UMEÅ UNIVERSITET

From birth to death: Statistical analysis of life courses

Xavier de Luna, Umeå University, Sweden

Department of Statistics – USBE @ Umeå University

Lab for Statistics and Register Studies

Umeå SIMSAM Lab & ALC

Swedish Statistical Society, Stockholm, March 2014

Outline

- Life course, trajectory, biography; sequence of events

- Life course as the unit of study: models and methods

  - Family life trajectories and retirement decisions

    Joint work with Ingrid Svensson, Emma Lundholm and Gunnar Malmberg (Ageing and Living Condition program, Umeå)

  - Effect of early retirement on health

    Joint work with Nicola Barban (Groningen), Francesco Billari (Oxford), Ingrid Svensson and Emma Lundholm (Umeå)

Life biographies data

State-space: $\mathcal{S} = \{S(ingle), M(arried), C(ohabiting)\}$

For individual i we observe $s_{it} \in \mathcal{S}$ for $t = 1, 2, \ldots, T$

Two life biographies:

$$SSSSCCCCCMMMMMMMMMMM$$

$$SSSSSSSSSSCCCCCCCCCCCCCCCC$$

Unit of study

- Transition between states (event history analysis, Markov Chains)

OR

- State trajectories

Two life biographies:

$$SSSSCCCCCMMMMMMMMMMM$$

$$SSSSSSSSSSCCCCCCCCCCCCCCC$$

Sequence analysis

- A collection of algorithm for classifying life trajectories

- Originally developed for genetic analyses

- Used increasingly by demographers & sociologists

Optimal matching algorithm

- OMA: a family of classification algorithm for sequences
- Compute a distance between 2 sequences as a function of the amount of edit operation needed to transform one sequence into the other
- Three operations
  - Insertion
  - Deletion
  - Substitution
- Cost defined for each operation: distance is sum of costs

Example

SATURDAY

SUNDAY

SATURDAY

S~~AT~~URDAY (2 x Deletion)

S UNDAY (1 x Substitution)

OM algorithm: model

$S = \{S_1, \ldots, S_T\}$ a vector random variable with state space $\Sigma = \{\sigma_1, \ldots, \sigma_K\}$

Realizations:

Biography for individual i is $s_i = \{s_{i1}, \ldots, s_{iT}\}$

(unit of study)

OM algorithm

$\omega : \Sigma \rightarrow \Sigma$ (operators)

$\omega \in \Omega$, where $\Omega = \{\text{ins, del, sub}\}$ (operator set)

$c(\omega) : \Omega \rightarrow \mathcal{R}^+$

Two biographies s_1 and s_2 such that

$$s_2 = \omega_1 \circ \omega_2 \circ \cdots \circ \omega_J(s_1) = \omega_.(s_1)$$

OM algorithm

Two biographies s_1 and s_2 such that

$$s_2 = \omega_1 \circ \omega_2 \circ \cdots \circ \omega_J(s_1) = \omega_.(s_1)$$

$$c(\omega_.) = \sum_{j=1}^{J} c(\omega_j) \qquad \text{(cost of the operation)}$$

$$\mathcal{D}(s_1, s_2) = \min_{\omega_.}\{c(\omega_.) \text{ s.t. } s_2 = \omega_.(s_1)\}$$

(distance between s_1 and s_2)

OM algorithm

Clustering algorithms using

$$\mathcal{D}(s_1, s_2) = \min_{\omega_\cdot}\{c(\omega_\cdot) \text{ s.t. } s_2 = \omega_\cdot(s_1)\}$$

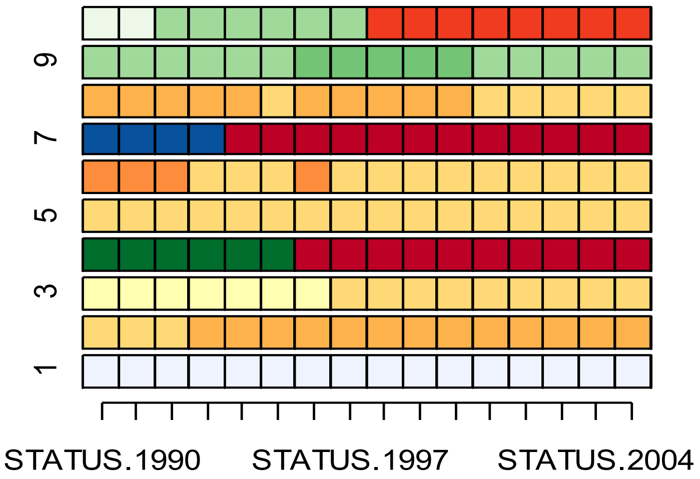
(distance between s_1 and s_2)

