Issues and options in designing a study

Jim Davis

jamesdav@hawaii.edu

Biostatistics Core

Issues in study design

How many people are needed?

Create a codebook

Field #	Variable name	Description	Value Labels	Validation Rules	Comments
1	diab	Diabetes	1=Has Diabetes 2=No Diabetes	Single value	-9 if unchecked
2	diab_age	Age at Diagnosis of Diabetes	1=before 20 yrs old 2=20-29 yrs old 3=30-39 yrs old 4=40-49 yrs old 5=50 yrs or older 6=don't know	Single Value	-9 if unchecked
3	HTN	Blood Pressure	1=High Blood Pressure 2=NO High Blood Pressure	Single Value	-9 if unchecked
4	HTN_age	Age at Diagnosis of High Blood Pressure	1=before 20 yrs old 2=20-29 yrs old 3=30-39 yrs old 4=40-49 yrs old 5=50 yrs or older 6=don't know	Single Value	-9 if unchecked
5	Pct_Hwn	Percent Hawaiian	___ ___ ___	Value 0 to 100	-9 if missing or unknown

Choice of variable names

Bad Variable Names

Patient ID	Score %	Ht (in.)	Test-Elisa	12 You had treatmnt of coronary vessels w/balloon angioplasty?
1	83	61	Pos	1
2	56	66	Neg	0
3	77	58	Neg	0
4	67	64	Pos	1

Good Variable Names

PatientID	Score_Pct	Ht_In	TestElisa	balloon_angioplasty
1	83	61	Pos	1
2	56	66	Neg	0
3	77	58	Neg	0
4	67	64	Pos	1

Simple rules for variable names

- Begin names with a letter
- Only use letters, numbers, and underscores
- Put numbers at the end

Letters aren’t numbers

Weight	Education
177	1
146	3
183	Refused
160	2
135	3
> 300	2

Numbers aren't dates

DateOfBirth	DxDate
January 5, 1983	February 9, 2011
0	June 15, 2011
July 13, 1972	Missing
March 3, 1975	December 7, 2011
August 24, 1978	April 22, 2011
99	September 4, 2011

And neither is text.

Limit: one data element per cell

PatientID	Conditions
1	History of Psychiatric Disorders; Chronic Drug Abuse
2	Hypertension
3	Pre-existing Anemia; Rheumatoid Arthritis
4	Hypertension; Coumadin Therapy; History of Psychiatric Disorder
5	Hypertension; Chronic Dementia
6	History of Psychiatric Disorder; Asthma
7	None
8	History of Cardiac Surgery; Coronary Artery Disease
9	Hypertension; Peptic Ulcer Disease
10	Hypertension; Coronary Artery Disease

Consider a long format

PatientID	Conditions
1	History of Psychiatric Disorders
1	Chronic Drug Abuse
2	Hypertension
3	Pre-existing Anemia
3	Rheumatoid Arthritis
4	Hypertension
4	Coumadin Therapy
4	History of Psychiatric Disorder
5	Hypertension
5	Chronic Dementia
6	History of Psychiatric Disorder
6	Asthma

Don’t hide columns from other users

Complete data

Weight	Height	BMI	Age	Gender
36.9	1.47	17.19	11.42	F
71.3	1.56	29.30	11.50	F
76.8	1.58	30.76	11.50	F
43.3	1.40	22.25	11.58	F
35.1	1.41	17.78	11.58	F
75.3	1.50	33.47	11.58	F
62.5	1.51	27.59	11.58	F
41.4	1.47	19.16	11.67	F
33.1	1.47	15.32	11.75	F

Hidden columns

Weight	Height	Gender
36.9	1.47	F
71.3	1.56	F
76.8	1.58	F
43.3	1.40	F
35.1	1.41	F
75.3	1.50	F
62.5	1.51	F
41.4	1.47	F
33.1	1.47	F

Don’t do analyses with data you share

id	score1	score2
1	3	0
2	5	0
3	5	0
Average	3.31	-0.10
Standard Deviation	1.11	0.77
4	3	0
5	1.5	1
6	2.5	1
7	5	0
8	4	0
9	2.5	0
10	5	0
Average	3.50	0.18
Standard Dev	1.16	0.60

Pass clean data files

- No Pie or other charts
- No analyses on the spreadsheet
- No calculated columns

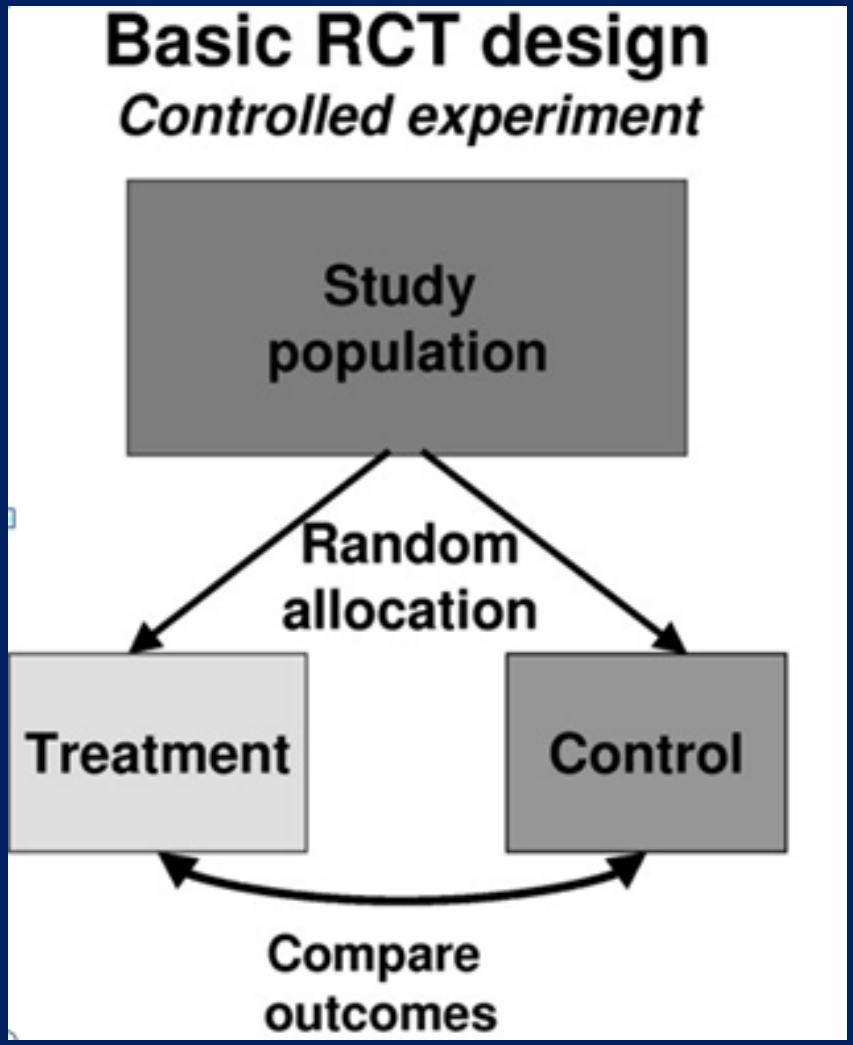
Have results before making plans to present them

