Research Centers in Minority Institutions

RESEARCH FORWARD

VIRTUAL APPLIED DATA SCIENCE TRAINING INSTITUTE

Module 3: Statistical Distributions, Sampling, & Hypotheses Testing

Bayesian Analysis : Alternative to Frequentists

03/05/2021

Paul Kolm, PhD

Outline

- Frequentist (classical) review.
- Bayesian perspective.
- Examples:

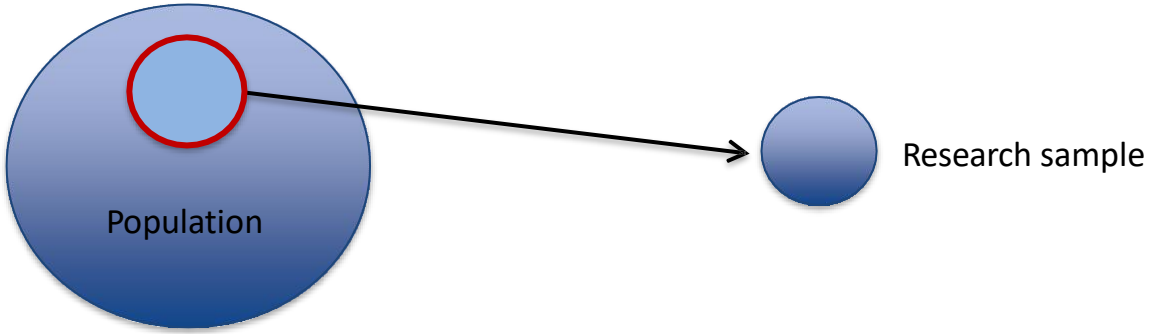
  Diagnostic tests: positive predictive values.

  Randomized trial of medications for epilepsy.

  Probabilistic sensitivity analysis (PSA).

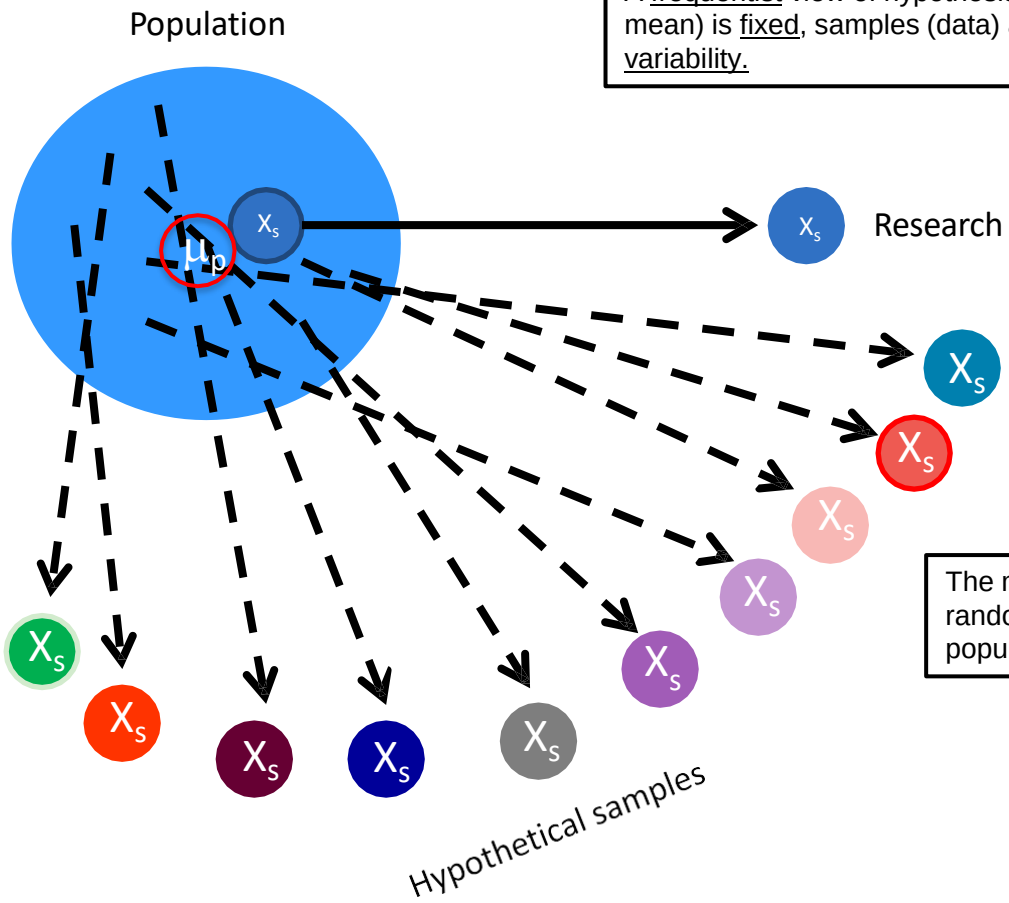
Populations and samples

- Population – everyone in the universe (of interest)
- Sample – subset of the population of interest (random or not)



- Is the sample representative of the population?

A frequentist view of hypothesis testing: population parameter (e.g., mean) is fixed, samples (data) are random – random sampling variability.



The mean of the means of an infinite number of random samples with size n will equal the population mean.