Family life trajectories

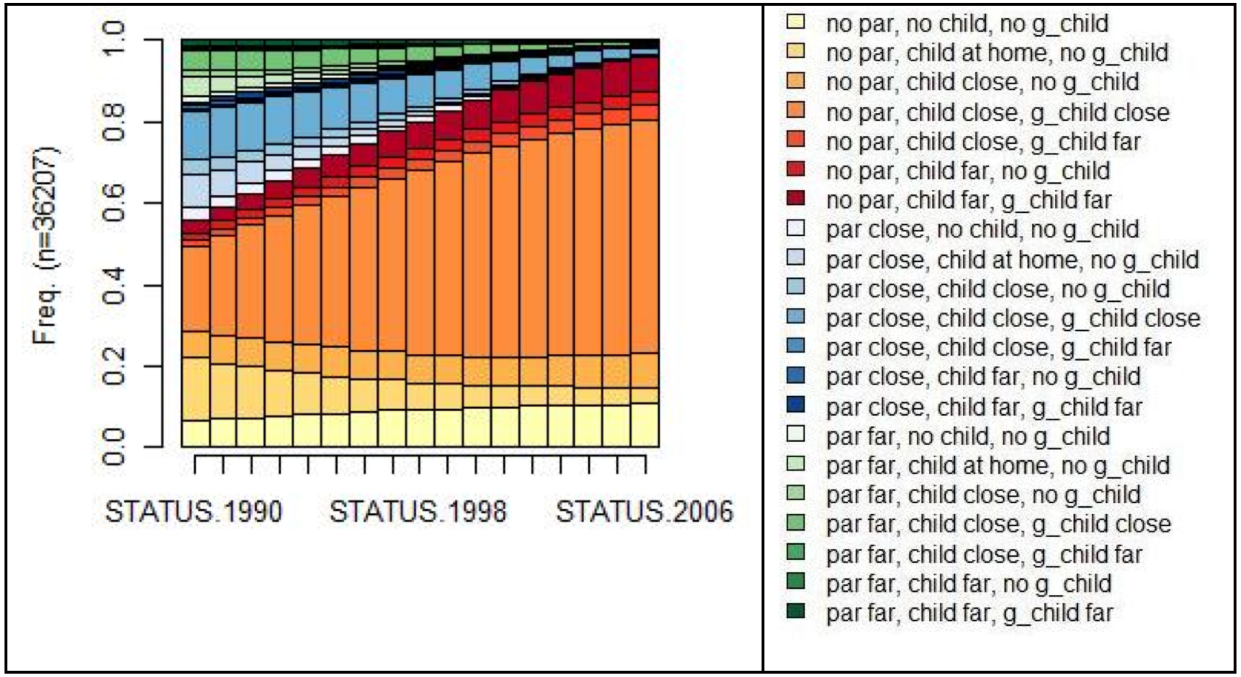
- all women born 1935 in Sweden
- Family trajectories for the period 1990-2006 (17 years)

🟨 no par, no child, no g_child
🟨 no par, child at home, no g_child
🟧 no par, child close, no g_child
🟧 no par, child close, g_child close
🟥 no par, child close, g_child far
🟥 no par, child far, no g_child
🟥 no par, child far, g_child far
⬜ par close, no child, no g_child
🟦 par close, child at home, no g_child
🟦 par close, child close, no g_child
🟦 par close, child close, g_child close
🟦 par close, child close, g_child far
🟦 par close, child far, no g_child
🟦 par close, child far, g_child far
⬜ par far, no child, no g_child
🟩 par far, child at home, no g_child
🟩 par far, child close, no g_child
🟩 par far, child close, g_child close
🟩 par far, child close, g_child far
🟩 par far, child far, no g_child
🟩 par far, child far, g_child far

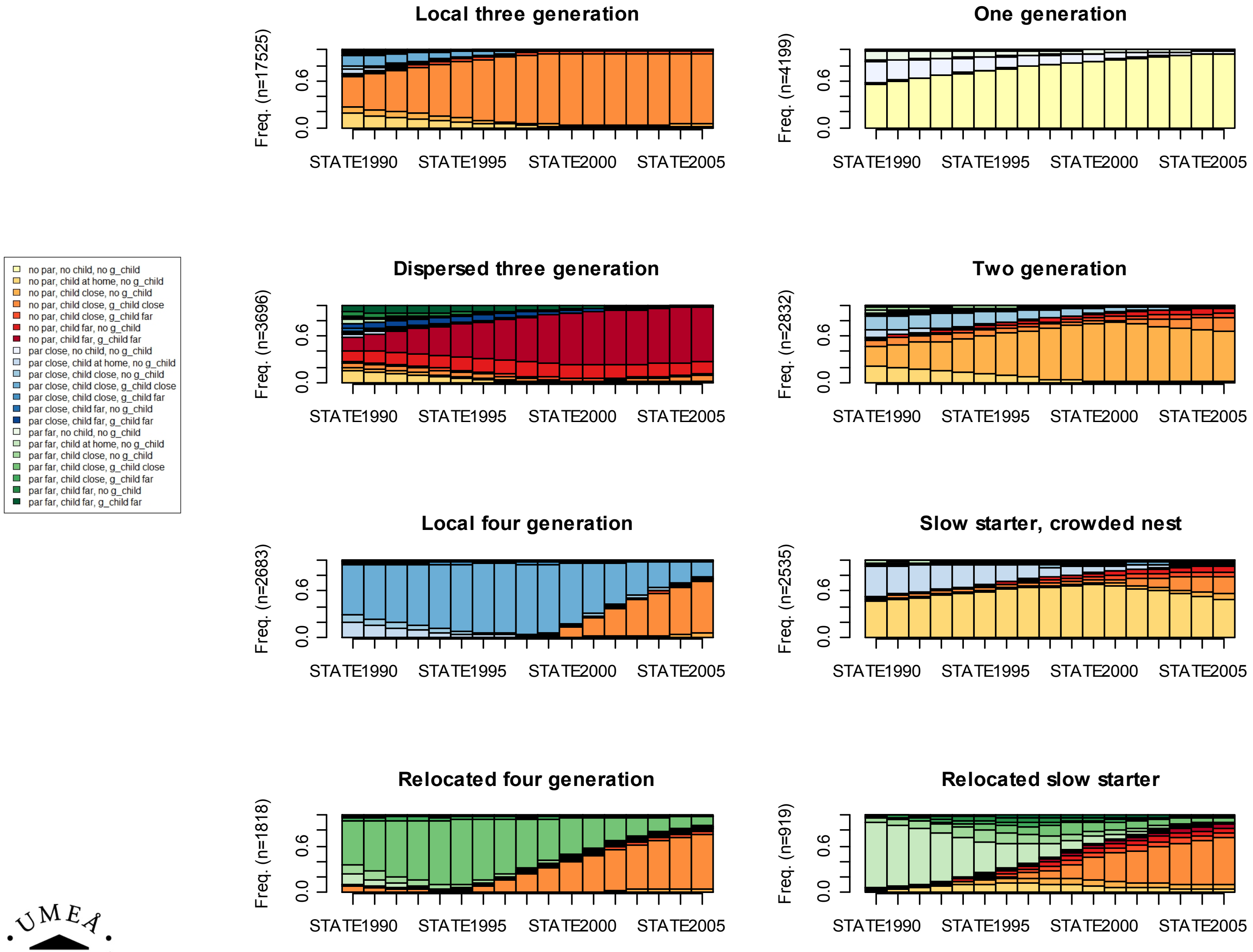
10 first sequences



Family life courses: state frequencies



OMA: clusters of family life trajectories



Description and inference

- Description (dimension reduction)

- Inference: life trajectories are either
  outcomes, covariates or control variables

  → I give examples of all these

- TriMineR package is used for OM

Categories as outcome

Table 2: Results from multinomial regression model of variables associated with family life course categories. Local three generation used as reference category.

