# Causal Inference and Propensity Scoring

Christine zu Eulenburg
Medical Statistics and Decision Making
UMCG

12.12.2017

# Help! Statistics! Lunchtime Lectures

What? frequently used statistical methods and questions in a manageable timeframe for all researchers at the UMCG
No knowledge of advanced statistics is required.

When? Lectures take place every 2nd Tuesday of the month, 12.00-13.00 hrs.

Who? Unit for Medical Statistics and Decision Making

| When? | Where? | What? | Who? |
|---|---|---|---|
| **Winter break January 2018** | | | |
| **Feb 13, 2018** | Room 16 | Regression towards the mean: Interpretational pitfalls | H. Burgerhof |
| **Mar 13 2018** | Room 16 | Sample size calculation | D. Postmus |
| **April 10, 2018** | Room 16 | Conditional Survival | Y. Chen |
| **May 8, 2018** | Room 16 | Missing data | S. la Bastide |
| **June 12, 2018** | Room 16 | Save the date! | ? |

# Content

- Introduction
- Definitions: causal effects, association
- Randomized experiments
- Observational studies
- Confounding
- Matching Methods
- Propensity Scoring
- Selecting confounders
- Assumptions, Limitations

# Introduction

Observational studies on life courses are increasingly common for estimating effects in non-experimental settings. Compared to randomized studies, the setting is more complicated, since the the truth in terms of causality is hard to define.

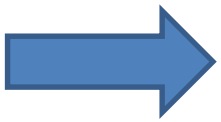
**Today, we will shed light on the problem of proving causality**, but also demonstrate some propensity scoring techniques to face this. These will be **PS matching, subclassification, and weighting**. Examples will come from non-experimental but also randomized studies.

**Stata codes** for implementing propensity score analyses will be described.

# What do we mean by a causal effect?

## Example

Zeus is waiting for a heart transplant. On January 1, he receives a new heart. 5 days later, he died.
Imagine we knew that (maybe in a parallel reality) if Zeus had not received the transplantation, but all other things in his life being equal, he would have been happy alive 5 days later.



The transplant has a **causal effect** on Zeus' survival!

(MA Hernan, J Epidemiol Community Health 2004; 58 :265–271)

# What do we mean by a causal effect?

In general:

**We compare the outcome when treatment T is present with the outcome when T is absent, all other conditions being equal. When the two outcomes differ, we say that T has a causal effect on the outcome.**

- The treatment T is also called **exposure variable**
  (1: exposed, 0 unexposed)
- The outcome Y can be dichotomous, >2 categories, continuous, or time-to-event

(MA Hernan, J Epidemiol Community Health 2004; 58 :265–271)

# Potential outcomes:

The **potential outcomes** that could be observed for each unit:

- Outcome under exposure: the outcome that would be observed if a subjects was exposed, Y(T=1)=Y(1)
- Outcome under non-exposure: the outcome that would be observed if a subject was unexposed, Y(T=0)=Y(0)

Zeus' Example: Y(1)=1 and Y(0)=0, because he died when exposed and would have survived if unexposed.

The exposure has a causal effect on a subject if $Y(1) \neq Y(0)$.

**Potential outcomes** are also called **counterfactual outcomes**.

# The fundamental problem of causal inference:

Individual causal effects are defined as the **difference of counterfactual outcomes**, of which only one can be observed.

**Problem:** Either a subject gets the treatment or he does not get the treatment, but we can not observe both outcomes and therefore we cannot measure their difference.

# *Definition: **Population** causal effects*

Define the probability P(Y(1)=1) as the proportion of all subjects that would have developed the outcome Y when being exposed.

*P(Y(1)=1) is the risk of Y(1).*

The exposure has a **causal effect *in the population*** if

*P(Y(1)=1) ≠ P(Y(0)=1)*

The population causal effect (in the latter referred to as "causal effect") can't be computed either in most cases,

But consistently estimated!

# The average treatment effect

For each subject, the treatment effect is defined to be $Y_i(1)-Y_i(0)$.

