Observational Studies
Advanced Materials

7 Key Aspects of Research Architecture

Almost Exact Matching / Fine Balance

Dealing with Time-Varying Covariates

Some Closing Thoughts

More Advanced Propensity Score Analyses

FEINSTEIN'S MODEL FOR RESEARCH ARCHITECTURE EXPANDED BY NEAL DAWSON

7 Key Aspects of Research Architecture*:
Making Fair Judgments about Causation

- Possibility of distorted assembly – sample doesn't reflect the population to which the results will be generalized.

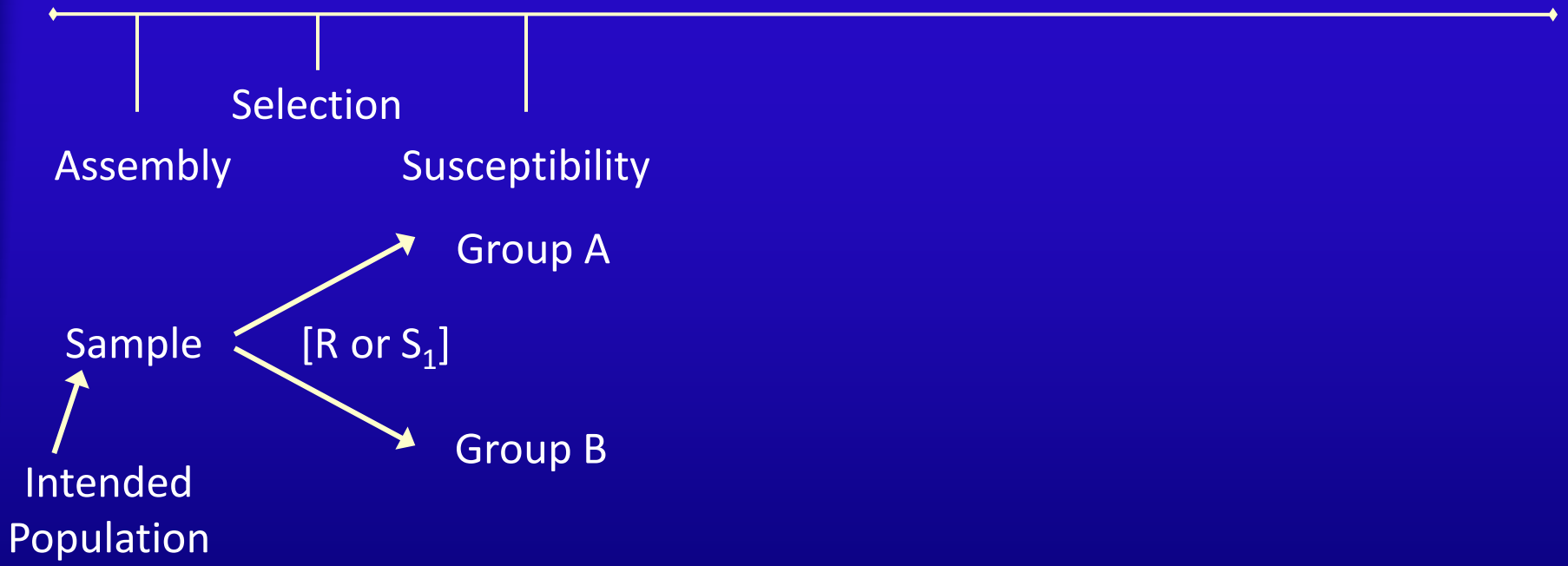
*Adapted by Neal Dawson from Alvan Feinstein's intellectual model (5 key aspects)

7 Key Aspects of Research Architecture*: Making Fair Judgments about Causation

- Selection Bias – who receives the exposure? Basis: (possibly unmeasured) covariates linked to outcomes? Why randomize?

*Adapted by Neal Dawson from Alvan Feinstein's intellectual model (5 key aspects)

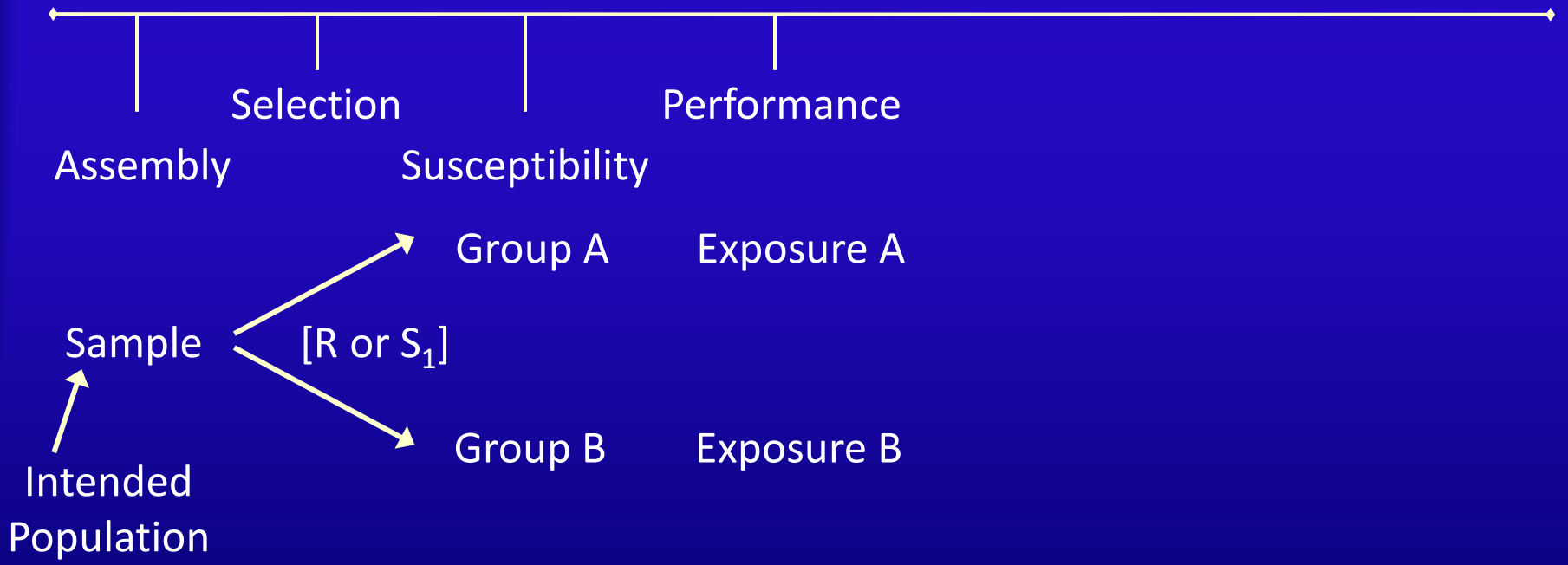
7 Key Aspects of Research Architecture*: Making Fair Judgments about Causation



- Are there importantly different expectations at baseline, for the eventual outcomes? Susceptibility reflects covariate differences.

*Adapted by Neal Dawson from Alvan Feinstein's intellectual model (5 key aspects)

