Triplet Matching:

Propensity Score Matching with 3 Groups

Jaron Arbet

Outline

1. Review of propensity score matching
2. Triplet matching
3. Alternative approaches

Propensity Score Review

- **Goal**: causal inference on a treatment/exposure variable X, observational study

- For binary X, PS is "the probability of treatment assignment conditional on observed baseline characteristics" - **Austin 2011**
  - PS typically estimated by logistic regression, although machine learning can be used

- "Balancing score": conditional on PS, distribution of baseline covariates should be balanced between treatment groups, similar to an RCT

Austin, Peter C. "An introduction to propensity score methods for reducing the effects of confounding in observational studies." *Multivariate behavioral research* 46.3 (2011): 399-424.

4 main PS methods

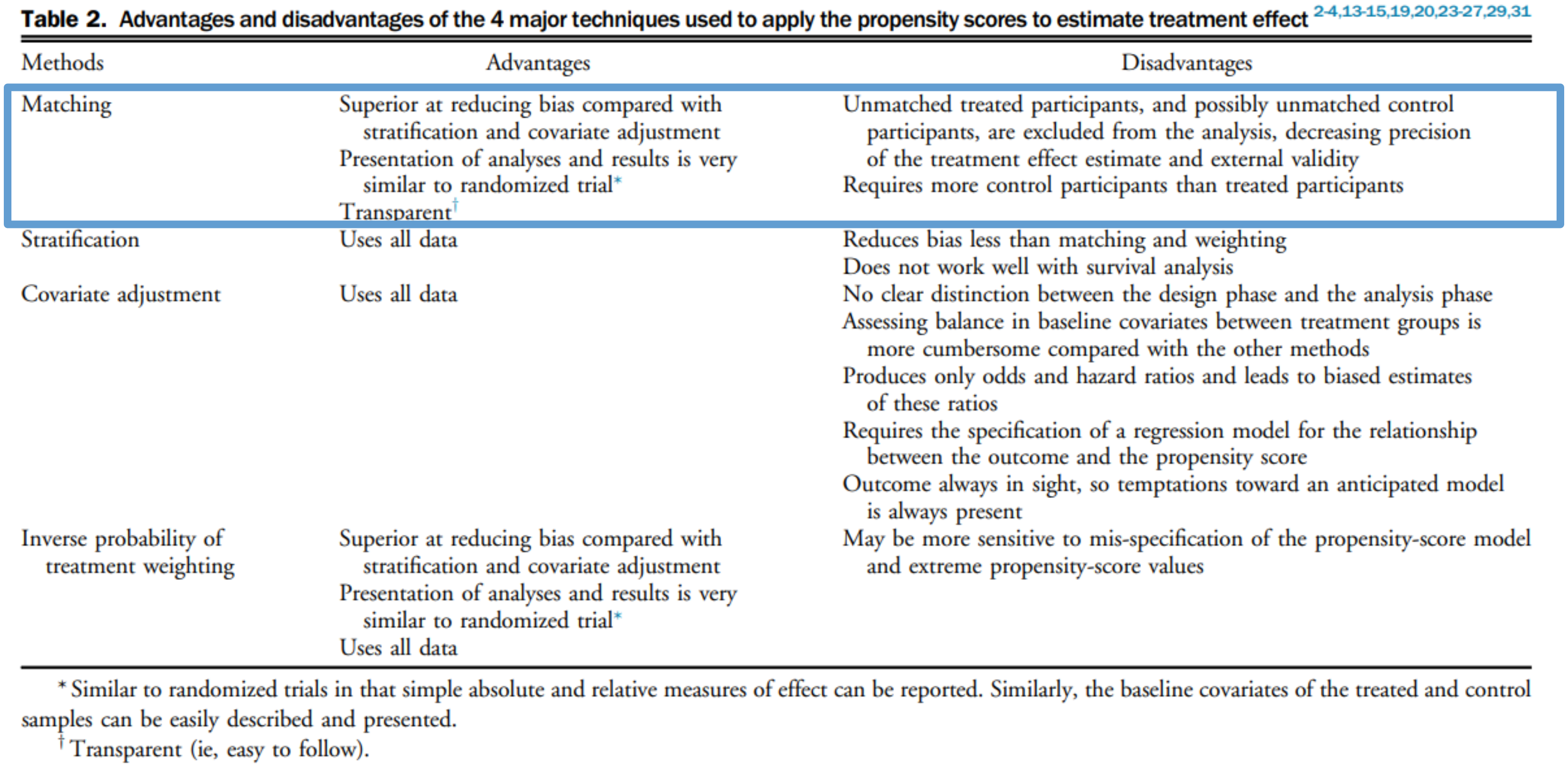


Table 2. Advantages and disadvantages of the 4 major techniques used to apply the propensity scores to estimate treatment effect[2-4,13-15,19,20,23-27,29,31]

Methods	Advantages	Disadvantages
Matching	Superior at reducing bias compared with stratification and covariate adjustment Presentation of analyses and results is very similar to randomized trial[*] Transparent[†]	Unmatched treated participants, and possibly unmatched control participants, are excluded from the analysis, decreasing precision of the treatment effect estimate and external validity Requires more control participants than treated participants
Stratification	Uses all data	Reduces bias less than matching and weighting Does not work well with survival analysis
Covariate adjustment	Uses all data	No clear distinction between the design phase and the analysis phase Assessing balance in baseline covariates between treatment groups is more cumbersome compared with the other methods Produces only odds and hazard ratios and leads to biased estimates of these ratios Requires the specification of a regression model for the relationship between the outcome and the propensity score Outcome always in sight, so temptations toward an anticipated model is always present
Inverse probability of treatment weighting	Superior at reducing bias compared with stratification and covariate adjustment Presentation of analyses and results is very similar to randomized trial[*] Uses all data	May be more sensitive to mis-specification of the propensity-score model and extreme propensity-score values