- The ***average treatment effect*** **(ATE)** is defined to be $E[Y_i(1)-Y_i(0)]$. The ATE is the average effect, at population level, of moving an entire population from untreated to treated.

- The ***average treatment effect of the treated*** **(ATT)** is defined to be $E[Y_i(1)-Y_i(0)|T=1]$. The ATT is the average effect of those who definitely received the treatment.

**In RCTs, ATT and ATE are equal, because those receiving the treatment do not differ from the general population, due to randomization.**

**In observational studies, these measures differ. Researchers should decide which measure answers their reseach question better!**

# *Definition:* ***Association***

Define the probability P(Y=1|T=1) as the proportion of subjects that developed the outcome Y ***among those being exposed*.**

*P(Y=1|T=1) is the risk of Y given T=1.*

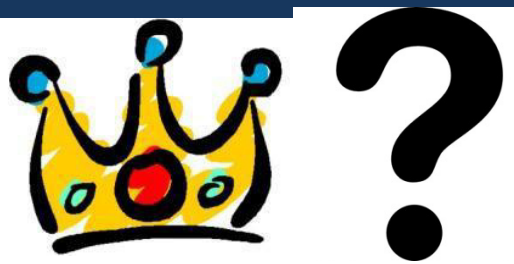
Exposure and Outcome are ***associated*** if

*P(Y=1|T=1) ≠ P(Y=1|T=0)*

**Note:** *Association* is here defined by **two disjoint risk sets**, exposed and unexposed subjects.
In contrast, *causation* was defined **on the same risk set with counterfactual exposures!**

(MA Hernan, J Epidemiol Community Health 2004; 58 :265–271)

# Randomization



- The gold standard for causal inference!
- In a randomized experiment, subjects are randomly assigned to (let's say) 2 groups, treatment and control.
- On average, the only difference between the two groups is then whether or not they receive the treatment.
- Thus, any difference in outcomes must be due to the treatment and **not to any other pre-existing differences between the groups**.

*The* **two disjoint risk sets** are only randomly unequal, that's why we can interprete **association = causation!**

# Complications of randomization

- Randomization not always feasible (e.g. genetics)
- Randomization not always ethical (e.g. smoking)
- People don't do what they're told (-> noncompliance)
- Long (expensive) studies: Randomize and wait 20 years?
- Selected sample, no general conclusions
- **Life-course epidemiology: (long-term) effects of congenital characteristics**

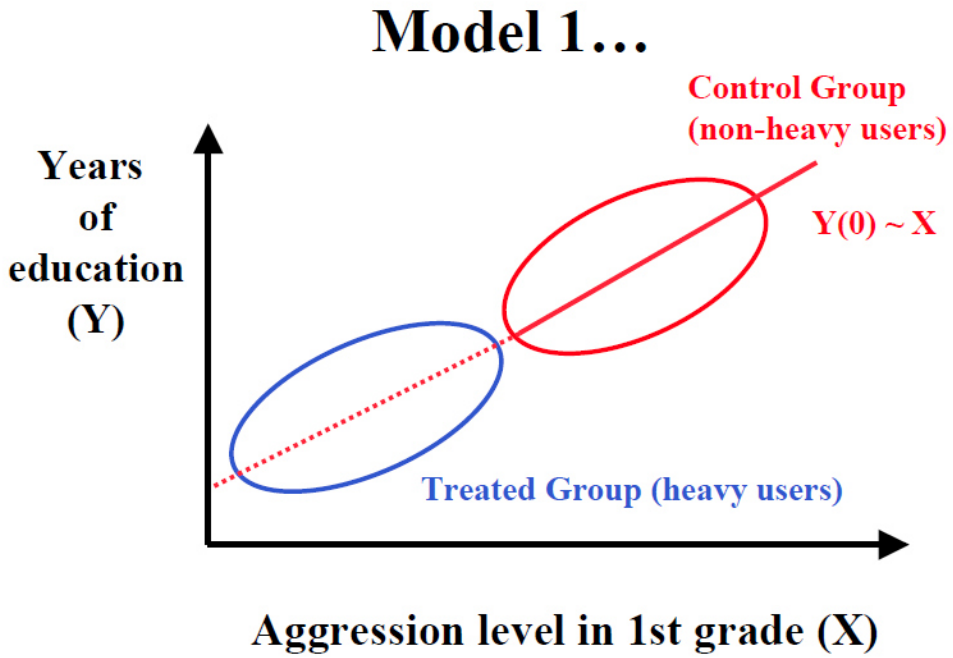
Instead:

