Evidence-based medicine: critical interpretation of clinical trials

BHS Course 23/04/2022

KM Wissing MD PhD MSc
Department of Nephrology
Universitaire Ziekenhuis Brussel

Universitair Ziekenhuis Brussel

Vrije Universiteit Brussel

What is medical evidence

- Evidence-based medicine: Medical decision making based on adequately designed and conducted research
  - → Elaboration of EB guidelines and policies
  - → Clinical decision making
  - → Medical education
- Guideline development
  - → Systematic retrieval of evidence on a specific question
  - → Critical appraisal (bias, confounding, sample size, treatment effects, clinical relevance etc….)

Universitair Ziekenhuis Brussel

Vrije Universiteit Brussel

Critical reading of articles is necessary for all physicians.

- Some key issues every doctor has to know
  - → Sample size
  - → Statistical power
  - → The significance of statistical "non-significance"
  - → What is bias ?
  - → What is confounding ?
  - → Does the study population represent the general population
  - → Can the study results be extrapolated to specific populations

Universitair Ziekenhuis Brussel

Vrije Universiteit Brussel

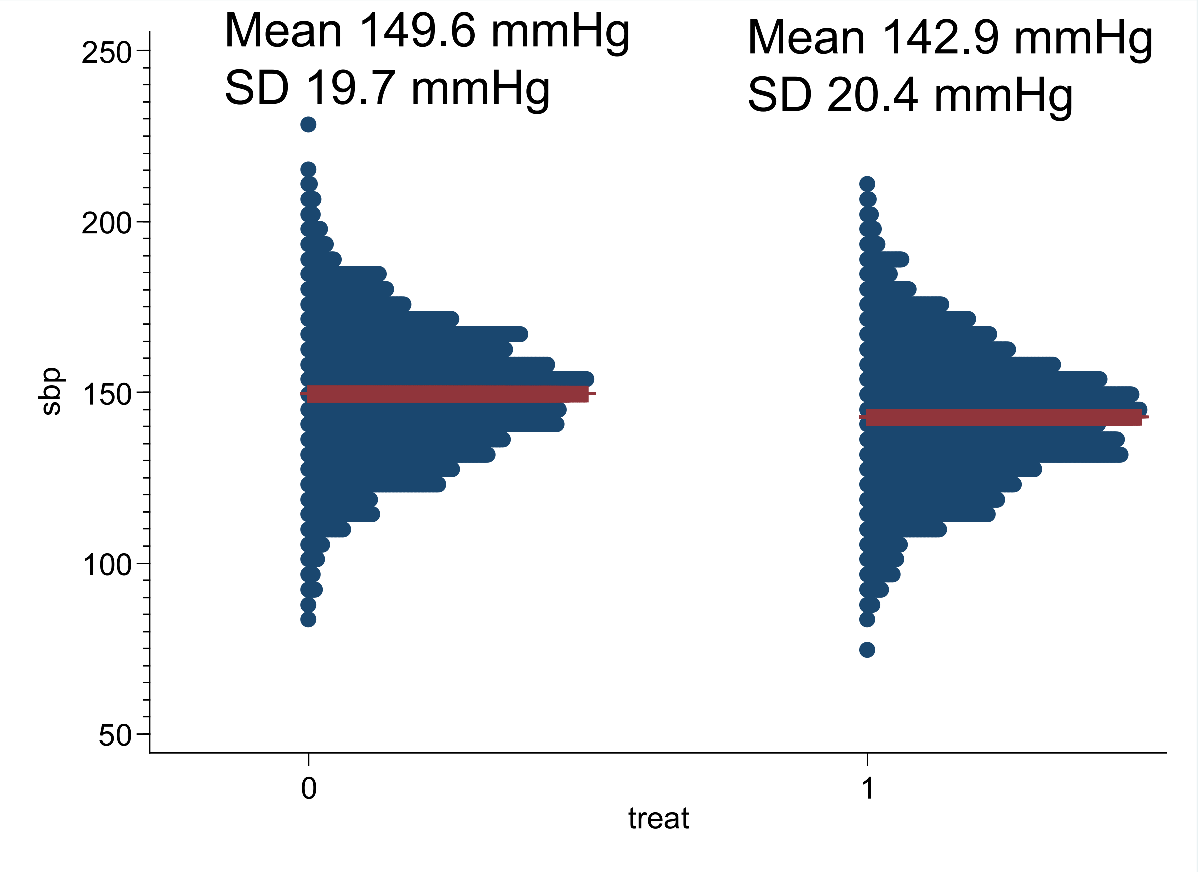
Sample properties, confidence intervals and hypothesis testing

- To what extend a sample reflects the characteristics of the underlying population
- How can we estimate the true effect size of an intervention based on a sample of patients
- To what extend sample properties differ due to random effect or differences in the characteristics of subjects ?
- To what extend lack of significant difference between samples reflects “equivalence”?

Universitair Ziekenhuis Brussel

Vrije Universiteit Brussel

Randomly generated populations of patients with arterial hypertension



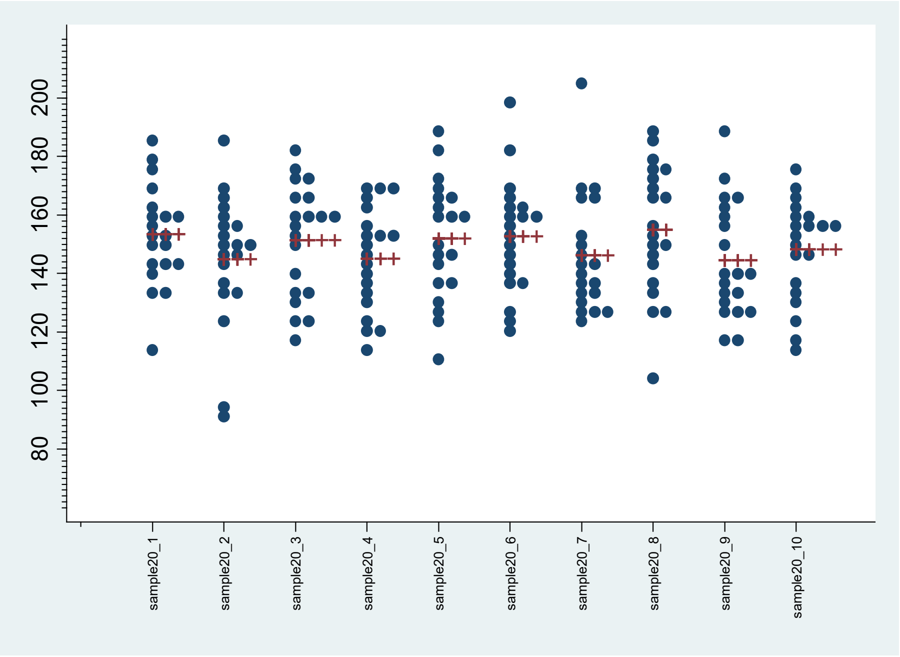
Randomly generated populations of 2500 with normal distribution and SD of 20