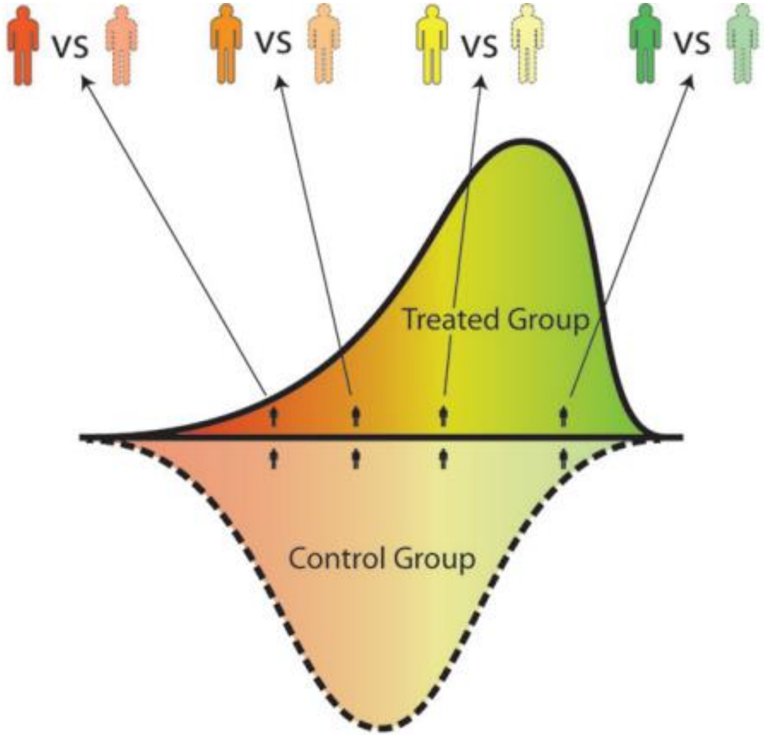
[*] Similar to randomized trials in that simple absolute and relative measures of effect can be reported. Similarly, the baseline covariates of the treated and control samples can be easily described and presented.

[†] Transparent (ie, easy to follow).

For review:

- Deb, Saswata, et al. "A review of propensity-score methods and their use in cardiovascular research." *Canadian Journal of Cardiology* 32.2 (2016): 259-265.
- Austin, Peter C. "An introduction to propensity score methods for reducing the effects of confounding in observational studies." *Multivariate behavioral research* 46.3 (2011): 399-424.

1:1 PS Matching



- Can think of PS as a univariate composite summary of all baseline covariates
- Matched patients must have similar PS value

McDonald, Robert J., et al. "Behind the numbers: propensity score analysis—a primer for the diagnostic radiologist." *Radiology* 269.3 (2013): 640-645.

Matching: many choices

Austin 2014 performed extensive simulations and made the following **recommendations**:

- Optimal vs **greedy nearest neighbor matching**?
- **Caliper** or no caliper?
- Match with or **without replacement**?
- Order in which treated subjects are selected (e.g. lowest to highest propensity score, highest to lowest propensity score, best match first, or **random order**)

Austin, Peter C. "A comparison of 12 algorithms for matching on the propensity score." Statistics in medicine 33.6 (2014): 1057-1069.

Triplet Matching

Generalized propensity score (GPS)

- Extend propensity score for multicategory, ordinal, or continuous treatments

- I will focus on the unordered 3 group case (A, B, C)

- **GPS$_i$** = $[\Pr(X_i = A|Z_i), \Pr(X_i = B|Z_i), \Pr(X_i = C)|Z_i]$ for baseline covariates Z
  - Can estimate using multinomial logistic regression or machine learning methods

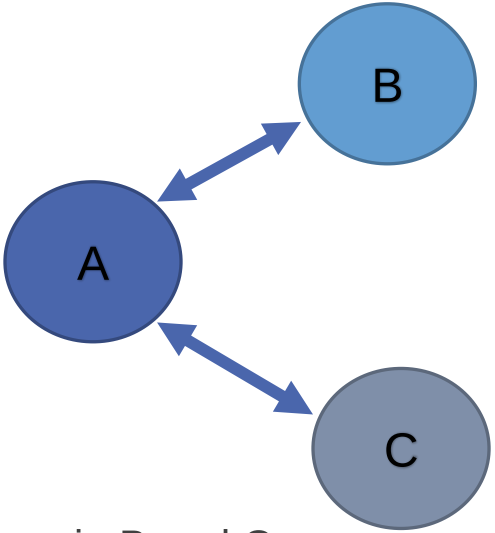
- Matched triplets: all 3 patients must have similar GPS *vectors.* Example:

Subject	True Group	Pr(X=A|Z)	Pr(X=B|Z)	Pr(X=C|Z)
1	A	0.30	0.60	0.10
2	B	0.29	0.59	0.12
3	C	0.31	0.61	0.08

Imai, Kosuke, and David A. Van Dyk. "Causal inference with general treatment regimes: Generalizing the propensity score." Journal of the American Statistical Association 99.467 (2004): 854-866.

