Statistics in Clinical Research

Prof. Jonathan Rabinowitz, PhD
Jonathan.Rabinowitz@biu.ac.il
Bar Ilan University
Israel

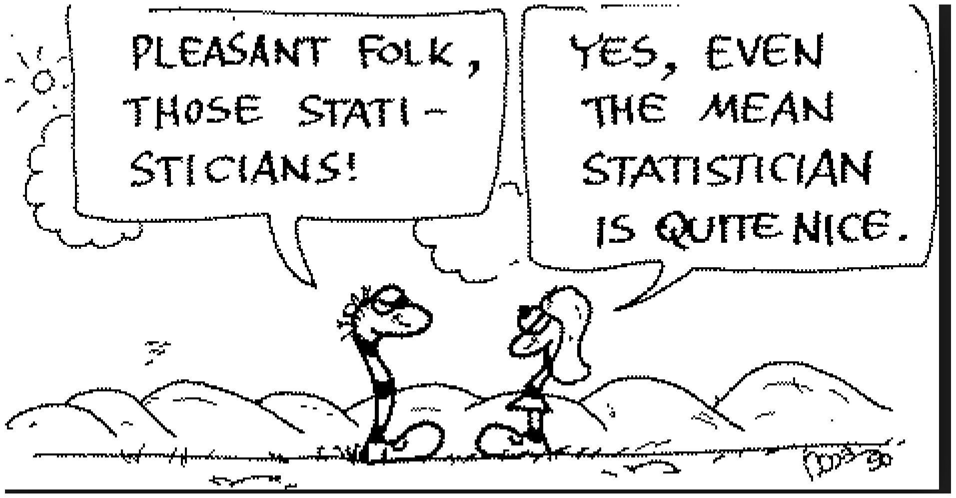
Disclosures

Jonathan Rabinowitz, PhD, has received research grant support and/or travel support and/or speaker fees and/or consultancy fees from Amgen, Janssen, JNJ, Clexio, Lundbeck, Minerva, Eli Lilly, Pfizer, Intra-Cellular Therapies, F. Hoffmann-La Roche, Neumarker AI, Genetikaplus, Takeda and Piere-Fabre.

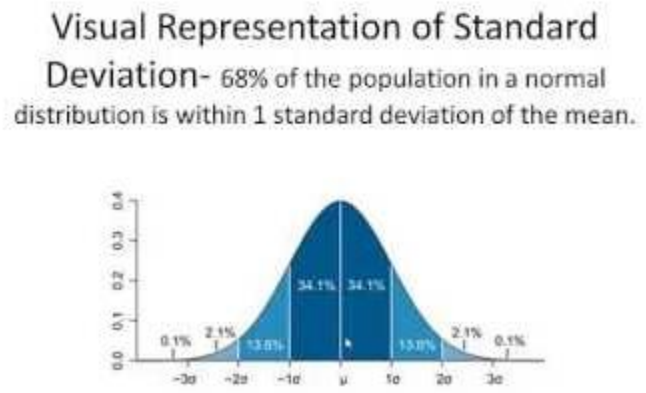
Why do we have statistics in research?

- (a) to confuse the public
- (b) to help statisticians earn a living
- (c) as a barrier of entry to research to people who get scared when they see numbers
- (d) to see the extent to which results are unlikely to be chance findings
- (e) to describe and summarize the data collected

Descriptive statistics



- Mean-average (a+b+c+d)/4
- Standard deviation-distance between each value and the mean
- Median-middle point by rank order
- Quartiles- 0-25%, 25%-50%, 50%-75%, 75%-100%
- Range- hi and low



Inferential statistics

- Infer from the sample data what the population looks like.
- Make judgments of the probability that an observed difference between groups is:

