Sample size, cluster randomization and (perhaps) precision

Henrik Støvring
Steno Diabetes Center Aarhus - Denmark
hersto@rm.dk

SDCA, Feb 22, 2024

Overview

- A basic example
- Rationale and statistical terminology
- Some more advanced topics
  - cluster sampling
  - when power is irrelevant
  - stochastic simulations (basic concepts revisited)

- A shopping list when planning your next study

Example: A new weight-loss pill

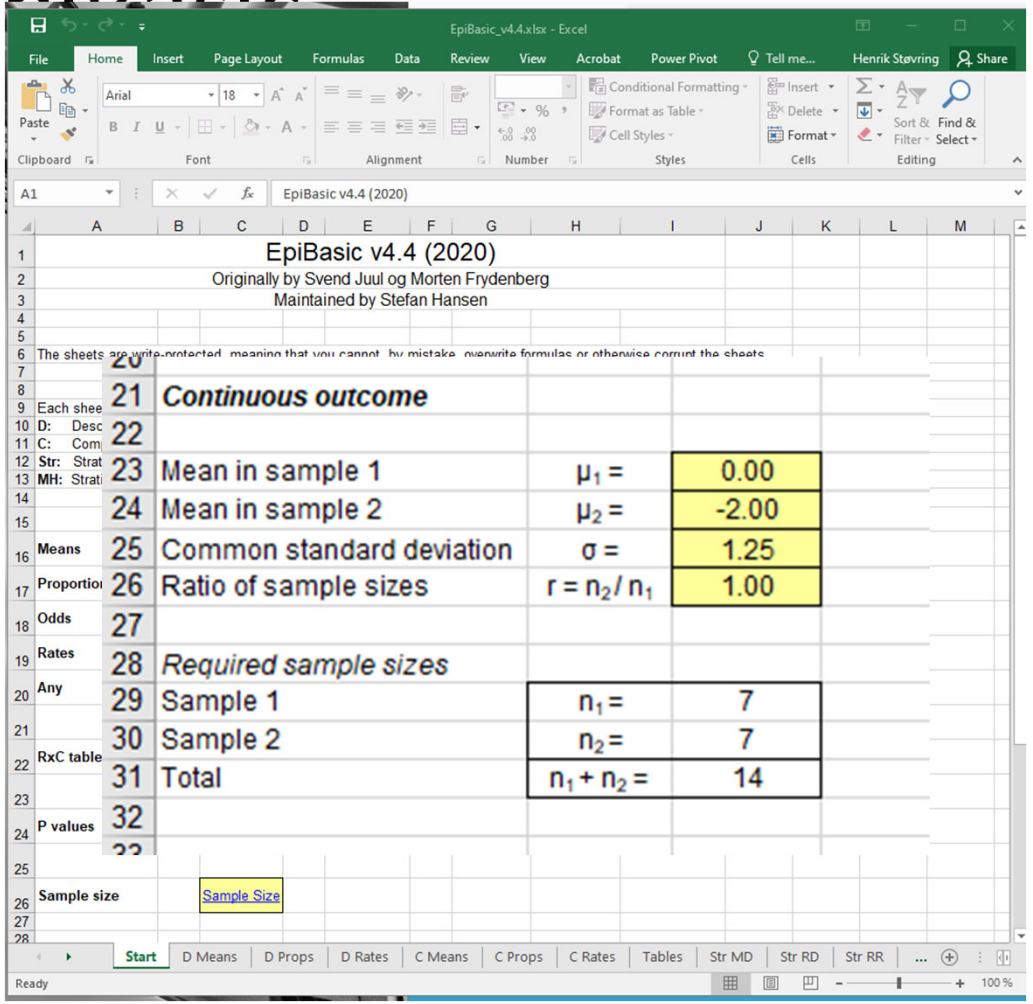
- Company NN has discovered an unexpected side-effect to one of their medications
- In trial of main effect of the pill, patients lost 2 kg in six months
- How large a study would be needed to conclude the pill has an effect different from a placebo effect?
- Study design:
  - Include patients at baseline, randomize to new pill or placebo
  - Measure weight at baseline and after six months of follow-up
- How many patients should be included?

Example cont'd: Assumptions

- Significance level 5%
- Effect size
  - No change on average for placebo
  - 2 kg reduction in intervention group
- Variation in weight change
  - Same variation in the two groups
  - In placebo group 95% of all will have individual change between ± 2.5 kg,
    i.e. SD(weight change) ≈ 1.25 kg
- KEY ASSUMPTION: Independent weight change between individuals
- We want to be 80% sure to detect this effect

Calculation using EpiBasic

- Download Epibasic from https://ph.medarbejdere.au.dk/undervisning/software
- Last sheet provides sample size calculations
- For our example:
  We need 14 patients
  (7 in each group)



Calculation using Stata

- Use command -power twomeans-
- For our example: We need 16 patients (8 in each group)
- Uses t-test instead of z-test (EpiBasic)
- Better to use t-test, but it should not matter much

```
. power twomeans -2 0, sd(1.25) power(.8)

Performing iteration ...

Estimated sample sizes for a two-sample means test
t test assuming sd1 = sd2 = sd
H0: m2 = m1  versus  Ha: m2 != m1

Study parameters:

        alpha =    0.0500
        power =    0.8000
        delta =    2.0000
           m1 =   -2.0000
           m2 =    0.0000
           sd =    1.2500

Estimated sample sizes:

            N =        16
  N per group =         8

.
```

Why do a sample size calculation?

