# Transporting established insights from classical experimental design to address real-life causal questions

D.B. Rubin
(joint work with M.-A. Bind)

Northwestern University

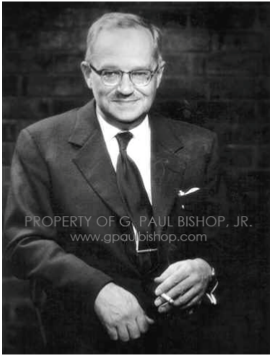
June 2017

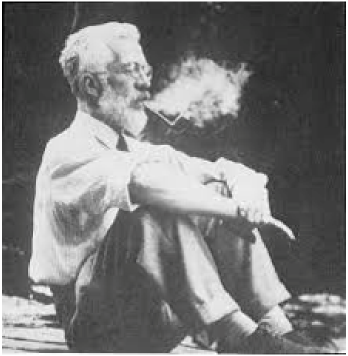
HARVARD
Faculty of Arts and Sciences
DEPARTMENT OF STATISTICS

# Once upon a time... many people smoked

# Once upon a time... many people smoked

# Once upon a time... many people smoked



# Today : how would you investigate whether parental smoking has an impact on children ?

*Plate 11. Mrs. Fisher 1938, with daughters, in order of age, Margaret (top right), Joan (bottom right), Phyllis (top left), Elizabeth (bottom left), Rose standing beside her chair, and June in her lap.*

# Does parental smoking have an impact on children's lung function ?

*Plate 11. Mrs. Fisher 1938, with daughters, in order of age, Margaret (top right), Joan (bottom right), Phyllis (top left), Elizabeth (bottom left), Rose standing beside her chair, and June in her lap.*

Parental smoking ⟶ Lower children's lung function

Goal : Quantify the impact of smoking / benefit of smoking reduction

# Four stages to address causality

- 1) A ***conceptual*** stage
- 2) A ***design*** stage
- 3) A ***statistical analysis*** stage
- 4) A ***summary*** stage

# First two stages to address causality

- 1) A ***conceptual*** stage that involves the precise formulation of the causal question (and related assumptions) using potential outcomes and described in terms of a hypothetical randomized experiment where the exposure is randomly assigned to units ; this description includes the timing of random assignment and defines the target population ; no computation is needed at this stage.
- 2) A ***design*** stage that attempts to reconstruct (or approximate) the design of a randomized experiment before any outcome data are observed (that is, with unconfounded assignment of exposure using the observed background and treatment assignment data) ; typically, heavy use of computing is needed at this stage, e.g., for multivariate matched sampling and extensive balance diagnostics.

# Last two stages to address causality

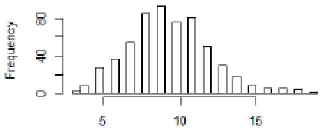
- 3) A ***statistical analysis*** stage defined in a protocol explicated before seeing any outcome data, comparing the outcomes of interest in similar (e.g., hypothetically randomly divided) exposed and non-exposed units of the hypothetical randomized experiment ; this stage is the one that most closely parallels the standard model-based analyses but uses more flexible methods.
- 4) A ***summary*** stage providing conclusions about statistical evidence for the sizes of possible causal effects of the exposure ; no computing is required at this stage, just thoughtful summarization, e.g., focusing on what actual world interventions.

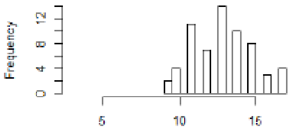
# First stage : formulation of the causal question in terms of a hypothetical randomized experiment using potential outcomes

| i | Age | Height | Sex | Parental smoking | FEV-1(0) | FEV-1(1) |
|---|---|---|---|---|---|---|
| 1 | | | | | | |
| ... | | | | | | |
| N=654 | | | | | | |

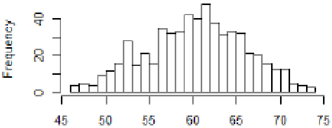
# Non-exposed vs. exposed children

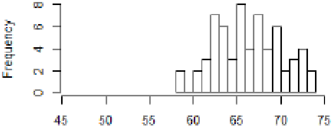
- What can you say about the distributions of ***age***?



