StatScience

Stats Power

Propensity Score Method

Lei Li, PhD

Disclaimer

The opinions and information in this presentation are those of the authors and do not represent the views and/or policies of the any regulatory agencies.

Outline

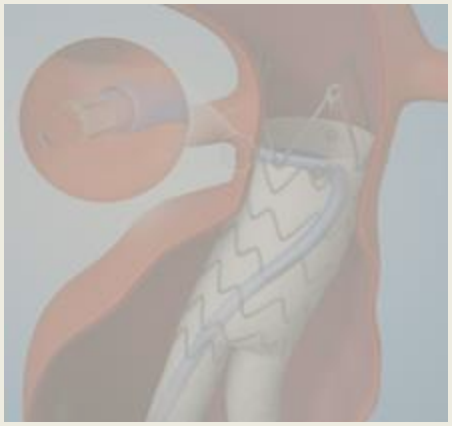
StatScience
Stats Power

- Background and Motivation Example
- Propensity Score Method Introduction, Considerations, and Example
- Practical Issues: Missingness issue and Collinearity Issue in small sample size study
- Summary

Background

StatScience
Stats Power

- 21st Century Cures Act
- Guidance issued to clarify how Real-World Evidence (RWE) may be used to support regulatory decisions.
- Practice in Center of Device and Radiation Health (FDA/CDRH):

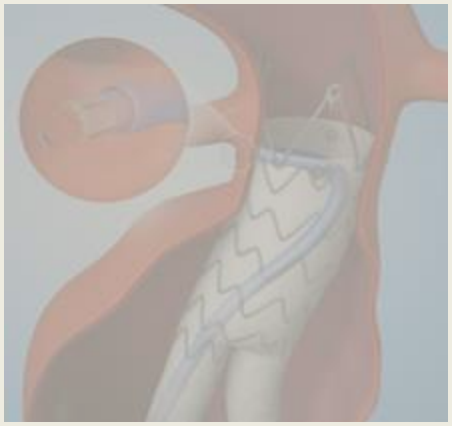


- *More Examples of Real-World Evidence (RWE) Used in Medical Device Regulatory Decisions:* Examples of Real-World Evidence (RWE) Used in Medical Device Regulatory Decisions

Background

StatScience
Stats Power

- 21st Century Cures Act
- Guidance issued to clarify how Real-World Evidence (RWE) may be used to support regulatory decisions.
- Practice in Center of Device and Radiation Health (FDA/CDRH):

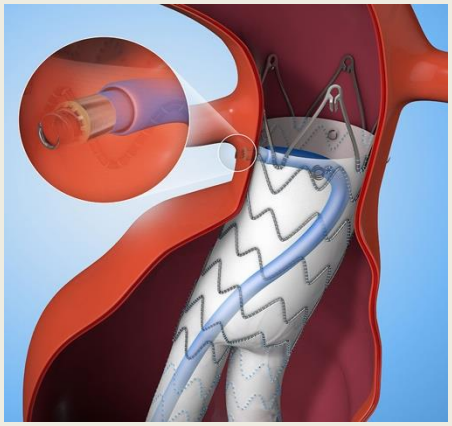


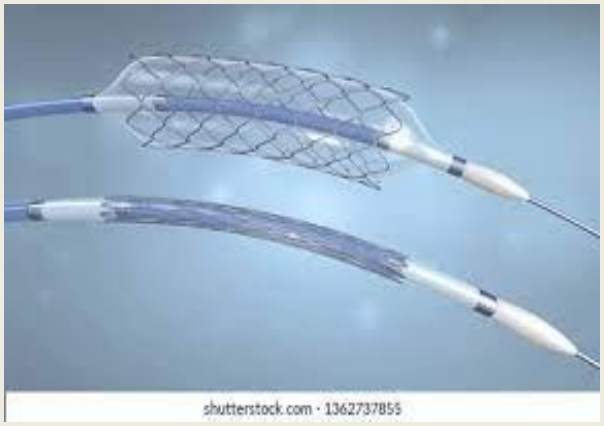
- *More Examples of Real-World Evidence (RWE) Used in Medical Device Regulatory Decisions:* Examples of Real-World Evidence (RWE) Used in Medical Device Regulatory Decisions

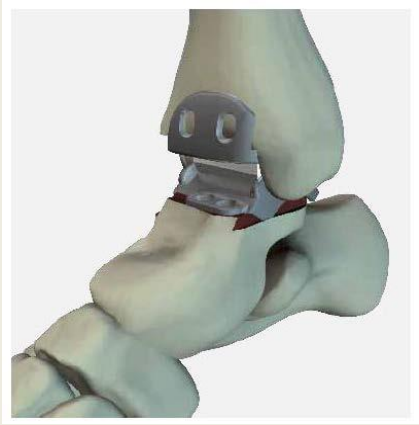
Background

StatScience
Stats Power

- 21st Century Cures Act
- Guidance issued to clarify how Real-World Evidence (RWE) may be used to support regulatory decisions.
- Practice in Center of Device and Radiation Health (FDA/CDRH):



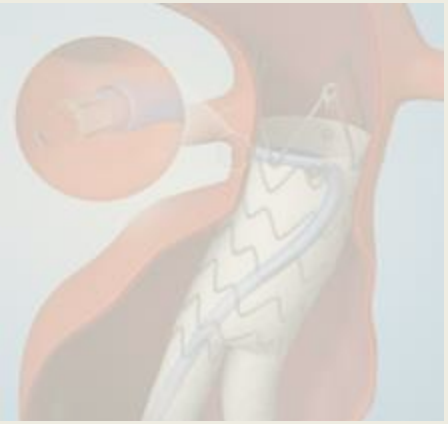


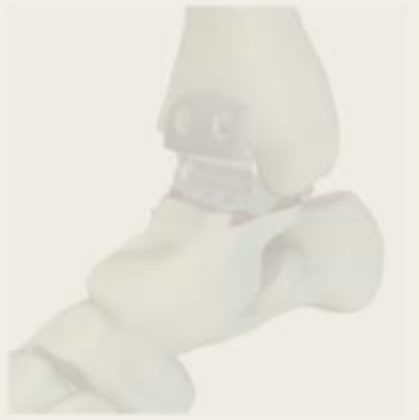


- *More Examples of Real-World Evidence (RWE) Used in Medical Device Regulatory Decisions:* Examples of Real-World Evidence (RWE) Used in Medical Device Regulatory Decisions

Background

- 21st Century Cures Act
- Guidance issued to clarify how Real-World Evidence (RWE) may be used to support regulatory decisions.
- Practice in Center of Device and Radiation Health (FDA/CDRH):



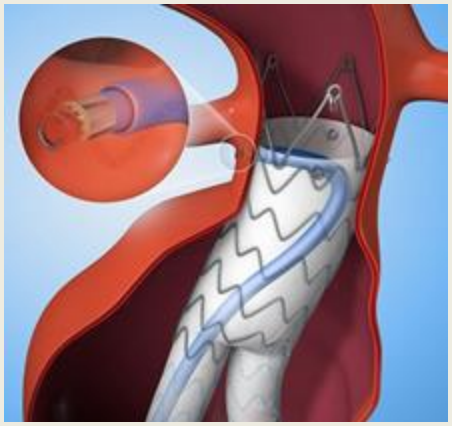


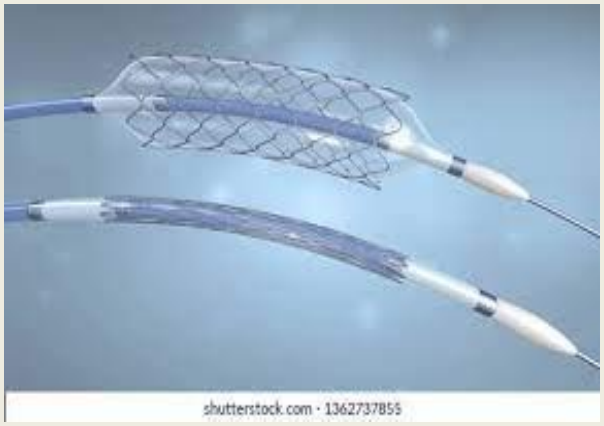
- *More Examples of Real-World Evidence (RWE) Used in Medical Device Regulatory Decisions:* [Examples of Real-World Evidence (RWE) Used in Medical Device Regulatory Decisions]

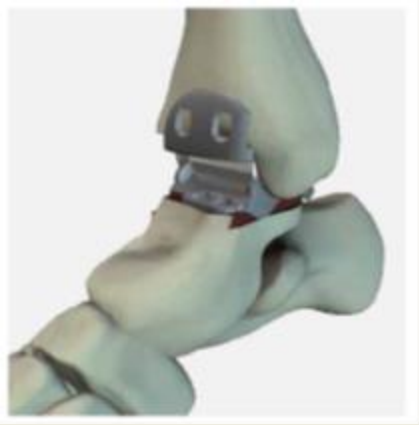
Background

StatScience
Stats Power

- 21st Century Cures Act
- Guidance issued to clarify how Real-World Evidence (RWE) may be used to support regulatory decisions.
- Practice in Center of Device and Radiation Health (FDA/CDRH):







- *More Examples of Real-World Evidence (RWE) Used in Medical Device Regulatory Decisions:* [Examples of Real-World Evidence (RWE) Used in Medical Device Regulatory Decisions]

Background Cont.

StatScience
Stats Power

Orthopedic and Surgical Device Area: Hypothetical Example

- Pivotal Clinical Trial:
  - Prospective, single-arm, minimum 24 months follow-up, multi-center study
  - Sample Size (typical case):
    - 100 ~ 200 subjects per treatment arm
    - 50 ~ 70 subjects for HDE (Humanitarian Device Exemption) case
  - Effectiveness Assessment: Average Treatment Effect (ATE) or Average Treatment Effect on the Treated Arm (ATT)

- External Data Sources:
  - Historical clinical trials and/or Registry data

- External Data Utilization:
  - Construct and/or Augment the control group

- Common Statistical Method: Propensity Score (PS) Subclassification/Stratification, PS Matching.

Background Cont.

StatScience
Stats Power

Orthopedic and Surgical Device Area: Hypothetical Example

- Pivotal Clinical Trial:
  - Prospective, single-arm, minimum 24 months follow-up, multi-center study
  - Sample Size (typical case):
    - 100 ~ 200 subjects per treatment arm
    - 50 ~ 70 subjects for HDE (Humanitarian Device Exemption) case
  - Effectiveness Assessment: Average Treatment Effect (ATE) or Average Treatment Effect on the Treated Arm (ATT)

- External Data Sources:
  - Historical clinical trials and/or Registry data

- External Data Utilization:
  - Construct and/or Augment the control group

- Common Statistical Method: Propensity Score (PS) Subclassification/Stratification, PS Matching.

Background Cont.

StatScience
Stats Power

Orthopedic and Surgical Device Area: Hypothetical Example

- Pivotal Clinical Trial:
  - Prospective, single-arm, minimum 24 months follow-up, multi-center study
  - Sample Size (typical case):
    - 100 ~ 200 subjects per treatment arm
    - 50 ~ 70 subjects for HDE (Humanitarian Device Exemption) case
  - Effectiveness Assessment: Average Treatment Effect (ATE) or Average Treatment Effect on the Treated Arm (ATT)

- External Data Sources:
  - Historical clinical trials and/or Registry data

- External Data Utilization:
  - Construct and/or Augment the control group

- Common Statistical Method: Propensity Score (PS) Subclassification/Stratification, PS Matching.

Background Cont.

StatScience
Stats Power

Orthopedic and Surgical Device Area: Hypothetical Example

- Pivotal Clinical Trial:
  - Prospective, single-arm, minimum 24 months follow-up, multi-center study
  - Sample Size (typical case):
    - 100 ~ 200 subjects per treatment arm
    - 50 ~ 70 subjects for HDE (Humanitarian Device Exemption) case
  - Effectiveness Assessment: Average Treatment Effect (ATE) or Average Treatment Effect on the Treated Arm (ATT)

- External Data Sources:
  - Historical clinical trials and/or Registry data

- External Data Utilization:
  - Construct and/or Augment the control group

- Common Statistical Method: Propensity Score (PS) Subclassification/Stratification, PS Matching.

Background Cont.

StatScience
Stats Power

Orthopedic and Surgical Device Area: Hypothetical Example

- Pivotal Clinical Trial:
  - Prospective, single-arm, minimum 24 months follow-up, multi-center study
  - Sample Size (typical case):
    - 100 ~ 200 subjects per treatment arm
    - 50 ~ 70 subjects for HDE (Humanitarian Device Exemption) case
  - Effectiveness Assessment: Average Treatment Effect (ATE) or Average Treatment Effect on the Treated Arm (ATT)

- External Data Sources:
  - Historical clinical trials and/or Registry data

- External Data Utilization:
  - Construct and/or Augment the control group

- Common Statistical Method: Propensity Score (PS) Subclassification/Stratification, PS Matching.

- Q: How do we know the borrowed historic data are comparable?

Propensity Score Method

StatScience
Stats Power

- **Propensity Score (PS):** the probability e_i that a subject receives investigational treatment (T) (rather than control (C)) conditioning on observed baseline covariates.

$$e_i = \Pr(Z_i = 1 | X_i)$$

Propensity Score Method

StatScience
Stats Power

- **Propensity Score (PS):** the probability that a subject receives investigational treatment (T) (rather than control (C)) conditioning on observed baseline covariates.

$$e_i = \Pr(Z_i = 1 | X_i)$$

- PS is a **balancing score**: conditional on the propensity score, the distribution of measured baseline covariates is similar between treated (T) and untreated (C) subjects, under

  No Unmeasured Confounding Assumption (Rosenbaum and Rubin D.B. 1983).