Population of diastolic BP values (mean = 76)

73	42	73	67	84	75	80	78	64	73	70	92
59	77	65	69	91	71	69	62	76	89	91	79
67	76	66	80	98	70	74	71	74	59	84	73
68	74	70	72	84	65	81	54	67	83	87	63
68	93	92	71	95	69	98	74	90	100	60	82
75	91	72	82	67	71	61	94	75	76	66	76
97	58	72	57	74	65	79	73	91	79	63	80
73	92	62	82	73	90	73	83	64	70	60	79
94	73	78	99	71	90	82	74	82	77	71	64
68	73	86	89	100	77	88	72	56	85	90	88
55	68	73	85	80	69	62	73	78	73	84	70
65	81	85	90	89	64	81	77	91	95	86	73

Sampling distribution of diastolic BP

Sample (n = 5)					Mean
63	73	61	73	66	67
90	92	56	68	60	73
69	67	85	72	73	73
90	80	55	73	77	75
69	74	91	73	66	75
64	71	80	76	94	77
62	91	79	89	62	77
81	76	91	77	63	78
95	91	73	65	72	79
72	82	73	90	81	80
83	75	69	92	79	80
89	90	77	80	82	84

Sampling distribution

95% confidence intervals for dBP from hypothetical samples (n = 5)
Population mean dBP = 76

Sample	Mean	SEM	LB	UB	56666666666777777777788888888889999999012345678901234567890123456789012345
72 82 73 90 81	**80**	3	70	89	(-----x----------)
89 90 77 80 82	**84**	3	77	91	x(----------)
95 91 73 65 72	**79**	6	63	95	(----------x----------)
83 75 69 92 79	**80**	4	69	90	(------x----------)
64 71 80 76 94	**77**	5	63	91	(----------x----------)
63 73 61 73 66	**67**	2	60	74	(----------) x
69 67 85 72 73	**73**	3	64	82	(----------x-----)
62 91 79 89 62	**77**	6	59	94	(----------x----------)
81 76 91 77 63	**78**	5	65	90	(----------x----------)
69 74 91 73 66	**75**	4	63	87	(----------x----------)

95% confidence interval: over repeated sampling, the CI will include the true population mean 95% of the time.

Hypothesis testing

Treatment: 80 88 76 77 84 Mean = 81

Control: 84 90 86 77 93 Mean = 86

Null hypothesis: Whatever difference there is between the means of the Treatment and Control groups occurred by random sampling variability – or chance.

Alternative hypothesis: There is a difference between means of the Treatment and Control groups that was unlikely to have occurred just by chance.

Assume the null hypothesis is true.

Based on our statistical test, we will reject the null hypothesis or fail to reject the null hypothesis .

Statistical decisions

- α and β are **arbitrary**.
- $p < 0.05$ is an entrenched tradition.
- If $p < 0.01$ is set as the error rate, then $p = 0.05$ is not "significant" (fail to reject the null hypothesis).
- If p is < 0.01, reject the null hypothesis, but would be wrong 1 time out of every 100 random samples.

p values

- **What p values are**:
  Preset level for rejecting the null hypothesis. Scaled as a probability (0,1).

- **What p values are not:**
  probability that the null hypothesis is false.
  probability that the alternative hypothesis is true.
  an index of importance or meaningfulness.

Interpretation of p values

- If $p < 0.05$ (or 0.01, 0.001 or 0.06), we reject the null hypothesis (having assumed it was true) and conclude that the observed sampling statistic (z, t, χ^2) was unlikely (5/100; 1/1000) to have occurred by random sampling variability (chance).

- If $p > 0.05$, we fail to reject the null hypothesis and conclude – "Found no evidence for a difference."

Bayesian and frequentist perspectives

- Frequentist: Probability is the long-run frequency of random events over repeated sampling.
- Frequentist: Data are random; population parameter is fixed.
- Bayesian: Probability is an index of confidence an investigator has about the occurrence of an event.
- Bayesian: Data are fixed; population parameter (e.g., mean) is random.

Bayesian analysis

- Prior belief (event, condition, diagnosis).
- Get new information (data).
- Revise prior belief with posterior (new) belief.

- Example: Physician updates initial medications based on current lab results.

Bayes' Theorem (conditional probability)

Probability of A given B:

$$P(A|B) = \frac{P(A)\,P(B|A)}{P(B)}$$

Thomas Bayes (1701 – 1761) was an English statistician, philosopher and Presbyterian minister.

Diagnostic test

		Disease +	Disease -
Test	+	TP	FP
	-	FN	TN

Sensitivity = TP/(TP+FN) **Specificity = TN/(TN+FP)**

PPV = TP/(TP+FP) **NPV = TN/(TN+FN)**

FP rate = FP/(FP+TN) **FN rate = FN/(FN+TP)**

Test statistics

- Sensitivity: proportion of TP correctly identified by test
- Specificity: proportion of TN correctly identified by test
- Positive Predictive Value (PPV): proportion of patients with + test that have the disease
- Negative Predictive Value (NPV): proportion of patients with – test that don't have the disease

- Prevalence: proportion of disease in a given population.
- Sensitivity and specificity of a test are independent of prevalence.
- PPV and NPV dependent on prevalence.