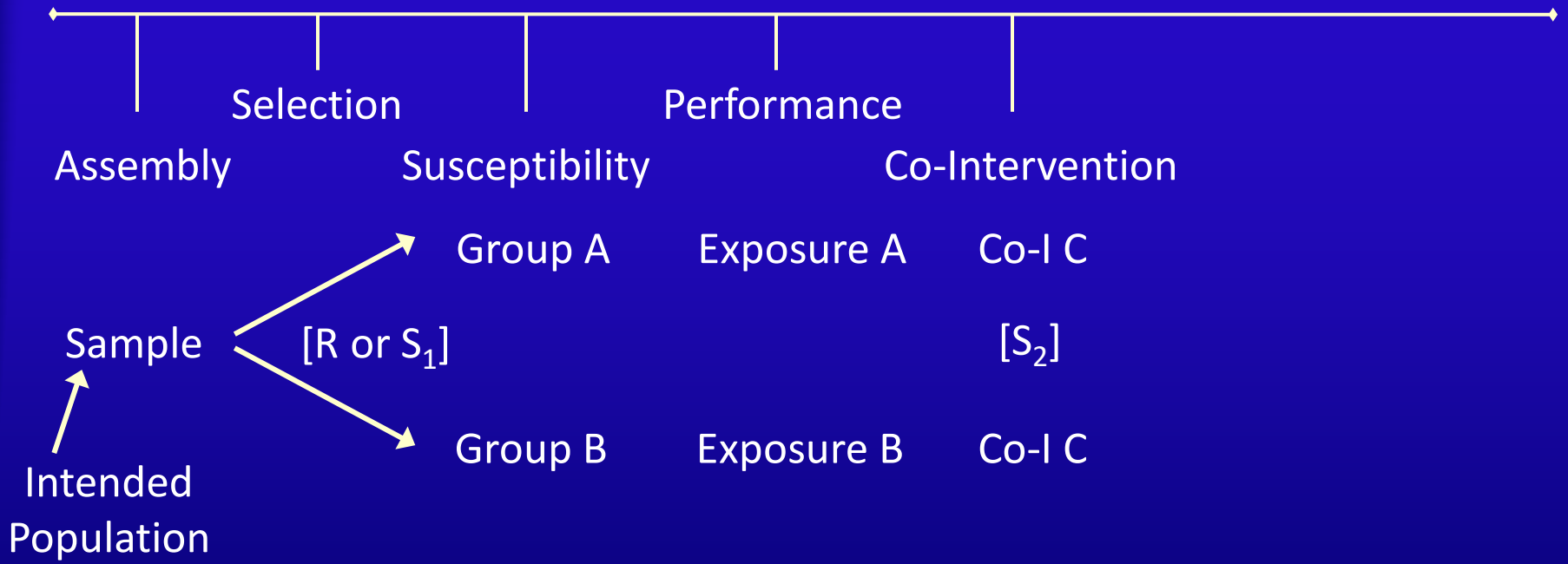
7 Key Aspects of Research Architecture*: Making Fair Judgments about Causation



- Are exposures applied with the same proficiency? How "well" do pts receive the exposures (dosage schedules, compliance)?

*Adapted by Neal Dawson from Alvan Feinstein's intellectual model (5 key aspects)

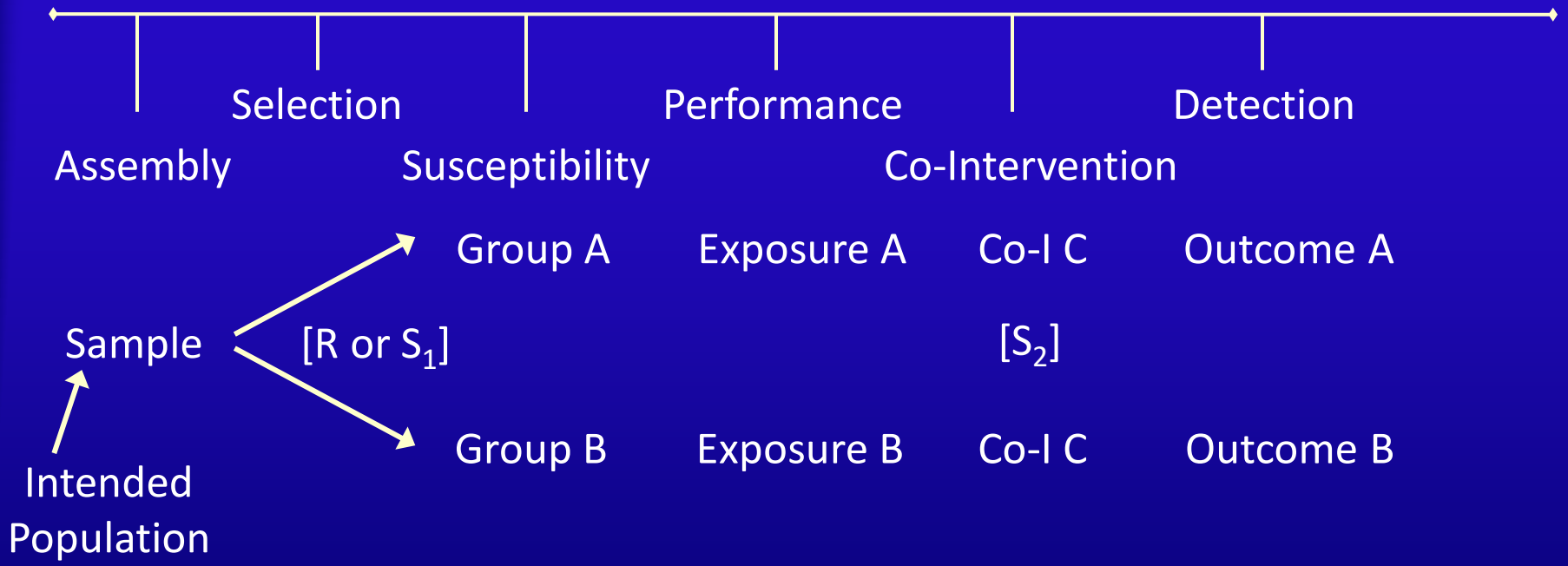
7 Key Aspects of Research Architecture*: Making Fair Judgments about Causation



- Additional selection opportunities – co-interventions (beyond exposure of interest) may influence likelihood of outcomes.

*Adapted by Neal Dawson from Alvan Feinstein's intellectual model (5 key aspects)

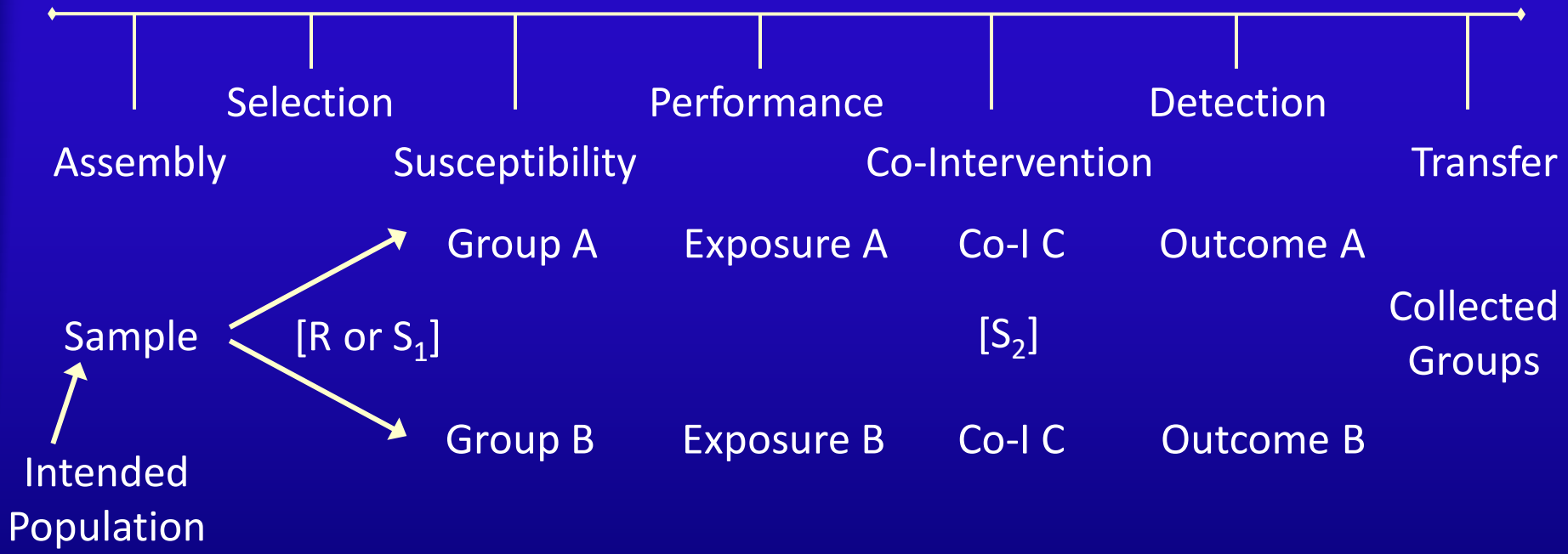
7 Key Aspects of Research Architecture*: Making Fair Judgments about Causation



- Is process for determining outcomes applied unequally? Differences in surveillance, diagnostic testing, or interpretation?

