Propensity Score Methods

Thomas Bøjer Rasmussen

Department of Clinical Epidemiology
Aarhus University Hospital

2020-09-09

Slide notes

- Made with `xaringan` package in R.
- pdf version of slides, SAS syntax and macros can be found in the knowledge bank [link]
- In HTML version press "p" to see presenter notes, "h" to see all options.
- Source code including HTML slides sat [github link]

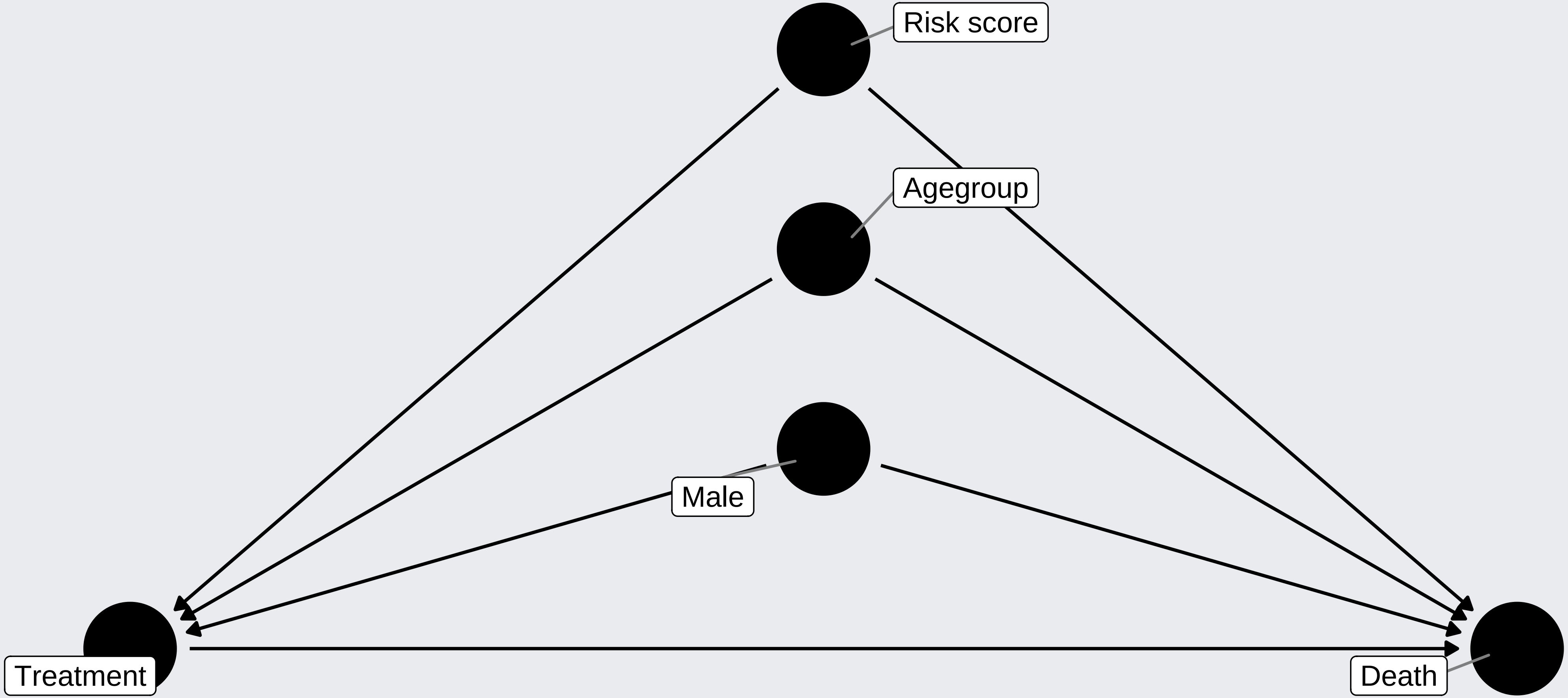
Motivation

- RCT considered the gold standard. Often infeasible and/or unethical
- Observational studies more practical.
- PS methods lets us mimic an RCT
- Focus on time-to-event-outcomes

Running example

- Will use simulated data for running example
- SAS example at the end of presentation.

Study hypothesis



Data excerpt

id	treatment	male	risk_score	agegroup	time	death
1	1	0	0.4738441	65+	365	0
2	0	1	0.2095327	19-64	46	1
3	1	1	0.9251847	19-64	66	1
4	1	1	1.7973623	19-64	237	1
5	0	0	-0.7082472	19-64	81	1
6	0	0	0.7085028	19-64	66	1
7	1	0	1.0341387	0-18	170	1
8	1	1	2.0787144	0-18	118	1
9	0	0	0.3129034	19-64	127	1
10	0	0	-1.7239237	65+	365	0