Matching with 3 groups

- See **Lopez 2017** for review



Pairwise or "common referent matching" (CRM):

1. Define reference group A (usually smallest group)
2. Match A to B
3. Match A to C
4. Form triplets by only keeping patients in A that had matches in B and C.

- Typically perform pairwise 2-group matching using 2 separate logistic regression models to estimate $\Pr(X = A|Z)$, although one could match on multinomial estimate of $\Pr(X = A|Z)$

Lopez, Michael J., and Roee Gutman. "Estimation of causal effects with multiple treatments: a review and new ideas." *Statistical Science* (2017): 432-454.

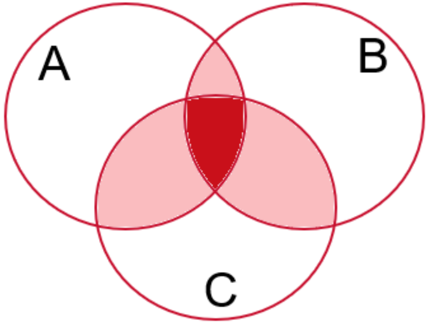
Pairwise matching

Pros:

- Straightforward to implement using existing 2-group matching software (e.g. SAS PSMATCH, or R packages MatchIt, matching)

Cons:

- **Lopez 2017** argues that “transitive property” not guaranteed: even if covariates are balanced for A vs B and A vs C, it’s possible that B vs C is unbalanced



Lopez, Michael J., and Roee Gutman. "Estimation of causal effects with multiple treatments: a review and new ideas." *Statistical Science* (2017): 432-454.

Example

- “Real World Outcomes of TAVR with the SAPIEN-3 Valve in Intermediate Risk Patients: Comparison of Data from the TVT Registry with PARTNER S3 Studies” 2018 https://www.tctmd.com/slide/real-world-outcomes-tavr-sapien-3-valve-intermediate-risk-patients-comparison-data-tvt

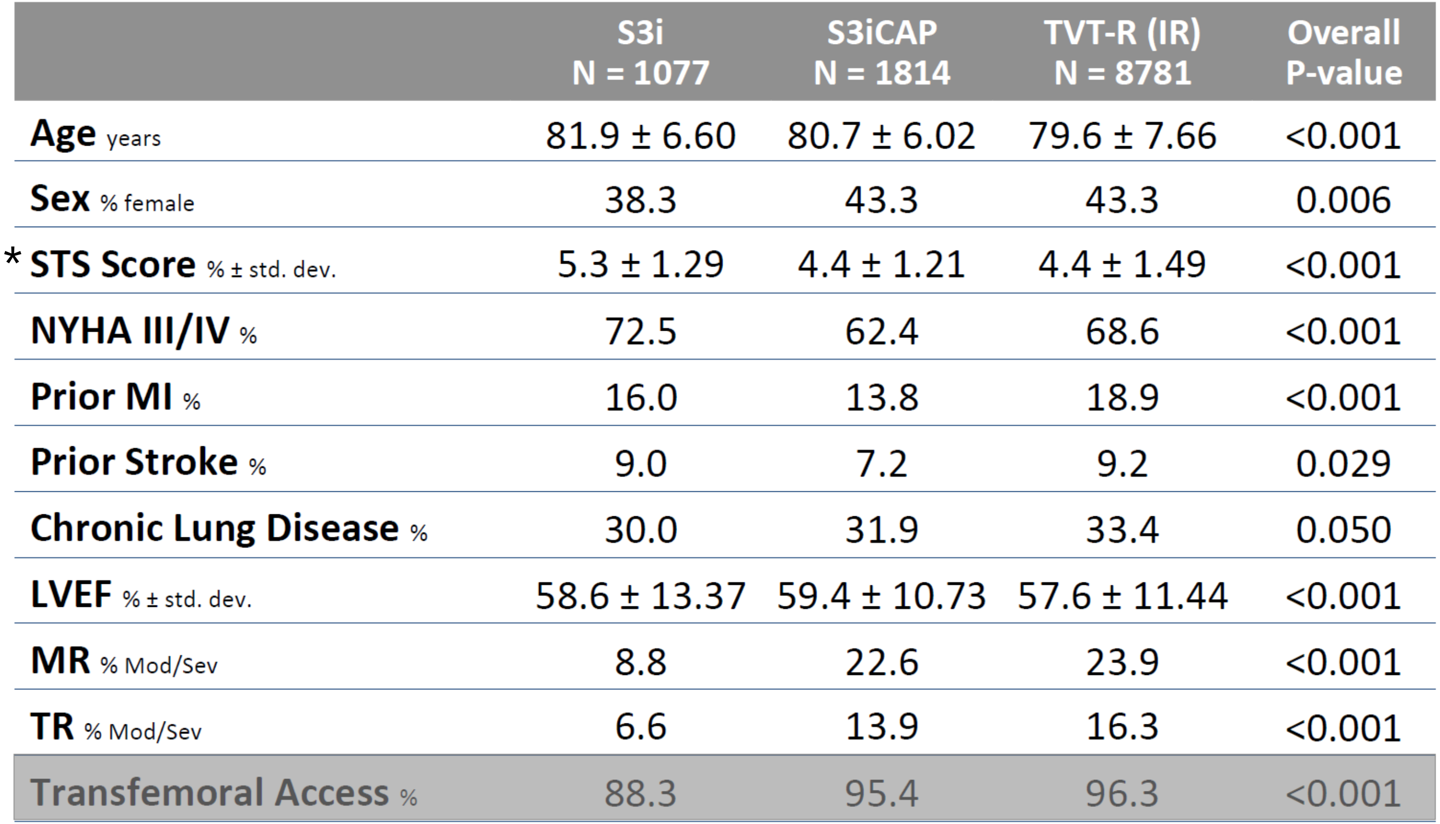
- Patients with severe aortic stenosis, intermediate surgical risk (IR), treated with SAPIEN 3 TAVR
  - The PARTNER II S3i trial demonstrated safety and efficacy

