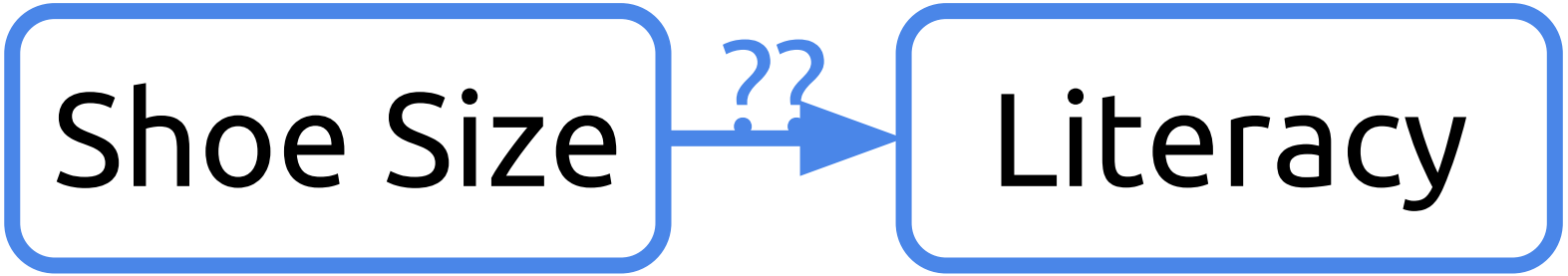
Study design, batch effects, and confounding

Jeff Leek

@jtleek

What is confounding?


