Unobserved Heterogeneity and the Robustness of Estimates of the Effects of Income on Life Expectancy – An Illustrative and Hypothetical Example

M. C. Wolfson and V. Bijelic, University of Ottawa

Basic Question

- what if, instead of the usual kinds of assumptions made when estimating effect sizes for risk factors, we made explicit assumptions about unobserved heterogeneity (UH) in the population under study; how would our results differ?

Outline of the Analysis

- how to frame the concept of unobserved heterogeneity (UH) in the context of survival curves, income effects, and life expectancy (LE)
- results from conventional Cox proportional hazards regressions
  - assess proportional hazards assumption
  - a simple approach to address violation
- build on Vaupel and Yashin (1985) and assume binary UH
- compare results of intervention "thought experiments"

Exploring the Effects of Income on Mortality

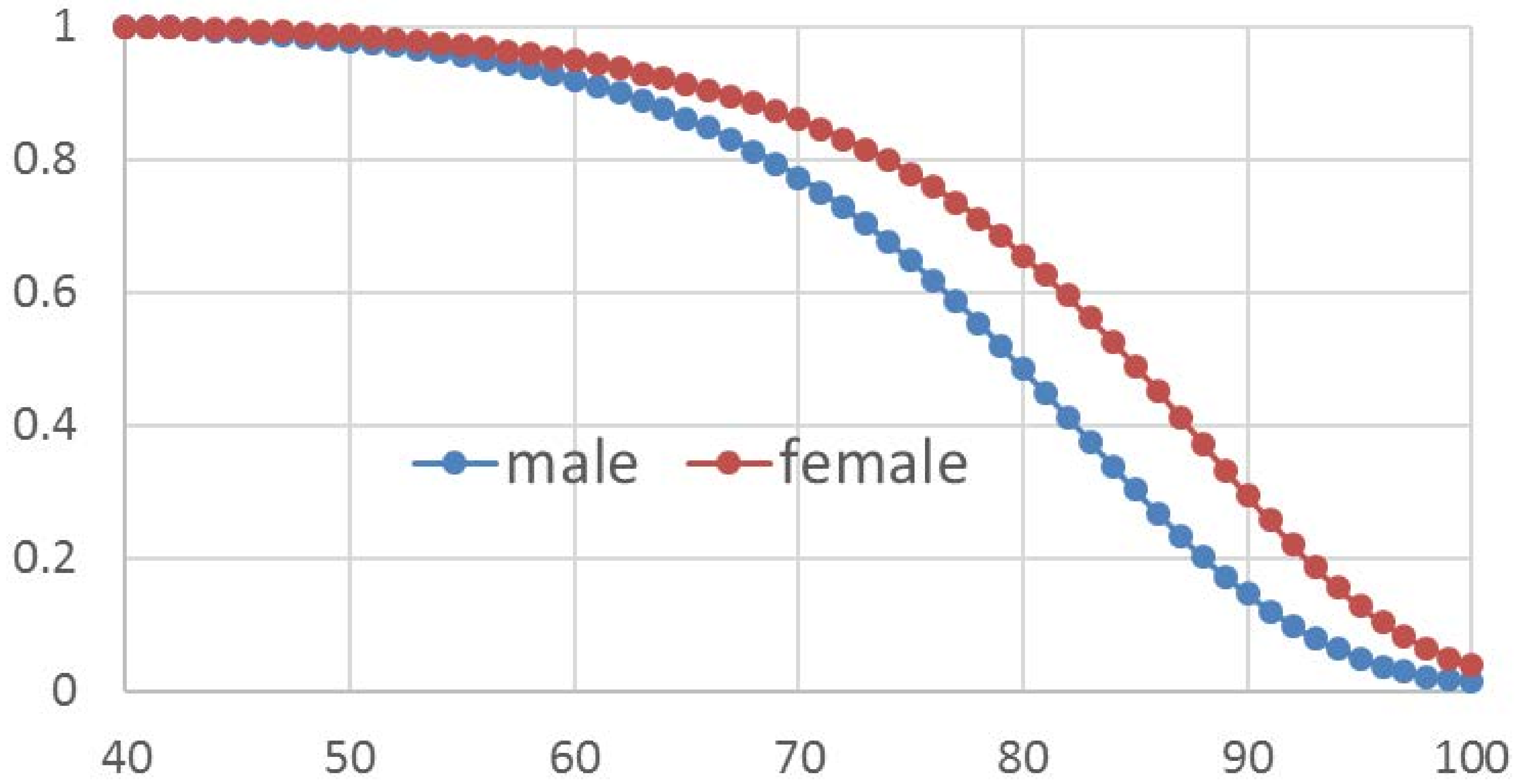
- use 1991 population census linked to mortality to 2006
- focus on individuals using family income expressed as a ratio to Statistics Canada’s Low Income Cut-Off (LICO, de facto poverty line))
- create five income groups after triming off bottom 1% and top 1% (for confidentiality)
    - very low (1^{st} to 10^{th} percentile)
    - low (10^{th} to 25^{th})
    - middle (25^{th} to 75^{th})
    - high (75^{th} to 90^{th})
    - very high (90^{th} to 99^{th})
- apply these percentiles to each sex and 5-year age group separately
- thereby creating age-group-specific income dummy variables
    - n.b. essential given large differences in income distributions by age

Resulting LICO (Low Income Cut-Off = “Poverty Line”) Ratios for 5-Year Age Groups by Population Percentiles