- It remains unknown whether TAVR with the S3 valve can be performed with similar safety and efficacy in real-world practice

- **Goal:** compare 30-day outcomes in IR patients from 3 data sources: **S3i** trial, S3i continued access registry (**S3iCAP**), ACC/STS Transcatheter Valve Therapeutics Registry (**TVT-R**).

  Use pairwise triplet matching to adjust for baseline covariates

Baseline characteristics

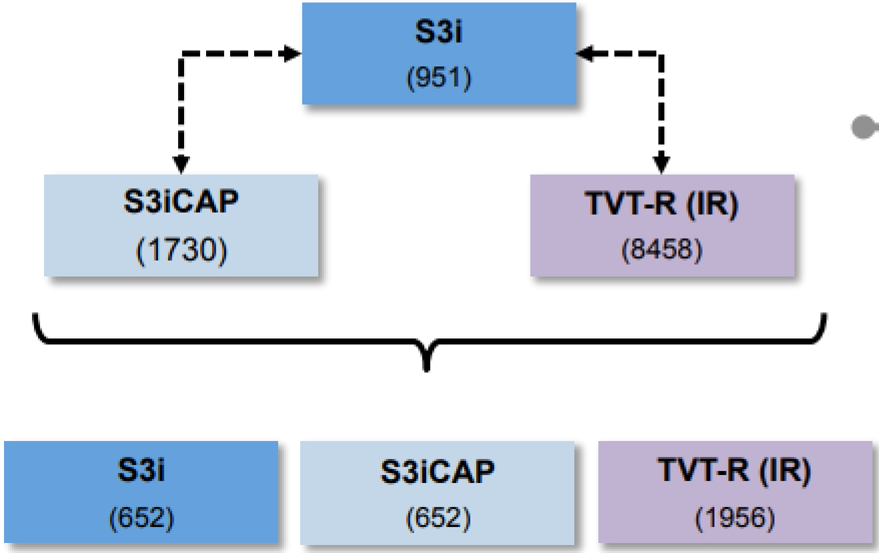


	S3i N = 1077	S3iCAP N = 1814	TVT-R (IR) N = 8781	Overall P-value
Age years	81.9 ± 6.60	80.7 ± 6.02	79.6 ± 7.66	<0.001
Sex % female	38.3	43.3	43.3	0.006
* **STS Score** % ± std. dev.	5.3 ± 1.29	4.4 ± 1.21	4.4 ± 1.49	<0.001
NYHA III/IV %	72.5	62.4	68.6	<0.001
Prior MI %	16.0	13.8	18.9	<0.001
Prior Stroke %	9.0	7.2	9.2	0.029
Chronic Lung Disease %	30.0	31.9	33.4	0.050
LVEF % ± std. dev.	58.6 ± 13.37	59.4 ± 10.73	57.6 ± 11.44	<0.001
MR % Mod/Sev	8.8	22.6	23.9	<0.001
TR % Mod/Sev	6.6	13.9	16.3	<0.001
Transfemoral Access %	88.3	95.4	96.3	<0.001

* did not match on STS Score, since one of the outcomes of interest was 30 day O:E ratio (observed 30 day mortality rate / STS score)

https://www.tctmd.com/slide/real-world-outcomes-tavr-sapien-3-valve-intermediate-risk-patients-comparison-data-tvt

Pairwise Triplet Matching



S3i
(951)

S3iCAP
(1730)

TVT-R (IR)
(8458)

S3i
(652)

S3iCAP
(652)