*Adapted by Neal Dawson from Alvan Feinstein's intellectual model (5 key aspects)

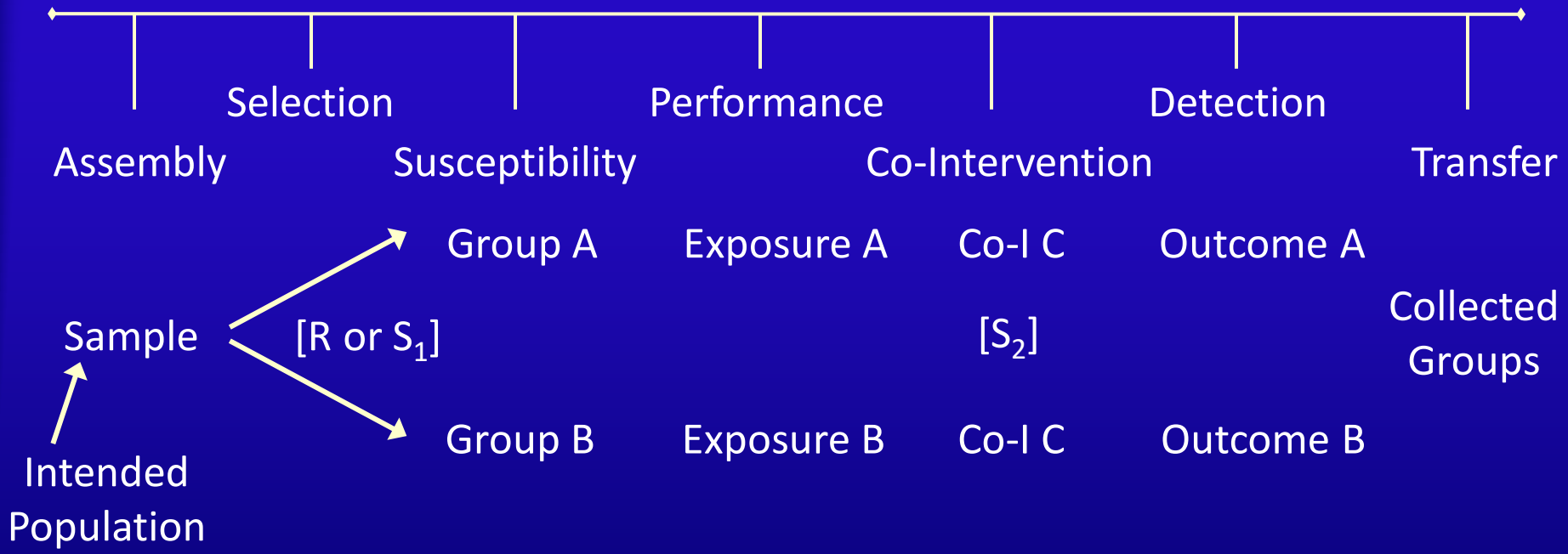
7 Key Aspects of Research Architecture*: Making Fair Judgments about Causation



- Comparison of members of original cohorts of A and B – dropouts, in-study exclusions, crossovers, dealing with missing data…

*Adapted by Neal Dawson from Alvan Feinstein's intellectual model (5 key aspects)

7 Key Aspects of Research Architecture*: Making Fair Judgments about Causation



- Goal: Comparability of groups who did and did not receive the exposure (except for the actual receipt of the exposure)

*Adapted by Neal Dawson from Alvan Feinstein's intellectual model (5 key aspects)

More Advanced Propensity Score Analyses

ALMOST EXACT MATCHING
FINE BALANCE
OPTIMAL FULL MATCHING

Is Regression Adjustment for Observational Studies Obsolete?

- Matching and stratification are old and trusted methods of adjustment for observational studies, but the difficulty of implementing them led earlier practitioners to prefer regression.
- It is now possible to perform optimal full matching, and to achieve levels of bias reduction that were previously unattainable.

Hansen (2004), Hosman et al. (2010), Rosenbaum (2010)

General Approaches to Optimal or Near-Optimal Constrained Matching

- Calculate propensity scores
- Establish distance matrix
  - Just a table with one row for each treated subject and one column for each potential control
  - Distances could be squared differences in propensity scores between the patients, or …
  - To use calipers – set to ∞ all cells in the table corresponding to a propensity difference > caliper
  - Can use other distance measures (Mahalanobis)

Rosenbaum (2010)

Who Gets Matched?

Treated Patient	Control 1 [M]	Control 2 [F]	Control 3 [F]	Control 4 [M]	Control 5 [F]	Control 6 [F]
1 [M]	.23	.47	.39	∞	.51	.35
2 [F]	.45	∞	.28	.31	.42	∞
3 [F]	∞	.35	∞	.27	.44	.28
4 [F]	.31	.26	.51	.29	∞	.24

- Treated Patient 1 matches to Control 1
- Treated 2 matches to Control 3
- Treated 3 to Control 4 (or 6 – sex important?)
- Treated 4 to Control 6 or 2 or 4 – hmmm…

Almost Exact Matching

- Suppose a few of the covariates are of enormous importance – want to match exactly on them wherever possible
- Adding a penalty to the distance matrix when the specified covariates don't match is the main approach here –
- Adding 2 to the Mahalanobis distance for mismatches roughly doubles the importance of that covariate as compared to the others.

Rosenbaum 2010

Fine Balance in Matching

- Constrain optimal matching that forces a nominal variable to be balanced, without restricting who is matched to whom
  - Useful with a nominal variable with many levels
  - Useful if you have a rare binary variable that is difficult to control using a distance
  - Useful when focusing on the interaction of several nominal variables
- Possible to get specific imbalance patterns.

Rosenbaum 2010 Chapter 10

Fine Balance
Initial Distance Matrix

Treated Patient	Control 1 [M]	Control 2 [F]	Control 3 [F]	Control 4 [M]	Control 5 [F]	Control 6 [F]
1 [M]	.23	.47	.39	∞	.51	.35
2 [F]	.45	∞	.28	.31	.42	∞
3 [F]	∞	.35	∞	.27	.44	.28
4 [F]	.31	.26	.51	.29	∞	.24

- Want optimal balance on propensity score while matching perfectly on gender margins
  - 4 treated patients (1 male, 3 female)
  - 6 potential controls (2 male, 4 female)
  - So we need to remove 1 male, 1 female in match

Augment Distance Matrix

Treated Patient	Control 1 [M]	Control 2 [F]	Control 3 [F]	Control 4 [M]	Control 5 [F]	Control 6 [F]
1 [M]	.23	.47	.39	∞	.51	.35
2 [F]	.45	∞	.28	.31	.42	∞
3 [F]	∞	.35	∞	.27	.44	.28
4 [F]	.31	.26	.51	.29	∞	.24
Extra 1	0	∞	∞	0	∞	∞
Extra 2	∞	0	0	∞	0	0

- Add 2 rows to the matrix, then run the match.
  - Extra 1 pulls away one male control.
  - Extra 2 pulls away one female control.
- Gender will be perfectly balanced, but the pairs will **not** be exactly matched for gender.

Fine Balance General Procedure

- To get the minimum distance match with fine balance (on nominal covariate, say **RACE**)

1. Cross tabulate **RACE** with treatment indicator
2. Determine # of controls to remove from each category of **RACE** to achieve perfect balance
3. Add one row for each control that must be removed, with 0 distance to its own category and infinite distance to all others
4. Find an optimal match for this square matrix
5. Discard extra rows and their matched controls

Full Matching in a study of coaching for the SAT

- In the past, it has been tough to implement full matching in observational studies, even though it is appealing in principle:
- Alignment of comparable treated and control subjects is as good as any alternate method, and potentially much better
- Hansen modifies full matching with modifications to minimize variance as well as bias
  - Optimal full matching removes as much as 99% of the bias along a PS on which treated and control means are separated by 1.1 SD's.
  - Reduces to insignificance biases along **27** covariates, while making use of more, not less, of the data than regression based analyses.

Hansen 2004

SAT Coaching Study

- Survey of a random sample of 1995-1996 SAT test takers about their preparation
- 12% of respondents had completed extracurricular test preparation courses
- Matching looked unattractive to the original researchers due to significant reduction in sample size – but they only considered 1:1 matching.
- Do 1:k matching options look better?