Propensity Score Method

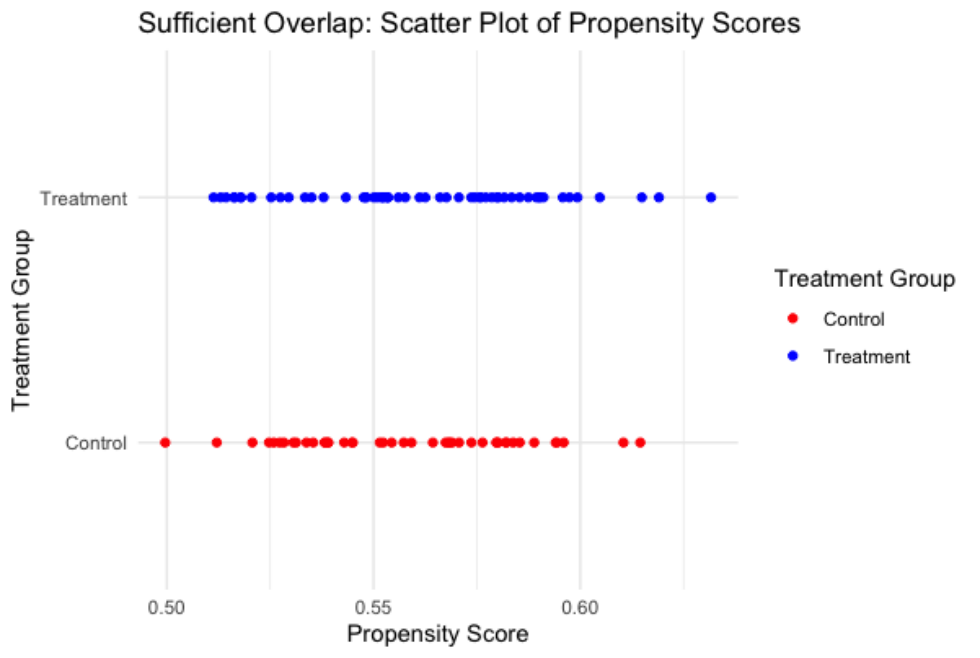
StatScience
Stats Power

- **Propensity Score (PS):** the probability e_i that a subject receives investigational treatment (T) (rather than control (C)) conditioning on observed baseline covariates.

$$e_i = \Pr(Z_i = 1 | X_i)$$

- PS is a **balancing score**: conditional on the propensity score, the distribution of measured baseline covariates is similar between treated (T) and untreated (C) subjects, under ***No Unmeasured Confounding Assumption.***

- PS design Simultaneously balance many observed covariates between two treatment groups, thereby eliminating the bias due to imbalance in baseline covariates



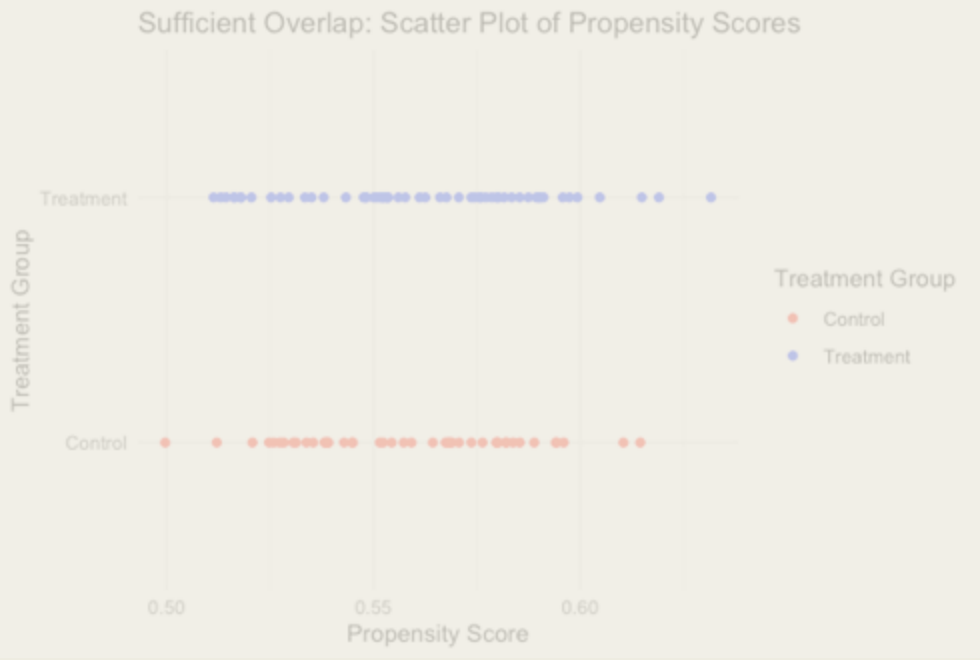
Propensity Score Method

StatScience
Stats Power

- **Propensity Score (PS):** the probability e_i that a subject receives investigational treatment (T) (rather than control (C)) conditioning on observed baseline covariates.

$$e_i = \Pr(Z_i = 1 | X_i)$$

- PS is a **balancing score**: conditional on the propensity score, the distribution of measured baseline covariates is similar between treated (T) and untreated (C) subjects, under ***No Unmeasured Confounding Assumption.***



- PS design Simultaneously balance many observed covariates between two treatment groups, thereby eliminating the bias due to imbalance in baseline covariates

- PS is an outcome-free design: NO outcome is needed in the design stage and hence reduce bias from post hoc analysis (mimic RCT)

Propensity Score Method

StatScience
Stats Power

- **Propensity Score (PS):** the probability that a subject e_i receives investigational treatment (T) (rather than control (C)) conditioning on observed baseline covariates.

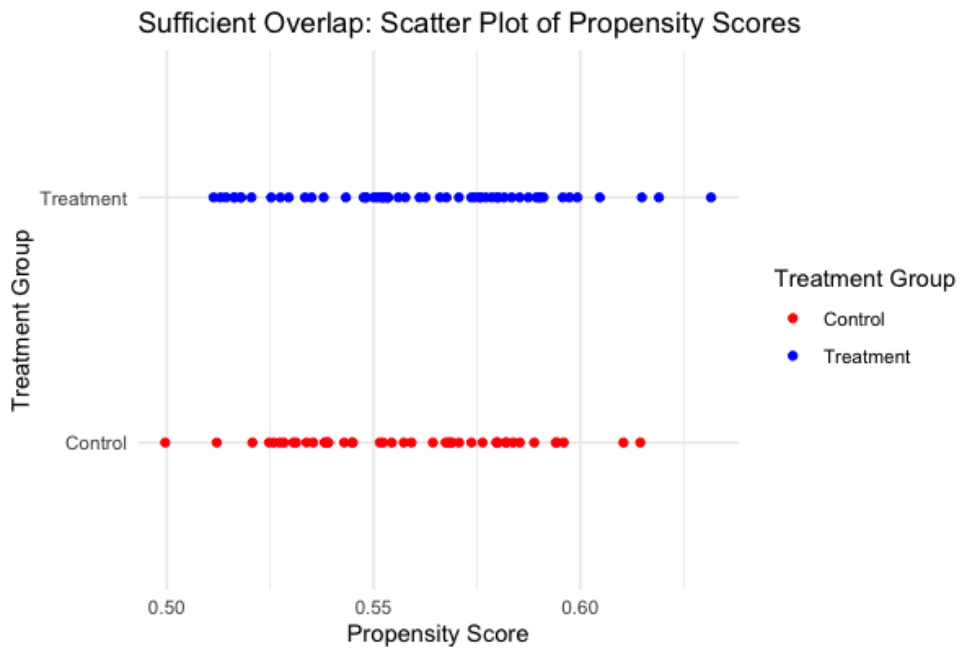
$$e_i = \Pr(Z_i = 1 | X_i)$$

- PS is a **balancing score**: conditional on the propensity score, the distribution of measured baseline covariates is similar between treated (T) and untreated (C) subjects, under

  No Unmeasured Confounding Assumption (Rosenbaum and Rubin, 1983).

- PS design Simultaneously balance many observed covariates between two treatment groups, thereby eliminating the bias due to imbalance in baseline covariates

- PS is an outcome-free design: NO outcome is needed in the design stage and hence reduce bias from post hoc analysis (mimic RCT)



Propensity Score Method Cont.

- **PS modeling is an Iterative Procedure:**

Propensity Score Method Cont.

StatScience
Stats Power

- **PS modeling is an Iterative Procedure:**
- **Step 1.** Pre-determine the baseline prognostic variables, identify appropriate historical control for comparison, pre-define outcome interest (ATE vs ATT).

		Baseline Covariates			
Subject	**Trt.**	**age**	**gender**	**BMI**	**…**
1	T	60	M	27.8	
2	T	45	F	24	
3	T	70	M	28.1	
…					
10	C	65	F	25.6	
11	C	50	F	22.3	

Propensity Score Method Cont.

StatScience
Stats Power

- **PS modeling is an Iterative Procedure:**
- **Step 1.** Pre-determine the baseline prognostic variables, identify appropriate historical control for comparison, pre-define outcome interest (ATE vs ATT).
- **Step 2.** Perform logistic regression, the response is the binary variable treatment arm (e.g., 1) and control arm (e.g., 0), and the covariates are the prognostic variables obtained from step 1, then calculate the propensity scores.

		Baseline Covariates			
Subject	**Trt.**	**age**	**gender**	**BMI**	**…**
1	T	60	M	27.8	
2	T	45	F	24	
3	T	70	M	28.1	
…					
10	C	65	F	25.6	
11	C	50	F	22.3	

Logistic Regression →

Subject	**Trt.**	**PS**
1	T	0.17
2	T	0.5
3	T	0.13
…		
10	C	0.48
11	C	0.15

Propensity Score Method Cont.

StatScience
Stats Power

- **PS modeling is an Iterative Procedure:**
- **Step 1.** Pre-determine the baseline prognostic variables, identify appropriate historical control for comparison, pre-define outcome interest (ATE vs ATT).
- **Step 2.** Perform logistic regression, the response is the binary variable treatment arm (e.g., 1) and control arm (e.g., 0), and the covariates are the prognostic variables obtained from step 1, then calculate the propensity scores.
- **Step 3.** Using PS matching or stratification method, identify group of patients in each arm for comparison. Then evaluate the PS distribution, baseline covariates distribution (along with summary stats, e.g., SMD, standard mean difference). If they are all in pre-defined regions, then we are done, ow go to step 1 and re-do these steps (e.g., adding interaction terms).

		Baseline Covariates			
Subject	Trt.	age	gender	BMI	…
1	T	60	M	27.8	
2	T	45	F	24	
3	T	70	M	28.1	
…					
10	C	65	F	25.6	
11	C	50	F	22.3	

Logistic Regression

Subject	Trt.	PS
1	T	0.17
2	T	0.5
3	T	0.13
…		
10	C	0.48
11	C	0.15

PS Stratification

Subject	Trt.	PS	PS Strata
1	T	0.17	1
3	T	0.13	
11	C	0.15	
…			
10	C	0.48	2
2	T	0.5	

Sufficient PS overlaps?

No

Yes

Done

Propensity Score Method Cont.

StatScience
Stats Power

- **PS modeling is an Iterative Procedure:**
- **Step 1.** Pre-determine the baseline prognostic variables, identify appropriate historical control for comparison, pre-define outcome interest (ATE vs ATT).
- **Step 2.** Perform logistic regression, the response is the binary variable treatment arm (e.g., 1) and control arm (e.g., 0), and the covariates are the prognostic variables obtained from step 1, then calculate the propensity scores.
- **Step 3.** Using PS matching or stratification method, identify group of patients in each arm for comparison, evaluate the PS distribution, baseline covariates distribution (along with summary stats, e.g., SMD, standard mean difference). If they are all in pre-defined regions, then we are done, ow go to step 1 and re-do these steps (e.g., adding interaction terms).

		Baseline Covariates			
Subject	Trt.	age	gender	BMI	…
1	T	60	M	27.8	
2	T	45	F	24	
3	T	70	M	28.1	
…					
10	C	65	F	25.6	
11	C	50	F	22.3	

Logistic Regression

Subject	Trt.	PS
1	T	0.17
2	T	0.5
3	T	0.13
…		
10	C	0.48
11	C	0.15

PS Stratification

Subject	Trt.	PS	PS Strata
1	T	0.17	1
3	T	0.13	
11	C	0.15	
…			
10	C	0.48	2
2	T	0.5	

Sufficient PS overlaps?

No

Yes

Done

Propensity Score Method Discussions

1. **PS Estimation:** One does not have to use **MLE** to estimate the propensity scores. Other methods such as machine learning tree-based methods or robust estimation method can also be used.

Propensity Score Method Discussions

StatScience
Stats Power

1. **PS Estimation:** One does not have to use **MLE** to estimate the propensity scores. Other methods such as machine learning tree-based methods or robust estimation method can also be used.
2. **Choice of Covariates:** Mainly depend on clinical team's opinion, once determined, cannot arbitrarily add or drop covariates for the sake of PS balance.