Universitair Ziekenhuis Brussel

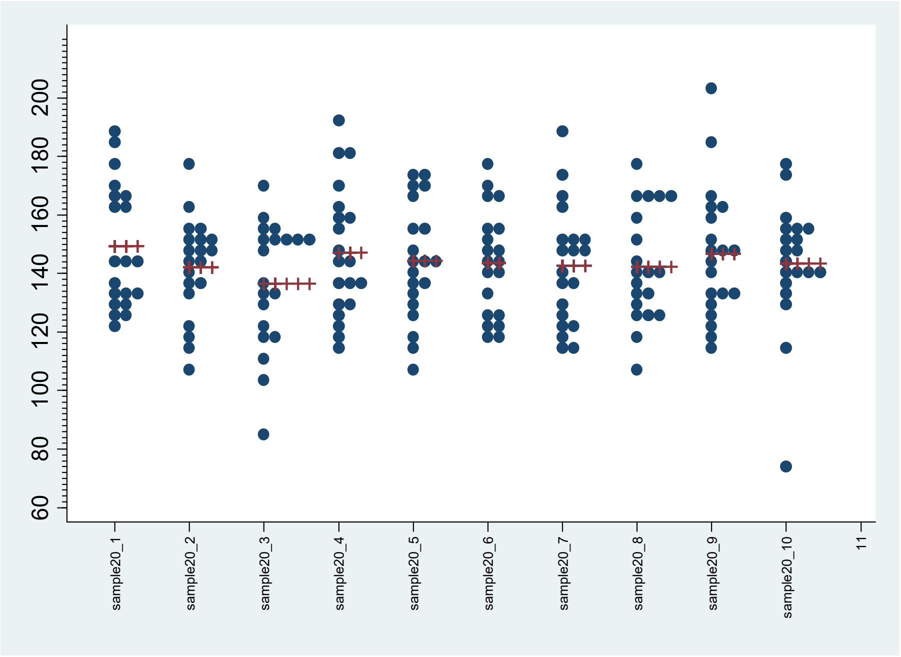
Vrije Universiteit Brussel

10 random samples of 20 form each of the two populations

No treatment

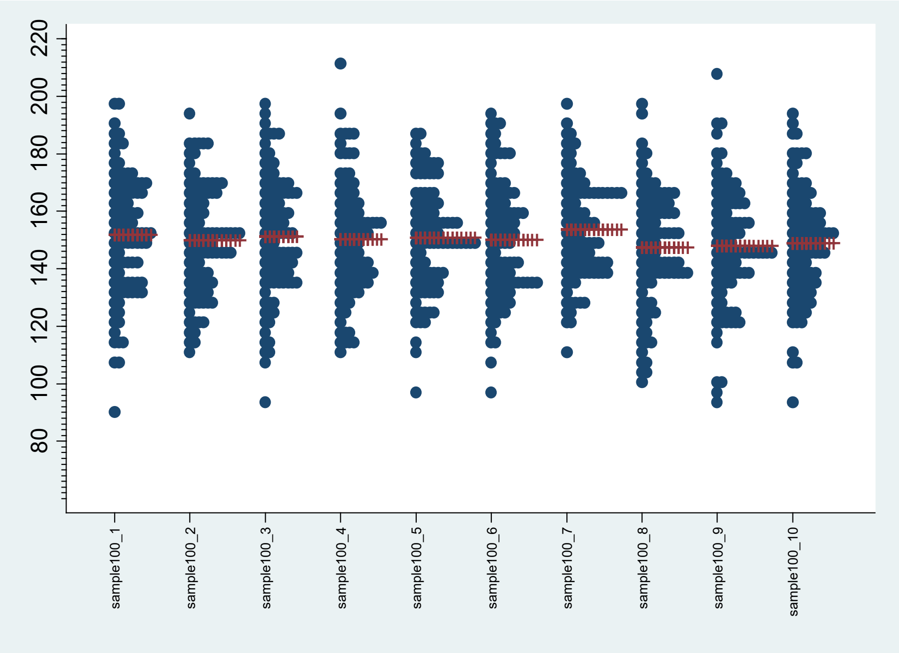


Treatment

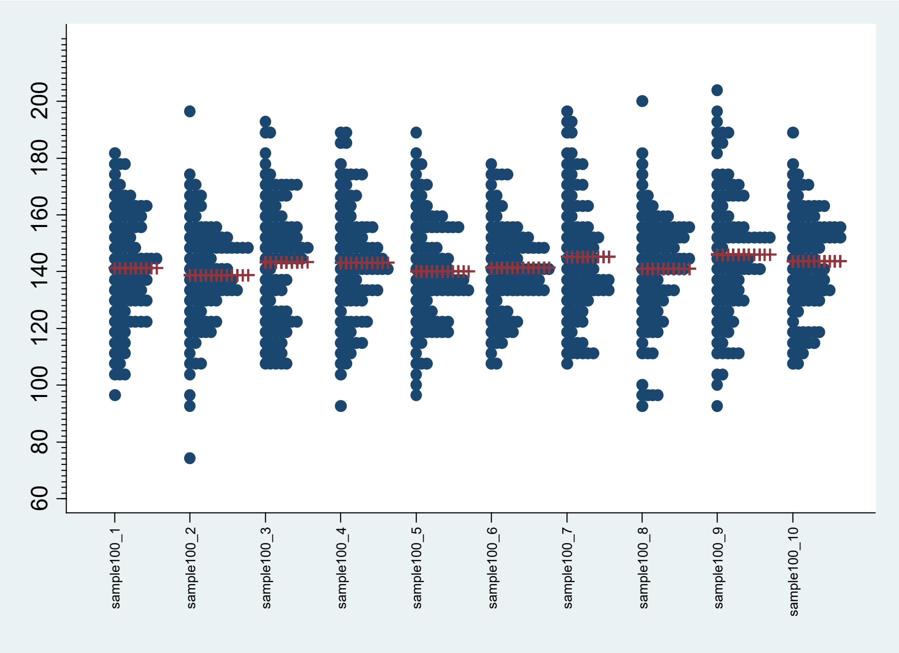


10 random samples of 100 form each of the two populations

No treatment



Treatment



Universitair Ziekenhuis Brussel

Vrije Universiteit Brussel

Effect of sample size on effect size estimates and the variability between estimates

Samples of 20

	No treatment		Treatment		Delta
	mmHg	SD	mmHg	SD	mmHg
sample 1	153.5	17.2	149.5	21.8	4.0
sample 2	144.9	22.9	142.2	16.7	2.7
sample 3	151.5	20.0	136.5	21.4	14.9
sample 4	145.1	17.5	147.2	22.0	-2.1
sample 5	151.9	20.3	144.4	19.9	7.6
sample 6	152.7	19.6	143.7	17.9	9.0
sample 7	146.1	20.2	142.8	21.0	3.4
sample 8	155.1	22.1	142.3	19.1	12.7
sample 9	144.5	19.7	146.8	22.3	-2.3
sample 10	148.2	17.1	143.4	21.7	4.8
Mean	149.4	19.7	143.9	20.4	5.5