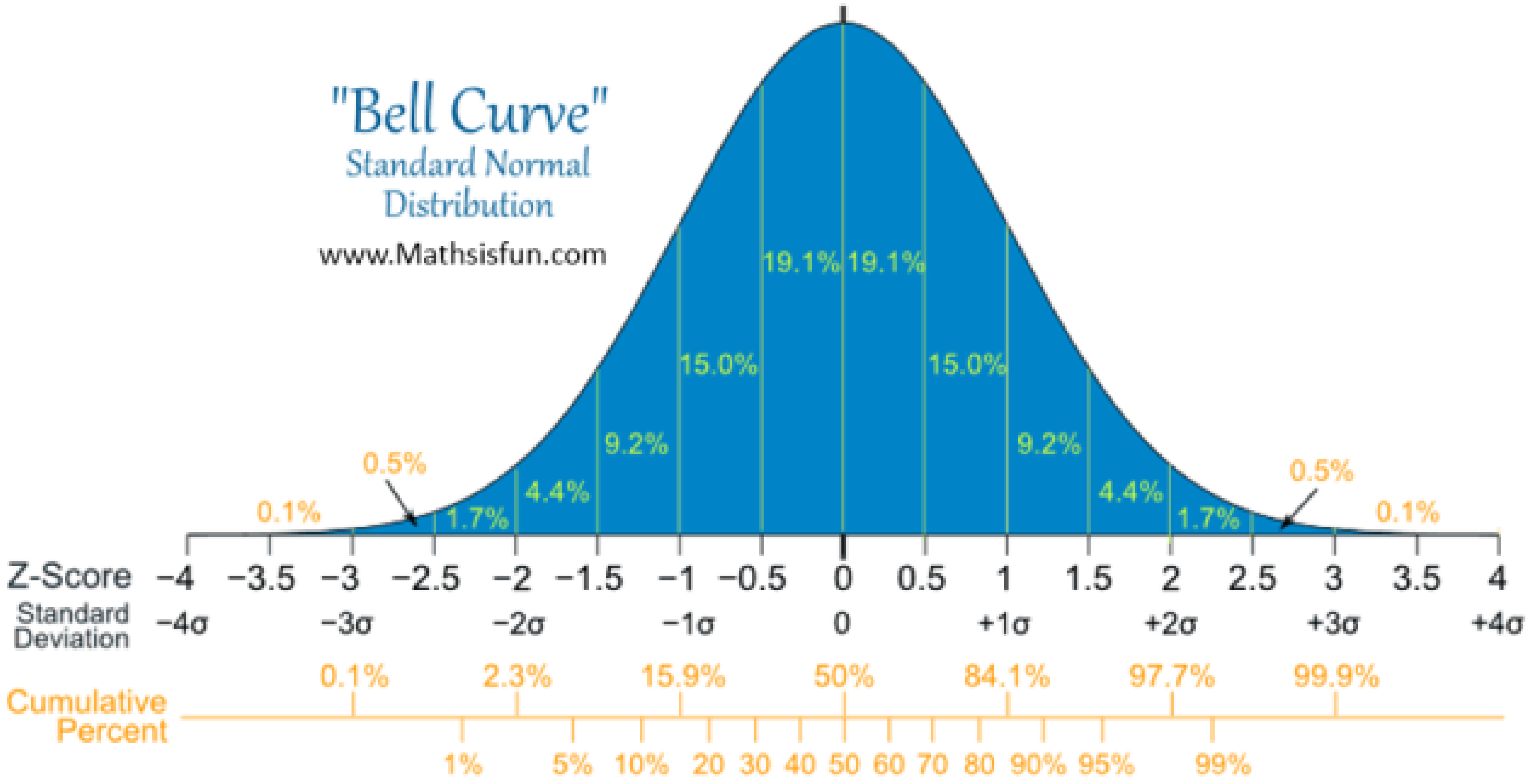
  dependable ($p<.05$) ☺

  OR

  might have happened by chance ($p>.05$) ☹

Central Limit Theorem: a fundamental theorem of statistics.

The sum of a sufficiently large number of independent identically distributed random variables approximately follows a **normal distribution**.



Standard deviation σ: How much variation there is away from the "average" (mean).

We have $N = 4$ because there are four data points:

$$x_1 = 5 \quad x_2 = 6 \quad x_3 = 8 \quad x_4 = 9$$

Replacing N with 4 Replacing $\overline{x}$ with 7

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2} \qquad \sigma = \sqrt{\frac{1}{4}\sum_{i=1}^{4}(x_i - \overline{x})^2}$$

$$\sigma = \sqrt{\frac{1}{4}\sum_{i=1}^{4}(x_i - 7)^2}$$

$$\sigma = \sqrt{\frac{1}{4}\left[(x_1 - 7)^2 + (x_2 - 7)^2 + (x_3 - 7)^2 + (x_4 - 7)^2\right]}$$

$$\sigma = \sqrt{\frac{1}{4}\left[(5 - 7)^2 + (6 - 7)^2 + (8 - 7)^2 + (9 - 7)^2\right]}$$

$$\sigma = \sqrt{\frac{1}{4}\left((-2)^2 + (-1)^2 + 1^2 + 2^2\right)} \quad \sigma = \sqrt{\frac{1}{4}(4 + 1 + 1 + 4)} \quad \sigma = \sqrt{\frac{10}{4}}$$

$$\sigma = \sqrt{\frac{5}{2}}$$

$\sigma = 1.5811$ This is the standard deviation.

Standard error

When Adam ate the apple?

Standard error or SE = $\frac{\sigma}{\sqrt{n}}$ or standard deviation / square root of n

Larger the n the smaller the standard error.

Confidence interval. Typical 95%, this is mean + and – (1.96 x standard error). (p=.05)

1.96 corresponds to almost 2 standard deviation difference, typical level of statistical significance.

For 99.7% confidence interval would be +/- 3 SE (p=.0027)

Plain language summary of standard error (SE)

SE = standard deviation / square root of n

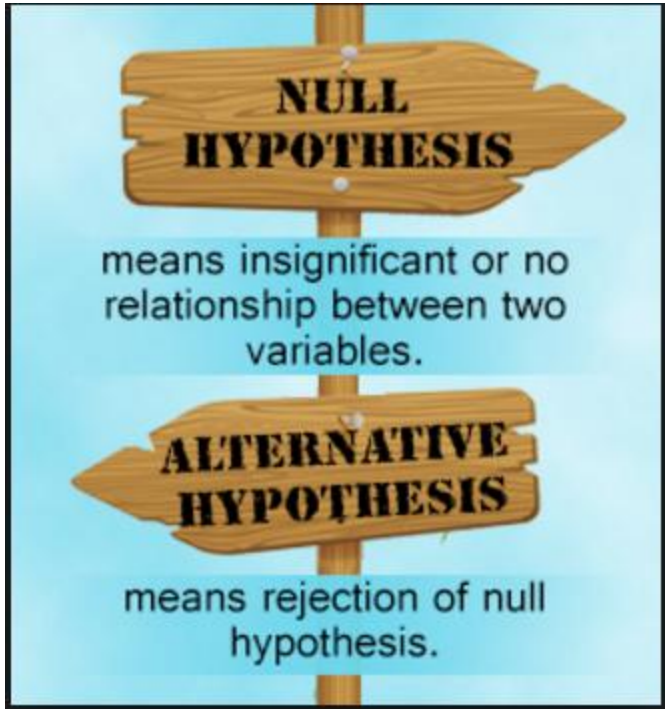
The larger the n the smaller the standard error.
The smaller the standard error smaller differences are significant.

Stats in research

- Documented in study protocol and statistical analysis plan.
- Hypothesis driven
- Choose the best measures
  - Primary
  - Secondary
  - Exploratory measures
- Sample size (power)
- Pre-specified analytic plan
  - How will the information collected be coded and scored?
  - How will those scores be used to test the study hypothesis?
  - What value on those tests will support/refute the hypothesis?
  - Will there be an interim analysis?
  - How will dropouts be handled?
  - How will multiple comparisons be dealt with?

Hypothesis driven

- Examples
  - Significantly fewer patients treated with treatment “A” will relapse as compared to placebo.
  - There will be significantly more reduction on the symptom measure (e.g., PANSS, MADRS etc.) among patients treated with “A” than those on placebo.
  - Symptom remission will be significantly greater among patients treated with “A” as compared to “B.”
  - Symptom improvement will be greater among patients treated with 10 mg of “A” then placebo and greater yet among patients treated with 20 mg.
  - Treatment “A” will not be inferior to treatment “B” by a margin of .10 effect size.