Propensity Score Method Discussions

StatScience
Stats Power

1. **PS Estimation:** One does not have to use **MLE** to estimate the propensity scores. Other methods such as machine learning tree-based methods or robust estimation method can also be used.
2. **Choice of Covariates:** Mainly depend on clinical team's opinion, once determined, cannot arbitrarily add or drop covariates for the sake of PS balance.
3. **How to check sufficient overlap:**

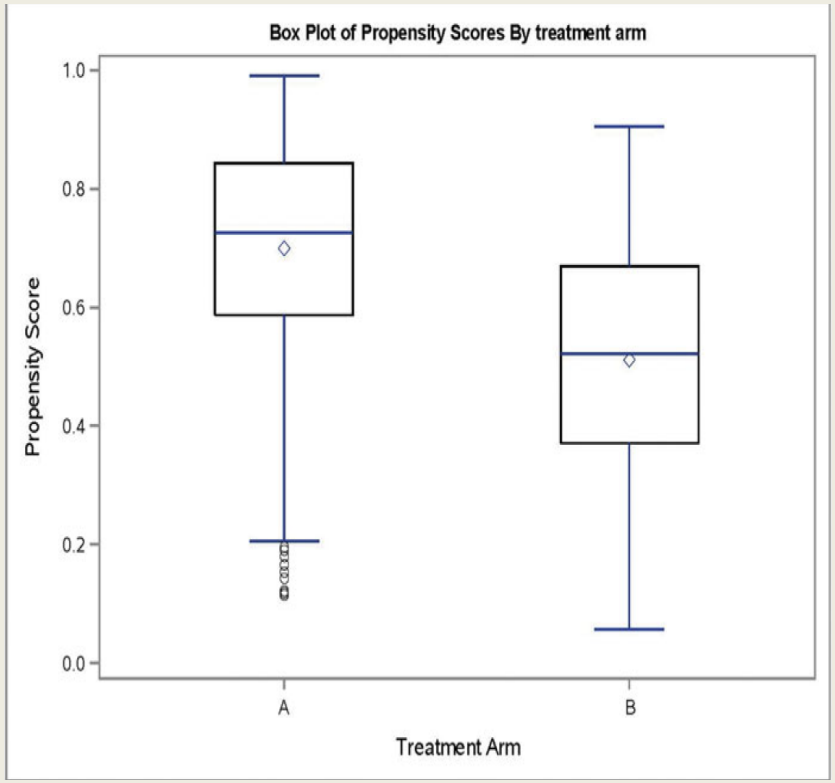
   Boxplot & Standardized Mean Difference (SMD): |SMD| <= 0.1/0.25 ➔ sufficient overlap

   - **3.1 Overall check on propensity scores through boxplot & average standardized difference (ASD)**
   - **3.2 For continuous covariate:**

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

   - **3.3 For dichotomos covariate:**

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



Propensity Score Method Discussions

StatScience
Stats Power

1. **PS Estimation:** One does not have to use **MLE** to estimate the propensity scores. Other methods such as machine learning tree-based methods or robust estimation method can also be used.
2. **Choice of Covariates:** Mainly depend on clinical team's opinion, once determined, cannot arbitrarily add or drop covariates for the sake of PS balance.
3. **How to check sufficient overlap:**

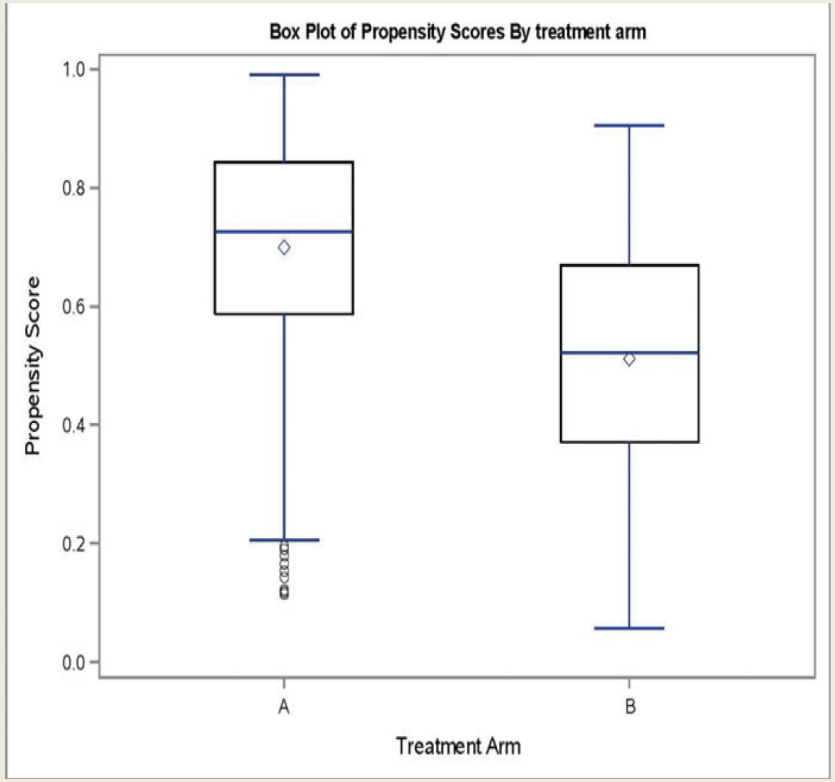
   Boxplot & Standardized Mean Difference (SMD): |SMD| <= 0.1/0.25 ➔ sufficient overlap

   - **3.1 Overall check on propensity scores through boxplot & average standardized difference (ASD)**
   - **3.2 For continuous covariate:**

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

   - **3.3 For dichotomos covariate:**

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



Propensity Score Method: Clinical Practice Considerations

StatScience
Stats Power

1. It is strongly recommended to hire **independent** statistician to conduct PS model design. The independent statistician should remain blinded until the PS model is fixed.

Propensity Score Method: Clinical Practice Considerations

StatScience
Stats Power

1. It is strongly recommended to hire **independent** statistician to conduct PS model design. The independent statistician should remain blinded until the PS model is fixed.
2. Regarding the PS modeling, generally one **CANNOT** purely add or drop variables to achieve PS balance.

Propensity Score Method: Clinical Practice Considerations

StatScience
Stats Power

1. It is strongly recommended to hire **independent** statistician to conduct PS model design. The independent statistician should remain blinded until the PS model is fixed.
2. Regarding the PS modeling, generally one **CANNOT** purely add or drop variables to achieve PS balance.
3. One **CANNOT** exclude the subjects in order to achieve PS balance, because such exclusion may change the indicated population.

Propensity Score Method: Clinical Practice Considerations

StatScience
Stats Power

1. It is strongly recommended to hire **independent** statistician to conduct PS model design. The independent statistician should remain blinded until the PS model is fixed.
2. Regarding the PS modeling, generally one **CANNOT** purely add or drop variables to achieve PS balance.
3. One **CANNOT** exclude the subjects in order to achieve PS balance, because such exclusion may change the indicated population.
4. It does **NOT** matter when the logistic regression model is mis-specified (Peter Austin 2011: “The distribution of the mis-specified propensity score was similar to that of the correctly specified propensity score”).

Propensity Score Method: Clinical Practice Considerations

StatScience
Stats Power

1. It is strongly recommended to hire **independent** statistician to conduct PS model design. The independent statistician should remain blinded until the PS model is fixed.
2. Regarding the PS modeling, generally one **CANNOT** purely add variables to achieve PS balance.
3. One **CANNOT** exclude the subjects in order to achieve PS balance, because such exclusion may change the indicated population.
4. It does **NOT** matter when the logistic regression model is mis-specified (Peter Austin 2011: “The distribution of the mis-specified propensity score was similar to that of the correctly specified propensity score”).
5. The PS modeling is different between ATE and ATT.

Propensity Score Method: Clinical Practice Considerations

StatScience
Stats Power

1. It is strongly recommended to hire **independent** statistician to conduct PS model design. The independent statistician should remain blinded until the PS model is fixed.
2. Regarding the PS modeling, generally one **CANNOT** purely add variables to achieve PS balance.
3. One **CANNOT** exclude the subjects in order to achieve PS balance, because such exclusion may change the indicated population.
4. It does **NOT** matter when the logistic regression model is mis-specified (Peter Austin 2011: "The distribution of the mis-specified propensity score was similar to that of the correctly specified propensity score").
5. The PS modeling is different between ATE and ATT.
6. Different PS methods (e.g., stratification method, matching method) may lead to different conclusions in causal inference, it's strongly recommended to pre-specify the primary PS method (others could be sensitivity analysis).

Propensity Score Method: Clinical Practice Considerations

StatScience
Stats Power

1. It is strongly recommended to hire **independent** statistician to conduct PS model design. The independent statistician should remain blinded until the PS model is fixed.
2. Regarding the PS modeling, generally one **CANNOT** purely add variables to achieve PS balance.
3. One **CANNOT** exclude the subjects in order to achieve PS balance, because such exclusion may change the indicated population.
4. It does **NOT** matter when the logistic regression model is mis-specified (Peter Austin 2011: "The distribution of the mis-specified propensity score was similar to that of the correctly specified propensity score").
5. The PS modeling is different between ATE and ATT.
6. Different PS methods (e.g., stratification method, matching method) may lead to different conclusions in causal inference, it's strongly recommended to pre-specify the primary PS method (others could be sensitivity analysis).

Table: PS Stratification Illustration - Number of Subjects in each arm

	Stratum 1	Stratum 2	Stratum 3	Stratum 4	Stratum 5	Total
Treatment Arm	50	90	100	120	150	510
Control Arm	100	80	60	40	30	310

Propensity Score Method: Clinical Practice Considerations

StatScience
Stats Power

1. It is strongly recommended to hire **independent** statistician to conduct PS model design. The independent statistician should remain blinded until the PS model is fixed.
2. Regarding the PS modeling, generally one **CANNOT** purely add variables to achieve PS balance.
3. One **CANNOT** exclude the subjects in order to achieve PS balance, because such exclusion may change the indicated population.
4. It does **NOT** matter when the logistic regression model is mis-specified (Peter Austin 2011: "The distribution of the mis-specified propensity score was similar to that of the correctly specified propensity score").
5. The PS modeling is different between ATE and ATT.
6. Different PS methods (e.g., stratification method, matching method) may lead to different conclusions in causal inference, it's strongly recommended to pre-specify the primary PS method (others could be sensitivity analysis).

Table: PS Stratification Illustration - Number of Subjects in each arm

	Stratum 1	Stratum 2	Stratum 3	Stratum 4	Stratum 5	Total
Treatment Arm	50	90	100	120	150	510
Control Arm	100	80	60	40	30	310

Causal Inference: After Propensity Score

- **0. Remove Blindness:** After reaching PS balance and agreement between the FDA and sponsor, the independent statistician's blindness could be removed.

Causal Inference: After Propensity Score

StatScience
Stats Power

- **0. Remove Blindness:** After reaching PS balance and agreement between the FDA and sponsor, the independent statistician's blindness could be removed.

- **1. Estimation of ATE & ATT:** Y is the response, Z = 1 indicates the subject is treated, and Z = 0 indicates the subject is not treated.

  ATE = E[Y(1)-Y(0)] **ATT = E[Y(1)-Y(0) | Z = 1]**

  For PS Stratification: $\boldsymbol{K}$ denotes the number of stratum, $\overline{\boldsymbol{Y_k}}(\mathbf{1}), \overline{\boldsymbol{Y_k}}(\mathbf{0})$ denote the outcome from the kth treated group and control group, respectively; $\boldsymbol{w_k}$ denotes the weight for the kth stratum.

  $\widehat{\boldsymbol{ATE}} = \sum_{k=1}^{K} \boldsymbol{w_k}[\overline{\boldsymbol{Y_k}}(\mathbf{1}) - \overline{\boldsymbol{Y_k}}(\mathbf{0})]$ $\widehat{\boldsymbol{ATT}} = \sum_{k=1}^{K} \boldsymbol{w_{kT}}[\overline{\boldsymbol{Y_{kT}}}(\mathbf{1}) - \overline{\boldsymbol{Y_{kT}}}(\mathbf{0})]$

Causal Inference: After Propensity Score

StatScience
Stats Power

- **0. Remove Blindness:** After reaching PS balance and agreement between the FDA and sponsor, the independent statistician's blindness could be removed.

- **1. Estimation of ATE & ATT:** Y is the response, Z = 1 indicates the subject is treated, and Z = 0 indicates the subject is not treated.

  ATE = E[Y(1)-Y(0)] **ATT = E[Y(1)-Y(0) | Z = 1]**

  For PS Stratification: K denotes the number of stratum, $\overline{Y_k}(1), \overline{Y_k}(0)$ denote the outcome from the kth treated group and control group, respectively; w_k denotes the weight for the kth stratum.

  $$\widehat{ATE} = \sum_{k=1}^{K} w_k[\overline{Y_k}(1) - \overline{Y_k}(0)] \qquad \widehat{ATT} = \sum_{k=1}^{K} w_{kT}[\overline{Y_{kT}}(1) - \overline{Y_{kT}}(0)]$$

  For other PS approach (e.g., PS matching), An alternative estimation method, the **inverse probability weighting (IPW)** method could also used.

  $$\widehat{ATE} = \frac{1}{n}\sum_{i=1}^{n}\frac{Z_i Y_i}{\hat{e}_i} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-Z_i)Y_i}{1-\hat{e}_i} \qquad \widehat{ATT} = \frac{1}{n}\sum_{i=1}^{n} Y_i\left(Z_i + \frac{(1-Z_i)\hat{e}_i}{1-\hat{e}_i}\right)$$

Causal Inference: After Propensity Score

StatScience
Stats Power

- **0. Remove Blindness:** After reaching PS balance and agreement between the FDA and sponsor, the independent statistician's blindness could be removed.

