	2.77	0.27	3.80	0.00
smoking25=quit<1dose60+	0.44	1.55	0.71	0.62	0.54
smoking26=currentdose1-10	0.86	2.37	0.03	28.53	0.00
smoking27=currentdose11-20	1.12	3.06	0.02	45.59	0.00
smoking28=currentdose21-30	1.23	3.44	0.03	37.52	0.00
smoking29=currentdose31-40	1.39	4.01	0.05	28.82	0.00
smoking30=currentdose41-60	1.45	4.26	0.10	15.07	0.00
smoking31=currentdose60+	1.53	4.60	0.24	6.26	0.00

smoking32=unknown/missing	0.40	1.49	0.04	10.51	0.00
physical2=rarely	-0.11	0.90	0.03	-3.78	0.00
physical3=1-3permonth	-0.24	0.78	0.03	-8.03	0.00
physical4=1-2perweek	-0.26	0.77	0.03	-9.06	0.00
physical5=3-4perweek	-0.31	0.74	0.03	-10.66	0.00
physical6=5+perweek	-0.27	0.77	0.03	-8.62	0.00
physical7=unknown/missing	-0.18	0.84	0.07	-2.69	0.01
alcohol1	-0.18	0.84	0.02	-11.06	0.00
alcohol2	-0.11	0.90	0.03	-3.86	0.00
alcohol3	-0.06	0.94	0.05	-1.28	0.20
alcohol4	0.15	1.16	0.05	2.72	0.01
alcohol5	0.00	1.00	0.11	0.00	1.00
alcohol6	0.05	1.05	0.08	0.64	0.52
alcohol7	0.01	1.01	0.17	0.03	0.98
alcohol8	0.37	1.44	0.08	4.48	0.00
health2=verygood	0.12	1.13	0.02	4.99	0.00
health3=good	0.34	1.41	0.02	14.24	0.00
health4=fair	0.67	1.95	0.03	22.76	0.00
health5=poor	1.10	3.01	0.05	20.60	0.00
health6=unknown	0.52	1.69	0.05	10.26	0.00
marriage2=widowed	0.15	1.16	0.02	8.47	0.00
marriage3=divorced	0.12	1.13	0.02	6.61	0.00
marriage4=separated	0.10	1.11	0.06	1.76	0.08
marriage5=nevermarried	0.27	1.31	0.03	9.45	0.00
marriage6=unknown	0.20	1.22	0.07	2.83	0.00
diabetes=yes	0.62	1.85	0.02	27.60	0.00
age	-0.09	0.91	0.01	-7.38	0.00
height	0.05	1.05	0.01	4.67	0.00

How We Tied Together Parameters In Model M_3

In this section we will describe how we fit a quadratic response function to our data without incurring the huge parameter-count penalty that a näive fit would entail. We start with a total of 68 linear effects from 7 categorical variables: Smoking has a total of 31 linear parameters (32 levels); physical activity has 6 (7 levels); education has 7; race, 6; alcohol, 8; self-reported health, 5; and marital status, 5. We also have linear effects of four continuous variables—BMI, Age, Height, and Diabetes—which we will interact with the categorical variables. Thus there are $68 + 4 = 72$ parameters needed to describe all the linear effects of the categorical variables and the continuous variables. Adding parameters for the square and cube of BMI (which we will not interact with any other variables) results in 74 parameters, or model degrees of freedom, in model M_2, described in the first section of this document. The näive quadratic response function including pairwise interactions between all seven of the categorical variables and four continuous variables would then end up adding a total of $72 + \binom{72}{2} = 72 + 2556 = 2628$ more parameters, all of which would need to be estimated. We will reduce the extraordinary number of parameters down to just $11 + \binom{11}{2} = 66$ parameters. The way we will do this is by tying the values of many of these parameters together.