Variable	Values	"One generation" versus "Local Three generation"	"Dispersed three generation" versus "Local Three generation"	"Two generation" versus "Local Three generation"	"Local four generation" versus "Local Three generation	"Slow starter crowded nest" versus "Local Three generation	"Relocated four generation" versus "Local Three generation	"Relocated slow starter" versus "Local Three generation"
Region	Sparsely populated (ref)	-	-	-	-	-	-	-
	Accessible countryside	0,742*	0,434**	1,110	1,619*	0,471**	1,010	0,856
	Urban	0,775*	0,314**	1,607*	1,537*	0,380**	1,505	0,884
Marital status	Single (ref)		-	-	-	-	-	-
	Married	0,247**	0,777**	0,737**	0,956	1,001	0,953	1,258*
Education	Low (ref)		-	-	-	-	-	-
	Medium	1,111*	1,483**	1,124*	0.860**	0,947	1,374**	1,908**
	High	1,977**	3,483**	1,318**	0,725**	1,468**	1,851**	5,452**
Income	No (ref)	-	-	-	-	-	-	-
	Yes	0,527**	0.793**	0,712**	1,101	0,536**	0,525**	0,816

** Significant at 1%**
* Significant at 5%

Categories as covariates

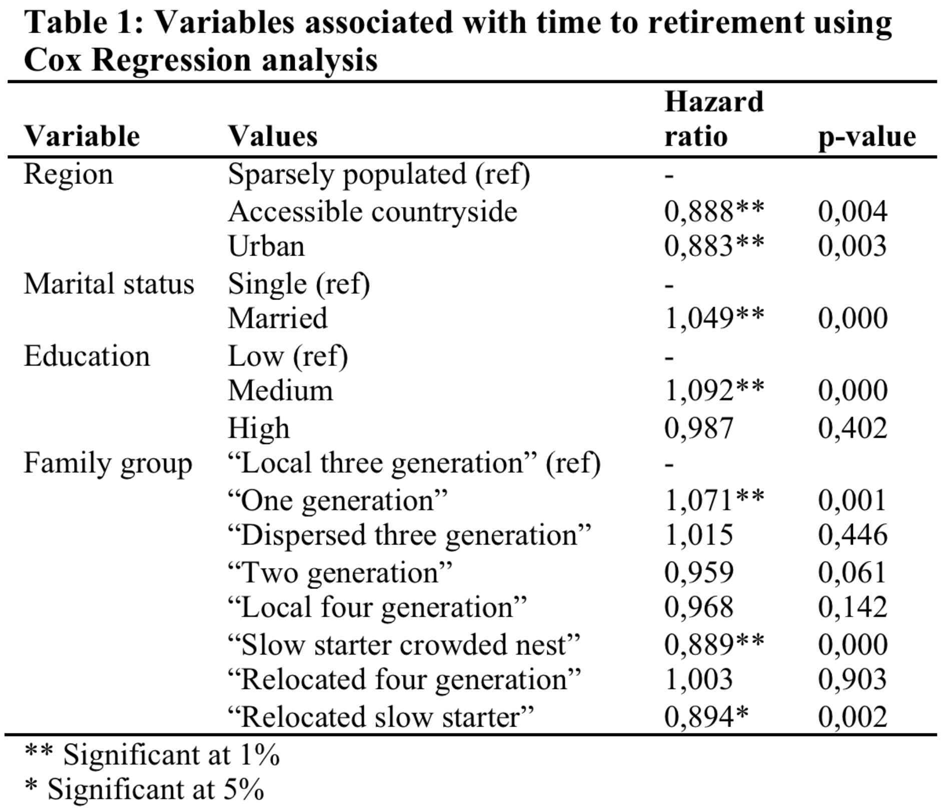


Table 1: Variables associated with time to retirement using Cox Regression analysis

Variable	**Values**	**Hazard ratio**	**p-value**
Region	Sparsely populated (ref)	-	
	Accessible countryside	0,888**	0,004
	Urban	0,883**	0,003
Marital status	Single (ref)	-	
	Married	1,049**	0,000
Education	Low (ref)	-	
	Medium	1,092**	0,000
	High	0,987	0,402
Family group	“Local three generation” (ref)	-	
	“One generation”	1,071**	0,001
	“Dispersed three generation”	1,015	0,446
	“Two generation”	0,959	0,061
	“Local four generation”	0,968	0,142
	“Slow starter crowded nest”	0,889**	0,000
	“Relocated four generation”	1,003	0,903
	“Relocated slow starter”	0,894*	0,002

** Significant at 1%

* Significant at 5%

Effect of retirement timing on health

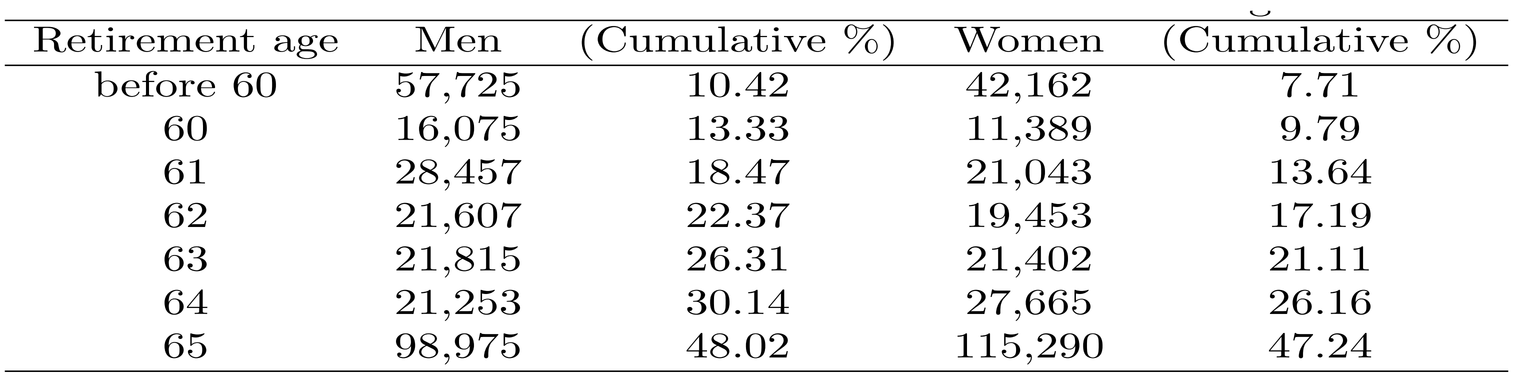
- Difficult to study because health before retirement affect both decision to retire AND health after retirement