Samples of 100

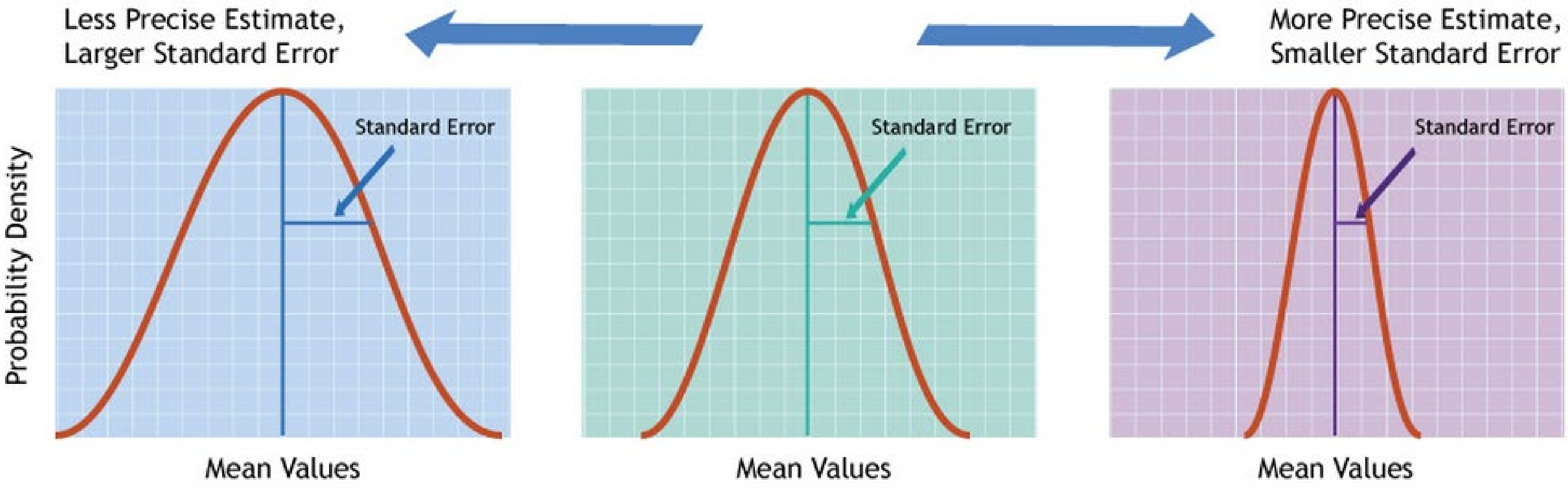
	No treatment		Treatment		Delta
	mmHg	SD	mmHg	SD	mmHg
sample 1	151.8	21.0	141.1	19.7	10.7
sample 2	149.8	18.5	138.7	19.1	11.1
sample 3	151.3	20.8	143.3	21.3	8.0
sample 4	150.2	20.4	143.1	20.5	7.1
sample 5	150.9	18.3	140.1	19.1	10.8
sample 6	150.1	20.7	141.4	15.6	8.7
sample 7	153.7	17.7	145.2	21.6	8.5
sample 8	147.4	20.3	141.1	19.7	6.3
sample 9	148.0	20.5	146.1	23.2	1.9
sample 10	148.9	19.0	143.6	18.0	5.2
Mean	150.2	19.7	142.4	19.8	7.8

Universitair Ziekenhuis Brussel

Vrije Universiteit Brussel

What conditions the precision of the estimate of the population mean

- Dispersion of the sample distribution (SEM):
  → Variabilty of the underlying population (estimated form sample SD)
  → Size of the sample (N)
  → SEM= SD/sqrt(N)



http://www.hcup-us.ahrq.gov/tech_assist/standarderrors/508/508course.html

Standard error of the mean

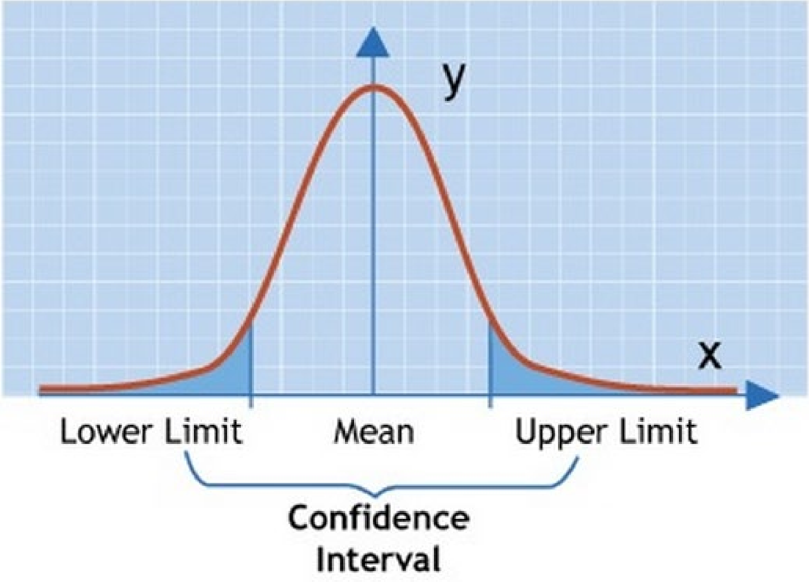
- Distribution of sample means
  - → Mean value identical to population mean
  - → Generally normally distributed even if the underlying population is not normally distributed
  - → Standard deviation of sampling distribution = standard error of the mean (SEM)
- **Standard deviation**: measures the dispersion of individual measures around the population (sample) mean
- **Standard error of mean (SEM)**: measures dispersion of sample means around the population mean

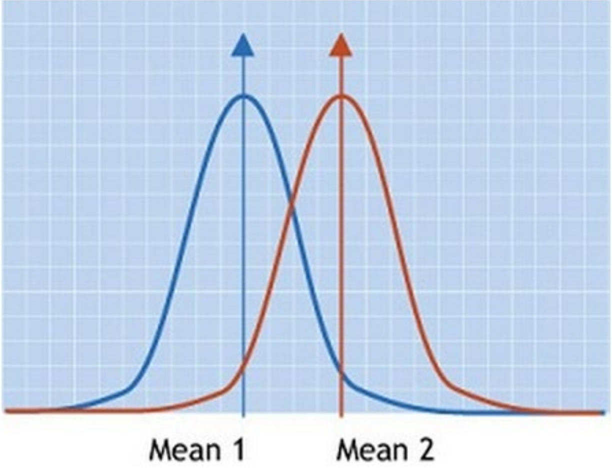
Universitair Ziekenhuis Brussel

Vrije Universiteit Brussel

Standard error of the mean

- Essential for descriptive statistics and hypothesis testing
- Used to calculate the 95% (99%) confidence interval
- Used to assess whether two samples are unlikely to differ due to chance alone (hypothesis testing)





Hypothesis testing

- Are there differences between the two samples?
- Null hypothesis is tested in most statistical tests.
  - → Underlying hypothesis: the two samples are drawn from the same population and variability is due to random error.
  - → P: probability of the null hypothesis
  - → Rejection of null hypothesis based on type I error (probability to reject null hypothesis although the two samples originate from the same population.
  - → P ≥ cut-off: Probability of null hypothesis too high to excluded that differences in samples are due to chance
  - → P< cut-off: Differences in samples are unlikely to be due to chance alone.

