Advancing Evidence Generation in Biomedical Research Using Natural Hermite and Propensity Score Indices: Applications to External Control Arms

Javier Cabrera

Dept of Statistics & Dept of Medicine,
Rutgers University

Student, Collaborators: Tray Beavers, Mingshi Cui, Mahan Dastgiri, Yajie Duan, Zern Ke, Davit Sargsyan, Sofia Weigle, G Moran, D Alemayehu

Outline

- Background- Matching distributions
- Distance Measures to compare distributions
  - Differential Natural Hermite Index (DNHI)
  - Propensity Scores Index (PSI)
- Algorithms
- Applications:
  - Animal studies
  - Augmenting control groups
  - Adaptive clinical studies
  - Translating clinical studies to real world
- Concluding Remarks

Background

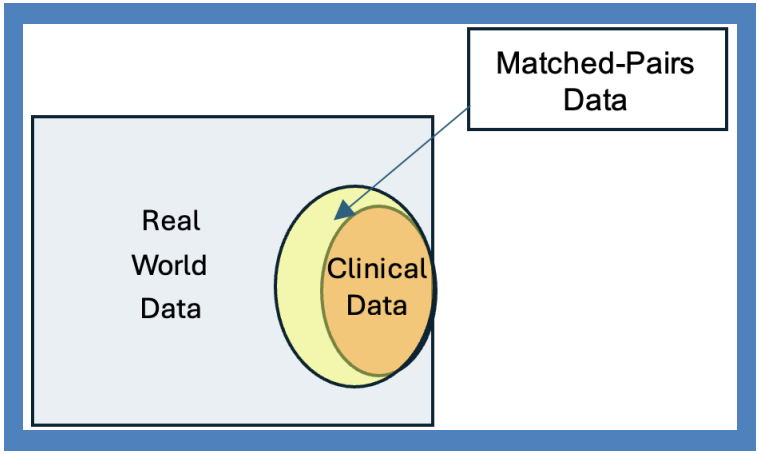
- Randomization and in particular randomized controlled trials (RCTs) considered gold standard for generating evidence
- Often, RCTs are challenging, due to operational constraints, or, when using placebo, ethical concerns.
- Large real-world datasets and historical data—such as claims data, electronic health records (EHRs), old clinical or animal studies—exist and contain historical information on the standard treatment. Use of external control arms (ECAs) are an attractive option.
- In small animal studies, poor randomization can lead to unreliable experiments.
- However, there are significant methodological and practical challenges
    - One concern: Ensuring the comparability of patients in external control and trial arms.

Matching Populations

- Matching or matched-pairs, a standard technique to estimate treatment effects
- Various approaches have been proposed in the literature
  - Bayesian borrowing methods
  - Propensity score (PS) matching techniques
- The R package *Matchit* provides alternative matching algorithms:
  - Optimal, Exact, Coarsed, …
  - Distances such as Euclidean, Propensity,…
- However, no single method has gained universal acceptance
  - Inherent limitations of each approach in guaranteeing comparability of the data sets without loss of information.

Drawbacks of standard matching approaches

- Matching often discards observations.
  - Often not all observations are matched
  - Especially true for small studies.



- When matching RCTs with RWD, there maybe usual lack of external validity, since clinical studies tend to use strict inclusion/exclusion criteria.

- For very large datasets, matching could be very expensive computationally

We study the concept of matching in distributions based on dissimilarity indices

Indices to Measure Dissimilarity

Natural Hermite Index (Cook, Buja, Cabrera, 1993)

A measure of the dissimilarity between a standardized density $f(\mathbf{y})$ and the standard normal density.

$$I^N = \int_{\mathbb{R}^d} \{f(\mathbf{y}) - \phi(\mathbf{y})\}^2 \phi(\mathbf{y}) d\mathbf{y}$$

It was first implemented as a projection pursuit index.

The idea was to measure dissimilarity between a multivariate distribution or multivariate data and the multivariate standard normal.

Differential Natural Hermite Index

- Analyzing experimental data, we often deal with differential experiments.
  - Research objective: Whether a test drug "works better than control" or another treatment.

- Differential Projection Pursuit:
  - Find projections that maximize the difference between two or more distributions or datasets.
  - These indices may be generalized to measure the dissimilarity between two distributions or datasets,

  (e.g., treatment *vs* control)

Differential Natural Hermite Index for *k P*opulations

- Let $f_1(x), \ldots, f_k(x)$ be a set of *k* density functions, and Let

$$f(x) = \frac{w_1 f_1(x) + \cdots + w_k f_k(x)}{w_1 + \cdots + w_k}$$

In many cases $w_1 = \cdots = w_k = 1$

- For every pair of densities $f_i(x). f_j(x)$ with respect to $f(y)$:

Differential Natural Hermite Index

$$d_f(f_i, f_j) = \left| \int_{\mathbb{R}^d} [f_i(x) - f_j(x)]^2 f(x) dx \right|^{\frac{1}{2}}$$

Differential Natural Hermite index for *k* populations

Proposition 1: The Differential Natural Hermite dissimilarity has the properties of a distance

(i) $d_f(f_i, f_j) \geq 0$

(ii) $d_f(f_i, f_j) = 0$ is zero when $f_i(x) = f_j(x)$ for all x, except in a set of probability zero under f.

Suppose that S be a set where $f_i(x) \neq f_j(x)$ and $\int_S f(x)dx > 0$.

Then, for all $x \in S$ $[f_i(x)-f_j(x)]^2 > 0$ and $d_f(f_i, f_j) > 0$

(iii) $d_f(f_i, f_j) = d_f(f_j, f_i)$

(iv) $d_f(f_i, f_j) \leq d_f(f_j, f_k) + d_f(f_k, f_j)$

For all x, by the triangle inequality $|f_i(x) - f_j(x)| \leq |f_j(x) - f_k(x)| + |f_k(x) - f_j(x)|$. Therefore, the proof of (iv) follows from by proof of the standard triangular inequality. Therefore d_f is a distance.

Comparing Multiple Populations

- For comparison of $k>2$ populations, we define the criterion

$$C = \sum_{i<j} d_f^2(f_i, f_j)$$

  - Would require the evaluation of $k(k\text{-}1)/2$ integrals.
- However, in Weigle, Cabrera (2023) the following was shown:

WLOG, assume $w_1 = \cdots = w_k = 1$. Given $f_1(y), \ldots, f_k(y)$ and $f(y) = \frac{f_1(y)+\cdots+f_k(y)}{k}$. Then,

$$\sum_{i<j} d_f^2(f_i, f_j) = k \sum_i d_f^2(f_i, f)$$

The proof is very similar to showing that

$$\sum_{i \neq j} (x_i - x_j)^2 = 2k \sum_i (x_i - \bar{x})^2$$

where $x_1, \ldots, x_k$, is a random sample, and $\bar{x}$, is the sample variance

Propensity Index and Other Indices

- Another simple way of defining an index is to use some function of the propensity scores (PS).
- Suppose we have the variable Treatment (Control=0, Treated=1) and a vector of covariates *X*.
- Let $h_{PS}(X)$ be the propensity score function:

$$h_{PS}(X) = P(Treatment = 1|X)$$

We introduce a Propensity Index as a function of $h_{PS}(X)$, e.g.,

$$I_{PS}(X) = Var(h_{PS}(X))$$

  - NB: $I_{PS}(X)$ is zero when the propensity score function is constant and hence the two populations are identical.
- $h_{PS}(X)$ is often estimated using a Super-Learner or in simple cases logistic regression
  - Other similar indices could be derived from LDA, SVM, deep learning,…
  - It is related to the Clever Covariate concept of TMLE.

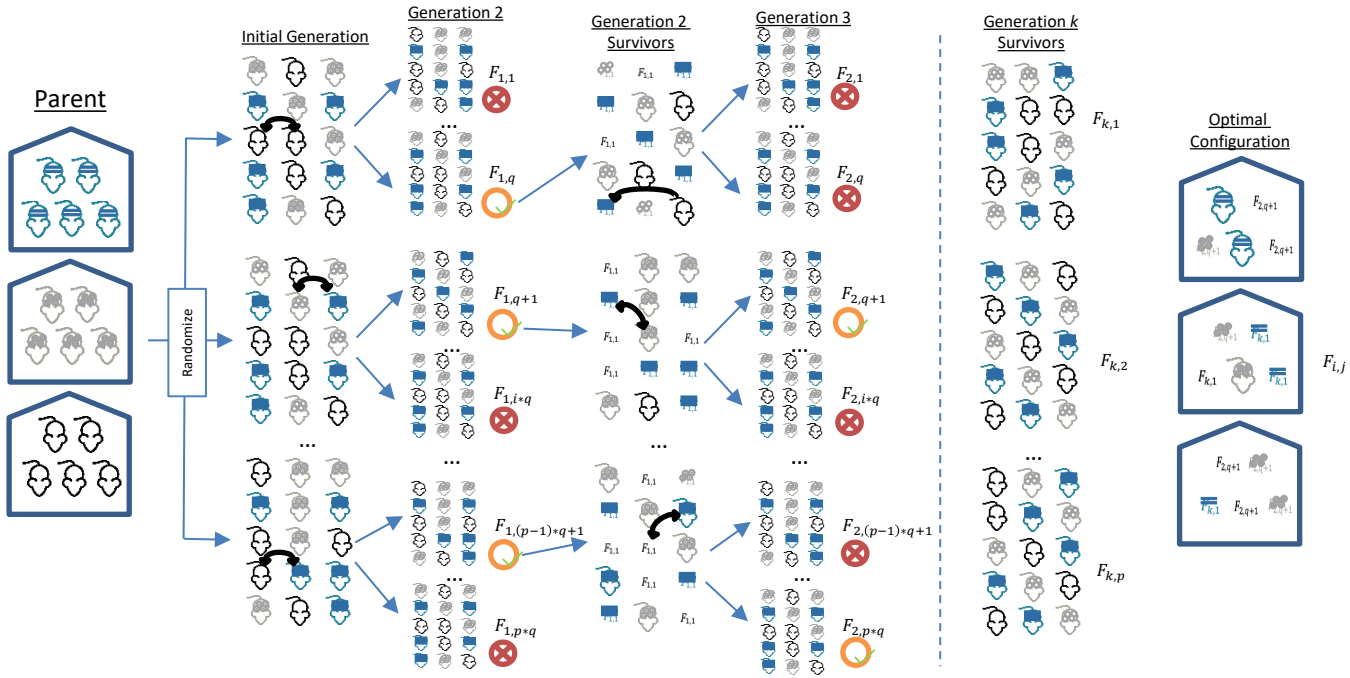
Applications of GA, DNHI and PSI
I. Randomization of Animal Studies

Objective: *N* Animals to be randomized into *k* groups optimizing a criteria.

