PONTIFICIA
UNIVERSIDAD
CATÓLICA
DE CHILE

iibm
Instituto de Ingeniería Biológica y Médica

IBM3104.- Statistical Methods for Biological and Medical Engineering

María Rodríguez Fernández

marodriguezf@uc.cl

http://marodriguezf.sitios.ing.uc.cl/

PONTIFICIA
UNIVERSIDAD
CATÓLICA
DE CHILE

iibm
Instituto de Ingeniería Biológica y Médica

IBM3104: Statistical Methods for BME

UNIT 5: REGRESSION ANALYSIS

María Rodríguez Fernández

marodriguezf@uc.cl

http://marodriguezf.sitios.ing.uc.cl/

06. Logistic regression

07. Cox Regression

INTRODUCTION TO COX REGRESSION

Outcome Variable	**Are the observation groups independent or correlated?**		**Modifications if assumptions violated:**
	independent	correlated	
Time-to-event (e.g., time to fracture)	**Rate ratio** (2 groups) **Kaplan-Meier statistics** (2 or more groups) **Cox regression** (multivariate regression technique)	**Frailty model** (multivariate regression technique)	**Time-varying effects**

INTRODUCTION TO COX REGRESSION

- Also called proportional hazards regression
- Multivariate regression technique where time-to-event (taking into account censoring) is the dependent variable.
- Estimates adjusted hazard ratios:
  - A hazard ratio is a ratio of rates (hazard rates)

RECALL: HAZARD RATIOS

- A hazard ratio is similar to a rate ratio, but it is the ratio of instantaneous incidence rates
- Since hazard ratios come from a regression, they are usually multi-variable adjusted

