Adjustment for confounding in practice

Raphaël Porcher

University Paris Descartes

MiRoR, Ghent

Causal effect

- We are interested in determining the effect of some "treatment" A on the outcome Y
- "Treatment" = not only a drug, but any exposure
- Effect is intended as compared to some control condition
- This is the aim of RCTs!
- But RCTs are not always feasible

Causal inference in observational studies

- Usually regarded as not providing unbiased estimates of the causal effect
- Because of confounding, as you have see previously
- Confounding:
  - $Y(a)$ likely to depends on L, and A as well
  - So $\{Y(1), Y(0)\}$ is no more independent of A
  - It is easy to estimate $\mathrm{E}\{Y(a)|A = a, L\}$ but not $\mathrm{E}\{Y(a)\}$
- Some solutions exist under various assumptions regarding the distribution of (A, L)

Usual analysis options

- Stratification and matching
- Regression analysis (adjustment)
- Propensity scores
- Some other methods (IV, . . .)

Balance

- Distribution of confounders similar in treated and untreated patients
- Can be assessed by looking at the distribution of confounders in both groups
- Imbalance can cause confounding

Example of confounding: Simpson's paradox

- NRS comparing treatments to remove kidney stones[1]
- Compare open surgery (A) *vs* percutaneous nephrolithotomy (B)

Population	A	B	Difference (95% CI)
Overall, *N*	350	350	
Success	273 (78%)	289 (83%)	-5% (-10 to $+1$)
Stones $<$ 2 cm, *N*	87	270	
Success	81 (93%)	234 (87%)	$+6\%$ (-2 to $+12$)
Stones $\geq$ 2 cm, *N*	263	80	
Success	192 (73%)	55 (69%)	$+4\%$ (-6 to $+16$)

[1] Charig et al. *BMJ* 1986;292: 879–82; Julious & Mullee *BMJ* 1994;309:1480

Noncollapsibility ($\neq$ confounding)

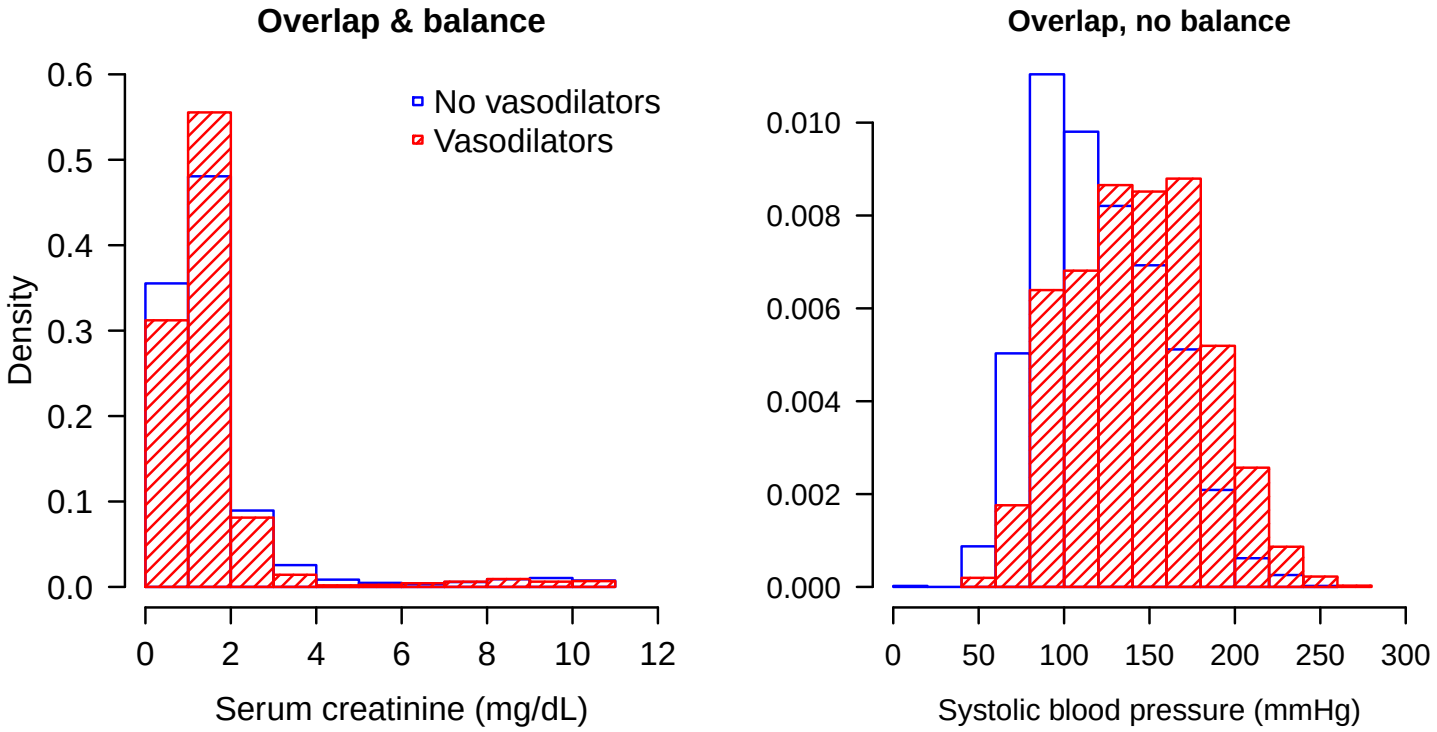
- Take a RCT of A *vs* B stratified on L