Types of Hypothesis

- **Null Hypothesis (H_0)**
- **Alternative Hypothesis (H_a or H_1)**

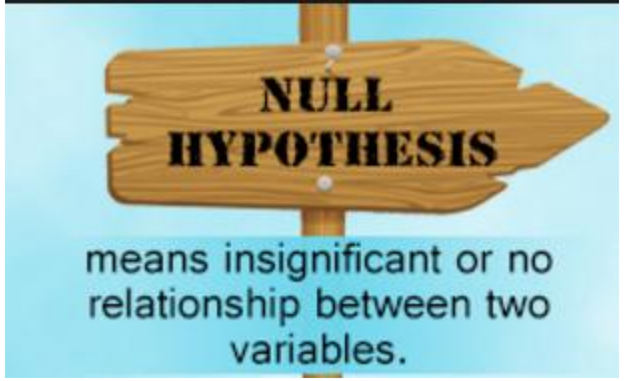
Each of the following statements is an example of a null hypothesis and alternative hypothesis.

$H_0 : \mu = \mu_0$	$H_a : \mu \neq \mu_0$
$H_0 : \mu \leq \mu_0$	$H_a : \mu > \mu_0$
$H_0 : \mu \geq \mu_0$	$H_a : \mu < \mu_0$

fppt.com

Remember: If P is Low, the Null must GO!

Type I & II error



		Truth (for population studied)	
		Null Hypothesis True	Null Hypothesis False
Decision (based on sample)	Reject Null Hypothesis	*Type I Error*	*Correct Decision*
	Fail to reject Null Hypothesis	*Correct Decision*	*Type II Error*

		Truth	
		Not Guilty	Guilty
Verdict	Guilty	Type I Error -- Innocent person goes to jail (and maybe guilty person goes free)	Correct Decision
	Not Guilty	Correct Decision	Type II Error -- Guilty person goes free

Significant Warnings in Effect

- Statistical significance is not the same as practical significance.
- Dichotomizing into significant and non-significant results encourages the dismissal of observed differences in favor of the usually less interesting null hypothesis of no difference.

How many subjects do you need?

- As in optics, statistics also has power, the greater the power the greater likelihood you will find a significant effect.
- If you are looking for skyscrapers

- Your naked eyes or glasses will do

Effect size

- Standardized difference
- Common Cohen's d
  - d=(mean change in group 1-mean change in group 2)/standard deviation
    - Typically the pooled standard deviation or overall standard deviation.
      - Can be obtained by getting the total mean and SD.
      - Can be obtained:
      - =Sqr [(standard deviation group 1)^{2} + (standard deviation group 2)^{2}]/2
  - Example:

    Mean change from baseline to endpoint

    Placebo 30 SD 10

    Active treatment 50 SD 10

    Effect size d= (30-50)/10 or .50

Small=.20 to .50
Example: .20 is average difference in heights of 15 and 16 year-old girls.
>>SSRI's =.35
>>second generation antipsychotics ES=.40

Medium=.50 to .80
Example: .50 is average difference in height between 14 and 18 year old girls.

Large> .80
Example: .80 is average difference in height between 13 and 18 year old girls.

Effect size: dichotomous variable

People on treatment are 2.3 times more likely to respond (e.g., at least 30% decline on symptom scale) than placebo.

	Response Yes	Response No
Treatment n=100	a 50	b 50
Placebo n=100	C 30	D 70

Odds ratio:

OR = (a/b) /(c/d) = ad/bc

(50x70)/(50x30)=2.33

3500/1500=2.33

Equivalent to a Cohen's d of .46
Log (2.33) * sqrt (3)/pi

Where to get effect size?

- Literature- good meta-analysis

Gene V. Glass, Inventor of meta-analysis

- Pragmatically, what effect size would make the tested treatment feasible (cost, risk etc.)

Gene v. Glass

- Glass, G. V, & Smith, M. L. (1978). Reply to Eysenck. *American Psychologist, 33,* 517- 519. Reprinted in T. Cook (Ed.). (1978). *Evaluation Studies Review Annual (Vol. 3)*. Beverly Hills: SAGE.
- Smith, M.L. & Glass, G. V (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, 32, 752-60. Reprinted in Kiesler, C. & Cummings, N., Psychology and National Health Insurance. Washington D.C.: American Psychological Association, 1978.) . Reprinted in I. Lehmann & W. Mehrens (Eds.). (1979). *Educational Research in Focus*. New York: Holt
- Smith, M.L.; Glass, G.V.; & Miller, T. (1980). *The Benefits of Psychotherapy*. Baltimore, MD: Johns Hopkins Univ. Press.

History of Meta-Analysis

- 1952-- Debate begins--Hans J. Eysenck
  - No favorable effects of psychotherapy

- 1952-1970’s—Continued research 100’s of studies no conclusion.

- 1978 --To prove Eysenck wrong, Gene V. Glass statistically aggregates findings of 375 psychotherapy outcome studies.
  - Redemption: Glass concludes psychotherapy works.
  - Glass calls his method **Meta-Analysis**

The Effects of Psychotherapy: An Evaluation
H. J. Eysenck (1952)
Institute of Psychiatry, Maudsley Hospital
University of London
First published in *Journal of Consulting Psychology*, *16*, 319-324.

Table 1
Summary of Reports of the Results of Psychotherapy

	N	Cured; much improved		Improved	Slightly improved		Not improved; died; left treatment	% Cured; much improved; improved
(A) *Psychoanalytic*								
1. Fenichel [13, pp. 28-40]	484	104		84	99		197	39
2. Kessel & Hyman [24]	34	16		5	4		9	62
3. Jones [22, pp. 12-14]	59	20		8	28		3	47
4. Alexander [1, pp. 30-43]	141	28		42	23		48	50
5. Knight [25]	42	8		20	7		7	67
All cases	760		335			425		44%
(B) *Eclectic*								
1. Huddleson [20]	200	19		74	80		27	46
2. Matz [30]	775	10		310	310		145	41
3. Maudsley Hospital Report (1931)	1721	288		900		533		69
4. Maudsley Hospital Report (1935)	1711	371		765		575		64
5. Neustatter [32]	46	9		14	8		15	50
6. Luff & Garrod [27]	500	140		135	26		199	55
7. Luff & Garrod [27]	210	38		84	54		34	68
8. Ross [34]	1089	547		306		236		77
9. Yaskin [40]	100	29		29		42		58
10. Curran [7]	83		51			32		61
11. Masserman & Carmichael [29]	50	7		20	5		18	54
12. Carmichael & Masserman [4]	77	16		25	14		22	53
13. Schilder [35]	35	11		11	6		7	63
14. Hamilton & Wall [16]	100	32		34	17		17	66
15. Hamilton *et al.* [15]	100	48		5	17		32	51
16. Landis [26]	119	40		47		32		73
17. Institute Med. Psychol. (quoted Neustatter)	270	58		132	55		25	70
18. Wilder [39]	54	3		24	16		11	50
19. Miles *et al.* ([31])	53	13		18	13		9	58