Bayes' Theorem

estimating predictive values for specific prevalence.

$$P(D|+) = \frac{P(+|D) \times P(D)}{P(+)}$$

Prior probability - prevalence

Posterior probabilities: PPV and 1 - NPV are revised estimates of the prior for patients positive and negative to the test.

$$P(D|+) = \frac{P(+|D) \times P(D)}{P(+|D) \times P(D) + P(+|ND) \times P(ND)}$$

P(D)	= Prevalence
P(D|+)	= PPV
P(+|D)	= Sensitivity
P(+|ND)	= 1 - Specificity

$$PPV = \frac{\text{sensitivity} \times \text{prevalence}}{\text{sensitivity} \times \text{prevalence} + (1 - \text{specificity}) \times (1 - \text{prevalence})}$$

By a similar argument:

$$NPV = \frac{\text{specificity} \times (1 - \text{prevalence})}{(1 - \text{sensitivity}) \times \text{prevalence} + \text{specificity} \times (1 - \text{prevalence})}$$

		Pathology		
		Abnormal (+)	Normal (−)	Total
Test	Abnormal (+)	231	32	263
	Normal (−)	27	54	81
	Total	258	86	344

Sensitivity = 231/258 = 0.895
Specificity = 54/86 = 0.628
PPV = 231/263 = 0.878
NPV = 54/81 = 0.667
Prevalence = 258/344 = 0.750

Solve for Prevalence = 0.50

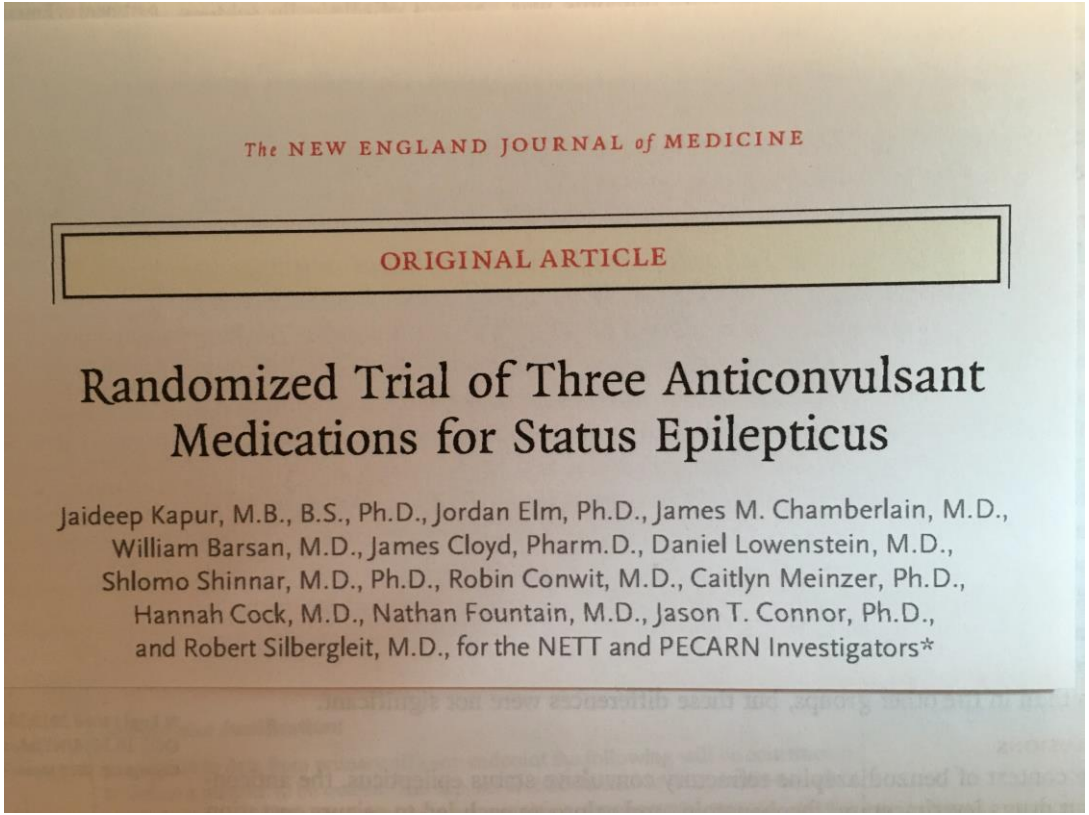
$$PPV = \frac{0.895*.50}{0.895*.50 + (1 - 0.628)*(1 - 0.50)}$$

$$= 0.706$$

Solve for Prevalence = 0.25

$$PPV = \frac{0.895*.25}{0.895*.25 + (1 - 0.628)*(1 - 0.25)}$$

$$= 0.445$$



The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Randomized Trial of Three Anticonvulsant Medications for Status Epilepticus

Jaideep Kapur, M.B., B.S., Ph.D., Jordan Elm, Ph.D., James M. Chamberlain, M.D., William Barsan, M.D., James Cloyd, Pharm.D., Daniel Lowenstein, M.D., Shlomo Shinnar, M.D., Ph.D., Robin Conwit, M.D., Caitlyn Meinzer, Ph.D., Hannah Cock, M.D., Nathan Fountain, M.D., Jason T. Connor, Ph.D., and Robert Silbergleit, M.D., for the NETT and PECARN Investigators*

METHODS

In a randomized, blinded, adaptive trial, we compared the efficacy and safety of three intravenous anticonvulsive agents — levetiracetam, fosphenytoin, and valproate — in children and adults with convulsive status epilepticus that was unresponsive to treatment with benzodiazepines. The primary outcome was absence of clinically evident seizures and improvement in the level of consciousness by 60 minutes after the start of drug infusion, without additional anticonvulsant medication. The posterior probabilities that each drug was the most or least effective were calculated. Safety outcomes included life-threatening hypotension or cardiac arrhythmia, endotracheal intubation, seizure recurrence, and death.