Small shoes
Not literate
Big shoes
Somewhat literate

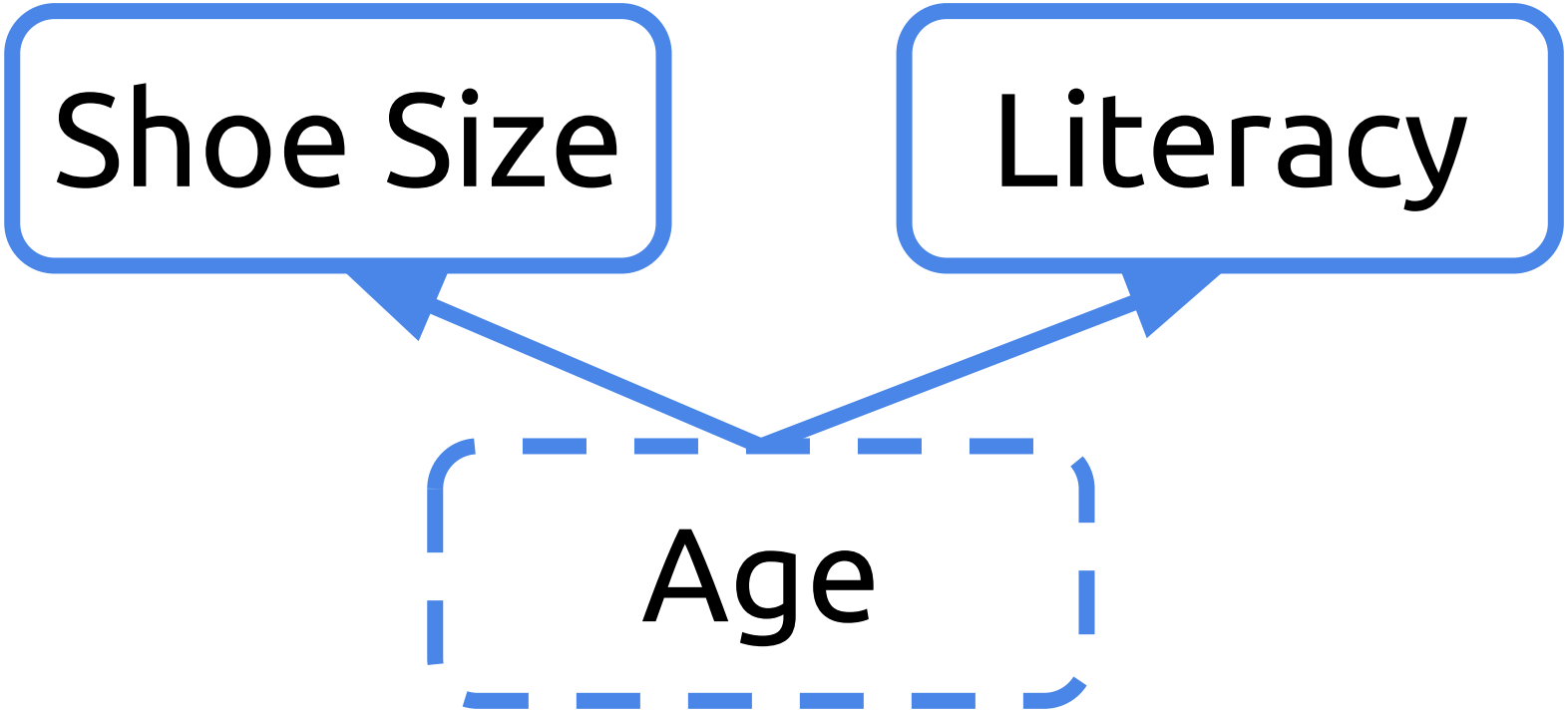

Shoe Size
??
Literacy

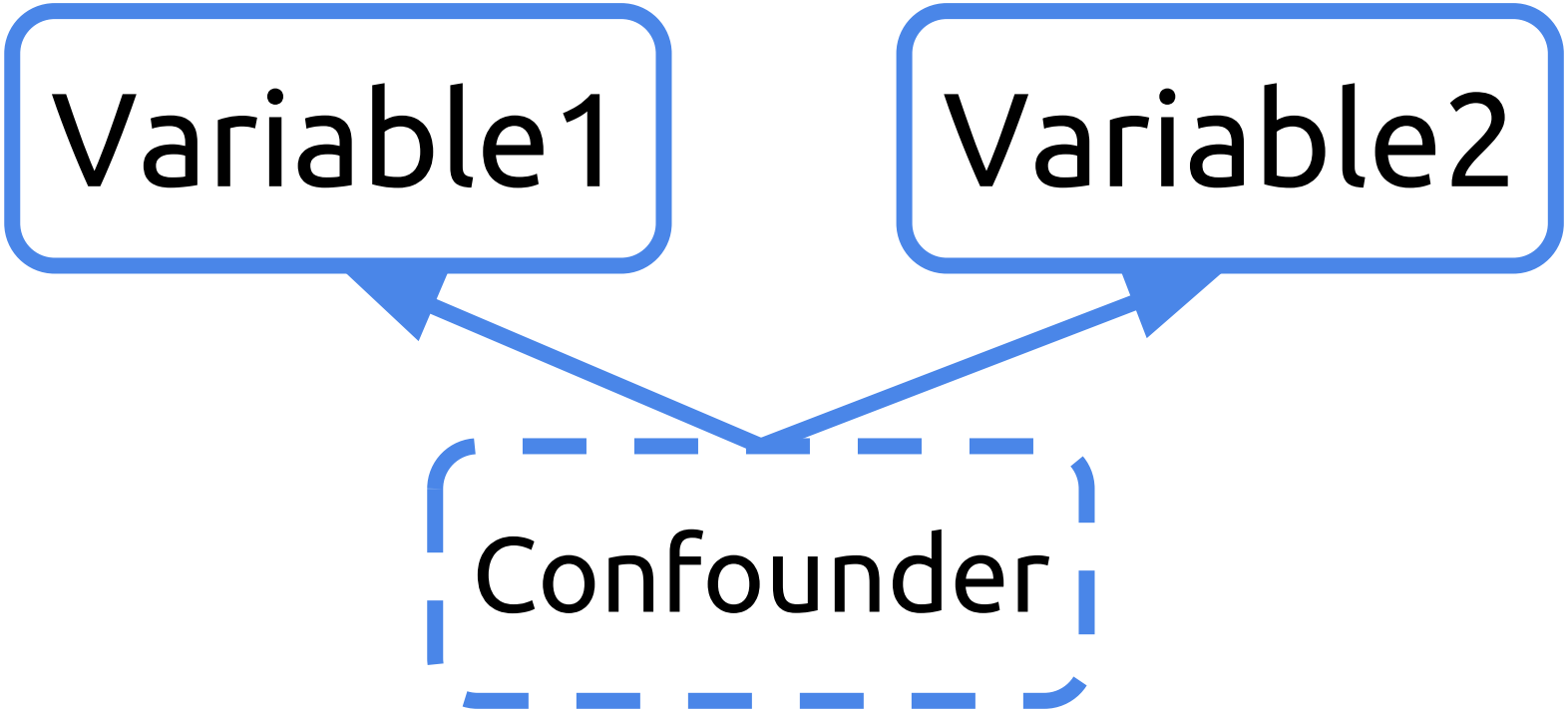

Small shoes
Not literate
Young
Big shoes
Somewhat literate
Middle aged

Shoe Size

Literacy

Age

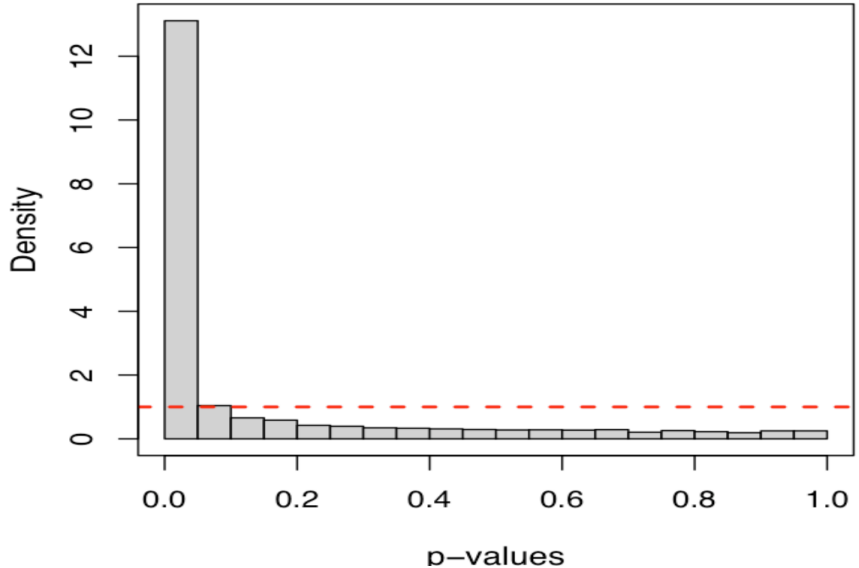

Shoe Size
Literacy
Age


Variable1
Variable2
Confounder

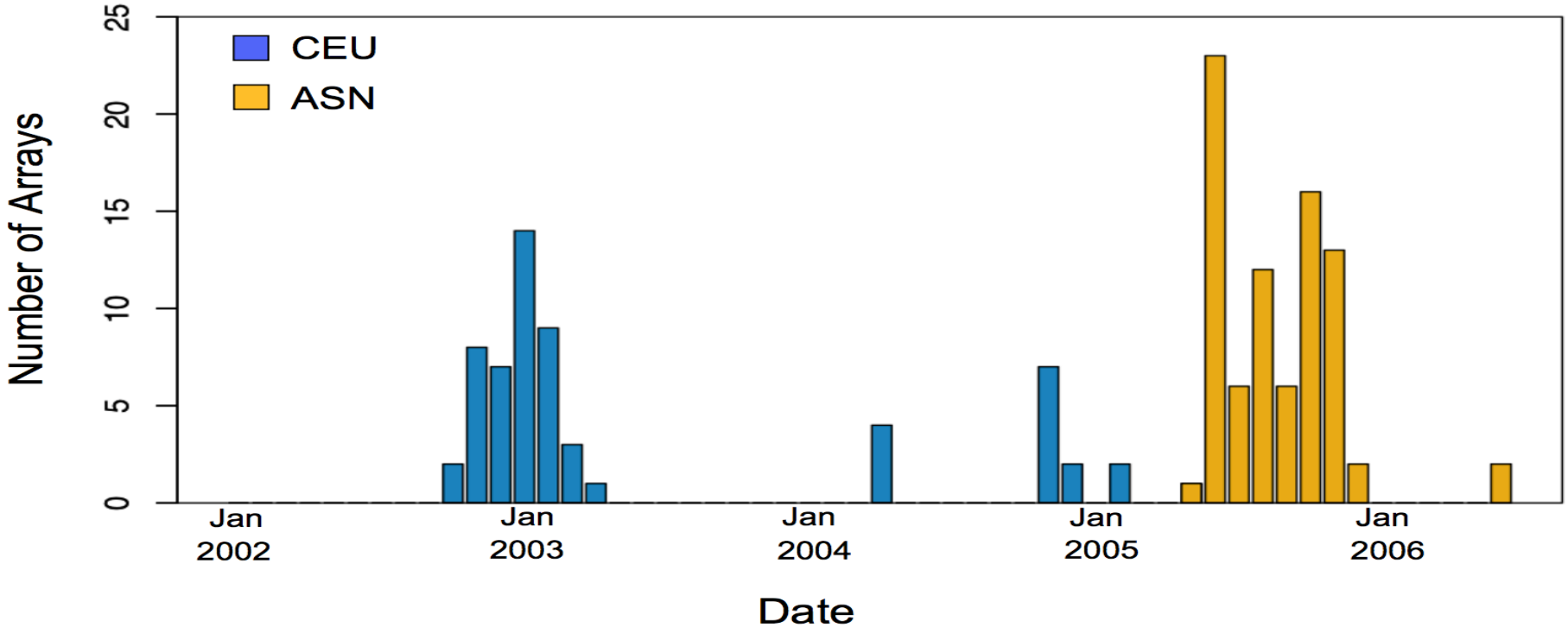
Most common confounder: batch effects

Common genetic variants account for differences in gene expression among ethnic groups

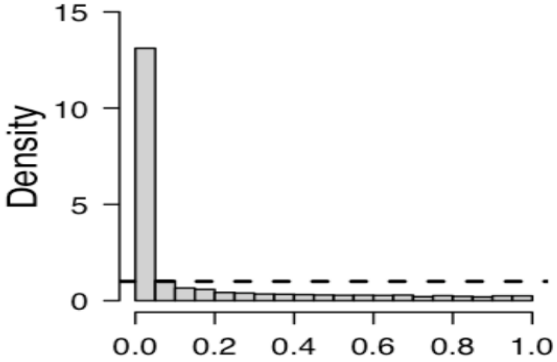