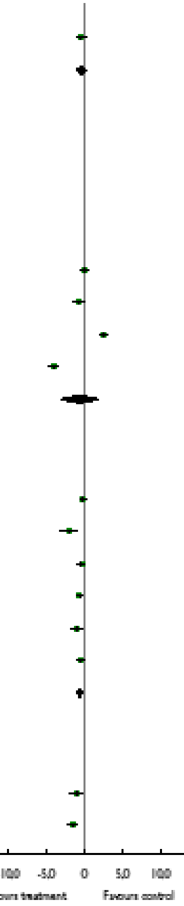
Analysis 01.01. Comparison 01 STPP vs wait-list/TAU/minimal treatment, Outcome 01 Reduction in general psychiatric symptoms: short-term

Review: Short-term psychodynamic psychotherapies for common mental disorders

Comparison: 01 STPP vs wait-list/TAU/minimal treatment

Outcome: 01 Reduction in general psychiatric symptoms: short-term

Study	Treatment N	Treatment Mean(SD)	Control N	Control Mean(SD)	Weight (%)	Standardised Mean Difference (Random) 95% CI
01 anxiety disorders						
Brom 1989	26	169.60 (57.90)	20	193.30 (67.70)	7.8	-0.37 [-0.96, 0.21]
Subtotal (95% CI)	26		20		7.8	-0.37 [-0.96, 0.21]
Test for heterogeneity: not applicable						
Test for overall effect z=1.25 p=0.2						
02 depressive disorders						
Subtotal (95% CI)	0		0		0.0	Not estimable
Test for heterogeneity: not applicable						
Test for overall effect: not applicable						
03 somatoform disorders						
Hamilton 2000	37	0.67 (0.48)	31	0.67 (0.52)	7.9	0.0 [-0.48, 0.48]
Monsen 2000	20	0.40 (0.26)	20	0.66 (0.44)	7.7	-0.71 [-1.35, -0.06]
Sjodin 1986	50	8.89 (0.82)	53	7.07 (0.65)	7.9	2.45 [1.93, 2.97]
Svedlund 1983	50	7.96 (0.62)	50	11.34 (0.99)	7.6	-4.06 [-4.76, -3.37]
Subtotal (95% CI)	157		154		31.0	-0.57 [-3.03, 1.89]
Test for heterogeneity chi-square=222.23 df=3 p=<0.0001 I² =98.7%						
Test for overall effect z=0.46 p=0.6						
04 mixed disorders						
Guthrie 1999	52	1.82 (0.79)	49	1.92 (0.77)	8.0	-0.13 [-0.52, 0.26]
Maina 2005	10	2.40 (1.30)	10	4.50 (0.50)	6.8	-2.04 [-3.17, -0.92]
Marmar 1988	31	0.93 (0.66)	30	1.15 (0.64)	7.9	-0.33 [-0.84, 0.17]
Piper 1990	47	0.60 (0.50)	57	0.90 (0.50)	8.0	-0.60 [-0.99, -0.20]
Shefler 1995	16	64.31 (10.33)	16	72.31 (6.84)	7.5	-0.89 [-1.62, -0.16]
Sloane 1975	30	5.27 (2.70)	33	6.45 (2.84)	7.9	-0.42 [-0.92, 0.08]
Subtotal (95% CI)	186		195		46.1	-0.56 [-0.90, -0.21]
Test for heterogeneity chi-square=12.42 df=5 p=0.03 I² =59.7%						
Test for overall effect z=3.14 p=0.002						
05 personality disorders						
Abbass 2006	14	27.10 (22.60)	13	58.60 (36.50)	7.4	-1.02 [-1.83, -0.21]
Winston 1994	25	5.77 (2.90)	26	9.57 (2.04)	7.7	-1.50 [-2.12, -0.87]



(Continued . . .)

Basics of planning sample size

- **Power: Chance of finding a difference if it exists (absence of type II error).**
- **Depends on size of the difference.**
  - Often expressed as "effect size":
    - Difference/standard deviation
      - Examples:
        - » .20 difference in heights of 15- and 16-year-old girls.
          Needs 527 per group (90% power)
        - » .80 difference in height between 13- and 18-year-old girls.
          Needs 30 per group (90% power)
- **Depends on level of significance (generally $p<.05$)**
- **Larger the effect size, greater the power & smaller sample size needed.**

Calculating a sample size short-cut

For 2 equal size groups total sample size

For 80% power $n=32/d^2$

For 90% power $n=42/d^2$

Example: For effect size of d=.40

80% power: 200 subjects in total are needed, 100 per arm

$32/(.40*.40)=32/.16=200$

90% power: 262 subjects in total are needed, 131 per arm

$42/(.40*.40)=42/.16=262$

Reporting sample size estimation for comparison of means (t-test)

- A sample size of 100 in each group will have 90% power to detect an effect size of 0.46 using a two group t-test with a 0.050 two-sided significance level.

For 90% power, p=.05, 2-tailed

- Too big: Not likely to be of clinical significance

1500 per arm

Effect size: .12

5241 per arm

10% vs. 12% (odds ratio 1.23)

If I only have x patients what can I test

Per Arm	Effect size	Example
20	0.91 (1.05)	Effects rarely seen in clinical trials.
30	0.74 (.85)	
60	0.51 (.60)	
100	0.40 (.46)	
130	0.35 (.40)	Typical effect size in placebo controlled anti-depressant studies is .35
170	0.30 (.35)	
500	.18 (.21)	
1000	.13 (.14)	
2000	.09 (.10)	Effects not likely to be of interest.