Outcomes

The primary outcome was an absence of clinically apparent seizures and improving responsiveness at 60 minutes after the start of trial drug infusion, without additional anticonvulsant medication, including medication used for endotracheal intubation.

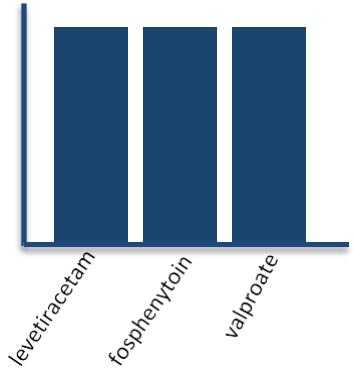
The primary safety outcome was the composite of life-threatening hypotension or cardiac arrhythmia within 60 minutes after the start of trial drug infusion. Data on serious adverse events were collected through the end of participation in the trial (hospital discharge or 30 days, whichever came first) for every patient. Data on adverse events were collected through the first 24 hours after enrollment.

We used a response-adaptive comparative-effectiveness design. Patients were randomly assigned to receive one of the three trial drugs, initially in a 1:1:1 ratio. After 300 patients were assigned to a treatment group, response-adaptive randomization was initiated on the basis of previously defined decision rules, with the goal of maximizing the likelihood of identifying the most effective treatment. Interim analyses were planned after the enrollment of 400, 500, 600, and 700 patients, at which times the trial could be stopped early for success or futility, the rules for which are contained in the protocol.

At each interim analysis, the randomization assignment probabilities were updated. The maximum sample was 795 patients. Randomization was stratified according to age category (2 to 17 years, 18 to 65 years, and >65 years) at the targeted assignment probabilities. **Before assessment of the trial results, all three drugs were considered to be equally likely to be the most effective or least effective treatment.**

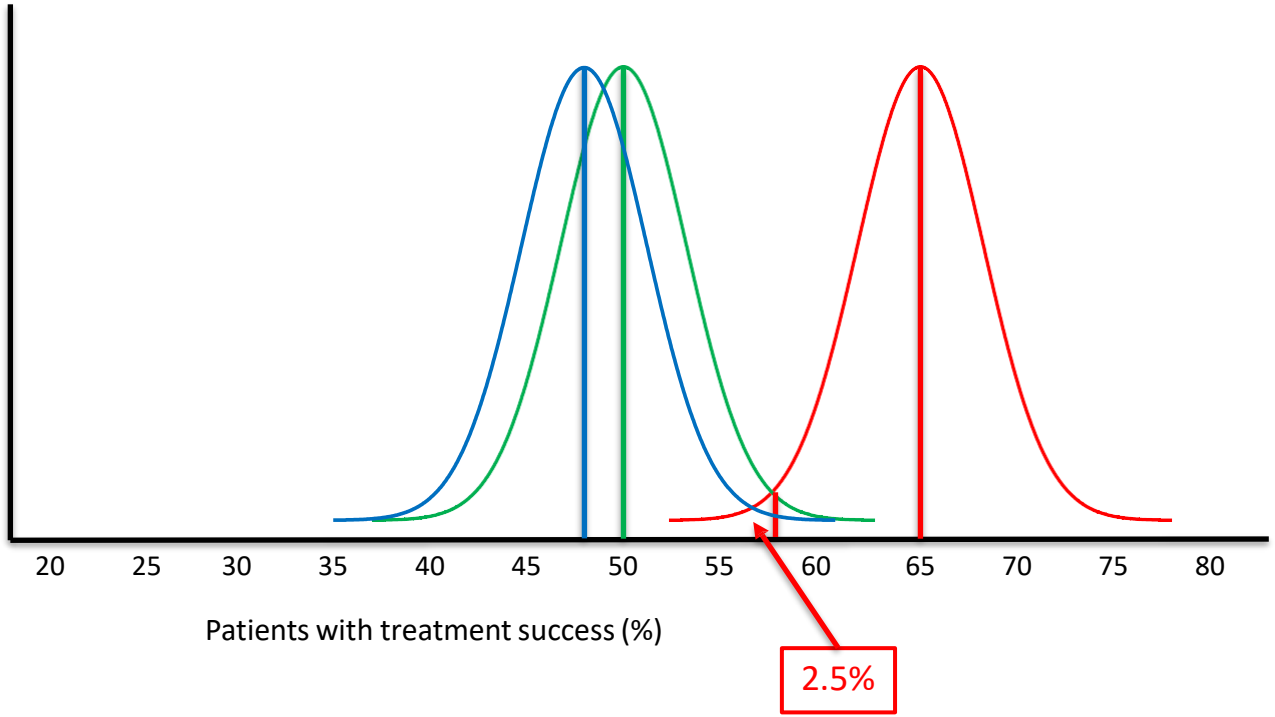
Response rates in each of the treatment groups were modeled independently with the use of Bayesian analysis. The percentage of patients with treatment success in each group was calculated starting with a **uniform(0,1) prior probability distribution** (which allows the treatment success to take any value between 0 and 100%) and was **updated on the basis of the observed binomial data with the use of a conjugate beta-binomial model.** From these three posterior distributions, the probability that each treatment was the most or least effective treatment was calculated as described previously.

Uniform distribution is a **non-informative** prior.