Good practice

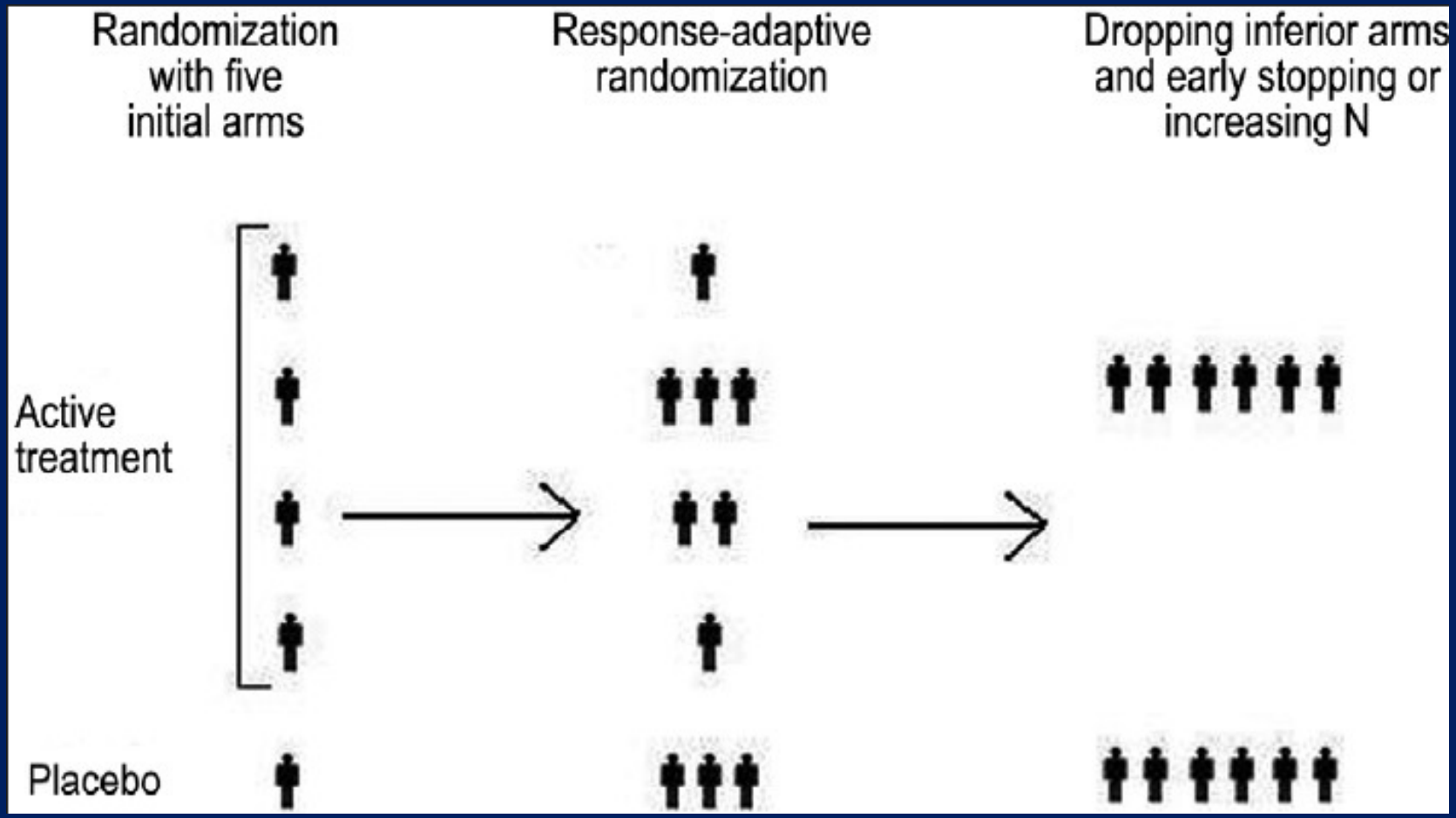
- Know the sample size you need
- Create a codebook
- Follow the rules for variable names
- Keep one data type per column
- Keep one data element per cell
- Don’t hide columns if sharing data
- Share clean (unmodified) data files
- Enjoy (don’t rush) your analyses

Options in study design

Experimental design



Adaptive designs

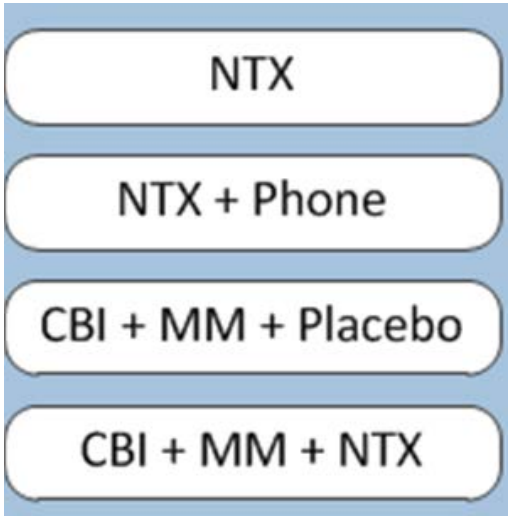


Possible adaptive decisions

- Modify of decide sample size
- Change the dose
- Alter timing of scheduled visits
- Change inclusion criteria
- Modification randomization ratios to treatments

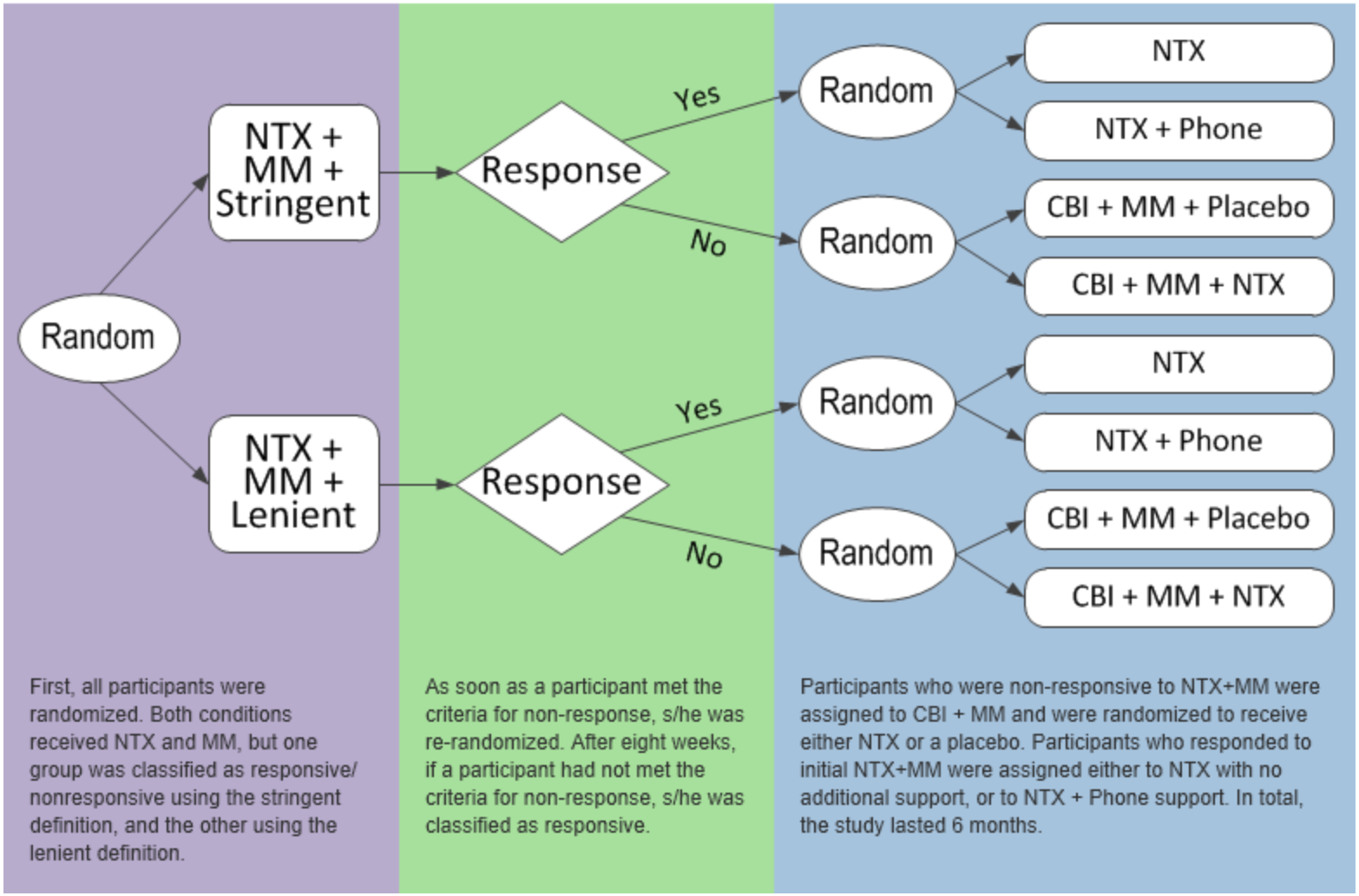
Randomize to combinations of interventions