Descriptive summary

	Untreated	Treated
N (%)	4,991 (49.9)	5,009 (50.1)
Male, N (%)	2,239 (44.9)	2,755 (55.0)
Risk score, median (Q1;Q3)	-0.2 (-0.8;0.5)	0.2 (-0.5;0.9)
Agegroup, N (%)		
0-18	1,343 (26.9)	1,106 (22.1)
19-64	2,578 (51.7)	2,517 (50.2)
65+	1,070 (21.4)	1,386 (27.7)
Time to event, median (Q1;Q3)	160.0 (101.0;236.0)	214.0 (135.0;316.0)
Deaths, N (%)	4,705 (94.3)	4,154 (82.9)

The potential outcomes framework

Treatment indicator Z:

- $Z = 0$: untreated
- $Z = 1$: treated

Pair of potential outcomes:

- Time-to-event under no treatment: Y_0
- Time-to-event under treatment: Y_1

Observed time-to-event: $Y = ZY_1 + (1 - Z)Y_0$

The potential outcomes framework

id	treatment	male	risk_score	agegroup	time	death	time_0	death_0	time_1	death_1
1	1	0	0.4738441	65+	365	0	NA	NA	365	0
2	0	1	0.2095327	19-64	46	1	46	1	NA	NA
3	1	1	0.9251847	19-64	66	1	NA	NA	66	1
4	1	1	1.7973623	19-64	237	1	NA	NA	237	1
5	0	0	-0.7082472	19-64	81	1	81	1	NA	NA
6	0	0	0.7085028	19-64	66	1	66	1	NA	NA
7	1	0	1.0341387	0-18	170	1	NA	NA	170	1
8	1	1	2.0787144	0-18	118	1	NA	NA	118	1
9	0	0	0.3129034	19-64	127	1	127	1	NA	NA
10	0	0	-1.7239237	65+	365	0	365	0	NA	NA

Average treatment effect (ATE)

Absolute effect: risk difference at time t: $F_{Y_1}(t) - F_{Y_0}(t)$

Relative effect: Hazard ratio at time t: $\lambda_{Y_1}(t)/\lambda_{Y_0}(t)$

Average treatment effect in the treated (ATT)

- Average treatment effect among patient who got the treatment
- RCT: ATE $\neq$ ATT.

ATT or ATE?

So what should we estimate? Depends on the context.

Example:

Effect of treatment on mortality.

- If the purpose of the study is to investigate whether or not the treatment works in patients that get it, the ATT is probably the effect of interest.
- If the purpose of the study is to investigate if the treatment should be applied to everyone with the disease, then the ATE is probably of more interest.

Marginal and conditional treatment effects

Conditional treatment effect: effect on individual level

- Regression models

Marginal treatment effect: effect on population level

- RCT
- PS methods

Non-collapsible effect estimator: marginal effect $\neq$ conditional effect

The hazard ratio and odds-ratio are non-collapsible.