Algorithms: Exhaustive search, Genetic Alg.
Fitness function: Irini (Ave CV^{-1}), Hermite distance

Genetic Algorithm:

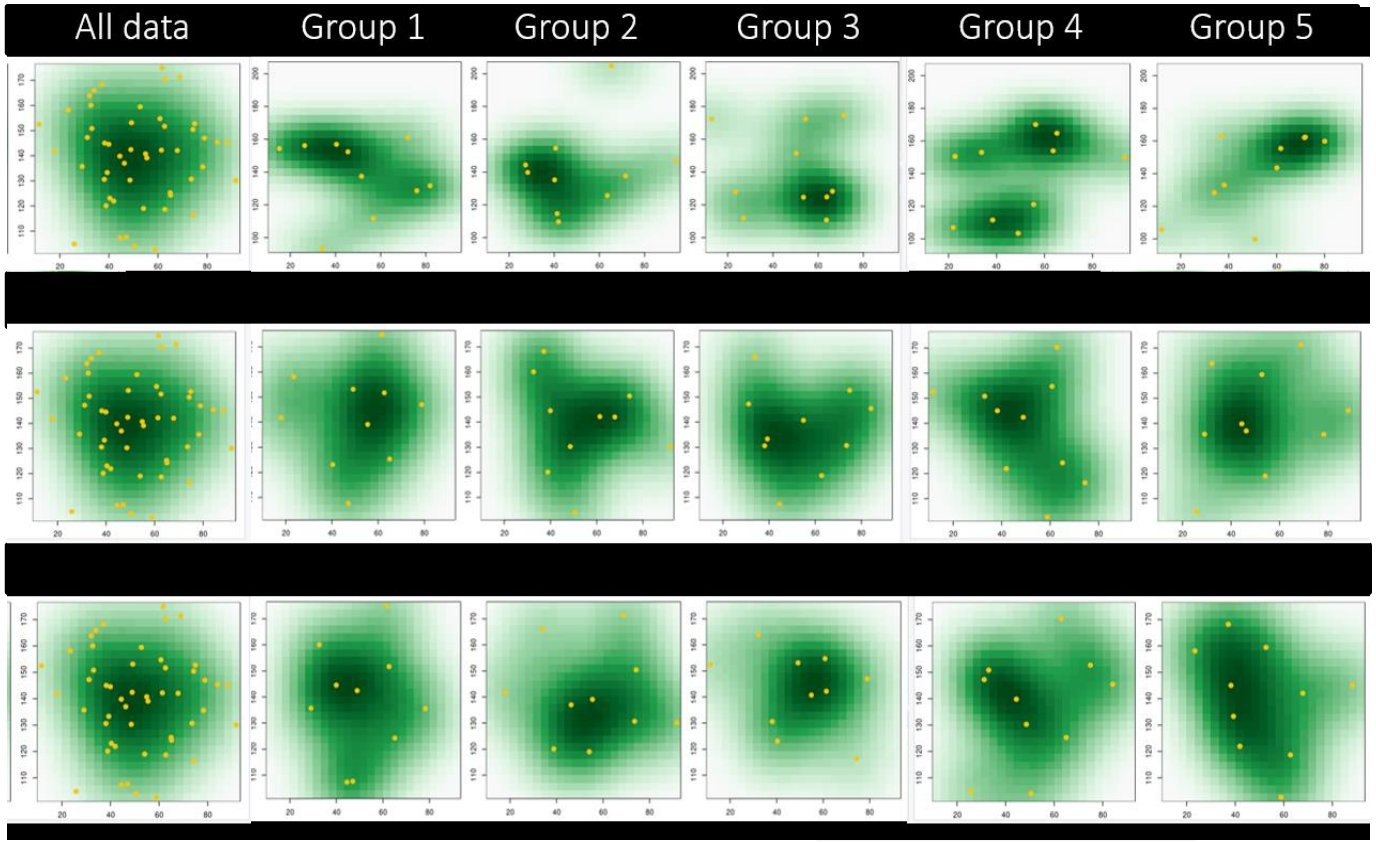
Multiple generation + Survival of the fittest



Randomization in Pre-Clinical Studies: When Evolution Theory Meets Statistics Pharmaceutical Statistics 24. S Weigle,e.a2025

Algorithm for sequential randomization to rebalance clinical studies

Example of results with Bivariate Data



Example: Simple Simulation

Dataset: 20 mice with two covariates Cov_1 and Cov_2, to be randomized into two groups of mice of size 10 each. Cov_1 and $Cov_2 \sim N(0,1)$

Method 1. Random allocation

Method 2. Genetic algorithm optimizing the Natural Hermite index

Model for generating *response:*

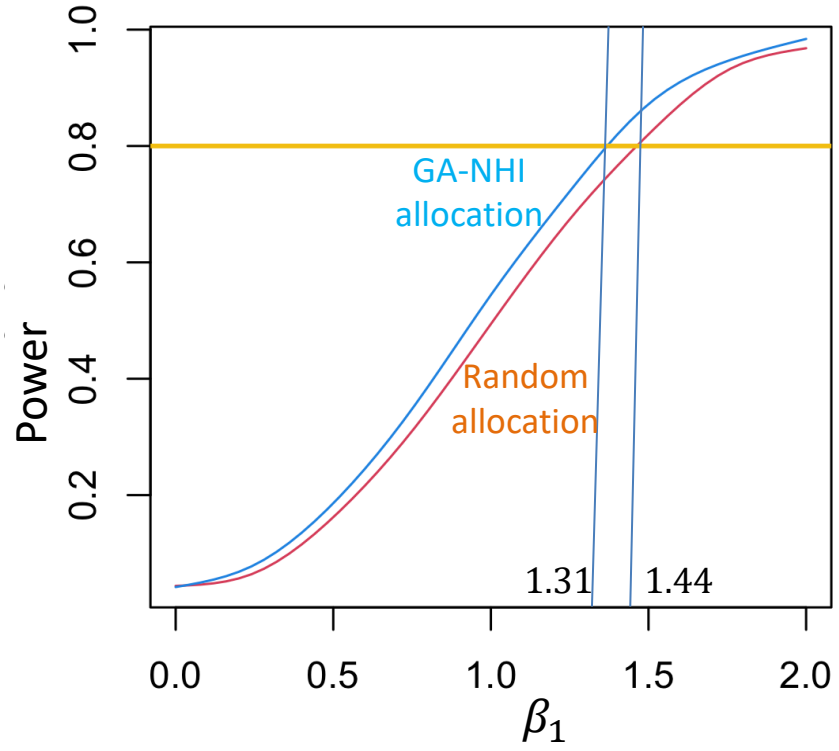
$$Y = \beta_1 Treatment + \beta_2 Cov_1 + \beta_3 Cov_2 + \varepsilon$$

Where

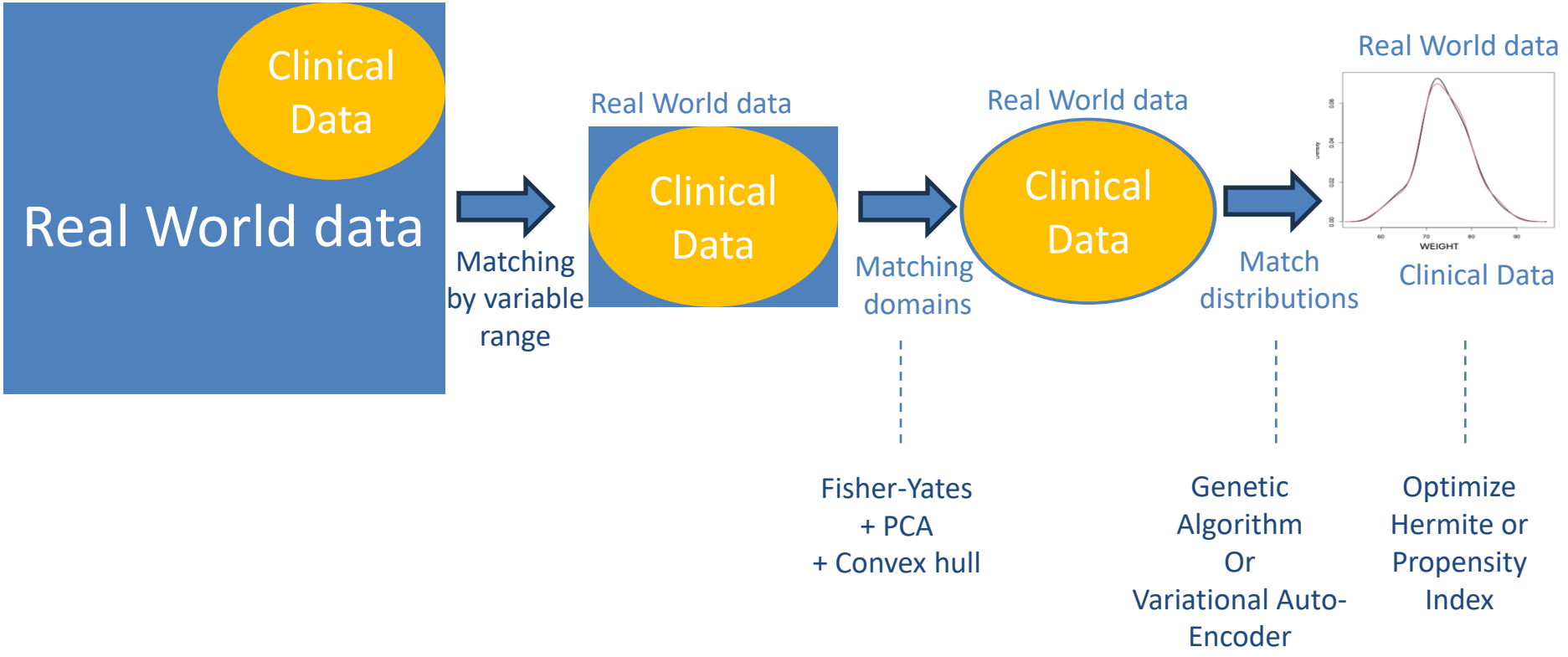
$\beta_1 = 0,0.25,0.5,0.75,1,1.25,1.5,1.75,2$

$\beta_2 = 1, \beta_3 = -1 \; and \; \sigma_\varepsilon = 1$

No of Simulations = 500 per value of β_1



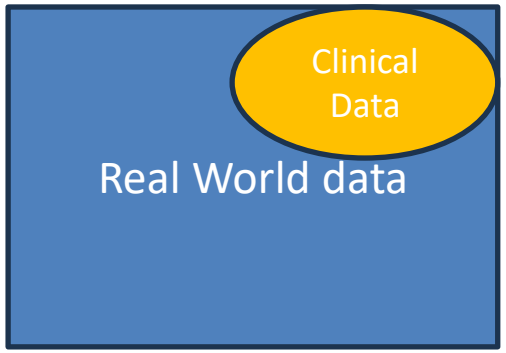
Matching Distributions
II. Augmenting clinical data with Real World



Advancing Evidence Generation in Biomedical Research Using Natural Hermite and Propensity Score Indices: Applications to External Control Arms. Cabrera Alemahehu & Weigle (submitted)

Preprocessing Real-World Data

- We want to augment RCT with real-world controls
- Instead of matched pairs , match distributions by minimizing the PS index or NH index.
- The RWD is trimmed to a subset with the same domain as the clinical data, using the following algorithm :



(i) Exclude real-world observation outside the range of any of the the clinical variables.

(ii) Apply Fisher-Yates transformation to the variables.

(iii) Do a PCA and extract the first p ($p \leq 7$) principal components.

(iv) Apply again Fisher-Yates to the principal components.

Steps (i)-(iii) can be expressed as a transformation T: $\mathbb{R}^p \rightarrow \mathbb{R}^d$, where $z=T(x)$ is approximately normally distributed.

(iv) Next the transformation $T(x)$ obtained from steps 1-3 is applied to the real-world data $z^*= T(x)$.

(v) Compute the convex hull C_H of the transformed clinical study dataset and use it to discard the transformed real-world data that follows outside of C_H

Preprocessing RWD (cont.)