We randomly and repeatedly (10^6 iterations) drew from these three posterior probabilities to calculate the probability that a given treatment was better than the other two. The same approach was taken for the potentially worst treatment. The criterion for declaring a most or least effective treatment was a **probability greater than 0.975.** The threshold of 0.975 was chosen by convention (analogous to an alpha of 0.025 in a one-sided comparison) and because a simulation study showed that with this threshold and trial design, the type I error rate was controlled. Unlike a trial in which success can be achieved in a number of different ways (e.g., multiple treatments vs. a control), only one treatment could be identified as best.

A maximum sample of 720 unique patients from 795 enrollments provided 90% power to identify the **most effective treatment when one treatment group had a true response rate of 65% and the true response rate was 50% in the other two groups (an absolute difference of 15 percentage points).** We report the percentage response in each treatment group with **95% credible intervals**. The primary analysis was based on the intention-to-treat population and included all unique patients who underwent randomization, regardless of the amount of treatment that was actually received.


20
25
30
35
40
45
50
55
60
65
70
75
80
Patients with treatment success (%)
2.5%

In this prospective, randomized, double-blind, adaptive comparative-effectiveness trial involving patients with benzodiazepine-refractory status epilepticus, we found no significant difference in the percentage of patients with seizure cessation among the levetiracetam group, fosphenytoin group, and valproate group.

Treatment success and 95% Credible Interval	
Levetiracetam:	47% (39% - 55%)
Fosphenytoin:	45% (36% - 54%)
Valproate:	46% (38% - 55%)

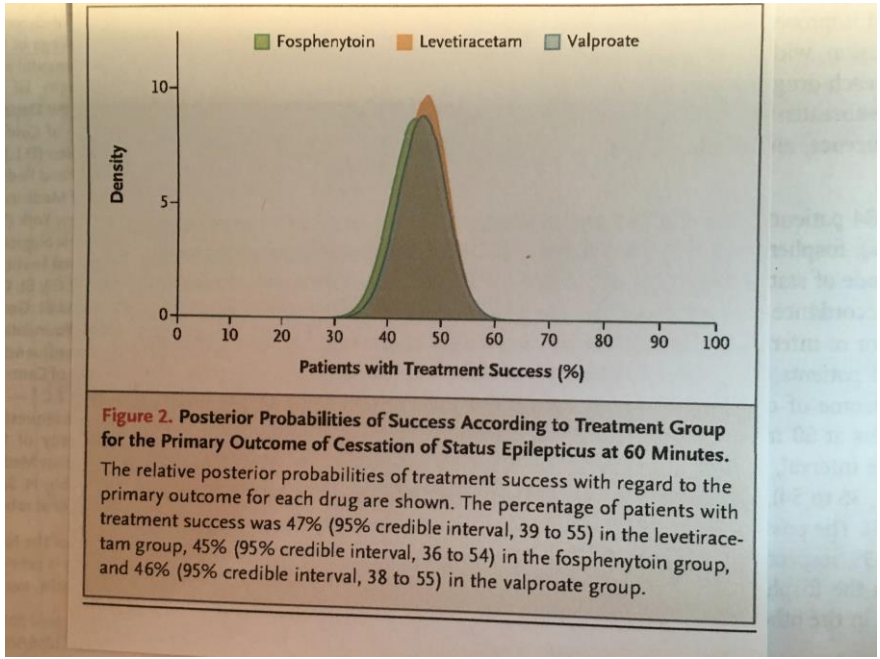


Figure 2. Posterior Probabilities of Success According to Treatment Group for the Primary Outcome of Cessation of Status Epilepticus at 60 Minutes.

The relative posterior probabilities of treatment success with regard to the primary outcome for each drug are shown. The percentage of patients with treatment success was 47% (95% credible interval, 39 to 55) in the levetiracetam group, 45% (95% credible interval, 36 to 54) in the fosphenytoin group, and 46% (95% credible interval, 38 to 55) in the valproate group.

Bayesian CI:
95% credible interval: 95% probability that the interval includes the population parameter.

Frequentist CI:
95% confidence interval: over repeated sampling, the interval will include the population parameter 95% of the time.

Probabilistic (Bayesian) sensitivity analysis

- Multiple input variables to estimate multiple parameters.
- Parameter estimation sensitive to changes in input variables.
- Prior and posterior distributions (e.g., normal, gamma, beta) the same for a given variable.
- Different distributions for other variables.