Universitair Ziekenhuis Brussel

Vrije Universiteit Brussel

Choice of cut-off for type I error

- Arbitrary decision and a compromise between two types of false conclusions
  - → Cut-off (P) high (ex $P<0.1$): implies higher risk to consider chance events as significant (equivalent type I error)
  - → Cut-off (P) low ($P<0.01$): implies a higher risk to consider differences in samples as chance effects although this is not the case (equivalent to type II error – lack of power)

Universitair Ziekenhuis Brussel

Vrije Universiteit Brussel

Impact of sample size on power to detect an existing difference in population means

Samples of 20 / Samples of 100

	No treatment		Treatment		Delta	P	No treatment		Treatment		Delta	P
	mmHg	SEM	mmHg	SEM	mmHg		mmHg	SEM	mmHg	SEM	mmHg	
sample 1	153.5	3.9	149.5	4.9	4.0	0.52	151.8	2.1	141.1	2.0	10.7	0.0003
sample 2	144.9	5.1	142.2	3.7	2.7	0.67	149.8	1.9	138.7	1.9	11.1	<0.0001
sample 3	151.5	4.5	136.5	4.8	14.9	0.03	151.3	2.1	143.3	2.1	8.0	0.008
sample 4	145.1	3.9	147.2	4.9	-2.1	0.74	150.2	2.0	143.1	2.1	7.1	0.015
sample 5	151.9	4.5	144.4	4.4	7.6	0.24	150.9	1.8	140.1	1.9	10.8	0.0001
sample 6	152.7	4.4	143.7	4.0	9.0	0.14	150.1	2.1	141.4	1.6	8.7	0.001
sample 7	146.1	4.5	142.8	4.7	3.4	0.61	153.7	1.8	145.2	2.2	8.5	0.003
sample 8	155.1	4.9	142.3	4.3	12.7	0.06	147.4	2.0	141.1	2.0	6.3	0.027
sample 9	144.5	4.4	146.8	5.0	-2.3	0.73	148.0	2.1	146.1	2.3	1.9	0.53
sample 10	148.2	3.8	143.4	4.9	4.8	0.44	148.9	1.9	143.6	1.8	5.2	0.048
Mean	149.4	4.4	143.9	4.6	5.5		150.2	2.0	142.4	2.0	7.8	

Universitair Ziekenhuis Brussel

Vrije Universiteit Brussel

Concept of type II error and statistical power

- Type II error: Probability to accept the null hypothesis of no difference between the two samples although the samples originate from different populations.
- This corresponds to the probability of being unable to detect an existing difference.
- Previous example:
  - → Samples of 20: type II error of 80%
  - → Samples of 100: type II error of 10%
- Power: 100% - type II error

Universitair Ziekenhuis Brussel

Vrije Universiteit Brussel

A particular issue for secondary endpoints

- Primary efficacy endpoint: The treatment effect that is used to determine the sample size
- Secondary endpoints: Important treatment effects that are assessed without adaptation of sample size
- Most studies are underpowered for serious complication and side effects which are fortunately rare events.

Universitair Ziekenhuis Brussel

Vrije Universiteit Brussel

Example: Hypertension trial

- Primary efficacy endpoint: Detection of a reduction of blood pressure by at least 7 mmHg
- Sample size of 130 per group to obtain 80% power with type I error of 5%
- Deaths 7/130 in treatment group and 3/130 in placebo group; P=0.33
- Hospitalisation with thrombotic events 6/130 in treatment group and 3/130 in placebo group; P=0.5
- **Conclusion: significant reduction in blood pressure. No significant differences in serious complications during 12 months follow up.**

Universitair Ziekenhuis Brussel

Vrije Universiteit Brussel

Lack of significance does not mean absence of effect

- One of the most frequent spurious interpretations:
  - → P=NS interpreted as “no difference between groups”
  - → “We can not exclude a chance effect to explain the variability in samples ” ≠ “ Values of the two samples are the same”
  - → Non-rejection of the null hypothesis corresponds to saying: “We can not exclude a chance effect with a sufficient degree of confidence but this does exclude that another reasons for variability between the two samples exist”

Universitair Ziekenhuis Brussel

Vrije Universiteit Brussel

Bias

- What is bias
  - → "Any systematic error in the design, conduct or analysis of a study that results in a mistaken estimate of an exposure's effect on the risk of disease"
- Two main categories of bias
  - → **Selection bias**: patients in control group differ in a way that makes evaluation of exposure of interest impossible
  - → **Information bias**: Systematic differences in the collection of exposure and outcome data in different patient categories.

Universitair Ziekenhuis Brussel

Vrije Universiteit Brussel

Bias example: Effect of smoking on the incidence of ESRD a case control study

- Dr. Wissing is interested in the effect of smoking on ESRD.
- Insufficient time and money to conduct a cohort study
- Case control design: Hospital unit which shares patients from Nephrology and Pneumology
- 100 last ERSD patients (GFR<15 ml/min) compared with 100 pneumology admissions with normal renal function randomly selected during the same period. Smoking history collected in medical files.

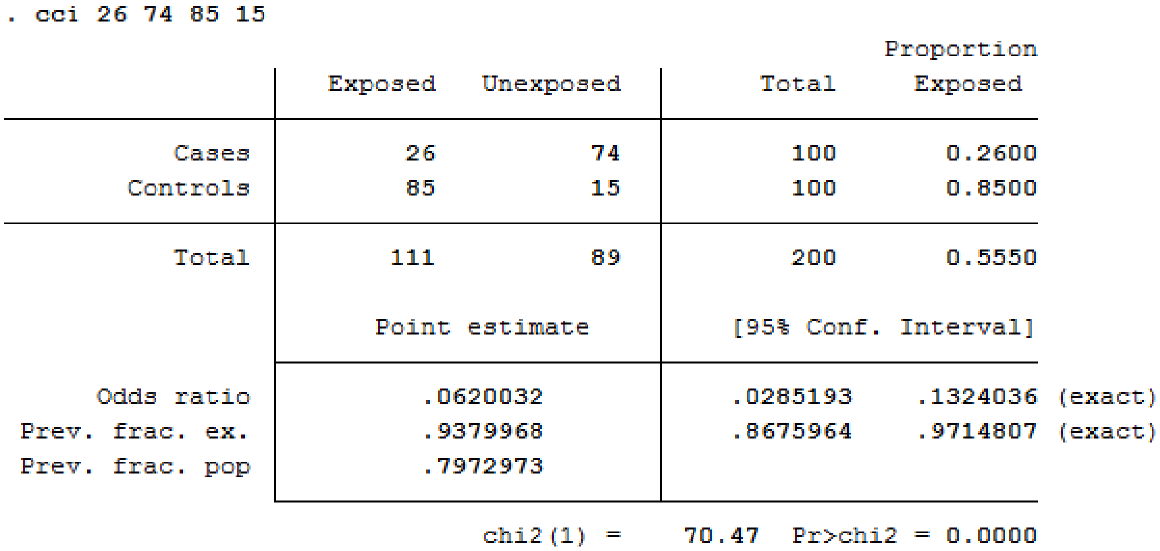
Universitair Ziekenhuis Brussel

Vrije Universiteit Brussel

Bias: “Smoking is a highly efficient protection against the development of ESRD”

	ESRD (Nephro)	Control (Pneum)	
Smoking	26	85	111
Non-smoking	74	15	89
	100	100	

Exposure Odds ratio: 26x15/85x74= 0.062 (95%CI 0.03 to 0.13; P<0.0001)



```
. cci 26 74 85 15
                                                          Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+------------------------
           Cases |        26          74  |        100       0.2600
        Controls |        85          15  |        100       0.8500
-----------------+------------------------+------------------------
           Total |       111          89  |        200       0.5550
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
      Odds ratio |         .0620032       |    .0285193     .1324036 (exact)
 Prev. frac. ex. |         .9379968       |    .8675964     .9714807 (exact)
 Prev. frac. pop |         .7972973       |
                 +-------------------------------------------------
                               chi2(1) =     70.47  Pr>chi2 = 0.0000
```