- We propose to control for health biographies before retirement timing

Data

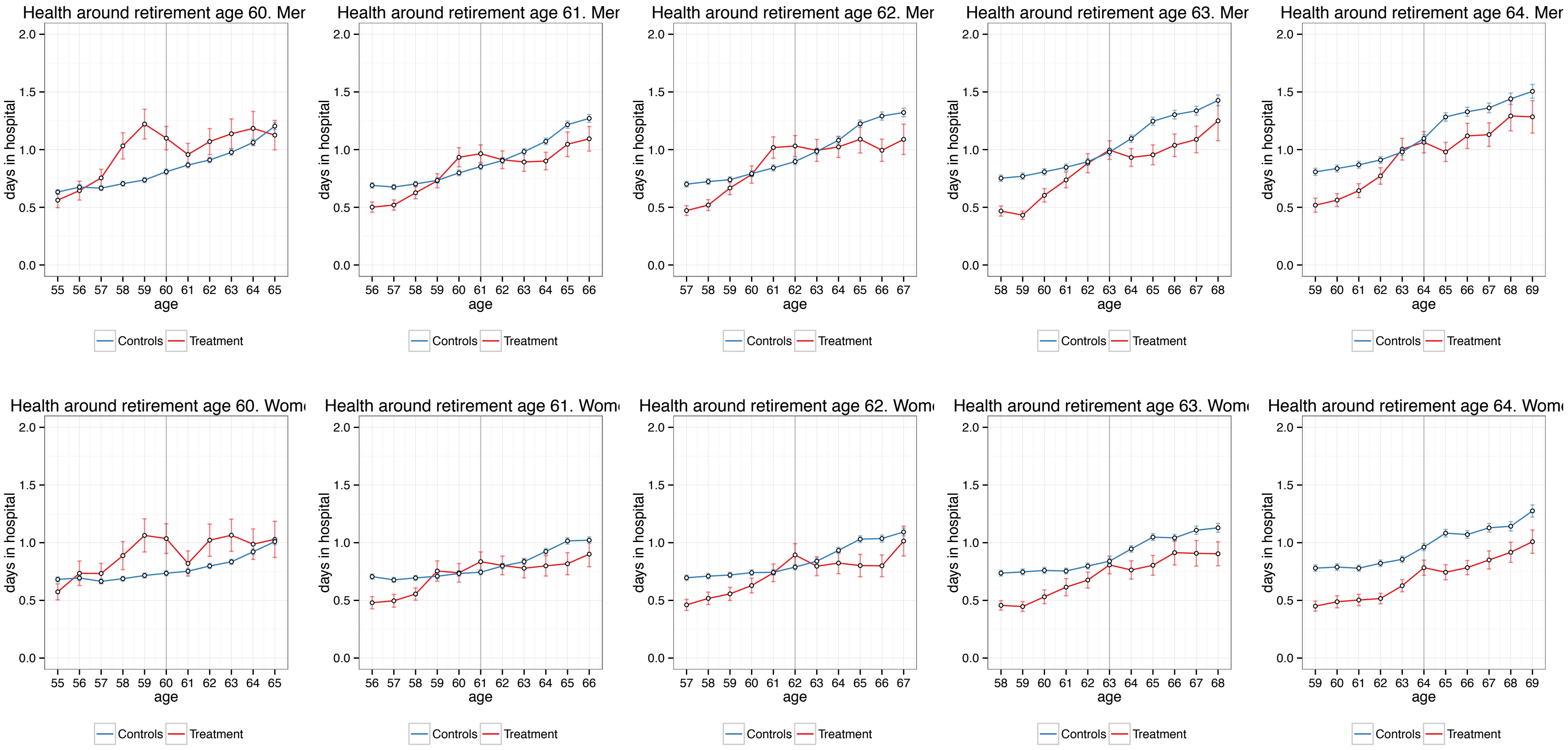
- Sample: born in Sweden 1935→1946 and resident in Sweden 1990
- Follow-up period: 1990 to 2006
- E.g.: 86’054 individuals born 1935

Retirement timing *T*



Retirement age	Men	(Cumulative %)	Women	(Cumulative %)
before 60	57,725	10.42	42,162	7.71
60	16,075	13.33	11,389	9.79
61	28,457	18.47	21,043	13.64
62	21,607	22.37	19,453	17.19
63	21,815	26.31	21,402	21.11
64	21,253	30.14	27,665	26.16
65	98,975	48.02	115,290	47.24

Health: # of days in hospital



Parameter and identification

$T = 1$: retires at age 61
$T = 0$: retires later

Potential outcomes: $Y(T \leftarrow 1)$ and $Y(T \leftarrow 0)$

Average causal effect: $\boxed{\tau = E(Y(1) - Y(0))}$

Parameter and identification

$T = 1$: retires at age 61
$T = 0$: retires later

Potential outcomes: $Y(T \leftarrow 1)$ and $Y(T \leftarrow 0)$

Average causal effect: $\tau = E(Y(1) - Y(0))$

If $\boxed{Y(0), Y(1) \perp\!\!\!\perp T | \mathbf{X} \text{ and } 0 < \Pr(T = 1 \mid \mathbf{X}) < 1}$

then $\boxed{\tau \text{ is identified}}$ from p.d.f.$(Y, T, \mathbf{X})$

$$Y = TY(1) + (1 - T)Y(0)$$

Design of a study by matching

Random sample from p.d.f.$(Y, T, \mathbf{X})$

n_1 retired, n_0 not retired (control)

For each treated unit $i = 1, \ldots, n_1$

pick up a control which has same $\mathbf{X}$.