Trial to reduce alcohol dependence



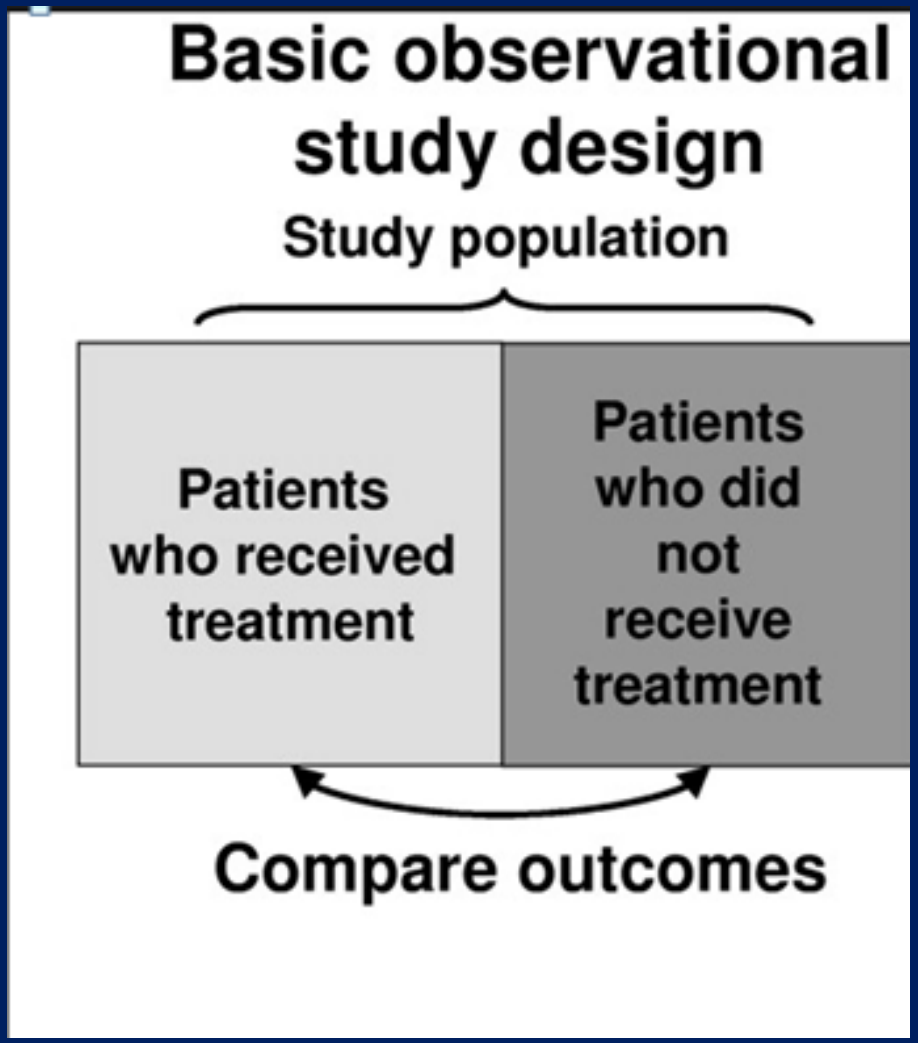
Ntx = naltrexone (medication)
MM = Medical management (face-to-face)
CBI = Combined behavioral intervention
Phone = Telephone disease management

Complete design



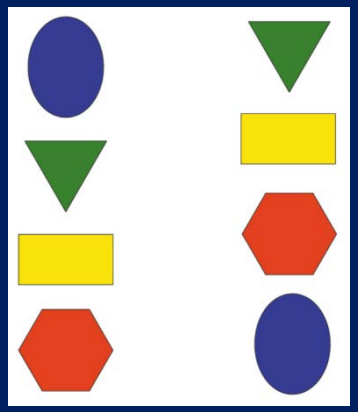
http://methodology.psu.edu/ra/adap-treat-strat/example

Observational studies



Control of confounding

Matching



Versus

Statistical adjustment

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

$$i = 1, 2, \ldots, n$$

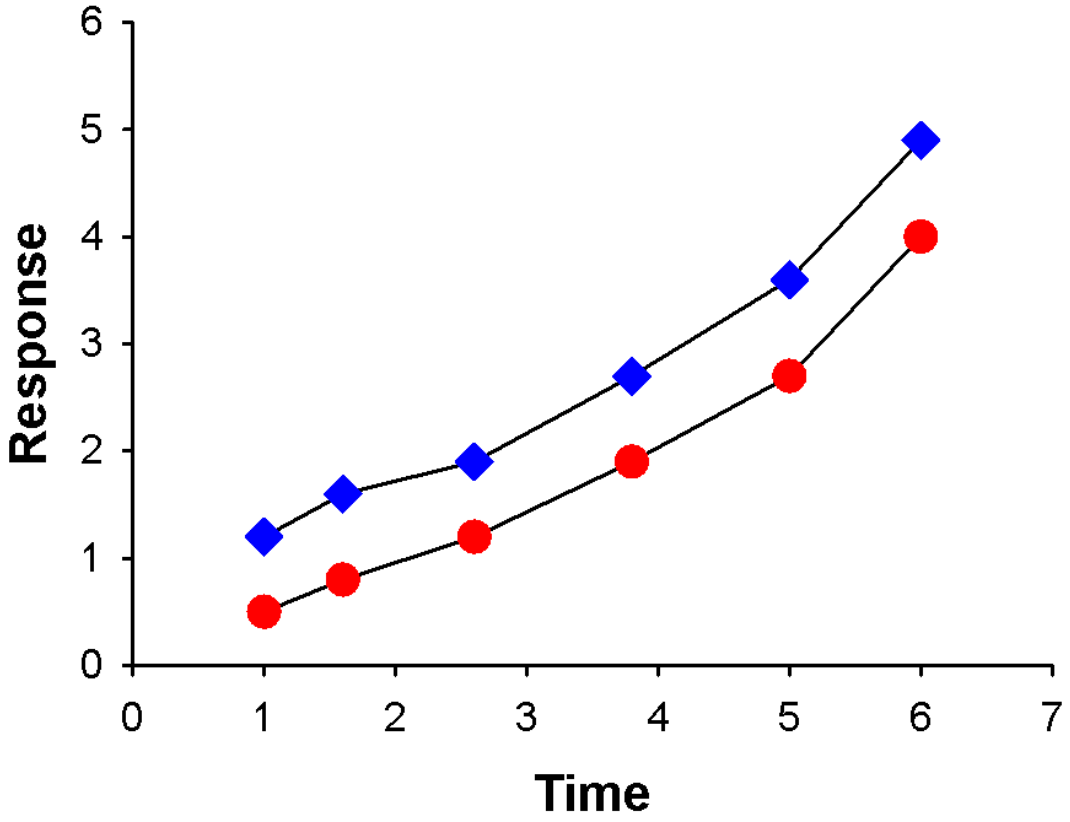
When matching may be useful

- When data for matching are available at the start
- When you want to collect additional data
- When having more intuitive results is important

Reasons to adjust

- When obtaining data to match on is expensive or difficult to obtain
- When you want to understand the importance of the factors you might match on
- When you want to understand complex interrelationships