- To satisfy funders' request

- Avoid futility
- Minimize harm to patients
- Maximize scientific value
- Avoid surprises when doing the study – necessitates thinking through the logistics of the study
- Analysis is predefined and so straightforward after the trial
- Very similar exercise to preparing a financial budget

The rationale of a sample size calculation

- **Objective:** $p < 5\%$
- How sure do we want to be? ← Power
- Power is
  Probability that study ends with a statistical significant finding ($p < 5\%$)
- Imagine you repeat a study 1,000 times – a power of 80% means that 800 studies will have a p-value below 5%
- Sample size is calculated from solving something like:

$$P\left(\frac{\text{Diff} - 0}{SE(\text{Diff})_{n}} > 1.96\right) = 80\%$$

- NB: $SE(\text{Diff})$ is a function of sample size:
  larger sample sizes gives smaller SE
- Can be expressed via Cohen's D: Diff/SD

The statistical testing approach

Follows the scientific *falsification approach:*

To show that pill A has a different effect than placebo, assume that

pill A is just as good as placebo

Then do the study, and (hopefully) ***reject this*** null-hypothesis ($p<5\%$)

We then conclude that pill A has an effect different from placebo

Logical consequence:

The objective is not to confirm pill A is better,
but to reject it has the same effect

Interpretation of $p>5\%$:

We cannot reject that pill A has no effect other than placebo (a triple negative?!)

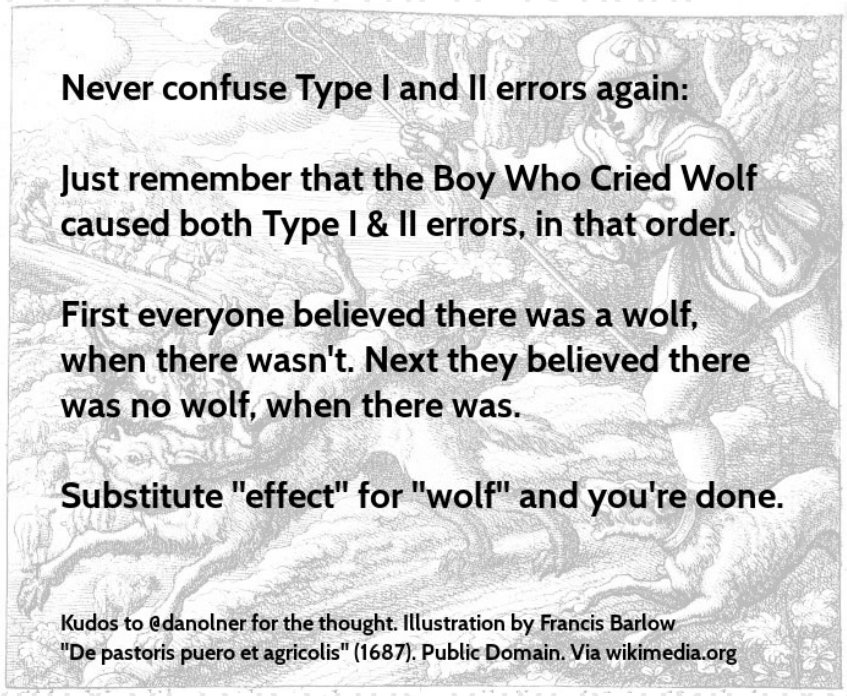
Type I or Type II error

Type I: **Reject null-hypothesis** even though it is true,
i.e. claim effect even though there is none

Probability of ty (almost always 5%)

Type II: **Not reje** it is false
i.e. not identify effect

Power is probab
= 1 – pro
= Probability of rejecting a false null hypothesis of no effect



What is needed for the most basic power calculation?

Continous and normal distributed outcomes (comparison of means)	Binary outcomes (comparison of proportions)
Significance level (5%)	Significance level (5%)
Intended power (80% or 90%)	Intended power (80% or 90%)
Expected difference in mean $\Delta = \mu_1 - \mu_0$	Expected risk difference, relative risk or odds ratio $\mathrm{RD} = \pi_1 - \pi_0$ $RR = \pi_1/\pi_0$ $OR = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_0}{1-\pi_0}}$
Standard Deviation of outcome in each group	(N/A)
Allocation ratio (for example 1:1)	Allocation ratio (for example 1:1)

How to guess-timate a standard deviation

The easy ones

- Knowledge
  - SD for systolic blood pressure (SBP) is 15 mmHg
  - (but what if outcome is change in SBP over 6 months?)