Richard S Spielman[1], Laurel A Bastone[2], Joshua T Burdick[3], Michael Morley[3], Warren J Ewens[4] & Vivian G Cheung[1,3,5]



78% of genes differentially expressed

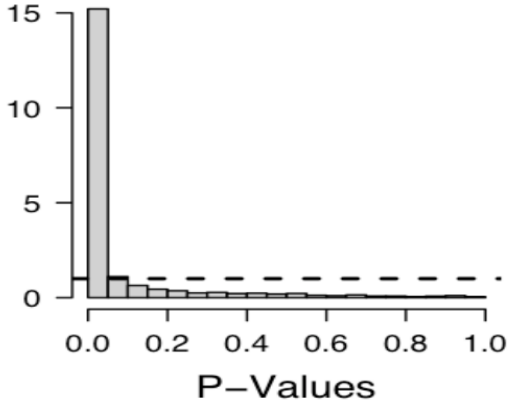

CEU
ASN
Number of Arrays
0
5
10
15
20
25
Jan
2002
Jan
2003
Jan
2004
Jan
2005
Jan
2006
Date

Between Population



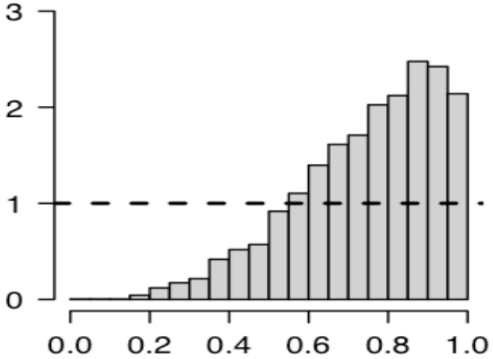
78% of genes estimated to be differentially

Between Years



96% of genes estimated to be differentially

Between Populations, Adjusting For Years



0% of genes estimated to be differentially

Extremely common

Article Views
- Abstract
- Full Text (PDF)
- Supporting Online Material

VERSION HISTORY
- science.1190532v4 (most recent)
- science.1190532v3
- science.1190532v2
- science.1190532v1

Published Online July 1 2010

Science DOI: 10.1126/science.1190532

< *Science* Express Index

REPORT

Genetic Signatures of Exceptional Longevity in Humans

Paola Sebastiani[1,*], Nadia Solovieff[1], Annibale Puca[2], Stephen W. Hartley[1], Efthymia Melista[3], Stacy Andersen[4], Daniel A. Dworkis[3], Jemma B. Wilk[5], Richard H. Myers[5], Martin H. Steinberg[6], Monty Montano[3], Clinton T. Baldwin[6,7] and Thomas T. Perls[4,*]

± Author Affiliations

*To whom correspondence should be addressed. E-mail: sebas@bu.edu (P.S.); thperls@bu.edu (T.H.P.)