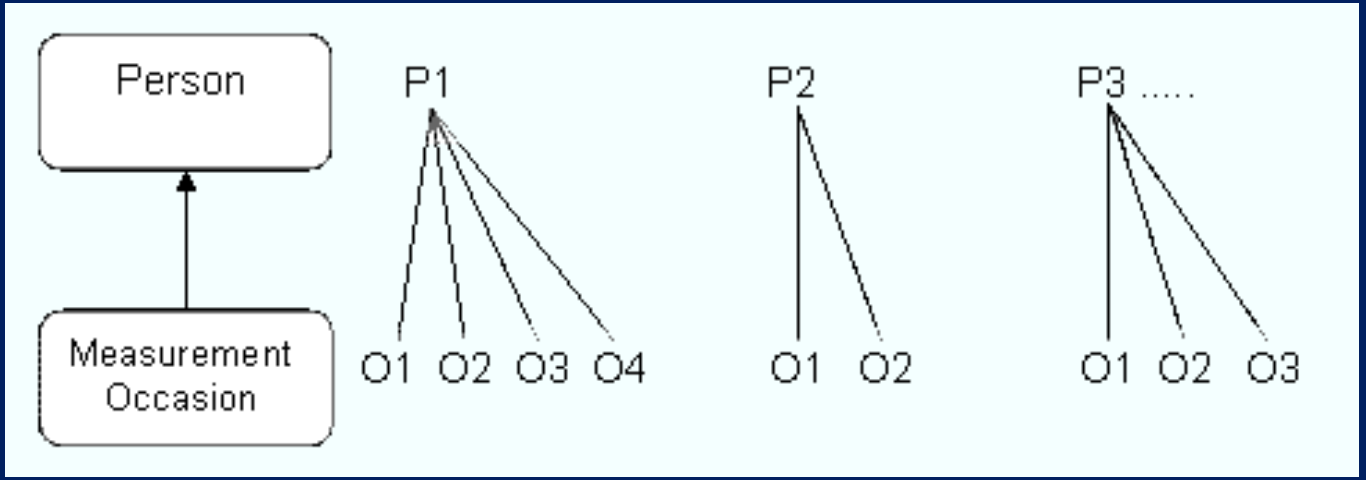
Longitudinal design



Decisions in longitudinal studies

Number and timing of measurements

- 2 to study change from a start to an end time
- 3 to study linear change
- 4 or more to study the pattern of change

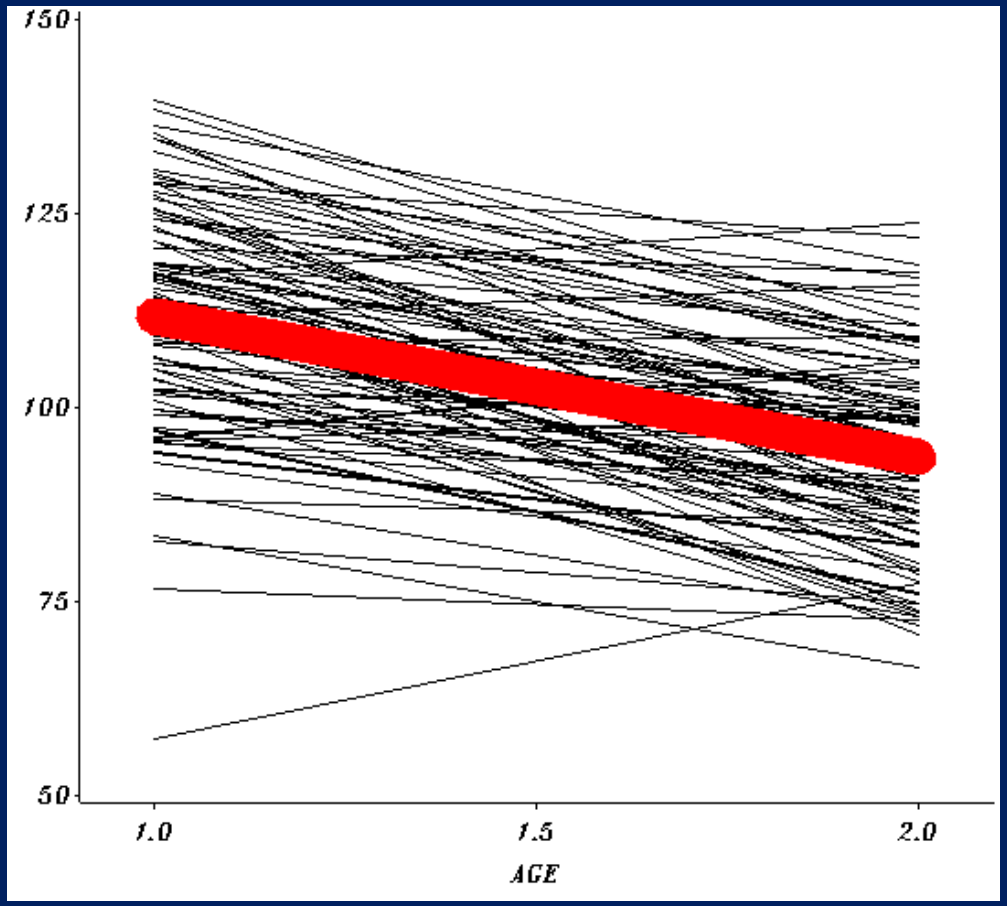


Time dependent variables

Participant characteristics that change over time

- Choice of which ones to measure
- Choice of how often to measure them

Interested in rate of change or variations in change?



Individual rates of change may differ substantially from the mean

Multilevel design

People within communities

- Community resources as well as the individual may be important

Patients in a physician practices

- Both patients and physicians may influence the outcome

Multilevel design

Outcome: Length of hospital stay

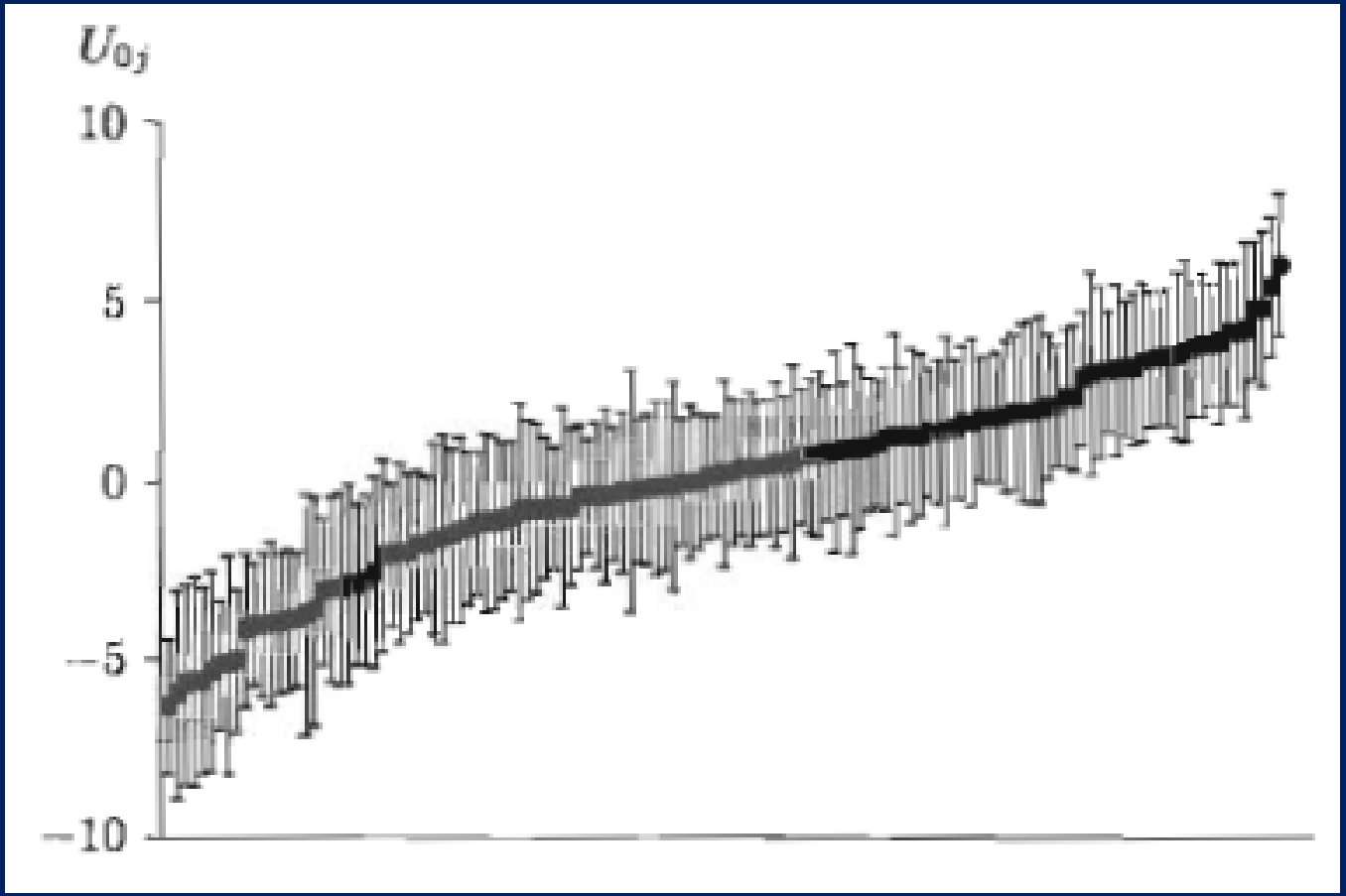
Patient predictors:

Age, gender, clinical history

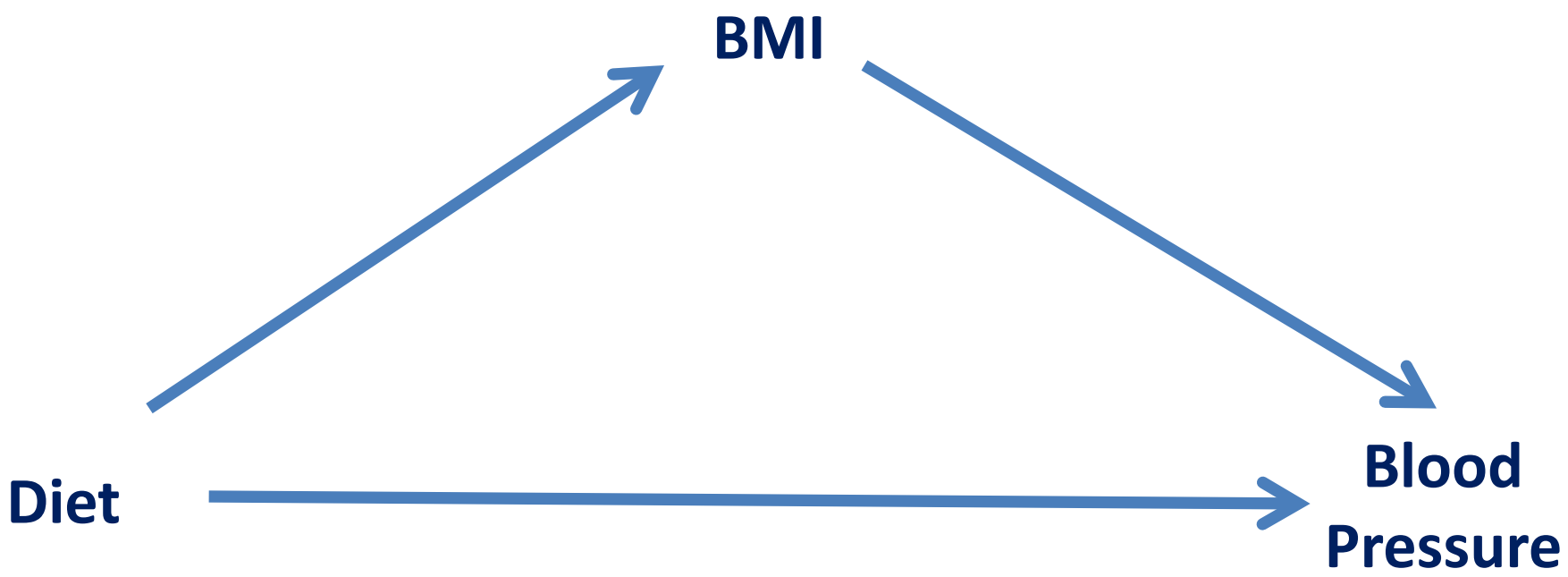
Hospital predictors:

Size of hospital, number of Patients treated

Caterpillar plot of hospitals

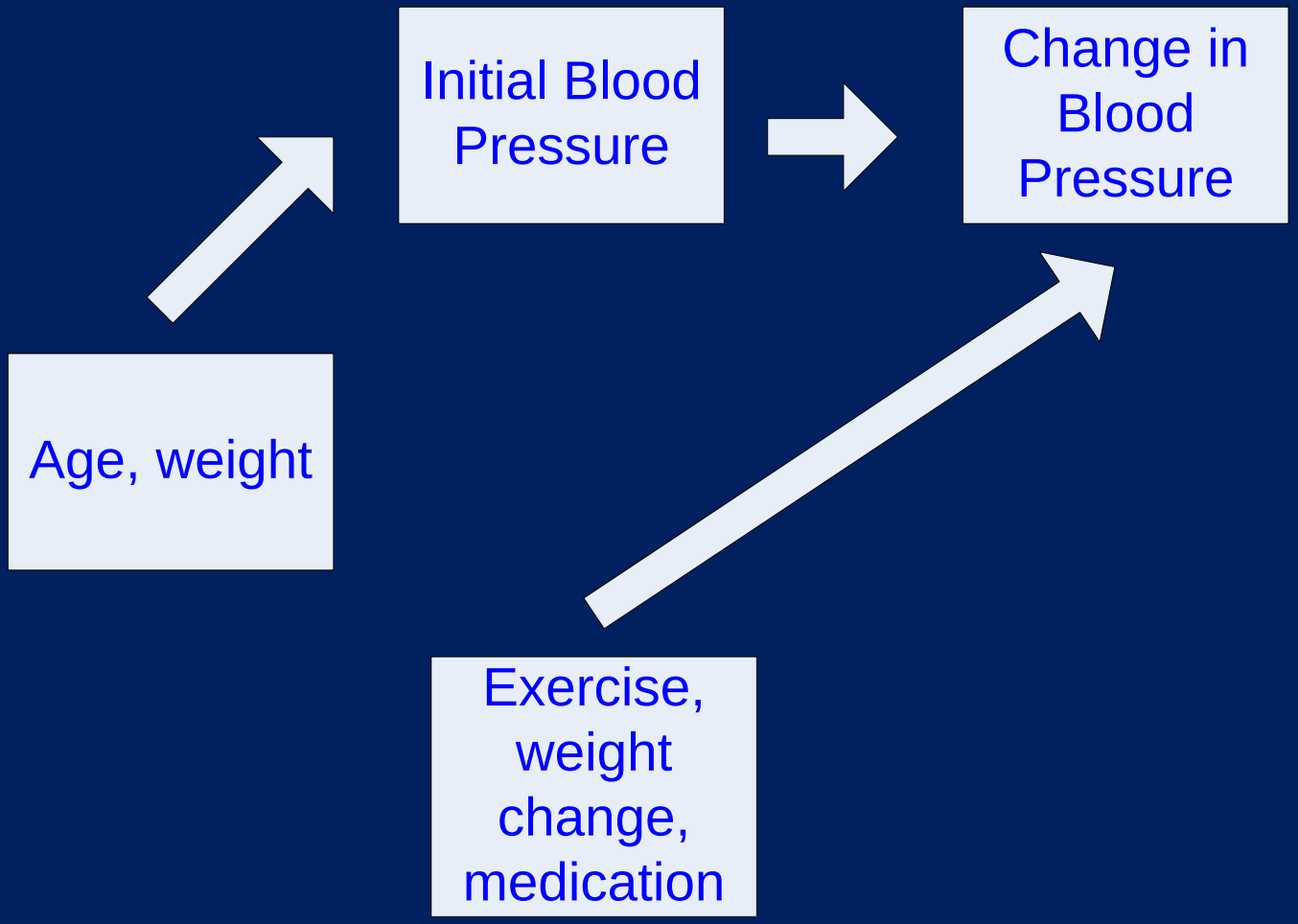


Mediation analysis



Timing of measurements again is crucial

Path analysis



Within person design

- Study design for outcomes that are affected by changes within a person over time
- Don’t require a comparison group
- Not biased by stable participant characteristics