- What can you say about the distributions of ***height*** ?



# Hypothetical experiments

- ***Hypothetical experiment A***

  A completely randomized experiment with $N_{Smoking}$=65 children with smoking parents and $N_{Non-Smoking}$=589 children with non-smoking parents.

- Suppose we intervene on smoking households before they have children and randomizing them to stop smoking with probability $\frac{9}{10}$, and thus with probability $\frac{1}{10}$ to continue to smoke.

# Hypothetical experiments

- ***Hypothetical experiment B***
  - Suppose we selected boundaries for the covariates *age* and *height*, and restricted the 361 children to fall within those boundaries.
  - For the restricted children, we completely randomized smoking households before they have children .
  - This strategy led to $N_{Smoking}$=61 children with smoking parents and $N_{Non-smoking}$=300 children with non-smoking parents.

- ***Hypothetical experiment C***
  - Randomized block experiment could have resulted in non-smoking parents with background covariates that are within certain strata defined by the background covariates of the smoking parents.
  - This formulation led to $N_{Smoking}$=57 children with smoking parents and $N_{Non-smoking}$=216 children with non-smoking parents.

# Hypothetical experiments

- Other hypothetical randomized experiments would also intervene on smoking parents before their child's conception ; we describe two such experiments.
  - First, ***Hypothetical experiment D.1***, a completely randomized experiment with balanced groups (e.g., creating two equal-sized groups of parents similar on background characteristics, that is, $N_{Smoking}$=$N_{Non-Smoking}$=63 children).
  - Or second, ***Hypothetical experiment D.2*** : a rerandomized experiment with two equal-sized groups of similar parents (with $N_{Smoking}$=$N_{Non-Smoking}$=63) for which the randomized allocations are allowed only when parents' covariates (e.g., *height*) mean differences between smokers and non-smokers are within some a priori defined calipers.

# Hypothetical experiments

- Another hypothetical randomized experiment, ***Hypothetical experiment E***, would intervene after the child's conception, from the point in time for which we know the child's gender, and would have a paired-randomized experiment where a coin flip determines which parents of a pair of two similar parents expecting a child with same gender is exposed to still-smoking parents, with $\text{N}_{Smoking} = \text{N}_{Non-Smoking} =$ 63 children).

# Second stage : Design stage that attempts to reconstruct the ideal conditions for a randomized experiment

- What type(s) of design do you know ?

# Second stage : Design stage that attempts to reconstruct the ideal conditions for a randomized experiment

| **Hypothetical experiment / Design stage methods** | **Number of children** |
|---|---|
| HYPOTHETICAL EXPERIMENT (A) / NO DESIGN (a) | 654 |
| HYPOTHETICAL EXPERIMENT (B) / TRIMMING (b) (Restriction to girls between 10 and 18 years old and height between 60 and 69 inches and to boys between 9 and 18 years and height between 58 to 72 inches) | 361 |
| HYPOTHETICAL EXPERIMENT (C) / STRATIFIED MATCHING (c) (*cem* R package) | 273 |
| HYPOTHETICAL EXPERIMENTS (D.1 and D.2) / PROPENSITY SCORE MATCHING (d) (caliper=1 standard deviation of the propensity score, *Matching* R package) | 126 |
| HYPOTHETICAL EXPERIMENT (E) / OPTIMAL PAIR MATCHING (e) (Minimum squared Mahalanobis distance, *optmatch* R package) | 126 |

# A few quotes on matching (Imbens and Rubin, 2016)

- Matching can be interpreted as reorganizing the data from an observational study in such a way that the assumptions from a randomized experiment hold, at least approximately.
- Unconfoundedness is not guaranteed (as it is in expectation for randomized experiment).
- Matching may be inexact, systematic differences in pre-exposure variables across the matched pairs may remain but can be subsequently adjusted in the analysis stage.