RANOLAZINE VS PLACEBO

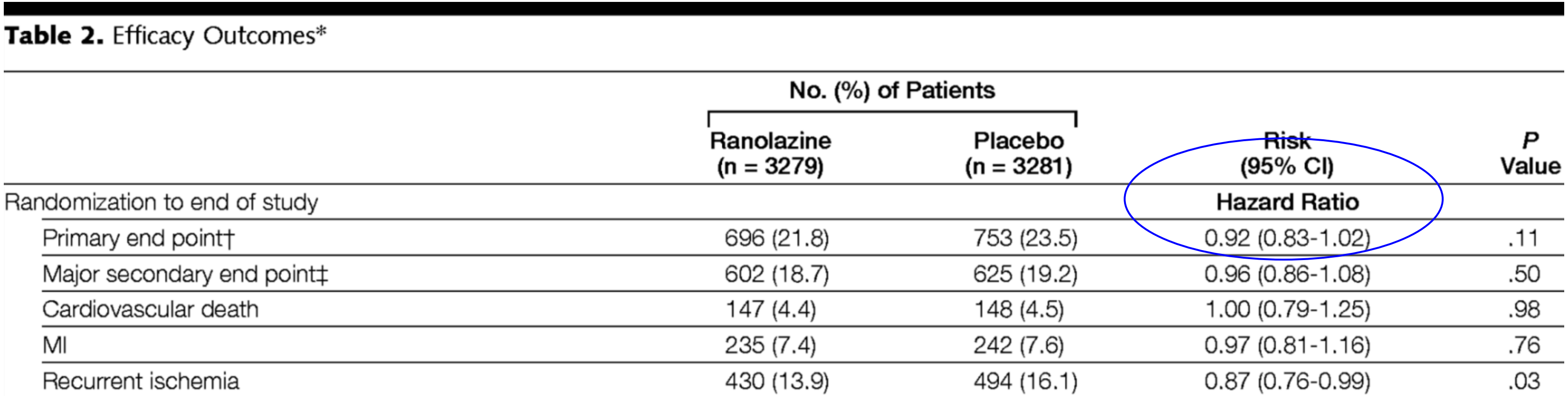


Table 2. Efficacy Outcomes*

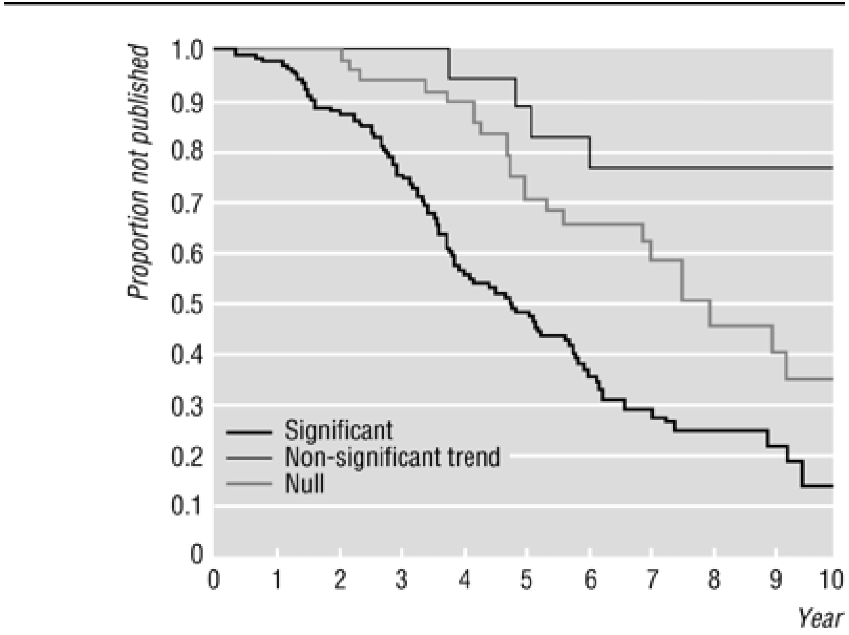
	No. (%) of Patients			
	Ranolazine (n = 3279)	Placebo (n = 3281)	Risk (95% CI)	*P* Value
Randomization to end of study			**Hazard Ratio**	
Primary end point†	696 (21.8)	753 (23.5)	0.92 (0.83-1.02)	.11
Major secondary end point‡	602 (18.7)	625 (19.2)	0.96 (0.86-1.08)	.50
Cardiovascular death	147 (4.4)	148 (4.5)	1.00 (0.79-1.25)	.98
MI	235 (7.4)	242 (7.6)	0.97 (0.81-1.16)	.76
Recurrent ischemia	430 (13.9)	494 (16.1)	0.87 (0.76-0.99)	.03

Interpretation: the rate of death, MI, or recurrent ischemia (primary end point) was reduced 8% in the ranolazine group compared with placebo (not significant).

Reproduced from: Morrow et al. Effects of Ranolazine on Recurrent Cardiovascular Events in Patients with Non-ST-Elevation Acute Coronary Syndromes. JAMA 2007; 297: 1775-1783.

EXAMPLE: STUDY OF PUBLICATION BIAS

Kaplan-Meier Curve:



No at risk						
Significant	144	127	77	36	15	2
Non-significant trend	20	20	19	14	4	3
Null	52	52	46	24	10	7

Reproduced from: Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects BMJ 1997;315:640-645

CORRESPONDING COX REGRESSION

$$ln\big(h(t)\big) = \alpha + \beta_{non-sign\ trend} + \beta_{sign\ results}$$

Table 4 Risk factors for time to publication using univariate Cox regression analysis

Characteristic	# not published	# published	Hazard ratio (95% CI)
Null	29	23	1.00
Non-significant trend	16	4	0.39 (0.13 to 1.12)
Significant	47	99	2.32 (1.47 to 3.66)

Reproduced from: Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects BMJ 1997;315:640-645

Interpretation: Significant results have a 2-fold higher incidence of publication compared to null results.