Intensive within person designs

Mobile technology

Types of designs

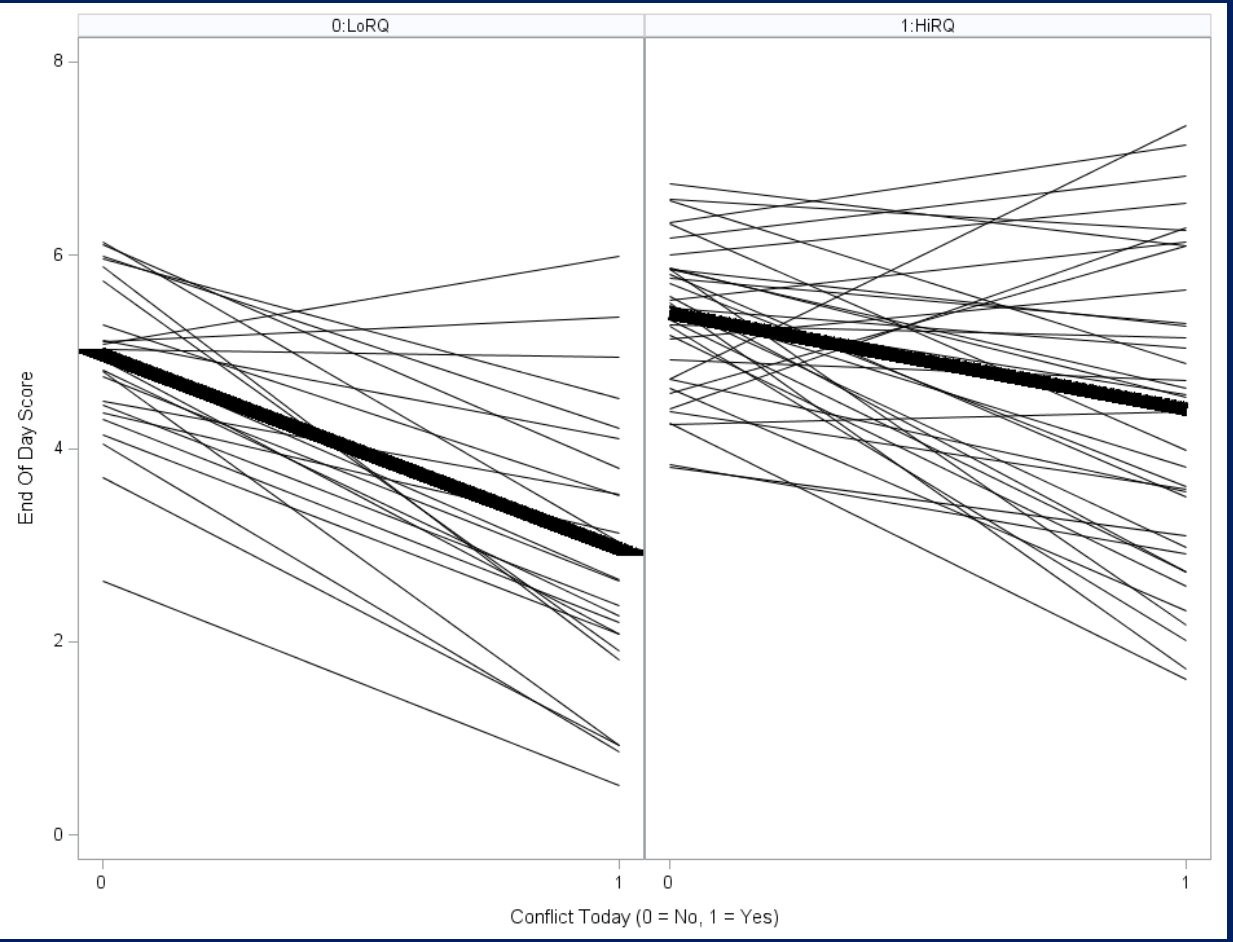
- Participants record experiences at regular predetermined intervals of time
- Participants record experiences in response to a signal from the researcher
- Participants report every time a predetermined event has taken place

Multiple device possibilities

- Self report as by text messaging
- Physiological indices (e.g., heart rate)
- Environmental indices (e.g., sounds, temperature, photographs)
- Spatial data (e.g., GPS information