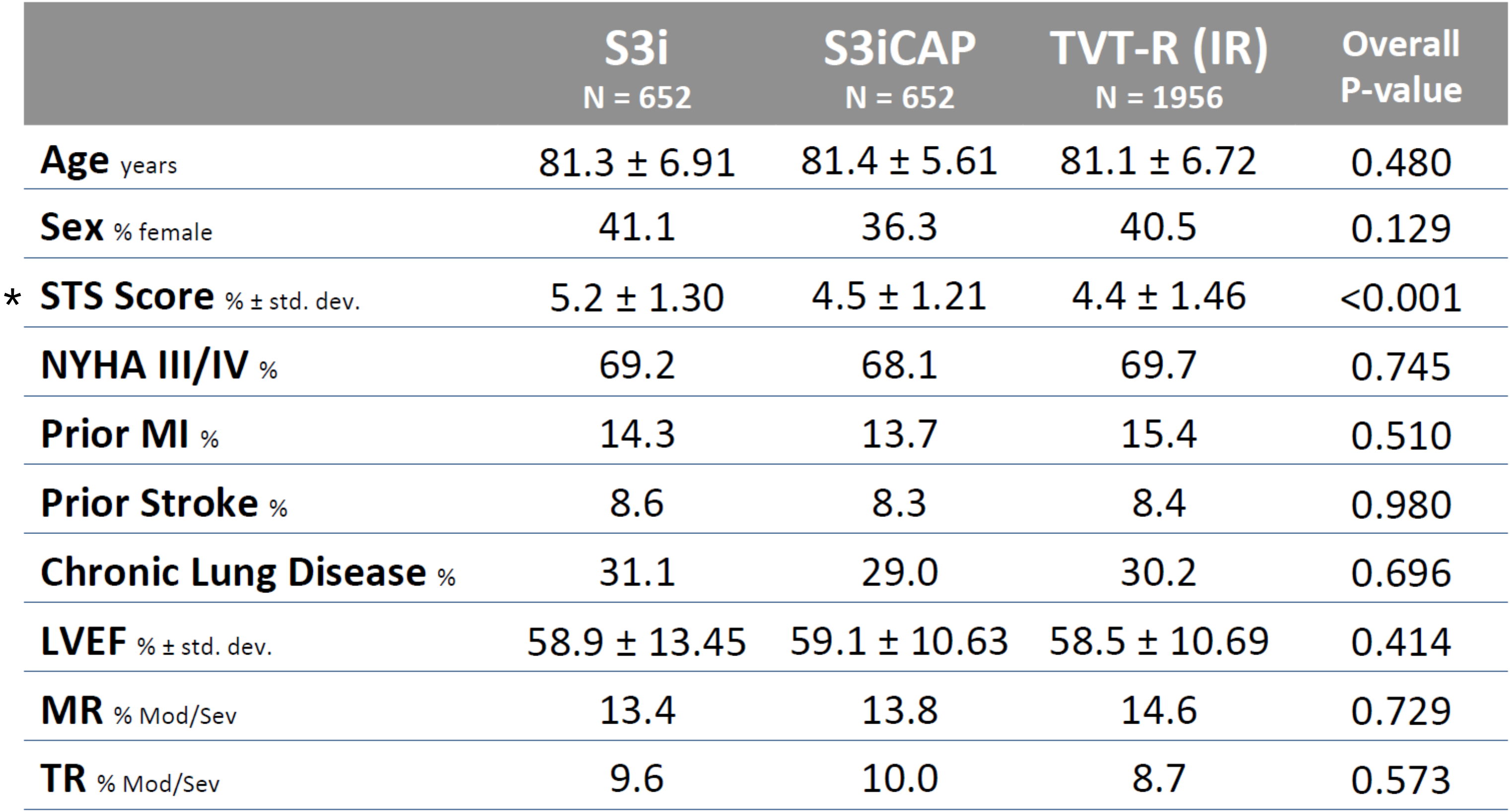
TVT-R (IR)
(1956)

Propensity Matched Analysis

- **1:1:3 subject selection**
- Transfemoral population
- **24 baseline covariates**
  - Missing values: imputed using Markov Chain Monte Carlo method
- **Logistic regression model**
- S3i used as control

https://www.tctmd.com/slide/real-world-outcomes-tavr-sapien-3-valve-intermediate-risk-patients-comparison-data-tvt