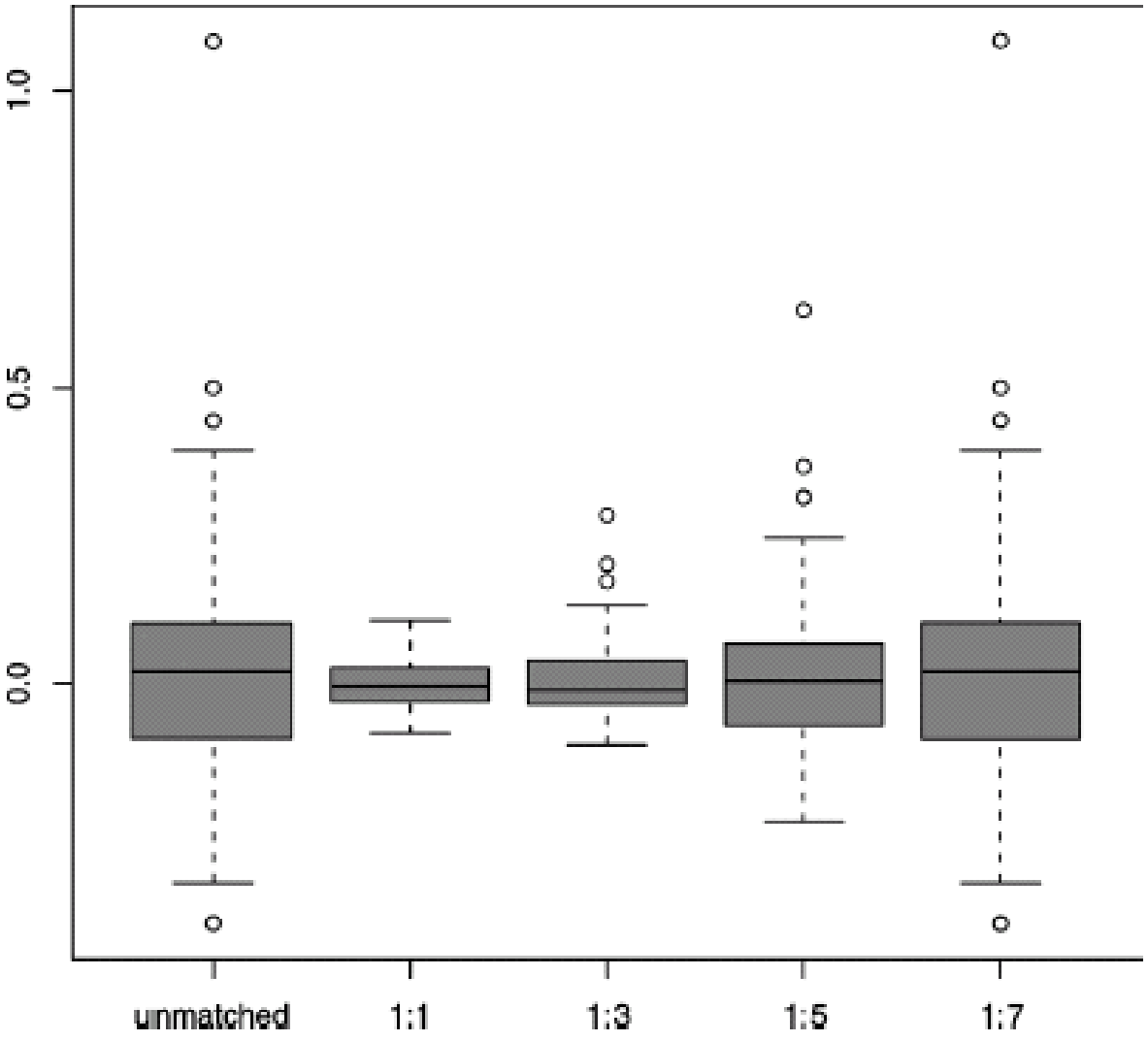


Figure 1. Covariate Imbalances in 1 : k Matching. Each boxplot represents standardized biases in the 99 categories of the 27 categorical covariates along with standardized bias in the propensity score (which in each plot is the uppermost outlier). Strictly speaking, the matching represented at far right is not a 1 : 7 matching but a blend of six 1 : 6 and 494 1 : 7 matched sets.

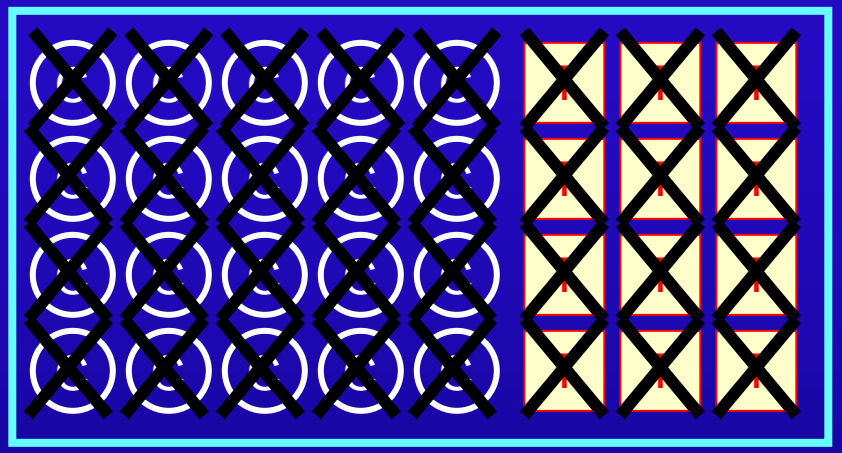
Covariate Imbalances in 1:k Matching

- In all of these cases, we're using less data
- Still some imbalance

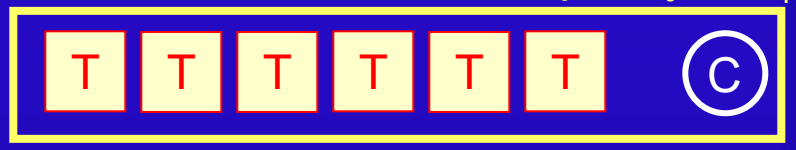
Hansen 2004

Optimal Full Matching

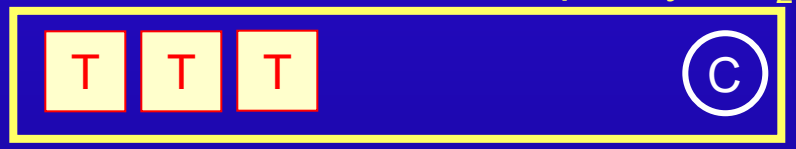
ORIGINAL SAMPLE



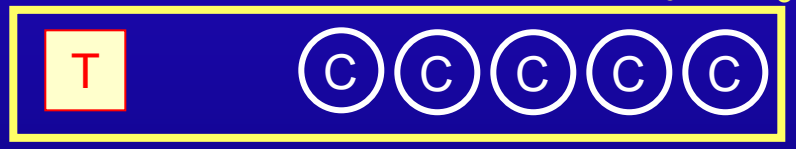
MATCHED SET 1: Discrepancy = D_1



MATCHED SET 2: Discrepancy = D_2

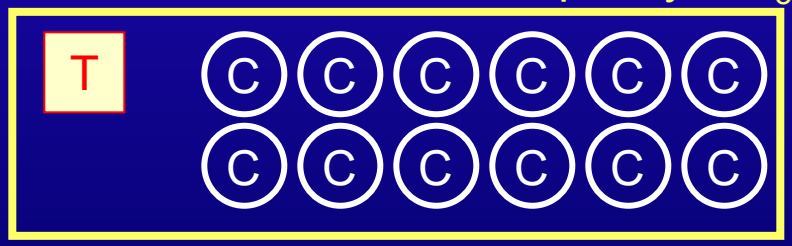


MATCHED SET 3: Discrepancy = D_3



MATCHED SET 4: Discrepancy = D_4

MATCHED SET 5: Discrepancy = D_5



- OFM minimizes propensity score distances (discrepancies) while using all treated and all control subjects (i.e. discarding no units).
- Here, infinite distances force matches on Race×Sex