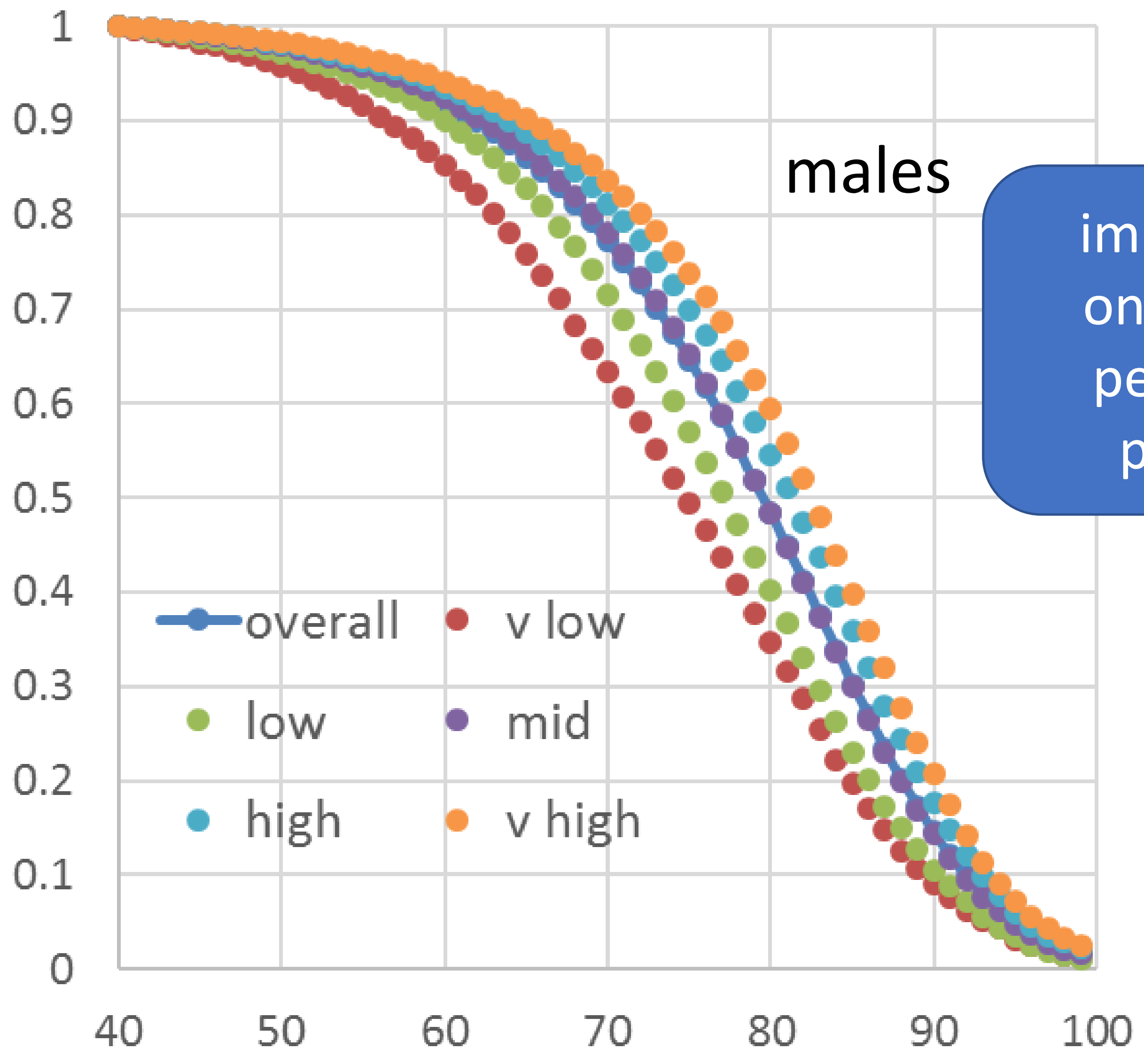
age	1st	10th	25th	75th	90th	99th
40-44	0.1	1.0	1.7	3.5	4.7	9.0
45-49	--	1.1	1.8	3.8	5.1	10.3
50-54	--	1.0	1.8	3.9	5.3	10.8
55-59	--	0.8	1.6	3.7	5.2	11.2
60-64	--	0.8	1.4	3.3	4.7	11.0
65-69	--	0.9	1.2	2.9	4.2	10.0
70-74	--	0.9	1.2	2.7	4.0	9.3
75-79	--	0.9	1.1	2.5	3.8	9.2
80-84	--	0.8	1.0	2.3	3.6	--
85-89	--	0.8	1.0	2.2	3.5	--
90-94	--	0.8	0.9	2.2	3.5	--
95-99	--	0.8	1.0	2.8	3.9	--