EXAMPLE 2: STUDY OF MORTALITY IN ACADEMY AWARD WINNERS FOR SCREENWRITING

Kaplan-Meier methods

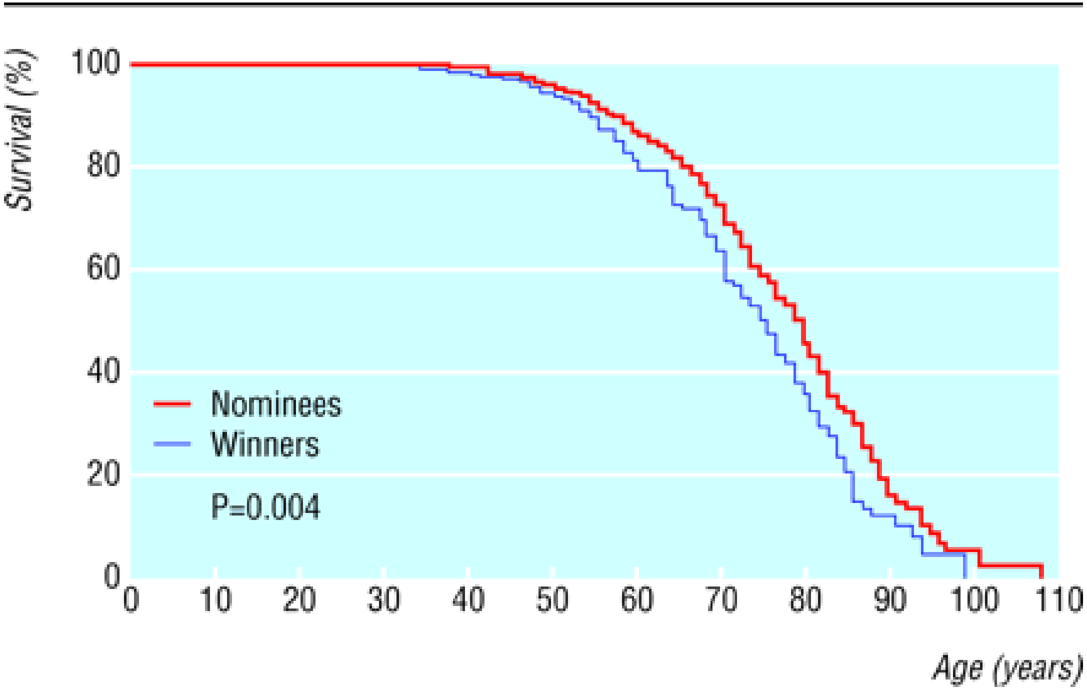


Figure 1 and Table 2 (next slide) were reproduced from: Redelmeier DA, Singh SM. Longevity of screenwriters who win an academy award: longitudinal study. *BMJ* 2001;323:1491-1496

Table 2. Death rates for screenwriters who have won an academy award.* Values are hazard ratios (95% confidence intervals) and are adjusted for the factor indicated

	Relative increase in death rate for winners
Basic analysis	1.37 (1.10 to 1.70)
Adjusted analysis	
Demographic:	
Year of birth	1.32 (1.06 to 1.64)
Sex	1.36 (1.10 to 1.69)
Documented education	1.39 (1.12 to 1.73)
All three factors	1.33 (1.07 to 1.65)
Professional:	
Film genre	1.37 (1.10 to 1.70)
Total films	1.39 (1.12 to 1.73)
Total four star films	1.40 (1.13 to 1.75)
Total nominations	1.43 (1.14 to 1.79)
Age at first film	1.36 (1.09 to 1.68)
Age at first nomination	1.32 (1.06 to 1.64)
All six factors	1.40 (1.11 to 1.76)
All nine factors	1.35 (1.07 to 1.70)

HR=1.37; interpretation: 37% higher incidence of death for winners compared with nominees

HR=1.35; interpretation: 35% higher incidence of death for winners compared with nominees even after adjusting for potential confounders

COX REGRESSION: MODEL DETAILS

THE HAZARD FUNCTION

$$h(t) = \lim_{\Delta t \longrightarrow 0} \frac{P(t \leq T < t + \Delta t / T \geq t)}{\Delta t}$$

In words: the probability that *if you survive to t*, you will succumb to the event in the next instant.

THE MODEL

Components:

•A baseline hazard function that is left unspecified but must be positive (=the hazard when all covariates are 0)

•A linear function of a set of k fixed covariates

Can take on any form!

$$\ln h_i(t) = \boxed{\ln h_0(t)} + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$$

HAZARD RATIO FOR A BINARY PREDICTOR

$$HR_{lung\ cancer\ /\ smoking} = \frac{h_i(t)}{h_j(t)} = \frac{\cancel{h_0(t)}e^{\beta_{smoking}(1)+\cancel{\beta_{age}(60)}}}{\cancel{h_0(t)}e^{\beta_{smoking}(0)+\cancel{\beta_{age}(60)}}} = e^{\beta_{smoking}(1-0)}$$

$$HR_{lung\ cancer\ /\ smoking} = e^{\beta_{smoking}}$$

This is the hazard ratio for smoking adjusted for age.

HAZARD RATIO FOR A CONTINUOUS PREDICTOR

$$HR_{\textit{lung cancer}\,/10-\textit{years}\text{ increase in age}} = \frac{h_i(t)}{h_j(t)} = \frac{\cancel{h_0(t)}e^{\cancel{\beta_{smoking}(0)}+\beta_{age}(70)}}{\cancel{h_0(t)}e^{\cancel{\beta_{smoking}(0)}+\beta_{age}(60)}} = e^{\beta_{age}(70-60)}$$