	Size < 2 cm ($L = 1$)		Size ≥ 2 cm ($L = 0$)		All (marginal)	
	$A = 1$	$A = 0$	$A = 1$	$A = 0$	$A = 1$	$A = 0$
$Y = 1$, N	80	60	40	20	120	80
$Y = 0$, N	20	40	60	80	80	120
Success rate	80%	60%	40%	20%	60%	40%
ARD	20%		20%		20%	
RR	1.33		2.00		1.50	
OR	2.67		2.67		2.25	

Overlap

- Overlap of the distributions (overlapping support)
- Lack of overlap implies extrapolating results
- Different from balance

Example of balance and overlap: ALARM study



Stratification

- Group together patients with same values of L
- Estimate the treatment effect in each subgroup
- Pool the results
- L is a vector of covariates $\rightarrow$ many subgroups
- Very difficult to use when there are many confounders

Regression model (for the outcome)

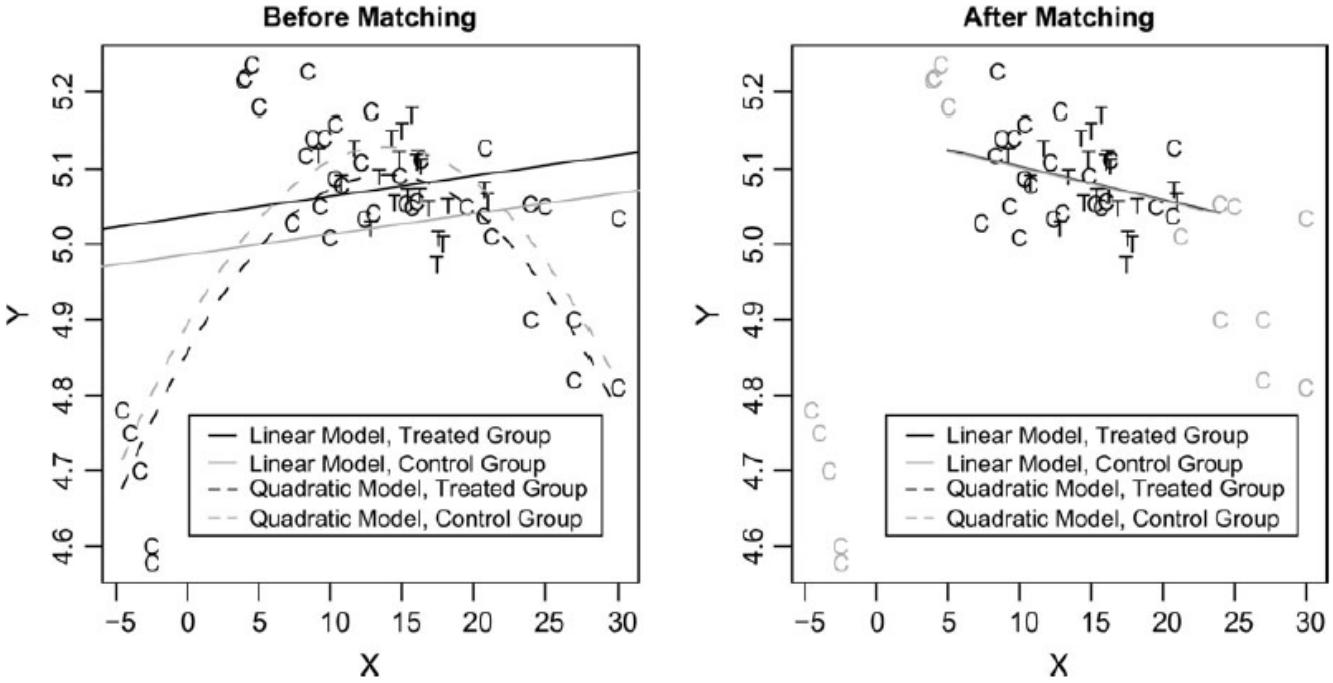
- Usually a linear regression model

$$\mathrm{E}(Y|A, L) = \beta_0 + \beta_1 A + \beta_2 L$$

- $\hat{\beta}_1$ is the estimate of the treatment effect
- What is behind?
  - Constant treatment effect, normally distributed residual errors, common slope on L
  - Estimates the conditional treatment effect
  - May be subject to curse of dimensionality (even more with nonlinear effects, interactions)
- Some assumptions may not hold, or may be unverifiable (in particular if the observed distributions of L do not overlap (extrapolation)

Effect of extrapolation (1)[2]

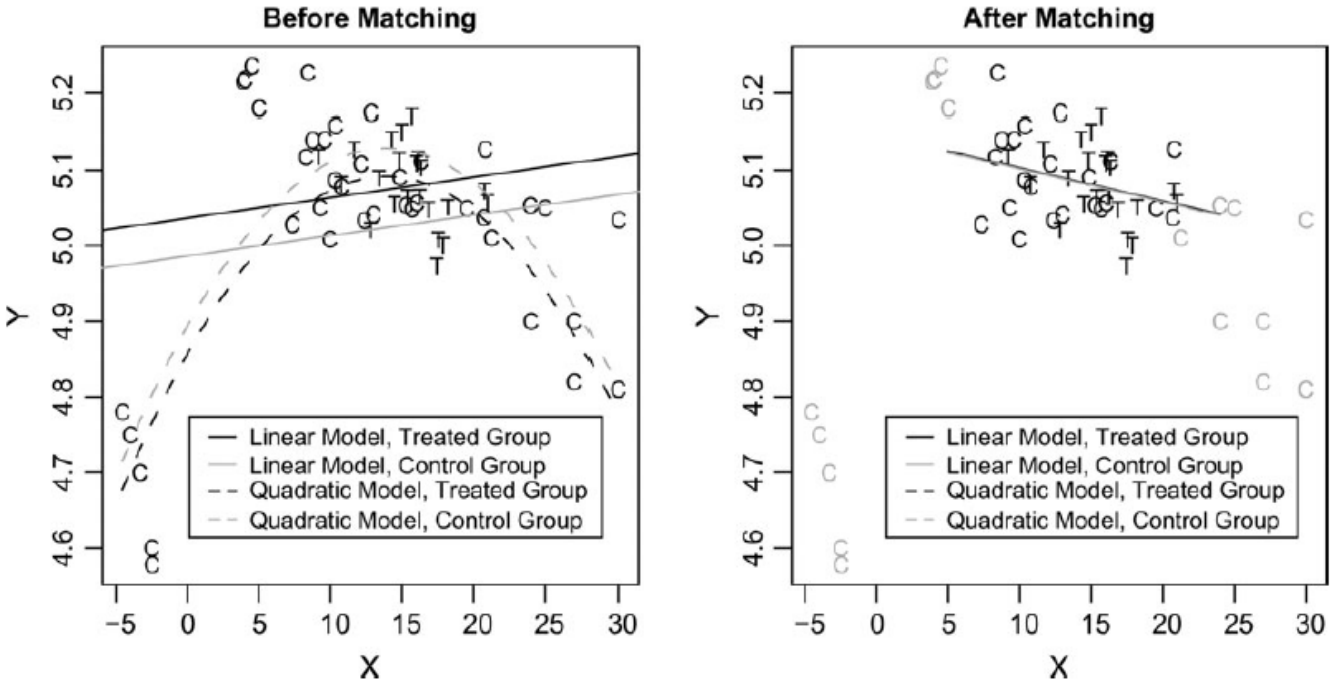
- Take whole data (left)
- Fit linear and quadratic models → different results