- Fisher-Yates: $y_i = FY(x_i) = \Phi^{-1}(\frac{Rank(x_i)}{n+1})$, where $\Phi(\mathrm{x})$ is the standard normal cdf.
- Suppose *X* a binary variable with 1% of 1's 99%0's *Y*=*X*/sd(X)

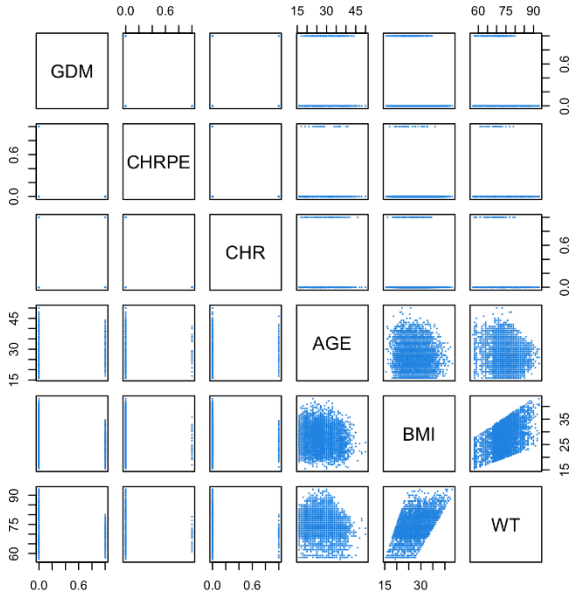
X=0	*X*=1	sd(*X*)	*Y*=*X*/sd(*X*)	*Z*=FY(*X*)	SD(*Z*)
99%	1%	0.0995	(0, 10.05)	(-0.031,2.23)	0.353
98%	2%	0.14	(0, 7.14)	(-0.038,2.16)	0.375
97%	3%	0.17	(0, 5.86)	(-0.044,2.09)	0.394
96%	4%	0.19	(0, 5.10)	(-0.050,2.04)	0.410

  - *X* appear to minimize the contribution of binary variables.
  - If transform *X* into *Y* some binary variables will create outliers or leverage points in the modeling.
  - *Z* seems acceptable as it does not create leverage points and does not minimize the contribution of binary variables.

A Real-World Example

A study on preventing placenta abruption in women in NJ (PACER,2023) . Want to use real world controls from pregnancy database of New Jersey hospital births. We have 18 variables (25 features) in common between the clinical study and the real-world database.

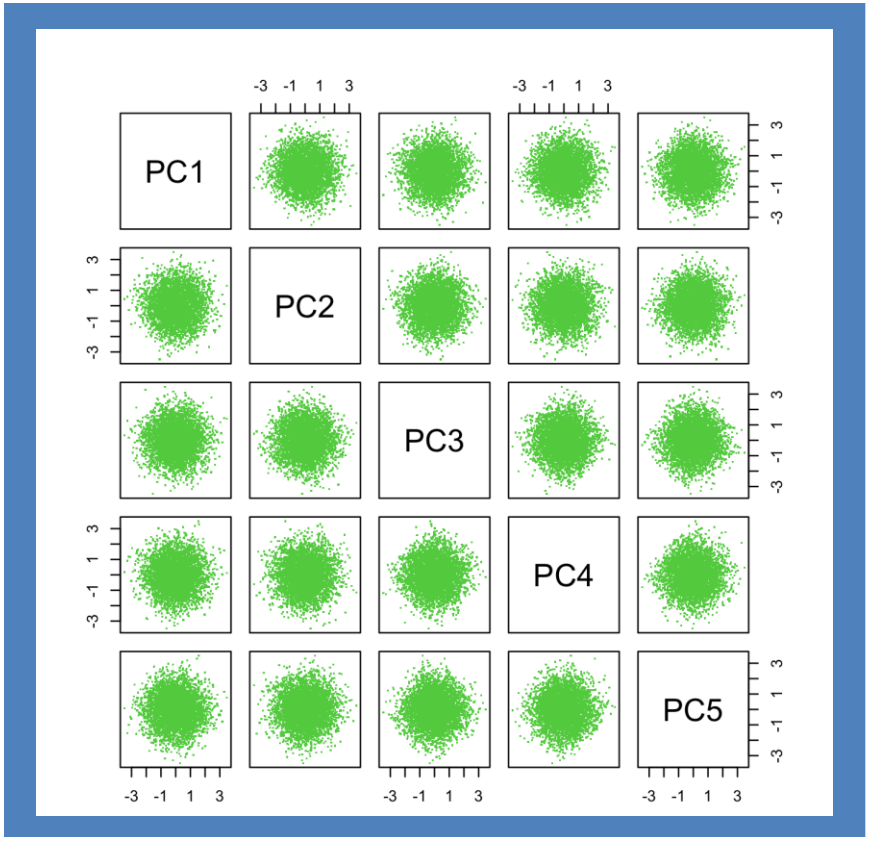
Scatter matrix of 6 variables of the 18 in data set.



Variables

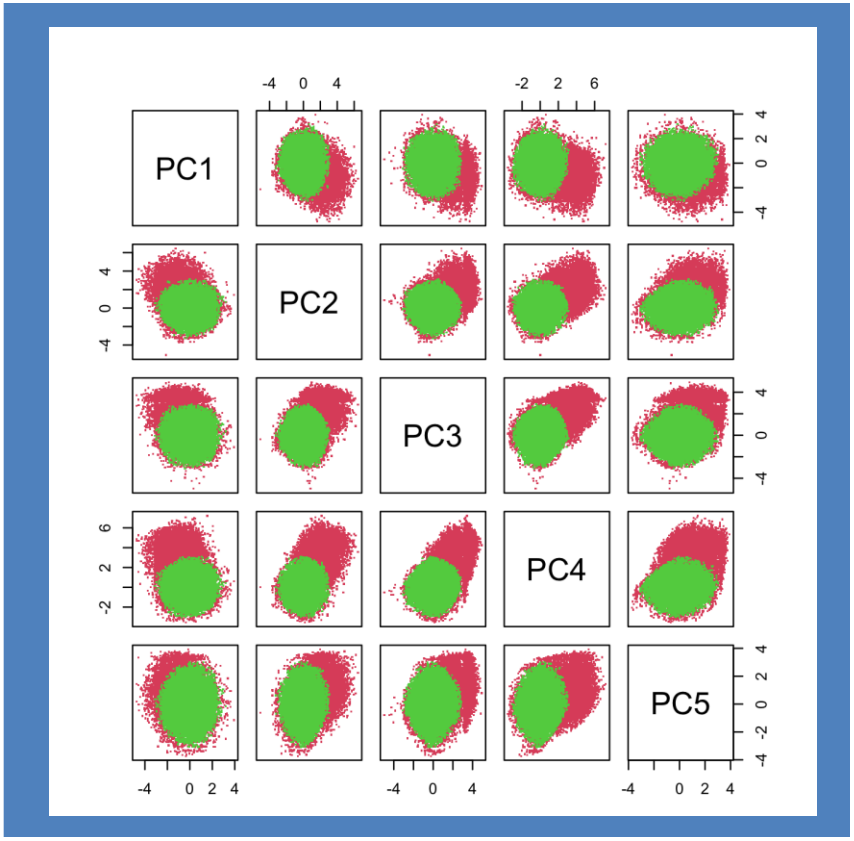
1. MONTH
2. PE_MILD
3. PE_SEVERE
4. REGION
5. RACE
6. PRE_DM
7. OLIGO
8. MARITAL
9. MULTIPLE
10. HOSPBEDR
11. HOSPOWN
12. GES_HYP
13. GDM
14. CHRPE
15. CHR
16. AGE
17. BMI
18. WT

Principal Components after Fisher-Yates Transformation



5 principal components after Fisher-Yates transformation

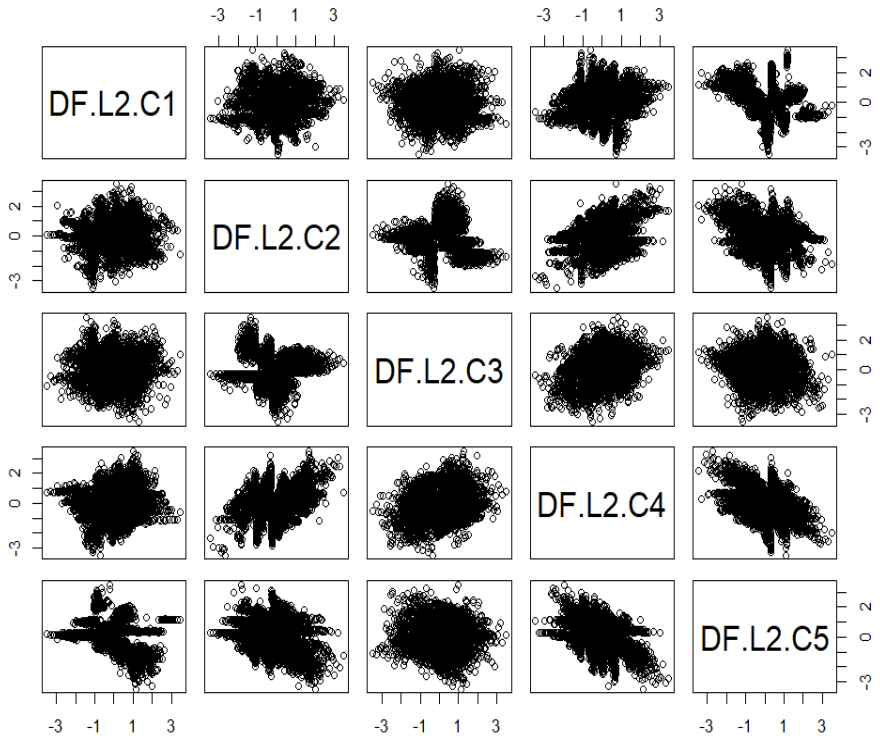
103K => 43K => 23K



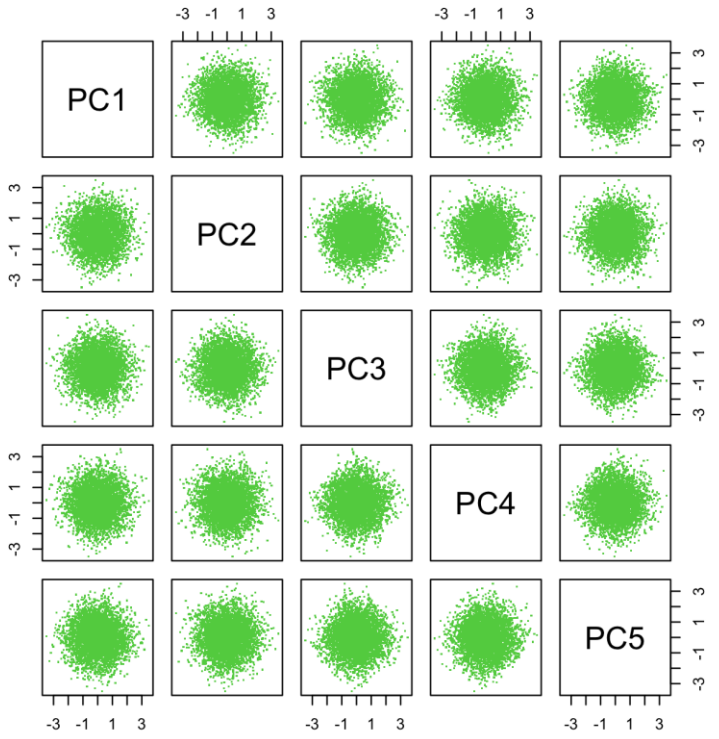
Same transformation applied to Real-World dataset.