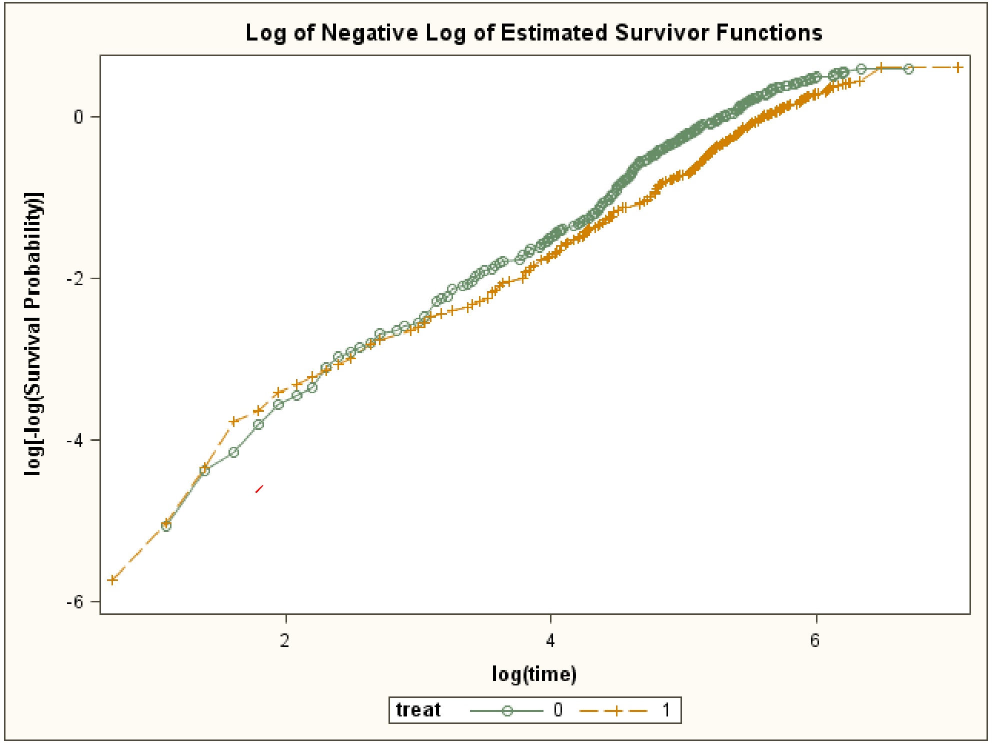
$$HR_{\textit{lung cancer}\,/10-\textit{years}\text{ increase in age}} = e^{\beta_{age}(10)}$$

This is the hazard ratio for a 10-year increase in age, adjusted for smoking.

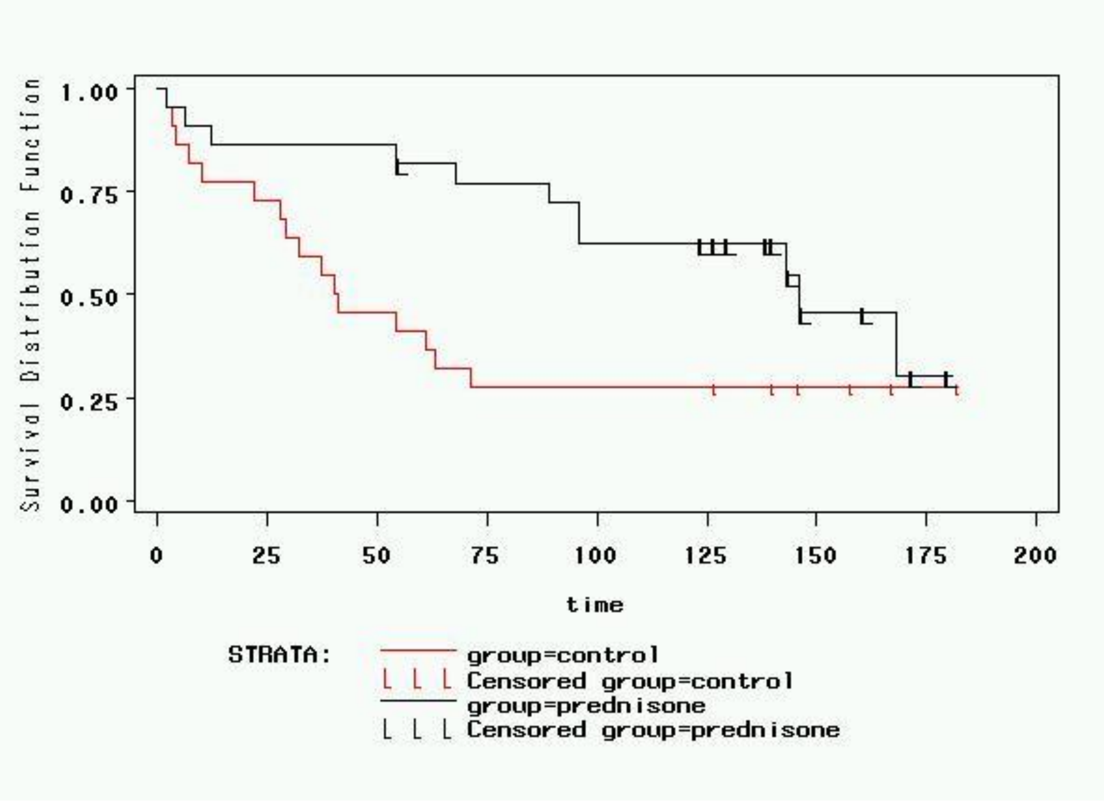
Exponentiating a continuous predictor gives you the hazard ratio for a 1-unit increase in the predictor.