No wrong or right, just different eff

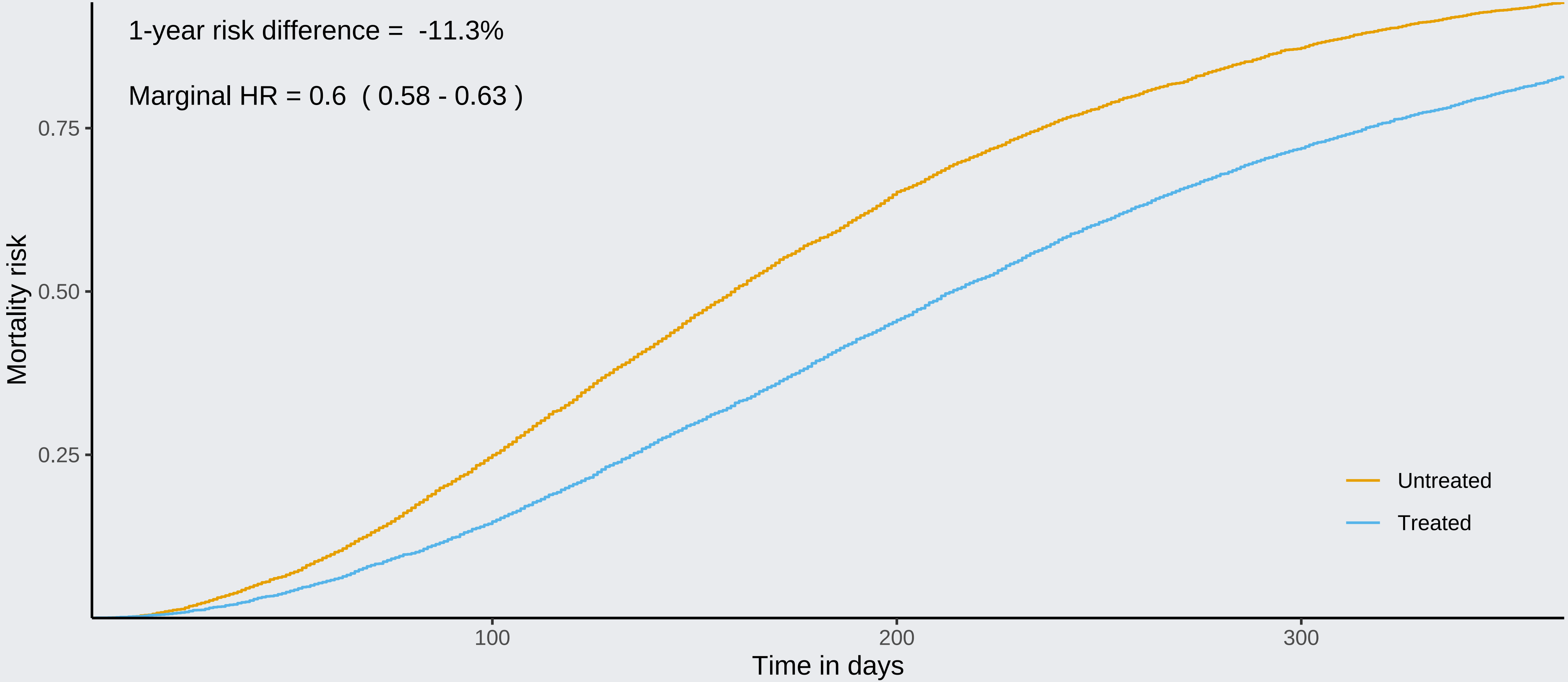
Estimate treatment effect in RCT

Randomization $=> F_{Y_i}(t) = F_{Y|Z=i}(t)$

$F(t) = 1 - S(t)$, estimate S(t) with KM estimate.

Do Cox regression on treatment to estimate relative rate of outcome

Analyze as RCT



1-year risk difference = -11.3%

Marginal HR = 0.6 (0.58 - 0.63)

Propensity scores

The propensity score (PS) is the conditional probability of treatment Z, given variables X:

$$PS = P(Z = 1|X)$$

The PS is a balancing score, ie

$$X \perp Z|PS$$

Estimation of PS's

- True propensity score
  - RCT: ps = 0.5
  - observation study: unknown
- Typically estimate using logistic regression

$$logit(P(Z=1|X)) = \alpha_0 + \alpha X$$

Include all (potential) confounders in model

Model is evaluated as whether or not it achieves balance.

Propensity score methods

We will concentrate on

- PS stratification 👎
- PS adjustment 👎
- PS matching 👍
- PS weighting 👍

PS matching

- 1:1 match on PS
- Nearest neighborhood matching with replacement on log(ps) using caliper (SAS macro)
- X balanced in matched population
- Matched cohort can be used to estimate ATT

PS matched population

Matched population

match	id	treatment	male	risk_score	agegroup	time	death	ps
1	9268	1	0	-0.30021524	19-64	240	1	0.4090981
1	9617	0	1	-1.31705570	19-64	282	1	0.4091391
2	9157	1	1	0.67779207	19-64	148	1	0.6155434
2	2952	0	1	-0.01347750	65+	41	1	0.6155073
3	2295	1	0	-0.09935415	65+	365	0	0.5017525
3	4265	0	1	-1.11627865	65+	285	1	0.5017862
4	2199	1	0	1.87860298	0-18	134	1	0.5918736
4	4349	0	1	-0.24857926	65+	183	1	0.5918785
5	9792	1	1	-0.31972265	65+	213	1	0.5846382
5	994	0	1	-0.31861877	65+	84	1	0.5847509

PS weighting

- Use PS to create weighted pseudo-population with covariate balance population
  - ATE weights: $\frac{Z}{ps} + \frac{1-Z}{1-ps}$
  - ATT weights: $Z + (1-Z)\frac{ps}{1-ps}$
- Weight trimming/truncating should be avoided.

ATE-weighted population

ATE-weighted population

id	treatment	male	risk_score	agegroup	time	death	ps	ate_weight
1	1	0	0.4738441	65+	365	0	0.5616492	1.780471
2	0	1	0.2095327	19-64	46	1	0.5680550	2.315110
3	1	1	0.9251847	19-64	66	1	0.6398295	1.562916
4	1	1	1.7973623	19-64	237	1	0.7193214	1.390199
5	0	0	-0.7082472	19-64	81	1	0.3683870	1.583248
6	0	0	0.7085028	19-64	66	1	0.5140339	2.057757
7	1	0	1.0341387	0-18	170	1	0.5042217	1.983254
8	1	1	2.0787144	0-18	118	1	0.7074906	1.413446
9	0	0	0.3129034	19-64	127	1	0.4725093	1.895768
10	0	0	-1.7239237	65+	365	0	0.3372429	1.508848

ATT-weighted population

ATT-weighted population

id	treatment	male	risk_score	agegroup	time	death	ps	ate_weight
1	1	0	0.4738441	65+	365	0	0.5616492	1.0000000
2	0	1	0.2095327	19-64	46	1	0.5680550	1.3151097
3	1	1	0.9251847	19-64	66	1	0.6398295	1.0000000
4	1	1	1.7973623	19-64	237	1	0.7193214	1.0000000
5	0	0	-0.7082472	19-64	81	1	0.3683870	0.5832479
6	0	0	0.7085028	19-64	66	1	0.5140339	1.0577565
7	1	0	1.0341387	0-18	170	1	0.5042217	1.0000000
8	1	1	2.0787144	0-18	118	1	0.7074906	1.0000000
9	0	0	0.3129034	19-64	127	1	0.4725093	0.8957682
10	0	0	-1.7239237	65+	365	0	0.3372429	0.5088484

PS analysis assumptions

- Consistency
- Conditional exchangeability
- SUTVA
- Positivity
- Correct specification of models

Consistency

Consistency: $A = a => Y_a = Y$

Untestable

ignorable treatment Assignment / No unmeasured confounding

No measured confounding

Untestable

Stable Unit Treatment Value Assumption (SUTVA)

No interference between subjects

Untestable

Positivity

Positivity: $0 < PS < 1$

Unmatched treated patients and large/small weights indicates problems

Positivity issues in matched population?