- **1. Estimation of ATE & ATT:** Y is the response, Z = 1 indicates the subject is treated, and Z = 0 indicates the subject is not treated.

  ATE = E[Y(1)-Y(0)] **ATT = E[Y(1)-Y(0) | Z = 1]**

  For PS Stratification: K denotes the number of stratum, $\overline{Y_k}(1), \overline{Y_k}(0)$ denote the outcome from the kth treated group and control group, respectively; w_k denotes the weight for the kth stratum.

  $$\widehat{ATE} = \sum_{k=1}^{K} w_k[\overline{Y_k}(1) - \overline{Y_k}(0)] \qquad \widehat{ATT} = \sum_{k=1}^{K} w_{kT}[\overline{Y_{kT}}(1) - \overline{Y_{kT}}(0)]$$

  For other PS approach (e.g., PS matching), An alternative estimation method, the **inverse probability weighting (IPW)** method could also used.

  $$\widehat{ATE} = \frac{1}{n}\sum_{i=1}^{n}\frac{Z_i Y_i}{\hat{e}_i} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-Z_i)Y_i}{1-\hat{e}_i} \qquad \widehat{ATT} = \frac{1}{n}\sum_{i=1}^{n} Y_i\left(Z_i + \frac{(1-Z_i)\hat{e}_i}{1-\hat{e}_i}\right)$$

- **2. Estimation of Variance of ATE & ATT:** For PS stratification method, the samples within each stratum could be treated as independent samples. Thus, two sample t-test could apply for continuous outcome and proportion test can be used for dichotomous outcome. In practice, the **bootstrap** method could also be used.

Causal Inference: After Propensity Score

StatScience
Stats Power

- **0. Remove Blindness:** After reaching PS balance and agreement between the FDA and sponsor, the independent statistician's blindness could be removed.

- **1. Estimation of ATE & ATT:** Y is the response, Z = 1 indicates the subject is treated, and Z = 0 indicates the subject is not treated.

  ATE = E[Y(1)-Y(0)] **ATT = E[Y(1)-Y(0) | Z = 1]**

  For PS Stratification: K denotes the number of stratum, $\overline{Y_k}(1), \overline{Y_k}(0)$ denote the outcome from the kth treated group and control group, respectively; w_k denotes the weight for the kth stratum.

  $$\widehat{ATE} = \sum_{k=1}^{K} w_k[\overline{Y_k}(1) - \overline{Y_k}(0)] \qquad \widehat{ATT} = \sum_{k=1}^{K} w_{kT}[\overline{Y_{kT}}(1) - \overline{Y_{kT}}(0)]$$

  For other PS approach (e.g., PS matching), An alternative estimation method, the **inverse probability weighting (IPW)** method could also used.

  $$\widehat{ATE} = \frac{1}{n}\sum_{i=1}^{n}\frac{Z_iY_i}{\hat{e}_i} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-Z_i)Y_i}{1-\hat{e}_i} \qquad \widehat{ATT} = \frac{1}{n}\sum_{i=1}^{n} Y_i\left(Z_i + \frac{(1-Z_i)\hat{e}_i}{1-\hat{e}_i}\right)$$

- **2. Estimation of Variance of ATE & ATT:** For PS stratification method, the samples within each stratum could be treated as independent samples. Thus, two sample t-test could apply for continuous outcome and proportion test can be used for dichotomous outcome. In practice, the **bootstrap** method could also be used.

- **3. Primary Estimand:** It is worth pointing out that even if the primary estimand is ATE, the agency (e.g., FDA) may still require to see ATT results. If the PS model are sufficiently overlapped, ATE and ATT estimates are usually similar.

Causal Inference: After Propensity Score

StatScience
Stats Power

- **0. Remove Blindness:** After reaching PS balance and agreement between the FDA and sponsor, the independent statistician's blindness could be removed.

- **1. Estimation of ATE & ATT:** Y is the response, Z = 1 indicates the subject is treated, and Z = 0 indicates the subject is not treated.

  ATE = E[Y(1)-Y(0)] **ATT = E[Y(1)-Y(0) | Z = 1]**

  For PS Stratification: K denotes the number of stratum, $\overline{Y_k}(1), \overline{Y_k}(0)$ denote the outcome from the kth treated group and control group, respectively; w_k denotes the weight for the kth stratum.

  $$\widehat{ATE} = \sum_{k=1}^{K} w_k[\overline{Y_k}(1) - \overline{Y_k}(0)] \qquad \widehat{ATT} = \sum_{k=1}^{K} w_{kT}[\overline{Y_{kT}}(1) - \overline{Y_{kT}}(0)]$$

  For other PS approach (e.g., PS matching), An alternative estimation method, the **inverse probability weighting (IPW)** method could also used.

  $$\widehat{ATE} = \frac{1}{n}\sum_{i=1}^{n}\frac{Z_i Y_i}{\hat{e}_i} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-Z_i)Y_i}{1-\hat{e}_i} \qquad \widehat{ATT} = \frac{1}{n}\sum_{i=1}^{n} Y_i\left(Z_i + \frac{(1-Z_i)\hat{e}_i}{1-\hat{e}_i}\right)$$

- **2. Estimation of Variance of ATE & ATT:** For PS stratification method, the samples within each stratum could be treated as independent samples. Thus, two sample t-test could apply for continuous outcome and proportion test can be used for dichotomous outcome. In practice, the **bootstrap** method could also be used.

- **3. Primary Estimand:** It is worth pointing out that even if the primary estimand is ATE, the agency (e.g., FDA) may still require to see ATT results. If the PS model are sufficiently overlapped, ATE and ATT estimates are usually similar.

Practical Issues

- **No Unmeasured Confounding Assumption**: assumes covariates that affect both the outcome and the treatment assignment have been measured:
  - **Problem:** cannot be verified generally
  - **Solution:** Include all possible relevant covariates into the Propensity Score model

Practical Issues

StatScience
Stats Power

- **No Unmeasured Confounding Assumption**: assumes covariates that affect both the outcome and the treatment assignment have been measured:
  - **Problem:** cannot be verified generally
  - **Solution:** Include all possible relevant covariates into the Propensity Score model
- **Practical Issues:**
  - Often some clinically relevant covariates unobserved, but these covariates could be **correlated** to each other:
    - e.g., different PRO instruments used (Oxford Score vs. Harris Score)
    - Oxford score, Harris score, weight, height, and BMI highly correlated
  - covariates observed but with certain percent **missing** values:
    - with limited External Data Sources, how significant the impact might be?

Practical Issues

StatScience
Stats Power

- **No Unmeasured Confounding Assumption**: assumes covariates that affect both the outcome and the treatment assignment have been measured:
  - **Problem:** cannot be verified generally
  - **Solution:** Include all possible relevant covariates into the Propensity Score model
- **Practical Issues:**
  - Often some clinically relevant covariates unobserved, but these covariates could be correlated to each other:
    - e.g., different PRO instruments used (Oxford Score vs. Harris Score)
    - Oxford score, Harris score, weight, height, and BMI highly correlated
  - covariates observed but with certain percent missing values:
    - with limited External Data Sources, how significant the impact might be?
- A few literature discuss about the missing impact (D'Agostino Jr. 2000 and Liu 2013) However,
  - the PS matching design and abundant external data source
  - How does this impact the causal inference?

Practical Issues

StatScience
Stats Power

- **No Unmeasured Confounding Assumption**: assumes covariates that affect both the outcome and the treatment assignment have been measured:
  - **Problem:** cannot be verified generally
  - **Solution:** Include all possible relevant covariates into the Propensity Score model
- **Practical Issues:**
  - Often some clinically relevant covariates unobserved, but these covariates could be correlated to each other:
    - e.g., different PRO instruments used (Oxford Score vs. Harris Score)
    - Oxford score, Harris score, weight, height, and BMI highly correlated
  - covariates observed but with certain percent missing values:
    - with limited External Data Sources, how significant the impact might be?
- A few literature discuss about the missing impact (D'Agostino Jr. 2000 and Liu 2013) However,
  - the PS matching design and abundant external data source
  - How does this impact the causal inference?

- We try to evaluate these practical issues in simulation settings

Practical Issues

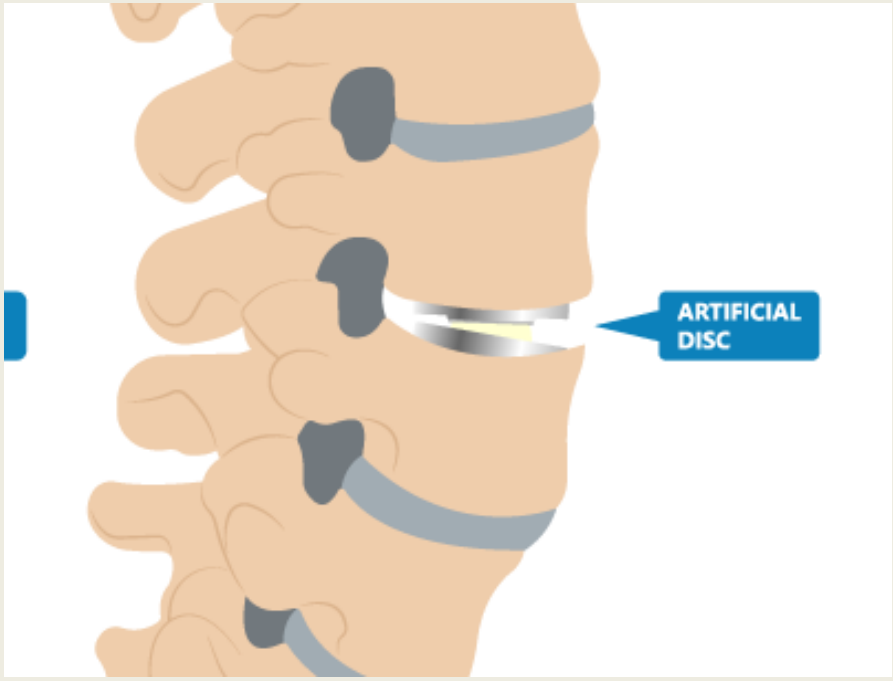
StatScience
Stats Power

- **No Unmeasured Confounding Assumption**: assumes covariates that affect both the outcome and the treatment assignment have been measured:
  - **Problem:** cannot be verified generally
  - **Solution:** Include all possible relevant covariates into the Propensity Score model
- **Practical Issues:**
  - Often some clinically relevant covariates unobserved, but these covariates could be correlated to each other:
    - e.g., different PRO instruments used (Oxford Score vs. Harris Score)
    - Oxford score, Harris score, weight, height, and BMI highly correlated
  - covariates observed but with certain percent missing values:
    - with limited External Data Sources, how significant the impact might be?
- A few literature discuss about the missing impact (D'Agostino Jr. 2000 and Liu 2013) However,
  - the PS matching design and abundant external data source
  - How does this impact the causal inference?
- We try to evaluate these practical issues in simulation settings

Hypothetical Example

StatScience
Stats Power

- The investigational device is a cervical artificial disc to maintain/improve motion of a functional spinal unit when replacing a diseased native disc.



Hypothetical Example

StatScience
Stats Power

- The investigational device is a cervical artificial disc to maintain/improve motion of a functional spinal unit when replacing a diseased native disc.

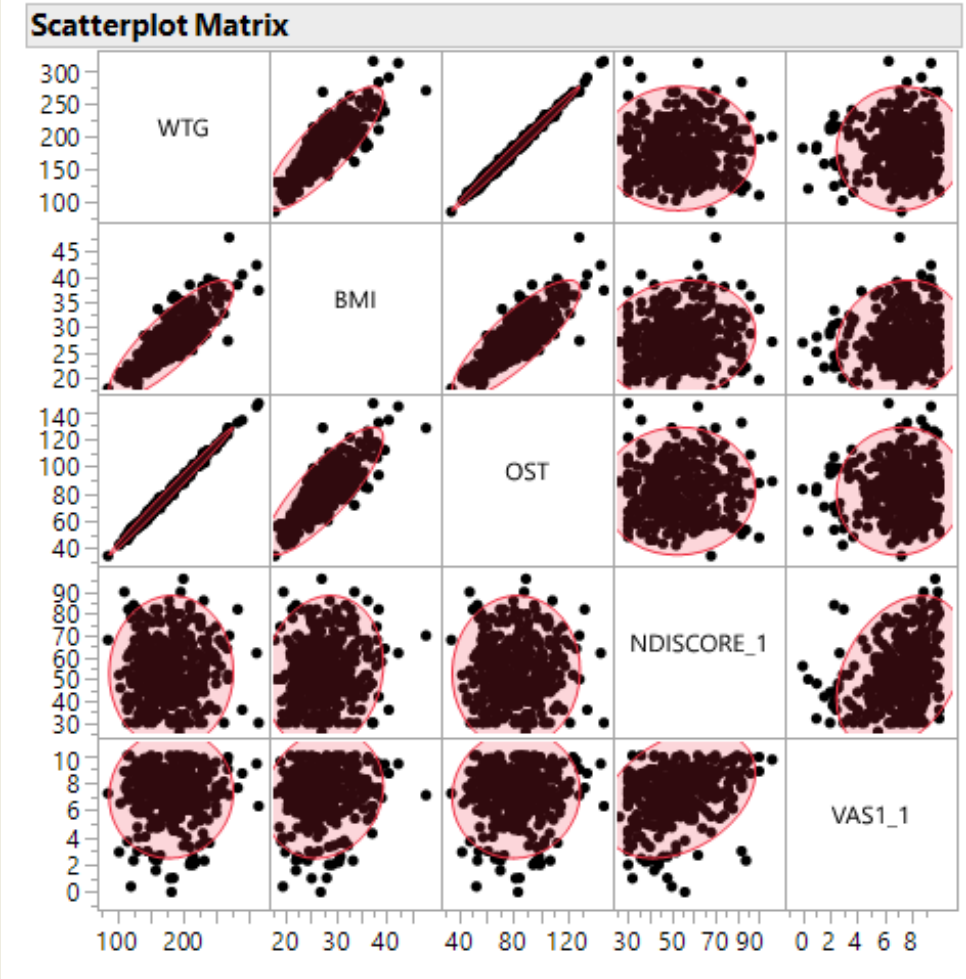
- Baseline Covariates (10):
  - Demographics:
    age, gender, BMI, weight,
    smoking status
  - Disease Characteristics:
    physical score,
    neck disability score,
    arm pain score,
    osteoporosis self-assessment score
    duration of symptoms

Hypothetical Example

StatScience
Stats Power

- The investigational device is a cervical artificial disc to maintain/improve motion of a functional spinal unit when replacing a diseased native disc.