- Previous papers
  - Score of self perceived stress, 0-40, mean 14, SD 6.4
  - Look for their Table 1 or 2

How to guess-timate a standard

The easy on

- Knowl
  - SD f
  - (but
    mon
- Previo
  - Scor
  - Look for their Table 1 or 2

	Total *N* = 9748 *n* (%) or mean (SD)	RWD *N* = 953 (9.8) *n* (%) or mean (SD)	Non-RWD *N* = 8795 (90.2) *n* (%) or mean (SD)	p
Self-rated health				< 0.001[b]*
Low (%)	2606 (26.7)	338 (35.5)	2268 (25.8)	
High (%)	6386 (65.5)	494 (51.8)	5892 (67.0)	
Missing (%)	756 (7.8)	121 (12.7)	635 (7.2)	
Gender				0.055 [b]
Female (%)	4973 (51.0)	458 (48.1)	4515 (51.3)	
Male (%)	4775 (49.0)	495 (51.9)	4280 (48.7)	
Alder, mean (SD)	15.8 (0.4)	16.0 (0.5)	15.8 (0.4)	< 0.001[c]*
Self-assessed SES				< 0.001[b]*
Low (%)	233 (2.4)	49 (5.1)	184 (2.1)	
Medium (%)	5411 (55.5)	552 (57.9)	4859 (55.3)	
High (%)	3529 (36.2)	248 (26.0)	3281 (37.3)	
Missing (%)	575 (5.9)	104 (10.9)	471 (5.4)	
Negative childhood events				< 0.001[b]*
0 events (%)	940 (9.6)	84 (8.8)	856 (9.7)	
1–3 events (%)	5769 (59.2)	433 (45.4)	5336 (60.7)	
4–7 events (%)	1616 (16.6)	181 (19.0)	1435 (16.3)	
8–11 events (%)	234 (2.4)	56 (5.9)	178 (2.0)	
Missing (%)	1189 (12.2)	199 (20.9)	990 (11.3)	
Loneliness				< 0.001[b]*
Not lonely (%)	7358 (75.5)	609 (63.9)	6749 (76.7)	
Lonely (%)	2379 (24.4)	341 (35.8)	2038 (23.2)	
Missing (%)	[illegible]	[illegible]	[illegible]	
Perceived stress [a], mean (SD)	14.3 (6.4)	16.5 (6.1)	14.1 (6.4)	< 0.001[c]*

[a] = Scale from 0 to 40; higher = more stress, [b] = chi², [c] = t-test, *statistical significant $p < 0.05$