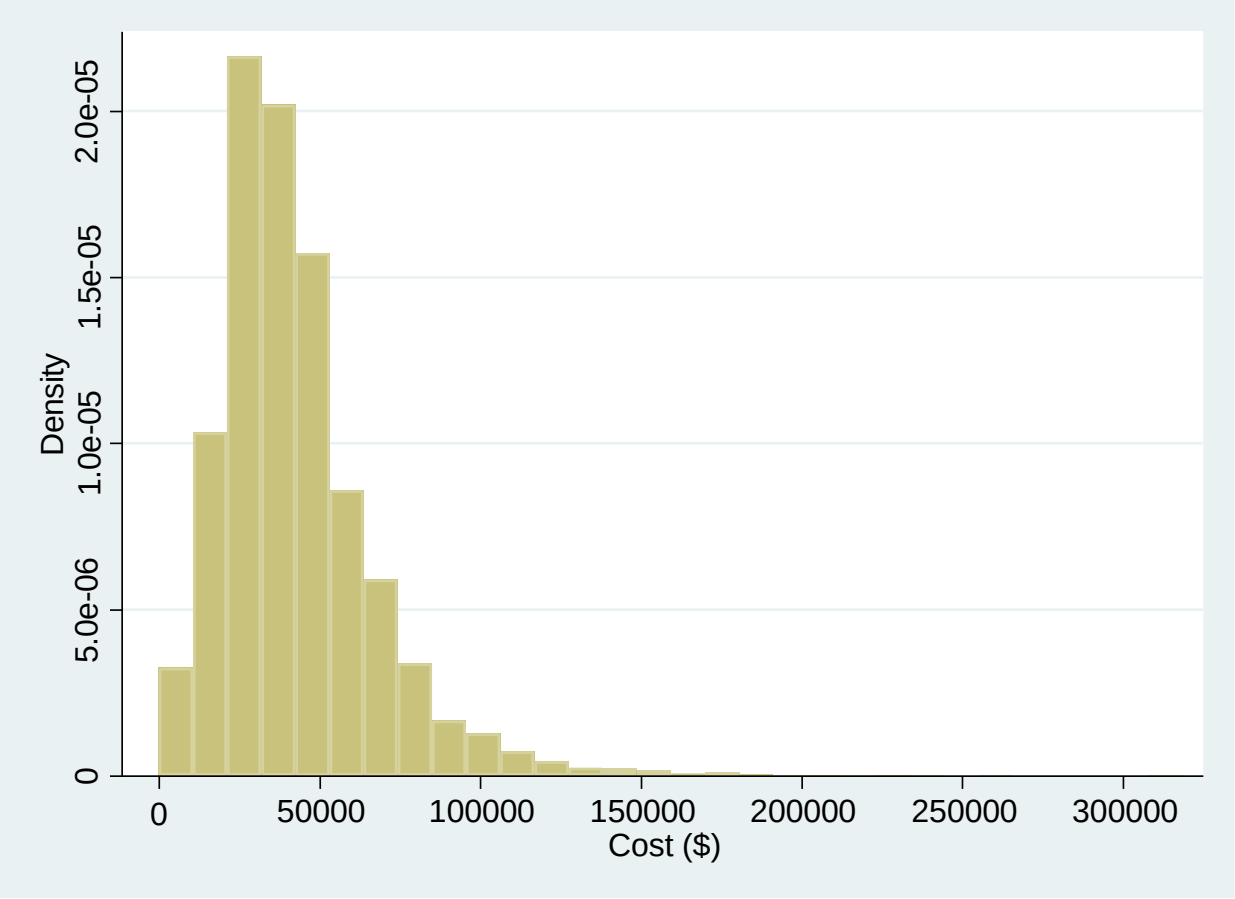
Patient Level Cost-effectiveness Analysis (CEA)

- Compare cost and effectiveness of two (or more) treatments.
- Patient level costs and measure of effectiveness.

Cost-effectiveness analysis

- **Outcome** from study -
  Survival, life years gained, quality-adjusted life years (QALYs)
- QALYs calculated by multiplying life years gained by utility, a measure of health status scaled from zero (death) to one (perfect health).
- **Cost** of treatment – procedures, physician, medications, hospitalizations, office visits, lab tests.

Table 2. Acute CVD Event Costs (Per Episode)						
Costs (2018 USD) by age group	Mean	Standard Deviation	Minimum	Maximum	Distribution	Reference
Fatal CVD event						
45-64	$18,940	$5,175	$11,921	$32,205	Gamma	(4,5)
65-84	$17,473	$4,567	$11,518	$29,420	Gamma	
≥85	$11,970	$3,394	$7,620	$20,926	Gamma	
Myocardial infarction						
45-64	$22,542	$3,908	$14,353	$29,672	Gamma	(4,5)
65-84	$22,410	$4,176	$14,241	$30,612	Gamma	
≥85	$14,159	$3,822	$8,288	$23,270	Gamma	


Density
0
5.0e-06
1.0e-05
1.5e-05
2.0e-05
0
50000
100000
150000
200000
250000
300000
Cost ($)

Supplemental Table 4. Cost-effectiveness analysis utility values.

Description	Value	Distribution	SD	Reference
Chronic Utility				
No history of CVD	1.0000	Beta	----	18,19
History of angina	0.9064	Beta	0.0223	18,19
History of MI	0.9648	Beta	0.0287	18,19
History of stroke	0.8835	Beta	0.0241	18,19
History of MI & stroke	0.8524	Beta	0.0359	18,19
Acute Disutility*				
Angina	-0.0078	Beta	0.0021	18,19
PCI	-0.0096	Beta	0.0028	18,19
CABG	-0.0192	Beta	0.0046	18,19
MI	-0.0079	Beta	0.0022	18,19
Stroke	-0.0113	Beta	0.0031	18,19

* Disutilities applied if patient experienced event or revascularization during the trial.
SD = standard deviation; CVD = cardiovascular disease; MI = myocardial infarction; PCI = percutaneous coronary intervention; CABG = coronary artery bypass grafting.