Baseline characteristics after matching

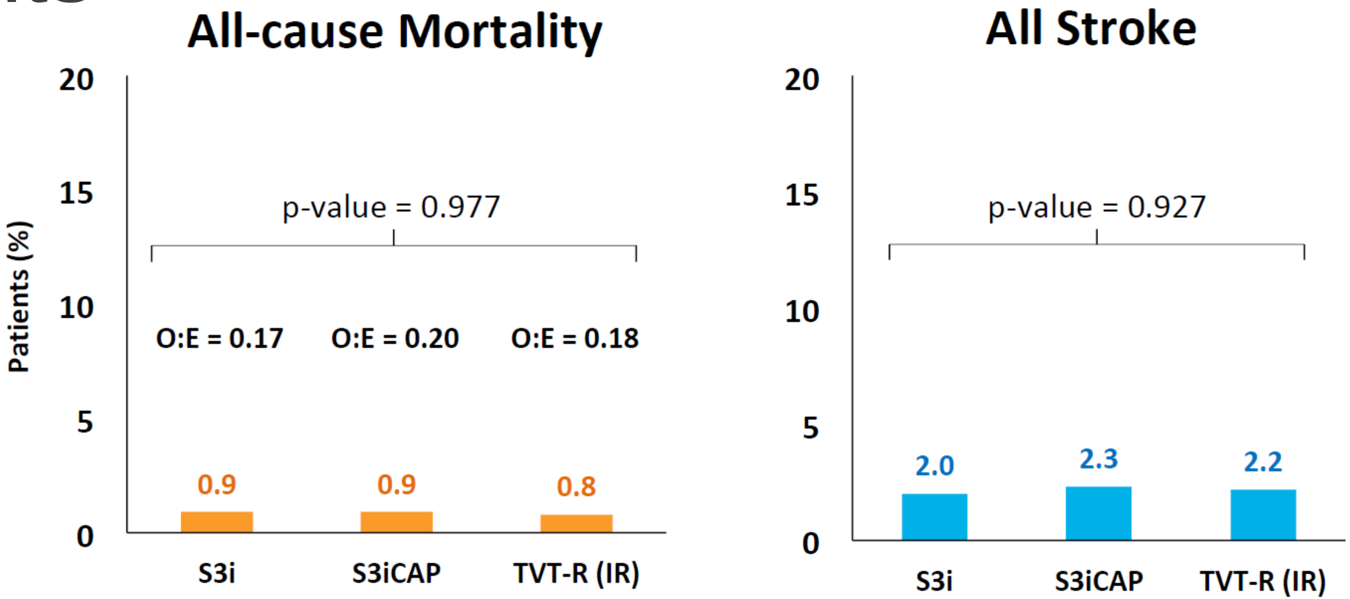


	S3i N = 652	S3iCAP N = 652	TVT-R (IR) N = 1956	Overall P-value
Age years	81.3 ± 6.91	81.4 ± 5.61	81.1 ± 6.72	0.480
Sex % female	41.1	36.3	40.5	0.129
* **STS Score** % ± std. dev.	5.2 ± 1.30	4.5 ± 1.21	4.4 ± 1.46	<0.001
NYHA III/IV %	69.2	68.1	69.7	0.745
Prior MI %	14.3	13.7	15.4	0.510
Prior Stroke %	8.6	8.3	8.4	0.980
Chronic Lung Disease %	31.1	29.0	30.2	0.696
LVEF % ± std. dev.	58.9 ± 13.45	59.1 ± 10.63	58.5 ± 10.69	0.414
MR % Mod/Sev	13.4	13.8	14.6	0.729
TR % Mod/Sev	9.6	10.0	8.7	0.573

* did not match on STS Score, since one of the outcomes of interest was 30 day O:E ratio (observed 30 day mortality rate / STS score)

https://www.tctmd.com/slide/real-world-outcomes-tavr-sapien-3-valve-intermediate-risk-patients-comparison-data-tvt

Results



STS	5.19	4.47	4.44			
# Patients	652	652	1956	652	652	1956
# Sites	51	60	453	51	60	453

Conclusion: after propensity matching, the 3 data sources (clinical and real world data) are comparable for the outcomes of interest

https://www.tctmd.com/slide/real-world-outcomes-tavr-sapien-3-valve-intermediate-risk-patients-comparison-data-tvt

Simultaneous matching

- Simultaneously form triplets that are close on all 3 elements of GPS vector

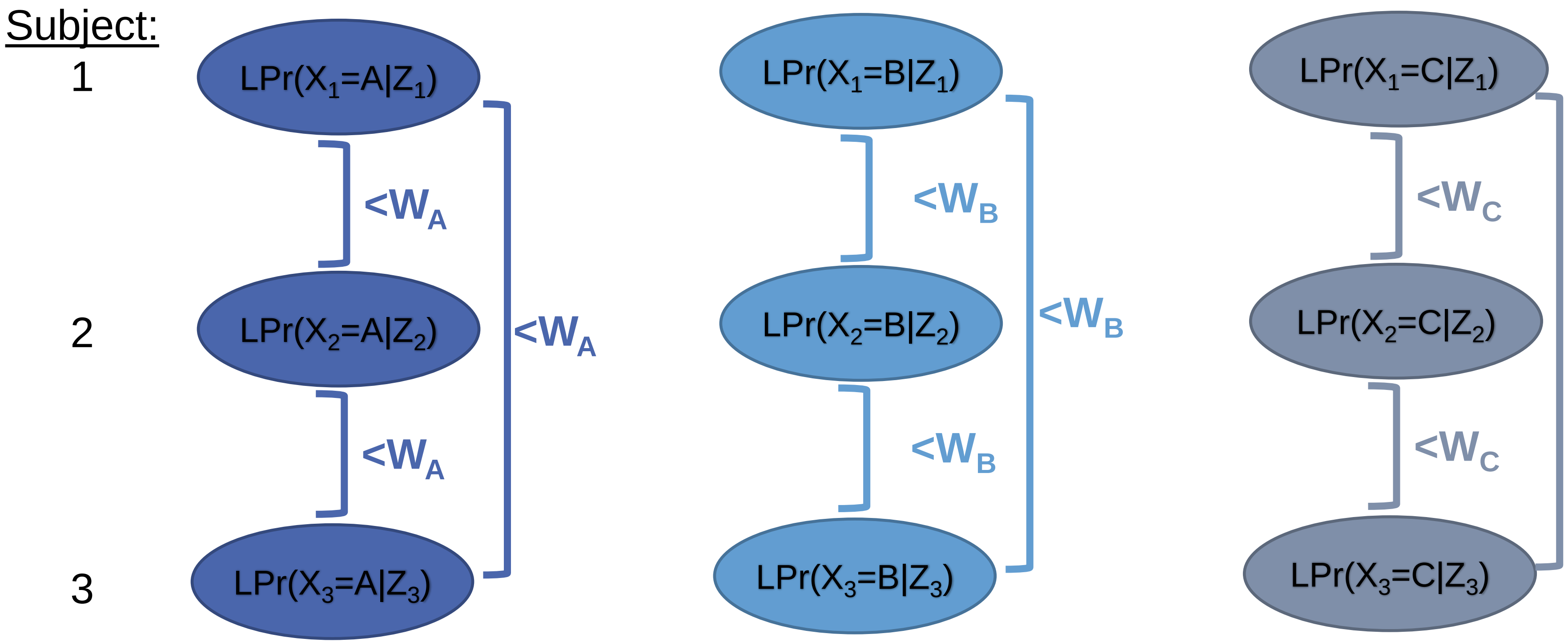
How to define “close”?