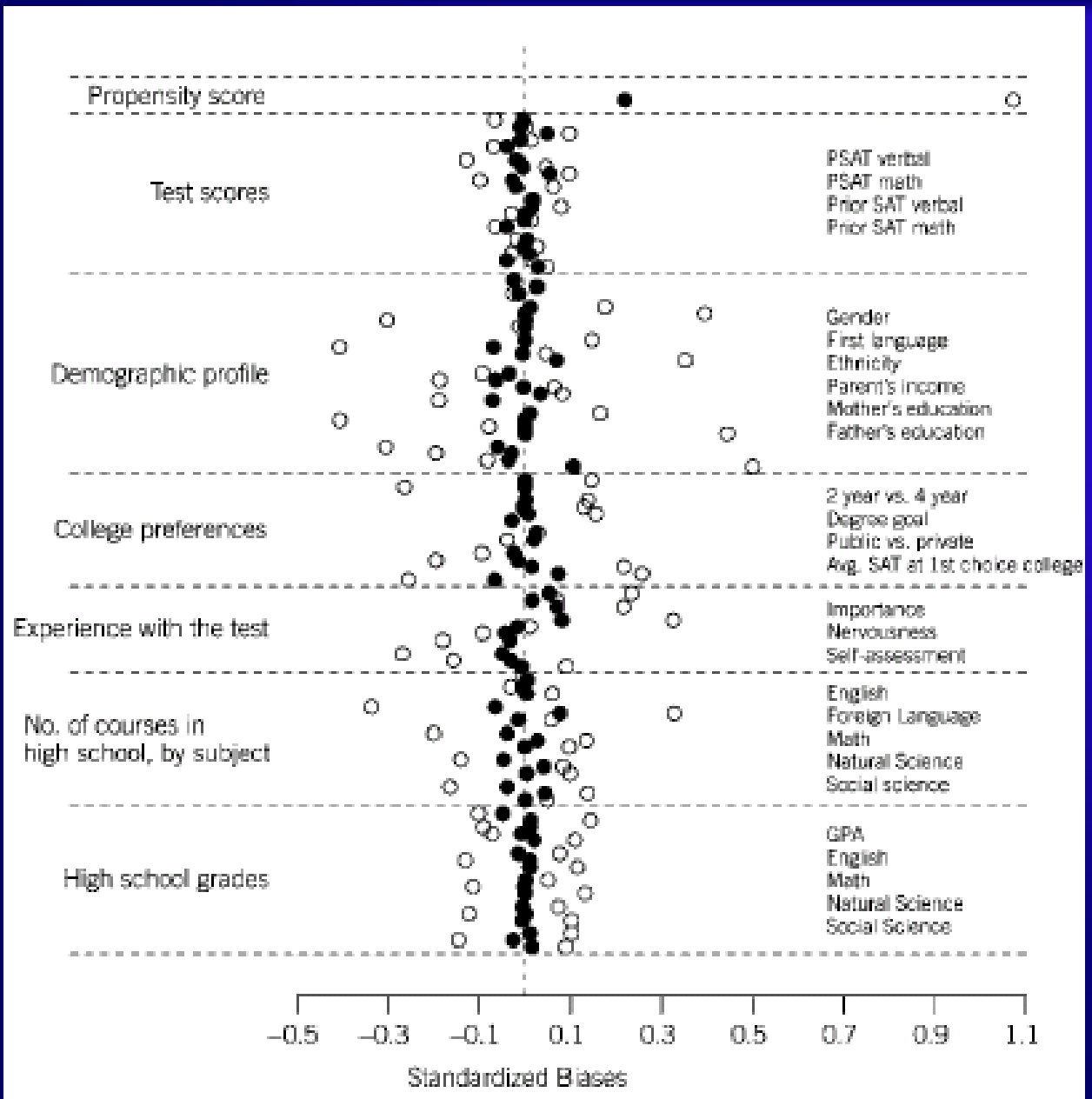


Figure 3. Standardized Biases Without Stratification or Matching, Open Circles, and Under the Optimal [.5, 2] Full Match, Shaded Circles.

Standardized Bias Plot

- Open circles are for standardized biases before matching
- Shaded circles describe results after full match

SAT Coaching Study Results

- Raw differences of treated and control group means were 41 points on Math and 9 on Verbal
- Full matching leads to aggregate contrasts of 26 points on Math and 1 point on the verbal.
  - Standard errors for these estimates are around 5 points.
- Surprised that Verbal effect is so small?
  - Control is not “no prep at all”
  - Estimated effect of treatment on the controls is 3 for Math and -8 on Verbal.
- Method doesn’t require homogeneity of coaching effects – whether and to what degree coaching is beneficial appears to vary greatly across students

Why not model outcome using all variables in the propensity model?

- Two stages: fit PS, then use PS in model
- One stage: just fit big outcome model
- Pros of two-stage approach:
  - Forces you to think hard about selection.
  - You don't care about parsimony in the PS, so you get maximum predictive value there.
  - You can fit a very complicated PS model first with interactions, higher order terms, splines, etc.
  - You can fit a smaller outcome model, which may let you assess its validity more accurately.

D'Agostino 1998 and Rubin 2001 provide some more details

Propensity Scores for More Than Two Possible Exposures

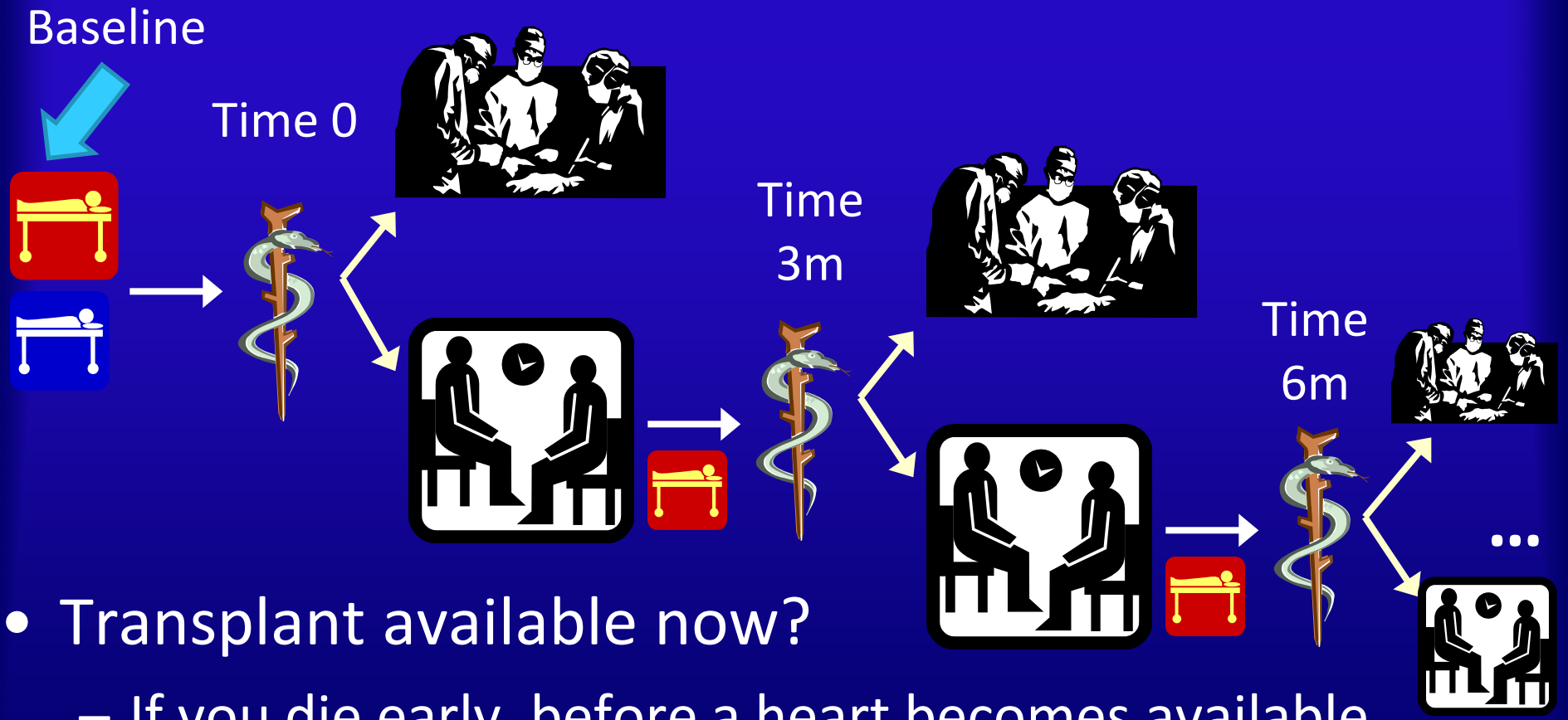
- Several generalizations have been proposed, which attempt to maintain the balancing properties of the usual propensity score...
- Role of the PS in estimating dose-response functions Imbens 2000
- Non-bipartite matching with doses in an OS of a media campaign against drug abuse Lu et al 2004
- Propensity Scores with Continuous Treatments Hirano and Imbens 2004; Elliott Zhang and Small 2015

DEALING WITH TIME-VARYING COVARIATES

When Do Time-Varying Covariates Arise?