80%, 90% power, alpha (p=.05), 2-Tails

Lab conditions vs. clinical trial

- Lab
  - Conditions controlled
    - Carefully bred animals
    - Controlling all stimuli
  - Well calibrated instruments
  - Very large effect sizes 1.50
- Clinical trial
  - Rating scales-hard to calibrate
  - Conditions controlled as best as possible
    - Inclusion and exclusion criteria insure that appropriate population of patients are enrolled
      - Whenever a test is available use it (e.g. pregnancy, heart irregularity, concomitant enrollment)
  - Small-medium effect sizes .30-.40

ReSpOnSe/clinical significance

- Differences in means tell you whether two groups differ, but this can be deceptive.
  - Study of income difference between college graduates and non-graduates.
    - Two random samples.
    - Bill Gates, who did not graduate college, is included in the study.
    - Results: The group who did not go to college makes more money.

	Less than 10%	**At least 10%**	**At least 20%**	**At least 30%**	**At least 40%**
Placebo (n=100)					10
Active treatment (n=100)					30

Adverse events: put your p values down!

- "One bad apple can spoil the whole bunch"
- Serious AE's

	Placebo (n=100)	**Active treatment (n=100)**
Myrocardial infarction-Yes	0	2
No	100	98

Chi-square= , df=1, P=

Common statistical methods

- Comparison of means
  - Two groups t-test
  - More than two groups ANOVA
    - Covariates or control variables can be added ANCOVA
      - For example control for possible differences between treatment groups.
        - Yes, even if groups are randomized you can have differences (e.g. mix of males and females, baseline scores etc.
      - Always adjust for baseline scores due to regression to the mean.
- Comparison on dichotomous outcome
  - Yes/No treatment response
  - Yes/No hospital discharge
- Time to event- survival analysis

Common statistical methods

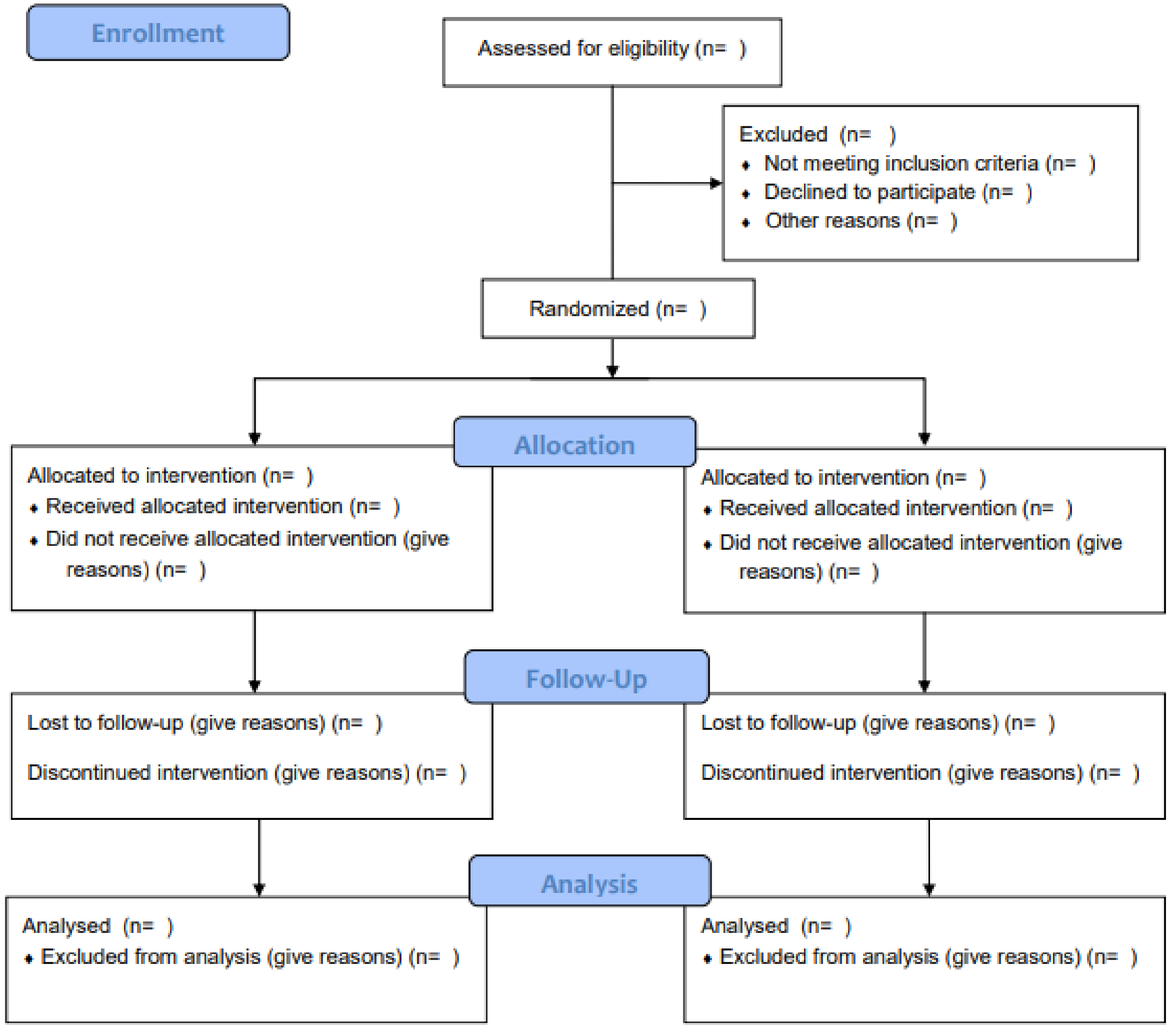
- <u>Parametric</u>
- A hypothesis test that presupposes a particular form of the distributions of underlying populations.
- E.g. **normal distribution**.
- T-test

- <u>Non-parametric</u>
- A hypothesis test where it is not necessary to specify the form of the distribution of the underlying population.
- Examples: rank order/**s**ign test, Wilcoxon test, Mann–Whitney test, Kruskal–Wallis test, and Kolmogorov–Smirnov (as good as the drink).

Can we ignore dropouts?