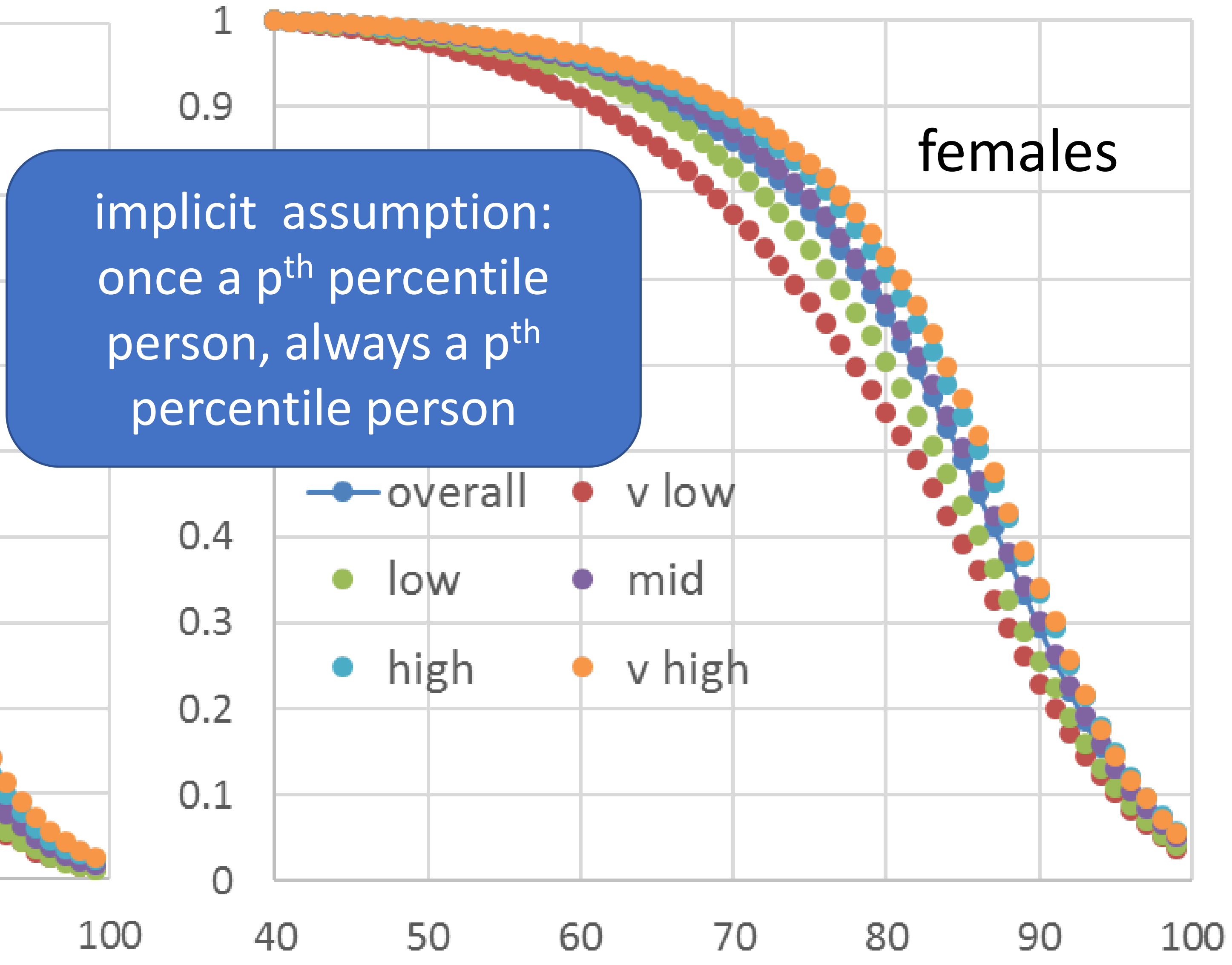
-- sample size too small

Overall Kaplan-Meier (K-M) Curves (conditional on survival to age 40)



K-M Curves by Sex and Five Income Groups

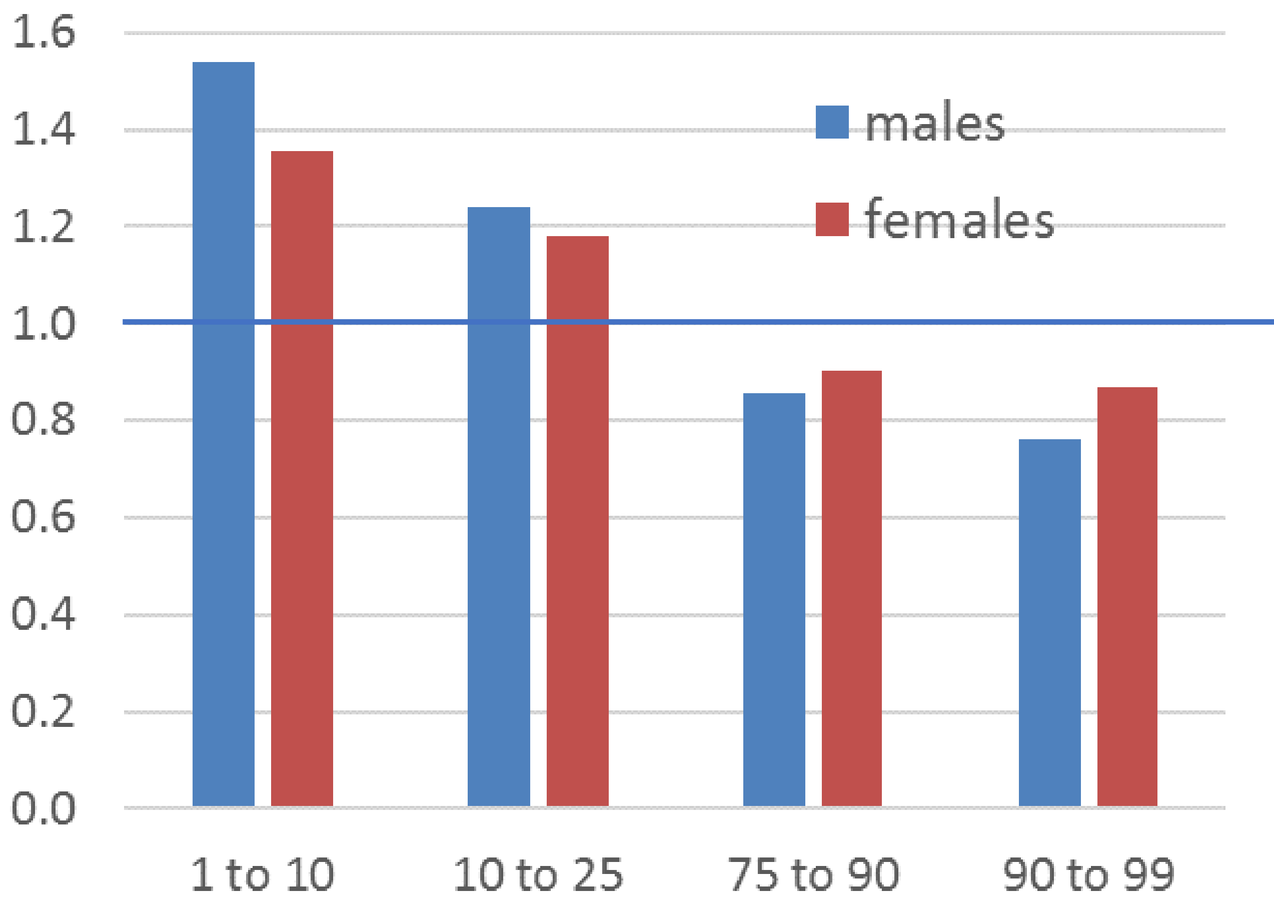




Cox Proportional Hazards Regression (Version 1)

income		hazard ratio	
group	pct'iles	males	females
v low	1 to 10	1.539	1.356
low	10 to 25	1.240	1.182
high	75 to 90	0.854	0.903
v high	90 to 99	0.762	0.870

- reference group = 25th to 75th percentiles
- Pr < Chi Sq .0001 for all coefficients



Inferring Life Expectancies (LEs) from Cox 1 Regression Results

- start with reference group mortality rates
  - mid = middle income group, 25th to 75th percentiles)
- scale vectors of sex- and age-specific mortality rates by each of the four sex-specific hazard rates
- derive / cumulate to obtain four corresponding survival curves
- compute LEs in usual way