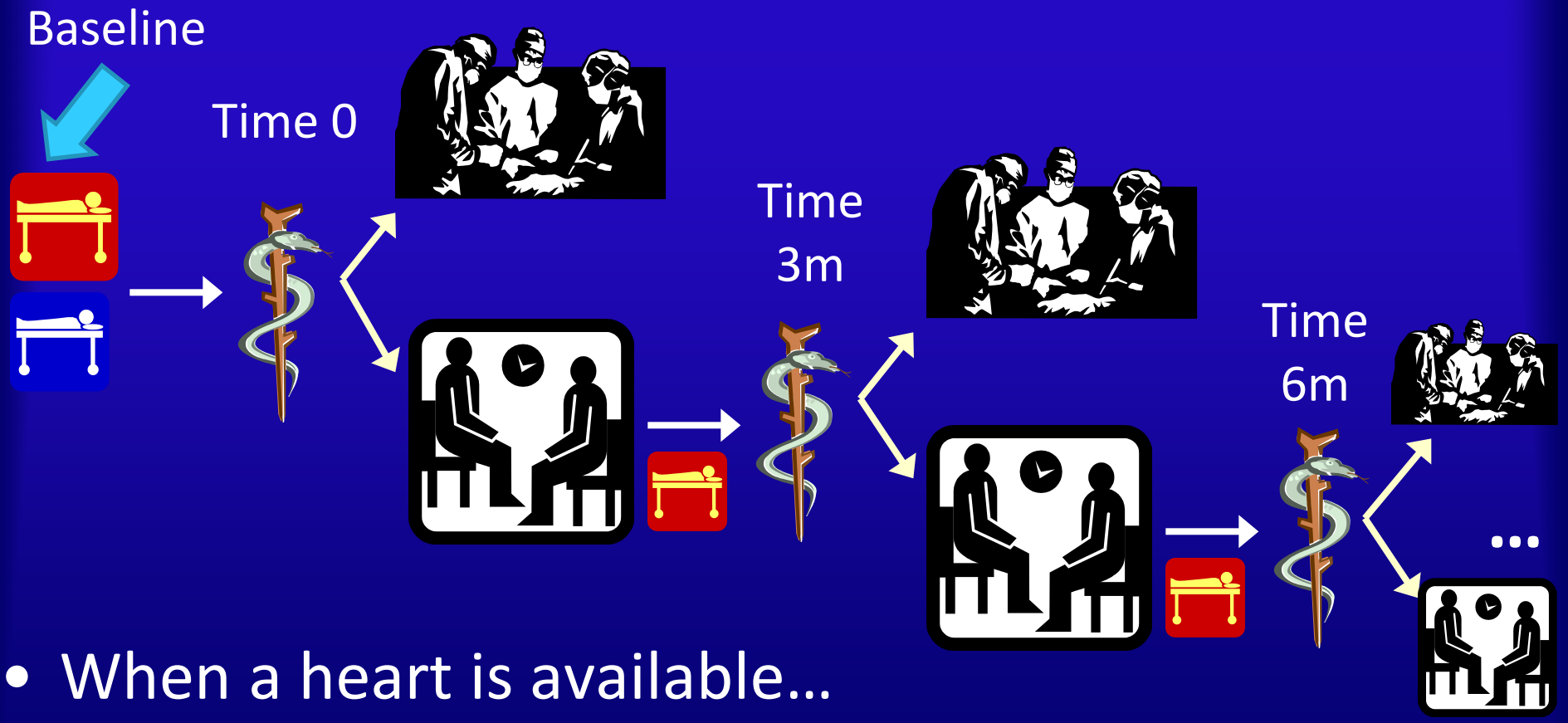
- When a treatment may be given at various times, we need to form matched pairs or sets in which subjects are similar prior to treatment, but avoid matching on events that followed treatment.
- In risk-set matching, a newly treated subject at time T is matched to controls who are not treated at time T based on covariate information describing subjects prior to time T.

Surgery Now vs. Waiting…



- Transplant available now?
  - If you die early, before a heart becomes available, you become a “control.”
  - If you live long enough to get a new heart, you’re treated.

Building Randomized Experiments



- When a heart is available…
  - Identify a pair of well-matched patients at that time
  - Randomly give heart to one, and compare survival…
  - Or keep “non-receiver” on list; investigate delay effects

Example 1. Risk-Set Matching in a Study of Surgery for Interstitial Cystitis

- Effects of surgery (cyctoscopy & hydrodistention) on symptoms of IC (chronic urologic disorder)
  - IC Database – patients can enter after at least six months of IC symptoms
  - Patients evaluated at entry, then every 3 months
  - Pain (0-9), Urgency (0-9), Nocturnal Voiding Freq.
  - Patients were treated periodically, not selected at random – those who find current symptoms unbearable will presumably opt for surgery

Li, Propert and Rosenbaum 2001, Rosenbaum 2010, Chapter 12

Matching in the IC Study

- Can't compare all "surgery" to all "no surgery" patients since if you never had surgery, your symptoms probably were never intolerable.
- Want to create pairs of patients who were similar up to the moment one had surgery
  - Matching makes pairs comparable pre-treatment
  - What happens after treatment is an outcome
  - Each new surgical patient is paired with a control who had similar symptoms up to the point of surgery for the surgical patient

Comparing Matched Pairs To Estimate the Impact of Decision

- We're estimating the effect of the choice that patients and their surgeons keep facing …

- Paired comparison estimates the effect of

  SURGERY NOW vs. DELAYING SURGERY

  with the delay being into the indefinite future

How The Matched Pairs Were Formed: Match on Observed History

- Patient A has surgery at time T, with covars. a
- Find a matched control by selecting from those potential controls who:
  - Have covariate history = a (or close to it)
  - Who did not receive the treatment at time T
- At time T, select a match for a patient getting surgery (with history a) from among the set of unmatched controls who have similar history.

Matching Algorithm

- Use a distance metric as we've seen previously, but with a change.
- Distance between patient with surgery at time T and a patient who had not yet received surgery as of time T computed using covariates for those two patients up to time T
- Used no information obtained after time T.