- Baseline Covariates (10):
  - Demographics:
    age, gender, BMI, weight,
    smoking status
  - Disease Characteristics:
    physical score,
    neck disability score,
    arm pain score,
    osteoporosis self-assessment score
    duration of symptoms



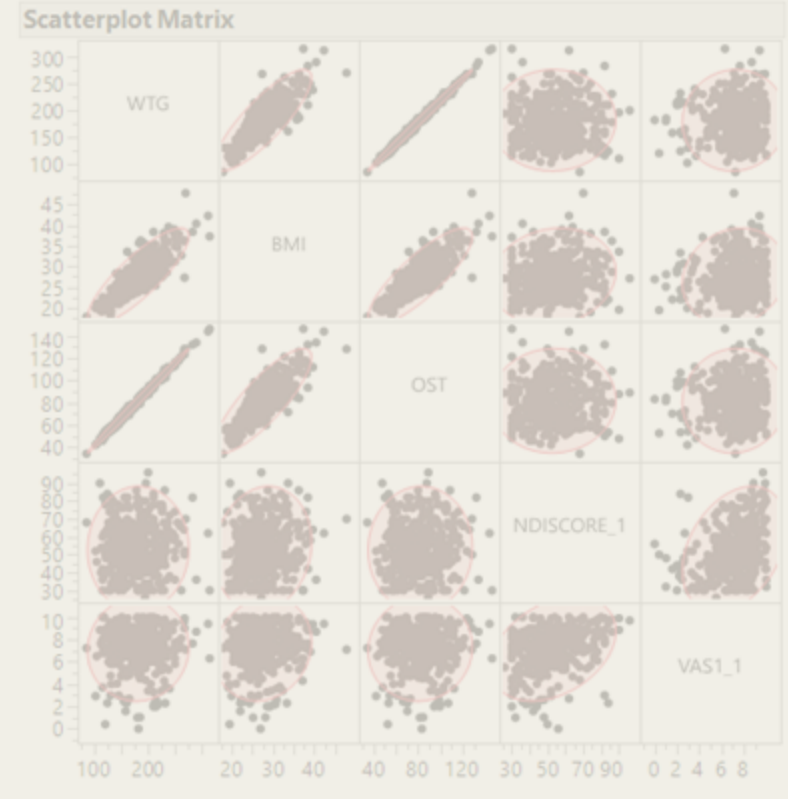
Hypothetical Example

StatScience
Stats Power

- The investigational device is a cervical artificial disc to maintain/improve motion of a functional spinal unit when replacing a diseased native disc.

- Baseline Covariates (10):
  - Demographics:
    age, gender, BMI, weight,
    smoking status
  - Disease Characteristics:
    physical score,
    neck disability score,
    arm pain score,
    osteoporosis self-assessment score
    duration of symptoms

- Outcomes:
  - **Primary**: binary Composite Endpoint at 24 months post-operation - Success/Failure
  - **Secondary**: continuous Spine Function Score



Hypothetical Example

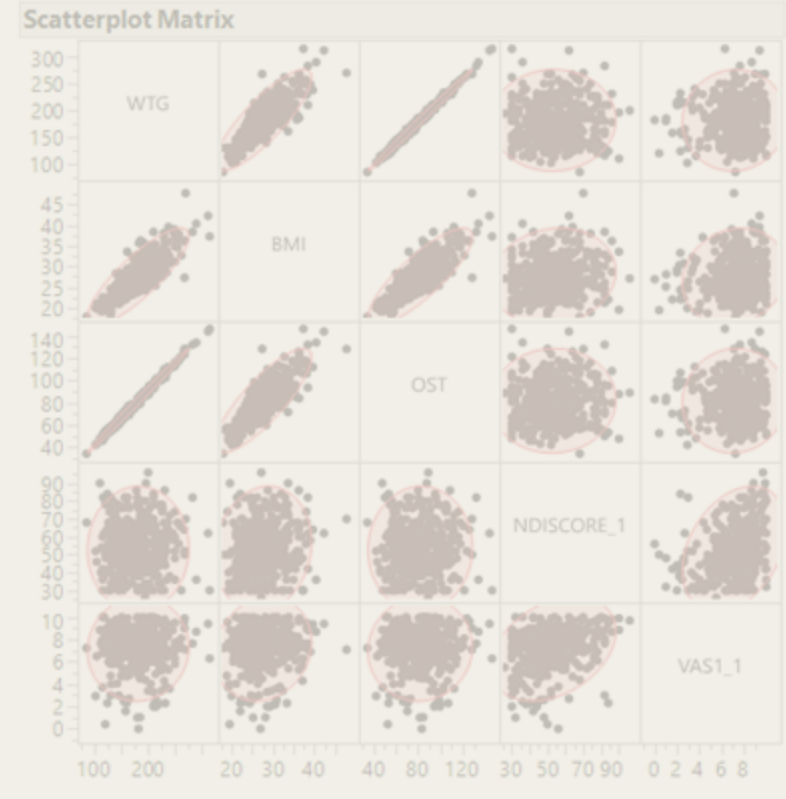
StatScience
Stats Power

- The investigational device is a cervical artificial disc to maintain/improve motion of a functional spinal unit when replacing a diseased native disc.

- Baseline Covariates (10):
  - Demographics:
    age, gender, BMI, weight,
    smoking status
  - Disease Characteristics:
    physical score,
    neck disability score,
    arm pain score,
    osteoporosis self-assessment score
    duration of symptoms

- Outcomes:
  - Primary: binary Composite Endpoint at 24 months post-operation - Success/Failure
  - Secondary: continuous Spine Function Score

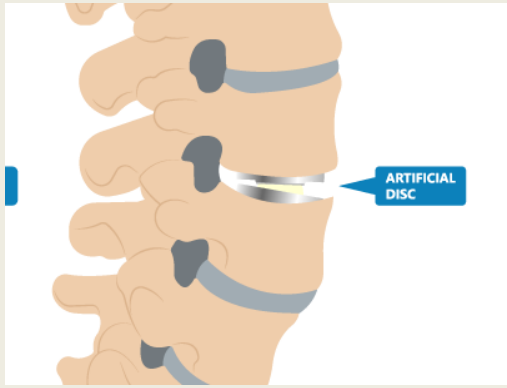
- Primary Hypothesis Test: **Non-Inferiority** Test on Difference in Proportion (Treat – Control)



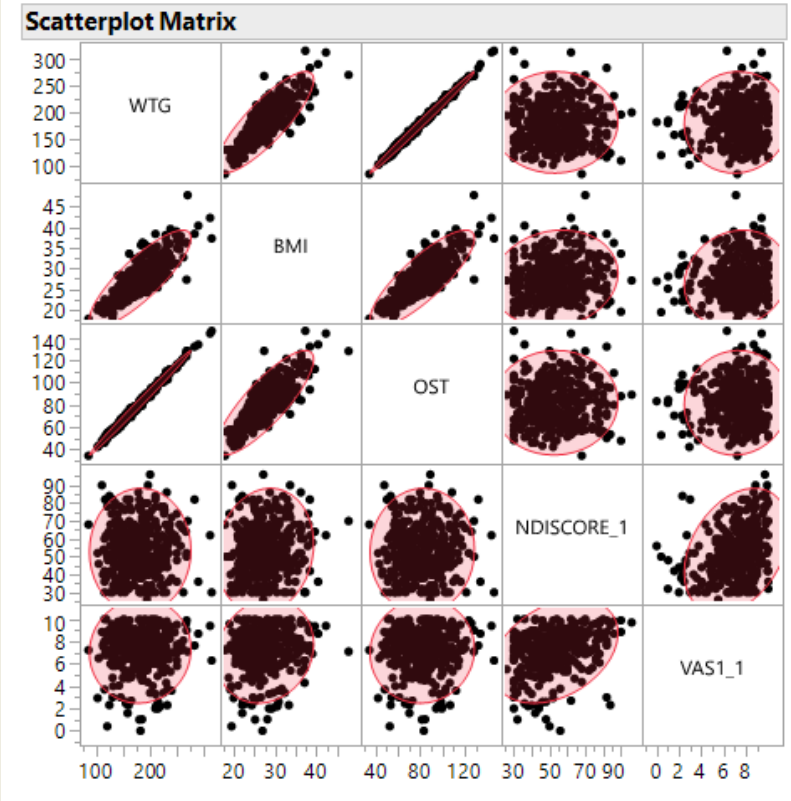
Hypothetical Example

StatScience
Stats Power

- The investigational device is a cervical artificial disc to maintain/improve motion of a functional spinal unit when replacing a diseased native disc.



- Baseline Covariates (10):
  - Demographics:
    age, gender, BMI, weight,
    smoking status
  - Disease Characteristics:
    physical score,
    neck disability score,
    arm pain score,
    osteoporosis self-assessment score
    duration of symptoms



- Outcomes:
  - Primary: binary Composite Endpoint at 24 months post-operation - Success/Failure
  - Secondary: continuous Spine Function Score

- Primary Hypothesis Test: **Non-Inferiority** Test on Difference in Proportion (Treat – Control)

Hypothetical Example Cont.

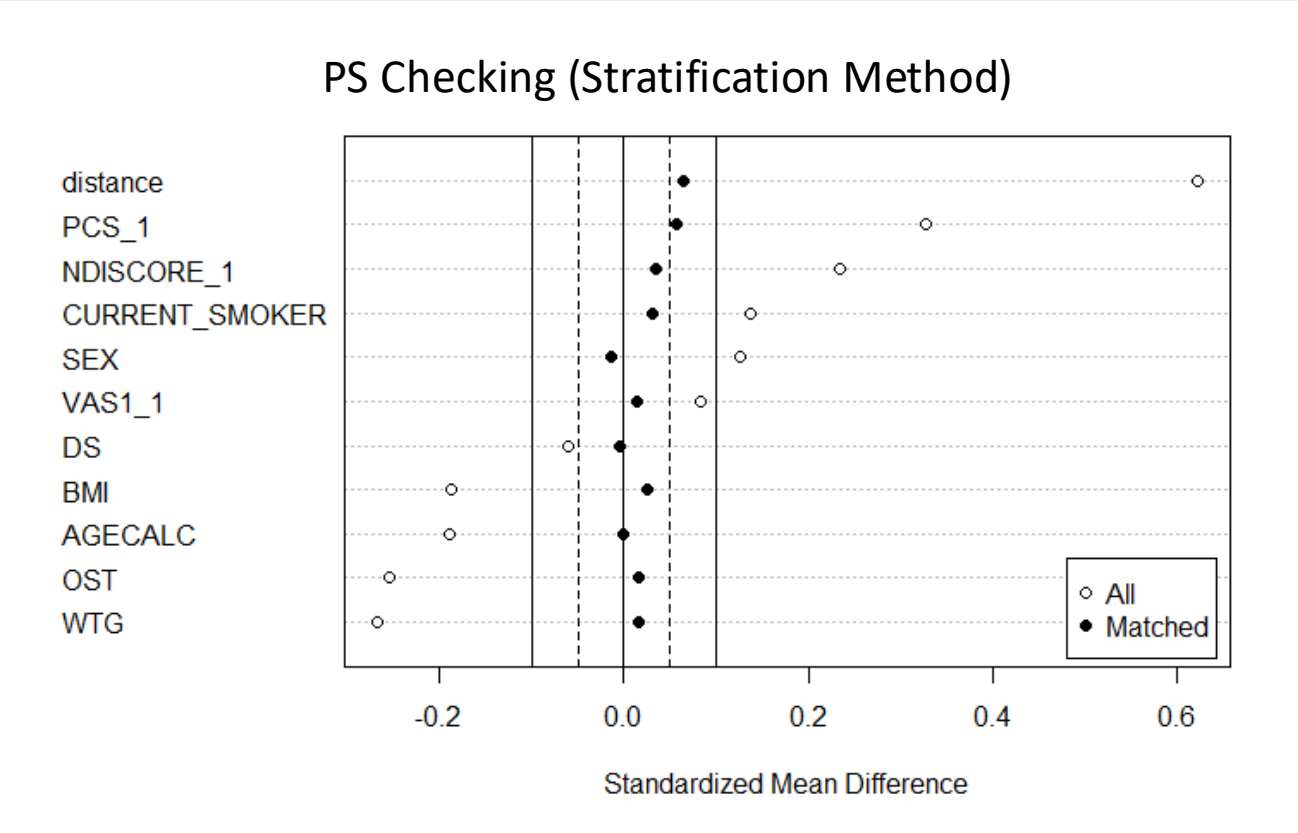
StatScience
Stats Power

- Base PS Model: Main-effect model with all 10 covariates
- With Complete Data: 100 trts, 150 ctls, true ATE: -0.063

Hypothetical Example Cont.