SD 6.4

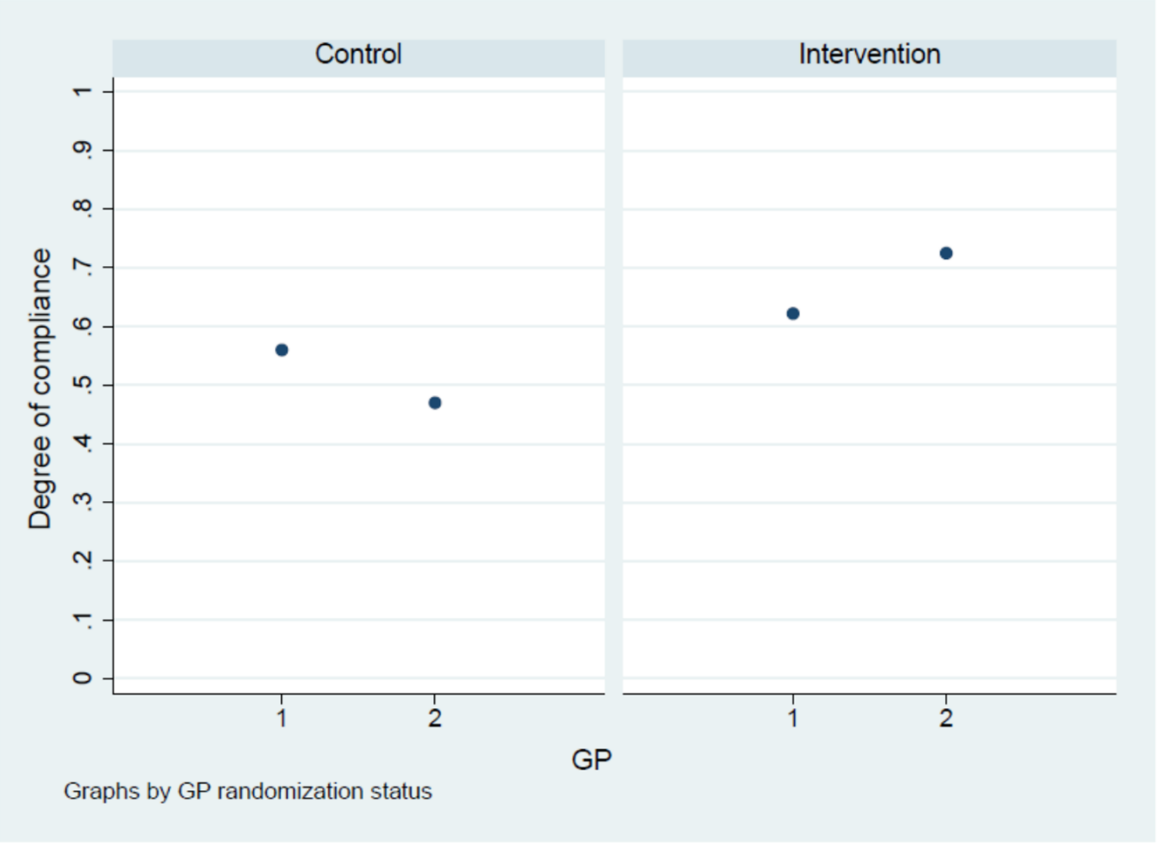
Some more advanced topics

- Trials with cluster effects
- When power is irrelevant
- Simulation approaches to power

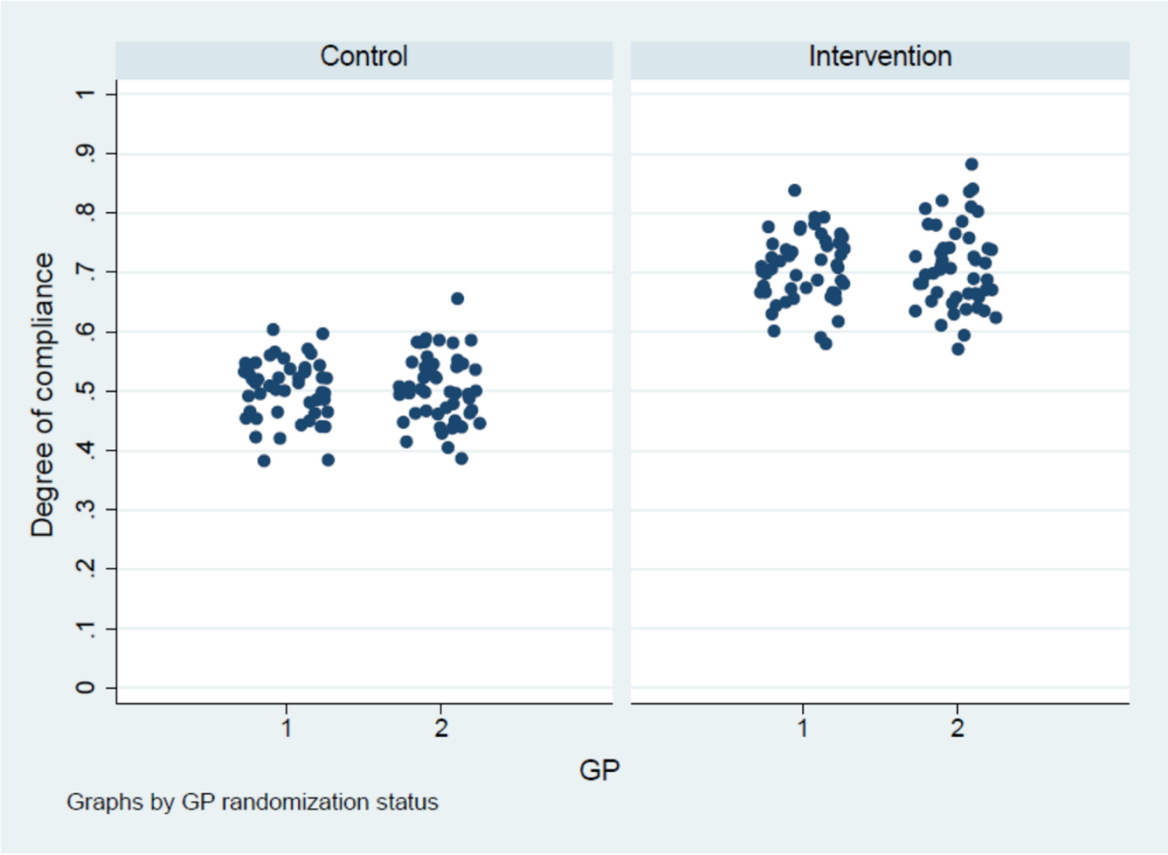
Example

- Imagine 4 GPs, each with 50 eligible patients
- Randomize GPs into two groups, two in each group
- Suppose outcome is compliance for a specific medication, rated from 0% to 100%
- Results:
  - Intervention group: 70% compliance ($SD = 5\%$)
  - Control group: 50% compliance ($SD = 5\%$)
- Naive p-value is (very nearly) zero