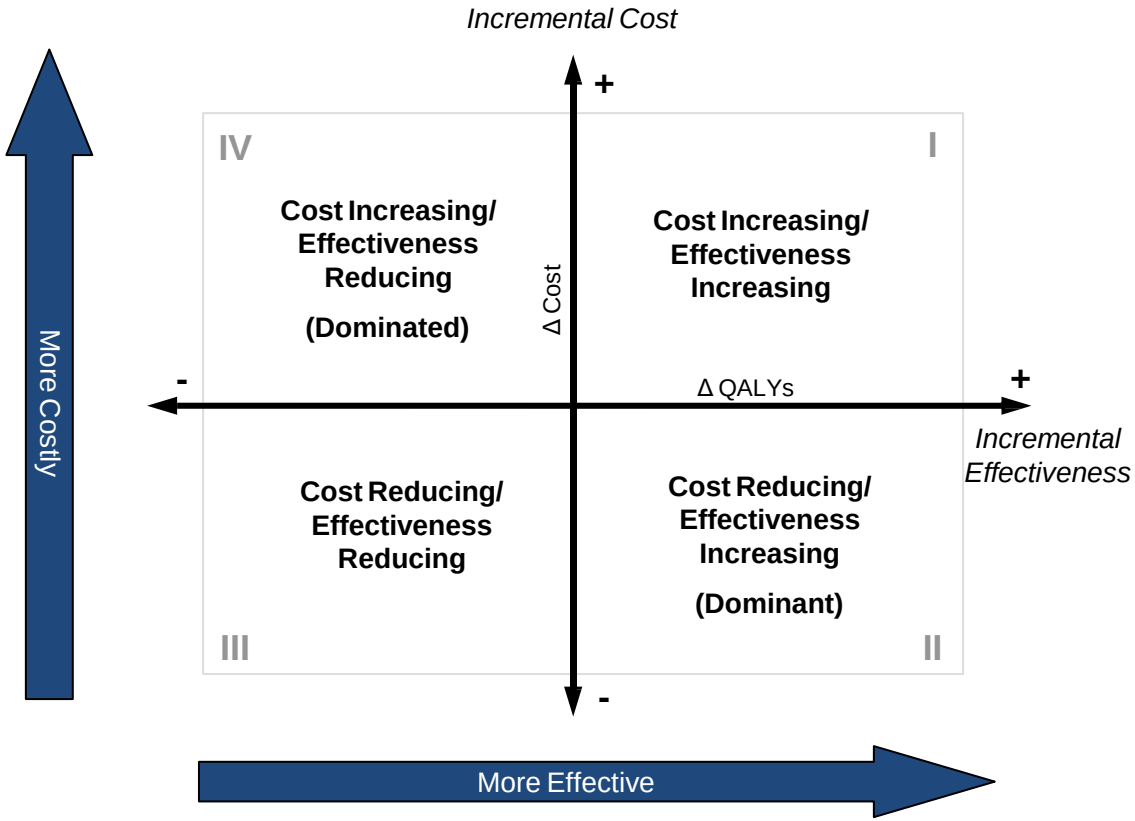
Incremental Cost-Effectiveness Ratio (ICER)