End of day score by conflict today and relationship quality

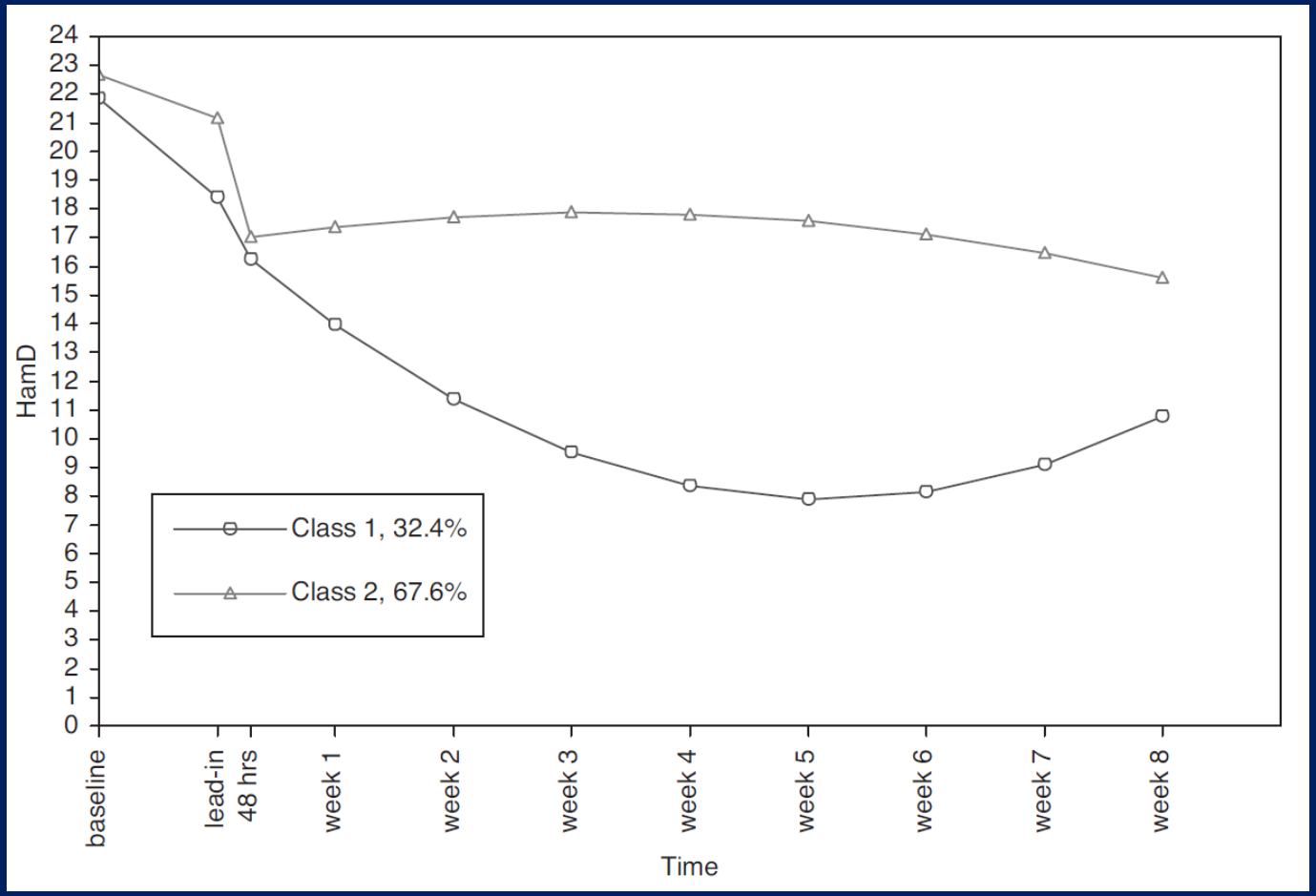
Low quality High quality



Designs for classification

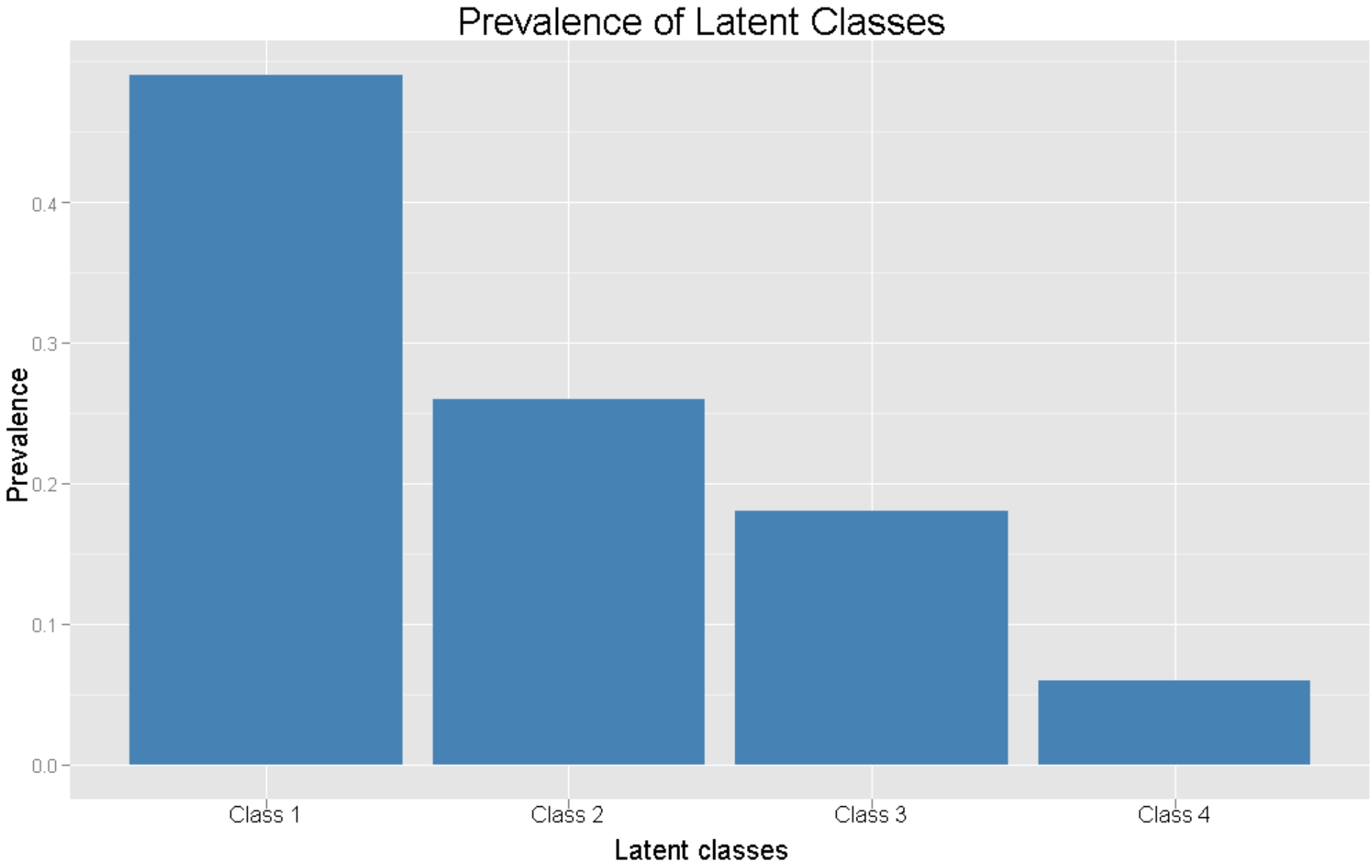
- Can separate people into groups based on their rates of change over time
- Can separate people into classes based upon their characteristics

Two class solution for placebo group in a study of depression



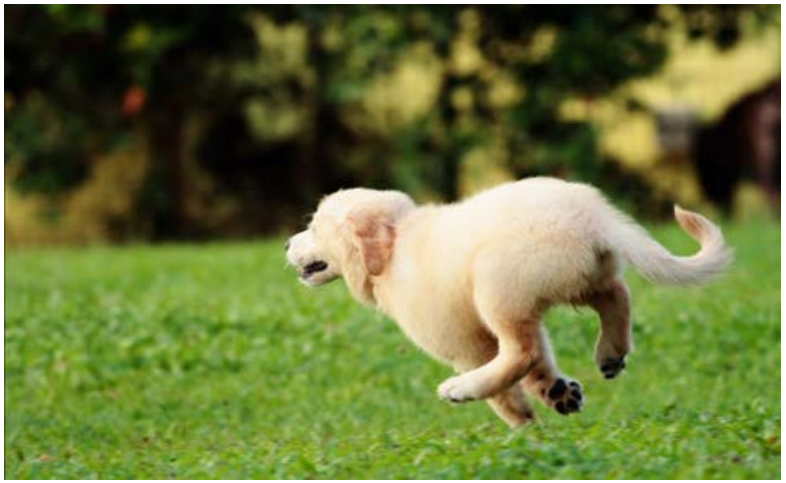
Classifying students in the tenth grade

	Latent Class			
	1	2	3	4
	Non-/Mild Delinquents	Verbal Antagonists	Shoplifters	General Delinquents
Question				
Lied to parents	0.33	0.81	0.78	0.89
Publicly loud/rowdy	0.2	0.82	0.62	1.00
Damaged property	0.01	0.25	0.25	0.89
Stolen something from store	0.03	0.02	0.92	0.88
Stolen something worth < $50	0.00	0.03	0.73	0.88
Taken part in group fight	0.04	0.31	0.24	0.44


Prevalence of Latent Classes
Prevalence
0.0
0.1
0.2
0.3
0.4
Class 1
Class 2
Class 3
Class 4
Latent classes