[2] Ho et al. *Political Analysis* 2007

Effect of extrapolation (2)

- Match treated and control patients with similar L (right)
- Gray units are discarded
- Similar treatement effect estimates by both models



Propensity score

- Probability of receiving the treatment A given the covariates L

$$\pi(L) = \Pr(A|L)$$

- Key properties of the PS
  - Balancing score
  - Under unconfoundedness, the difference between groups at a given value of $\pi(L)$ is an unbiased estimate of treatment effect at that value
  - Using sample estimates of $\pi(L)$ can produce sample balance on L
  - Heuristically, two individuals with the same PS only differ by the treatment they received

Uncounfoundedness

- Potential outcomes $\{Y(0), Y(1)\}$ do not depend on the treatment actually received given the covariates L
- $0 < \Pr(A = 1|L) < 1$ (positivity)
- Also termed 'ignorability'

Assumptions

- Impossible to know that no confounder was missed
- Rely on knowledge, draw DAGs
- Positivity can be looked at
- Balance has to be checked to verify that the PS model was successful

Key steps in building a PS model

1. Define the intervention and target population
2. Identify appropriate data
3. Select appropriate covariates (confounders)
4. Estimate the propensity score
5. Apply the PS ("use" it)
6. Assess balance (PS successfulness)
7. Analyze the outcome

Data sources

- Ad-hoc studies (perferably prospective)
  - ⊕ Collect appropriate data (confounders, outcomes)
  - ⊖ Need to collect data (time consuming, expensive . . .)
- Large (huge) administrative databases
  - ⊕ Data readily available for large no. of patients
  - ⊖ Representativeness, potentially missing important confounders
- Grouping of administrative databases
  - ⊕ Even more data, better representation
  - ⊖ Clustered missing confounders

Selection of covariates for the PS model

- "True" confounders (related to A and Y) should be included
- Better include more than less variables, if possible (sample size)
- For smaller sample size, concentrate on variables strongly related to the outcome rather than treatment
- Think that too many variables may lead to narrower common support (and information loss)
- Avoid colliders and IVs (DAGs, again)

PS estimation

- Any regression model for binary variable
- Logistic regression most commonly used
- Other options
  - CART
  - More recent: boosted CART, random forests
- The PS model itself is of little interest: the predictions $\pi(L_i)$ are just needed
  - The predictive ability of the model is not central
  - Neither are overfitting or collinearity
  - But should result in successfully balanced samples

"Conditioning" on the propensity score

- "Conditioning" can be intended different ways
- Matching
- Weighting
- Some other approaches have also been considered

Assessing balance

- Properties of the PS rely on balancing: the succesfulness of the PS model to achieve balance has to be assessed
- Not a matter of standard diagnostics for the PS model itself

- Somewhat beyond the scope of statistical testing between groups
- Show summary statistics for groups before/after matching/weighting
- As well as standardized differences (mean diff./pooled SD)

Matching

- Match m controls to n treated
- (m, n) are generally fixed (often with $m = 1$ and $n = 1$)
- Full matching: all controls and treated with "close" PS are matched together
- Controls (but also treated patients) on the "edge" of the PS distribution likely not to be matched
- Waste of data for some, asset for others
- Often estimates ATT, but sometimes arguable (when some treated cannot be matched)

Matching in practice

- Try to match each treated patient with the control with the closest PS, $\pi(L)$
- With or without replacement
- Within a range of PS values (*caliper*) or not
- Several algorithms for matching (e.g. optimal matching, ...)

Analysis

- Same type of analysis as would have been performed on the whole sample
- Preferably accounting for within-pairs correlation for variance estimation
- Weighted analyses if matching with replacement or if full matching

Example: Bilateral *vs* single-LT for IPF

- Patients with idiopathic pulmonary fibrosis
- Intervention = BLT *vs* SLT
- Outcome = survival
- UNOS registry, 3327 patients
- 1:1 matching without replacement within a 0.25 SD caliper

Baseline data

Table 1. **Main Baseline Patient Characteristics, by Type of Lung Transplantation**