Example (continued): Clustered (extreme)



Example (continued): Independence



Trials with cluster effects

- The fundamental problem
  "A new observation at a unit (cluster) with previous observations gives less information than a new observation for a unit with no previous observations"

- Degree of within-cluster dependence is measured by

$$ICC = \frac{SD^2_{between}}{SD^2_{between} + SD^2_{within}}$$

- Study size should be scaled with *Design Effect* given by (m is average cluster size, e.g. patients per GP)

$$D_{\text{eff}} = 1 + (m - 1)ICC$$

What values to use for ICC?

- Does a “small” ICC of 0.01 matter?
- Ordinary sample size computation: 800 patients required
- In the study period an ordinary hospital can include 160 patients, i.e. average cluster size is 160
- Design effect is

$$D_{eff} = 1 + (160 - 1)0.01 = 2.59$$

- Instead of 800 patients, we should enroll 2,072 patients (sic!!) corresponding to 13 hospitals
- Similar papers on ICC exist for studies done in general practice (ICC from 0.01 to 0.05 are typical)

Table 4 Intraclass correlation coefficients for the outcomes variables of study

	Usual care		Care pathways	
Outcome variables	**ICC**	**95% CI**	**ICC**	**95% CI**
LOS (days)*	0.020	0.000-0.184	0.063	0.007-0.311
Cost (€)*	0.046	0.001-0.265	0.001	0.000-0.107
In-hospital mortality†	0.001	0.000-0.003	0.001	0.000-0.003
Disease Severity at Discharge (NYHA)†	0.182	0.062-0.554	0.000	0.000-0.076
AOS†	0.203	0.059-0.436	0.069	0.003-0.155
Unscheduled readmission†	0.004	0.000-0.036	0.010	0.000-0.046

DE: Design effect, AOS: Appropriateness of the stay, NYHA: New York Heart Association, CI: confidence interval.
*Ordinal or Continuous variable.
†Binary variables.

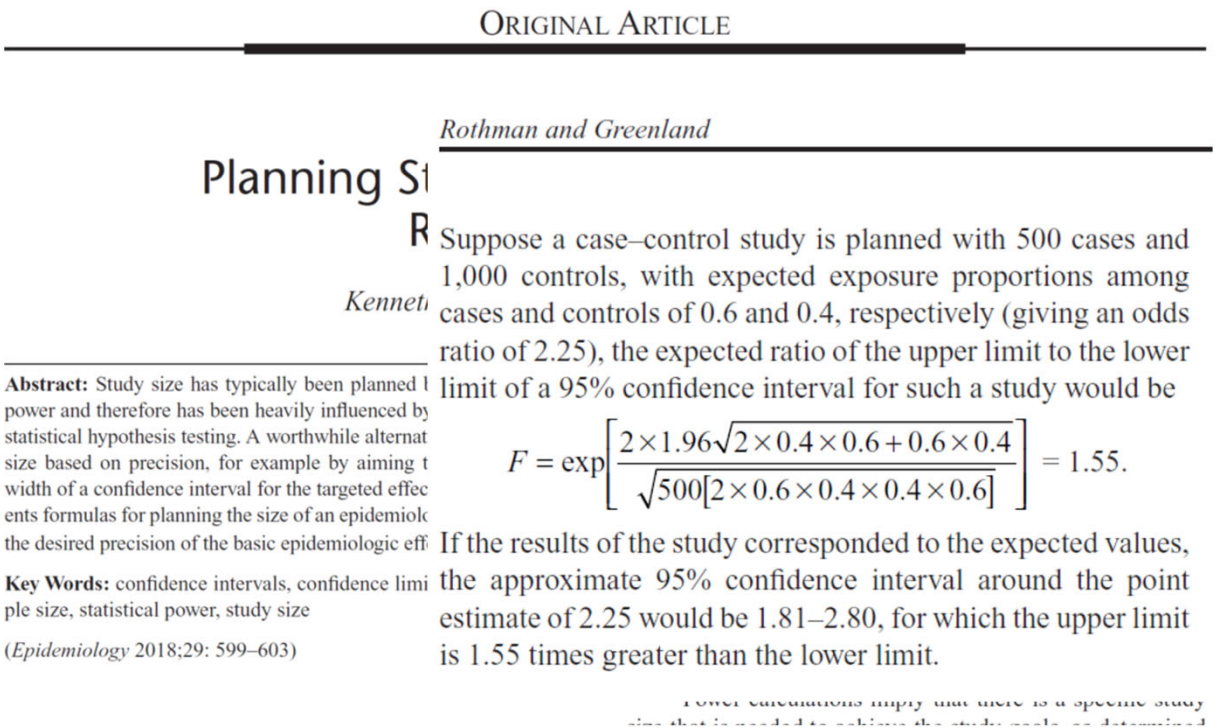
Background: Cluster randomized trials are increasingly being used in healthcare evaluation to show the effectiveness of a specific intervention. Care pathways (CPs) are becoming a popular tool to improve the quality of health-care services provided to heart failure patients. In order to perform a well-designed cluster randomized trial to demonstrate the effectiveness of Usual care (UC) and CP in heart failure treatment, the intraclass correlation coefficient (ICC) should be available before conducting a trial to estimate the required sample size. This study reports ICCs for both demographical and outcome variables from cluster randomized trials of heart failure patients in UC and care pathways.