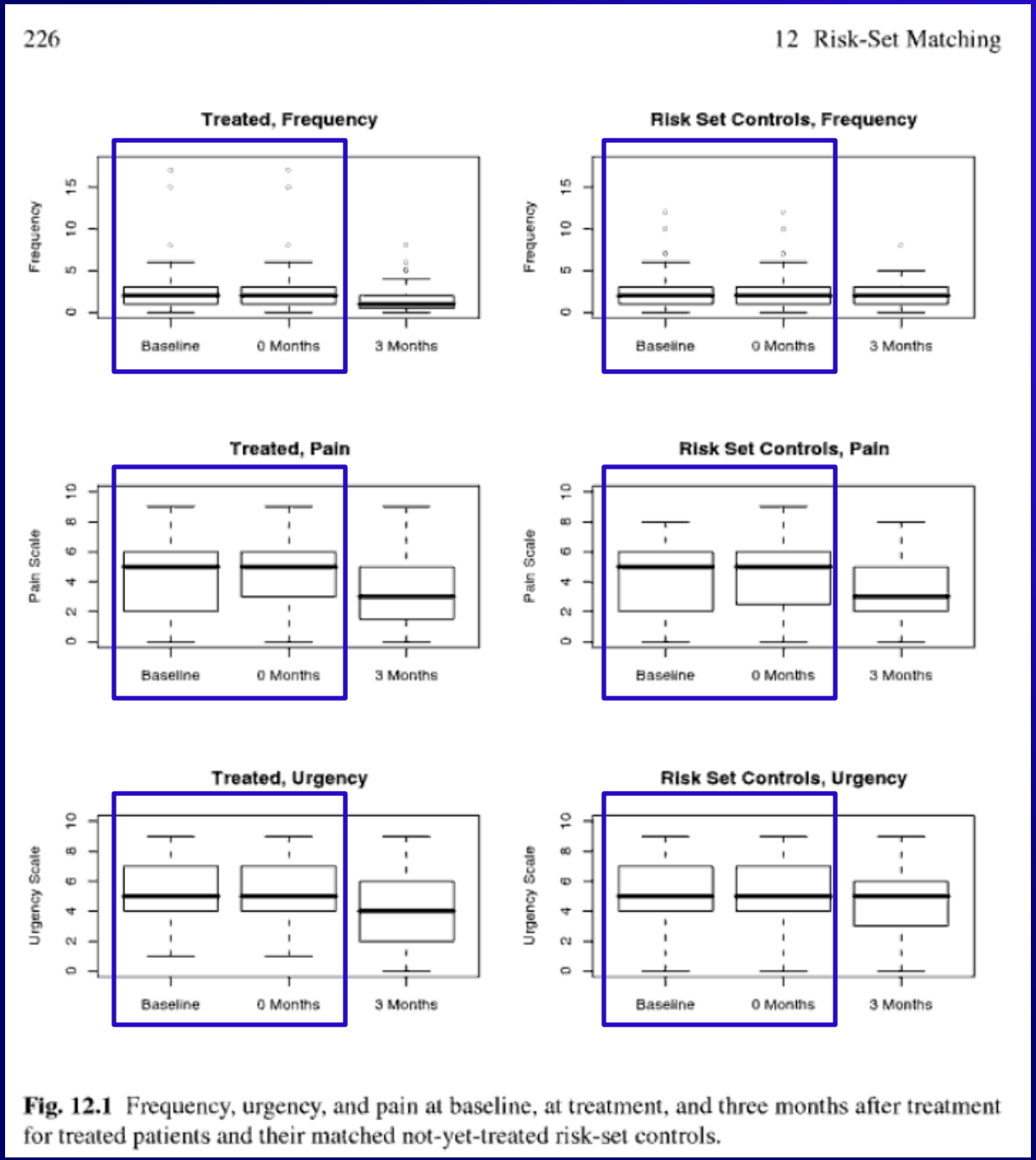
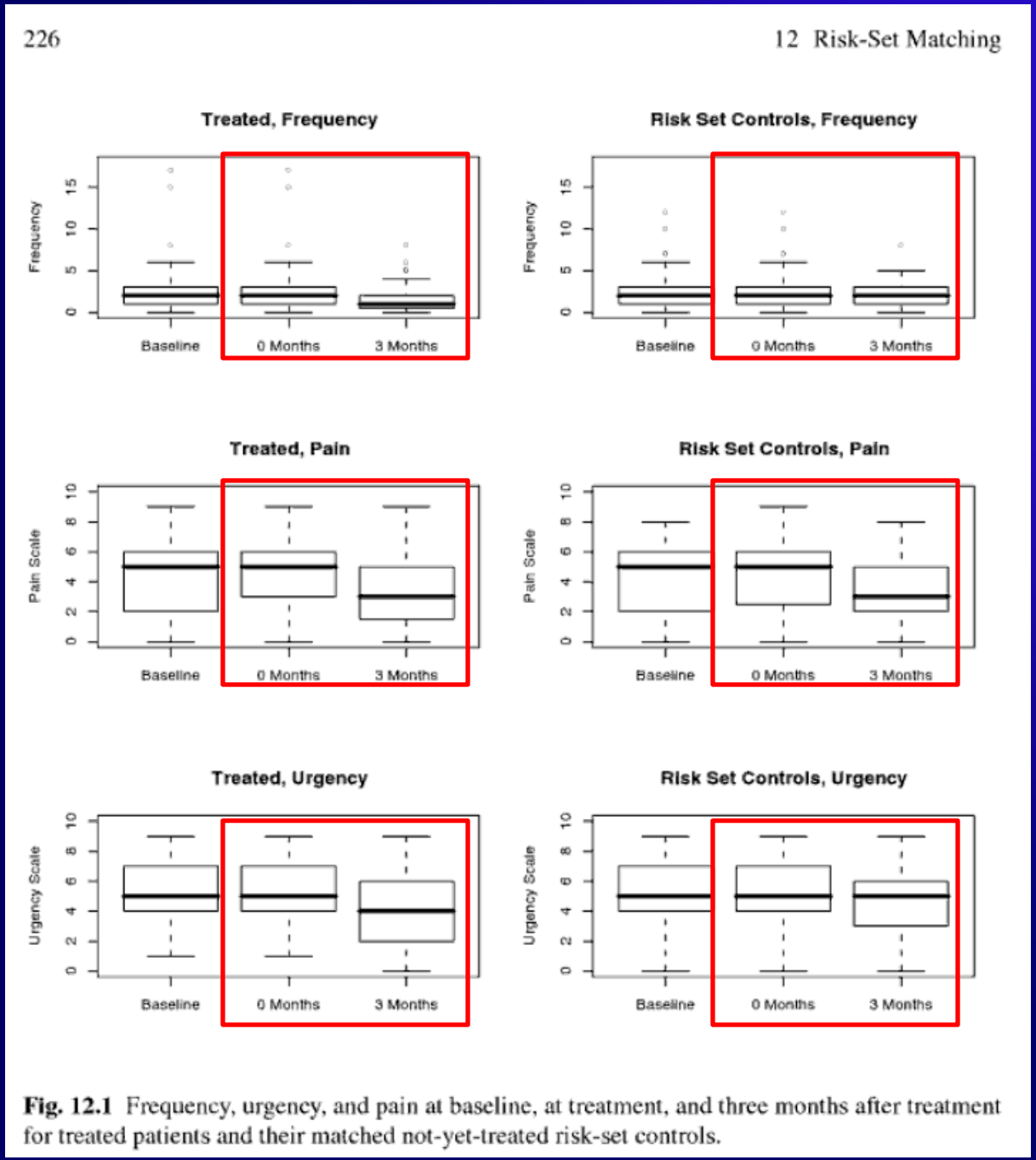


Fig. 12.1 Frequency, urgency, and pain at baseline, at treatment, and three months after treatment for treated patients and their matched not-yet-treated risk-set controls.

Effects of Matching

- Baseline
- 0 months (when treated)
- 3 months later
- Rank Corr's. > .9 at baseline and at 0 months btw treated & their matched controls

Rosenbaum 2010, p. 226



IC Conclusions

- Improved pain and urgency scores at 3 m for both treated and risk set controls.
- Frequency improved a bit for treated, not controls…

Rosenbaum 2010, p. 226

Second Example: Maturity at Discharge from Neonatal ICU

- Babies are kept in NICU until they have matured sufficiently to go home.
- Once babies look mature enough they stay around for “a few days” just to be sure.
- Does a longer stay in the NICU benefit those babies who receive it?

Silber et al. 2009, Rosenbaum 2010, Section 12.3

NICU Maturity at Discharge Study

- 1402 premature infants (gestational age ≤ 34 weeks) born at KP in Northern CA: 1998-2002
- Approach: **risk-set matching** using **time-dependent propensity scores** and an **optimal nonbipartite matching**
- Result: 1402 babies divided into 701 pairs so that one was an "early baby" and the other was a "late baby" but were similar at the postmenstrual age (to the day) that the "early baby" went home.

Silber et al. 2009, Rosenbaum 2010, Section 12.3

Available Covariate Information on These 1402 Premature Babies

- 20 fixed (not time-dependent) covariates
  - Gestational age at birth, weight at birth, gender
  - Baby's health history
  - Mom covariates
- 13 time-dependent (maturity) covariates
  - Postmenstrual age, Current weight
  - Propensity to discharge
  - Six maturity measures scored daily (1 if not achieved, 0 if achieved) – exponential smoothing applied to the binary variables…

Time-Dependent Propensity Score

- Fit a Cox proportional hazards model to describe the hazard of discharge on a particular day, using as predictors both the fixed and time-dependent variables
- The linear portion (log-hazard of discharge) is then taken as the time-dependent PS
- The PS will thus vary from day to day for each baby.

Lu (2005) describes the general approach

Optimal Non-Bipartite Matching

- Non-bipartite matching doesn't presuppose treated patients and controls
- Begin instead with a single group and form pairs so as to minimize the distance between paired subjects or sets of matched subjects
- Very flexible – permits looking at doses, for instance, or matching with several groups

Lu, Zanutto et al. 2002, Rosenbaum 2010, Chapter 11

Distance Matrix for Nonbipartite Matching

Table 11.1 A 6×6 distance matrix for nonbipartite matching for six units. Unlike treatment-control matching, every unit appears as both a row and a column of this distance matrix. The optimal nonbipartite matching (1,2), (3,6), (4,5) is shown in **bold** with a minimum total distance of 106+25+34 = 165.

ID	1	2	3	4	5	6
1	0	**106**	119	231	110	101
2	**106**	0	207	126	192	68
3	119	207	0	156	247	**25**
4	231	126	156	0	**34**	67
5	110	192	247	**34**	0	212
6	101	68	**25**	67	212	0

- So, in the NICU study, we'll have 1402 rows and 1402 columns (1402 babies), and we'll be looking for the 701 pairs of babies with minimum total distance, under constraints.