$$ICER = \frac{\mu CostA - \mu CostB}{\mu EffA - \mu EffB}$$

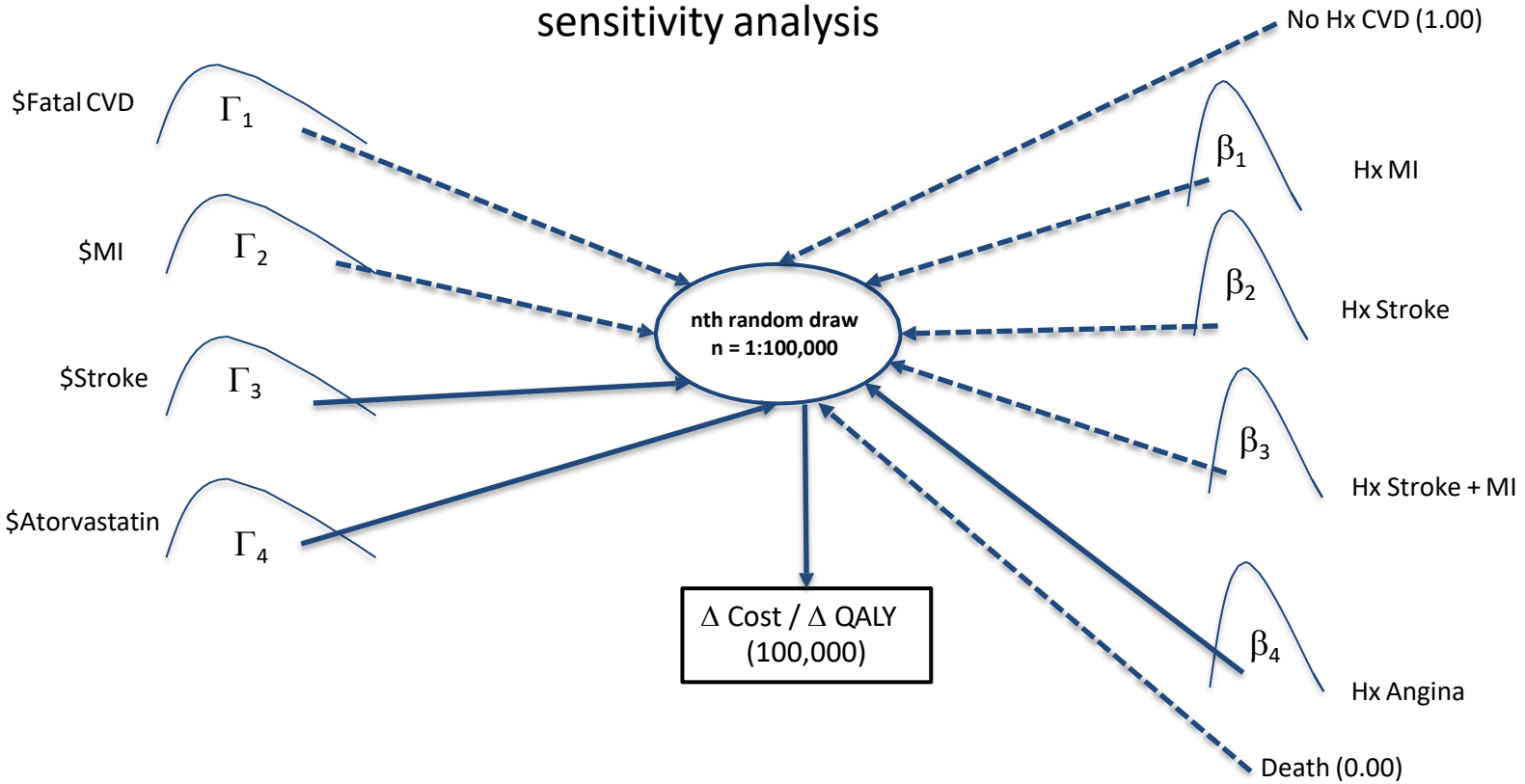
$$= \frac{\mu \Delta C}{\mu \Delta E}$$

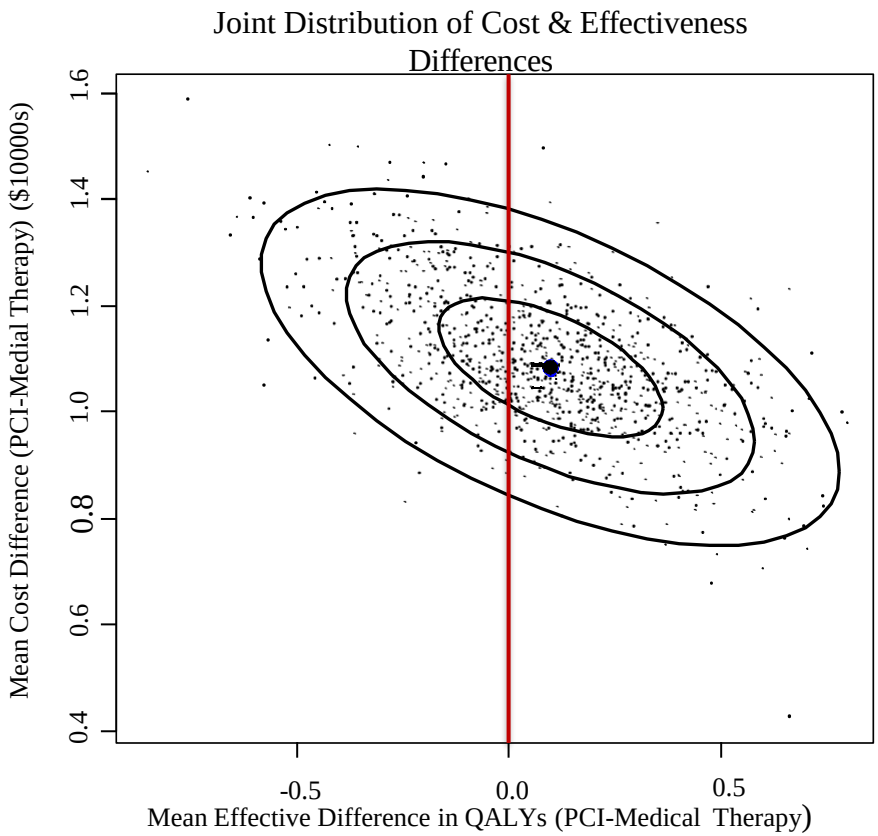
Spend $50,000 for every QALY gained.

Cost-effectiveness plane



Probabilistic (Bayesian) sensitivity analysis




Joint Distribution of Cost & Effectiveness Differences
Mean Cost Difference (PCI-Medial Therapy) ($10000s)
0.4
0.6
0.8
1.0
1.2
1.4
1.6
-0.5
0.0
0.5
Mean Effective Difference in QALYs (PCI-Medical Therapy)

Summary

- Frequentist: based on repeated sampling of data from population.
- Confidence interval: over repeated sampling, interval will
  - include population parameter X% of the time (95%, 99%).
- Bayesian: prior distribution → data → posterior distribution.
- Credible interval: X probability (e.g., .95) that interval includes population parameter.

Summary

- Frequentist methods used for simple or straightforward analyses.
- Bayesian methods can handle complex, multiparameter analyses difficult for frequentist methods.
- Bayesian simulation: first 10,000-50,000 iterations are “burn-in”.

References:

Kapur J, Elm J, Chamberlain, JM, et. al. Randomized Trial of Three Anticonvulsant Medications for Status Epilepticus. *New England Journal of Medicine* 2019;381:2103-2113.

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian Data Analysis, 3rd edition, CRC Press, Boca Raton, FL, 2014.

Gray K, Hampton B, Silveti-Falls T, McConnell A, Bausell C. Comparison of Bayesian Credible Intervals and Frequentist Confidence Intervals. *Journal of Modern Applied Statistical Methods* 2015;14:43-52.

Thank you

paul.kolm@medstar.net