Retired average outcome: $\bar{Y}_1$

Matched control: $\bar{Y}_0$

$$\boxed{\hat{\tau} = \bar{Y}_1 - \bar{Y}_0}$$

consistent because distr $\mathbf{X}$ is balanced among retirees and control

Dimension reduction

Definition:

$$b(\mathbf{X}) \text{ is a } \underline{\text{balancing score}} \text{ if } T \perp\!\!\!\perp \mathbf{X} | b(\mathbf{X})$$

Result (Rosenbaum & Rubin, 1983):

If $Y(0), Y(1) \perp\!\!\!\perp T | \mathbf{X}$ and $0 < \Pr(T = 1 \mid \mathbf{X}) < 1$

then

$$\boxed{Y(0), Y(1) \perp\!\!\!\perp T | b(\mathbf{X})}$$

Propensity score

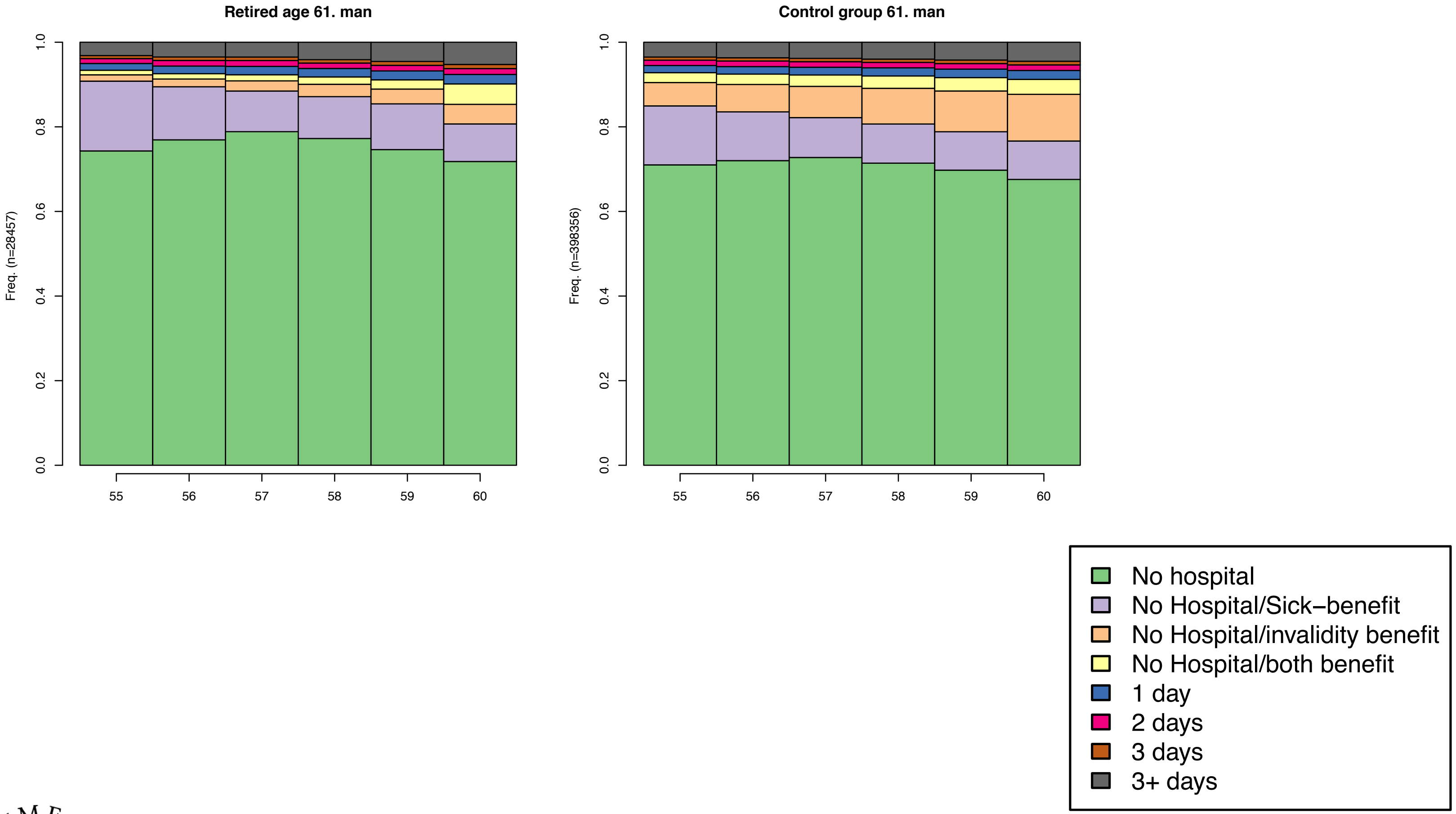
$\mathrm{e(X)}{=}\mathrm{Pr}(T = 1 \mid \mathbf{X})$ is a balancing score

(Rosenbaum & Rubin, 1983)

$\rightarrow$ match for $e(\mathbf{X})$ (a scalar)

need to be modelled and fitted

Health biographies



Design of the study

- We strive at balancing **X** for *T* = 1 and *T* = 0, where

$$\mathbf{X} = \{\mathbf{X}^b, \mathbf{S}\}$$

- We consider three designs by matching on either
  - $e(\mathbf{X}^b)$ (one-to-one matching on the propensity score)
  - $\mathbf{S}$ (one-to-one optimal matching)
  - Both (matching on combined distance)

Matching on combined distance

$\mathcal{D}_e$ and $\mathcal{D}_s$ (propensity score distance and OM distance)