Characteristic	Nonmissing Data, *n* (%)	Single-Lung Transplantation (*n* = 2146)	Bilateral Lung Transplantation (*n* = 1181)	Standardized Difference, %*
Recipient				
Mean age (SD), *y*	3327 (100)	57.1 (9.0)	54.0 (10.0)	32.1
Age distribution, *n (%)*	3327 (100)			
≤50 y		424 (19.8)	362 (30.7)	25.3
51–55 y		324 (15.1)	188 (15.9)	2.3
56–60 y		552 (25.7)	279 (23.6)	4.9
>60 y		846 (39.4)	352 (29.8)	20.3
Women, *n (%)*	3327 (100)	705 (32.9)	358 (30.3)	5.5
Functional status, *n (%)*†	2852 (85.7)			
Class I		464 (26.0)	213 (19.9)	14.6
Class II		928 (52.1)	483 (45.1)	13.9
Class III		390 (21.9)	374 (35.0)	29.3
Diabetes, *n (%)*	3044 (91.5)	279 (14.7)	186 (16.2)	4.1
Oxygen required at rest, *n (%)*	2498 (75.1)	1318 (76.1)	642 (83.8)	19.4
Mean FVC (SD), *% predicted*	3082 (92.6)	49.0 (16.0)	47.4 (17.9)	9.3
Mean pulmonary capillary wedge pressure (SD), *mm Hg*	2842 (85.4)	8.8 (5.9)	10.1 (6.1)	22.8
Mean pulmonary artery pressure (SD), *mm Hg*	2474 (74.4)	23.4 (8.8)	28.4 (11.5)	49.2
Mean body mass index (SD), *kg/m²*	3193 (96.0)	27.2 (4.5)	26.8 (4.3)	11.0
Donor				
Mean age (SD), *y*	3327 (100)	32.2 (13.6)	33.0 (14.9)	5.3
Female, *n (%)*	3327 (100)	775 (36.1)	532 (45.0)	18.3
Mean body mass index (SD), *kg/m²*	3146 (94.6)	24.8 (5.1)	25.0 (5.1)	3.2
Diabetes, *n (%)*	3060 (91.9)	76 (4.0)	46 (4.0)	0
Cause of death, *n (%)*	3149 (94.6)			
Anoxia		136 (6.8)	98 (8.5)	6.3
Stroke		740 (37.1)	445 (38.6)	3.0
Head trauma		1105 (55.4)	599 (51.9)	7.0
CNS tumor		14 (0.7)	12 (1.0)	3.6
Donor-to-recipient				
Cytomegalovirus status mismatches, *n (%)*	2361 (71.0)	610 (44.3)	434 (44.2)	0.2
Sex mismatches, *n (%)*	3327 (100)	616 (28.7)	418 (35.4)	14.4
Blood group mismatches, *n (%)*	3327 (100)	221 (10.3)	101 (8.6)	6.0
HLA mismatches, *n (%)*	2735 (82.2)	4.6 (1.1)	4.7 (1.1)	7.6

CNS = central nervous system.
* Mean difference divided by the pooled SD, expressed as a percentage.
† Ranges from class I to III, indicating that the patient performs activities of daily living with no, some, or total assistance, respectively.

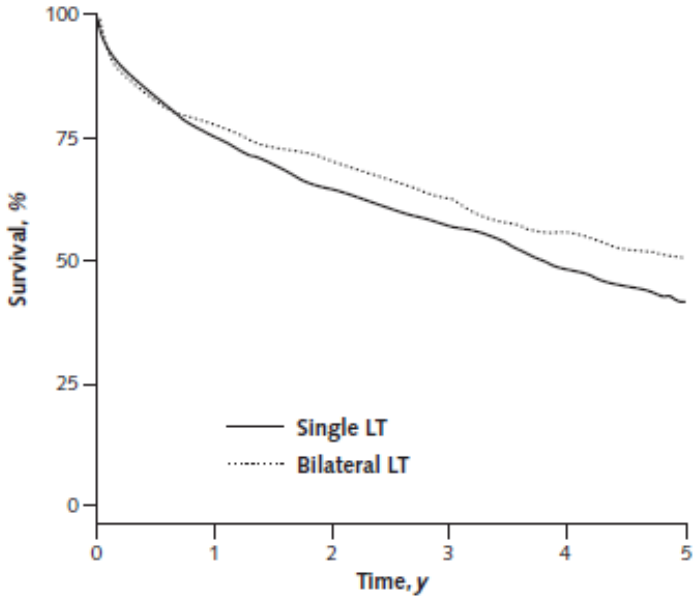
Matched data

Table 2. **Main Baseline Characteristics of Patients Matched by Propensity Score, by Type of Lung Transplantation**