Vrije Universiteit Brussel

Bias

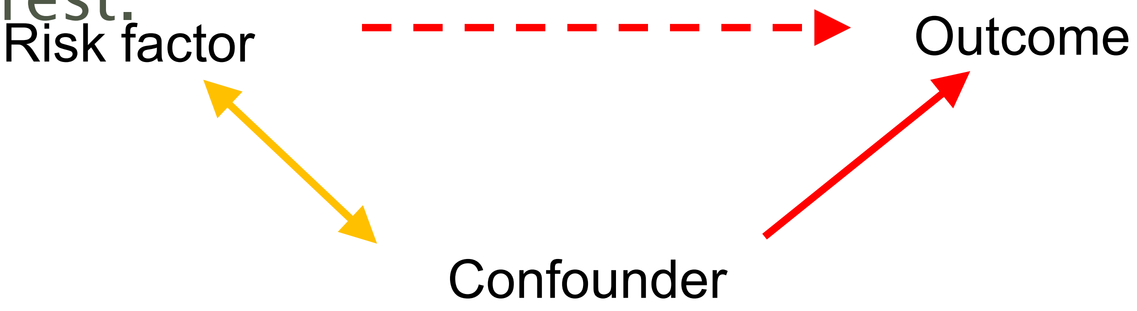
- Selection bias
  - → Choosing pneumology inpatients as controls artificially enriched the control population for the exposure of interest
- Information bias
  - → Pneumology patients are more likely to have been asked about their smoking habits and to have this information in their medical file. This can result in underestimation of smoking in the Cases (**misclassification**).
  - → Even if patients were asked pneumology patients are probably more prone to indicate smoking (**recall bias**)
- Bias cannot be recovered at the analysis stage: Creation of false evidence, waste of resources, study results to be discarded.
  - → Every study protocol has to be screened for potential sources of bias during the design phase.

Universitair Ziekenhuis Brussel

Vrije Universiteit Brussel

Confounding

- Very frequent in studies that are not randomized
- A confounder is a known risk factor for the outcome and associated with the exposure of interest.

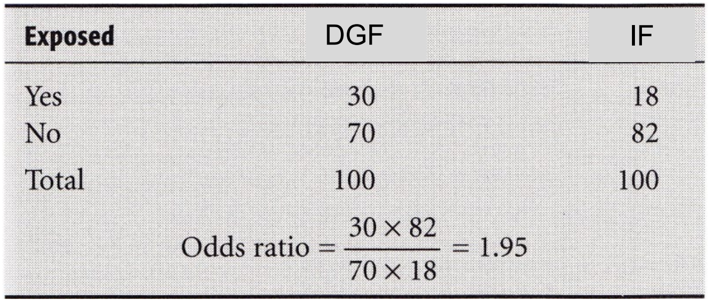


- Often spurious attribution of causal relationship in case of association

Confounding: Hypothetical example

Use of a new preservation fluid (exposure = Yes) is associated with a reduction in delayed graft function (DGF) ? Case-control study to address the question

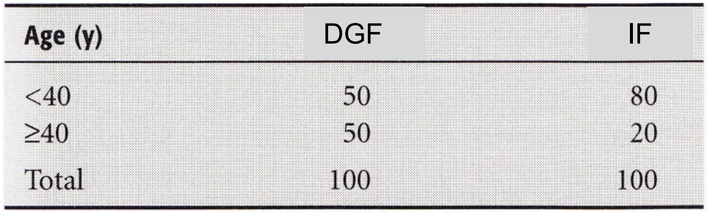
The odds of exposure is nearly 2 times higher in patients with DGF



Exposed	DGF	IF
Yes	30	18
No	70	82
Total	100	100

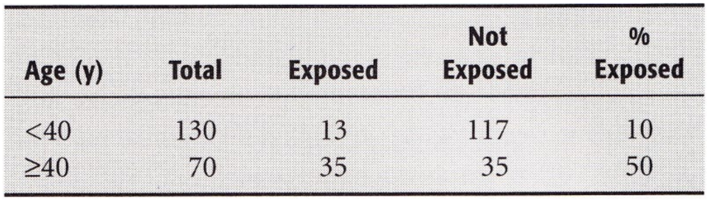
$$\text{Odds ratio} = \frac{30 \times 82}{70 \times 18} = 1.95$$

Donor age is the most important predictor of DGF



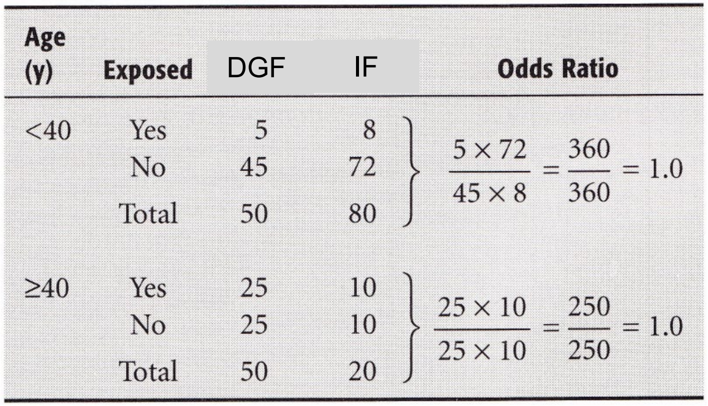
Age (y)	DGF	IF
<40	50	80
≥40	50	20
Total	100	100

Older donor age is associated with the use of the new preservation fluid



Age (y)	Total	Exposed	Not Exposed	% Exposed
<40	130	13	117	10
≥40	70	35	35	50

After stratification for donor age the preservation fluid is no longer associated with DGF



Age (y)	Exposed	DGF	IF	Odds Ratio
<40	Yes	5	8	$\frac{5 \times 72}{45 \times 8} = \frac{360}{360} = 1.0$
	No	45	72	
	Total	50	80	
≥40	Yes	25	10	$\frac{25 \times 10}{25 \times 10} = \frac{250}{250} = 1.0$
	No	25	10	
	Total	50	20	

A confounder is a known risk factor for the outcome and associated with the exposure of interest. Can be controlled for by stratification. You can only control for known confounders.

Post-hoc analysis and subgroup analysis

- Post-hoc analysis:
  - → Examining data for findings or effects that were not prespecified a priori but are analyzed after the study has been completed.
  - → "Data dredging": Multiple tests of association increase the risk to find a chance association below the threshold of type I error
  - → Often only positive results are shown and the reader does not know how many associations had to be tested to "obtain" the significant results

Universitair Ziekenhuis Brussel

Vrije Universiteit Brussel

Subgroup analysis

- Clinicians want to investigate how the data from large RCT can be applicated to the individual patient
  - → Effect in relation to disease severity
  - → Risk and efficacy in specific patient populations
  - → Dependence of effect on co-morbidities
  - → Dependence of effect on genetic variability
- Assess whether a treatment with limited efficacy in the overall population has benefits in sub-populations