Comparison of Life Expectancy Estimates by Sex, Income Group, and Method

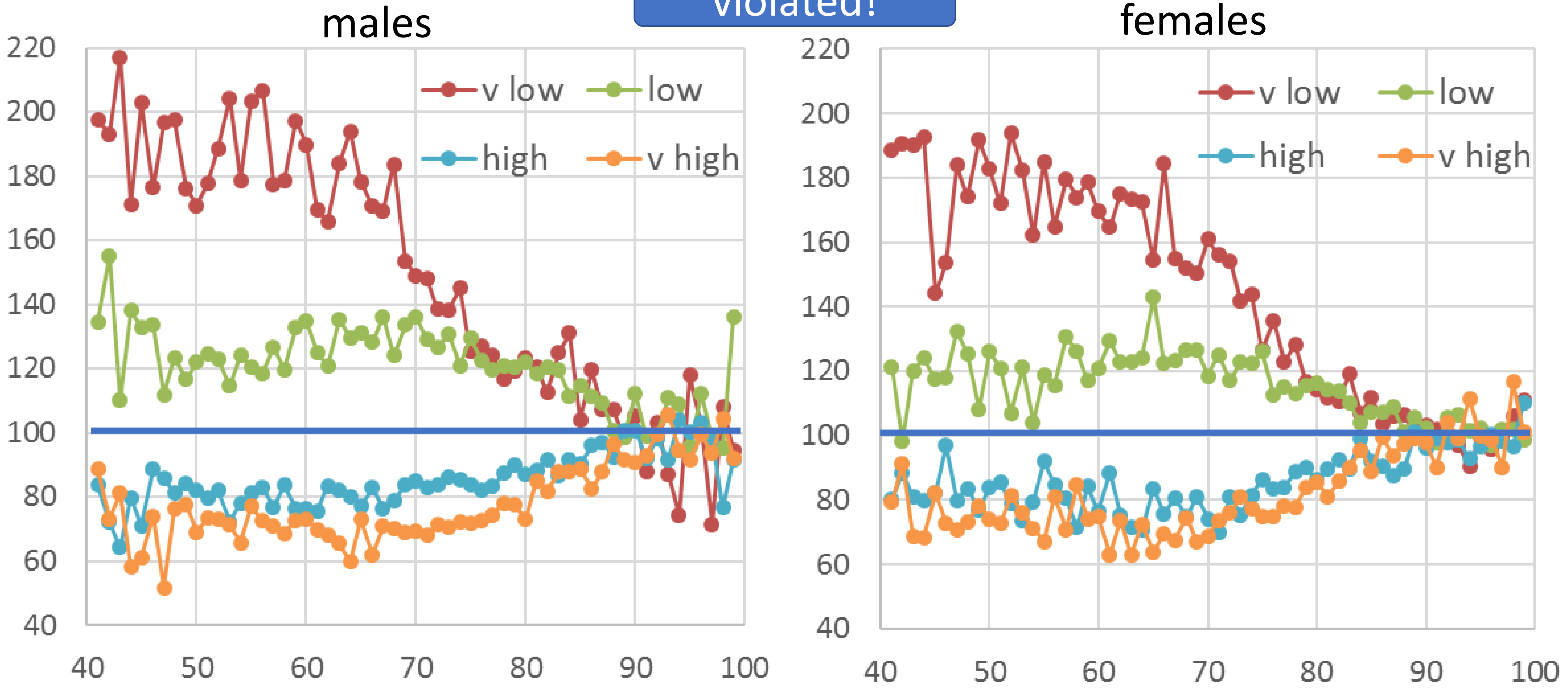
	Income Group	v low	low	mid	high	v high
	Income Percentiles	1 to 10	10 to 25	25 to 75	75 to 90	90 to 99
Males	from K-M	73.3	75.5	77.7	79.1	80.2
	from Cox 1	73.3	75.4	77.5	79.1	80.2
Females	from K-M	79.0	80.8	82.6	83.5	83.9
	from Cox 1	79.3	80.6	82.2	83.1	83.4

LEs close to or identical across sex and quantile income groups

Assessing Proportional Hazards Assumption

Ratios of Mortality Rates to Overall Rate (%; n.b. inferred)

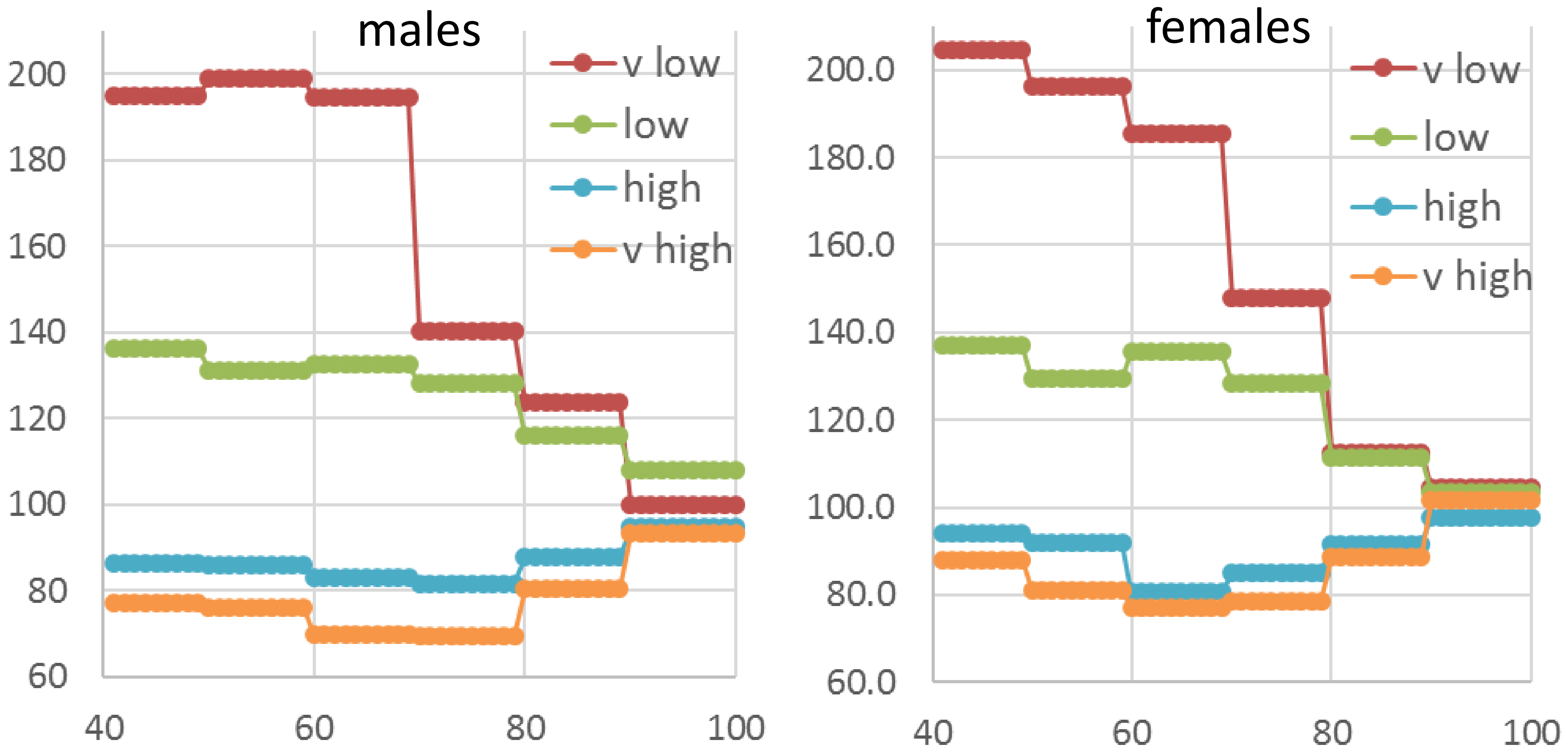
violated!