We will focus on the smoking $\times$ education interaction. All the others are modeled in a similar fashion. For $i = 1, \ldots, 31$, let S_i be the 31 different possible levels of smoking, and likewise for $j = 1, \ldots, 7$, let E_j be the 7 possible levels of education. Then a simple linear model would look like

$$Y = \beta_{S_1} \times S_1 + \cdots + \beta_{S_{31}} \times S_{31} + \beta_{E_1} \times E_1 + \cdots + \beta_{E_7} \times E_7 + \cdots + \epsilon$$

(where we have left off many of the other variables along with the constant; all of these can be thought of as being hidden in the last "$\cdots$"). The full näive model would add $31 \cdot 7 = 217$ interaction terms which look like $\gamma_{S_i E_j} S_i E_j$. This would give us the model

$$Y = \sum_{i=1}^{31} \beta_{S_i} S_i + \sum_{j=1}^{7} \beta_{E_j} + \sum_{i=1}^{31} \sum_{j=1}^{7} \gamma_{S_i E_j} S_i E_j + \cdots + \epsilon$$

We will reduce this entire collection of interactions to a single parameter η_{SE} by tying the parameters together:

$$\gamma_{S_i E_j} = \eta_{SE} \beta_{S_i} \beta_{E_j}$$

This will reduce the number of parameters fit for interactions from 217 down to just 1. Our model then looks like:

$$Y = \sum_{i=1}^{31} \beta_{S_i} S_i + \sum_{j=1}^{7} \beta_{E_j} E_j + \eta_{SE} \left(\sum_{i=1}^{31} \beta_{S_i} S_i \right) \left(\sum_{j=1}^{7} \beta_{E_j} E_j \right) + \cdots + \epsilon.$$

We can think of the factor $\sum_{i=1}^{31} \beta_{S_i} S_i$ as being a total linear effect due to smoking, and $\sum_{j=1}^{7} \beta_{E_j} E_j$ as being the total linear effect due to education. Using that language, we think of our model with tied coefficients as simply adding a single interaction between the total effect for smoking and the total effect for education.

Estimating the parameters of this nonlinear model cannot be done using traditional least-squares code, but would require a more general optimizer to find the best fit. Since the "Y" given in our model is really a hazard rate for a Cox proportional hazard model, it is even more complex. We use instead a simple algorithm, using off-the-shelf code, to approximate the above model. First, estimate the model assuming the η terms are all zero. Then construct the variables total-smoking, total-education, etc., forming the linear combinations generated by this initial fit. Now we can estimate a standard Cox model which has

- total-smoking,
- total-education,
- total-education $\times$ total-education,
- total-education $\times$ total-smoking,
- total-smoking $\times$ total-smoking,
- etc.

Coefficients of Model M_3

The estimated coefficients from model M_3 are listed below. In the following tables, the seven categorical variables race, edu, smoking, physical, alcohol, health, and marriage are "tied-together" variables, representing the linear effects of these input variables, as described in the section on tying the variables together. The variables BMI, age, diabetes, and height are scaled to have a mean of zero and a standard deviation of one.

Model M3: Men

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
BMI	0.03	1.03	0.01	2.97	0.00
BMI^2	0.09	1.09	0.01	14.38	0.00
BMI^3	-0.01	0.99	0.00	-6.46	0.00
age	-0.01	0.99	0.01	-0.78	0.44
diabetes	0.56	1.76	0.03	22.25	0.00
race	1.04	2.82	0.38	2.70	0.01
edu	1.20	3.33	0.12	9.77	0.00
smoking	1.07	2.93	0.03	35.37	0.00
physical	0.97	2.63	0.13	7.18	0.00
alcohol	1.10	3.01	0.13	8.41	0.00
health	1.10	3.00	0.04	24.92	0.00
marriage	1.41	4.08	0.22	6.31	0.00
height	0.07	1.07	0.01	5.18	0.00
race^2	0.35	1.42	1.95	0.18	0.86
edu^2	1.12	3.06	1.11	1.00	0.32
smoking^2	0.01	1.01	0.04	0.29	0.77
physical^2	0.17	1.18	0.69	0.24	0.81
alcohol^2	-0.11	0.89	1.11	-0.10	0.92
health^2	0.20	1.22	0.06	3.11	0.00