Universitair Ziekenhuis Brussel

Vrije Universiteit Brussel

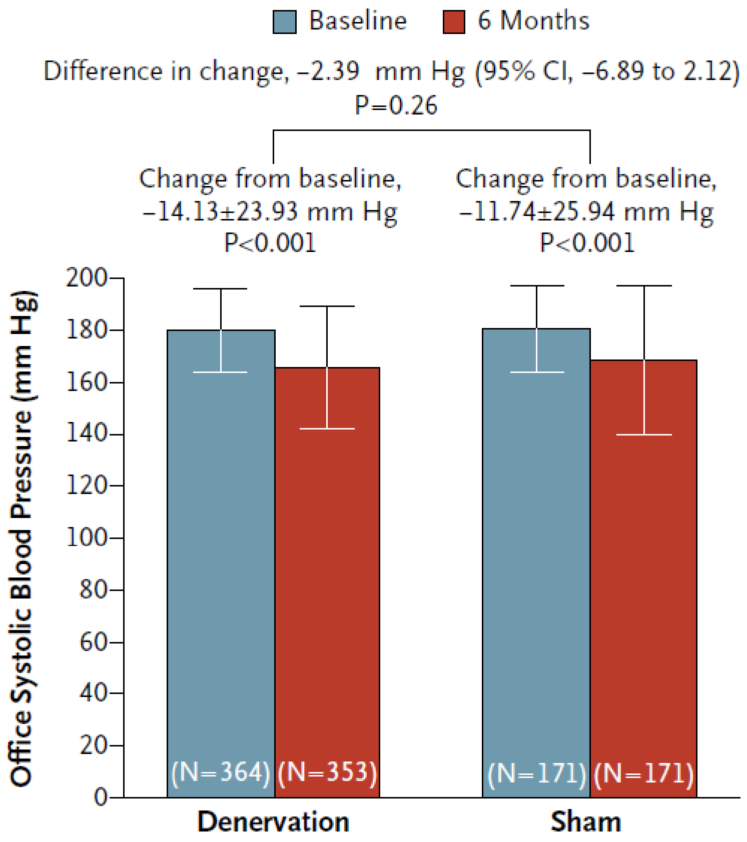
Rothwell PM Lancet 2005; 365:176

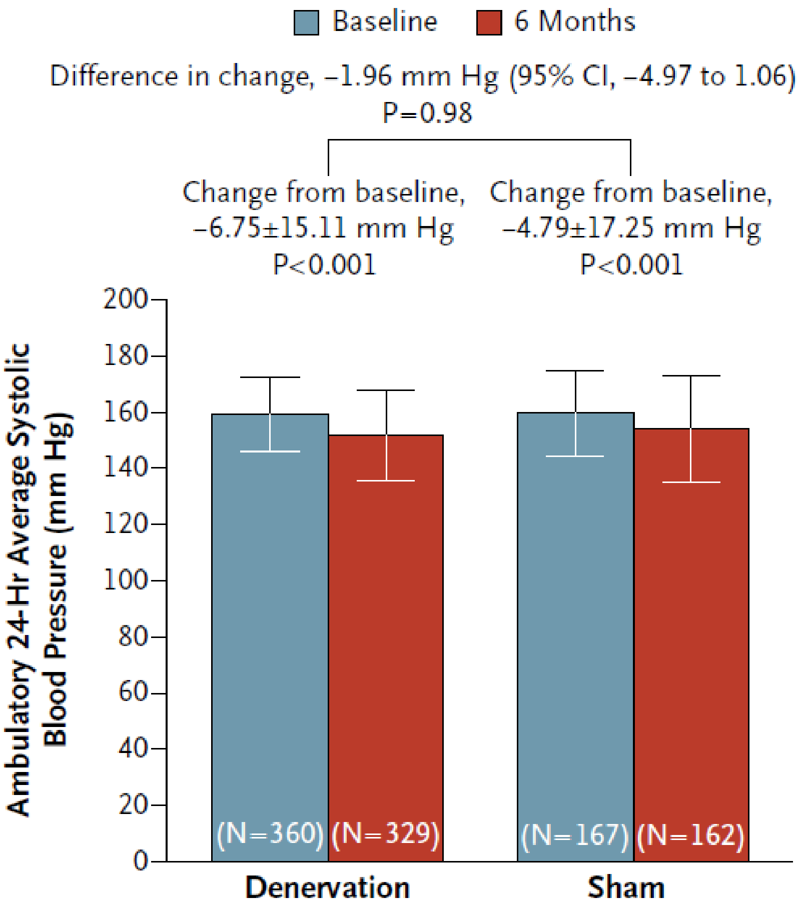
Simplicity HTN-3 trial

- Prospective and randomized, sham controlled, single-blind trial.
- Randomization 2:1 to renal denervation (N=364) versus sham procedure (N=171).
- Primary efficacy endpoint: change in office systolic blood pressure at 6 months.

Bhatt DL NEJM 2014; 370: 1393

Simplicity HTN-3 trial



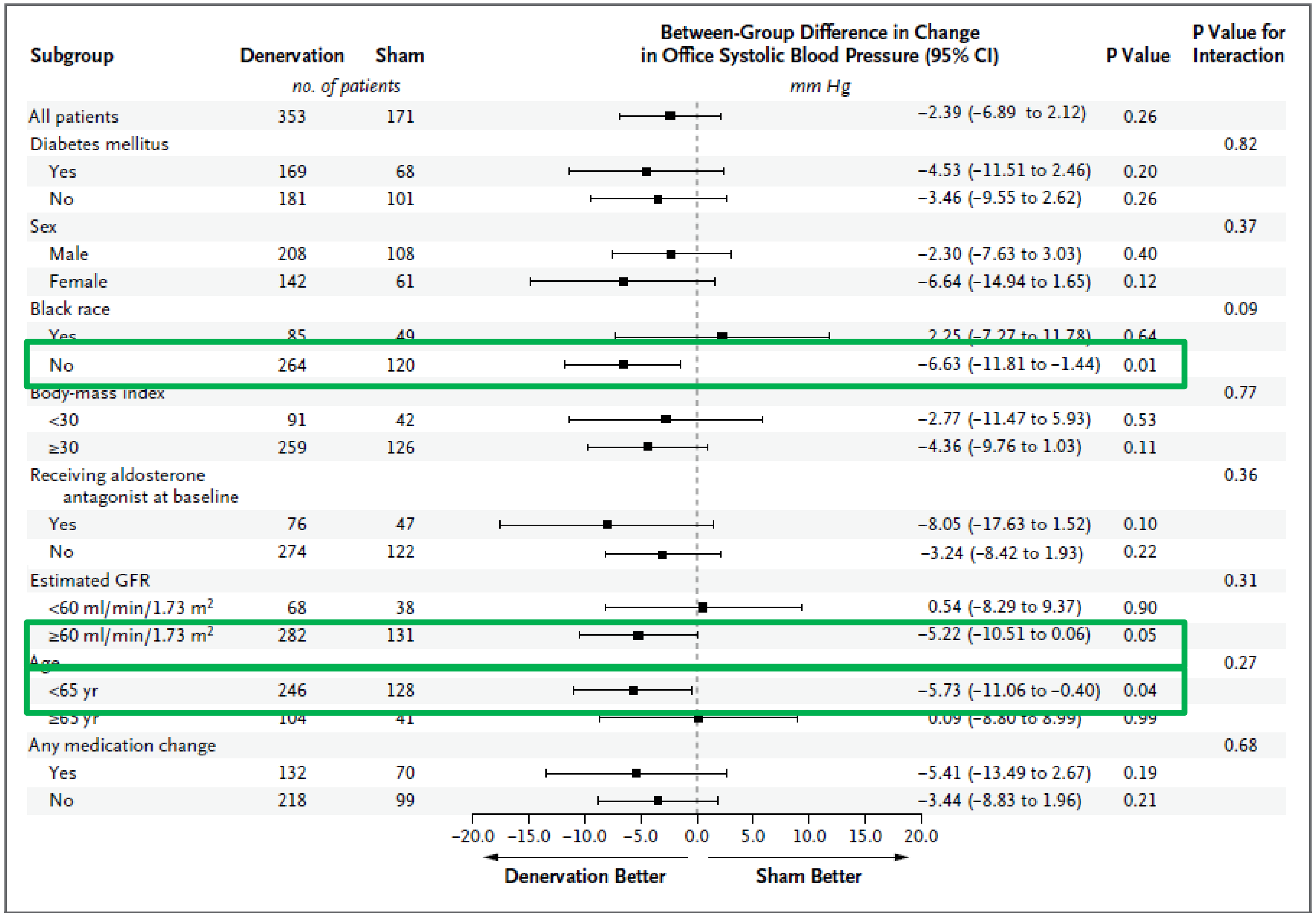


Universitair Ziekenhuis Brussel

Vrije Universiteit Brussel

Simplicity HTN-3 trial

The inevitable subgroup analysis…..