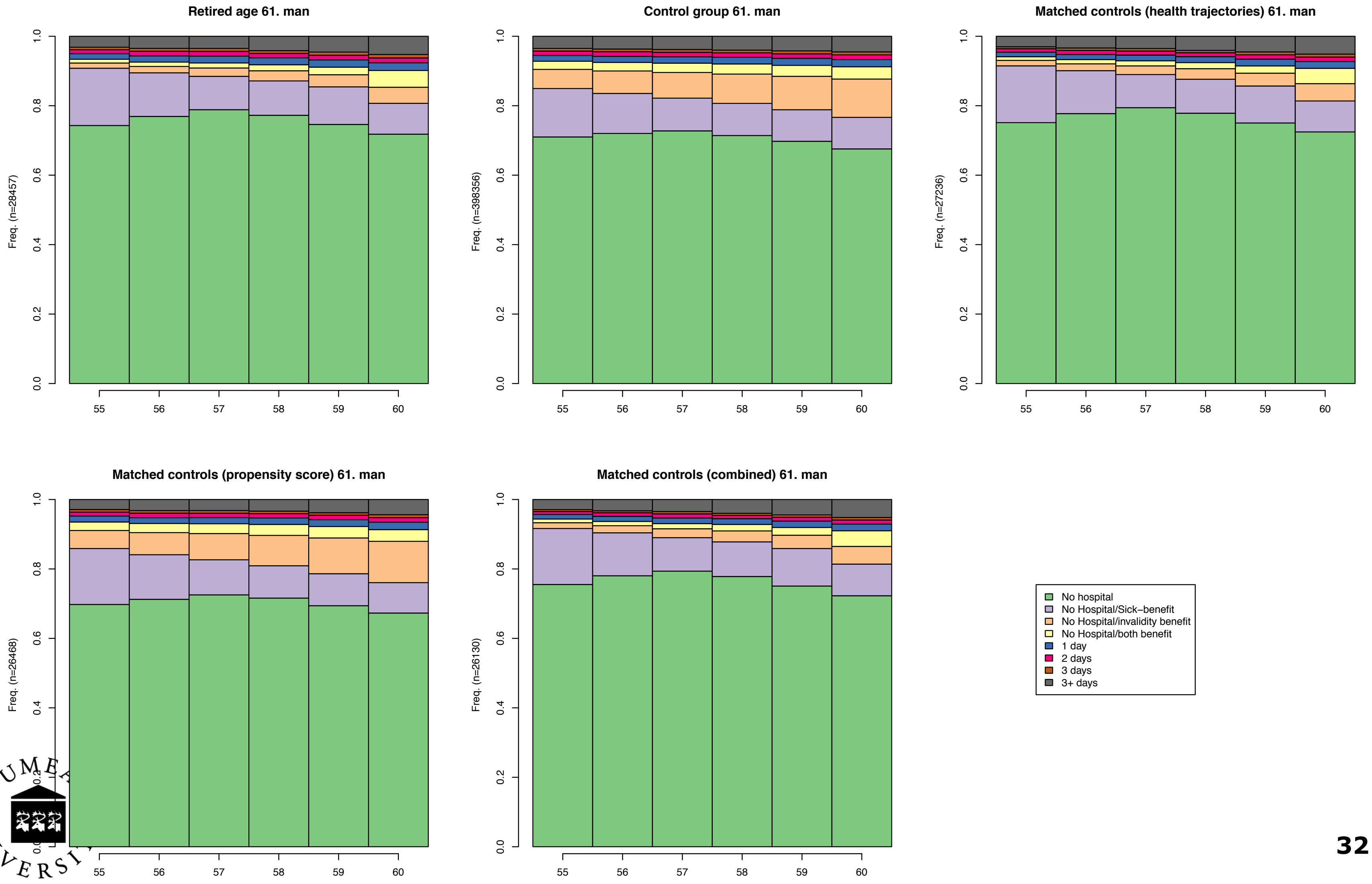
$$\mathcal{D}_c(x_1, x_2) = \frac{1}{\max_{k,l} \mathcal{D}_e(x_k^b, x_l^b)} \mathcal{D}_e(x_1^b, x_2^b) + \frac{1}{\max_{k,l} \mathcal{D}_s(s_k, s_l)} \mathcal{D}_s(s_1, s_2)$$