Cox Proportional Hazards Regression (Version 2)

- proportional hazards assumption is clearly violated
- one straightforward approach:
  - stratify population by 10-year age groups
  - run separate regressions for each sex and 10-year age group
- again start with reference group mortality rates (25^{th} to 75^{th} percentiles)
- but scale mortality rates by each of the four hazard rates separately for each 10-year age group
- construct four corresponding survival curves by piecing together each 10-year survival curve for both males and females, and each income
- compute LEs in usual way

Cox 2: Estimated Ratios of Mortality Rates (%) to Overall Rate by Income Using 10-Year Age Groups



Comparison of Life Expectancy Estimates by Sex, Income Group, and Method

	Income Group	v low	low	mid	high	v high
	Income Percentiles	1 to 10	10 to 25	25 to 75	75 to 90	90 to 99
Males	from K-M	73.3	75.5	77.7	79.1	80.2
	from Cox 1	73.3	75.4	77.5	79.1	80.2
	from Cox 2	72.6	75.2	77.5	79.1	80.3
Females	from K-M	79.0	80.8	82.6	83.5	83.9
	from Cox 1	79.3	80.6	82.2	83.1	83.4
	from Cox 2	78.1	80.2	82.2	83.2	83.7

LEs from Cox 2 quite different in lower income groups

Thought Experiment 1

- (n.b. ☹: assume correlation is causation)
- suppose we redistribute income so that one person moves from v low to low (lifetime) income
- then using Cox 1 regression results, their LE would increase by 2.1 years for males, and 1.3 years for females
- and almost the same changes in LE if we were to use K-M results

Thought Experiment 2

- again suppose we redistribute income so that one person moves from v low to low (lifetime) income
- then their LE would increase by 2.6 years (was 2.1 years) for males, and 2.1 years (was 1.3 years) for females, using Cox 2 regression results (compared to Cox 1)
- i.e. there is more than half a year of LE income effect size change when proportional hazards assumption is addressed

Beyond K-M and Semi-Parametric Regression: Assume Binary Unobserved Heterogeneity (UH)

- start with same K-M survival curves
- avoid Cox regression and (strong) proportional hazards assumption
- instead start simply with K-M survival curves
- and assume binary unobserved heterogeneity
- i.e. assume two “pure” but unobservable population types
    - frail = shorter LEs = “ultraviolet” (UV, shorter wave length)
    - robust = longer LEs = “infrared” (IR, longer wave length)
- such that all differences in survival curves between income groups are simply the result of nothing more than that their populations consist of different mixtures of the two underlying “pure types” of individuals, UV (frail) and IR (robust)

(the seminal paper)

Heterogeneity's Ruses: Some Surprising Effects of Selection on Population Dynamics