Methods: To calculate the degree of within-cluster dependence, the ICC and associated 95% confidence interval were calculated by a method based on analysis of variance. All analyses were performed in R software version 2.15.1.

When power is irrelevant

- p-values are irrelevant for observational studies
- → power is irrelevant for observational studies
- Often sample size cannot be changed in observational studies
- Alternative to power: *statistical precision*
- Can be expressed as the expected
  Size of standard error
  Width of 95% confidence interval

Reference for precision calculation vs power



ORIGINAL ARTICLE

Planning St
R

Kennet

Abstract: Study size has typically been planned l
power and therefore has been heavily influenced by
statistical hypothesis testing. A worthwhile alternat
size based on precision, for example by aiming t
width of a confidence interval for the targeted effec
ents formulas for planning the size of an epidemiolc
the desired precision of the basic epidemiologic eff

Key Words: confidence intervals, confidence limi
ple size, statistical power, study size

(*Epidemiology* 2018;29: 599–603)

Suppose a case–control study is planned with 500 cases and 1,000 controls, with expected exposure proportions among cases and controls of 0.6 and 0.4, respectively (giving an odds ratio of 2.25), the expected ratio of the upper limit to the lower limit of a 95% confidence interval for such a study would be

$$F = \exp\left[\frac{2\times 1.96\sqrt{2\times 0.4\times 0.6+0.6\times 0.4}}{\sqrt{500[2\times 0.6\times 0.4\times 0.4\times 0.6]}}\right] = 1.55.$$

If the results of the study corresponded to the expected values, the approximate 95% confidence interval around the point estimate of 2.25 would be 1.81–2.80, for which the upper limit is 1.55 times greater than the lower limit.

Example of precision calculation in pharmacoepi

- A typical 2x2 table in studies on rare adverse effects of a drug

- The smallest number of the table is users with AE (exposed cases), i.e. a
- a is called the "bottleneck count"
- The bottleneck count determines precision

$$SE(\ln(OR)) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \approx \sqrt{\frac{1}{a}}$$

	Adverse effect	No adverse effect
Users	a	b
Non-users	c	d

Bottleneck count and precision

Received: 30 June 2020 | Accepted: 13 January 2021
DOI: 10.1002/pds.5200

ORIGINAL ARTICLE

WILEY

Bottleneck analysis: Simple prediction of the precision of a planned case–control or cohort study based on healthcare registers

Jesper Hallas[1,2] | Morten Rix Hansen[1,

[1]Clinical Pharmacology, Pharmacy and Environmental Medicine, University of Southern Denmark, Odense, Denmark

[2]Department of Clinical Biochemistry and Pharmacology, Odense University Hospital, Odense, Denmark

[3]Department of Public Health—Biostatistics, Aarhus University, Aarhus, Denmark

Correspondence
Jesper Hallas, Clinical Pharmacology and Pharmacy, University of Southern Denmark, JB Winsløwsvej 19, 2, 5000 Odense C, Denmark.
Email: jhallas@health.sdu.dk

Funding information
University of Southern Denmark

Abstract

Purpose: In
usually deper
denote this tl
into an achiev
Methods: Firs
the BNC deter
a null-effect st
confidence lim
ture, analyzing
Results: The l
log(BNC). The
bleeding was
ture review showed an inverse linear relationship between the logarithmic BNC and
the log–log transformed ULCLR, which was largely independent of study design,

TABLE 3 Relationship between bottleneck count and predicted precision in a null result

	Theoretical optimum, prediction based on Equation (3)		Prediction based on simple empirical model in Table 2		
Bottleneck count	95% CI (null estimate)	ULCLR	95% CI (null estimate)	ULCLR	Relative increase in empirical ULCLR over theoretical optimum
5	(0.42; 2.40)	5.77	(0.37; 2.70)	7.27	1.26
10	(0.54; 1.86)	3.45	(0.49; 2.06)	4.25	1.23
20	(0.65; 1.55)	2.40	(0.59; 1.69)	2.87	1.20
50	(0.76; 1.32)	1.74	(0.71; 1.42)	2.00	1.15
100	(0.82; 1.22)	1.48	(0.78; 1.29)	1.66	1.12
200	(0.87; 1.15)	1.32	(0.83; 1.20)	1.45	1.10
500	(0.92; 1.09)	1.19	(0.89; 1.13)	1.28	1.07
1000	(0.94; 1.06)	1.13	(0.92; 1.09)	1.19	1.06