In red is the subset or the RW outside the clinical study domain

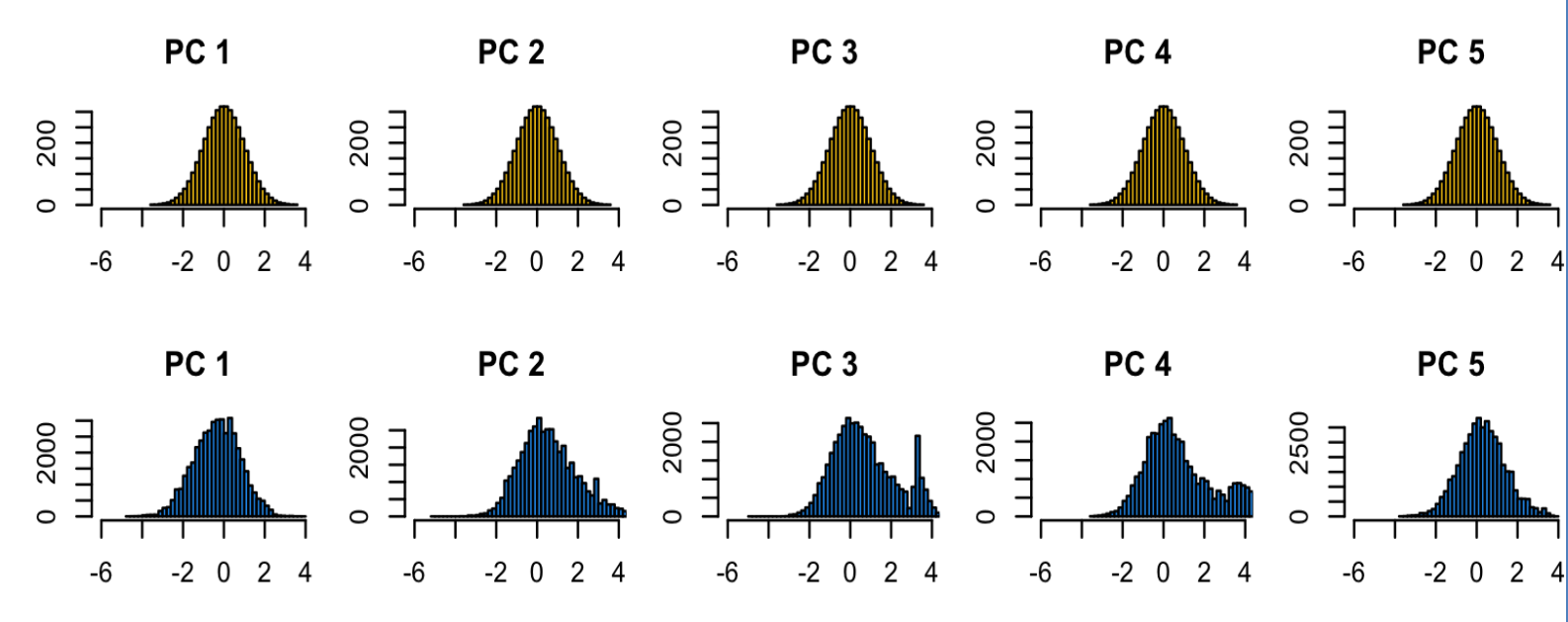
Autoencoder vs Principal Components/Fisher-Yates Transformation



Using Autoencoder

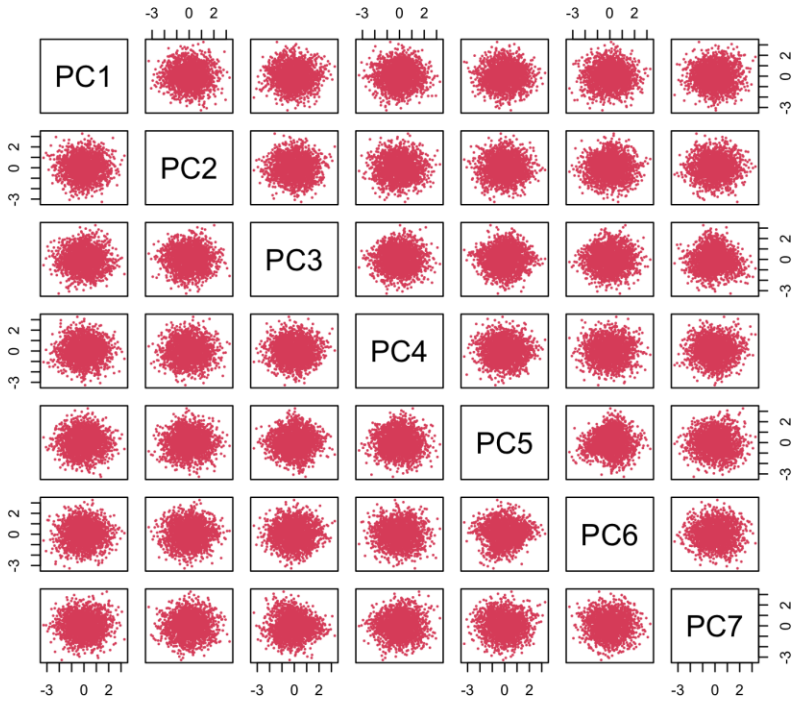


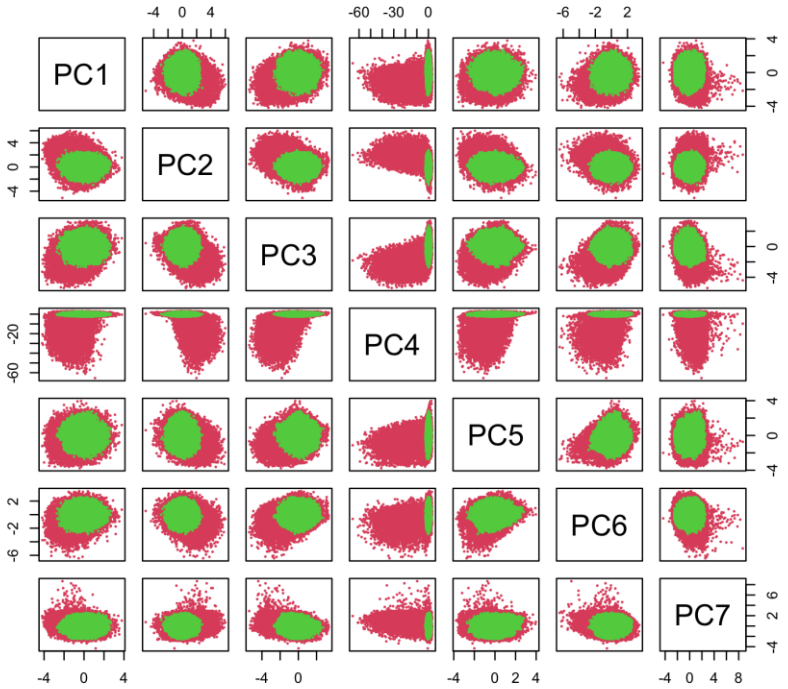
Using Fisher-Yates/ PCA



Histograms of transformed variables $T(x)$ generated by the dimension reduction algorithm using clinical and real-world data.

Principal Components after Fisher-Yates Transformation





7 principal components after Fisher-Yates transformation

103K => 43K => 19K

Same transformation applied to Real-World dataset.

In red is the subset or the RW outside the clinical study domain

A Genetic Algorithm for Augmenting Control Groups from RWD

- Assume a trial comparing a test drug (A) vs a control (B), with n_1 and n_2 subjects assigned to A and B, respectively ($n_1 >> n_2$)
- Suppose we have a RWD (*R*) set of size $M >> n_1$, and we wish to augment n_2 by drawing *m* observations from *R*.
- Let *Y* denote the outcome variable of interest, and *X* a d-dimensional vector of covariates.

- **Goal**: Find *m* observations from R that, combined with the n_2 controls in B, are 'similar' to the n_1 observations in A, the treatment group, according to some optimality criterion.

Genetic Algorithm (Cont.)

- In GA, the Index optimized by mutations forming new generations followed by natural selection.
- We need to define what corresponds to a mutation in our context.
- The simplest way is to switch a pair between R and B.
- Super-Learner variability is an issue with stopping rule.
  - Simple stopping rule: k generations in a row where the top configuration is the same and the index value was less than some threshold.

A Genetic Algorithm (cont.)

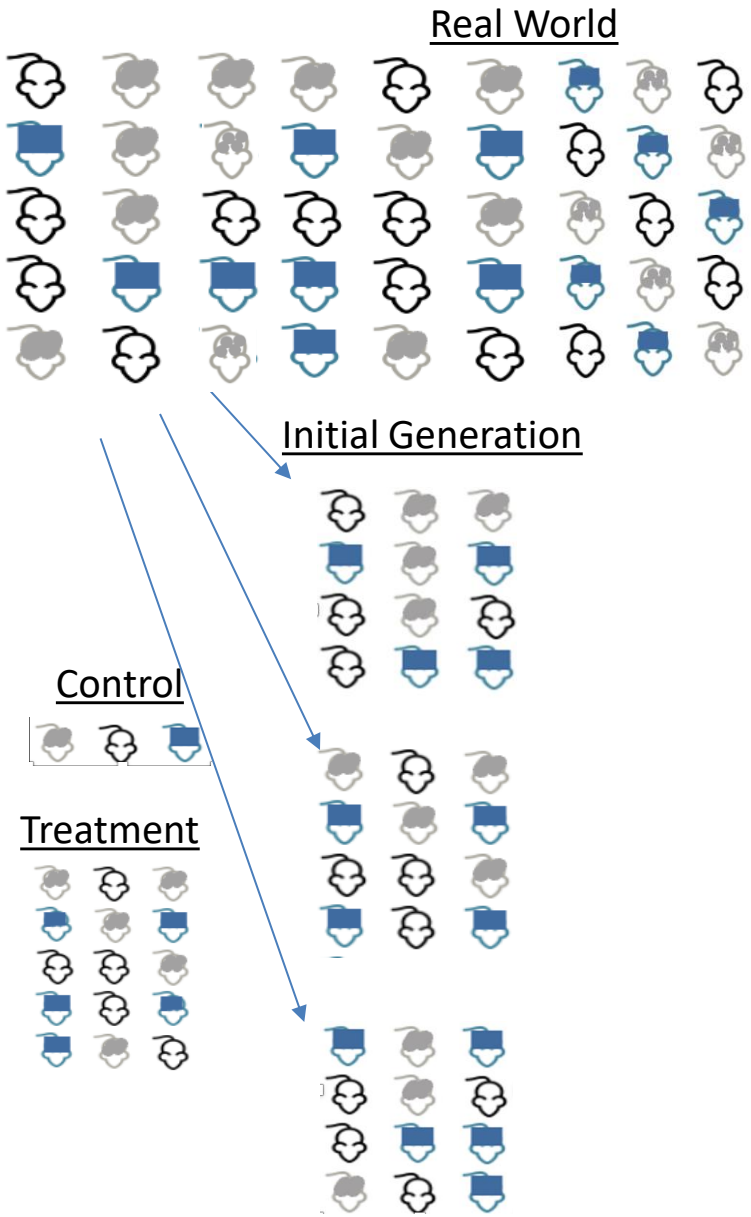
Real World

Genetic Alg. Fitness Criteria:
Propensity Index, Nat. Hermite Index, Ave CV^{-1},