JAMES W. VAUPEL and ANATOLI I. YASHIN*

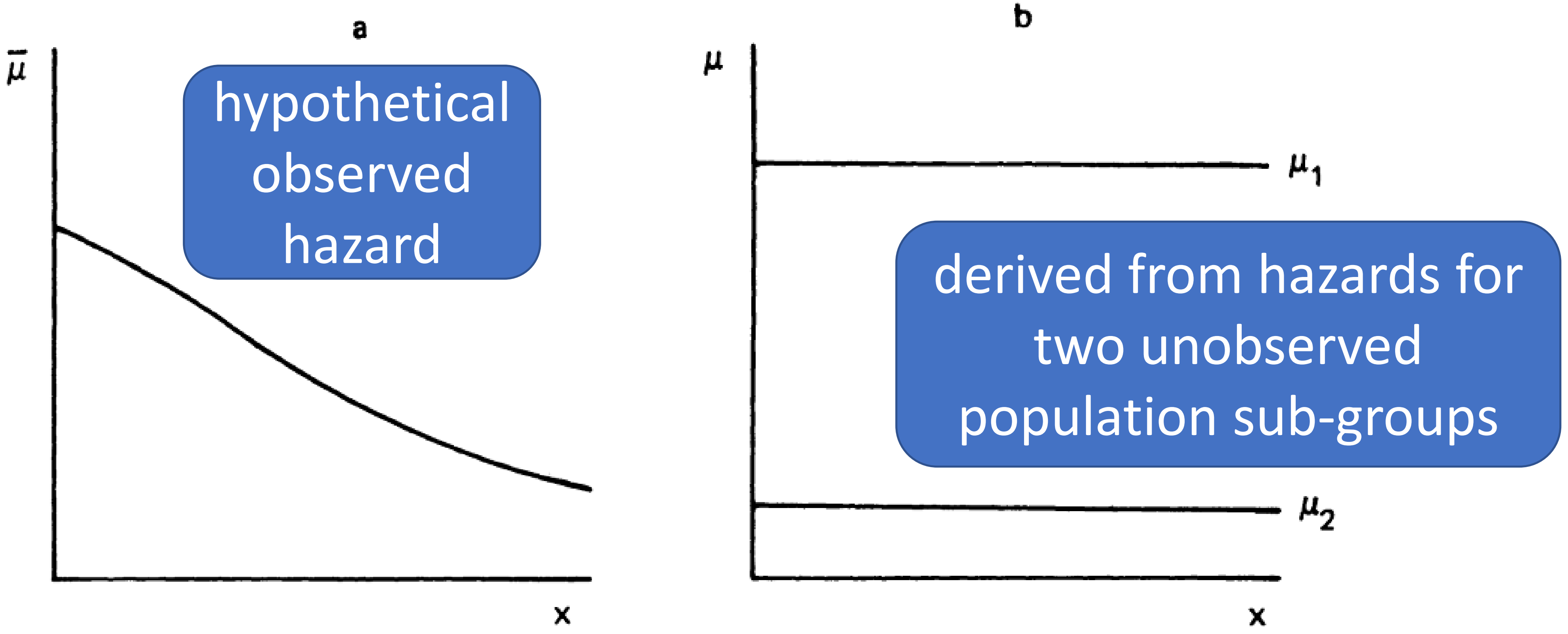


Figure 1. The observed hazard rate may decline even though the hazard rates for the two subcohorts are constant. The curve for $\bar{\mu}$ was calculated from (2), (3), and (4) using $\mu_1 = .06$, $\mu_2 = .01$, and $\pi(0) = .8$. The curves are shown for values of x from 0 to 75.

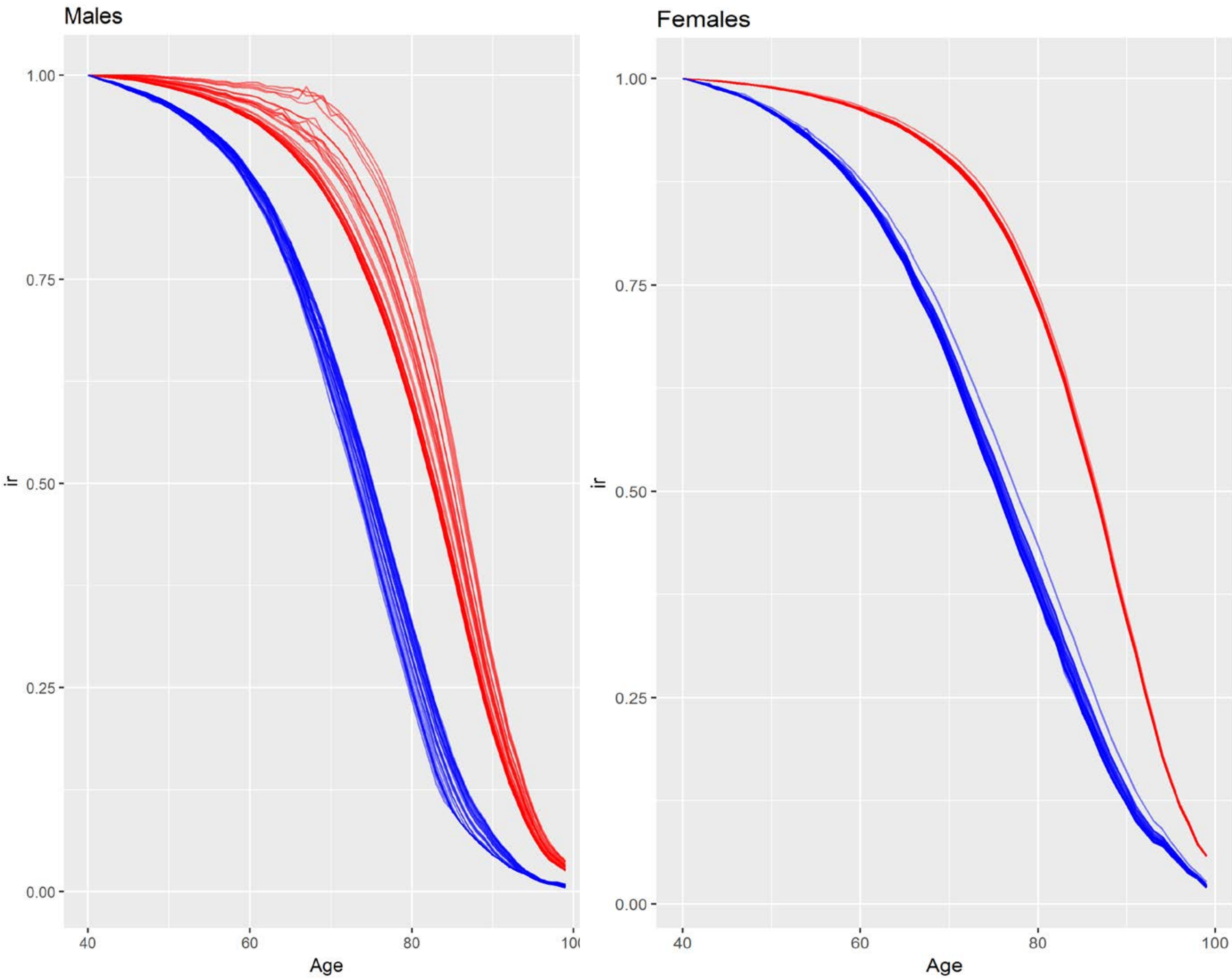
Inferring UV and IR Populations from Observed K-M Survival Curves by Income

- use non-parametric approach, i.e. as open as possible
- represent UV and IR survival curves as vectors of annual survival rates
- posit that each K-M income group survival curve is a simple linear combination = weighted average of the same overall UV and IR curves
- we then have for each sex 5 x 61 observed data points (i.e. 61 for each of the 5 income groups)
- and 2 x 61 + 5 unknowns (i.e. 61 for UV curve + 61 for IR curve + 5 weights)
- use simulated annealing to find best fitting UV, IR, and 5 weights
- where "fit" criterion is analogous to elastic net, i.e. a weighted average of a lasso (L1) + ridge regression (L2) normed distances summed across the 5 income groups
- repeat simulated annealing process for a series of random number seeds and "starting temperatures" to obtain 30 replicates