- Non-experimental ("observational") studies
- **Problems:**
  - Exposed and unexposed subjects might be **systematically different**, in both observed and unobserved ways ("Confounding")
  - It's not always clear whether X affects Y or Y affects X

# Example 1

The association between adolescent marijuana use and adult outcome (Woodlawn study)

We would like to predict what the heavy users' outcome (years of education) would have been if they had not been heavy users.




(EA Stuart, Dev Psychol 2008; 44(2) :395-406)

# Solutions?!

- **Adjustment for confounders through regression models**
  - **Drawbacks**:
  - When treatment groups have very different distributions of the confounders, this can lead to bias if model is misspecified.
  - No appropriate model checks
  - Might be: sample size issue when many confounders are considered

- **Matching** (Comparing only subjects with the same aggression level to achieve “balance”)
  - But how to deal with all the other confounders (education years of the parents, number of siblings, neighborhood,…)?

- **Propensity Score Methods (Rosenbaum / Rubin 1983)**



(PR Rosenbaum, DB Rubin; Biometrika 1983; 70; 1: 41-55)

# The Propensity Score

- Definition: The Propensity Score (PS) is defined as the probability of receiving the investigated therapy (of being exposed).

- Two-step procedure:

  1. Estimate the PS for each single observation

  (e.g. through logistic regression)

  2. Apply the PS to estimate the exposure effect (different ways!)

# The Propensity Score

**The propensity score is a balancing score**: At each value of the PS, the distribution of the considered covariates is the same in the treated and in the control group.

- If two individuals had the same probability of being exposed (become a heavy drug user), and one actually is and one not, this allocation can be seen as random.
- With similar propensity scores, exposed and unexposed subjects look only randomly different on the observed covariates.
- **Difference in outcomes within groups of same/similar PS gives unbiased estimate of the exposure effect.**

(E Stuart, Educational Researcher 2007;36;4:187-198)

# Step 1: Estimating the PS

**… through logistic regression**

- Dependent variable: exposure (1/0)
- Covariables (independent variables): patient characteristics measured at baseline (confounder)

**How to select the covariables to calculate the PS?**

- Variables that are associated with the outcome (not only those that differ across exposure groups)
- The more the better, parsimony is not an issue

# Step 2: Applying the PS

**There are in general three different methods:**

1. **PS Matching**
   - e.g. k to 1 nearest neighbor matching: For each treated unit select k controls with closest propensity scores
   - Many variations possible

2. **PS stratification**
   - Group individuals into groups with similar PS values.
   - Rubin recommends to use 5 groups (quintiles).
   - Pool estimates across strata

3. **Inverse-probability-of-treatment-weighting (IPTW)**
   - Compute the *inverse* PS
   - Weight the observations accordingly
   - Exclude heavy weights