StatScience
Stats Power

- Base PS Model: Main-effect model with all 10 covariates
- With Complete Data: 100 trts, 150 ctls, true ATE: -0.063



Simulation Scenarios

StatScience
Stats Power

- Sample Size:

Situation	Treat	External Data
HDE	60	100
Small	**100**	**150**
Moderate	180	250

- External Data Scenarios:
  - Missing in Baseline Covariates:
    - 20% and 40% missing in partial covariates or all covariates
  - Missing in Outcomes (binary outcome & continuous outcome):
    - 20% and 40% missing
  - Missing in both Covariates and Outcomes
  - Unobserved Covariates:
    - Unobserved covariates (in control) highly correlated to observed covariates (in treatment)
- External Data Utilization Method: PS Stratification and PS Weighting
- Estimands: ATT and ATE
- **Assumption:** Missing at Random (**MAR**)

Scenario: 20% Missing

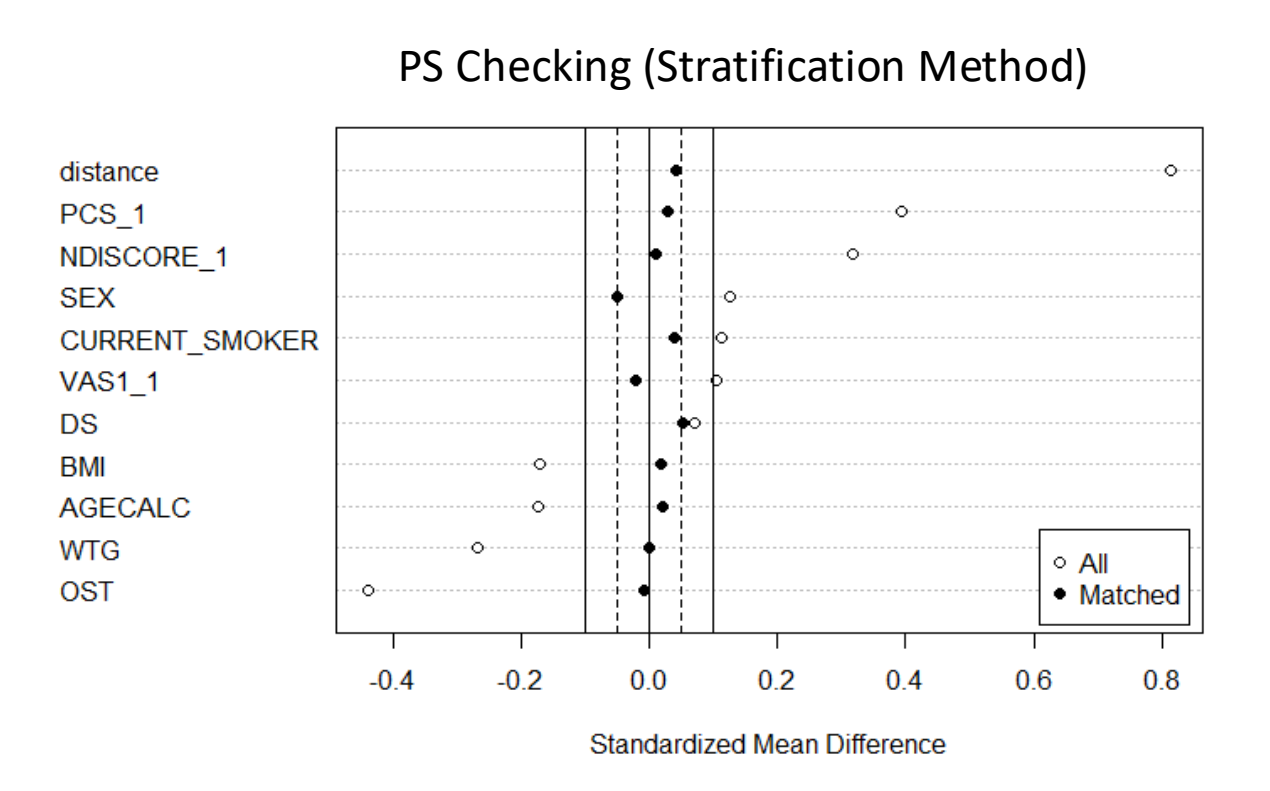
StatScience
Stats Power

- **Setting:** 100 trt, 150 ctl, 20% missing; Estimand: ATE; **Outcome**: proportion of success CCS6
- **Imputation Methods:**
  - Single Imputation: mean for continuous; module for categorical (commonly used in device application)
  - Multiple Imputation

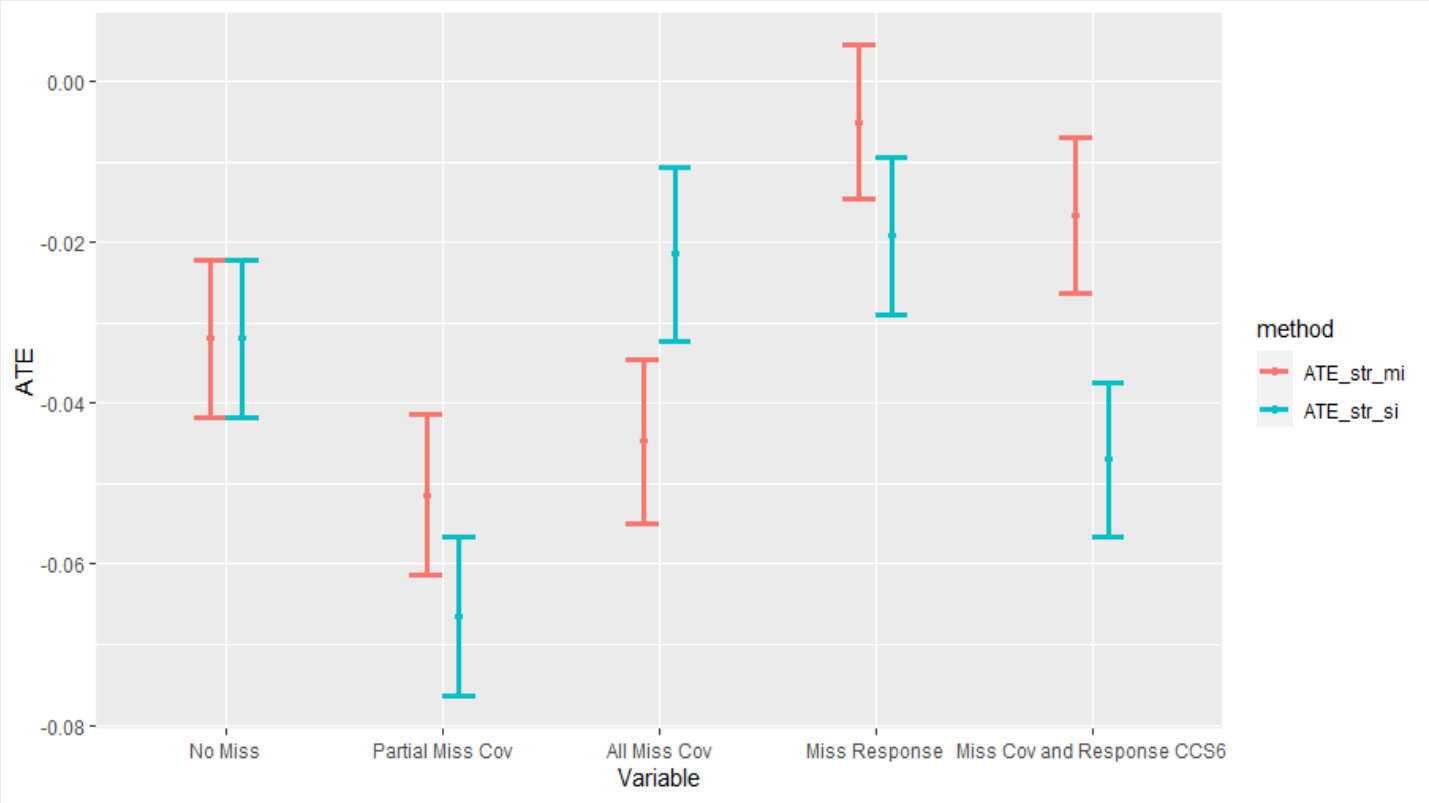
Scenario: 20% Missing

StatScience
Stats Power

- **Setting:** 100 trt, 150 ctl, 20% missing; **Estimand**: ATE; **Outcome**: proportion of success CCS6
- **Imputation Methods:**
  - Single Imputation: mean for continuous; module for categorical (commonly used in device application)
  - Multiple Imputation



Scenario: 20% Missing



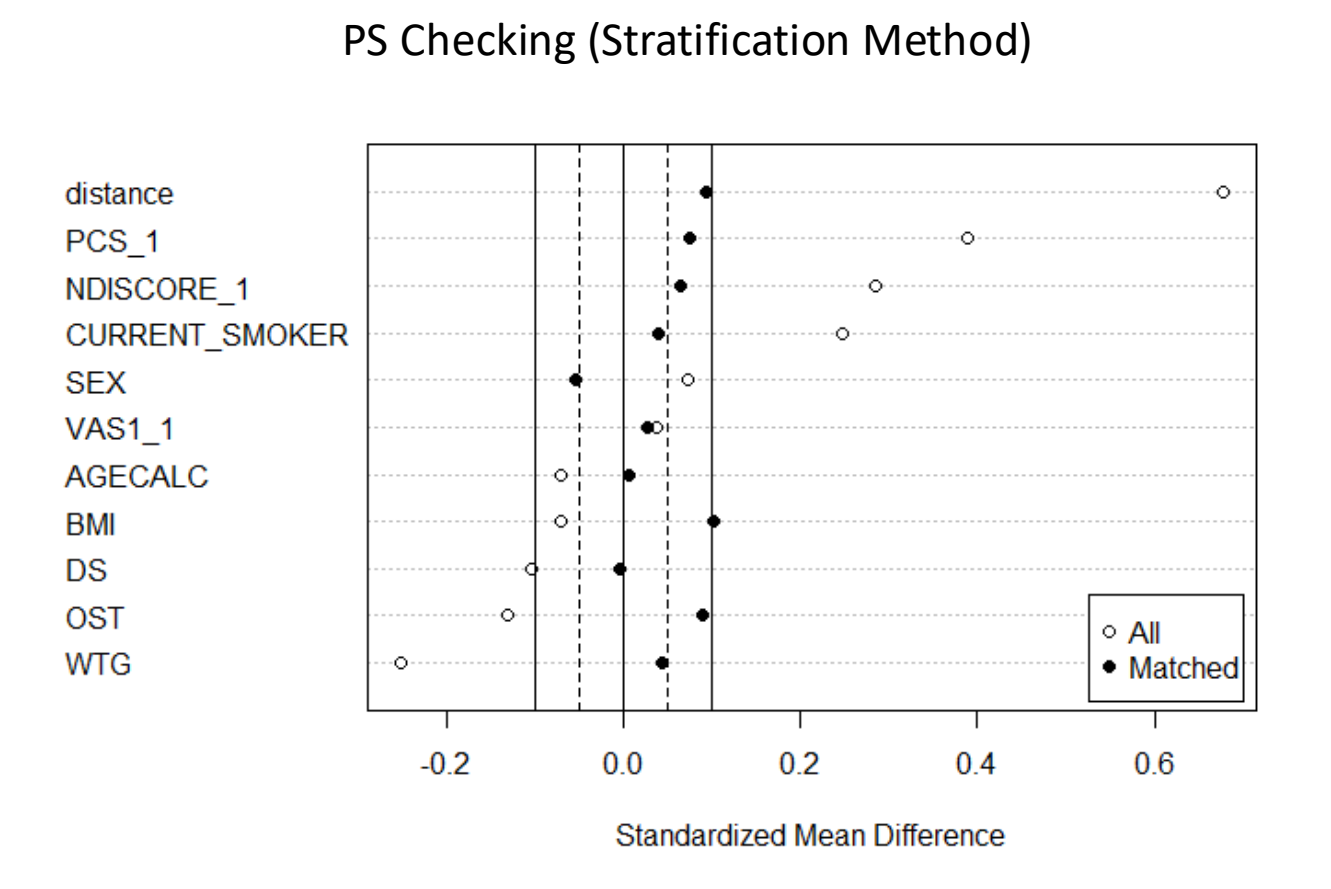
Observations: Given 20% missing,

- All produce biases but not too much
- MI slightly better than SI
- Overall, the resulted PS design is relatively reliable.