# One approach using propensity score matching strategy

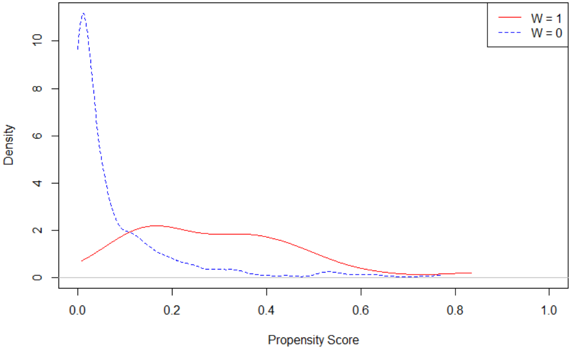
- Overall picture, compare "like with like".
- logit P(Smoking=1|Age, Height, Sex)
  $= \beta_0 + \beta_1$ Age $+ \beta_2$ Height $+ \beta_3$ Sex
- Fitted values $= \widehat{P}(Smoking = 1|Age, Height, Sex) =$ Propensity score

| i | Age | Height | Sex | Parental smoking | Propensity score |
|---|---|---|---|---|---|
| 1 | 9 | 58 | 1 | 1 | 0.01 |
| ... | ... | ... | ... | ... | ... |
| 654 | 9 | 58 | 1 | 0 | 0.01 |

- 1-1 matching with caliper on the estimated propensity score.

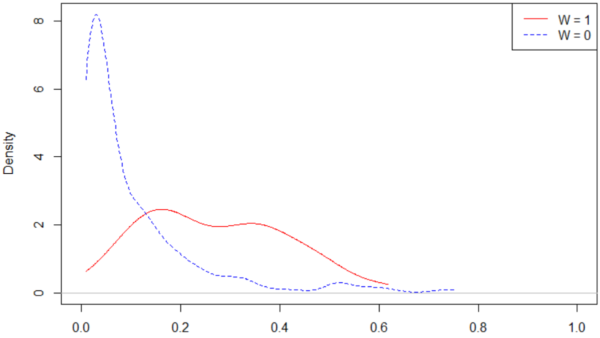
# Overlap

- 154 unexposed children were "unmatchable" to exposed children (i.e., outside of the range of the other exposed children in terms of covariates) and 2 exposed children were "unmatchable" to unexposed children.



# Overlap

- After trimming and refitting,



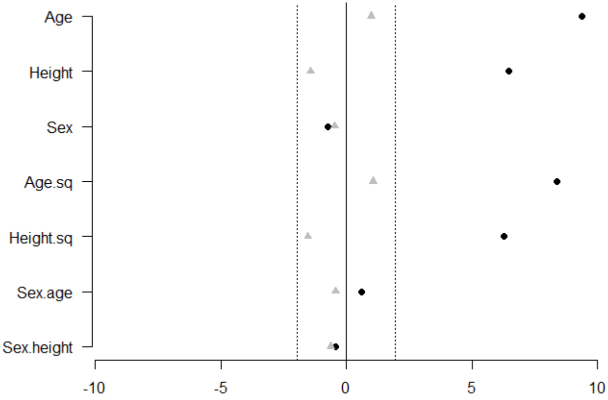
# Matched pairs based on propensity score

| Pair | Age | Height | Sex | Parental smoking | $\widehat{PS}$ |
|---|---|---|---|---|---|
| 1 | (9, 9) | (58, 58) | (1,1) | (1,0) | (0.01, 0.01) |
| ... | ... | ... | ... | ... | ... |
| 63 | (18, 16) | (70.5, 66.5) | (1,0) | (1,0) | (0.6, 0.6) |

- We ended up with N=126 children (i.e., $N_T = N_C = 63$ "similar" matched pairs).