marriage^2	-0.89	0.41	1.06	-0.84	0.40
age^2	-0.02	0.98	0.01	-2.13	0.03
height^2	0.01	1.01	0.01	1.52	0.13
BMI:age	-0.01	0.99	0.01	-1.80	0.07
BMI:diabetes	-0.01	0.99	0.01	-0.86	0.39
BMI:race	-0.32	0.73	0.16	-2.01	0.04
BMI:edu	0.10	1.11	0.09	1.16	0.25
BMI:smoking	-0.09	0.91	0.01	-6.16	0.00
BMI:physical	-0.15	0.86	0.06	-2.46	0.01
BMI:alcohol	-0.14	0.87	0.06	-2.35	0.02
BMI:health	-0.14	0.87	0.02	-5.82	0.00
BMI:marriage	0.02	1.02	0.05	0.28	0.78
BMI:height	0.04	1.04	0.01	5.81	0.00
age:diabetes	-0.08	0.92	0.02	-4.74	0.00
age:race	0.32	1.37	0.17	1.89	0.06
age:edu	-0.22	0.80	0.09	-2.43	0.02
age:smoking	-0.05	0.95	0.01	-3.31	0.00
age:physical	0.05	1.05	0.07	0.66	0.51
age:alcohol	-0.07	0.93	0.06	-1.14	0.26
age:health	-0.02	0.98	0.03	-0.56	0.57
age:marriage	-0.37	0.69	0.06	-6.30	0.00
age:height	0.02	1.02	0.01	1.98	0.05
diabetes:race	0.35	1.42	0.36	0.96	0.34
diabetes:edu	0.34	1.40	0.23	1.44	0.15
diabetes:smoking	-0.20	0.82	0.04	-4.84	0.00
diabetes:physical	0.21	1.23	0.17	1.24	0.21
diabetes:alcohol	-0.11	0.90	0.16	-0.67	0.50
diabetes:health	-0.17	0.85	0.06	-2.63	0.01

diabetes:marriage	0.08	1.09	0.16	0.52	0.60
diabetes:height	-0.02	0.98	0.02	-1.18	0.24
race:edu	2.79	16.32	2.27	1.23	0.22
race:smoking	0.97	2.63	0.46	2.12	0.03
race:physical	-1.41	0.24	1.59	-0.89	0.37
race:alcohol	0.46	1.59	1.62	0.28	0.78
race:health	-0.94	0.39	0.64	-1.48	0.14
race:marriage	-0.16	0.85	1.52	-0.11	0.92
race:height	-0.42	0.66	0.20	-2.11	0.03
edu:smoking	0.22	1.25	0.22	1.03	0.30
edu:physical	-0.72	0.49	1.00	-0.72	0.47
edu:alcohol	-1.00	0.37	0.87	-1.15	0.25
edu:health	-1.74	0.18	0.38	-4.60	0.00
edu:marriage	0.83	2.30	0.84	0.99	0.32
edu:height	-0.24	0.79	0.12	-2.04	0.04
smoking:physical	-0.35	0.71	0.16	-2.19	0.03
smoking:alcohol	0.13	1.14	0.14	0.89	0.37
smoking:health	-0.38	0.68	0.06	-6.07	0.00
smoking:marriage	-0.57	0.57	0.14	-4.20	0.00
smoking:height	-0.01	0.99	0.02	-0.61	0.54
physical:alcohol	0.03	1.03	0.66	0.04	0.97
physical:health	0.50	1.64	0.26	1.94	0.05
physical:marriage	0.41	1.51	0.61	0.68	0.50
physical:height	-0.04	0.96	0.09	-0.43	0.67
alcohol:health	-0.43	0.65	0.26	-1.67	0.10
alcohol:marriage	0.13	1.14	0.58	0.23	0.82
alcohol:height	-0.06	0.94	0.08	-0.71	0.48
health:marriage	-1.09	0.34	0.24	-4.50	0.00

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
health:height	-0.05	0.95	0.03	-1.52	0.13
marriage:height	0.07	1.07	0.08	0.87	0.39