Result: Simulated Annealing Survival Curves for UV (blue) and IR (red) "Pure Type" Populations

30 Replicates

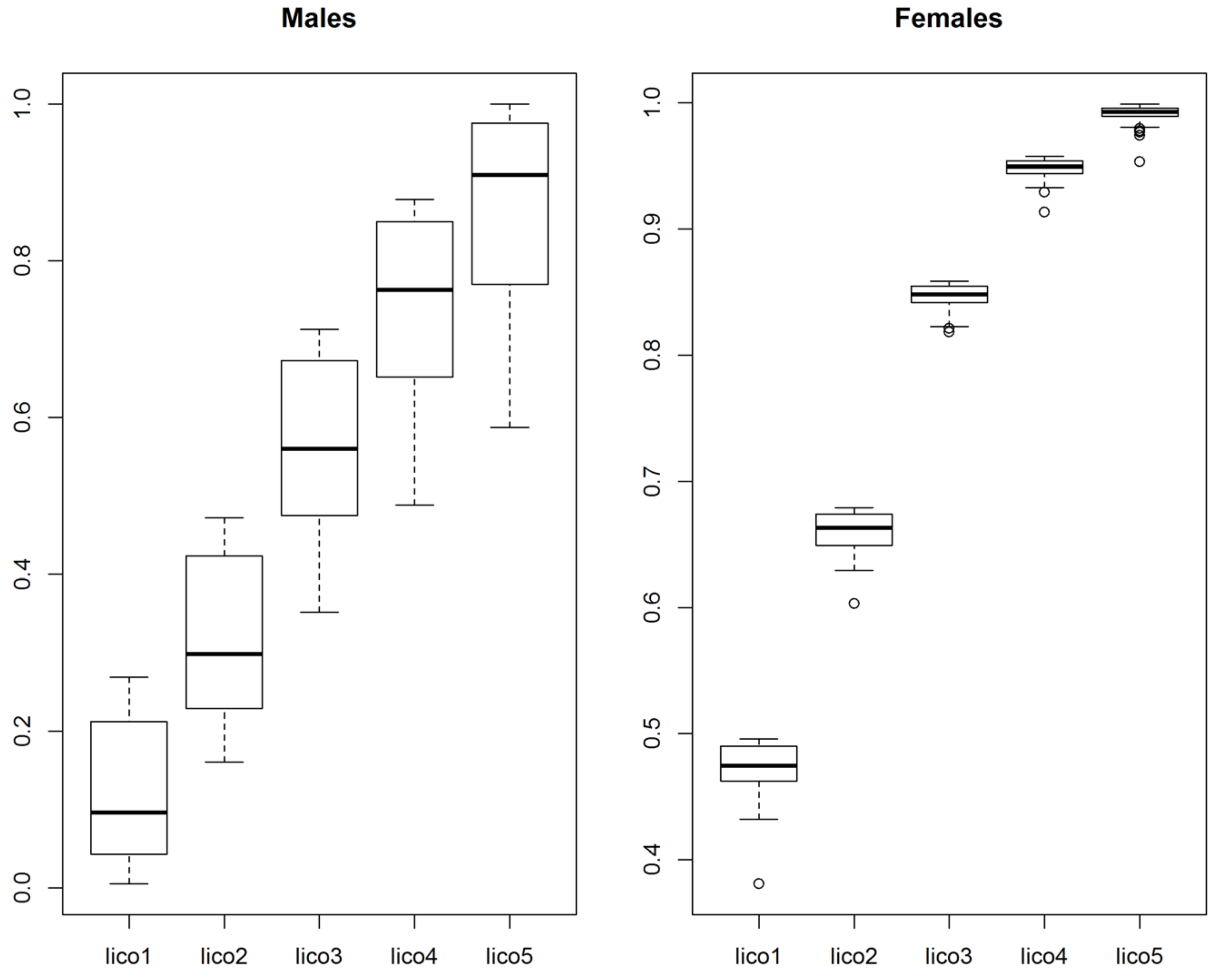
IR Males Most Variable



Result: Simulated Annealing Weights for UV and IR "Pure Type" Populations

30 Replicates

Males More Variable for all Income Groups



Simulated Annealing Results for UV and IR Life Expectancies (LEs), 10 Replicates

		ir	uv	v low	low	mid	high	v high
males	min	80.4	71.8	73.6	75.4	77.6	79.1	80.2
	max	84.6	73.6	73.8	75.6	77.7	79.1	80.2
	median	81.0	73.0	73.7	75.5	77.6	79.1	80.2
	average	81.7	72.8	73.7	75.5	77.6	79.1	80.2
females	min	83.9	74.3	79.0	80.8	82.5	83.5	83.9
	max	84.1	75.2	79.0	80.8	82.5	83.5	83.9
	median	83.9	74.4	79.0	80.8	82.5	83.5	83.9
	average	84.0	74.5	79.0	80.8	82.5	83.5	83.9

- even though UV and IR LEs by themselves are quite variable across replicates
- LE results by income group (i.e. combining weights and corresponding survival curves for each replicate) are very tight

Comparison of Life Expectancy Estimates by Sex, Income Group, and Method

	Income Group	v low	low	mid	high	v high
	Income Percentiles	1 to 10	10 to 25	25 to 75	75 to 90	90 to 99
Males	from K-M	73.3	75.5	77.7	79.1	80.2
	from Cox 1	73.3	75.4	77.5	79.1	80.2
	from Cox 2	72.6	75.2	77.5	79.1	80.3
	from uv & ir	73.7	75.5	77.6	79.1	80.2
Females	from K-M	79.0	80.8	82.6	83.5	83.9
	from Cox 1	79.3	80.6	82.2	83.1	83.4
	from Cox 2	78.1	80.2	82.2	83.2	83.7
	from uv & ir	79.0	80.8	82.5	83.5	83.9