Balancing properties

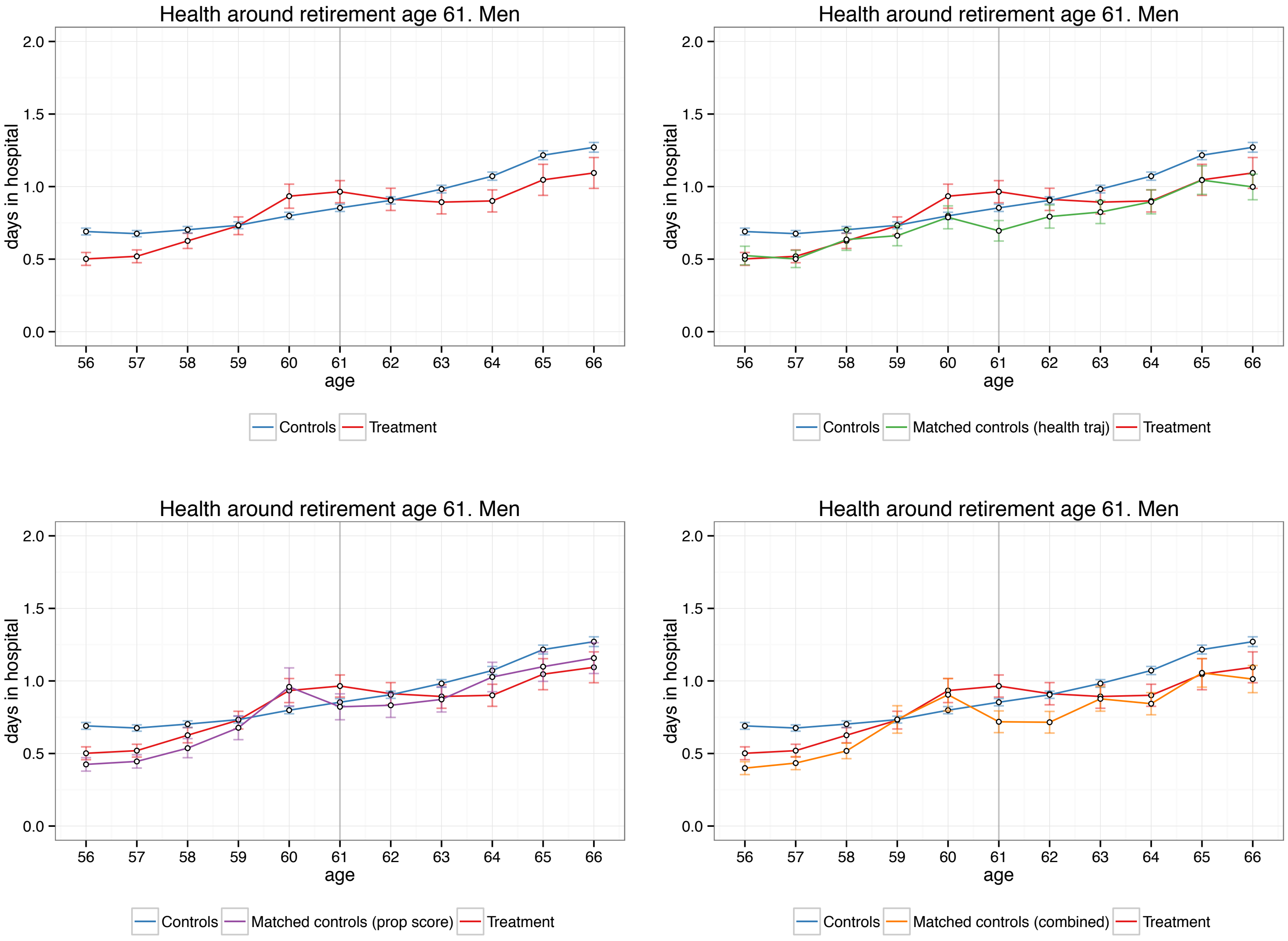
	retirees	controls	p-val	HB match	p-val	PS match	p-val	Comb	p-val
hosp t-5	0.501	0.69	0	0.525	0.561	0.425	0.019	0.399	0.001
hosp t-4	0.52	0.676	0	0.501	0.633	0.445	0.022	0.434	0.008
hosp t-3	0.626	0.703	0.008	0.635	0.841	0.537	0.038	0.518	0.005
hosp t-2	0.73	0.733	0.918	0.662	0.149	0.677	0.312	0.735	0.935
hosp t-1	0.934	0.798	0.002	0.788	0.013	0.959	0.749	0.905	0.683
unempl t-5	0.045	0.094	0	0.095	0	0.04	0.01	0.039	0.001
unempl t-4	0.05	0.104	0	0.109	0	0.048	0.355	0.046	0.089
unempl t-3	0.065	0.111	0	0.117	0	0.067	0.368	0.066	0.675
unempl t-2	0.097	0.115	0	0.124	0	0.115	0	0.112	0
unempl t-1	0.057	0.119	0	0.126	0	0.065	0	0.067	0
low education	0.301	0.453	0	0.437	0	0.317	0	0.322	0
med education	0.427	0.365	0	0.369	0	0.413	0.003	0.416	0.018
high education	0.272	0.182	0	0.194	0	0.269	0.53	0.262	0.017
married	0.715	0.7	0	0.727	0.006	0.732	0	0.74	0
partner retired	0.061	0.042	0	0.041	0	0.063	0.476	0.066	0.021
income*	2597.123	2414.627	0	2513.814	0	2378.221	0	2385.619	0

*(5 years before)