Subgroup	Denervation	Sham	Between-Group Difference in Change in Office Systolic Blood Pressure (95% CI)	P Value	P Value for Interaction
	no. of patients		*mm Hg*		
All patients	353	171	−2.39 (−6.89 to 2.12)	0.26	
Diabetes mellitus					0.82
Yes	169	68	−4.53 (−11.51 to 2.46)	0.20	
No	181	101	−3.46 (−9.55 to 2.62)	0.26	
Sex					0.37
Male	208	108	−2.30 (−7.63 to 3.03)	0.40	
Female	142	61	−6.64 (−14.94 to 1.65)	0.12	
Black race					0.09
Yes	85	49	2.25 (−7.27 to 11.78)	0.64	
No	264	120	−6.63 (−11.81 to −1.44)	0.01	
Body-mass index					0.77
<30	91	42	−2.77 (−11.47 to 5.93)	0.53	
≥30	259	126	−4.36 (−9.76 to 1.03)	0.11	
Receiving aldosterone antagonist at baseline					0.36
Yes	76	47	−8.05 (−17.63 to 1.52)	0.10	
No	274	122	−3.24 (−8.42 to 1.93)	0.22	
Estimated GFR					0.31
<60 ml/min/1.73 m^2	68	38	0.54 (−8.29 to 9.37)	0.90	
≥60 ml/min/1.73 m^2	282	131	−5.22 (−10.51 to 0.06)	0.05	
Age					0.27
<65 yr	246	128	−5.73 (−11.06 to −0.40)	0.04	
≥65 yr	104	41	0.09 (−8.80 to 8.99)	0.99	
Any medication change					0.68
Yes	132	70	−5.41 (−13.49 to 2.67)	0.19	
No	218	99	−3.44 (−8.83 to 1.96)	0.21	

Figure 3. Selected Subgroup Analyses.

Shown are between-group differences in the change in office systolic blood pressure from baseline to 6 months in selected subgroups. The body-mass index is the weight in kilograms divided by the square of the height in meters. GFR denotes glomerular filtration rate.

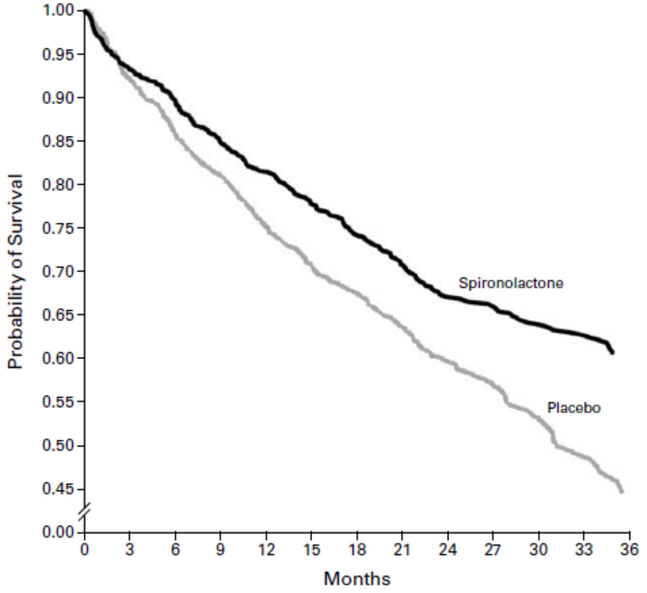
Can the study results be extrapolated to the general population

- RANDOMIZED ALDACTONE EVALUATION STUDY INVESTIGATORS (RALES) N Engl J Med 1999:341:709-17
  - → Effect of spironolactone on clinical outcomes in patients with severe heart failure
  - → LVEF<=35% NYHA class III-IV heart failure
  - → ACE inhibitor (if tolerated) and loop diuretic
  - → Exclusion criteria: valvular disease, renal failure and hyperkalemia
- Mean age in both groups (±800) 65 y

Universitair Ziekenhuis Brussel

Vrije Universiteit Brussel

RALES outcomes



No. at Risk

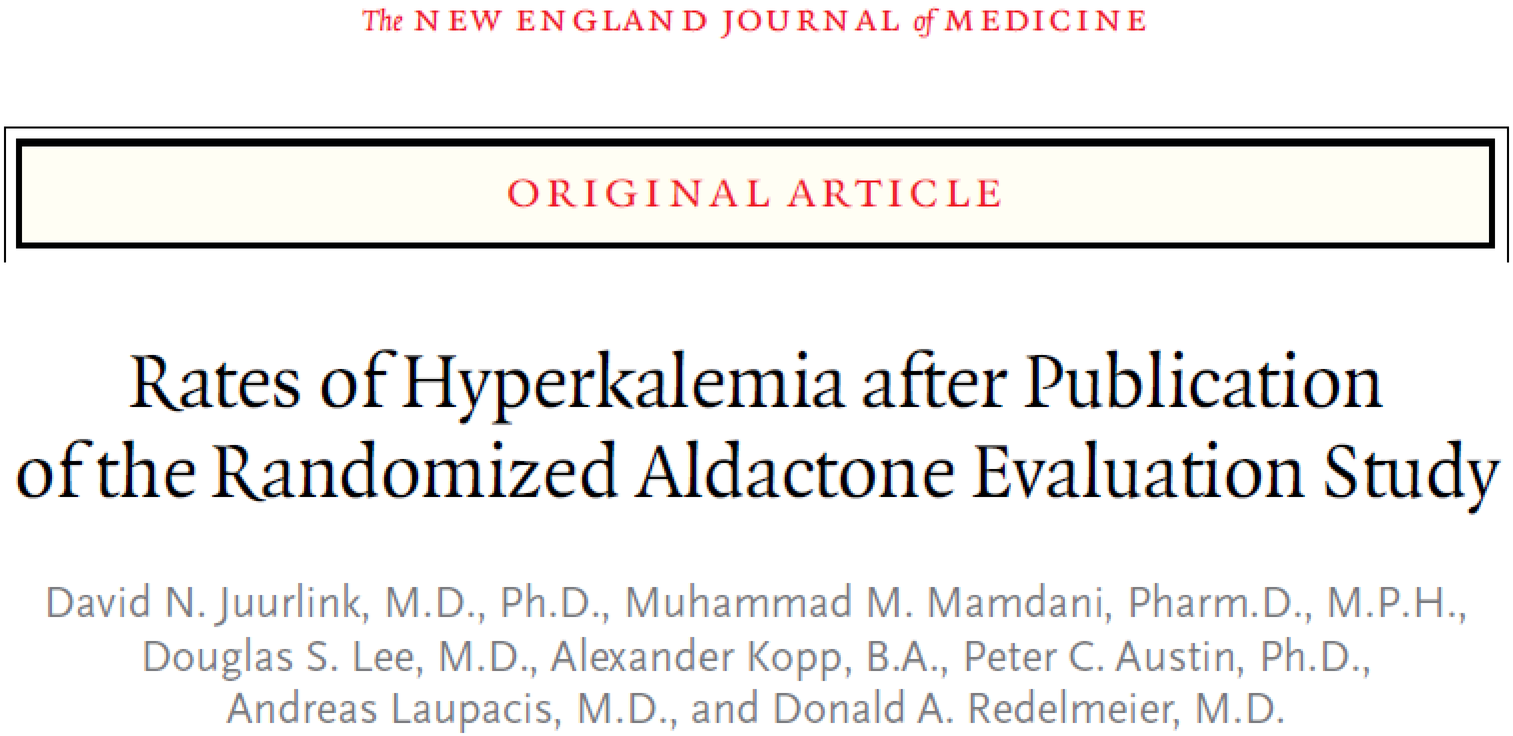
Placebo	841	775	723	678	628	592	565	483	379	280	179	92	36
Spironolactone	822	766	739	698	669	639	608	526	419	316	193	122	43