Article Views
- Abstract
- Full Text (PDF)
- Supporting Online Material

VERSION HISTORY
- science.1190532v4 (most recent)
- science.1190532v3
- science.1190532v2
- science.1190532v1

This article has been retracted

Published Online July 1 2010

Science DOI: 10.1126/science.1190532

< *Science* Express Index

REPORT

Genetic Signatures of Exceptional Longevity in Humans

Paola Sebastiani[1,*], Nadia Solovieff[1], Annibale Puca[2], Stephen W. Hartley[1], Efthymia Melista[3], Stacy Andersen[4], Daniel A. Dworkis[3], Jemma B. Wilk[5], Richard H. Myers[5], Martin H. Steinberg[6], Monty Montano[3], Clinton T. Baldwin[6,7] and Thomas T. Perls[4,*]

± Author Affiliations

*To whom correspondence should be addressed. E-mail: sebas@bu.edu (P.S.); thperls@bu.edu (T.H.P.)

THE LANCET
Volume 359, Issue 9306, 16 February 2002, Pages 572-577

doi:10.1016/S0140-6736(02)07746-2 | How to Cite or Link Using DOI
Permissions & Reprints

Fast track — Mechanisms of Disease

Use of proteomic patterns in serum to identify ovarian cancer

How Bright Promise in Cancer Testing Fell Apart

Michael Stravato for The New York Times

Keith Baggerly, left, and Kevin Coombes, statisticians at M. D. Anderson Cancer Center, found flaws in research on tumors.

Perspective

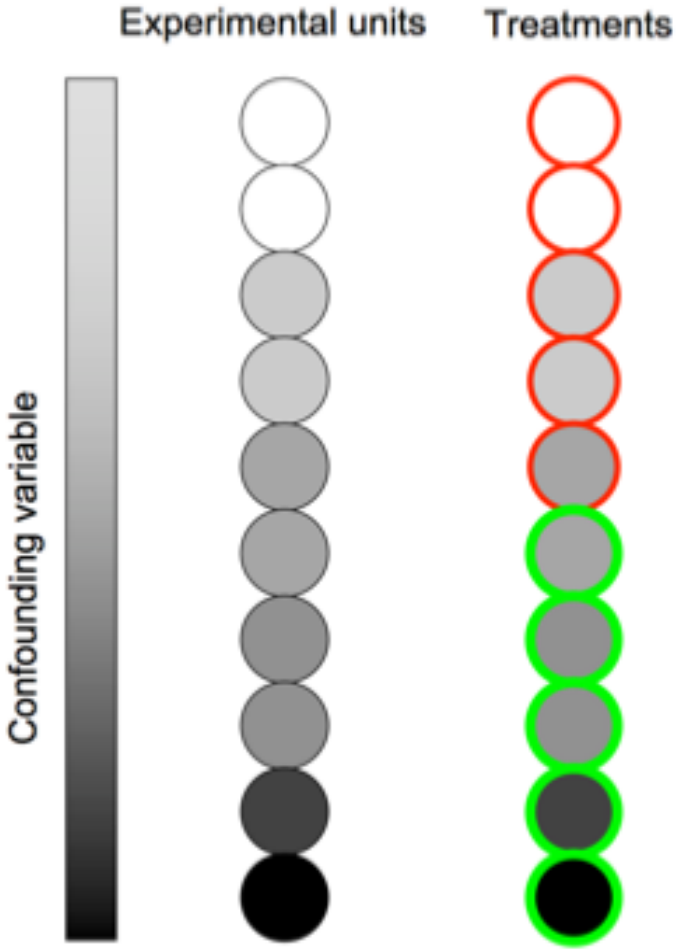
Nature Reviews Genetics **11**, 733–739 (1 October 2010) | doi:10.1038/nr

Tackling the widespread and critical impact of batch effects in high-throughput data