Population	# treated patients
Matched	5005
Original	5009

Positivity issues in weighted populations?

Population	Treatment	n	min	p1	p99	max	mean	stddev
ATE-weighted	Untreated	4991	1.1619737	1.2858763	3.789460	5.902530	2.002595	0.5173061
ATE-weighted	Treated	5009	1.1603368	1.2766289	3.664038	5.533381	1.996498	0.5113759
ATT-weighted	Untreated	4991	0.1619737	0.2858763	2.789460	4.902530	1.002595	0.5173061
ATT-weighted	Treated	5009	1.0000000	1.0000000	1.000000	1.000000	1.000000	0.0000000

Correct specification of models

Outcome models need to be correctly specified

Correct PS model specification is sufficient but not actually necessary!

Balance assessment

Need to assess if covariate balance have been achieved after matching/weighting

Fine-tune PS model

Review populaiton exlusion/inclusion criterias

Repeat until balance!

Assess balance

Assessment of balance

- "Table 1" of weighted / matched population
- Look at at ps-distribution
- Standardized differences
- Weight statistics
- Empirical CDF

Assessment in stratas of confounders.

Table 1 in matched/weighted sample

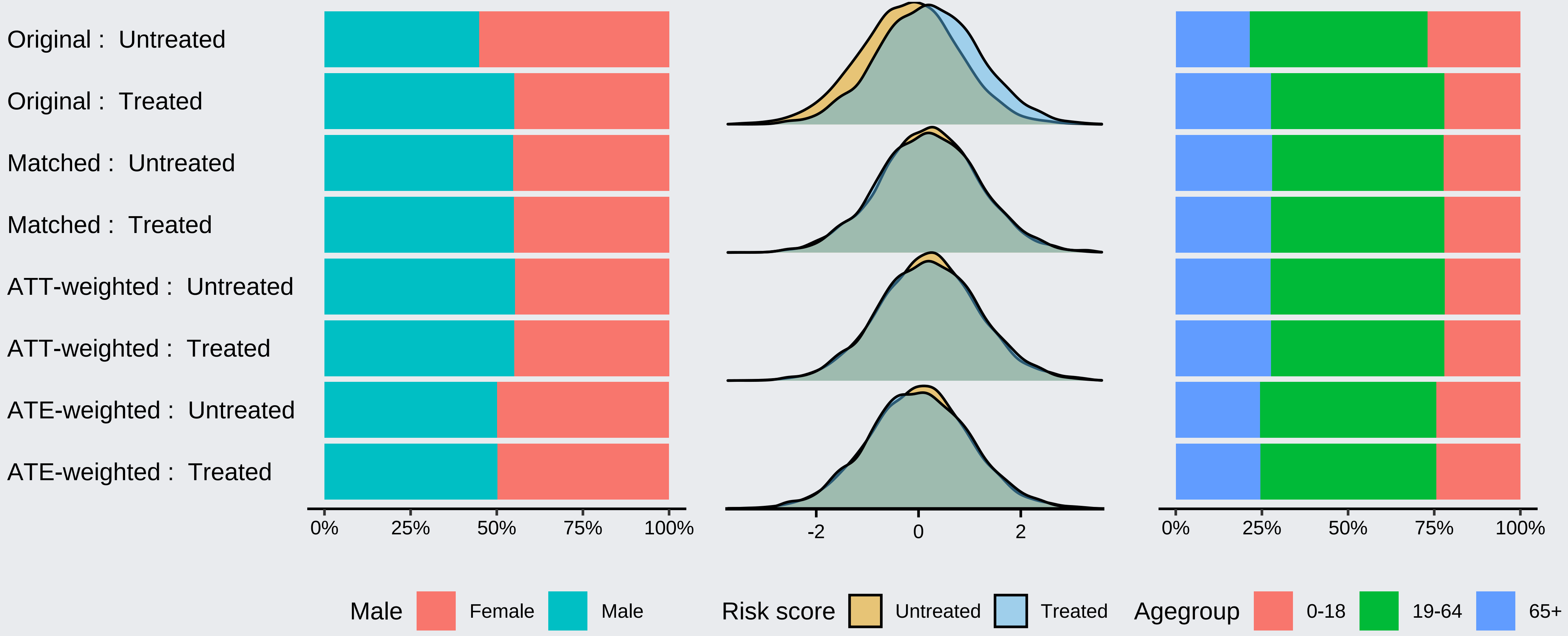
Highly subjective

If you calculate p-values to test balance we can't be friends.