n.b. results from K-M and UH (uv & ir) methods almost identical, as intended

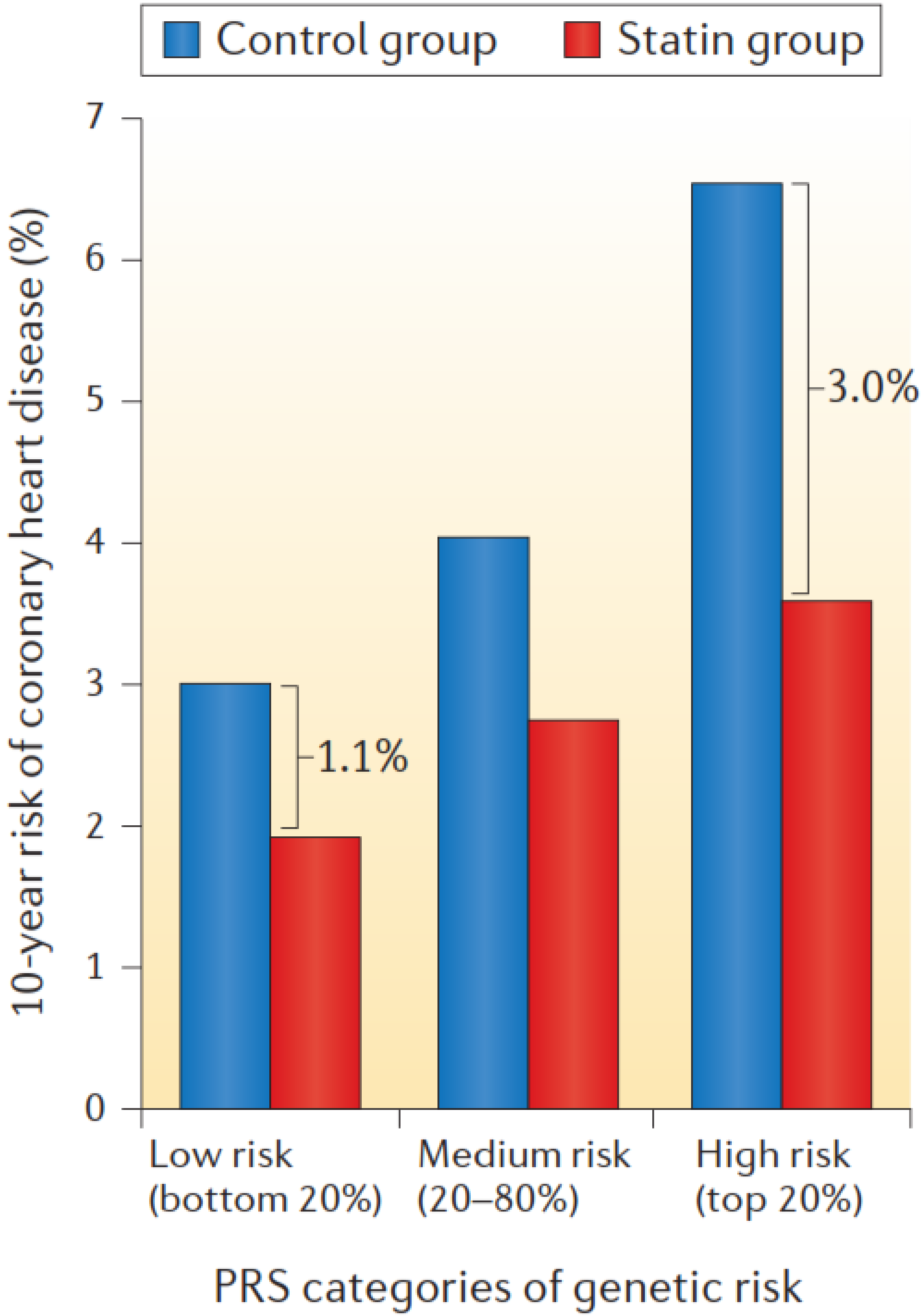
Thought Experiment 3

- suppose income redistribution only benefits frail = UV individuals
  - no matter what their income
  - no effect at all for robust = IR individuals
- if one UV individual could be helped to become like an IR individual
  - LE would increase about 8 years for males, and almost 10 years for females
  - compared to 1.3 to 2.6 years based on Cox regressions
  - i.e. a huge difference

Thought Experiment 4 – Analogy with Statins

- PRS = polygenic risk score
- until recent advances in genetics, PRS was a form of unobserved heterogeneity
- study results show clearly that "effect size" of a statin intervention on 10-year CHD events is almost triple in individuals with high compared to low risk PRS

Mega, J. L. *et al.* Genetic risk, coronary heart disease events, and the clinical benefit of statin therapy: an analysis of primary and secondary prevention trials. *Lancet* ***385, 2264–2271 (2015).***



Thought Experiment 4 – Result

- follow Thought Experiments 1 and 2, and suppose income redistribution raised everyone in the "v low" income group to the "low" income group
- recall that the v low group is 9% of the population, while the low income group is 15%
- also recall the increase in LE from the v low group to the low income group is about 2 years for males, and 1 year for females (whether K-M, Cox 1 or Cox 2)
- the proportion of UVs in the v low group is about 10% and 47% for males and females respectively, while in the low group it is about 30% and 66%
- so the difference in the proportion of UVs in the v low as compared to the low income group is about 20% for each sex
- the proportion of UVs in the new low income group would thus increase by about 10%
- IRs are assumed not to be affected at all by the income redistribution (admittedly an extreme case)
- thus improvements in LE would be about one-tenth the differences in LE between the v low and low income groups, i.e. about 0.13 to 0.26 years
- i.e. a very small effect

Conclusions

- empirical analysis of risk factors for important health outcomes typically makes a priori assumptions about statistical methods
- many of these assumptions are not tested, e.g. functional form (specification error), proportional hazards for Cox regression, and omitted variable bias
- an important form of omitted variable bias results from unobserved heterogeneity – which seems in principle unavoidable
- using a very large data set, the 1991 Canadian census, we have estimated the income gradient for mortality, and then explored the likely “effect sizes” for income redistribution
- we have also explored the implications of an explicit assumption of binary unobserved heterogeneity (n.b. not necessarily stronger than the widespread assumption of proportional hazards) using a novel non-parametric approach and simulated annealing
- the difference in effect sizes for hypothetical income redistribution between being able to identify the otherwise unobserved heterogeneity on the one hand, and assuming it is there but unidentifiable, is on the order of 100:1