- One could constrain the total distance between each pair of vectors, or the sum
- Instead, I used 3 separate “caliper widths” to ensure all patients within a triplet are close on each element of the GPS vector
- **Pro**: unlike Pairwise, one can directly control distance between each element of GPS vector, which may lead to better covariate balance
- **Con**: more computationally challenging to implement, limited available software

3 calipers widths:

$W_A = \tau * SD\{logit[Pr(X=A|Z)]\}$, $W_B = \tau * SD\{logit[Pr(X=B|Z)]\}$, $W_C = \tau * logit[Pr(X=A|Z)]$

$\tau = 0.2$ as recommended in Austin 2011 and Wang 2013



Subject:

1

2

3

$LPr(X_1=A|Z_1)$

$LPr(X_2=A|Z_2)$

$LPr(X_3=A|Z_3)$

$<W_A$

$<W_A$

$<W_A$

$LPr(X_1=B|Z_1)$

$LPr(X_2=B|Z_2)$

$LPr(X_3=B|Z_3)$

$<W_B$

$<W_B$

$<W_B$

$LPr(X_1=C|Z_1)$

$LPr(X_2=C|Z_2)$

$LPr(X_3=C|Z_3)$

$<W_C$

$<W_C$

Austin, Peter C. "Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies." *Pharmaceutical statistics* 10.2 (2011): 150-161.
Wang, Yongji, et al. "Optimal caliper width for propensity score matching of three treatment groups: a Monte Carlo study." *PloS one* 8.12 (2013): e81045.

Checking covariate balance

2 groups

- Can use “standardized differences” to assess balance between groups (see Austin 2011 and references therein)

Continuous variable: $$d = \frac{(\overline{x}_{treatment} - \overline{x}_{control})}{\sqrt{\frac{s^2_{treatment} + s^2_{control}}{2}}},$$

Binary variable:

$$d = \frac{(\hat{p}_{treatment} - \hat{p}_{control})}{\sqrt{\frac{\hat{p}_{treatment}(1 - \hat{p}_{treatment}) + \hat{p}_{control}(1 - \hat{p}_{control})}{2}}},$$

Cutoffs of ≤0.1 (Austin 2011) and ≤ 0.25 (Harder 2010) have been used

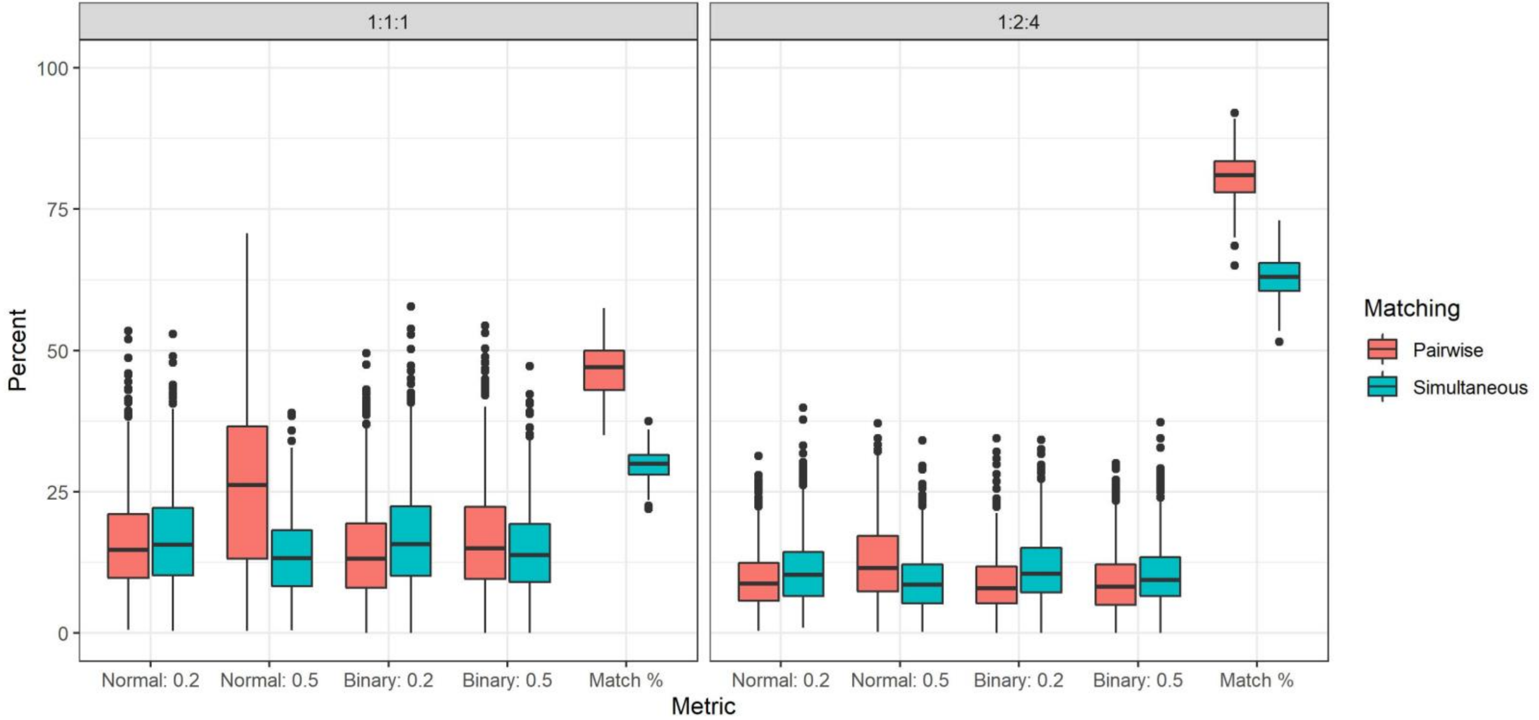
3+ groups

- Lopez 2017: calculate all pairwise standardized differences, then require the maximum standardized difference to be ≤ C

cobalt R package: easily check standardized differences before and after matching with >=2 groups (both tabular outputs and plots)