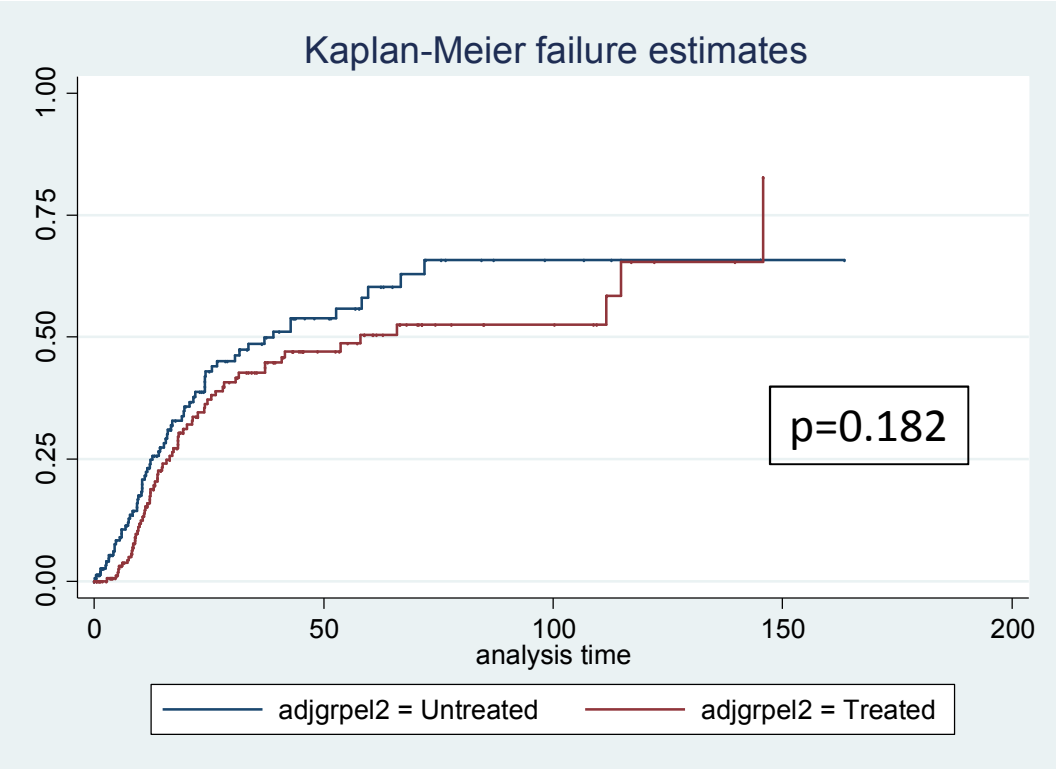
**The method that achieves the best balance should be selected!**

# Example: Radiation Therapy and survival in vulvar cancer

- It is unclear whether patients with advanced vulvar cancer benefit from adjuvant radiation therapy (RT). Because of the low incidence (2-4 diagnoses per 100.000 women per year), an RCT is not feasible.
- The **Care-1 study** is a retrospective cohort study on 1249 vulvar cancer patients treated at 29 gynecologic cancer center in Germany in 1998 – 2008.
- **346 patients with affected lymph nodes (N+) and measured FU**
- **Median FU 16 months**
- **Thereoff 164 with adjuvant RT (47%)**
- **Median DFS and OS: 15 and 43 months**

(C Eulenburg, Plos One 2016;11(11))

# Result from a Kaplan-Meier analysis Overall survival (OS) by treatment



# Patient characteristics (extract)

| | | without adjuvant treatment (n=164) | with adjuvant treatment (n=182) | How to check for balance? |
|---|---|---|---|---|
| Resection status | | | | |
| R0 | | 20 (12.2) | 35 (19.2) | |
| R1 | | 122 (74.4) | 134 (73.6) | |
| unknown | | 22 (13.4) | 13 (7.1) | |
| Positive LN | | | | |
| 1 | | 84 (51.2) | 59 (32. | |
| 2 | | 32 (19.5) | 47 (25. | |
| 3 | | 18 (11.0) | 27 (14. | |
| >3 | | 24 (14.6) | 40 (22. | |
| unknown | | 6 (3.7) | 9 (5. | |
| ECOG | | | | |
| 0 | | 32 (19.5) | 56 (30.8) | |
| 1 | | 24 (14.6) | 43 (23. | |
| 2 | | 21 (12.8) | 27 (14. | |
| 3 | | 13 (7.9) | 6 (3. | |
| 4 | | 1 (0.6) | 0 (0. | |
| unknown | | 73 (44.5) | 50 (27. | |
| | | | | |
| age years | | 67 (20-89) | 71 (30-87) | |
| Metastasis diameter mm | | 15 (0.3-50) | 23 (1-80) | |

p-values? depend on n!