Descriptive summary tables

	Original		Matched		ATT-weighted		ATE-weighted	
	Untreated	**Treated**	**Untreated**	**Treated**	**Unteated**	**Treated**	**Untreated**	**Treated**
N (%)	4,991 (49.9)	5,009 (50.1)	5,005 (50.0)	5,005 (50.0)	5,004 (50.0)	5,009 (50.0)	9,995 (50.0)	10,000 (50.0)
Male, N (%)	2,239 (44.9)	2,755 (55.0)	2,736 (54.7)	2,751 (55.0)	2,763 (55.2)	2,755 (55.0)	5,002 (50.0)	5,012 (50.1)
Risk score, median (Q1;Q3)	-0.2 (-0.8;0.5)	0.2 (-0.5;0.9)	0.2 (-0.4;0.9)	0.2 (-0.5;0.9)	0.2 (-0.4;0.8)	0.2 (-0.5;0.9)	0.0 (-0.7;0.7)	0.0 (-0.7;0.7)
Agegroup, N (%)								
0-18	1,343 (26.9)	1,106 (22.1)	1,115 (22.3)	1,106 (22.1)	1,099 (22.0)	1,106 (22.1)	2,442 (24.4)	2,440 (24.4)
19-64	2,578 (51.7)	2,517 (50.2)	2,492 (49.8)	2,516 (50.3)	2,529 (50.5)	2,517 (50.2)	5,107 (51.1)	5,109 (51.1)
65+	1,070 (21.4)	1,386 (27.7)	1,398 (27.9)	1,383 (27.6)	1,376 (27.5)	1,386 (27.7)	2,446 (24.5)	2,451 (24.5)

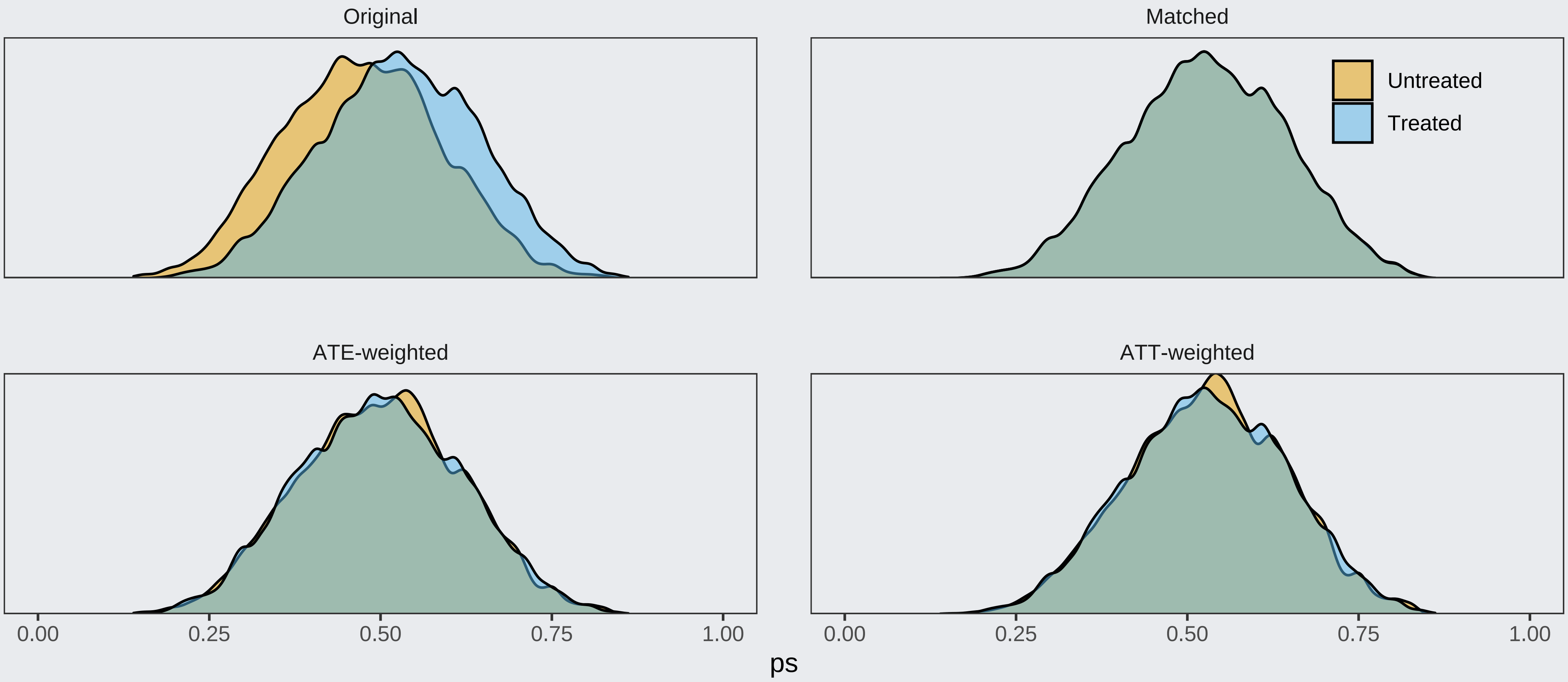
Descriptive summary graph



PS distribution

- Check lack of overlap in ps-distribution before matching/weighting
- Quickly assess if something is very wrong with ps-model
- Does not really say much about covariate balance