Model M3: Women

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
BMI	0.05	1.05	0.01	4.03	0.00
BMI^2	0.07	1.07	0.01	11.77	0.00
BMI^3	-0.01	0.99	0.00	-4.34	0.00
age	-0.05	0.95	0.02	-3.04	0.00
diabetes	0.68	1.97	0.04	15.34	0.00
race	0.39	1.47	0.38	1.01	0.31
edu	0.96	2.62	0.39	2.49	0.01
smoking	1.02	2.77	0.04	25.60	0.00
physical	1.03	2.81	0.18	5.67	0.00
alcohol	0.94	2.55	0.15	6.39	0.00
health	0.95	2.58	0.07	14.41	0.00
marriage	1.01	2.76	0.16	6.52	0.00
height	0.09	1.09	0.02	4.84	0.00
race^2	-0.39	0.68	1.17	-0.33	0.74
edu^2	1.71	5.52	4.66	0.37	0.71
smoking^2	-0.01	0.99	0.05	-0.24	0.81
physical^2	0.20	1.22	0.94	0.21	0.84
alcohol^2	-0.10	0.90	0.47	-0.21	0.83
health^2	0.22	1.25	0.10	2.27	0.02
marriage^2	-0.06	0.94	0.98	-0.06	0.95
age^2	-0.01	0.99	0.01	-1.42	0.16
height^2	0.01	1.01	0.01	1.65	0.10
BMI:age	0.01	1.01	0.01	1.15	0.25
BMI:diabetes	-0.02	0.98	0.02	-1.01	0.31

BMI:race	-0.02	0.98	0.15	-0.11	0.91
BMI:edu	0.15	1.16	0.18	0.84	0.40
BMI:smoking	-0.04	0.96	0.01	-3.20	0.00
BMI:physical	-0.13	0.87	0.06	-2.22	0.03
BMI:alcohol	-0.07	0.93	0.06	-1.10	0.27
BMI:health	-0.14	0.87	0.03	-5.69	0.00
BMI:marriage	-0.03	0.97	0.07	-0.49	0.62
BMI:height	0.03	1.03	0.01	2.93	0.00
age:diabetes	-0.12	0.89	0.03	-4.60	0.00
age:race	0.27	1.31	0.19	1.41	0.16
age:edu	-0.94	0.39	0.23	-4.04	0.00
age:smoking	-0.02	0.98	0.02	-1.40	0.16
age:physical	0.04	1.04	0.09	0.41	0.68
age:alcohol	-0.31	0.73	0.08	-3.81	0.00
age:health	-0.08	0.93	0.03	-2.15	0.03
age:marriage	-0.15	0.86	0.09	-1.57	0.12
age:height	0.01	1.01	0.01	0.95	0.34
diabetes:race	0.13	1.14	0.48	0.27	0.79
diabetes:edu	0.22	1.24	0.70	0.31	0.76
diabetes:smoking	-0.28	0.75	0.05	-5.40	0.00
diabetes:physical	0.78	2.18	0.22	3.50	0.00
diabetes:alcohol	-0.03	0.97	0.25	-0.13	0.90
diabetes:health	-0.21	0.81	0.09	-2.20	0.03
diabetes:marriage	0.77	2.16	0.27	2.85	0.00
diabetes:height	-0.03	0.97	0.03	-1.03	0.30
race:edu	-2.48	0.08	4.80	-0.52	0.60
race:smoking	0.27	1.32	0.45	0.62	0.54
race:physical	4.19	66.28	1.88	2.24	0.03