# Diagnostics for second stage : Love plots

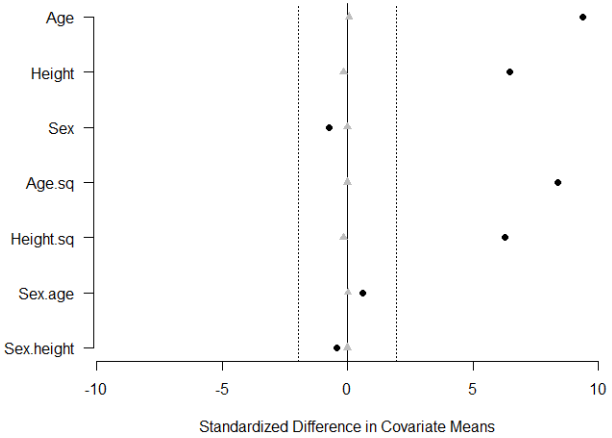
- Propensity score,




Black : before matching, grey : after matching

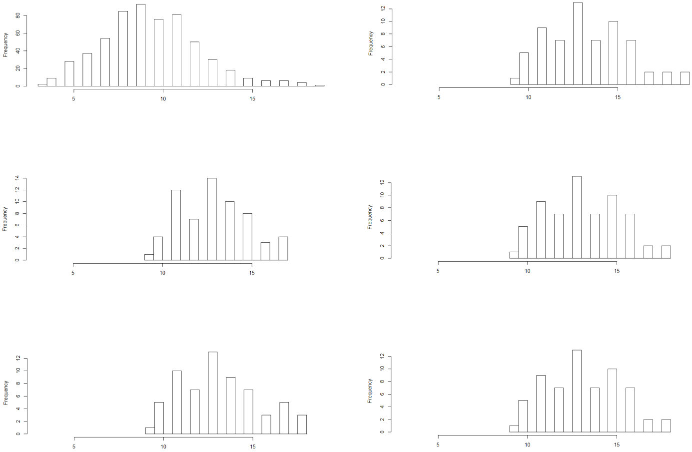
# Diagnostics for second stage : Love plots

- After optimal matching,



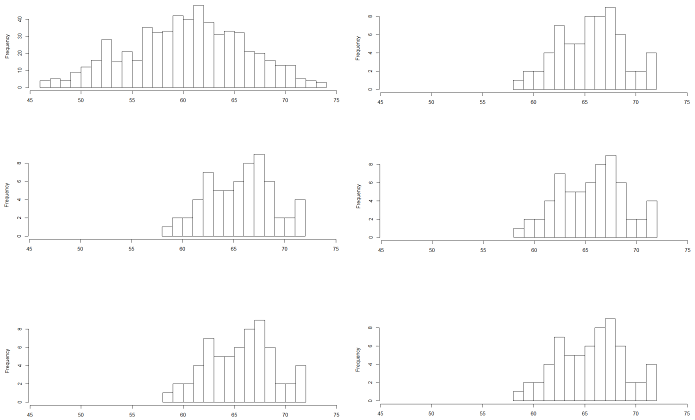

Black : before matching, grey : after matching

# Diagnostics for second stage : Age histograms in original, ps-matched, optimal paired datasets



KS p-values for : 1) before matching $= 10^{-16}$, 2) after matching, not significant

# Diagnostics for second stage : Height histograms in original, ps-matched, optimal paired datasets

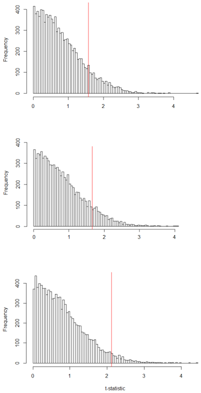



KS p-values for : 1) before matching $= 10^{-12}$, 2) after matching, not significant