PS distribution



Standardized Differences

$$SD = \frac{|\bar{x}_t - \bar{x}_c|}{\sqrt{\frac{s_t^2 + s_c^2}{2}}}$$

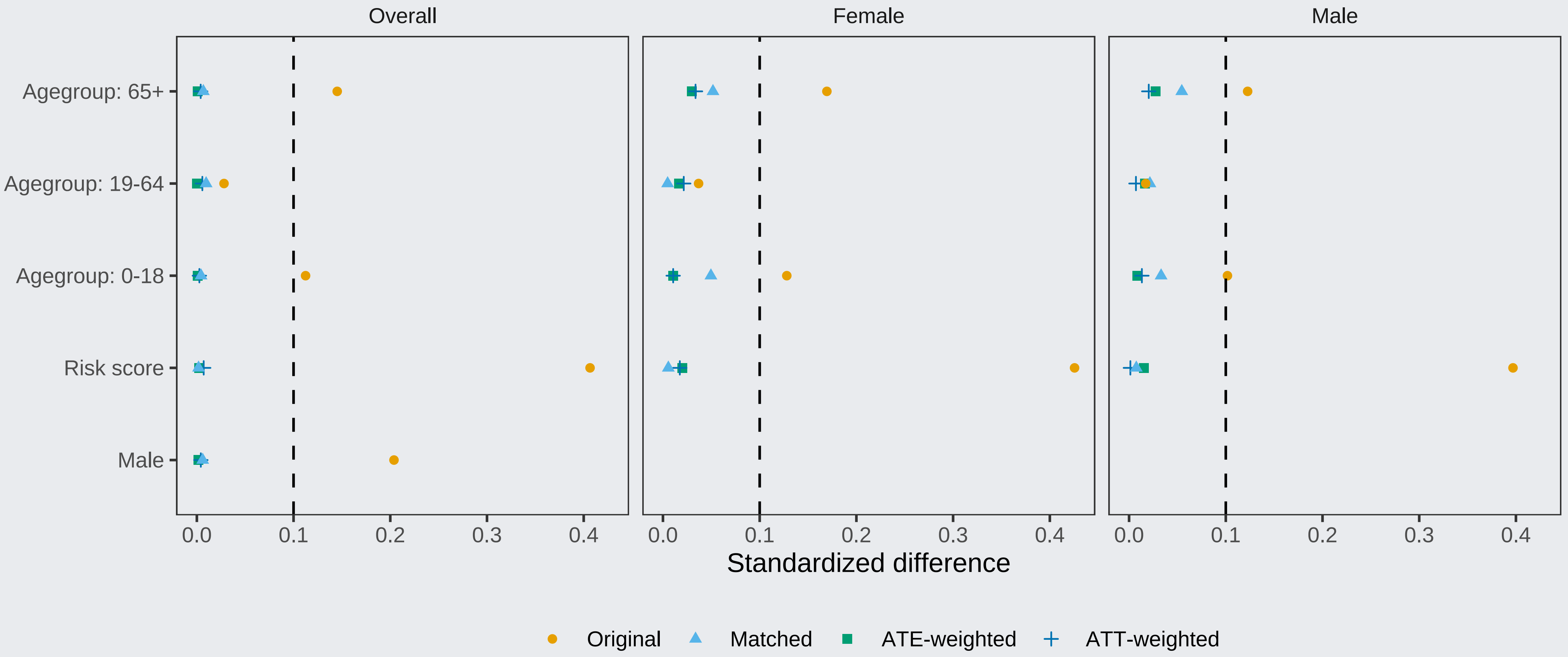
where

$$\bar{x} = \frac{1}{\sum_i w_i} \sum_i w_i x_i$$

$$s^2 = \frac{\sum_i w_i}{(\sum_i w_i)^2 - \sum_i w_i^2} \sum_i w_i (x_i - \bar{x})^2$$

- Not influenced by sample size.
- Can be used for both continuous and dichotomous variables.
- Categorical variable?
- Only compares means of distributions!

Standardized differences



Weight statistics

If PS-weighting is used, calculate weight statistics

Large weight are only problematic if they are large relative to the population size. For example:

- max weight = 10, N = 1 million. Irrelevant
- Max weight = 10, N = 100. Not good!