Note: ULCLR: ratio between upper and lower confidence limit

126 effect estimates from 57 publications in *Pharmacoepidemiology and Drug Safety*, 2015–2018

Power in non-simple settings

- Imagine the following study:

- An intervention will be given to diabetes patients in groups in an unknown number of municipalities
- Each group will have 10-12 patients
- Patients randomized to be controls (no intervention, usual care) are *not* in groups

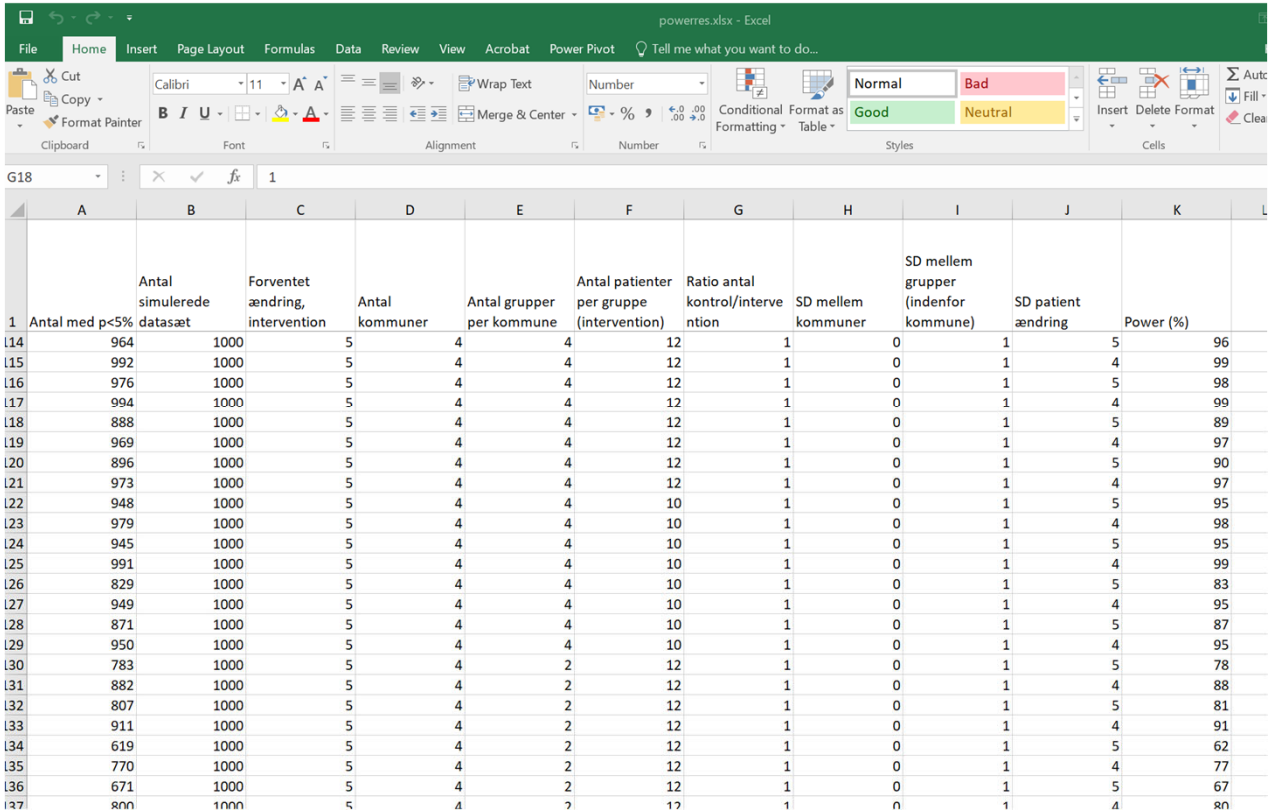
- Statistical concerns
  - Clustering due to groups in intervention, but NOT for controls
  - Clustering due to municipality for intervention AND for controls

Simulation of power

- Imagine that for change in SF-12, PC we know :
  - SD for an individual
  - SD for variation in mean change for groups
  - SD for variation in mean change for municipalities
  - allocation ratio
  - number of patients in each group
  - number of groups per municipality
  - number of municipalities
  - Expected average change in SF-12, PC, due to intervention
  - Assume change in SF-12, PC, follows a normal distribution

- Then we can generate datasets based on these assumptions
- Repeat 1,000 times for a setting and analyse each dataset
- Count number of analyses with $p < 5\%$