THE PROPORTIONAL HAZARDS ASSUMPTION

TESTING PROPORTIONAL HAZARDS: LOG-LOG PLOT



RECALL: HEPATITIS EXAMPLE

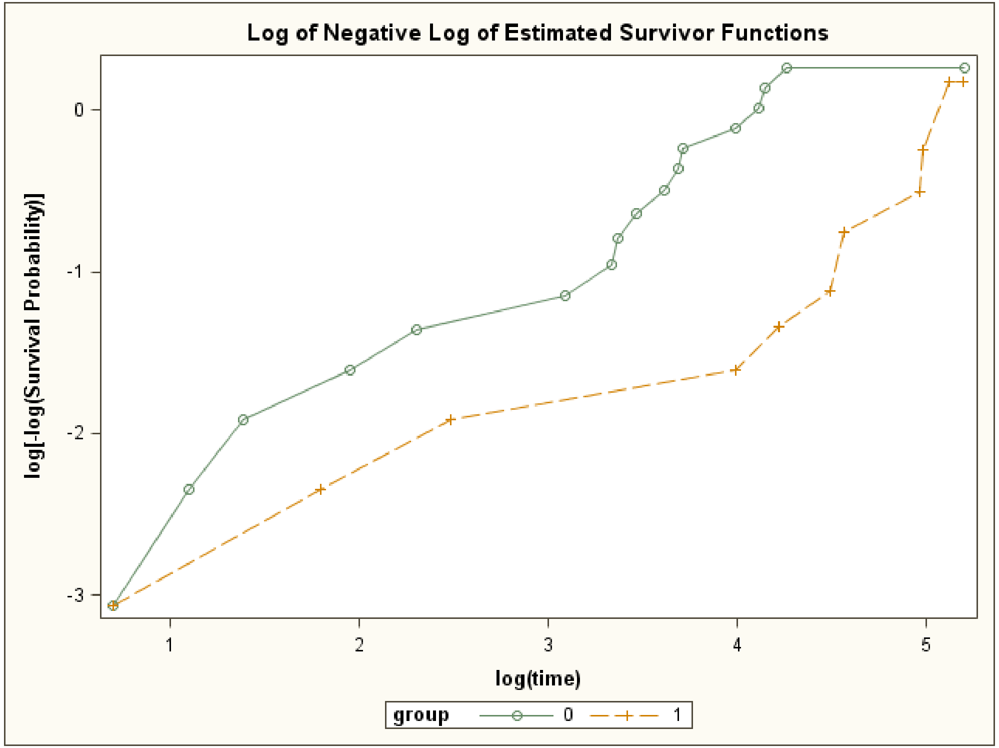


Data reproduced from: Bland and Altman. Time to event (survival) data. *BMJ* 1998;317:468.

CORRESPONDING COX REGRESSION

Analysis of Maximum Likelihood Estimates					
Parameter	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Treatment vs. Control	-0.83230	0.39739	4.3865	0.0362	0.435

TEST OF PROPORTIONAL HAZARDS ASSUMPTION: LOG-LOG PLOT



08. Transforming variables

09. Overfitting

10. Missing data

11. Residual confounding

Variable transformation may be useful for linear regression when:

1. My continuous outcome variable is not normally distributed (especially important for smaller samples, n<100.)

2. I have non-constant (non-homogenous) variances.

3. My predictor (independent) and outcome (dependent) variables do not have a linear relationship.

COMMON TRANSFORMATIONS

- Log
- Square root
- Reciprocal

UNIT 5: REGRESSION ANALYSIS

08. Transforming variables

09. Overfitting

10. Missing data

11. Residual confounding

OVERFITTING

- In multivariate modeling, you can get highly significant but meaningless results if you put too many predictors in the model.
- The model is fit perfectly to the quirks of your particular sample, but has no predictive ability in a new sample.

Rule of thumb: You need at least 10 subjects for each predictor variable in the multivariate regression model (and the intercept).

DECIDING ON THE IMPORTANT VARIABLES

- The most direct approach is called all subsets or best subsets regression: we compute the least squares fit for all possible subsets and then choose between them based on some criterion that balances training error with model size.
- However we often can not examine all possible models, since they are 2^p of them; for example when $p = 40$ there are over a billion models!
- Instead we need an automated approach that searches through a subset of them. We discuss two commonly used approaches next.