Control

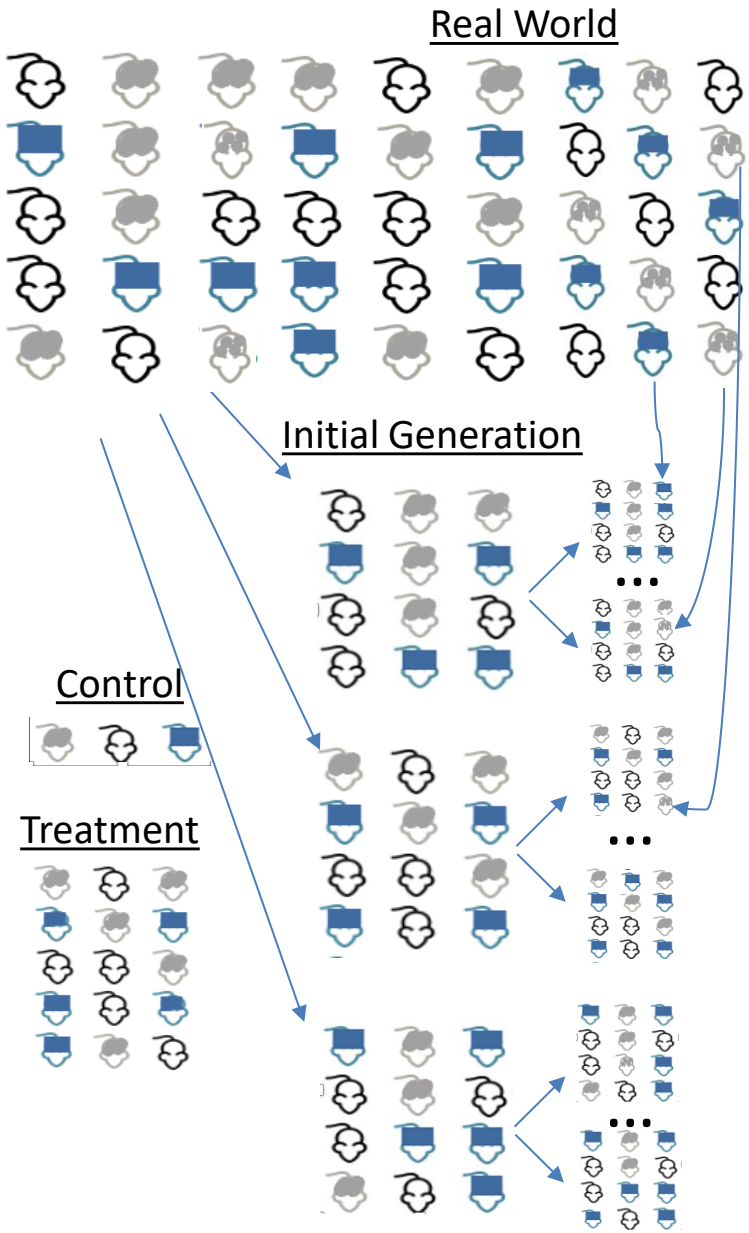
Treatment

A Genetic Algorithm (cont.)



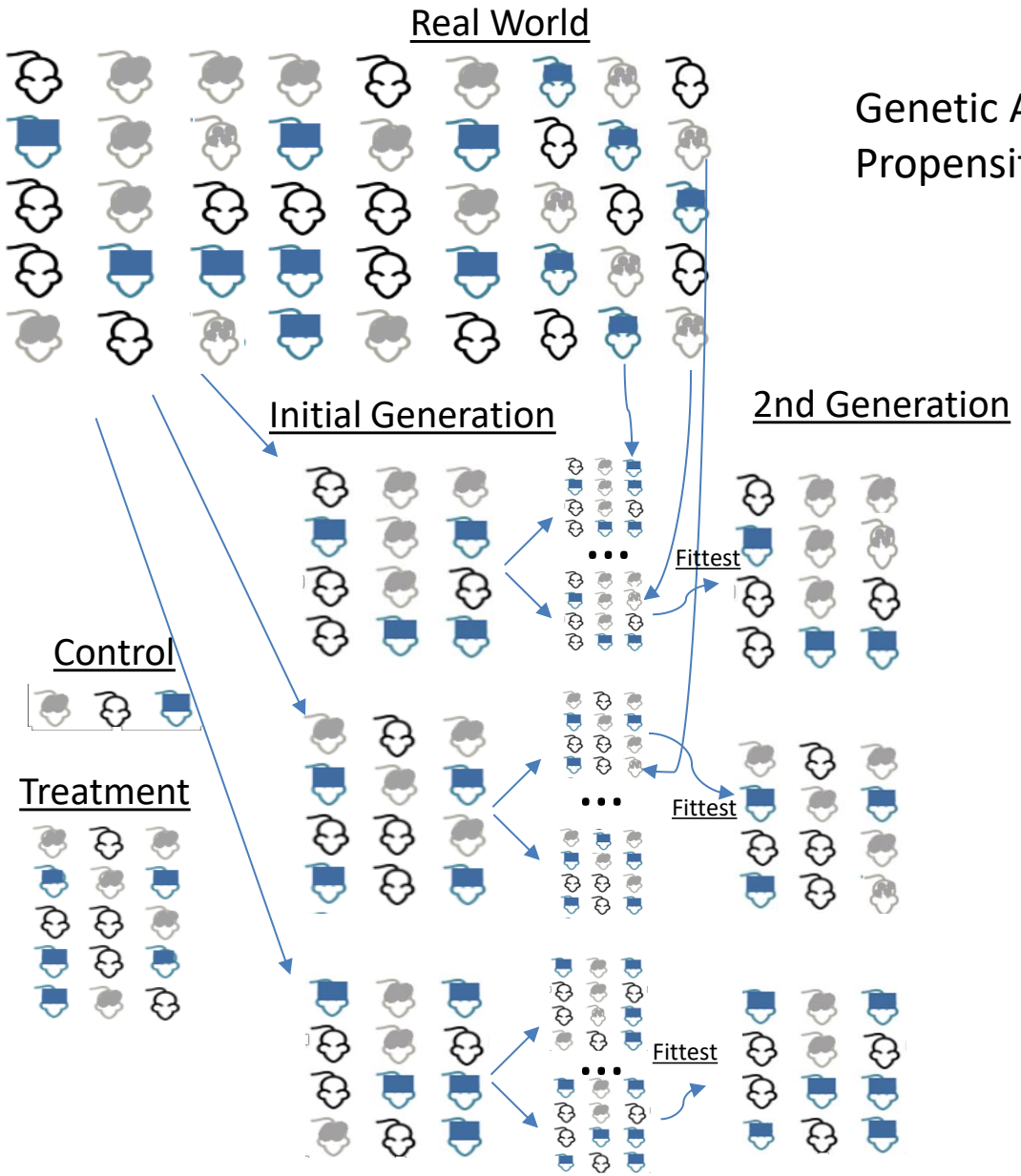
Genetic Alg. Fitness Criteria:
Propensity Indix, Nat. Hermite Index, Ave CV^{-1},

A Genetic Algorithm (cont.)



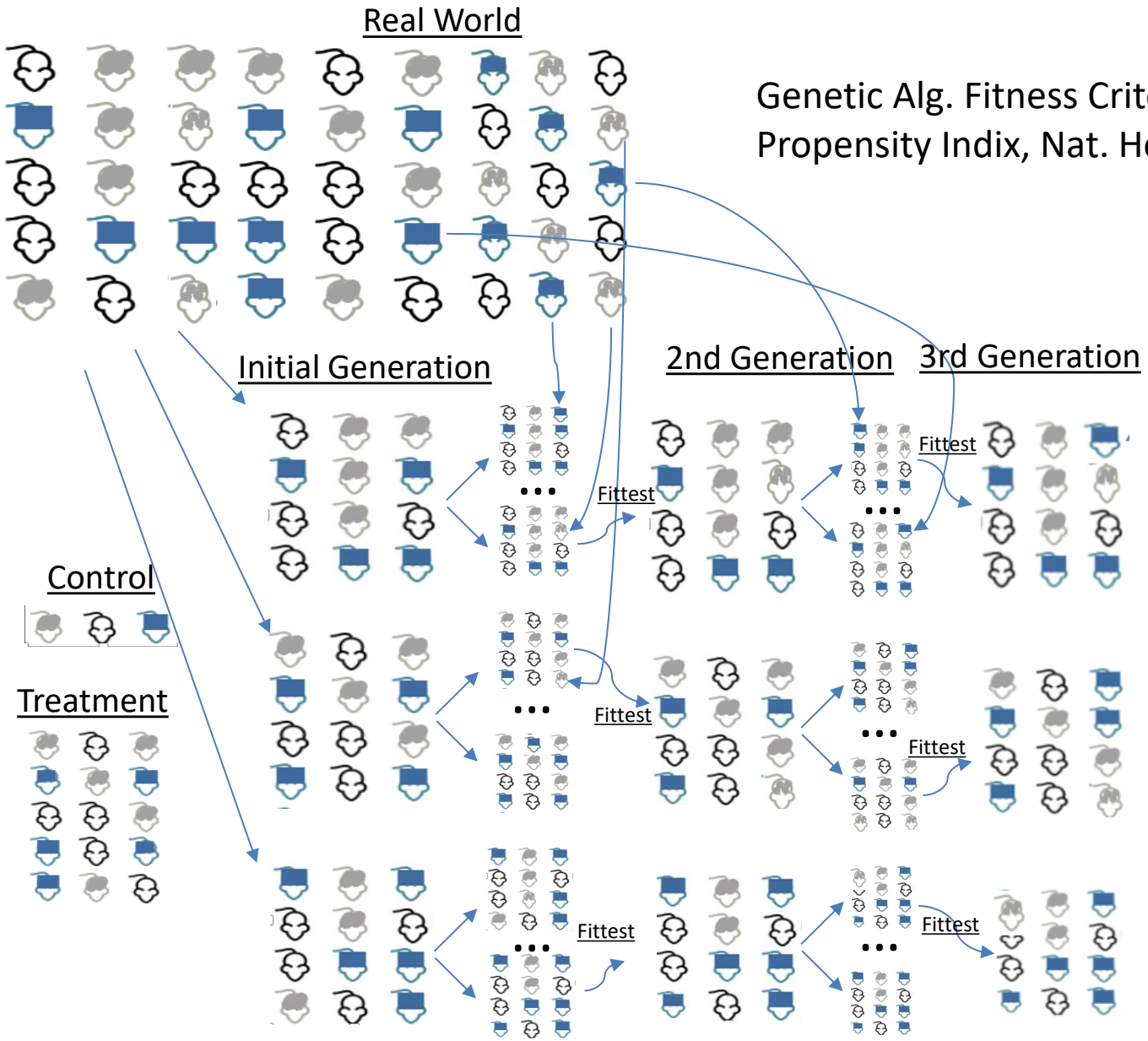
Genetic Alg. Fitness Criteria:
Propensity Index, Nat. Hermite Index, Ave CV^{-1},

A Genetic Algorithm (cont.)



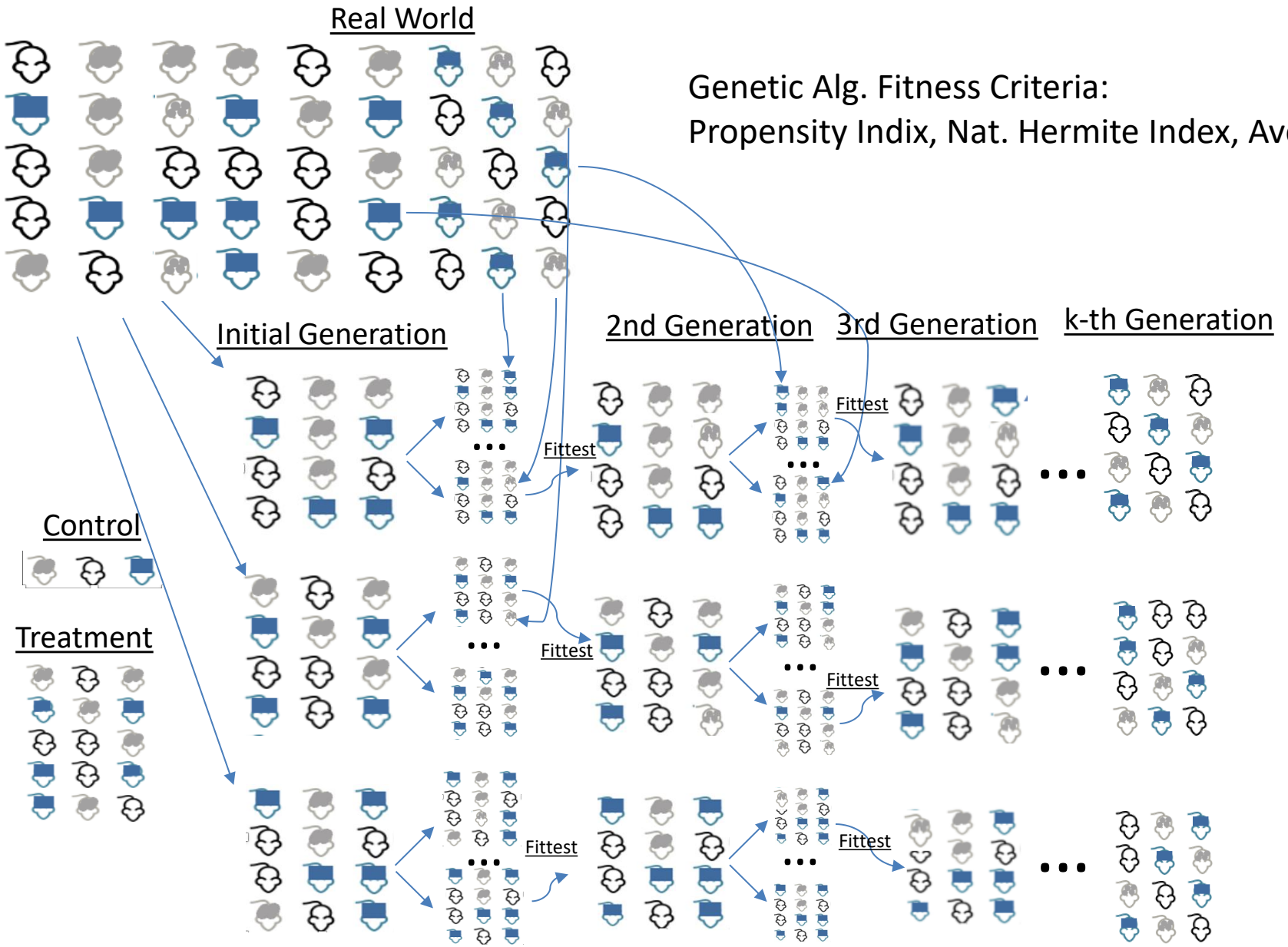
Genetic Alg. Fitness Criteria:
Propensity Index, Nat. Hermite Index, Ave CV^{-1},