Austin, Peter C. "An introduction to propensity score methods for reducing the effects of confounding in observational studies." Multivariate behavioral research 46.3 (2011): 399-424.
Harder, Valerie S., Elizabeth A. Stuart, and James C. Anthony. "Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research." Psychological methods 15.3 (2010): 234.
Lopez, Michael J., and Roee Gutman. "Estimation of causal effects with multiple treatments: a review and new ideas." Statistical Science (2017): 432-454.

Simulations: Pairwise vs Simultaneous matching



- Pairwise results in better Match % (% matched in smallest group)
- Other boxplots show standardized diffs after matching (lower=better) for Normal & Binary covariates, simulated with standardized diffs of 0.2 or 0.5. Simultaneous is better for Normal 0.5

Alternative approaches

Weighting:

- weightit R package supports 3+ groups and can check covariate balance using cobalt R package
- Lu 2019: “Propensity score inverse weighting and regression approaches… are generally not optimal for use in pre-market medical device applications…”

Stratification:

- How to stratify on a multivariate vector (GPS)? Imai 2004 and Brown 2020
- Lopez 2017 proposes a clustering + stratified pair matching approach (no code)

Brown, Derek W., et al. "A novel approach for propensity score matching and stratification for multiple treatments: Application to an electronic health record–derived study." Statistics in medicine 39.17 (2020): 2308-2323.
Imai, Kosuke, and David A. Van Dyk. "Causal inference with general treatment regimes: Generalizing the propensity score." Journal of the American Statistical Association 99.467 (2004): 854-866.
Lu, Nelson, Yunling Xu, and Lilly Q. Yue. "Good statistical practice in utilizing real-world data in a comparative study for premarket evaluation of medical devices." *Journal of biopharmaceutical statistics* 29.4 (2019): 580-591.
Lopez, Michael J., and Roee Gutman. "Estimation of causal effects with multiple treatments: a review and new ideas." Statistical Science (2017): 432-454.

Conclusion

- One can easily implement pairwise matching in practice and use when covariate balance is achieved (e.g. max standardized pairwise diff < 0.1 or 0.25)

- Simultaneous matching may be better in theory, but need more research and software development
  - Nattino 2021 seems promising

Nattino, Giovanni, et al. "Triplet matching for estimating causal effects with three treatment arms: a comparative study of mortality by trauma center level." Journal of the American Statistical Association 116.533 (2021): 44-53.

References

- Austin, Peter C. "Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies." Pharmaceutical statistics 10.2 (2011): 150-161.
- Austin, Peter C. "An introduction to propensity score methods for reducing the effects of confounding in observational studies." Multivariate behavioral research 46.3 (2011): 399-424.
- Austin, Peter C. "A comparison of 12 algorithms for matching on the propensity score." *Statistics in medicine* 33.6 (2014): 1057-1069.
- Brown, Derek W. Austin, Peter C. "A comparison of 12 algorithms for matching on the propensity score." Statistics in medicine 33.6 (2014): 1057-1069., et al. "A novel approach for propensity score matching and stratification for multiple treatments: Application to an electronic health record–derived study." Statistics in medicine 39.17 (2020): 2308-2323.
- Deb, Saswata, et al. "A review of propensity-score methods and their use in cardiovascular research." Canadian Journal of Cardiology 32.2 (2016): 259-265.
- Harder, Valerie S., Elizabeth A. Stuart, and James C. Anthony. "Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research." Psychological methods 15.3 (2010): 234.
- Imai, Kosuke, and David A. Van Dyk. "Causal inference with general treatment regimes: Generalizing the propensity score." Journal of the American Statistical Association 99.467 (2004): 854-866.
- McDonald, Robert J., et al. "Behind the numbers: propensity score analysis—a primer for the diagnostic radiologist." Radiology 269.3 (2013): 640-645.
- Nattino, Giovanni, et al. "Triplet matching for estimating causal effects with three treatment arms: a comparative study of mortality by trauma center level." Journal of the American Statistical Association 116.533 (2021): 44-53.
- Lopez, Michael J., and Roee Gutman. "Estimation of causal effects with multiple treatments: a review and new ideas." Statistical Science (2017): 432-454.
- Lu, Nelson, Yunling Xu, and Lilly Q. Yue. "Good statistical practice in utilizing real-world data in a comparative study for premarket evaluation of medical devices." Journal of biopharmaceutical statistics 29.4 (2019): 580-591.
  Lopez, Michael J., and Roee Gutman. "Estimation of causal effects with multiple treatments: a review and new ideas." Statistical Science (2017): 432-454.
- Ternacle, Julien, et al. "Impact of Predilation During Transcatheter Aortic Valve Replacement: Insights From the PARTNER 3 Trial." Circulation: Cardiovascular Interventions 14.7 (2021): e010336.
- Wang, Yongji, et al. "Optimal caliper width for propensity score matching of three treatment groups: a Monte Carlo study." PloS one 8.12 (2013): e81045.