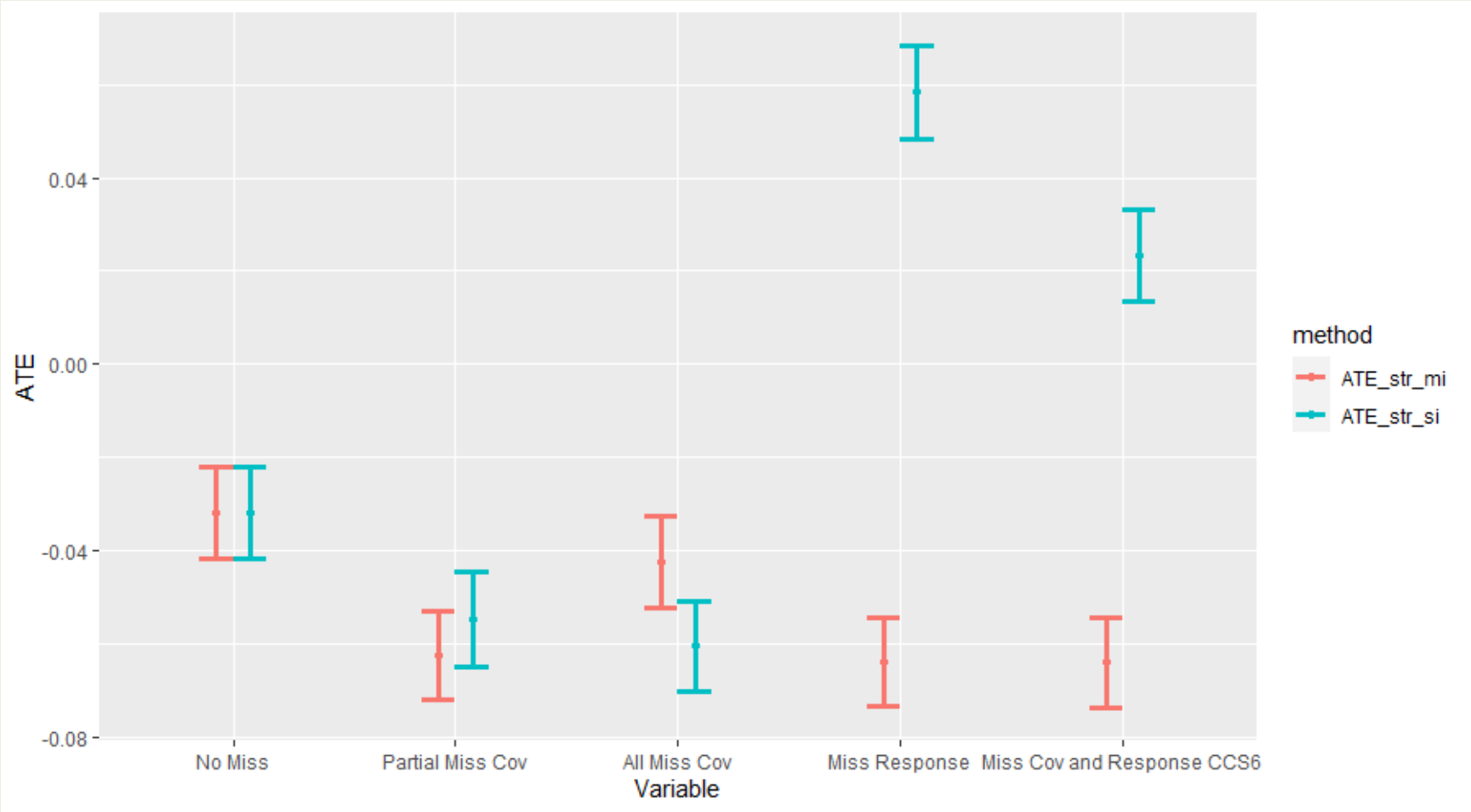
Scenario: 40% Missing

StatScience
Stats Power

Setting: 40% missing; **Estimand**: ATE; **Outcome**: proportion of success CCS6



Scenario: 40% Missing



Observations:

- Biases get larger with 40% missing, especially in response missing
- MI better than SI

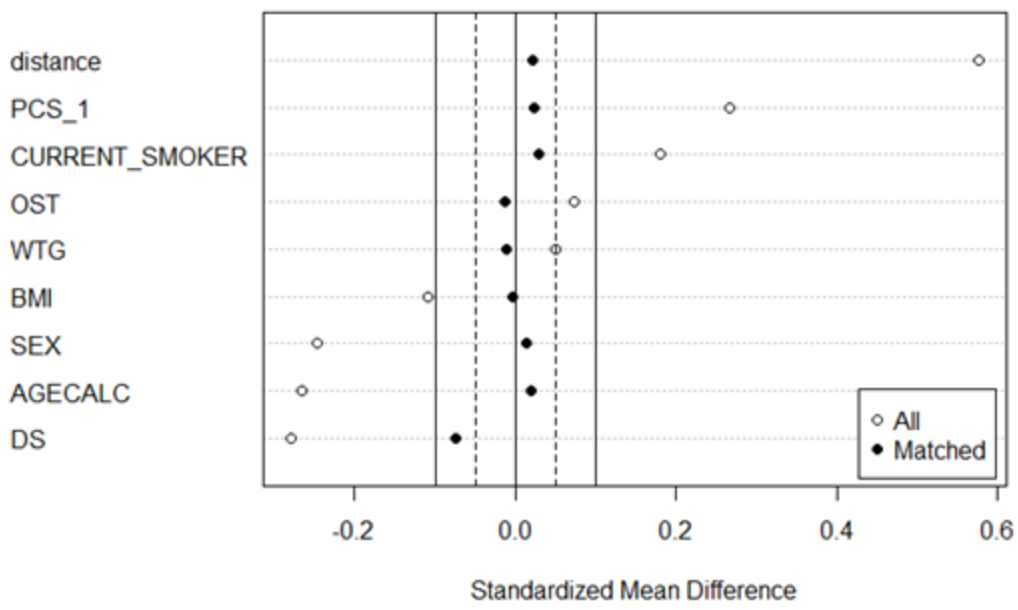
Scenario: Unobserved Covariates

StatScience
Stats Power

- **Scenario**: Covariates are often correlated to each other. For Unobserved Covariates but highly correlated with observed covariates
- In our exercise, weight and OST are highly correlated
- Simulation: one covariate unobserved, two covariates unobserved (i.e., NDISCORE and WAS)

Setting: 60 trt, 100 ctl;
Estimand: ATE;
Outcome: proportion of success CCS6

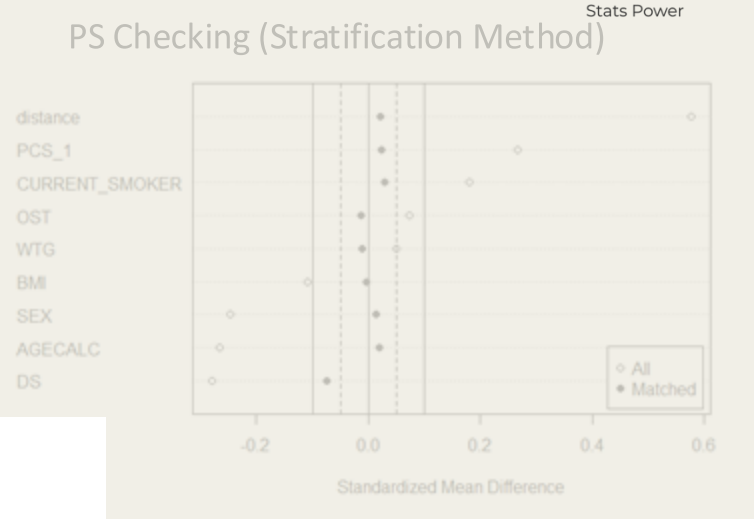
PS Checking (Stratification Method)

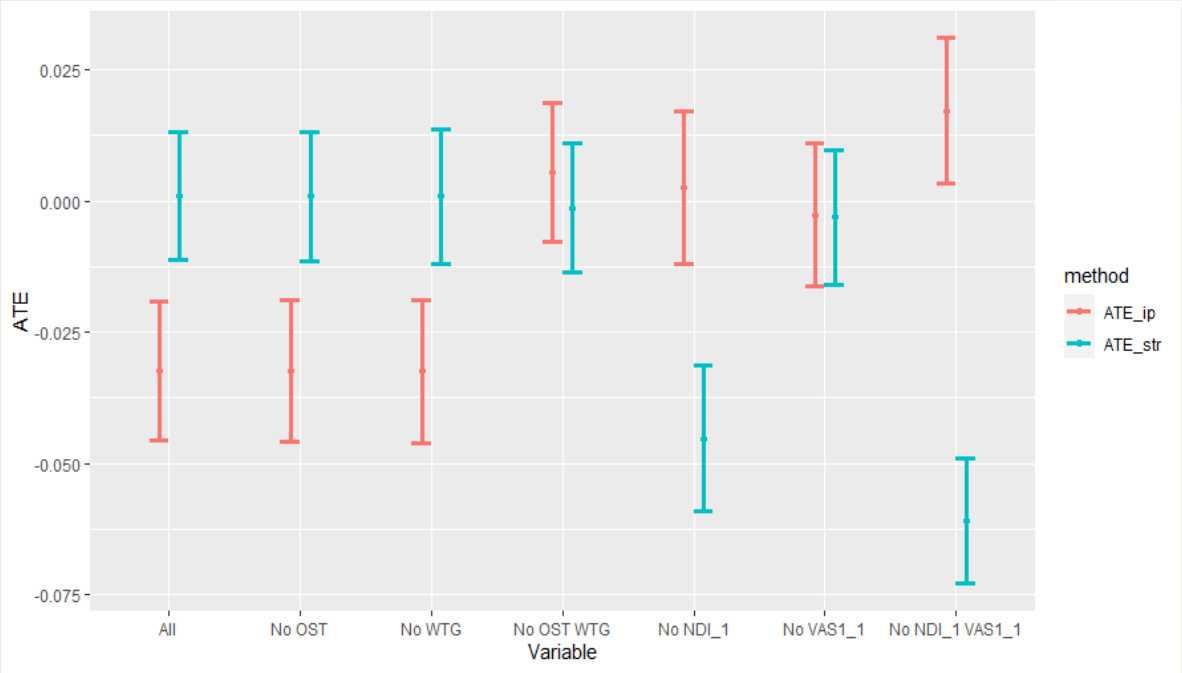


Scenario: Unobserved Covariates

StatScience
Stats Power

- **Scenario**: Covariates are often correlated to each other. For Unobserved Covariates but highly correlated with observed covariates
- In our exercise, weight and OST are highly correlated
- Simulation: one covariate unobserved, two covariates unobserved (i.e., NDISCORE and WAS)





Observations:
The resulted PS design is relatively reliable when only one covariate unobserved and such covariate is highly correlated with other observed covariates.

Take Away

- PS modeling is a powerful tool for observational studies, especially when RCT is not feasible.

Take Away

- PS modeling is a powerful tool for observational studies, especially when RCT is not feasible.
- Regarding **Sample Size**: Larger sample size makes inference more reliable

Take Away

- PS modeling is a powerful tool for observational studies, especially when RCT is not feasible.
- Regarding **Sample Size**: Larger sample size makes inference more reliable
- Regarding **Collinearity**: It is relatively safe to disregard variables in trt arm that are highly correlated to some variables in ctl arm and the PS inference remain similar

Take Away

StatScience
Stats Power

- PS modeling is a powerful tool for observational studies, especially when RCT is not feasible.
- Regarding **Sample Size**: Larger sample size makes inference more reliable
- Regarding **Collinearity**: It is relatively safe to disregard variables in trt arm that are highly correlated to some variables in ctl arm and the PS inference remain similar
- For **Missingness** Issue:
  - Categorical Missing variables (in either covariates or response) creates higher impact than Continuous Missing variables
  - Missing in response has higher impact than missing in covariates
  - Balance in PS model CANNOT yield unbiased estimate
  - Imputation could help with causal Inference:
    - For less missing variables (<=20%), simple imputation and multiple imputation are similar
    - For high missing variables (>20%), multiple method performs better than simple imputation method

Take Away

StatScience
Stats Power

- PS modeling is a powerful tool for observational studies, especially when RCT is not feasible.
- Regarding **Sample Size**: Larger sample size makes inference more reliable
- Regarding **Collinearity**: It is relatively safe to disregard variables in trt arm that are highly correlated to some variables in ctl arm and the PS inference remain similar
- For **Missingness** Issue:
  - Categorical Missing variables (in either covariates or response) creates higher impact than Continuous Missing variables
  - Missing in response has higher impact than missing in covariates
  - Balance in PS model CANNOT yield unbiased estimate
  - Imputation could help with causal Inference:
    - For less missing variables (<=20%), simple imputation and multiple imputation are similar
    - For high missing variables (>20%), multiple method performs better than simple imputation method
- Suggestion
  - Avoid Missing (especially in response)
  - Use Multiple Imputation in the existence of missing