Rosenbaum 2010, Chapter 11

Doing Risk-Set, Optimal Bipartite Matching in this Study

- Proportional hazards model gives time-dependent PS for each baby
- Built a 1402 x 1402 distance matrix comparing babies to one another.
- If the baby in row i and in column j were discharged on same day, used ∞ distance.
- Otherwise, value in cell i, j describes distance between babies on the day earlier baby was discharged, with a caliper on the PS.

Silber et al. 2009, Rosenbaum 2010, Section 12.3

Table 12.2 Balance on fixed and time-dependent covariates after risk-set matching for 1402 premature babies in 701 matched pairs. Matching ensured that paired babies were similar on the day the early baby was discharged home from the neonatal intensive care unit, but the late baby was more mature (older, heavier) on the later day of discharge for the late baby. Of course, the fixed covariates are the same on both days; only the time-dependent covariates change.

Covariate Group	Covariate	Early Baby at Early Baby Discharge	Late Baby at Early Baby Discharge	Late Baby at Late Baby Discharge
	Number of Babies	701	701	701
Baby at Birth (fixed)	Gestational Age (weeks) at birth	31.1	31.1	31.1
	Weight at birth (grams)	1669	1686	1686
	SNAP-II 20 to 59	0.15	0.13	0.13
	SNAP-II 10 to 19	0.18	0.20	0.20
	SNAP-II 0 to 9	0.67	0.67	0.67
	Male Sex	0.51	0.52	0.52
Baby's Health History (fixed)	Bronchopulmonary Dysplasia	0.09	0.11	0.11
	Necrotizing Enterocolitis	0.01	0.01	0.01
	Retinopathy Stage ≥ 2	0.06	0.06	0.06
	Intraventricular Hemorrhage ≥ 3	0.02	0.01	0.01
Mom (fixed)	Maternal Age (years)	29.9	30.3	30.3
	Marital Status Single	0.24	0.24	0.24
	Other children = 0	0.40	0.37	0.37
	Other children = 1	0.34	0.37	0.37
	Other children ≥ 2	0.26	0.26	0.26
	Income $	59,517	59,460	59,460
	White Race	0.47	0.48	0.48
	Black	0.10	0.09	0.09
	Asian	0.20	0.23	0.23
	Hispanic	0.22	0.18	0.18
Baby's Time Dependent Variables	Postmenstrual Age (days)	247.4	247.4	250.9
	Propensity to discharge	0.67	0.64	1.33
	Apnea smoothed score	0.04	0.05	0.03
	Brady smoothed score	0.06	0.07	0.04
	Methyl smoothed score	0.04	0.03	0.02
	Oxygen smoothed score	0.11	0.11	0.07
	Gavage smoothed score	0.22	0.23	0.10
	Incubator smoothed score	0.15	0.15	0.08
	Combined maturity score	0.62	0.63	0.34
	Current weight	2153	2148	2231
	Current weight < 1700 grams	0.02	0.03	0.01
	$1700 \leq$ weight < 1800	0.06	0.06	0.02
	Current weight ≥ 1800 grams	0.92	0.91	0.97

Balance Achieved via Risk-Set Matching across the 701 pairs

Larger Version of Bottom Section on Next Slide

Rosenbaum 2010, p. 229

Balance Achieved on Time-Dependent Covariates

Covariate Group	Covariate	Early Baby at Early Baby Discharge	Late Baby at Early Baby Discharge	Late Baby at Late Baby Discharge
	Number of Babies	701	701	701
Baby's Time Dependent Variables	Postmenstrual Age (days)	247.4	247.4	250.9
	Propensity to discharge	0.67	0.64	1.33
	Apnea smoothed score	0.04	0.05	0.03
	Brady smoothed score	0.06	0.07	0.04
	Methyl smoothed score	0.04	0.03	0.02
	Oxygen smoothed score	0.11	0.11	0.07
	Gavage smoothed score	0.22	0.23	0.10
	Incubator smoothed score	0.15	0.15	0.08
	Combined maturity score	0.62	0.63	0.34
	Current weight	2153	2148	2231
	Current weight $<$ 1700 grams	0.02	0.03	0.01
	1700 $\leq$ weight $<$ 1800	0.06	0.06	0.02
	Current weight $\geq$ 1800 grams	0.92	0.91	0.97

Rosenbaum 2010, p. 229

Absolute Standardized Differences Across 20 Fixed & 13 Time-Dependent Covariates for 701 pairs

Quantile	Min	25	50	75	Max
Fixed	0.00	0.01	0.04	0.06	0.09
Time-Dependent (at early discharge)	0.00	0.01	0.02	0.06	0.09
Time-dependent (at own discharge)	0.09	0.16	0.19	0.34	0.75

- Babies are quite similar on the day the early baby was discharged – but quite different on the day of their own discharge

Silber et al. 2009, Rosenbaum 2010, Section 12.3

So … Did the "late babies" benefit from a few more days to grow in the NICU?

- As it turns out, the "early" and "late" babies had similar experiences after discharge.
- No real strong indications of either benefit or harm for those with the extra time in NICU
- Not that NICUs aren't expensive places to be…

Silber et al. 2009, Rosenbaum 2010, Section 12.3

SOME CLOSING THOUGHTS

A Few Advantages of Propensity Methodology

- Results can be persuasive even to audiences with limited statistical training.
- Though estimating the PS requires some care, the comparability of treated and control patients can be verified simply.
- PS methods address selection bias well.
- PS methods may be combined with other sorts of adjustments.

Strategic Issues

- How can we make our investigations compelling to our intended audience?
- Why is this hard?
  - Audience is not focused on statistical techniques
  - Audience may have limited training in statistics
- Why is this important?
  - Who makes key policy decisions?
  - Who needs to be convinced by the evidence?

Strategic Issues in Observational Studies

- Design observational studies
  - Exert as much experimental control as possible, carefully consider the selection process, and anticipate hidden biases
- Focus on simple comparisons
  - Increase impact of results on consumers
- Compare subjects who looked comparable prior to treatment
- Use sensitivity analyses to delimit discussions of hidden biases due to unobserved covariates

Rosenbaum (2002)

Some Cautions and Limitations

- Hidden Bias: Beware unmeasured covariates which affect outcomes and/or assignment.
  - Sensitivity Analysis helps quantify the problem
- This is a reasonable method with fairly large samples.
  - Matching vs. stratification vs. adjustment methods
- Options narrow as an investigation proceeds.
  - Sadly, though OS work cries out for design, we're often working with secondary data, where we have fewer options

What should always be done in an OS … and often isn't?

1. Collect data so as to be able to model selection
2. Demonstrate need for adjustment - selection bias
3. Carefully record intervention time - adjust only for things present before or at time of intervention.
4. Ensure baseline characteristic overlap [comparability]
5. Check baseline characteristic balance after adjustment
6. Specify relevant post-adjustment population with care
7. Estimate treatment effect in light of adjustment
8. Estimate sensitivity of results to potential hidden bias

How Can We Avoid Being Misled by Observational Studies?

1. What differentiates an observational study from a randomized controlled trial?
   - One key element: potential for selection bias.
2. What is selection bias, and why should I care about it?
   - Baseline characteristics of comparison groups are different in ways that affect the outcome.
3. What can be done to deal with selection bias in observational studies?
   - Propensity score methods for overt bias.
   - Sensitivity analyses to deal with hidden bias.

Acknowledgments

- Center for Health Care Research and Policy, including …
  - Randall D. Cebul, MD
  - Neal V. Dawson, MD
  - Charles L. Thomas
  - Steven A. Lewis, MS
  - Scott S. Husak
  - Alfred Connors, MD
  - Irene Katzan, MD MS
  - Pat Murray, MD MS
  - Mark Votruba, PhD
- Students in CRSP 500
- Paul Rosenbaum, PhD
- Donald Rubin, PhD
- Therese Stukel, PhD
- Sharon-Lise Normand, PhD
- Rickey Mehta
- Ralph D'Agostino, Jr., PhD
- Ben Hansen, PhD
- Michael Posner, PhD
- Ali Ahmed, MD
- SMDM
- ASA Cleveland Chapter
- ASA Health Policy Statistics Section

Thank you very much.

Thomas.Love@case.edu