race:alcohol	1.46	4.32	1.89	0.77	0.44
race:health	0.68	1.98	0.72	0.95	0.34
race:marriage	3.37	29.00	2.04	1.65	0.10
race:height	-0.20	0.82	0.26	-0.76	0.45
edu:smoking	0.25	1.28	0.49	0.51	0.61
edu:physical	-4.39	0.01	2.47	-1.78	0.08
edu:alcohol	2.68	14.60	2.23	1.20	0.23
edu:health	-0.91	0.40	0.98	-0.93	0.35
edu:marriage	1.99	7.35	2.46	0.81	0.42
edu:height	-0.30	0.74	0.32	-0.94	0.35
smoking:physical	-0.05	0.95	0.17	-0.33	0.74
smoking:alcohol	-0.13	0.88	0.16	-0.78	0.43
smoking:health	-0.30	0.74	0.07	-4.26	0.00
smoking:marriage	-0.15	0.86	0.19	-0.75	0.45
smoking:height	-0.05	0.95	0.02	-2.19	0.03
physical:alcohol	0.71	2.04	0.78	0.92	0.36
physical:health	0.33	1.39	0.32	1.04	0.30
physical:marriage	-0.42	0.66	0.93	-0.45	0.65
physical:height	0.21	1.24	0.11	1.90	0.06
alcohol:health	-0.26	0.77	0.34	-0.75	0.45
alcohol:marriage	0.60	1.82	0.89	0.67	0.50
alcohol:height	-0.15	0.86	0.11	-1.40	0.16
health:marriage	-0.25	0.78	0.39	-0.64	0.52
health:height	-0.10	0.91	0.05	-2.00	0.05
marriage:height	0.07	1.07	0.13	0.56	0.58

Data for Participants in Figure 1

	a	b	c	d

sex	male	male	female	female
age	61	64	55	54
height (m)	1.90	1.70	1.93	1.72
weight (lbs)	225	172	125	267
race	white	white	white	black
education	somecollege	collegegrad	postgrad	vocation/techschool
smoking	nonsmoker	nonsmoker	quit10+dose1-10	quit5-9dose11-20
physical	3-4perweek	never	1-2perweek	never
alcohol	2	1	3	0
health	excellent	fair	excellent	fair
marriage	married	married	divorced	married
diabetes	no	no	no	no
BMI	28.1	26.9	15.2	40.6

How to Derive POB_2's and Confidence Intervals for Them

Model M_3 is a Cox proportional hazards regression model that models a participant's risk of death as a function (of a specific form) of, among other things, his or her BMI_2, BMI_2^2, BMI_2^3, and the interaction between his or her BMI_2 and ten other variables, such as smoking status, race, education, etc.

The model is as follows, where $h_i(t)$ is the hazard function for participant i at time t (with t measured in years since birth, i.e., age, and where BMI is assumed to be BMI_2):

$$\log h_i(t) = \log h_0(t) + C_i + \beta_1 BMI_i + \beta_2 BMI_i^2 + \beta_3 BMI_i^3 \quad (1)$$

$$+ BMI_i \times (\beta_4 X_i^{\text{age-at-entry}} + \beta_5 X_i^{\text{race}} + \beta_6 X_i^{\text{edu}} \quad (2)$$

$$+ \beta_7 X_i^{\text{smoking}} + \beta_8 X_i^{\text{health}} + \beta_9 X_i^{\text{physical-activity}} \quad (3)$$

$$+ \beta_{10} X_i^{\text{diabetes}} + \beta_{11} X_i^{\text{alcohol}} + \beta_{12} X_i^{\text{marriage}} \quad (4)$$

$$+ \beta_{13} X_i^{\text{height}}). \quad (5)$$

Here, C_i denotes, for participant i, the sum of the linear, quadratic, and two-way interaction effects associated with the variables other than BMI_2 (smoking status, race, education, etc.).

We computed, for almost every participant, his or her *personalized optimal* BMI_2, or POB_2, which is the BMI_2 value that minimizes his or her relative risk according to the fitted model, and a confidence interval for the POB_2. In this section we explain how we did so.