Results of simulation study



1	A Antal med p<5%	B Antal simulerede datasæt	C Forventet ændring, intervention	D Antal kommuner	E Antal grupper per kommune	F Antal patienter per gruppe (intervention)	G Ratio antal kontrol/intervention	H SD mellem kommuner	I SD mellem grupper (indenfor kommune)	J SD patient ændring	K Power (%)
14	964	1000	5	4	4	12	1	0	1	5	96
15	992	1000	5	4	4	12	1	0	1	4	99
16	976	1000	5	4	4	12	1	0	1	5	98
17	994	1000	5	4	4	12	1	0	1	4	99
18	888	1000	5	4	4	12	1	0	1	5	89
19	969	1000	5	4	4	12	1	0	1	4	97
20	896	1000	5	4	4	12	1	0	1	5	90
21	973	1000	5	4	4	12	1	0	1	4	97
22	948	1000	5	4	4	10	1	0	1	5	95
23	979	1000	5	4	4	10	1	0	1	4	98
24	945	1000	5	4	4	10	1	0	1	5	95
25	991	1000	5	4	4	10	1	0	1	4	99
26	829	1000	5	4	4	10	1	0	1	5	83
27	949	1000	5	4	4	10	1	0	1	4	95
28	871	1000	5	4	4	10	1	0	1	5	87
29	950	1000	5	4	4	10	1	0	1	4	95
30	783	1000	5	4	2	12	1	0	1	5	78
31	882	1000	5	4	2	12	1	0	1	4	88
32	807	1000	5	4	2	12	1	0	1	5	81
33	911	1000	5	4	2	12	1	0	1	4	91
34	619	1000	5	4	2	12	1	0	1	5	62
35	770	1000	5	4	2	12	1	0	1	4	77
36	671	1000	5	4	2	12	1	0	1	5	67
37	800	1000	5	4	2	12	1	0	1	4	80

Summary

- Power calculations are relevant for planning randomized trials
- Required ingredients are
  - study design
  - planned analysis model
  - effect size
  - for continuous outcomes: variation in outcome, SD
- Relevant input parameters typically require
  - clinical insight
  - knowledge of literature in the field
  - + some formula trickery from statistical theory

Summary (cont'd)

- Clustering effects should not be ignored
- Power calculations should not be done for observational studies
- Power calculations should not be done after the study is finished
- Expected statistical precision is relevant when planning an observational study
- Sometimes formulas are inadequate, but then a simulation approach can be considered

Summary (cont'd)

Not covered

- Time to event studies
- Non-inferiority or equivalence trials
  - Reject a null-hypothesis of an "important superiority" or "important difference"
- Early stopping or adaptive stopping strategies
- ...

A statistician's shopping list for (simple) power calculations

- Description of study design
  - what defines intervention group? control group?
  - allocation ratio?
- Outcome measure
  - continuous outcome or binary?
  - log-scale or not?
- Choice of analysis model for comparison
  - linear regression?
  - mixed model for repeated measures?
  - logistic regression?

A statistician's shopping list for (simple) power calculations

- Parameter(s) to be compared
  - mean difference?
  - odds ratio? hazard ratio? etc
- Assumed magnitude of parameter in control and intervention group if intervention works as intended
- For continuous outcomes: SD of "one observation" (could be SD of change from baseline in outcome)
- Desired power (80% or 90%) and intended significance level (5%)
- Are clustering effects likely?

Thank you for your attention – questions?

Some (perhaps?) useful formulas - general

- Finding SD from an SE of the mean

$$SE(mean) = \frac{SD}{\sqrt{n}} \quad \text{and so} \quad SD = \sqrt{n} \cdot SE(mean)$$

- Finding SE from a 95% confidence interval

$$95\%CI(\text{some true value}) \approx \left(\text{estimate} \pm 1.96 \times SE(\text{estimate})\right)$$

and so

$$SE(\text{estimate}) = \frac{\text{upper limit} - \text{lower limit}}{2 \times 1.96}$$

- If SD is equal in two independent groups

$$SE(\text{diff}) = \sqrt{SE_1^2 + SE_2^2} = \sqrt{\frac{SD^2}{n_1} + \frac{SD^2}{n_2}} = SD \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$SD = \frac{SE(\text{diff})}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Some (perhaps?) useful formulas - proportions

- Standard error of a proportion π

$$SE(\pi) = \frac{\pi(1-\pi)}{\sqrt{n}}$$

	outcome +	outcome -
Group 1	a	b
Group 0	c	d

- Standard error of $RD = (\pi_1 - \pi_0)$

$$SE(RD) = \sqrt{SE(\pi_1)^2 + SE(\pi_0)^2}$$

- Standard error for $\ln(RR)$

$$SE(\ln(RR)) = \sqrt{\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}}$$

- Standard error for $\ln(OR)$

$$SE(\ln(OR)) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

Thanks for your attention – questions welcome!

(Djursland, July 2015 – H Støvring)