# Third stage : Analysis stage that compares the outcomes of interest in the exposed versus non-exposed units of the hypothetical randomized experiment

| Hypothetical experiment / Design stage methods | Analysis method | Number of units | Estimate of the average causal effect (ACE) | 95% confidence interval |
|---|---|---|---|---|
| **HYPOTHETICAL EXPERIMENT (A) / NO DESIGN (a)** | Crude comparison | 654 | 0.71 | [0.50; 0.93] |
| | Standard linear regression with no interactions | 654 | -0.09 | [-0.20; 0.03] |
| **HYPOTHETICAL EXPERIMENT (B) / TRIMMING (b) (Restriction to girls between 10 and 18 years old and height between 60 and 69 inches and to boys between 9 and 18 years and height between 58 to 72 inches)** | Crude comparison | 361 | 0.18 | [-0.03; 0.39] |
| | Standard linear regression with no interactions | 361 | -0.16 | [-0.30; -0.03] |
| **HYPOTHETICAL EXPERIMENT (C) / STRATIFIED MATCHING (c) (*cem* R package)** | Crude comparison | 273 | -0.16 | [-0.37; 0.05] |
| | Standard linear regression with no interactions | 273 | -0.16 | [-0.30; -0.03] |
| **HYPOTHETICAL EXPERIMENTS (D.1 and D.2) / PROPENSITY SCORE MATCHING (d) (caliper=1 standard deviation of the propensity score, *Matching* R package)** | Crude comparison | 126 | -0.20 | [-0.43; 0.03] |
| | Standard linear regression with no interactions | 126 | -0.23 | [-0.46; -0.00] |
| **HYPOTHETICAL EXPERIMENT (E) / OPTIMAL PAIR MATCHING (e) (Mimimum squared Mahalanobis distance, *optmatch* R package)** | Crude comparison | 126 | -0.19 | [-0.46; 0.08] |
| | Standard linear regression with no interactions | 126 | -0.18 | [-0.35; -0.01] |

# Third stage : Randomization-based p-values in the completely randomized, rerandomized, and paired-randomized experiments




p-values=0.12 ; 0.10 ; 0.04

$$(\mathbf{Y(0)},\mathbf{Y(1)} \mid X_i, \beta_C, \beta_T, \sigma_C^2, \sigma_T^2) \sim \mathrm{N}\left( \begin{pmatrix} X_i\beta_C \\ X_i\beta_T \end{pmatrix}, \begin{pmatrix} \sigma_C^2 & 0 \\ 0 & \sigma_T^2 \end{pmatrix} \right)$$

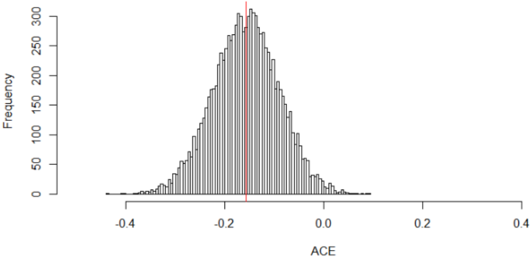
$$Y_i^{mis} \mid \mathbf{Y}^{obs}, \mathbf{W}, \mathbf{X}, \beta_C, \beta_T, \sigma_C^2, \sigma_T^2 \sim \mathrm{N}(W_i X_i \beta_C + (1-W_i) X_i \beta_T); W_i \sigma_C^2 + (1-W_i)\sigma_T^2)$$