Changes in classification

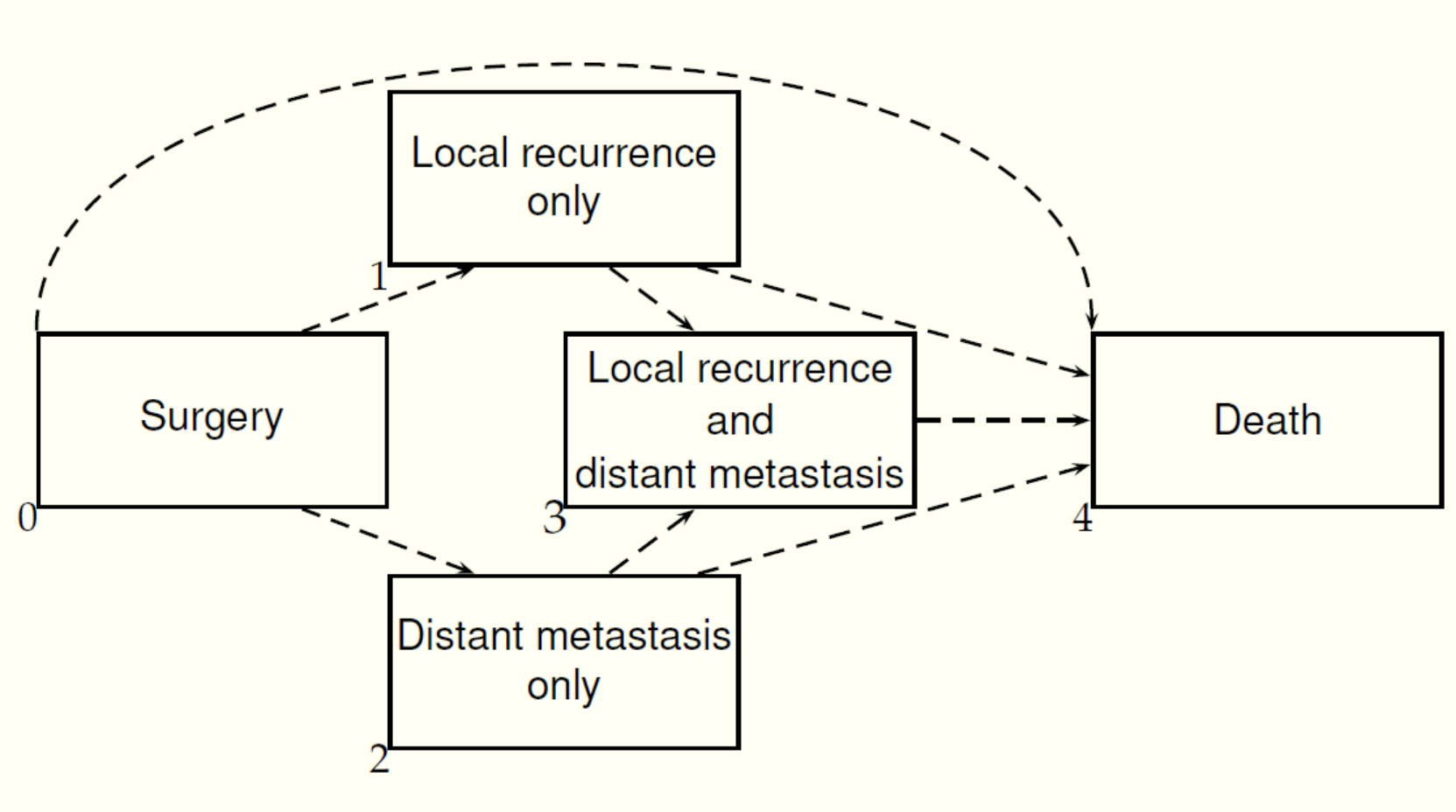
Movers and Stayers



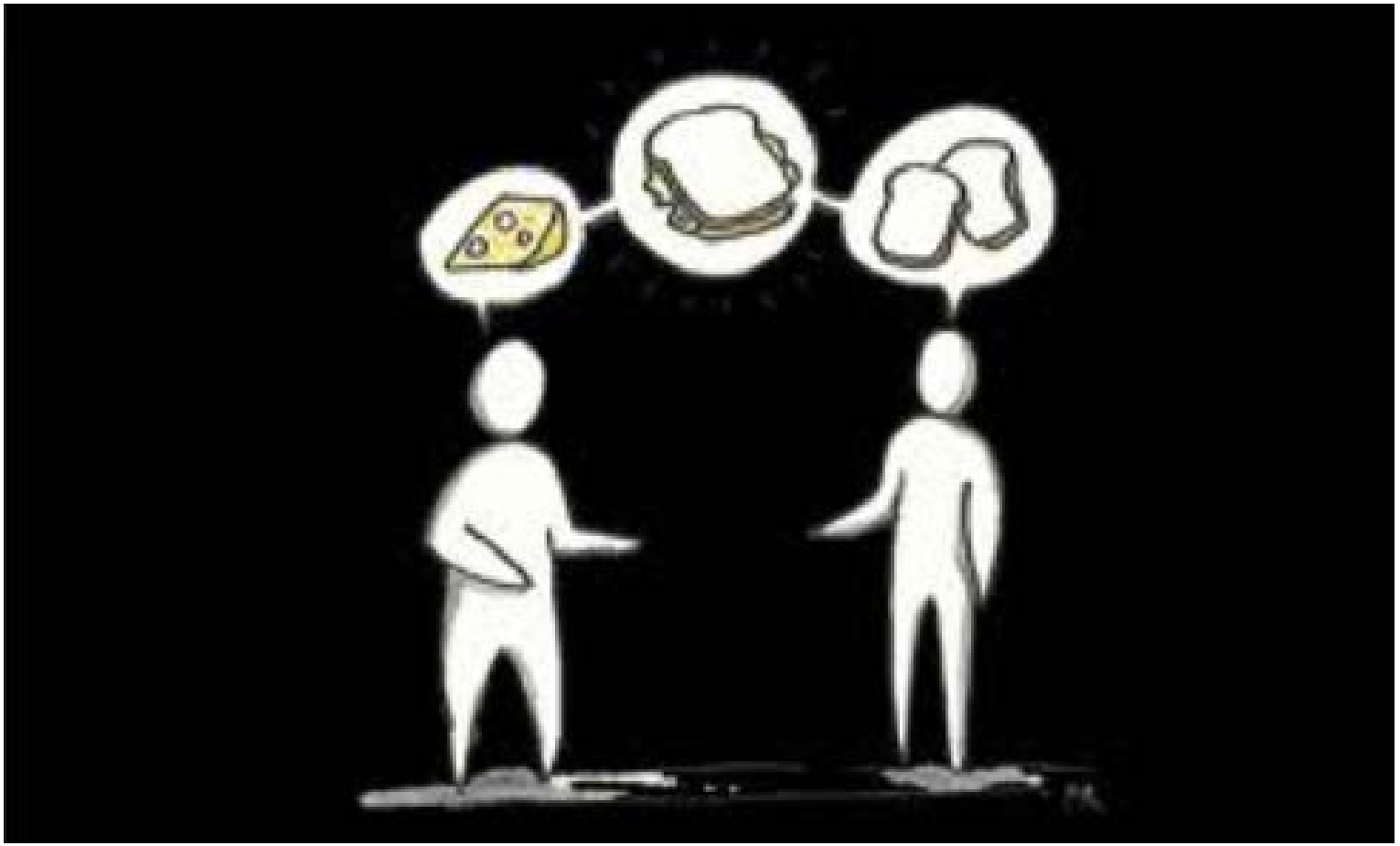
Mover

Stayer

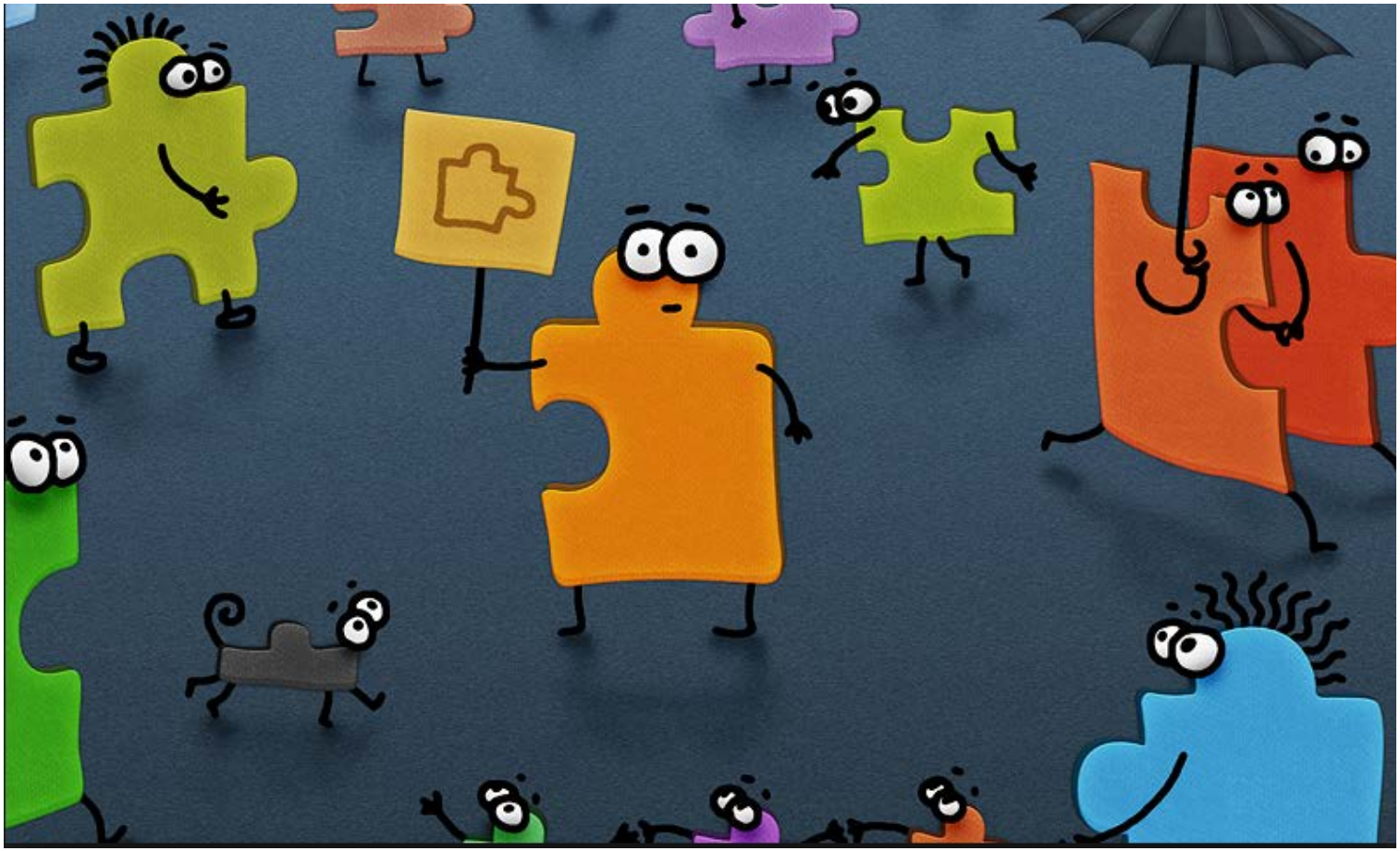
Changes across multiple states



Ideas and methods



Missing a piece?



Include a statistician on your team from the start