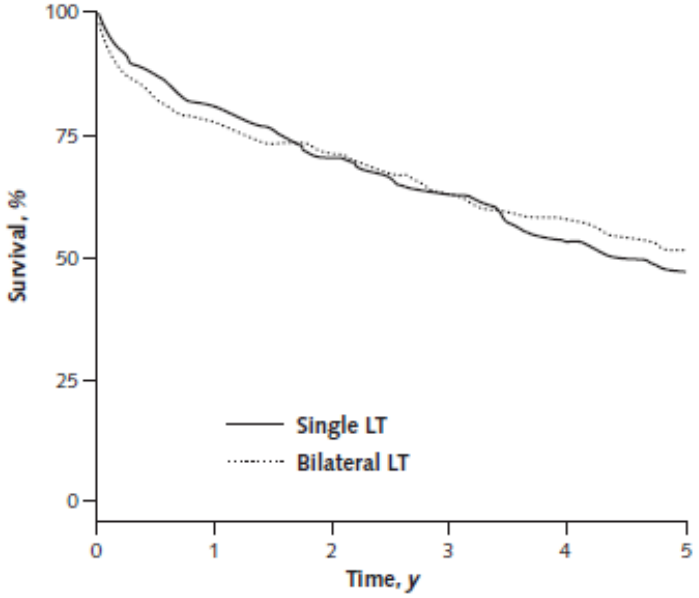
Characteristic	Single-Lung Transplantation (n = 795)	Bilateral Lung Transplantation (n = 795)	Standardized Difference, %*
Recipient			
Mean age (SD), *y*	56.0 (8.4)	55.9 (8.4)	0.7
Age distribution, *n (%)*			
≤50 y	172 (21.6)	180 (22.6)	2.4
51–55 y	134 (16.9)	126 (15.8)	2.7
56–60 y	218 (27.4)	214 (26.9)	1.1
>60 y	271 (34.1)	275 (34.6)	1.1
Women, *n (%)*	244 (30.7)	229 (28.8)	4.2
Functional status, *n (%)*†			
Class I	179 (22.5)	173 (21.8)	1.8
Class II	369 (46.4)	376 (47.3)	1.8
Class III	247 (31.1)	246 (30.9)	0.3
Diabetes, *n (%)*	143 (18.0)	125 (15.7)	6.0
Oxygen required at rest, *n (%)*	674 (84.8)	672 (84.5)	0.7
Mean FVC (SD), *% predicted*	48.9 (16.6)	48.5 (17.4)	2.4
Mean PCWP (SD), *mm Hg*	9.7 (6.0)	9.5 (5.6)	4.5
Mean pulmonary artery pressure (SD), *mm Hg*	24.8 (8.6)	24.7 (8.7)	0.3
Body mass index (SD), *kg/m²*	27.2 (4.4)	26.9 (4.2)	5.8
Donor			
Mean age (SD), *y*	32.9 (13.8)	33.3 (15.0)	2.5
Female, *n (%)*	334 (42.0)	329 (41.4)	1.3
Mean body mass index (SD), *kg/m²*	25.0 (5.2)	25.0 (5.1)	0.2
Diabetes, *n (%)*	31 (3.9)	36 (4.5)	2.0
Cause of death, *n (%)*			
Anoxia	64 (8.1)	68 (8.6)	1.8
Stroke	297 (37.4)	305 (38.4)	2.1
Head trauma	425 (53.5)	412 (51.8)	3.3
CNS tumor	9 (1.1)	10 (1.3)	1.2

Outcome analysis

- Cox PH model with time-dependent effect and robust variance (matched structure)