A Genetic Algorithm (cont.)

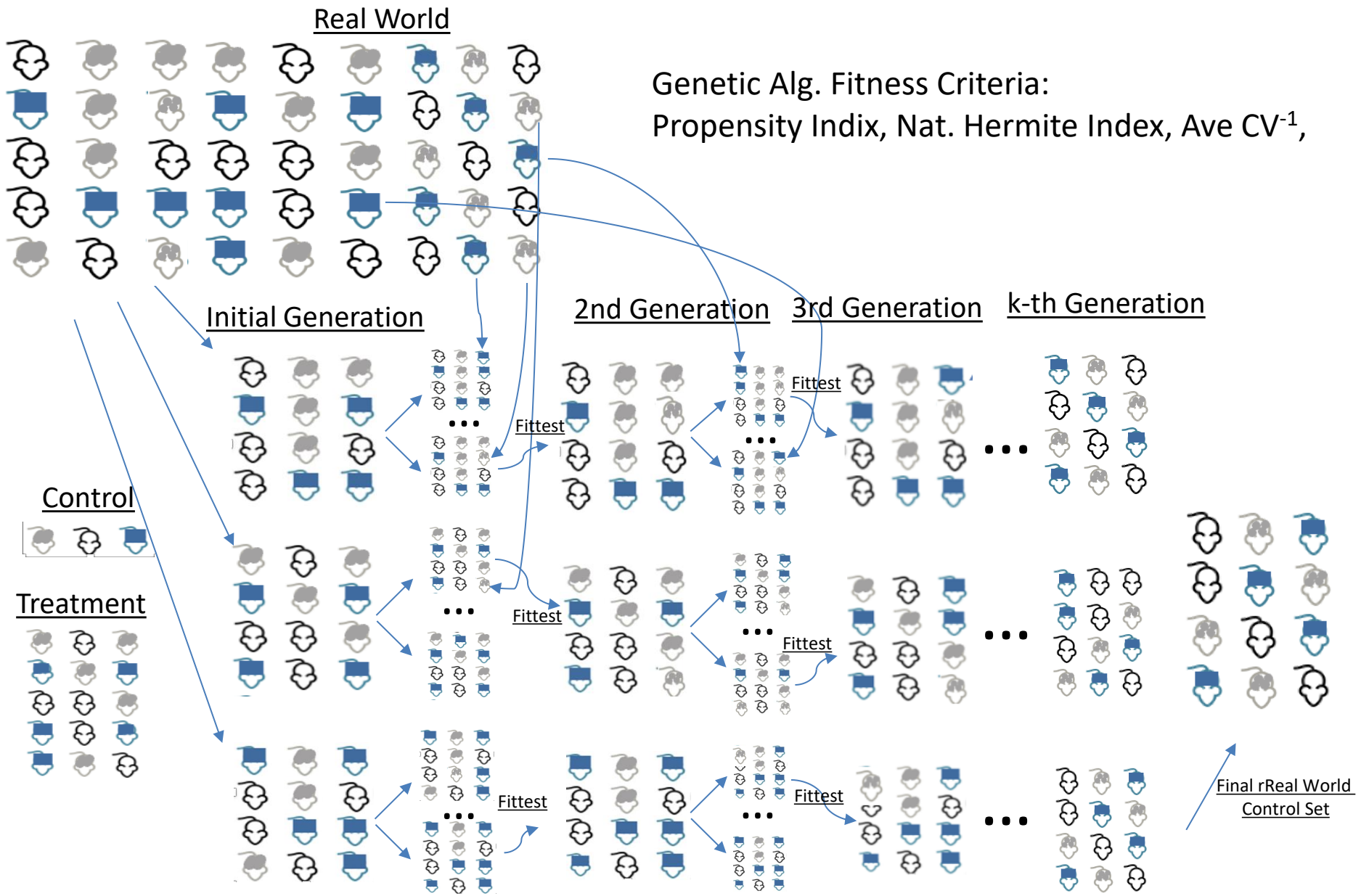


Genetic Alg. Fitness Criteria:
Propensity Index, Nat. Hermite Index, Ave CV^{-1},

A Genetic Algorithm (cont.)



A Genetic Algorithm (cont.)



A Genetic Algorithm (cont.)

The proposed approach consists of the following steps:

1. Draw an initial k (e.g., k=10) sets of m observations from R the RWD. Compute corresponding k index measures $\{c_i\}$. Also calculate

$$C_1 = \min_{1 \le i \le k} c_i$$

For g=2, …, G

2. Interchange one observation from each initial set of m observations by randomly drawn an observation from the the RWD.
3. Compute the index measure for each augmented datasets.
4. Repeat the above s times (e.g., s=50), forming $k \times s$ replicates of n_2+m observations and $k \times s$ index measures (c_{gij} , i=1, ..k; j=1,…,s).

5. Select the k datasets with the smallest c_{gij} and compute C_g the smallest $\{c_{gij}, C_{g-1}\}$

6. Repeat steps 2-5 G times until $C_{g+1-\delta}$- C_g correspond to the same configuration and are the lowest (we used $\delta = 3$) and less than 0.05

A Simulation Experiment

- 100 datasets generated simulating RCT, treat group of 50 and a control group of 10 subjects.
- RWD *R* had 50,000 subjects generated following the dataset available at: **https://www.khstats.com/blog/tmle/tutorial.**
- For the 50,000 subjects, the variables W_1, W_2, W_3, W_4 were generated from the following distributions:

$$W_1 \sim Bernoulli(p = 0.2), W_2 \sim Bernoulli(p = 0.5),$$

$$W_3 \sim Round(Unif(2,7)), W_4 \sim Round(Unif(0,4))$$

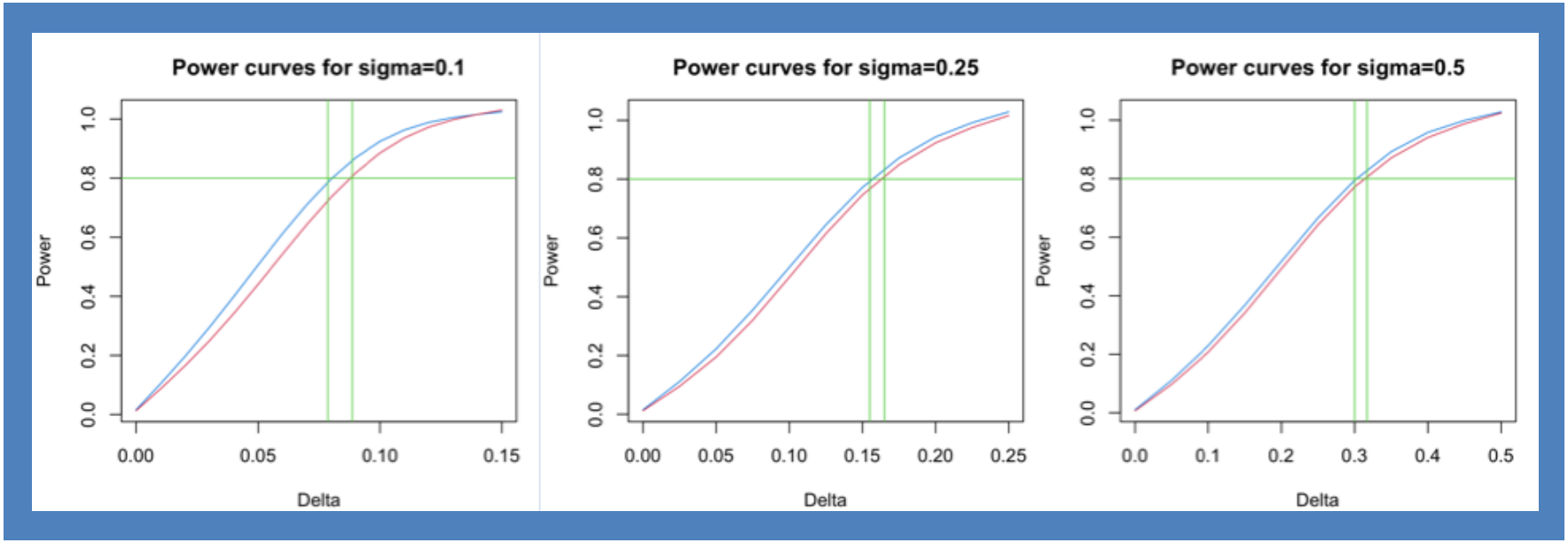
- Treatment group subjects were also generated from the same distribution as above.
- Control group of 10 subjects augmented by selecting 40 subjects from RWD by
  - Method 1 is standard randomization, and method 2 is using the genetic algorithm to find the subset which minimizes the propensity scores index.
- Response variable was generated 100 times for each dataset using the following formula:

$$Y = 5.5 + 0.2\,W_2 + log(0.1\,W_3) + 0.3\,W_4 + 0.2\,W_1\,W_4 + \delta\,Treat + \varepsilon,$$

where δ is the treatment effect size (0-5 range) and ε is normally distributed with zero mean and standard deviation σ_ε (0.1,0.25,0.5).

A Simulation Experiment

Power curves as a function of δ, obtained from the above model using the two augmentation methods. The three panels represent the power curves corresponding to the three values of σ_ε



Simulation results comparing random augmentation (red) to GA algorithm augmentation (blue) for 3 values of $\sigma\varepsilon$ (0.1,0.25,0.5).

In the case when the populations of treatment control and real-world are the same, it appears that the genetic algorithm improves power for detecting treatment effect versus control.

Back to the Real-World Example of Abruption

- Data was pre-processed using Algorithm 1, and the subset of the RWD that intersects the domain of the clinical dataset was selected.