**standardized differences**: Difference between means, divided by the pooled standard deviation

standardized differences <0.1 indicate sufficient balance (PC Austin, 2011)

(C Eulenburg, Plos One 2016;11(11))

# Patient characteristics (extract)

| | | without adjuvant treatment (n=164) | with adjuvant treatment (n=182) | Standardized differences observed data |
|---|---|---|---|---|
| Resection status | | | | |
| R0 | | 20 (12.2) | 35 (19.2) | **0.19** |
| R1 | | 122 (74.4) | 134 (73.6) | -0.02 |
| unknown | | 22 (13.4) | 13 (7.1) | **-0.21** |
| Positive LN | | | | |
| 1 | | 84 (51.2) | 59 (32.4) | **-0.39** |
| 2 | | 19.5) | 47 (25.8) | **0.15** |
| 3 | | 11.0) | 27 (14.8) | **0.12** |
| >3 | | 14.6) | 40 (22.0) | **0.19** |
| unknown | | (3.7) | 9 (5.0) | 0.06 |
| ECOG | | | | |
| 0 | | 19.5) | 56 (30.8) | **0.26** |
| 1 | | 14.6) | 43 (23.6) | **0.23** |
| 2 | | 21 (12.8) | 27 (14.8) | 0.06 |
| 3 | | 13 (7.9) | 6 (3.3) | **-0.20** |
| 4 | | 1 (0.6) | 0 (0.0) | **-0.11** |
| unknown | | 73 (44.5) | 50 (27.5) | **-0.36** |
| | | | | |
| age years | | 67 (20-89) | 71 (30-87) | **-0.24** |
| Metastasis diameter mm | | 15 (0.3-50) | 23 (1-80) | **0.44** |

High grade of imbalance: Treated patients have better ECOG, but are older and have more affected lymph nodes.

(C Eulenburg, Plos One 2016;11(11))

# Step 1: Estimating the PS

**… through logistic regression**

- Dependent variable: treat
- Confounders to consider: those variables significantly associated with the endpoint. In this case: overall survival

- Stata code: install ado 'pscore'

```
logistic treat c.alter i.catClassTumorPH i.catDiametTumor
i.catDepthInvTum i.catMaxDiaNode i.catR0resec
i.catGradeCancer i.catNoAffRiGroin i.catECOGPH
i.catTypSurgery i.catSNL
predict pscore if e(sample)
sum pscore, detail
```

# Step 1: Estimating the PS

**Result:**

```
. sum pscore, detail

                        Pr(adjgrpel2)
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .1290571       .1127941
 5%     .2186168       .1196162
10%     .2623469       .1260787       Obs                 346
25%     .4095996       .1290571       Sum of Wgt.         346

50%     .5453753                      Mean           .5260116
                       Largest        Std. Dev.      .1720358
75%     .6573181       .8096634
90%     .7507457        .812732       Variance       .0295963
95%     .7779717       .8133522       Skewness      -.3226202
99%     .8096634       .8344911       Kurtosis       2.244752
```

Should be around 50%

# Step 2: Applying the PS

**There are in general three different methods:**

1. **PS Matching** **nearest neighbor, 1:3, caliper (0.2*sd(log(pscore))**

   Caliper: the maximum tolerated difference between matched subjects.
   Outcome can be compared directly between matched samples.
   Stata package: psmatch2 (alternatives in R: MatchIt, Optmatch)

```
psmatch2 treat c.alter[...] i.catSNL, neighbor(3) logit caliper(0.153)
```

There are so many matching opportunities!
The one model that produces the **best balance** should be selected in the end!

**Results I:**
psmatch2 matches to each of the 182 treated patients up to three untreated patients. **Here: 123/164 patients (75%)** are used in total, some of them were selected 7 times!
Note: Matching works best when the control group is (3 times) bigger than the treatment group!
**Matching allows to estimate the ATT!**

# Step 2: Applying the PS

## There are in general three different methods:

### 2. PS Stratification (IPTW)

**Subjects are stratified into groups of similar PS**. Within each stratum, the distribution of covariates is then similar. The treatment effect can be estimated by group comparisons within each stratum and pooling the results across strata.