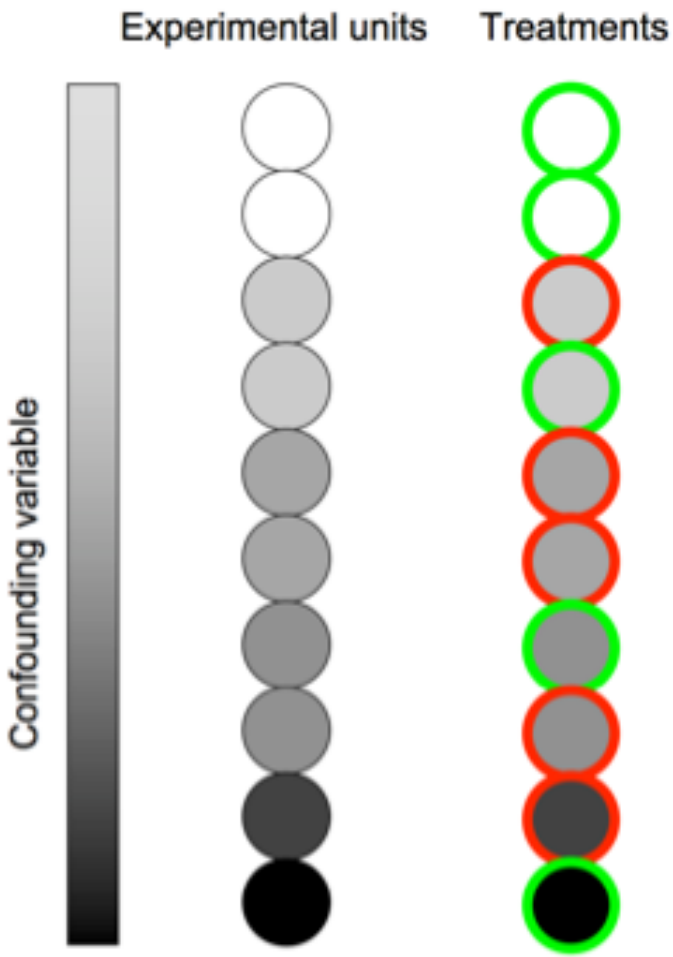
Jeffrey T. Leek , Robert B. Scharpf , Héctor Corrada Bravo , David Simcha , Benjamin Langmead , W. Evan Johnson , Donald Geman , Keith Baggerly & Rafael A. Irizarry

High-throughput technologies are widely used, for example to assay genetic variants, gene and protein expression, and epigenetic modifications. One often overlooked complication with such studies is batch effects, which occur because measurements are affected by laboratory conditions, reagent lots and personnel differences. This becomes a major problem when batch effects are correlated with an outcome of interest and lead to incorrect conclusions. Using both published studies and our own analyses, we argue that batch effects (as well as other technical and biological artefacts) are widespread and critical to address. We review experimental and computational approaches for doing so.

Randomization



Without randomization, the confounding variable differs among treatments



With randomization, the confounding variable does not differ among treatments

Stratification example

Example:

- 20 males and 20 females.
- Half to be treated; the other half left untreated.
- Can only work with 4 individuals per day.

Question:

How to assign individuals to treatment groups and to days?

A bad design

Week One					Week Two				
M	Tu	W	Th	F	M	Tu	W	Th	F
C	C	C	C	C	T	T	T	T	T
C	C	C	C	C	T	T	T	T	T
C	C	C	C	C	T	T	T	T	T
C	C	C	C	C	T	T	T	T	T

T = treated, C = control, pink = female, blue = male

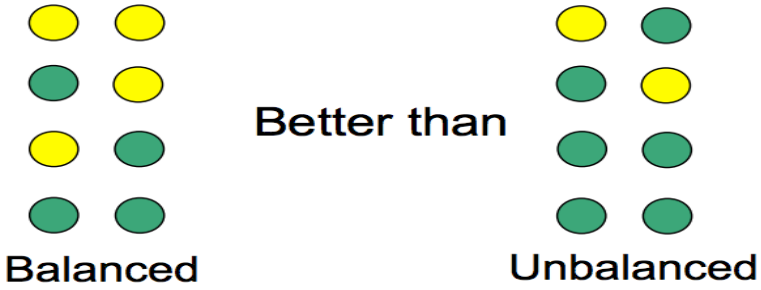
Stratifying

Week One					Week Two				
M	Tu	W	Th	F	M	Tu	W	Th	F
C	T	T	T	T	T	T	T	C	T
T	C	C	C	T	C	C	C	T	T
C	C	C	T	C	C	C	T	T	C
T	T	T	C	C	T	T	C	C	C

T = treated, C = control, pink = female, blue = male

More good study characteristics

- Balanced



- Replicated
- Has Controls