FORWARD SELECTION

- Begin with the null model — a model that contains an intercept but no predictors.
- Fit p simple linear regressions and add to the null model the variable that results in the lowest RSS.
- Add to that model the variable that results in the lowest RSS amongst all two-variable models.
- Continue until some stopping rule is satisfied, for example when all remaining variables have a p-value above some threshold.

BACKWARD SELECTION

- Start with all variables in the model.
- Remove the variable with the largest p-value — that is, the variable that is the least statistically significant.
- The new $(p - 1)$-variable model is fit, and the variable with the largest p-value is removed.
- Continue until a stopping rule is reached. For instance, we may stop when all remaining variables have a significant p-value defined by some significance threshold.

MODEL SELECTION - CONTINUED

- Later we discuss more systematic criteria for choosing an “optimal” member in the path of models produced by forward or backward stepwise selection.
- These include Mallow’s Cp, Akaike information criterion (AIC), Bayesian information criterion (BIC), adjusted R2 and Cross-validation (CV).

08. Transforming variables

09. Overfitting

10. Missing data

11. Residual confounding

ALWAYS CHECK YOUR N’S

- Most regression analyses automatically throw out incomplete observations, so if a subject is missing the value for just one of the variables in the model, that subject will be excluded.
- This can add up to lots of omissions!
- Always check your N’s!

UNIT 5: REGRESSION ANALYSIS

08. Transforming variables

09. Overfitting

10. Missing data

11. Residual confounding

RESIDUAL CONFOUNDING

- You cannot completely wipe out confounding simply by adjusting for variables in multiple regression unless variables are measured with zero error (which is usually impossible).
- Example: meat eating and mortality

Men who eat a lot of meat are unhealthier for many reasons!