CONSORT
TRANSPARENT REPORTING of TRIALS

CONSORT 2010 Flow Diagram



Last Observation Carried Forward (LOCF)

			Visit						
ID	Trt	BL	1	2	3	4	5	6	LOCF
1	1	22	20	18	16	14	12	10	-12
2	1	22	21	18	15	12	9	6	-16
3	1	22	22	21	20	*	*	*	-2
4	2	20	20	20	20	21	21	22	2
5	2	21	22	22	23	24	25	25	4
6	2	18	19	21	*	*	*	*	3

Mixed Models Repeated Measurements (MMRM)

			Visit					
ID	Trt	BL	1	2	3	4	5	6
1	1	22	20	18	16	14	12	10
2	1	22	21	18	15	12	9	6
3	1	22	22	21	20	*	*	*
4	2	20	20	20	20	21	21	22
5	2	21	22	22	23	24	25	25
6	2	18	19	21	*	*	*	*

MMRM sees that subject 3 was improving, but more slowly than subjects 1 and 2. Means at visits 4-6 are based on projected slow improvement of subject 3.

Trending methods: New and exciting

JAMA Guide to Statistics and Methods

Target Trial Emulation
A Framework for Causal Inference From Observational Data

Miguel A. Hernán, MD, DrPH; Wei Wang, PhD; David E. Leaf, MD, MMSc

JAMA | Original Investigation

Emulation of Randomized Clinical Trials With Nonrandomized Database Analyses
Results of 32 Clinical Trials

Shirley V. Wang, PhD, ScM; Sebastian Schneeweiss, MD, ScD; and the RCT-DUPLICATE Initiative

jama_wang_2023_oi_230035_1712598871.82677.pdf (silverchair.com)

Target Trial Emulation: A Framework for Causal Inference From Observational Data | Research, Methods, Statistics | JAMA | JAMA Network

Applying the target trial emulation framework in your research involves several key steps:
1. Define the Protocol of the Target Trial: Start by specifying the protocol of the hypothetical randomized trial you wish to emulate. This includes defining eligibility criteria, treatment strategies, assignment procedures, outcomes, follow-up periods, causal contrasts of interest, and the statistical analysis plan[1].

2. Use Observational Data: Since a target trial is not feasible, use available observational data to emulate the trial. Ensure that the data source is robust and can provide the necessary information for the components defined in your protocol[1].

3. Design the Emulation: Carefully design the observational study to follow the protocol of the target trial as closely as possible. This includes matching the eligibility criteria, treatment strategies, and outcomes to what would have been used in the target trial[2].

4. Adjust for Confounding: Implement statistical methods to adjust for confounding factors that could affect the validity of your causal inference. This is crucial because, unlike randomized trials, observational studies are prone to confounding bias[1].

5. Analyze the Data: Follow the statistical analysis plan you outlined in your protocol. Use appropriate statistical methods to estimate the causal effect of the intervention as if it were a randomized trial[3].

6. Interpret the Results: Interpret the results in the context of the target trial emulation framework. Be cautious about the limitations of observational data and the potential for residual confounding[4]

From: **Association Between Early Treatment With Tocilizumab and Mortality Among Critically Ill Patients With COVID-19**

JAMA Intern Med. 2021;181(1):41-51. doi:10.1001/jamainternmed.2020.6252

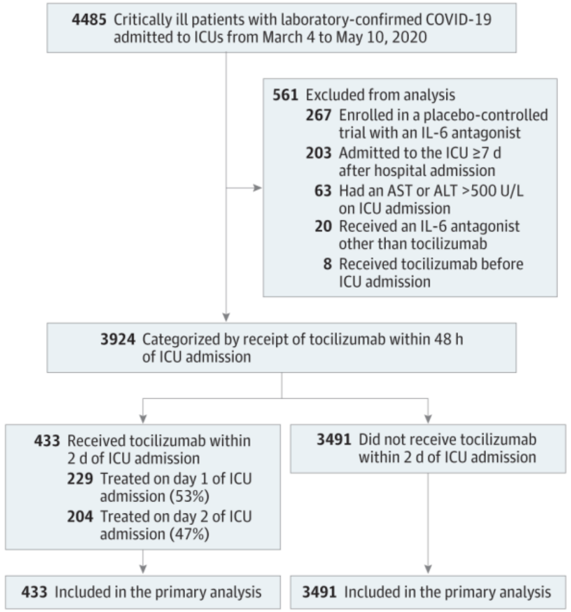


Figure Legend:

Study Cohort and Emulated Trial FlowAbbreviations: ALT, alanine aminotransferase; AST, aspartate aminotransferase; ICU, intensive care unit; and IL-6, interleukin 6. To convert ALT and AST to µkat/L, multiply by 0.0167.

Abstract

Importance: …..

Objective: To test whether tocilizumab decreases mortality in this population.

Design, setting, and participants: The data for this study were derived from a multicenter cohort study of 4485 adults with COVID-19 admitted to participating intensive care units (ICUs) ….. Critically ill adults with COVID-19 were categorized according to whether they received or did not receive tocilizumab in the first 2 days of admission to the ICU. Data were collected retrospectively until June 12, 2020. A Cox regression model with inverse probability weighting was used to adjust for confounding.

Exposures: Treatment with tocilizumab in the first 2 days of ICU admission.

Main outcomes and measures: Time to death, compared via hazard ratios (HRs), and 30-day mortality, compared via risk differences.

Results: Among the 3924 patients included in the analysis …..Patients treated with tocilizumab were younger (median age, 58 [IQR, 48-65] vs 63 [IQR, 52-72] years) and had a higher prevalence of hypoxemia on ICU admission (205 of 433 [47.3%] vs 1322 of 3491 [37.9%] with mechanical ventilation and a ratio of partial pressure of arterial oxygen to fraction of inspired oxygen of <200 mm Hg) than patients not treated with tocilizumab. After applying inverse probability weighting, baseline and severity-of-illness characteristics were well balanced between groups.

A total of 1544 patients (39.3%) died, including 125 (28.9%) treated with tocilizumab and 1419 (40.6%) not treated with tocilizumab. In the primary analysis, during a median follow-up ….patients treated with tocilizumab had a lower risk of death compared with those not treated with tocilizumab (HR, 0.71; 95% CI, 0.56-0.92). The estimated 30-day mortality was …….. (risk difference, 9.6%; 95% CI, 3.1%-16.0%).