Following Rosenbaum and Rubin (1983), 5 strata according to quintiles remove 90% of the bias.

**Pooled results are estimated by weighting stratum-specific results.**
Weighting by the total number in strata allows estimating the **ATE**.
Weighting by the number of treated within strata allows estimating the **ATT**.

```
xtile psquintile = pscore, nquantiles(5)

stcox  treat i.catECOGPH i.catClassTumorPH  alter  i.catDepthInvTum
i.catGradeCancer i.catNoAffRiGroin, strata(psquintile)
```

(Rosenbaum P, Rubin D, Biometria 1983; 70: 41-55)

# Step 2: Applying the PS

## There are in general three different methods:

### 3. Inverse Probability of Treatment weighting (IPTW)

**IPTW** weights subjects with respect to their PS. Doing so, an artificial sample is created, in which the distribution of covariates is equal between treatment groups.

***A subject's weight is equal to the inverse probability of receiving the treatment he actually received.***

Let $T_i$ be an indicator denoting whether the $i$th subject was treated and $e_i$ the PS for $i$. The IPT-weight can be estimated as $w_i = \frac{T_i}{e_i} + \frac{(1-T_i)}{(1-e_i)}$

```
gen iptweight= treat/pscore + (1-treat)/(1-pscore)

stset Monate [pweight =iptweight], failure(catCauseDeath==7) scale(1)

stcox  treat i.catECOGPH i.catClassTumorPH  alter  i.catDepthInvTum
i.catGradeCancer i.catNoAffRiGroin, vce(robust)
```



(Austin PC, Multivariate Behav Res 2011; 46(3): 399-424)

# Step 3: Balance check

The main criteria for a good PS model is whether balance was achieved. This can be tested through computing standardized differences again:

```
Stratification
pbalchk treat alter catR0resec_1  catR0resec_2 catR0resec_3
catGradeCancer1 catGradeCancer2 catGradeCancer3 catNoAffRiGroin_11
catNoAffRiGroin_12 catNoAffRiGroin_13 catNoAffRiGroin_14
catNoAffRiGroin_15 ecog_1 ecog_2 ecog_3 ecog_4 ecog_5 ecog_6,
strata(psquintile)
```

```
Matching
pbalchk treat alter catR0resec_1  catR0resec_2 catR0resec_3
catGradeCancer1 catGradeCancer2 catGradeCancer3 catNoAffRiGroin_11
catNoAffRiGroin_12 catNoAffRiGroin_13 catNoAffRiGroin_14
catNoAffRiGroin_15 ecog_1 ecog_2 ecog_3 ecog_4 ecog_5 ecog_6, wt(_weight)
```

```
IPTW
pbalchk treat alter catR0resec_1  catR0resec_2 catR0resec_3
catGradeCancer1 catGradeCancer2 catGradeCancer3 catNoAffRiGroin_11
catNoAffRiGroin_12 catNoAffRiGroin_13 catNoAffRiGroin_14
catNoAffRiGroin_15 ecog_1 ecog_2 ecog_3 ecog_4 ecog_5 ecog_6,
wt(iptweight)
```

Alternatives in stata: pstest, psmatch2

# Standardized differences (extract)

| | Observed data | PS stratification | PS matching 1:3 | IPTW |
|---|---|---|---|---|
| **Resection status** | | | | |
| **R0** | **0.194** | 0.003 | -0.066 | 0.046 |
| **R1** | **-0.017** | 0.024 | 0.071 | -0.034 |
| **unknown** | **-0.207** | -0.038 | -0.024 | -0.007 |
| **Positive LN** | | | | |
| **1** | **-0.387** | -0.006 | -0.087 | 0.000 |
| **2** | **0.151** | -0.008 | 0.017 | 0.022 |
| **3** | **0.115** | 0.012 | 0.038 | 0.003 |
| **>3** | 0.063 | 0.001 | 0.027 | 0.021 |
| **unknown** | **0.190** | 0.005 | 0.043 | -0.036 |
| **ECOG** | | | | |
| **0** | **0.261** | -0.016 | **0.115** | 0.032 |
| **1** | **0.229** | -0.008 | -0.070 | -0.041 |
| **2** | 0.059 | 0.065 | -0.064 | 0.034 |
| **3** | **-0.202** | -0.024 | 0.056 | -0.017 |
| **4** | **-0.110** | -0.076 | -0.033 | -0.061 |
| **unknown** | **-0.360** | -0.006 | -0.023 | -0.005 |
| | | | | |
| **age years** | **-0.239** | 0.001 | 0.082 | 0.049 |
| **Metastasis diameter mm** | **0.437** | **0.322** | **0.414** | **0.113** |