We now denote the following value of BMI_2 for participant i to be his or her POB_2 (where we are suppressing the dependence of POB_2 on i):

$$POB_2 = \arg\min_x \Big(\hat\beta_3 x^3 + \hat\beta_2 x^2 + x \times (\hat\beta_1 + \hat\beta_4 X_i^{\text{age-at-entry}} + \hat\beta_5 X_i^{\text{race}} + \cdots + \hat\beta_{13} X_i^{\text{height}})\Big) \tag{6}$$

$$= \frac{-b + \sqrt{b^2 - 4ac_i}}{2a}, \tag{7}$$

where $a = 3\hat\beta_3$, $b = 2\hat\beta_2$, and $c_i = \hat\beta_1 + \hat\beta_4 X_i^{\text{age-at-entry}} + \hat\beta_5 X_i^{\text{race}} + \cdots + \hat\beta_{13} X_i^{\text{height}}$. (Note that $a < 0$, so this is the smaller root.) In other words, to minimize the cubic function of BMI_2, we take the derivative, which is a quadratic function of BMI_2, and set it to zero, and solve using the quadratic formula. If there is a local minimum, we set POB_2 to the local minimum. (This is the smaller root, if there are two.) If there is no local minimum (i.e., if $b^2 - 4ac_i < 0$, which happens for 79 men in our study and 5 women in our study), then we simply don't report a POB_2 for that individual.

Next, we define a confidence interval for a POB_2 as the set S of all values $z \in U$ such that we cannot reject the hypothesis that the derivative of the relative risk curve for individual i at z equals zero. In other words, defining $U := \{15.0, 15.1, 15.2, ..., 49.9, 50.0\}$, we want to test for participant i and all values of $z \in U$, the null hypothesis $H_{0,i}(z)$ against the alternative hypothesis, $H_{A,i}(z)$, where:

$$H_{0,i}(z): \ 3\beta_3 z^2 + 2\beta_2 z + \beta_1 + \beta_4 X_i^{\text{age-at-entry}} + \beta_5 X_i^{\text{race}} + \cdots + \beta_{13} X_i^{\text{height}} = 0. \tag{8}$$

$$H_{A,i}(z): \ 3\beta_3 z^2 + 2\beta_2 z + \beta_1 + \beta_4 X_i^{\text{age-at-entry}} + \beta_5 X_i^{\text{race}} + \cdots + \beta_{13} X_i^{\text{height}} \neq 0. \tag{9}$$

Another way to express this pair of hypotheses is as a linear function of the unknown regression coefficients β:

$$H_{0,i}(z) : d_{1,i}(z)\beta_1 + d_{2,i}(z)\beta_2 + \cdots + d_{13,i}(z)\beta_{13} = 0, \tag{10}$$

$$H_{A,i}(z) : d_{1,i}(z)\beta_1 + d_{2,i}(z)\beta_2 + \cdots + d_{13,i}(z)\beta_{13} \neq 0. \tag{11}$$

where the $d_{j,i}(z)$'s define a linear combination of the regression coefficients, and

$$d_{1,i}(z) = 1 \tag{12}$$

$$d_{2,i}(z) = 2z \tag{13}$$

$$d_{3i}(z) = 3z^2 \tag{14}$$

$$d_{4i}(z) = X_i^{\text{age-at-entry}} \tag{15}$$

$$d_{5i}(z) = X_i^{\text{race}} \tag{16}$$

$$\cdots \tag{17}$$

$$d_{13,i}(z) = X_i^{\text{height}}. \tag{18}$$

The test statistic for the hypothesis test is:

$$T_i^*(z) = \frac{d_{1,i}(z)\hat{\beta}_1 + d_{2,i}(z)\hat{\beta}_2 + \cdots + d_{13,i}(z)\hat{\beta}_{13}}{\text{se}(d_{1,i}(z)\hat{\beta}_1 + d_{2,i}(z)\hat{\beta}_2 + \cdots + d_{13,i}(z)\hat{\beta}_{13})} \sim t_{n-13-1}, \tag{19}$$

where

$$\text{se}(d_{1,i}(z)\hat{\beta}_1 + d_{2,i}(z)\hat{\beta}_2 + \cdots + d_{13,i}(z)\hat{\beta}_{13}) = \sqrt{\mathbf{d_i(z)}\Sigma\mathbf{d_i(z)}^T}, \tag{20}$$

where $\mathbf{d_i(z)}$ is a 1×13 matrix, and where Σ is the 13×13 covariance matrix of the estimated regression coefficients $\beta_1, \beta_2, ..., \beta_{13}$.