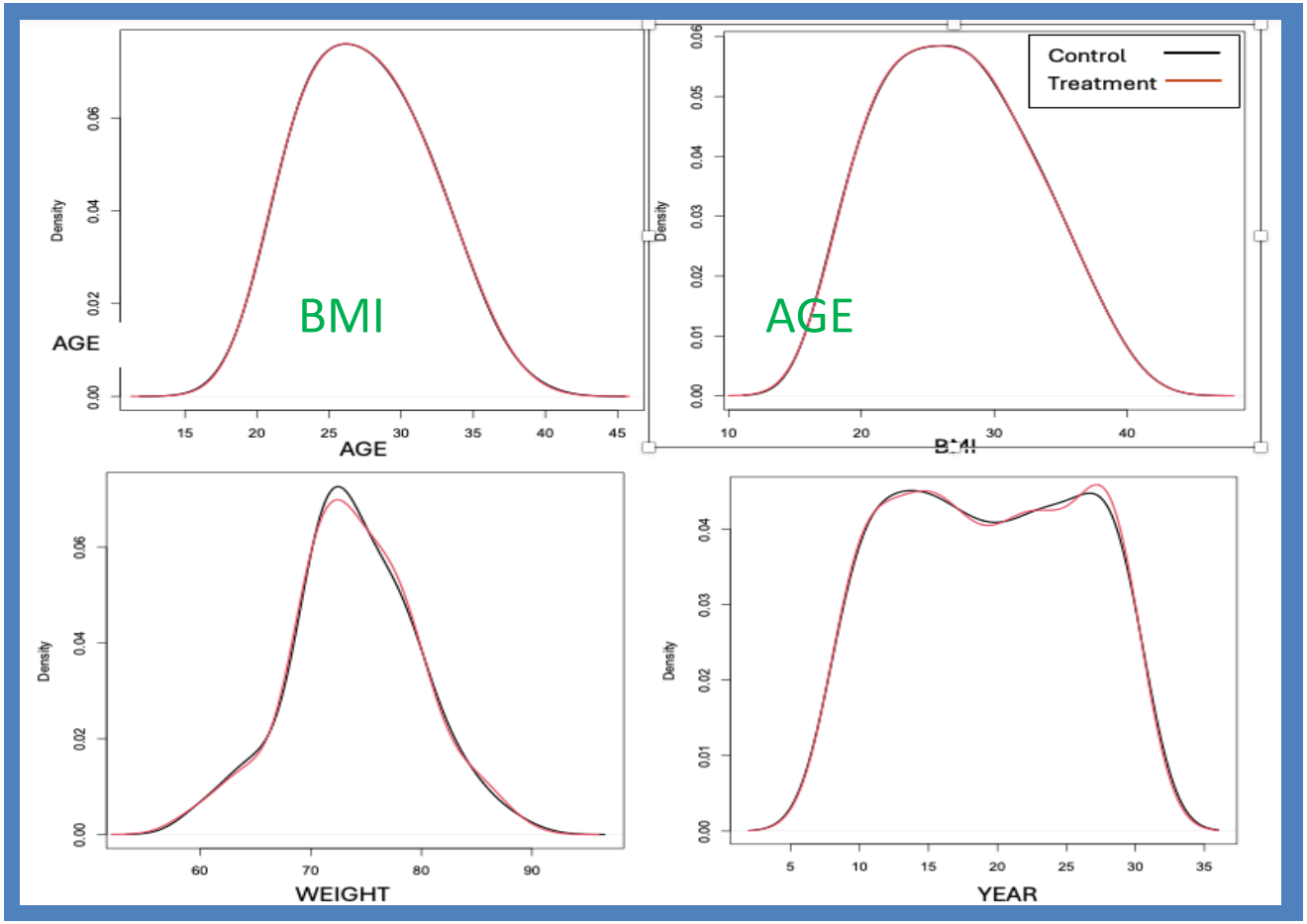
- This control data set of M=18,992 had the same domain as the clinical data, but the distribution was not the same.

- The discrete/categorical variables were all binarized resulting in 21 binary variables. After running the genetic algorithm, the proportion of 1s of each of the 21 binary variables was identical between the treatment and real-world controls and the difference in counts of 1s was zero for all the variables.

Results of the Genetic Algorithm Applied to RWD

After preprocessing: RW= 18992 records; clinical study: 1857 records.

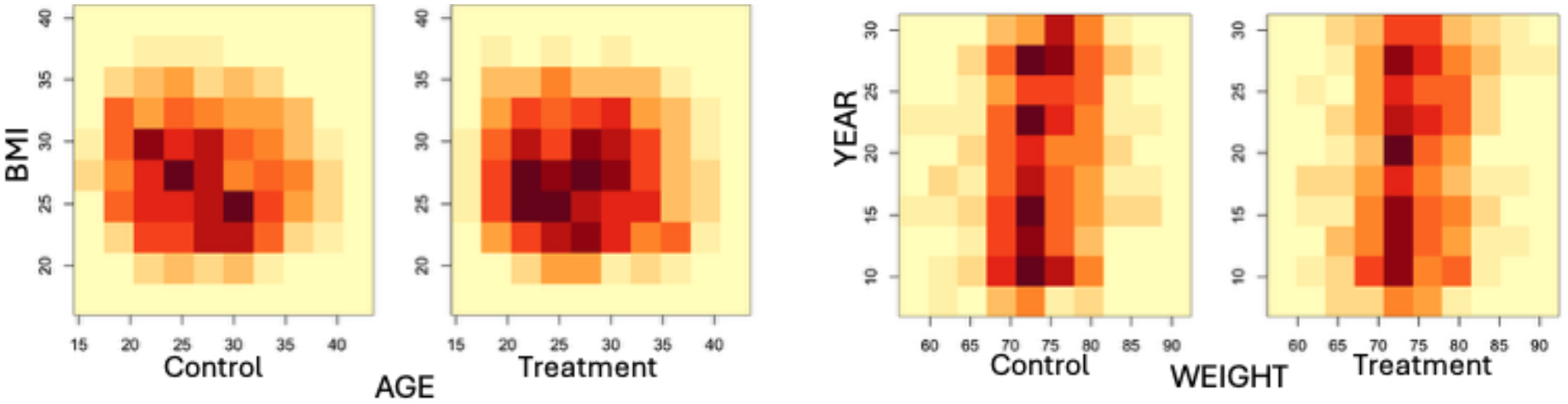
Results of GA for continuous vars



For Binary vars the marginals were identical

Results of the Genetic Algorithm Applied to RWD

Density estimators of the distributions of four continuous predictors in the treatment and control datasets.



The controls dataset is a subset of the real-world dataset of controls.

Algorithm for sequential randomization to rebalanced clinical studies

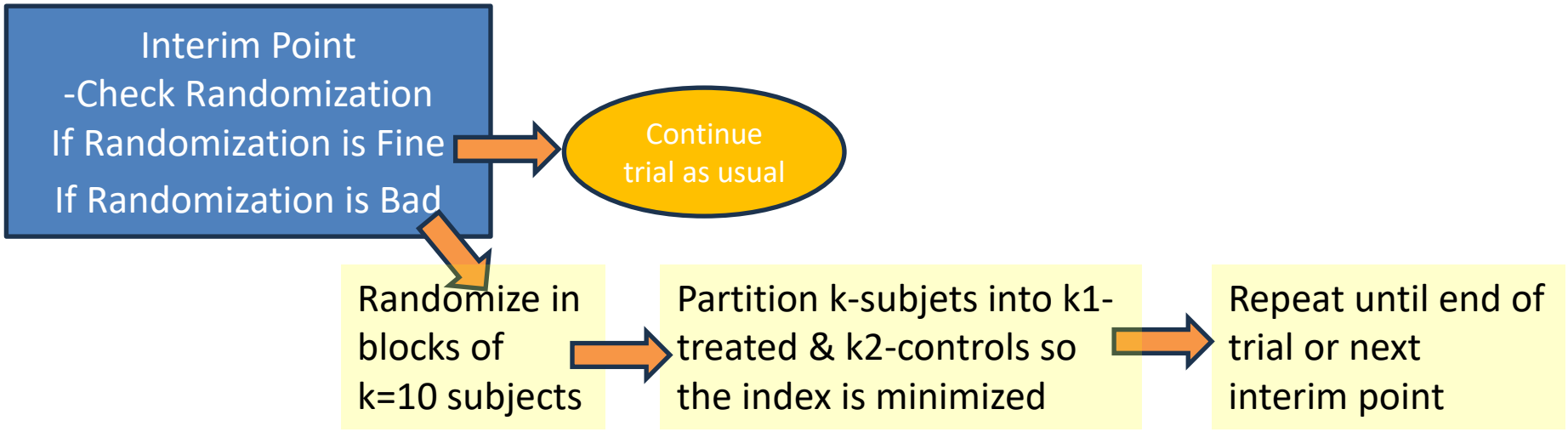
Clinical Trials get unbalanced (often loss of 25% of the subjects)

Loss usually not at random so Treatment and Control groups are unbalanced.

1. At Interim point check the samples balance.

- Check PS-index or NH index between treatment and control
- If the value is large compared to random split start Balancing pseudo-randomization

Balancing pseudo-randomization:



A Simulation Experiment

For each subject entering the clinical study variables W_1, W_2, W_3, W_4 were generated from the following distributions:

$$W_1 \sim Bernoulli(p = 0.2)$$
$$W_2 \sim Bernoulli(p = 0.5)$$
$$W_3 \sim Round(Unif(2,6))$$
$$W_4 \sim Round(Unif(0,4))$$

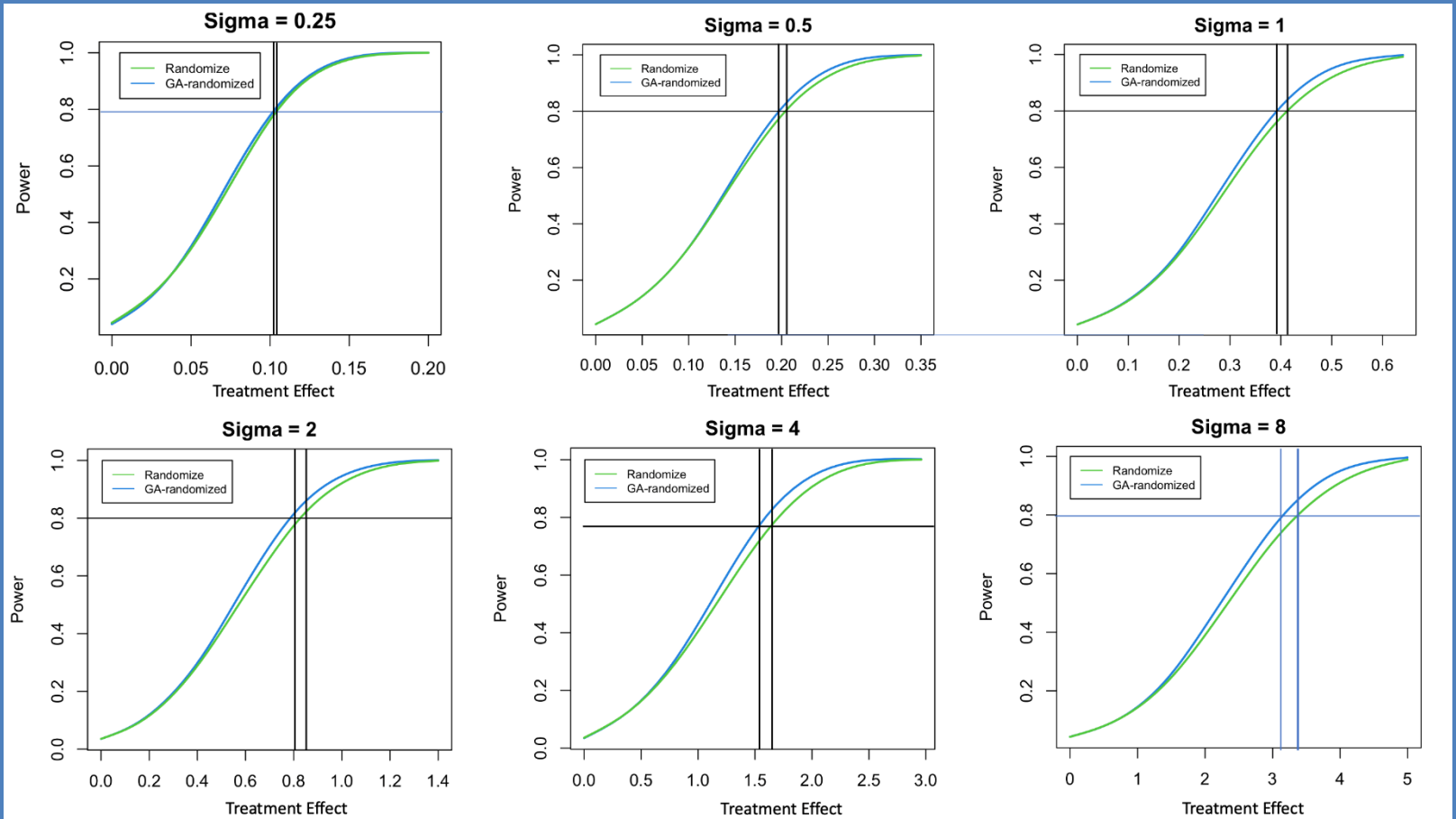
The first M observations are generated and randomized to treatment and control.
After the interim point the new observations are randomized in groups of 6 at the time and some are assigned to treatment or control to minimize the index and equating the number of observations in the treatment and control groups until completion.