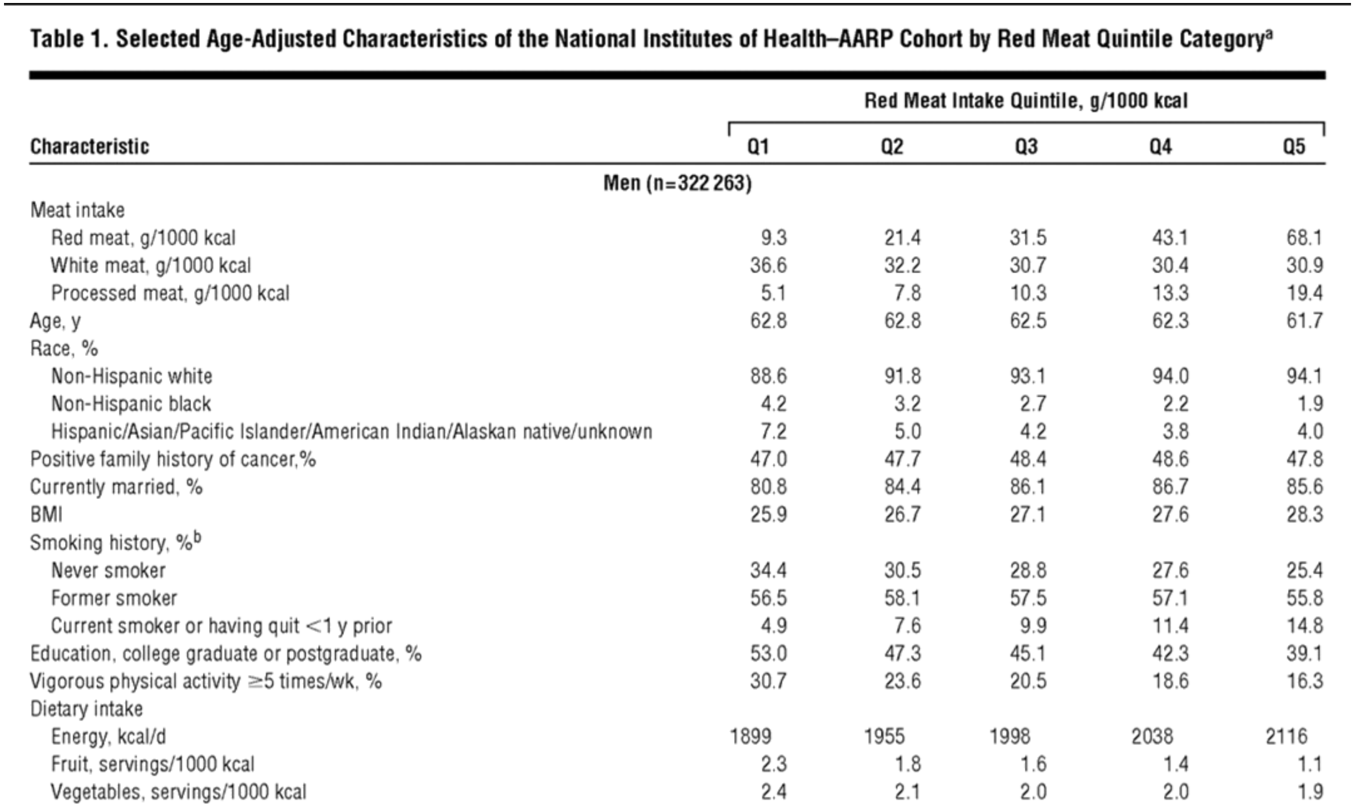


Table 1. Selected Age-Adjusted Characteristics of the National Institutes of Health–AARP Cohort by Red Meat Quintile Category[a]

Characteristic	Red Meat Intake Quintile, g/1000 kcal				
	Q1	Q2	Q3	Q4	Q5
Men (n=322 263)					
Meat intake					
Red meat, g/1000 kcal	9.3	21.4	31.5	43.1	68.1
White meat, g/1000 kcal	36.6	32.2	30.7	30.4	30.9
Processed meat, g/1000 kcal	5.1	7.8	10.3	13.3	19.4
Age, y	62.8	62.8	62.5	62.3	61.7
Race, %					
Non-Hispanic white	88.6	91.8	93.1	94.0	94.1
Non-Hispanic black	4.2	3.2	2.7	2.2	1.9
Hispanic/Asian/Pacific Islander/American Indian/Alaskan native/unknown	7.2	5.0	4.2	3.8	4.0
Positive family history of cancer,%	47.0	47.7	48.4	48.6	47.8
Currently married, %	80.8	84.4	86.1	86.7	85.6
BMI	25.9	26.7	27.1	27.6	28.3
Smoking history, %[b]					
Never smoker	34.4	30.5	28.8	27.6	25.4
Former smoker	56.5	58.1	57.5	57.1	55.8
Current smoker or having quit <1 y prior	4.9	7.6	9.9	11.4	14.8
Education, college graduate or postgraduate, %	53.0	47.3	45.1	42.3	39.1
Vigorous physical activity ≥5 times/wk, %	30.7	23.6	20.5	18.6	16.3
Dietary intake					
Energy, kcal/d	1899	1955	1998	2038	2116
Fruit, servings/1000 kcal	2.3	1.8	1.6	1.4	1.1
Vegetables, servings/1000 kcal	2.4	2.1	2.0	2.0	1.9

Reproduced from: Sinha R, Cross AJ, Graubard BI, Leitzmann MF, Schatzkin A. Meat intake and mortality: a prospective study of over half a million people. *Arch Intern Med* 2009;169:562-71