Conclusions and relevance: Among critically ill patients with COVID-19 in this cohort study, the risk of in-hospital mortality in this study was lower in patients treated with tocilizumab in the first 2 days of ICU admission compared with patients whose treatment did not include early use of tocilizumab. However, the findings may be susceptible to unmeasured confounding, and further research from randomized clinical trials is needed.

JAMA | Original Investigation

Emulation of Randomized Clinical Trials With Nonrandomized Database Analyses
Results of 32 Clinical Trials

Shirley V. Wang, PhD, ScM; Sebastian Schneeweiss, MD, ScD; and the RCT-DUPLICATE Initiative

Key Points

Question Are database studies that are explicitly designed to emulate past and ongoing randomized clinical trials (RCTs) of medications able to generate similar causal conclusions?

Findings In this highly selected, nonrepresentative sample, real-world evidence studies generally reached similar conclusions as RCTs (Pearson correlation *r*, 0.82; 75% statistical significance agreement, 66% estimate agreement, 75% standardized difference agreement). In a post hoc, exploratory stratified analysis, agreement was higher in RCT-database pairs classified as having closer emulation of the RCT design.

Meaning Selected database studies can complement RCT evidence to enhance understanding of how medications work in clinical practice. Emulation differences, chance, and residual confounding can contribute to divergence in results and are difficult to disentangle.

JAMA.2023;329(16):1376-1385. doi:10.1001/jama.2023.422

Garbage in garbage out: insuring data quality

- Choose most appropriate measures.
- Insure measures are used correctly.
  - Proper training and supervision.
  - Ongoing review.
- Careful adherence to inclusion/exclusion criteria.

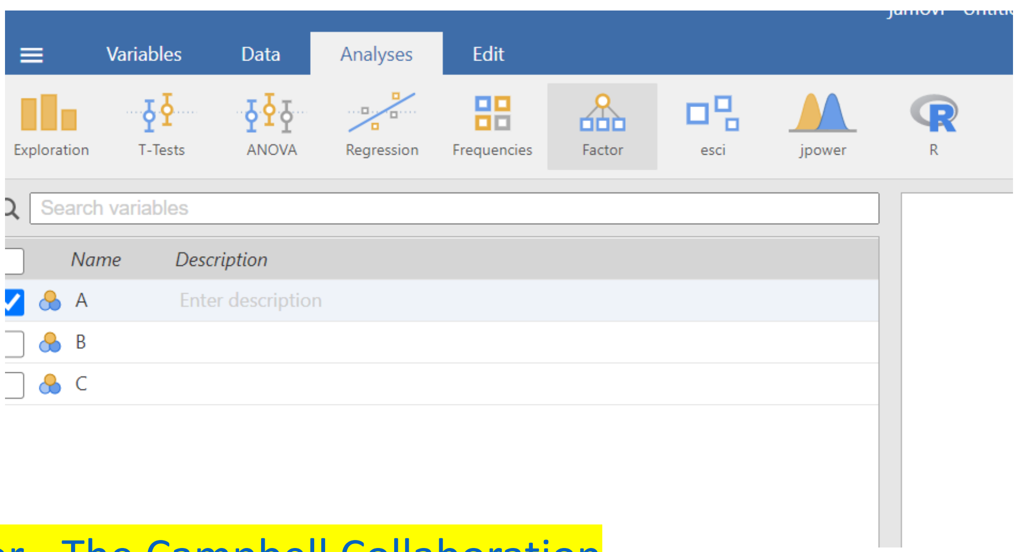
Stats in clinical research while standing on one foot

- Documented in study protocol and statistical analysis plan.
- Hypothesis driven
- Choose the best measures
  - Primary
  - Secondary
  - Exploratory measures
- Sample size (power)
- Pre-specified analytic plan
  - How will the information collected be coded and scored?
  - How will those scores be used to test the study hypothesis?
  - What value on those tests will support/refute the hypothesis?
  - Will there be an interim analysis?
  - How will dropouts be handled?
  - How will multiple comparisons be dealt with?

Conclusion- Great free online resources

1. Excellent free online text on statistics: https://www.bmj.com/about-bmj/resources-readers/pubdlications/statistics-square-or
2. Amazing clear concise video presentations e.g., statistical power, large language models, Bayesian statistics, censoring, missing data [Stats, STAT! | NEJM Evidence]

3. Excellent free software package: https://www.jamovi.org



4. [Effect size calculator - The Campbell Collaboration]

Reporting sample size estimation: proportion

- A two group continuity corrected chi-square test with a 0.050 two-sided significance level will have 90% power to detect the difference between a Group 1 proportion, p_1, of 0.300 and a Group 2 proportion, p_2, of 0.500 (odds ratio of 2.333) when the sample size in each group is 134.

Reporting sample size estimation: Survival analysis

When the sample size in each group is 58, with a total number of events required, E, of 33, a 0.050 level two-sided log-rank test for equality of survival curves will have 90% power to detect the difference between a Group 1 proportion p_1 at time t of 0.500 and a Group 2 proportion p_2 at time t of 0.800 (a constant hazard ratio of 3.106); this assumes no dropouts before time t.

Sample size estimation for non-inferiority study

- When the sample size in each group is 1714, a two group 0.050 one-sided t-test will have 90% power to <u>reject the null hypothesis of non-equivalence </u>in favor of the alternative hypothesis that the means of the two groups are equivalent, assuming that the difference between the equivalence limit and the expected difference divided by the common standard deviation is 0.100 or greater.