The response was generated from the model

$$Y = 5.5 + \beta_1 Treatment + 0.2W_2 + \log(0.1W_3) + 0.3W_4 + 0.2W_1W_4 + \varepsilon$$

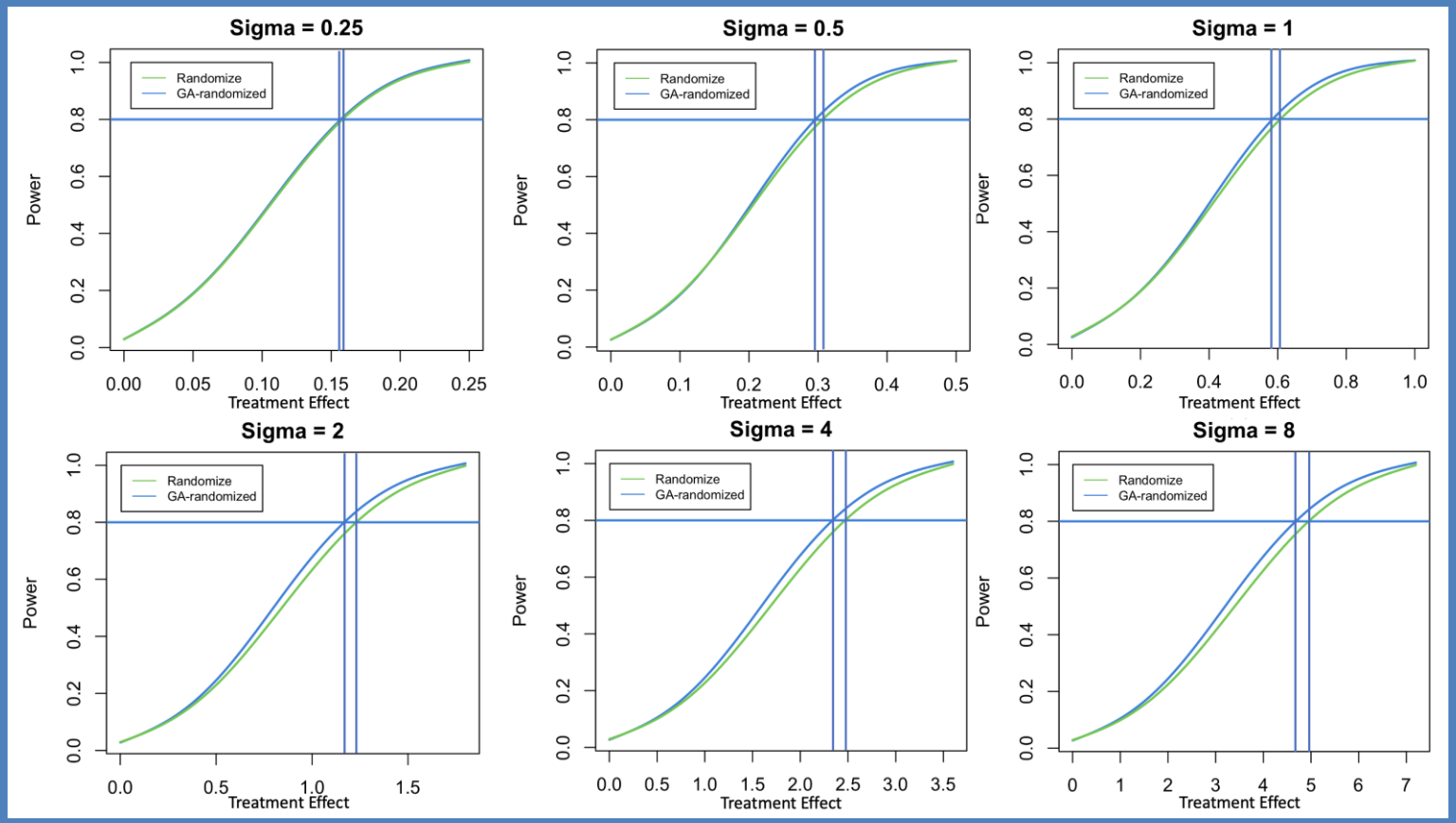
Simulation results comparing random (GREEN) to GA algorithm (BLUE) forvalues of $\sigma\varepsilon$. =0.25,0.5,1,2,4,8).

A Simulation Experiment: N=200



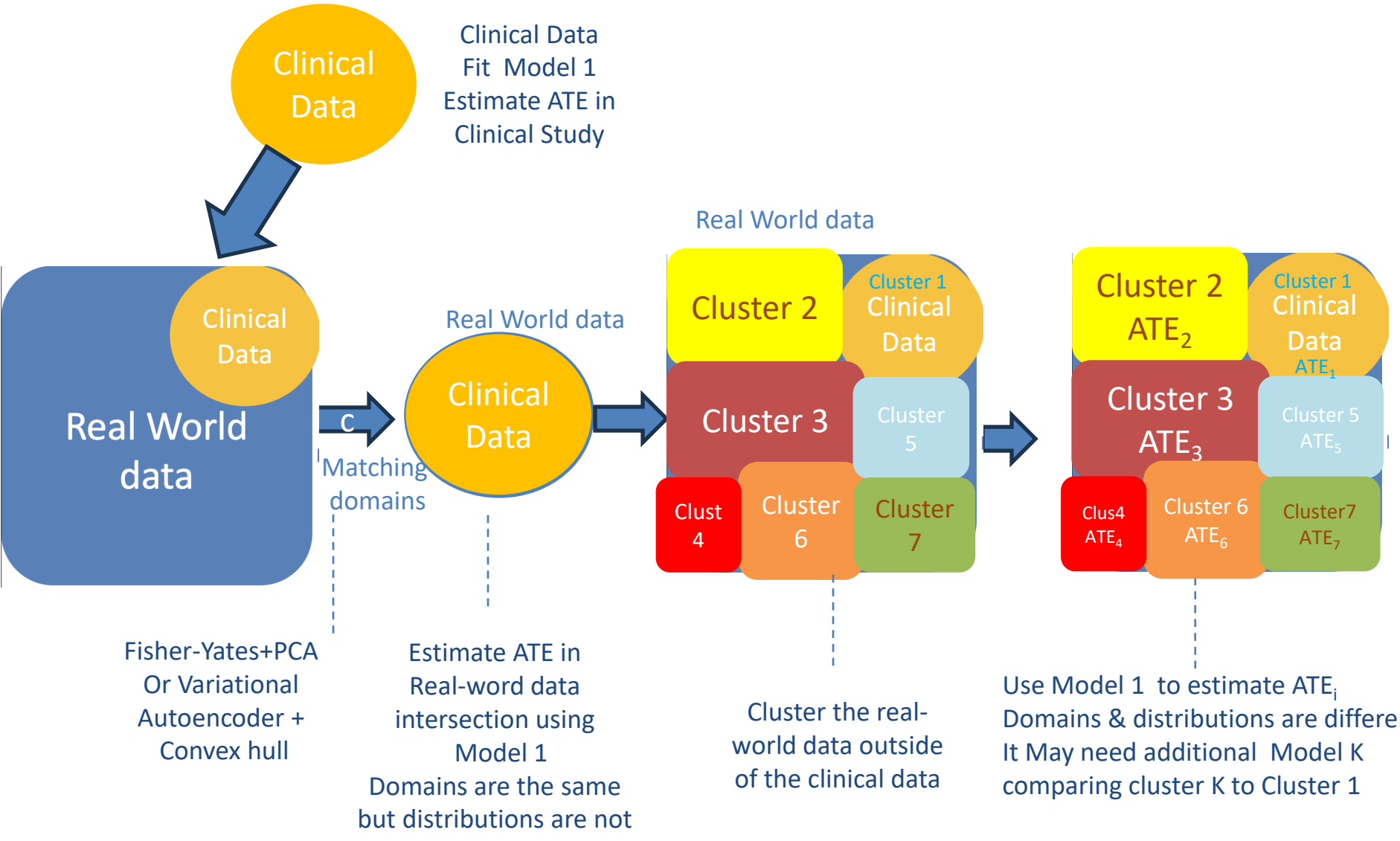
Power function of treatment as a function of effect β_1, for σ_ε=0.25,0.5,1,2,4,8. The study size is 200 subjects and the interim point is at 50 subjects. The green and blue curves are the power for treatment effect under standard randomization and Propensity scores index randomization respectively. The model for Y was given by

A Simulation Experiment: N=100



Power function of treatment as a function of effect β_1, for σ_ε=0.25,0.5,1,2,4,8. The study size is 100 subjects and the interim point is at 28 subjects. The green and blue curves are the power for treatment effect under standard randomization and Propensity scores index randomization respectively. The model for Y was given by

Extrapolating Model from clinical study to real-word data



Conclusions and Further Work

- External control arms, practical alternatives where RCTs are not feasible.
- Conventional matched pairs methods exhibit notable limitations, including data attrition and systematic bias arising from variations in patient health status and specificity
- Distribution matching methodologies introduced as viable options
- The Natural Hermite Index and its differential projection pursuit extension are presented as multivariate criteria for quantifying dissimilarity between distributions
- A measure based on the variance of estimated propensity scores also proposed
- A genetic algorithm utilized to iteratively refine subsets of real-world data to augment clinical trial control groups
- Simulation results indicate that the genetic algorithm demonstrates substantial efficacy in identifying control subsets whose distributions closely mirror those of the treatment group
- The outcomes are promising, revealing that the treatment and control group distributions exhibit near indistinguishability.
- Future research will encompass larger-scale simulations to further validate the effectiveness of the proposed distribution-matching methodologies.

Selected References

- Duan Y., Cabrera J., Emir B. (2024). A New Projection Pursuit Index for Big Data. JCGS,
- Dastgiri M., Duan Y., Cabrera J.(2025). Novel Machine Learning Approach to Differential Flow Cytometry Analysis base on differential Projection Pursuit. JBS
- Weigle S., Cabrera J., Sargsyan D.(2025) Randomization in pre-clinical studies: when evolution theory meets statistics. (under revision). JBS
- Cabrera J., Alemayehu D., Weigle S. Advancing Evidence Generation in Biomedical Research Using Natural Hermite and Propensity Score Indices: Applications to External Control Arms. (submitted)