All four methods yield sufficient balance here!



(C Eulenburg, Plos One 2016;11(11))

# Summarized results

Is there an association (a causal effect?) between radiation therapy and survival in vulvar cancer patients?

| Model | HR | 95%CI | p-value |
|---|---|---|---|
| **Endpoint: OS** | | | |
| **Cox regression model** | 0.63 | (0.43-0.91) | **0.015** |
| **PS stratification** | 0.64 | (0.41-1.09) | **0.103** |
| **Matching 1:1** | 0.81 | (0.50-1.31) | **0.384** |
| **Caliper Matching 1:3** | 0.69 | (0.44-1.08) | **0.102** |
| **IPTW** | 0.65 | (0.44-0.98) | **0.04** |

In this study, all approaches show similar results: Treated patients lived longer.

# Appropriateness of the approaches

1. **PS Matching**
   - ATT, not ATE
   - 25% of observations are omitted
   - Requires 2-3-fold larger control group than treatment group*
2. **PS stratification**
   - ATT and ATE, depends on weighting
   - How many strata?
   - Requires enough treated and untreated within every stratum
   - Often important to do additional regression adjustment due to differences in subclasses
3. **Inverse-probability-of-treatment-weighting (IPTW)**
   - ATT or ATE, depends on weighting
   - Extreme weights can yield unstable results (solution: trimming, stabilized weights)

**General assumption in PS methods: No unmeasured confounding**

In the CaRE-1 study we used as example here, the IPTW results were considered to provide the best fit. The example here differs a bit from the original analysis for didactic reasons.

(*RB D'Agostino, Stat Med 1998;17(19):2265-81)

# Summary

- **Most important goal in propensity scoring: achieving balance across treatment groups!**
- Computing the PS using logistic regression (Alternative: CART)
- Use covariables associated with outcome
- Sufficient overlap in covariable distributions needed!
- Try a variety of different covariable sets to compute the PS and check for balance
- Decide on whether to estimate ATT or ATE
- If ATT: Matching is the best approach, when C:T>3:1, otherwise stratification
- If ATE: IPTW is the most accurate approach, but stratification also works well with additional adjustment.
- Try a variety of different matching / weighting approaches and compare resulting balance
- Covariates can be included in PS model and outcome model!

# PS method versus Regression analysis

- PS method mimics an RCT through its' two-step procedure: For estimating the PS, the study outcome is not considered. -> study design is separated from the outcome
- Forces you to check balance
- PS can take more covariables into account than a regression model (no risk of overfitting)
- Whenever estimating causal effects from observational data, PS should always be used!

# Advanced topics

- Missing data

Complete-case analyses are generally inappropriate / biased. Better: multiple imputation followed by PS

- Time-varying treatment

If treatment changes over time, the methods mentioned here are inappropriate. The marginal structural model is an alternative here. It bases on IPTW (Cole, Hernan; 2003)

- Multilevel data

# Thanks for your attention ☺

| When? | Where? | What? | Who? |
|---|---|---|---|
| **Winter break January 2018** | | | |
| **Feb 13, 2018** | Room 16 | Regression towards the mean: Interpretational pitfalls | H. Burgerhof |
| **Mar 13 2018** | Room 16 | Sample size calculation | D. Postmus |
| **April 10, 2018** | Room 16 | Conditional Survival | Y. Chen |
| **May 8, 2018** | Room 16 | Missing data | S. la Bastide |
| **June 12, 2018** | Room 16 | Save the date! | ? |