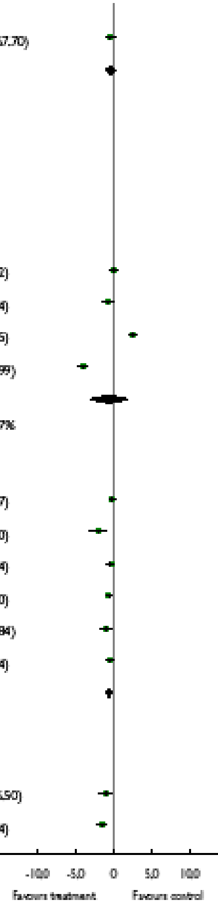
Analysis 01.01. Comparison 01 STPP vs wait-list/TAU/minimal treatment, Outcome 01 Reduction in general psychiatric symptoms: short-term

Review: Short-term psychodynamic psychotherapies for common mental disorders

Comparison: 01 STPP vs wait-list/TAU/minimal treatment

Outcome: 01 Reduction in general psychiatric symptoms: short-term

Study	Treatment N	Mean(SD)	Control N	Mean(SD)	Standardised Mean Difference (Random) 95% CI	Weight (%)	Standardised Mean Difference (Random) 95% CI
01 anxiety disorders							
Brom 1989	26	169.60 (57.90)	20	193.30 (67.70)		7.8	-0.37 [-0.96, 0.21]
Subtotal (95% CI)	26		20			7.8	-0.37 [-0.96, 0.21]
Test for heterogeneity: not applicable							
Test for overall effect z=1.25 p=0.2							
02 depressive disorders							
Subtotal (95% CI)	0		0			0.0	Not estimable
Test for heterogeneity: not applicable							
Test for overall effect: not applicable							
03 somatoform disorders							
Hamilton 2000	37	0.67 (0.48)	31	0.67 (0.52)		7.9	0.0 [-0.48, 0.48]
Monsen 2000	20	0.40 (0.26)	20	0.66 (0.44)		7.7	-0.71 [-1.35, -0.06]
Sjodin 1986	50	8.89 (0.82)	53	7.07 (0.65)		7.9	2.45 [1.93, 2.97]
Svedlund 1983	50	7.96 (0.62)	50	11.34 (0.99)		7.6	-4.06 [-4.76, -3.37]
Subtotal (95% CI)	157		154			31.0	-0.57 [-3.03, 1.89]
Test for heterogeneity chi-square=222.23 df=3 p=<0.0001 I² =98.7%							
Test for overall effect z=0.46 p=0.6							
04 mixed disorders							
Guthrie 1999	52	1.82 (0.79)	49	1.92 (0.77)		8.0	-0.13 [-0.52, 0.26]
Maina 2005	10	2.40 (1.30)	10	4.50 (0.50)		6.8	-2.04 [-3.17, -0.92]
Marmar 1988	31	0.93 (0.66)	30	1.15 (0.64)		7.9	-0.33 [-0.84, 0.17]
Piper 1990	47	0.60 (0.50)	57	0.90 (0.50)		8.0	-0.60 [-0.99, -0.20]
Shefler 1995	16	64.31 (10.33)	16	72.31 (6.84)		7.5	-0.89 [-1.62, -0.16]
Sloane 1975	30	5.27 (2.70)	33	6.45 (2.84)		7.9	-0.42 [-0.92, 0.08]
Subtotal (95% CI)	186		195			46.1	-0.56 [-0.90, -0.21]
Test for heterogeneity chi-square=12.42 df=5 p=0.03 I² =59.7%							
Test for overall effect z=3.14 p=0.002							
05 personality disorders							
Abbass 2006	14	27.10 (22.60)	13	58.60 (36.50)		7.4	-1.02 [-1.83, -0.21]
Winston 1994	25	5.77 (2.90)	26	9.57 (2.04)		7.7	-1.50 [-2.12, -0.87]



(Continued . . .)

Improving Fidelity of Instruments

- Problems with measurement

Appendix. The Activity Rating Scale

Please indicate how often you performed each activity in your healthiest and most active state, **in the past year.**

	Less than one time in a month	**One time in a month**	**One time in a week**	**2 or 3 times in a week**	**4 or more times in a week**
Running: running while playing a sport or jogging					
Cutting: changing directions while running					
Decelerating: coming to a quick stop while running					
Pivoting: turning your body with your foot planted while playing a sport; For example: skiing, skating, kicking, throwing, hitting a ball (golf, tennis, squash), etc.					

Cheng, D.M. and J.W. Hogan, *The Sense and Sensibility of Sensitivity Analyses. New England Journal of Medicine, 2024.* ***391(11): p. 972-974.***

First, an assessment of sensitivity to key choices and assumptions should be summarized in a table comparing results from the primary analysis and each sensitivity analysis.

Second, comparisons of treatment-effect findings should incorporate statistical uncertainty; comparing point estimates alone is neither appropriate nor scientifically meaningful.

Third, investigators should explain the assumptions made with each method and whether testable assumptions are supported by the data.

Fourth, investigators should frame the key conclusions in terms of the primary analyses and use the sensitivity analysis to represent the degree of uncertainty attributable to modeling choices and untestable assumptions.

Table 1. Eight Ideas for Limiting Missing Data in the Design of Clinical Trials.

Target a population that is not adequately served by current treatments and hence has an incentive to remain in the study.
Include a run-in period in which all patients are assigned to the active treatment, after which only those who tolerated and adhered to the therapy undergo randomization.
Allow a flexible treatment regimen that accommodates individual differences in efficacy and side effects in order to reduce the dropout rate because of a lack of efficacy or tolerability.
Consider add-on designs, in which a study treatment is added to an existing treatment, typically with a different mechanism of action known to be effective in previous studies.
Shorten the follow-up period for the primary outcome.
Allow the use of rescue medications that are designated as components of a treatment regimen in the study protocol.
For assessment of long-term efficacy (which is associated with an increased dropout rate), consider a randomized withdrawal design, in which only participants who have already received a study treatment without dropping out undergo randomization to continue to receive the treatment or switch to placebo.
Avoid outcome measures that are likely to lead to substantial missing data. In some cases, it may be appropriate to consider the time until the use of a rescue treatment as an outcome measure or the discontinuation of a study treatment as a form of treatment failure.

Table 2. Eight Ideas for Limiting Missing Data in the Conduct of Clinical Trials.

Select investigators who have a good track record with respect to enrolling and following participants and collecting complete data in previous trials.
Set acceptable target rates for missing data and monitor the progress of the trial with respect to these targets.
Provide monetary and nonmonetary incentives to investigators and participants for completeness of data collection, as long as they meet rigorous ethical requirements.[15,16]
Limit the burden and inconvenience of data collection on the participants, and make the study experience as positive as possible.
Provide continued access to effective treatments after the trial, before treatment approval.
Train investigators and study staff that keeping participants in the trial until the end is important, regardless of whether they continue to receive the assigned treatment. Convey this information to study participants.
Collect information from participants regarding the likelihood that they will drop out, and use this information to attempt to reduce the incidence of dropout.
Keep contact information for participants up to date.