| Age | Fev | Height | Sex | Smoking | Y0 | Y1 |
|---|---|---|---|---|---|---|
| | | | | | Potential outcomes | |
| 11 | 3.104 | 67.5 | 0 | 1 | ? | 3.104 |
| 13 | 4.045 | 69.0 | 1 | 1 | ? | 4.045 |
| 14 | 4.763 | 68.0 | 1 | 1 | ? | 4.763 |
| 10 | 2.100 | 58.0 | 1 | 0 | 2.100 | ? |
| 11 | 3.069 | 65.0 | 0 | 1 | ? | 3.069 |
| 11 | 2.785 | 69.0 | 1 | 0 | 2.785 | ? |
| 15 | 4.284 | 70.0 | 1 | 0 | 4.284 | ? |
| 15 | 4.506 | 71.0 | 1 | 1 | ? | 4.506 |
| 18 | 2.906 | 66.0 | 0 | 0 | 2.906 | ? |
| 19 | 5.102 | 72.0 | 1 | 0 | 5.102 | ? |
| 19 | 3.519 | 66.0 | 0 | 1 | ? | 3.519 |

# Third stage : Bayesian approach and ACE

$$(\mathbf{Y(0)},\mathbf{Y(1)} \mid X_i, \beta_C, \beta_T, \sigma_C^2, \sigma_T^2) \sim \mathrm{N}\left( \begin{pmatrix} X_i\beta_C \\ X_i\beta_T \end{pmatrix}, \begin{pmatrix} \sigma_C^2 & 0 \\ 0 & \sigma_T^2 \end{pmatrix} \right)$$

$$Y_i^{mis} \mid \mathbf{Y}^{obs}, \mathbf{W}, \mathbf{X}, \beta_C, \beta_T, \sigma_C^2, \sigma_T^2 \sim \mathrm{N}(W_i X_i \beta_C + (1-W_i) X_i \beta_T); W_i \sigma_C^2 + (1-W_i)\sigma_T^2)$$

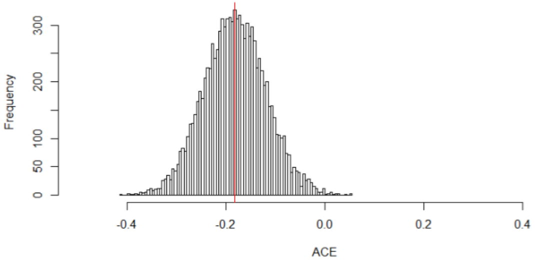



After PS matching, posterior mean= -0.16 [95% interval : -0.29 to -0.03].

# Third stage : Bayesian approach and ACE

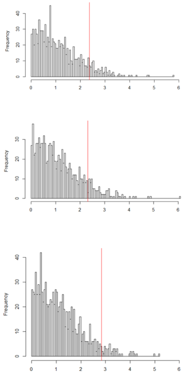
$$(\mathbf{Y}(0),\mathbf{Y}(1) \mid X_i, \beta_C, \beta_T, \sigma_C^2, \sigma_T^2) \sim \mathrm{N}\left( \begin{pmatrix} X_i\beta_C \\ X_i\beta_T \end{pmatrix}, \begin{pmatrix} \sigma_C^2 & 0 \\ 0 & \sigma_T^2 \end{pmatrix} \right)$$

$$Y_i^{mis} \mid \mathbf{Y}^{obs}, \mathbf{W}, \mathbf{X}, \beta_C, \beta_T, \sigma_C^2, \sigma_T^2 \sim \mathrm{N}(W_iX_i\beta_C + (1-W_i)X_i\beta_T); W_i\sigma_C^2 + (1-W_i)\sigma_T^2)$$




After optimal pairing, posterior mean= -0.18, [95% interval : -0.30 to -0.06].

# Third stage : Mixing Bayesian and Fisherian approaches in the completely randomized, rerandomized, and paired-randomized experiments




p-values=0.09 ; 0.10 ; 0.04

# Discussion

- Our approach complements the use of associational models by adding a design phase to the analyses that aim to address causal questions.
- Strengths and Limitations
  - Assumptions may not be plausible, but at least are transparent.
  - Covariate balance is key to address causality in hypothetical interventions.
  - Classical experimental design insights solidify statistical analyses that aim to propose policy interventions.