Figure 1. Kaplan–Meier Analysis of the Probability of Survival among Patients in the Placebo Group and Patients in the Spironolactone Group.

The risk of death was 30 percent lower among patients in the spironolactone group than among patients in the placebo group (P<0.001).

- RR death 0.7
- RR Worsening HF 0.64
- RR Sudden death 0.71
- Increase in K by 0.3 mEq
- Serious hyper K 2% vs 1% (P=0.42)

N Engl J Med 1999:341:709-17



The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

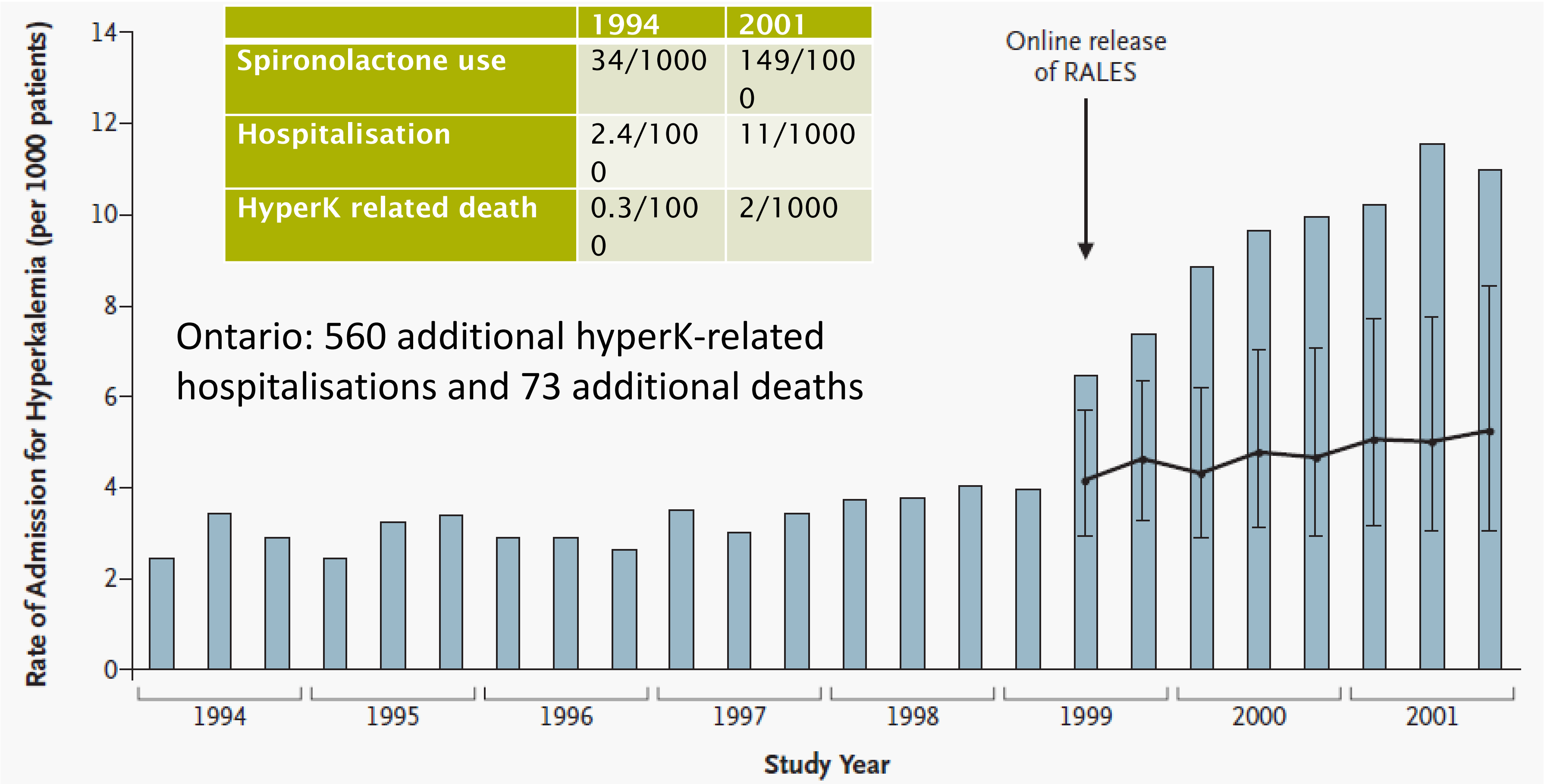
Rates of Hyperkalemia after Publication of the Randomized Aldactone Evaluation Study

David N. Juurlink, M.D., Ph.D., Muhammad M. Mamdani, Pharm.D., M.P.H., Douglas S. Lee, M.D., Alexander Kopp, B.A., Peter C. Austin, Ph.D., Andreas Laupacis, M.D., and Donald A. Redelmeier, M.D.

Population-based time-series analysis to examine trends in the rate ofspironolactone prescriptions and the rate of hospitalization for hyperkalemia in ambulatory patients before and after the publication of RALES

N Engl J Med 2004;351:543-51

Increase in spironolactone-realted hyperkalemia



	1994	2001
Spironolactone use	34/1000	149/1000
Hospitalisation	2.4/1000	11/1000
HyperK related death	0.3/1000	2/1000

Ontario: 560 additional hyperK-related hospitalisations and 73 additional deaths

N Engl J Med 2004;351:543-51

Take home messages

- Statistical tests evaluate the probability that two samples of patients originate from the same background population.
- The standard deviation expresses the variability between individuals in the sample.
- The standard error of the mean expresses the variability of sample means in relation to the population mean.
- The standard error of the mean depends both on between-subject variability and sample size.

Universitair Ziekenhuis Brussel

Vrije Universiteit Brussel

Take home messages

- A P value above the significance threshold means that differences in sample properties could be observed by chance. This does not exclude that sample properties might be different for other reasons.
- Statistical power expresses the probability that the null hypothesis will be rejected in case two samples originate indeed from different populations.
- Bias is a systematic error in the design of the study that results in mistaken estimates of the effect of an exposure (risk factor) on outcome. Bias cannot be controlled for at the analysis stage.
- A confounder is a known risk factor for the outcome that is associated with the exposure of interest. Confounding can be controlled for by stratification.

Universitair Ziekenhuis Brussel

Vrije Universiteit Brussel