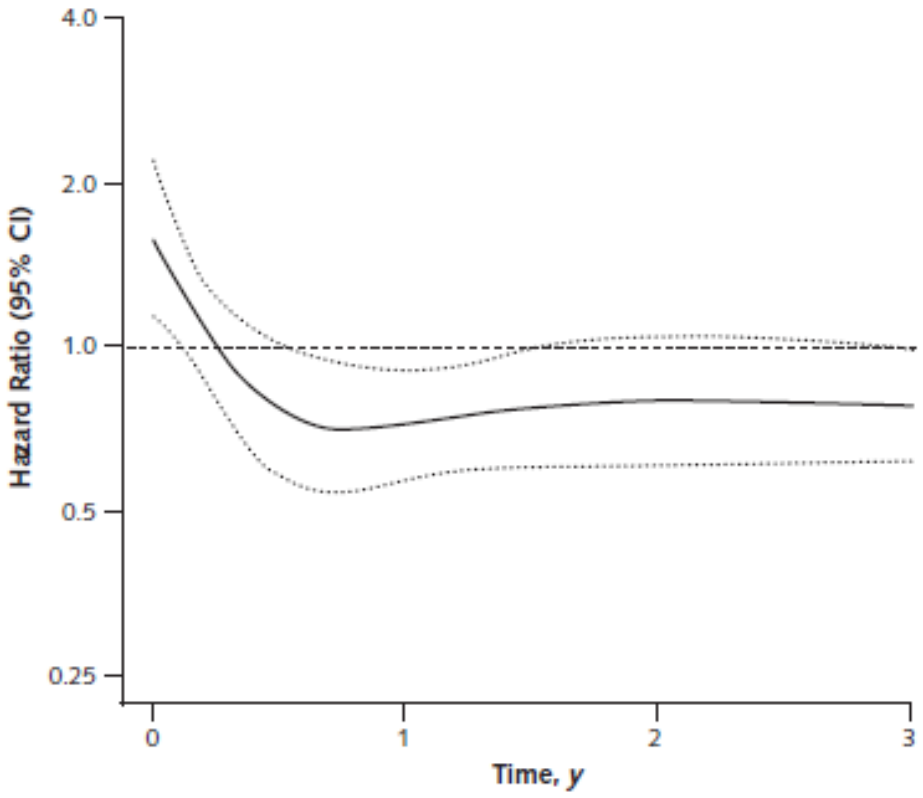


At risk, *n*						
Single LT	2146	1378	1003	765	551	404
Bilateral LT	1181	648	435	284	199	129
Total	3327	2026	1438	1049	750	533



At risk, *n*						
Single LT	795	499	333	235	153	112
Bilateral LT	795	456	325	225	164	105
Total	1590	955	658	460	317	217

Time-dependent treatment effect



IPTW

- Inverse probability of treatment weighting
- Weights inversely proportional to probability of receiving the treatment actually recieved
- Tries to reconstruct a population with similar structure in both groups

Inverse probability of treatment weighting (IPTW)

- Linked to Horvitz-Thompson weighting in survey sampling
- Treated patients are weighted by $1/\hat{e}(L_i)$
- Control patients are weighted by $1/[1 - \hat{e}(L_i)]$
- Overweights patients who had low probability of receiving the treatment they actually received
  - Compensates the larger no. of patients of the other group with similar $\pi(L)$
- Estimates ATE (weighting up to full population)
- But weigths for ATT can also be used
  - Weight for treated is 1
  - Weight for controls is $\hat{e}(L_i)/[1 - \hat{e}(L_i)]$ (the odds)

Some choices in practice

- Extreme weights may yield unstable results
- Some solutions are
  - Stabilized weights: multiply the weights by the marginal probability of the treatment actually received
  - Truncation (or trimming): fix a maximum value for weights
- Truncation produces bias but variance will be lower
- Still important to check balance (weighted analysis)

Outcome analysis

- Use weighted analysis (weighted t-test, weighted regression, . . .)
- Use 'robust' variance estimator
- Or use bootstrap

Conclusion on propensity scores

- Different methods, different effect measures
- Makes sense to use several methods as sensitivity analyses
- Estimate marginal effects
- Rely on unconfoundedness: cannot balance on unobserved counfounders $\rightarrow$ remaining bias

References

- Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**:41–55.
- D'Agostino RB. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine* 1998; **17**:2265–2281. URL `http://www.ncbi.nlm.nih.gov/pubmed/9802183`. PMID: 9802183.
- Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 2007; **15**:199–236. URL `http://pan.oxfordjournals.org/content/15/3/199.abstract`.
- Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine* 2004; **23**:2937–2960.
- Thabut G, Christie JD, Ravaud P, Castier Y, Dauriat G, Jebrak G, Fournier M, Leseche G, Porcher R, Mal H. Survival after bilateral versus single-lung transplantation for idiopathic pulmonary fibrosis. *Ann Intern Med* 2009; **151**:767–774.

Methods in Research on Research

A presentation delivered at the

first MiRoR training event
October 19-21, 2016
Ghent, Belgium

This project has received funding from the EU Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant Agreement #676207