Balancing properties



Balancing and effects



Health around retirement age 61. Men

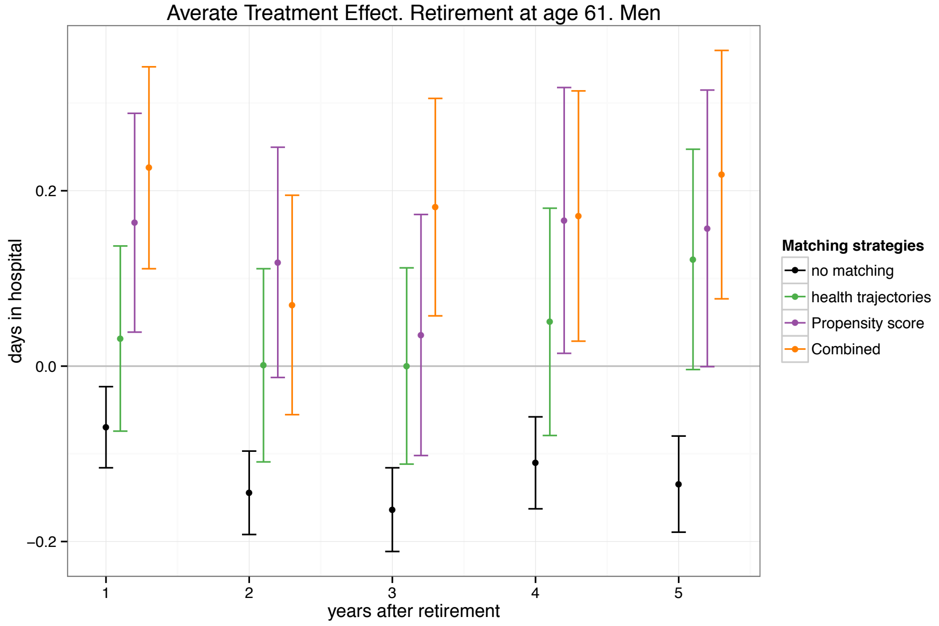
Health around retirement age 61. Men

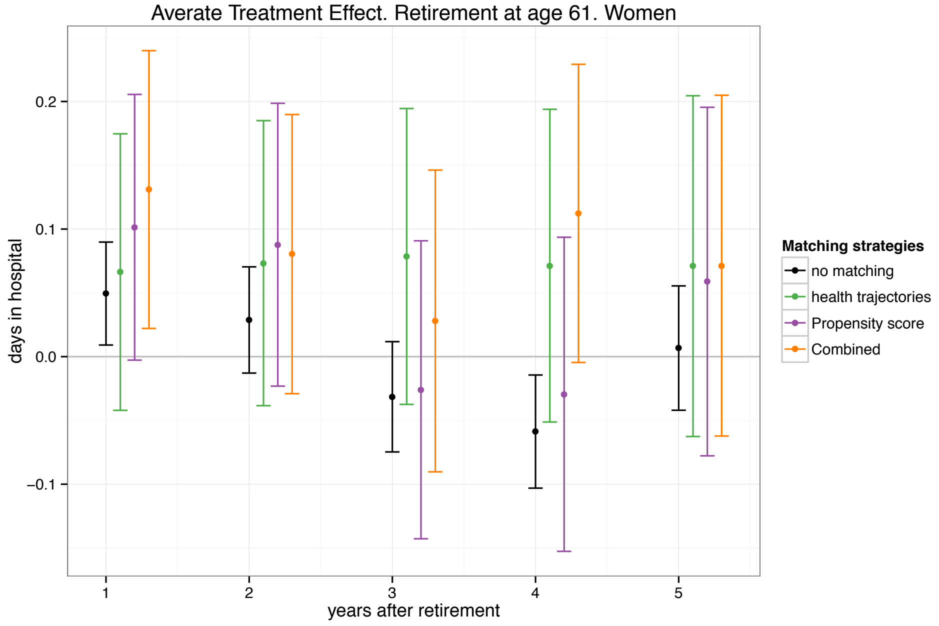
Health around retirement age 61. Men

Health around retirement age 61. Men

No effect on #days in hospital

(DD after matching)

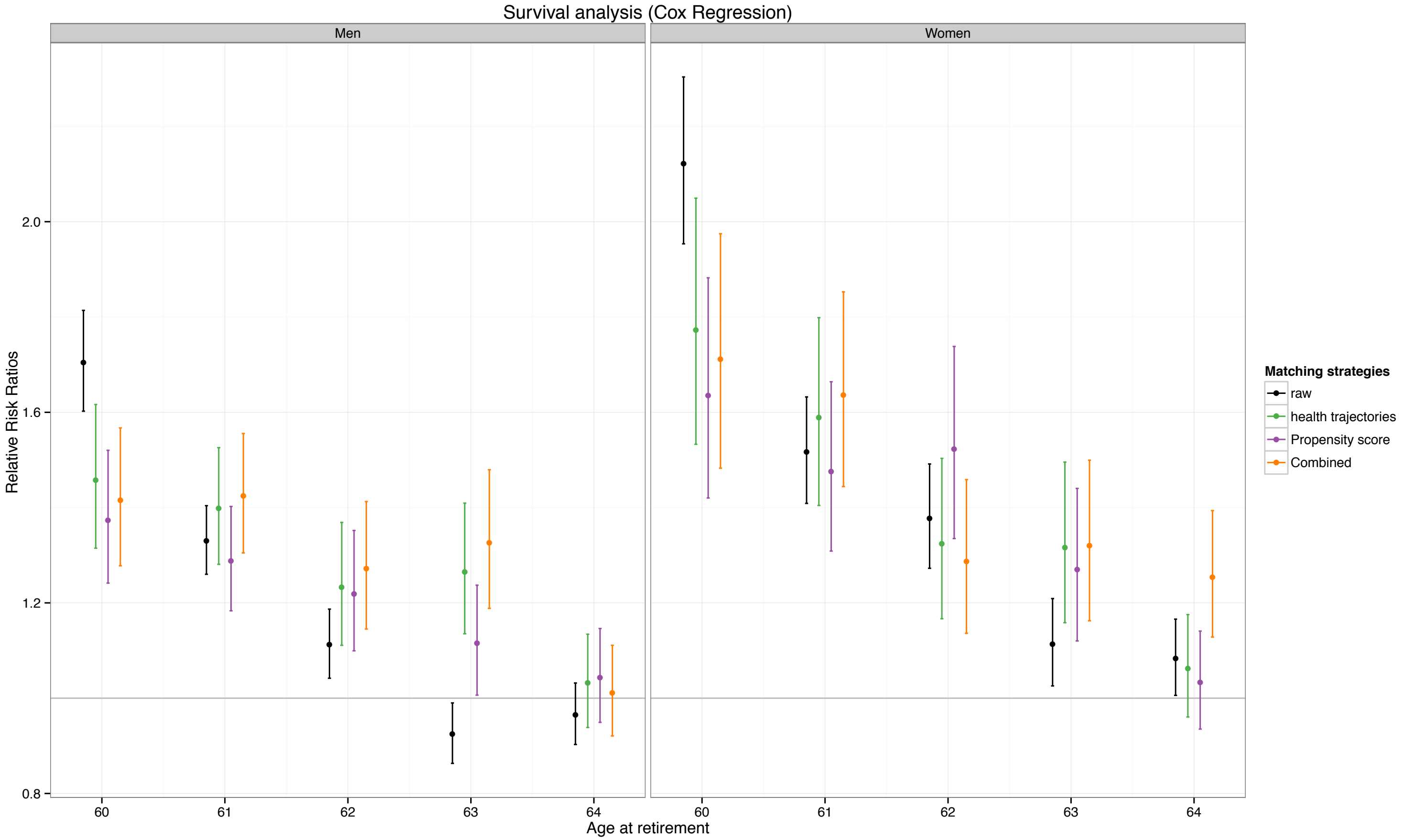




Censoring

- Sequences of different length due to
  - Censoring by beginning or end of follow up
    - Alignment of sequences
    - Truncate sequences to same length when matching
  - Censoring by death
    - Health outcomes not defined after death
    - Analysis on survivors AND survival analysis

Effect on survival



Concluding

Large and rich longitudinal micro-data:

New opportunities:

complex unit of studies: Here biographies

Challenges for statisticians:

descriptive and visualization tools

dimension reduction methods

inference: models and theory