We perform this test for each participant $i = 1, ..., n$ (where recall that for men, $n = 235,546$, and for women, $n = 176,953$), and for a grid of values of $z \in U$, which is the observed range of BMI_2 in this sample (since we excluded those with BMI_2's outside this interval), for a total of 351 tests for each participant. The hypothesis test is two-sided, so we set the critical value of $T_i^*(z)$ to 1.96 for an $\alpha = 0.05$ significance level.

The result is that, for each respondent, we estimate the POB_2 as well as provide a confidence interval for it.

Let S be the set of all $z \in U$ such that we did not reject the hypothesis that the derivative of the relative risk curve is 0 at $\mathrm{BMI}_2 = z$.

There are five cases:

1. $S = \emptyset$. In this case, we assign no confidence interval (and there is no POB_2 in this case).

2. There is a $\gamma \in U$ such that $S = [\gamma, \infty] \cap U$. (Informally, S consists of one interval, "unbounded" on the right.) In this case, we do not define a confidence interval.

3. There are $\gamma_1, \gamma_2 \in U$ with $\gamma_1 \leq \gamma_2 \leq 49.9$ such that $S = [\gamma_1, \gamma_2] \cap U$. (Informally, S consists of one closed interval.) In this case, we assign the confidence interval I to be $[\gamma_1, \gamma_2]$.

4. There are $\gamma_1 \leq \gamma_2 \leq \gamma_3 \in U$ with $\gamma_3 \leq 49.9$ and $\gamma_2 \leq \gamma_3 - 0.2$ such that $S = ([\gamma_1, \gamma_2] \cup [\gamma_3, \infty]) \cap U$. (Informally, S consists of two disjoint intervals, the right one "unbounded.") Here again we assign the confidence interval I to be $[\gamma_1, \gamma_2]$.

5. There are $\gamma_1 \leq \gamma_2 \leq \gamma_3 \leq \gamma_4 \in U$ with $\gamma_2 \leq \gamma_3 - 0.2$ and $\gamma_4 \leq 49.9$ such that $S = ([\gamma_1, \gamma_2] \cup [\gamma_3, \gamma_4]) \cap U$. (Informally, S consists of two disjoint intervals, both bounded.) In this case once again we assign the confidence interval I to be $[\gamma_1, \gamma_2]$.

The reason we assign I to be the left interval in the last two cases is that in these cases, the right interval is the confidence interval for the *larger* root (which is a local maximum) of the quadratic.

There were precisely 2 people in the first case and 381 in the second, men and women combined. For all remaining people, a confidence interval was defined.

As mentioned earlier, for 79 men and five women, the POB_2 is undefined, because there is no local minimum of the relative risk curve for these respondents in $[15, 50]$. Furthermore, for fewer than 0.1% of the participants (across men and women), the confidence interval for one's POB_2 is not defined (cases 1 and 2 above), because their relative risk curves are nearly flat for BMI_2 values up to 50.

For these individuals, we provide no confidence intervals and we recommend further study.

Of the more than 99.9% of participants for whom we define confidence intervals (i.e., cases 3-5), the average interval width was about 2.5 for women and about 1.7 for men.

References

[1] Anne Thiébaut and Jacques Bénichou, "Choice of Time-Scale in Cox's Model Analysis of Epidemiologic Cohort Data: A Simulation Study," *Statistics in Medicine* 23, 2004, pp. 3803-3820.

[2] Prabhakar Chalise, Eric Chicken, and Daniel McGee, "Time Scales in Epidemiological Analysis: An Empirical Comparison," *Statistics in Medicine* 00, 2009, pp. 1-13.

[3] Kenneth Adams, Arthur Schatzkin, Tamara Harris, Victor Kipnis, Traci Mouw, Rachel Ballard-Barbash, et al., "Overweight, Obesity, and Mortality in a Large Prospective Cohort of Persons 50 to 71 Years Old," *The New England Journal of Medicine*, August 24, 2006, pp. 763-778.