Weight distribution

pop	treatment	n	min	p1	p99	max	mean	stddev
ATE-weighted	Untreated	4991	1.1619737	1.2858763	3.789460	5.902530	2.002595	0.5173061
ATE-weighted	Treated	5009	1.1603368	1.2766289	3.664038	5.533381	1.996498	0.5113759
ATT-weighted	Untreated	4991	0.1619737	0.2858763	2.789460	4.902530	1.002595	0.5173061
ATT-weighted	Treated	5009	1.0000000	1.0000000	1.000000	1.000000	1.000000	0.0000000

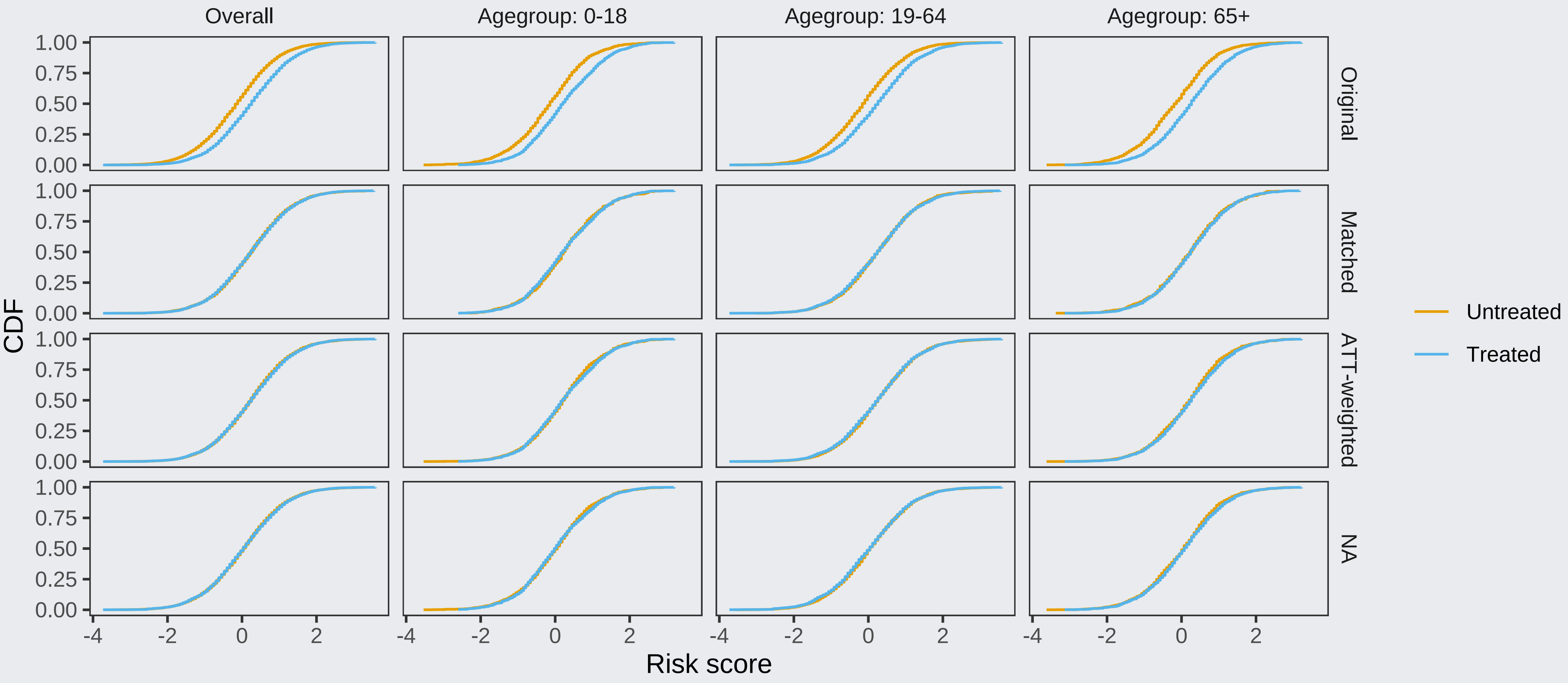
Empirical CDF

- Balance in mean is not enough. Should also look at other moments of the distribution, eg the variance.
- Empirical CDFs are a straight forward way to compare the distribution of continuous variables.

$$CDF(x) = \frac{1}{\sum_i w_i} \sum w_i I(x_i \leq x)$$

where w_i is the weight and I the indicator function.

Empirical CDF



Separation of design and analysis of study

- Only when we are satisfied with the covariate balance do we move on to estimating the treatment effect
- Do analyses in matched/weighted populations and treat as RCT

Variance estimation

- After matching/weighting observations are no longer independent
- Should be accounted for in analyses. Different approaches. We will use bootstrapping here.

Bootstrapping

Resample method.