References

StatScience
Stats Power

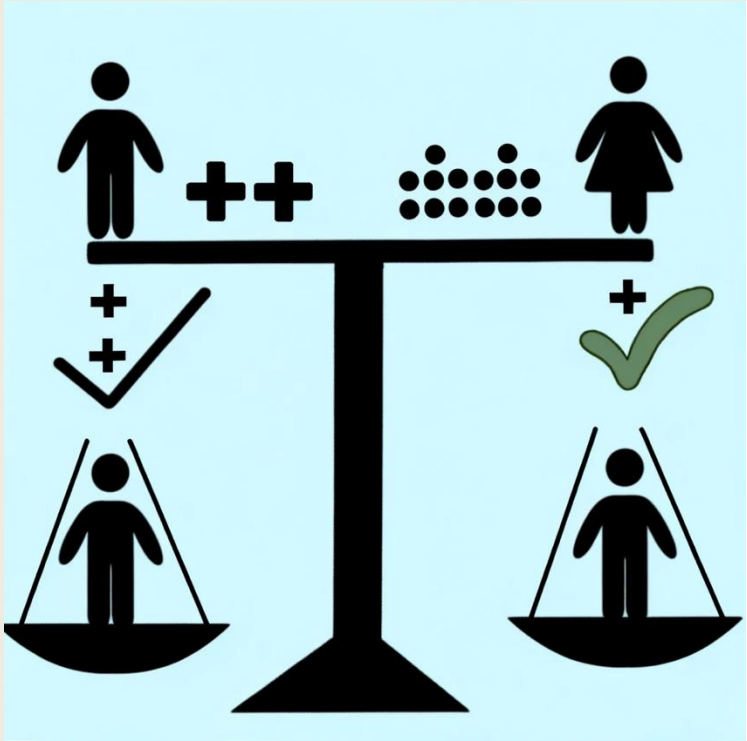
- Austin, P. (2011a), “An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies,” Multivariate Behavioral Research, 46, 399–424.
- D’Agostino, R. B., Jr., and Rubin, D. B. (2000), “Estimating and Using Propensity Scores With Partially Missing Data,” Journal of the American Statistical Association, 95, 749–759
- Li, H., Mukhi, V., Lu, N., Xu, Y., and Yue, Q. L. (2016), “A Note on Good Practice of Objective Propensity Score Design for Premarket Nonrandomized Medical Device Studies With an Example,” Statistics in Biopharmaceutical Research, 8, 282–286
- Liu, W., Kuramoto, S. J., and Stuart, E. A. (2013). “An Introduction to Sensitivity Analysis for Unobserved Confounding in Non-experimental Prevention Research.” Prevention Science 14:570–580.
- Lu, N., Xu, Y., and Yue, Q. L. (2019), “Some Considerations on Design and Analysis Plan on a Nonrandomized Comparative Study Using Propensity Score Methodology for Medical Device Premarket Evaluation”, Statistics in Biopharmaceutical Research, 12:2, 155-163.
- Yan, X., Lee, S., and Li, N. (2009), “Missing Data Handling Methods in Medical Device Clinical Trials,” Journal of Biopharmaceutical Statistics,19, 1085–1098.
- Yue, L.Q. (2012), “Regulatory Considerations in the Design of Comparative Observational Studies Using Propensity Scores,” Journal of Biopharmaceutical Statistics, 22, 1272–1279.

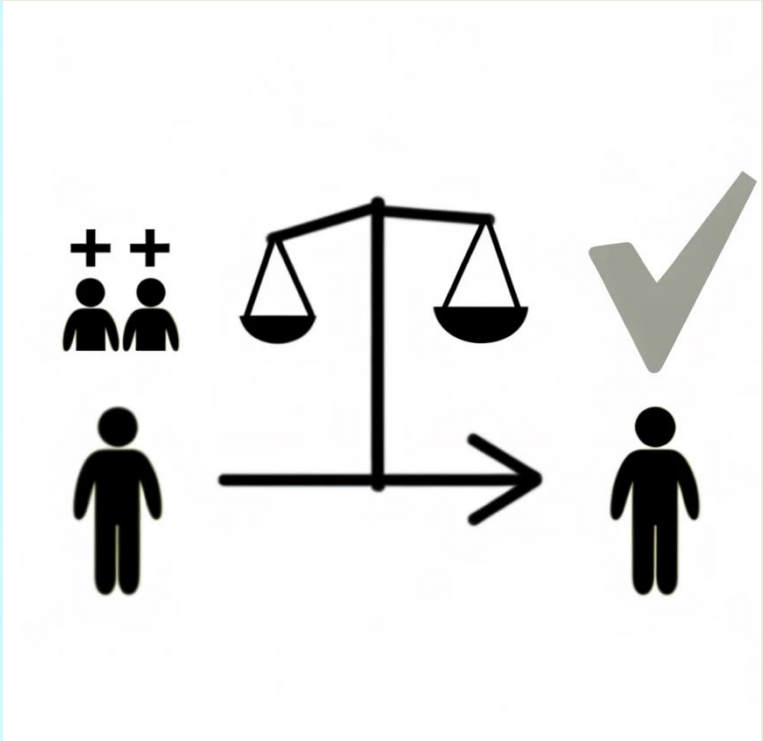
StatScience

Stats Power

Thank you ☺

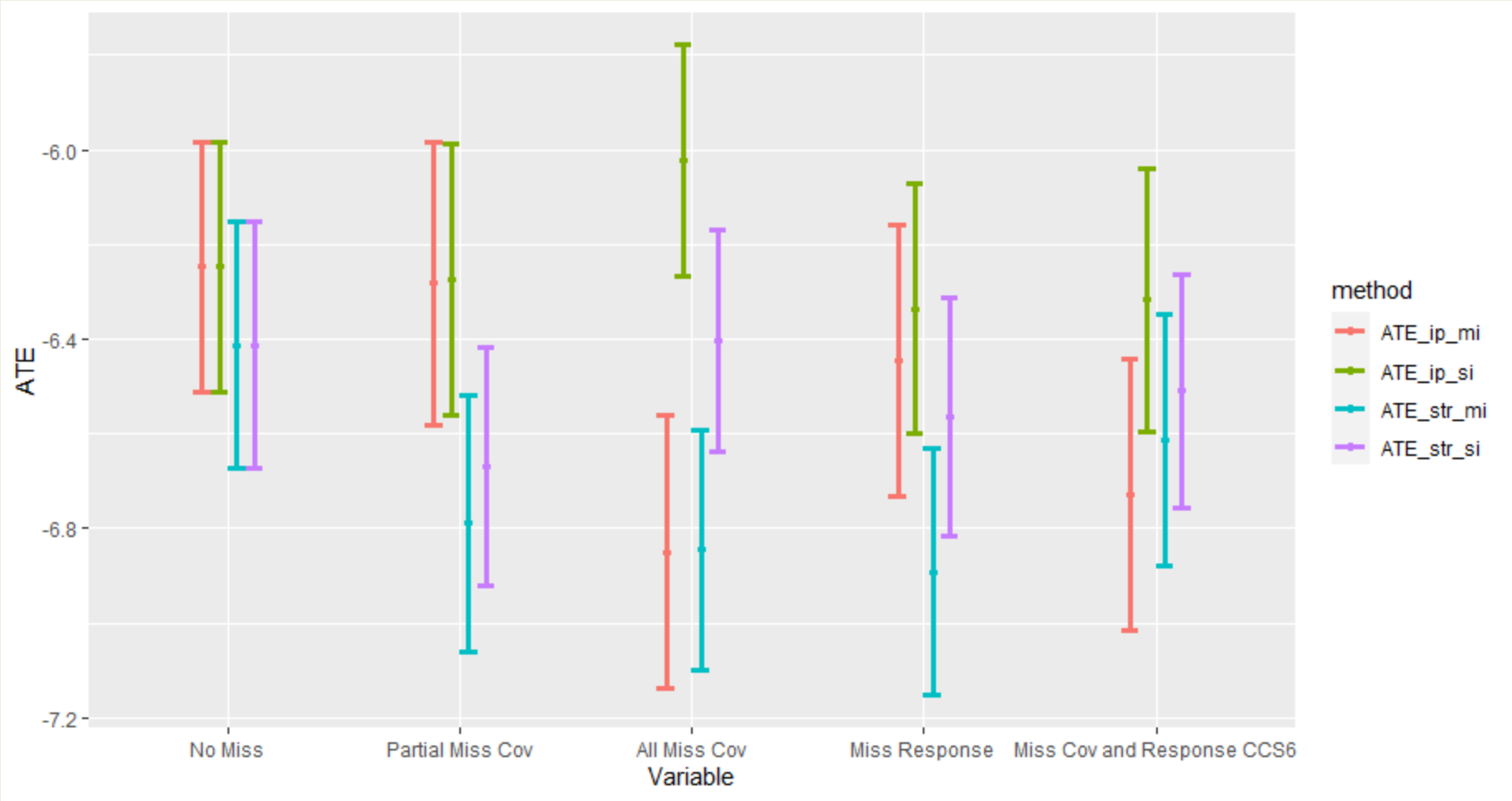
Back Up: Propensity Score Illustration





Back up: Continuous response.

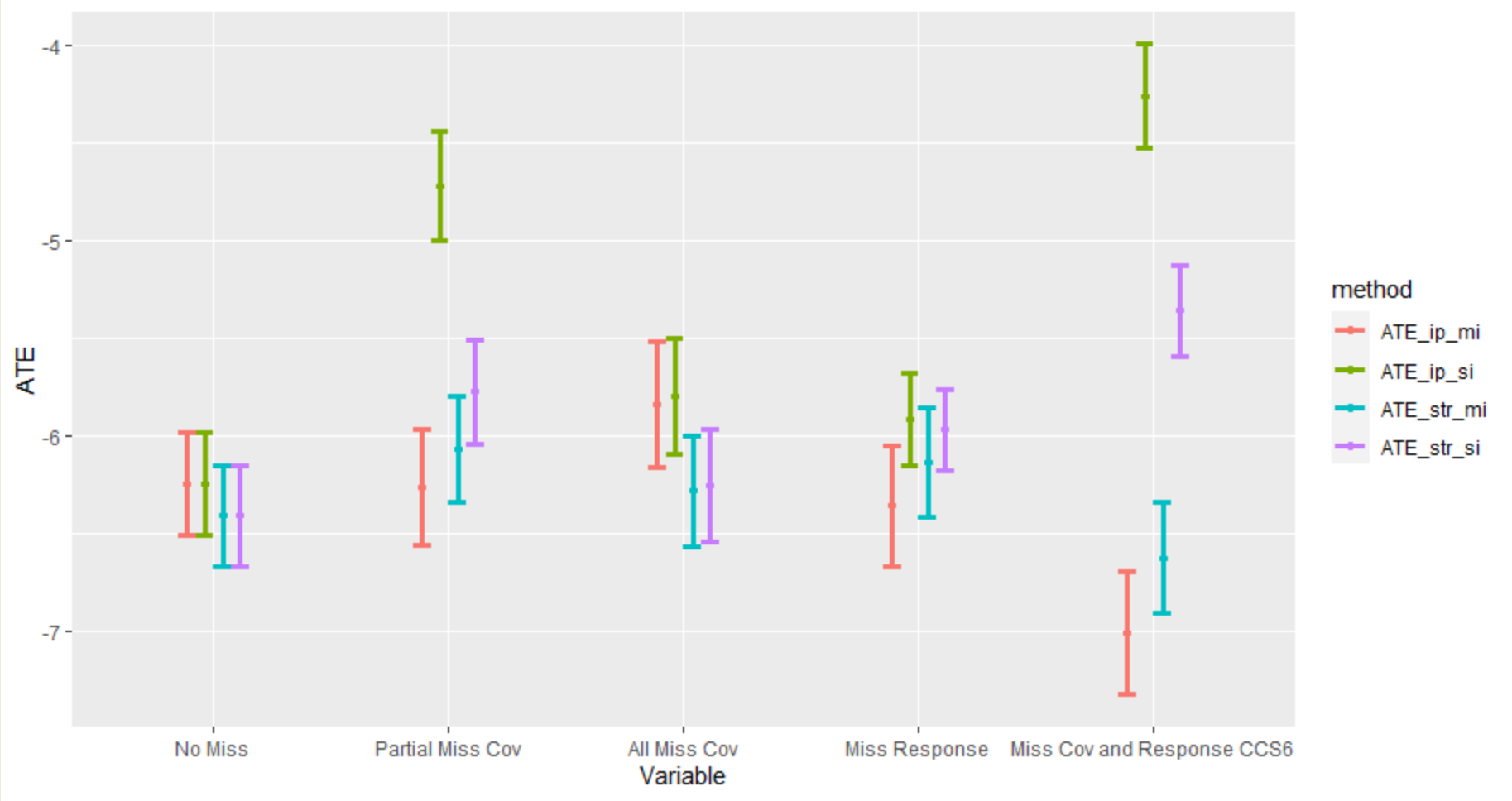
Setting: Response: NDISCORE_6, 100 trt, 150 ctls, ATE, 20% miss
Indication: MI slightly better than SI



Back up: Continuous response.

StatScience
Stats Power

Setting: Response: NDISCORE_6, 100 trt, 150 ctls, ATE, 40% miss
Indication: MI better than SI;



Missing Mechanism

StatScience
Stats Power

- **Missing Completely At Random (MCAR)**: missingness does not depend on the observed or unobserved measurements (covariates or outcomes).

- **Missing At Random (MAR)**: missingness depends only on the observed values, not on the unobserved measurements (covariates or outcomes):
  - the behavior of the post dropout observations can be predicted from the observed variables.

- **Missing Not At Random (MNAR)**: neither MCAR nor MAR, i.e., missingness depends on the unobserved measurements.

Related Literatures

StatScience
Stats Power

- Generalized Location Method with EM
  - the applied method is more statistically complicated and less commonly adopted
  - computational complicated
  - proposed under Propensity Score Matching Design

- Sensitivity Analysis methods for PS matching Design
- tipping point that negates the statistical significance of the outcome-treatment association
- derives the point estimate of the true outcome-treatment association with a 95% confidence interval