MORTALITY RISKS

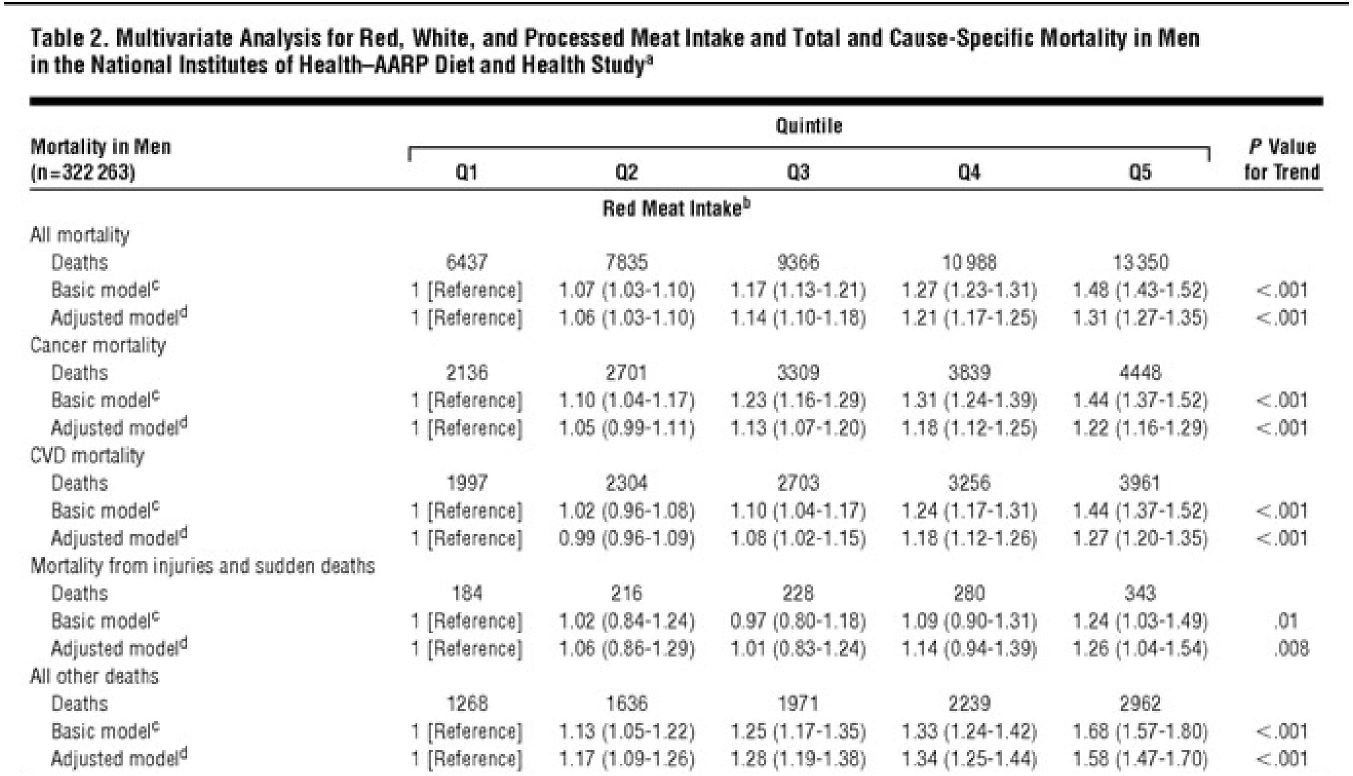


Table 2. Multivariate Analysis for Red, White, and Processed Meat Intake and Total and Cause-Specific Mortality in Men in the National Institutes of Health–AARP Diet and Health Study[a]

Mortality in Men (n=322 263)	Quintile					*P* Value for Trend
	Q1	Q2	Q3	Q4	Q5	
			Red Meat Intake[b]			
All mortality						
Deaths	6437	7835	9366	10 988	13 350	
Basic model[c]	1 [Reference]	1.07 (1.03-1.10)	1.17 (1.13-1.21)	1.27 (1.23-1.31)	1.48 (1.43-1.52)	<.001
Adjusted model[d]	1 [Reference]	1.06 (1.03-1.10)	1.14 (1.10-1.18)	1.21 (1.17-1.25)	1.31 (1.27-1.35)	<.001
Cancer mortality						
Deaths	2136	2701	3309	3839	4448	
Basic model[c]	1 [Reference]	1.10 (1.04-1.17)	1.23 (1.16-1.29)	1.31 (1.24-1.39)	1.44 (1.37-1.52)	<.001
Adjusted model[d]	1 [Reference]	1.05 (0.99-1.11)	1.13 (1.07-1.20)	1.18 (1.12-1.25)	1.22 (1.16-1.29)	<.001
CVD mortality						
Deaths	1997	2304	2703	3256	3961	
Basic model[c]	1 [Reference]	1.02 (0.96-1.08)	1.10 (1.04-1.17)	1.24 (1.17-1.31)	1.44 (1.37-1.52)	<.001
Adjusted model[d]	1 [Reference]	0.99 (0.96-1.09)	1.08 (1.02-1.15)	1.18 (1.12-1.26)	1.27 (1.20-1.35)	<.001
Mortality from injuries and sudden deaths						
Deaths	184	216	228	280	343	
Basic model[c]	1 [Reference]	1.02 (0.84-1.24)	0.97 (0.80-1.18)	1.09 (0.90-1.31)	1.24 (1.03-1.49)	.01
Adjusted model[d]	1 [Reference]	1.06 (0.86-1.29)	1.01 (0.83-1.24)	1.14 (0.94-1.39)	1.26 (1.04-1.54)	.008
All other deaths						
Deaths	1268	1636	1971	2239	2962	
Basic model[c]	1 [Reference]	1.13 (1.05-1.22)	1.25 (1.17-1.35)	1.33 (1.24-1.42)	1.68 (1.57-1.80)	<.001
Adjusted model[d]	1 [Reference]	1.17 (1.09-1.26)	1.28 (1.19-1.38)	1.34 (1.25-1.44)	1.58 (1.47-1.70)	<.001

Reproduced from: Sinha R, Cross AJ, Graubard BI, Leitzmann MF, Schatzkin A. Meat intake and mortality: a prospective study of over half a million people. *Arch Intern Med* 2009;169:562-71

RESIDUAL CONFOUNDING

- For a binary predictor, incomplete of confounding can plausibly generate spurious relative risks in the range of 0.6 to 1.6.
- In addition to creating spurious associations, residual confounding can also obscure relationships, leading researchers to miss associations.

iibm

Instituto de Ingeniería Biológica y Médica