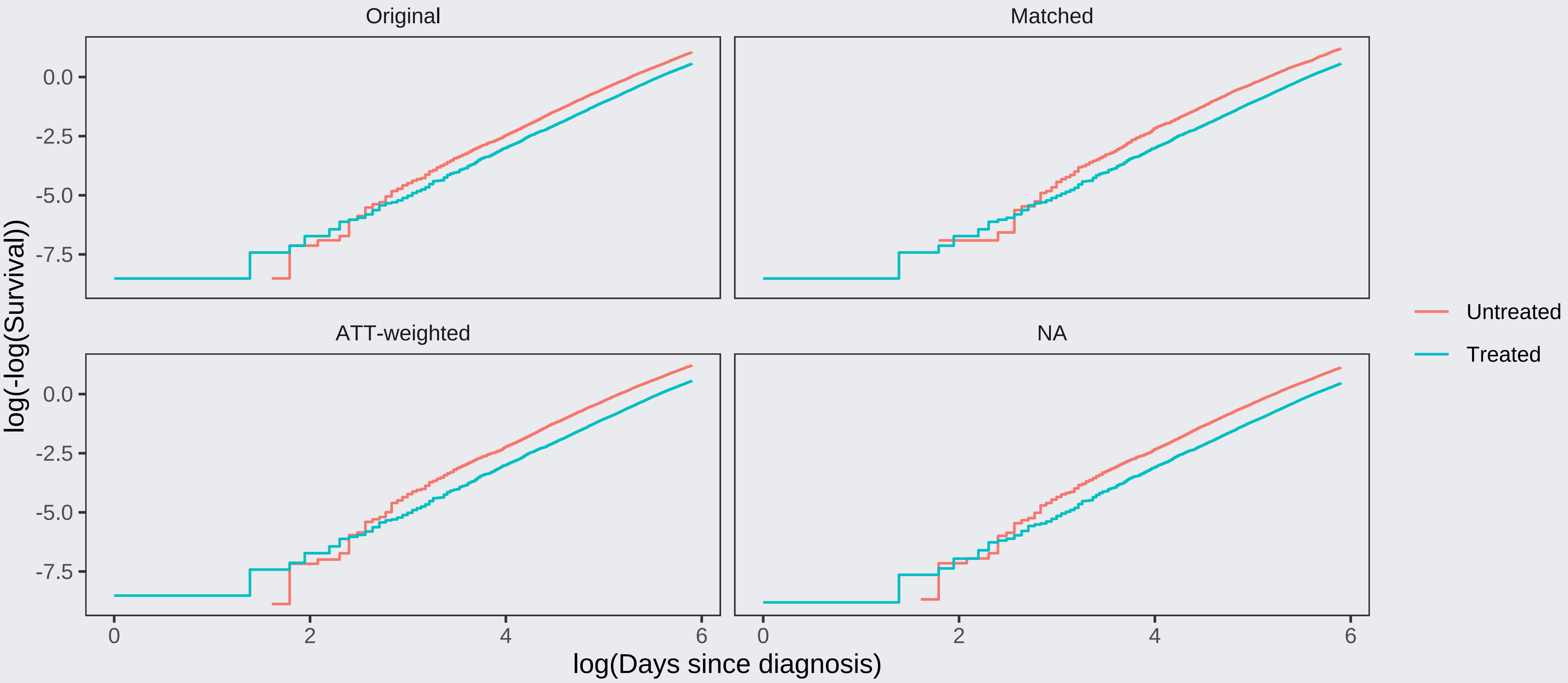
Pros:

- Easy
- Proper CI coverage
- Can always be done no matter the analysis.

Cons

- Can be done in many different variations.
- Infeasable in practice if many analyses are done.

Assess PH assumption



Estimate relative treatment effect

pop	effect	hr_ci
ATE-weighted	Marginal	0.5 (0.48 - 0.52)
ATT-weighted	Marginal	0.5 (0.48 - 0.52)
Matched	Marginal	0.5 (0.48 - 0.53)
Original	Conditional	0.45 (0.43 - 0.47)
Original	Marginal	0.6 (0.58 - 0.63)

Estimate absolute treatment effect

pop	cum_inc_ci
ATE-weighted	-0.16 (-0.17 ; -0.14)
ATT-weighted	-0.14 (-0.15 ; -0.13)
Matched	-0.14 (-0.15 ; -0.12)
Original	-0.11 (-0.12 ; -0.1)

PS-matching pros and cons

Pros:

- Easy to analyze
- Easier to report analysis/results to collaborators and readers

Cons:

- Non-trivial to make the matched population.
- Can also be used to estimate ATE but it is less straight forward.
- Removal of patients from the population => ATT not estimated

PS weighting pro/cons

Pros:

- Easy to calculate weights to estimate eg ATT and ATE
- Efficient
- Has to explicitly handle extreme weight instead of simply removing the problem

Cons:

- Software might not have weight options for the analysis that has to be made => catastrophe!

PS methods vs regression analyses

PS methods allows for estimating effects as in RCT

blinding oneself to the outcomes

If more exposures than outcomes

More obvious if populations are even comparable. Regression will extrapolate to make things work, eg you wont necesarily notice that only men in one tretment and only women in the other.

Advanced topics

- Alternative methods to estimae PS's (GBM etc, decision trees, other GLMS etc)
- More than two treatments generalization (pair-wise compare with ref)
- double robust approaches (adjust for covariates again in outcome model)
- Advanced matching/stratification approaches (and why it is probably not worth it 99% of the time).
- Stabilized ATE weights.

Common misconceptions

- PS adjusts for confounding better than regression. marginal vs conditional fallacy.

References

Austin, P. C. (2014). "The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments".

Austin, P. C. and E. A. Stuart (2015). "Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies".

Rosenbaum, P. R. and D. B. Rubin (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects".