Refer to the article "**Randomized Controlled Trial of Routine Cervical Examinations in Pregnancy**" by Buekens P. et al in The Lancet Vol 344 pages 841-814 September 24, 1994.

1 (Statistical Methods, paragraph 1)

(i) Verify the sample size requirement of 7000 per group, and the calculation for the reduced sample trial of 3000 per group.

(ii) Given the huge budget and 'not easy to replicate' nature of this trial, one could argue for demanding higher power. Redo the calculations using a beta of 10%.

(iii) For an alpha of 0.05 and a given delta, what is the percentage increase in sample size as one goes from beta=0.2 to beta=0.1?

2 (Statistical Methods, paragraph 2)

(i) Illustrate, using as an example the first outcome in Table 3, that statistical comparisons of the frequencies of binary outcomes yield the same p-value whether carried out with the chi-square test for a 2x2 table [your choice of computation method] or the z-test for 2 proportions. Use enough decimal places to be sure they really are the same.

(ii) Repeat both calculations, but this time using the continuity correction

$$x^2 = \sum \frac{\{|o–e|–0.5\}^2}{e} \quad \text{or}$$

$$\frac{N\ \{\ |ad - bc\ |\ –N/2\ \}^2}{r_1\ r_2\ c_1\ c_2} \quad \text{or}$$

$$\frac{\{\ |\ a - E[a]\ |\ - 0.5\}^2}{r_1\ r_2\ c_1\ c_2\ /\ N^3} \quad \text{or}$$

$$z = \ = \frac{p_1{*} - p_2{*}}{\sqrt{p\{1–p\}\{1/n_1\ +\ 1/n_2\}}} \quad \text{or}$$

where, if p_1 is the larger and p_2 the smaller of the two proportions,

$$p_1{*} = \frac{y_1\ - 0.5}{n_1} \ \text{ and } \ p_2{*} = \frac{y_2\ + 0.5}{n_2}$$

(Colton explains this correction in the z-test for proportions on page 165 but then leaves a typographical error in the z formula)

Note a propos the CI's for RRs: The Taylor series CI for the ratio of proportions is described in §15.2.1 of Kleinbaum, Kupper and Morgenstern's text "Epidemiologic Research".

3 (Results: Characteristics of study population)

"The trial groups were similar in age, education level and baseline obstetric history (Table 1)"

(i) One would expect with such large sample sizes that the balance would be excellent; but just how close should the means be? For example, should the n's of ≈2750 and ≈2750 "guarantee*" that the average age in the two groups would not differ by more than 0.1 year or 0.5 years or 1.5 years? Assume some justifiable standard deviation of individual ages and calculate the possibilities for the difference between the average age of one random half of the subjects and that of the other half. Sketch a frequency distribution to illustrate the results of your calculations.

Hint: the question concerns the sampling variation of $\bar{y}_1 - \bar{y}_2$ calculated assuming randomization. [* nothing can be "guaranteed" but use as an operational definition "95% sure"]

(ii) Do the same for the difference in the frequency of primapara.

Note: While these types of calculations are the basis for significance tests for <u>outcomes</u>, they should not be used to carry out formal tests of hypotheses on <u>baseline</u> data from RCT's (Table 1 in most clinical trials). There is no great reason to calculate p-values for baseline differences unless one wishes to check if they carried out the ranoomization correctly. A much more important question than whether any imbalances are statistically significant is the magnitude of the differences and how much distortion these imbalances actually make to the comparison i.e. . it's a question of "embarrassing" rather than "statistically significant" differences.

4 (Results: second paragraph)

The authors report a total of 20+0+23+1 = 44 multiple births among 2719+2721 = 5440 women followed up until delivery.

(i) Calculate a 95%CI for the frequency of multiple births per 1000 women. What confidence do you have in this interval estimate as an estimate of the frequency of multiple births in general in these countries?

(ii) The authors calculated preterm rates per 100 total births. They used the chi-square [= z^2] tests to compare them. Are all of the statistical requirements for such tests met? If not, will the reported p-value be too big or too small?

5 **(Results: Cervical examinations and interventions)**

(i) What is the purpose of reporting and comparing the number of cervical examinations in the two groups? Is it advisable to perform a formal test of significance for this comparison?

(ii) Why are medians rather than means used in the third paragraph?

(iii) From the point of view of budget people, why is the mean more relevant than the median?

(iv) Are t-tests on mean numbers of visits or mean lengths of bed rest stay or hospital stay contra-indicated by your answer to (ii)?

(v) The authors used the Mann-Whitney test (= Wilcoxon Rank Sum Test) for comparisons of "distributions" (third paragraph). If they had consulted you about between parametric vs. non-parametric tests for these, what would you have advised and why?

(vi) How does one obtain a p-value for the Rank Sum test when the n's are so high?

6 **(Outcomes)**

Throughout, the authors are more concerned with <u>ratios</u> of proportions (what they call risk ratios or RR's) than with <u>differences</u> in proportions (what we might call risk differences or RD's). If we wanted to test that RR=0, using alpha = 0.05 two sided; for the outcomes in Table 3, how could we infer the results of such tests without actually carrying them out?

7 (i) Make a rough plot the 7 CI's for the RR in the top half of Table 4 in the graphical style used in meta-analyses.

Why are the CI's not symmetrical about the point estimate?

(ii) Sketch what a graphical display of the results would look like if presented on a RD rather than an RR scale.

8 (i) Do you agree with the need for the Bonferroni correction (i.e. using alpha = 0.05/14 rather than 0.05) when interpreting the RR's for the different countries? see M& M p 742- for methods for 'correcting for' multiple comparisons; they concentrate on comparisons of several treatments; but the issue can also be raised concerning comparisons of the same two treatments in several subgroups of the same dataset;. The Bonferroni correction (see page 844) involves dividing the overall alpha (in this case 0.05) by the number of tests carried out (in this case 14), and using this stricter alpha for each separate test.

(ii) If one performs 14 independent tests of a null hypothesis using an alpha of 0.05 for each one, and the null hypothesis is indeed true, what is the probability of at least one false rejection among the 14? (as is often the case with calculations involving "at least one", it is easier to calculate it as 1 minus the probability of all 14 test being negative)

(iii) If -- again when Ho is true -- one uses an alpha of 0.05/14 for each test, what is the probability of at least one false positive test? (It will not come out exactly to 0.05 but it will be close)

(iv) Can you make a case why -- even if you agree with the principle of correcting for multiple independent tests -- dividing by 14 in the Bonferroni correction may be overly stringent in this example and why a smaller divisor -- somewhere between 7 and 14 --might be more reasonable? (Think of the two tests done for each country)

(v) Do you believe that the results in Spain and Portugal could be chance variations and the significant results are a consequence of overtesting (overfishing?) or do you believe they are real?

(vi) How much of your interpretation comes from the fact that these occurred in Spain and Portugal? Would your interpretation have been different if the countries in Table 4 were blinded and referred to only as country 'A' to country 'F'?

(vii) What does you answer say about relying solely on the p-value in judging whether a difference is 'real' or not?

9 (i) What test is appropriate for testing the hypothesis that the frequency of preterm deliveries differs among the countries? Do not carry out the test, but point to a similar example in Moore and McCabe.

(ii) What are the alternative and alternative hypotheses tested?

(ii) If the results were significant, what conclusions could one safely draw?

10 Are the results 'definitely negative' or simply 'inconclusive'? Advocates of repeated cervical examinations will surely point to the fact that the study did not have sufficient power to detect a risk reduction of 20% in the preterm rate i.e. an RR of 0.8.

(i) Calculate the power of a study with 2750 and 2750, with alpha =0.05 two sided as before, to detect such a reduction? Hint: solve for Z_{beta} in the formula linking n, alpha beta and delta.

(ii) Such power calculations are relevant only for planning purposes and are of little relevance after the data are in. Instead, one should use the calculated CI: in this case the 95% CI for the RR was 0.85 to 1.29.

(iii) Suppose you are involved in a panel discussion about the study. One panelist focuses on the inadequate power but does a poor job of communicating in non-technical language what statistical power is. Provide a non-technical but correct 'translation'.

(iv) Another panelist uses the CI from the study rather than the pre-study calculations of power, but again does a poor job of conveying the meaning of a CI. Give an understandable and at the same time technically correct translation of the CI.

(v) A third panelist says "we should all be much more Bayesian about all of this". Again, translate "Bayesian".

11 The trialists considered a risk reduction of 20%, or from 5% to 4%, as clinically significant.

(i) What factors go into deciding what is clinically significant?

To many, what is relevant is the RD rather than the RR, since one can directly calculate the number required to treat in order to prevent one bad outcome as 1/RD. Proponents such as Sackett call this the "Number Required to Treat" (NRT). So if the RD was 1.1% (7.7% minus 6.6%), one would need to intervene on NRT=100/1.1 = 91 pregnancies to prevent one low birthweight birth.

(ii) However the estimate of RD, and thus of NRT, has some sampling variability. Use the 95% CI for RD for low birthweight to find a 95% CI for NRT.

Answers RCT of Routine Cervical Examinations in Pregnancy

1i π_E =0.04, π_C= 0.05, Δ =0.01. Several options for n/group, ranging from the most exact (Colton p 168)

$$\frac{\{z_{\alpha/2}\ SE[p_E - p_C \mid H_0] + z_\beta\ SE[p_E - p_C \mid H_{alt}]\}^2}{\Delta^2} \qquad [1]$$

$$= \frac{\{z_{\alpha/2}\ \sqrt{\pi_C[1-\pi_C] + \pi_C[1-\pi_C]} + z_\beta\ \sqrt{\pi_E[1-\pi_E] + \pi_C[1-\pi_C]}\ \}^2}{\Delta^2}$$

to the rougher

$$\frac{\{z_{\alpha/2}\ \sqrt{\bar\pi[1-\bar\pi] + \bar\pi[1-\bar\pi]} + z_\beta\ \sqrt{\bar\pi[1-\bar\pi] + \bar\pi[1-\bar\pi]}\ \}^2}{\Delta^2}$$

$= 2\{z_{\alpha/2} + z_\beta\}^2 \dfrac{\bar\pi[1-\bar\pi]}{\Delta^2}$ [2]. Here $\bar\pi = 0.045$.

If α= **0.05** & ß= **0.2**, $z_{\alpha/2}$ = 1.96 & z_β =0.84, so $2\{z_{\alpha/2} + z_\beta\}^2$ = 15.68, and

if we round 15.68 to 16, we get the rougher $16 \dfrac{\bar\pi[1-\bar\pi]}{\Delta^2}$

where $\bar\pi[1-\bar\pi]$ is the variance of individual observations on a 0/1 scale (see article by _______ reproduced in Material on Chapter 8).

If π_E =**0.035**, π_C= 0.05, Δ =0.015, then $\bar\pi$ = 0.0425.

ii If α= **0.05** & ß= **0.1**, $z_{\alpha/2}$ = 1.96 & z_β =1.28, so $2\{z_{\alpha/2} + z_\beta\}^2$ = 20.9952, we can round it to 21, for a fairly exact $21 \dfrac{\bar\pi[1-\bar\pi]}{\Delta^2}$

iii If only change ß, and use [2] above, then change $\{z_{\alpha/2} + z_\beta\}^2$ to 20.9952 from 15.68 or an increase of 34%.

2i Data p_E = 169/2521 and p_C =161/2520, p = 330/5041. Or if you are an $\begin{matrix} a & b \\ c & d \end{matrix}$ person who uses columns for groups being contrasted and rows for outcomes, then

```
                  G R O U P
                  E         C        All
OUTCOME
Bad      a =  169  b=  161 | r1=  330
Good     c = 2352  d= 2359 | r2= 4711
All      c1= 2521 c2= 2520 |  N= 5041
```

Several options for uncorrected X^2,

definitional formula $\sum \dfrac{\{o-e\}^2}{e}$ or $\dfrac{N\{ad-bc\}^2}{r_1\ r_2\ c_1\ c_2}$ or $\dfrac{\{a-E[a \mid H_0]\}^2}{var[a \mid H_0]}$

with var[a | H_0)=$r_1\ r_2\ c_1\ c_2 / N^3$

Using the 2nd form gives X^2 = **0.20414221..** or $\sqrt{X^2}$ = **0.4518..**

For z test for 2 proportions,

$$Z = \frac{p_E - p_C}{\sqrt{p\{1-p\}\ \{\dfrac{1}{2521} + \dfrac{1}{2520}\}}} = \mathbf{0.4518..}$$

For z **test** for 2 π's, must use **common p** [H_0 stipulates $\pi_E = \pi_C$]; for a **CI** for Δ of 2 π's, the SE of the difference uses separate p's, since there is no H_0 to say anything about what the two underlying π's should be. X^2 and Z^2 will not match exactly if use separate p's

ii With continuity correction, X^2 = **0.15592817.., Z = 0.3948**.. = $\sqrt{X^2}$ [the numerator of Z changes from 0.003148.. to 0.002751.. but nothing else is affected]

3i At issue is the sampling variability of the **statistic** $\bar y_E - \bar y_C$. With the large n's involved this should be Gaussian around a mean of zero, no matter what the SD_{ind} of **individual** y's (CLT). The SD or SE of the Gaussian distribution of the statistic will be the square root of the sum of the squares of the SEM's of the two component statistics ie. SE = $\sqrt{\dfrac{SD_{ind}^2}{2521} + \dfrac{SD_{ind}^2}{2520}}$

= $SD_{ind} \sqrt{\dfrac{1}{2521} + \dfrac{1}{2520}}$ = $0.028SD_{ind}$. Thus there is a 95% probability that possible $\bar y_E - \bar y_C$'s will be within the range –1.95SE to +1.96SE or between **$-0.055SD_{ind}$ and $+0.055SD_{nd}$.**

The variation of individual ages is probably not Gaussian and so SD_{ind}. cannot be used with the Z table to describe the limits of individual variation; however it can be used for the limits of the statistic $\bar{y}_E - \bar{y}_C$. If we thought that maternal ages varied from 15 to 45, the SD_{ind} cannot be more than half the range i.e. (45–15)/2 =15. Since the ages have a strong central tendency, SD_{ind} is probably closer to **5**. Substituting this in the limits above gives **±0.275 years**. In other words there is a good chance that the difference between the average age of 2 randomly chosen samples of maternal ages will not be more than 0.3 years.

ii For the difference in the proportion of primipara's, the only difference is that we are dealing with a mean of a 0/1 variable. Since 45% of the y's are 1 and 55% are 0's, the $\mathbf{SD_{ind}}$ or these 0's and 1's is $\sqrt{0.45x0.55} \approx \mathbf{0.5}$. So the difference in the two proportions will again have a Gaussian distribution with mean zero and 95% limits of ±0.055(0.5) i.e. **±0.0275 or ±2.75%.**

A large number of students spoke about CI's. **CI's are for parameters**. Here we are dealing with **distribution of a statistic**. i.e.

$$\bar{y}_E - \bar{y}_C \sim N(\mathbf{0},\ SD_{ind}\sqrt{\frac{1}{2521} + \frac{1}{2520}}\ )$$

4i p=44/5400 = 0.00809 or 8.09 per 1000, based on a sample of n=5440. Large sample 95%CI for π, the frequency of multiple births, based on Binomial variation, given by $p \pm 1.96\sqrt{p[1-p]/5440}$ = 0.00809±0.00238 or 5.71 to 10.47 per 1000. [Some of you got the scales mixed up: better to stay in (0,1) proportion scale until end and then convert to rate per 1000. Also, some of you based CI on n = 1000 rather than 5440].
CI constructed from margin of error, and a strict funcction of n and p; any biases in sampling are not included in the calculation. Some high risk pregnancies were excluded, and this could well mean that multiple births are therefore under-represented (other exclusions too). Also, should be suspicious of 'national' samples based on two to 6 clinics or hospitals per country.

ii Chi-square and z tests of 2 proportions are (equivalent) large sample tests that try to approximate binomial sampling variation by Gaussian variation. For some of the country-specific comparisons, the <u>expected</u> numbers of bad events are below 5, a threshold below which users are often warned about the accuracy of the approximations. Cannot always be sure of direction of inaccuracies: I usually calculate by large sample and 'exact' methods and find close agreement even when E's are well < 5.
A more fundamental issue in this example is whether it is reasonable to count the outcomes in 2 twins as independent observations [the binomial, just like every SE involving √n, is built on independent observations]. The outcomes are likely to be positively correlated and so using n= no. of infants gives a falsely small SE, and a more extreme p-value than we should have. However, the multiple births counted in the n is not large and the 'false inflation' of n is inconsequential.

5i When intervention is left to caregivers, authors must assure us that it was carried out well. If it was only halfheartedly done, that might explain why no effect was seen. It is a little like in a drug trial checking that the patients filled their prescriptions and actually took their medications. It does not make any sense to compare this process variable in the two groups by a formal statistical test. The null hypothesis is not of interest; the issue is whether the intervention was implemented <u>intensively enough</u> that we might expect a result; a small difference in the number of examinations could be statistically significant if the n's were large, but simply telling the reader that the difference was "statistically significant" would be of no use; the reader wants to be sure that the intervention was given in a big enough 'dose' to work.

ii Because the distributions of the variables were skewed, and a median is more descriptive of central tendency, and is robust ('resistant') to extreme observations than a mean.

iii Total numbers of bed days or lengths of stay can be reconstructed/estimated from the mean but not from the median.

iv Even with very skewed distributions of individual observations, with the large sample sizes involved, the sampling variability of the sample mean (a statistic) is quite Gaussian (CLT).

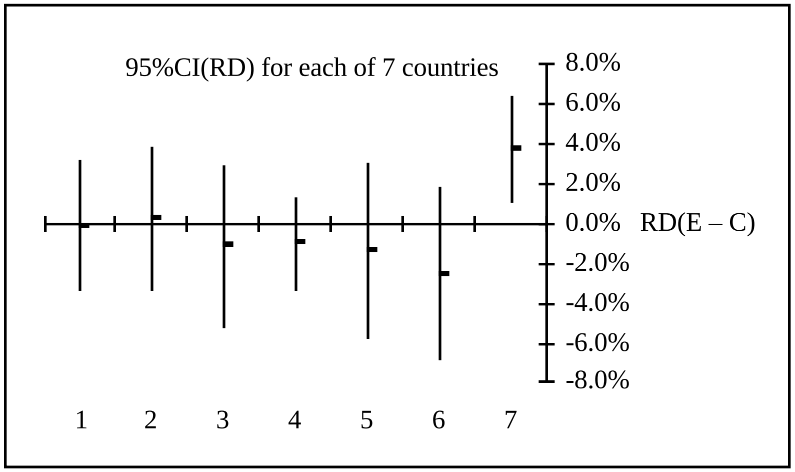
v Both are reasonable in this large-sample example. However, if the natural parameter is the mean, why not test means by parametric tests.

vi Use Gaussian approximation to distribution of Sum of Ranks in smaller sample. The null distribution of this statistic is very Gaussian even for sample sizes as low as 10. The SD of the Sum of Ranks, needed for using the Gaussian table, is simply a function of the two n's involved (thats why these tests are called 'distribution-free') and the formula is given in the material from A&B.

6 A zero value of the RD parameter corresponds to a unity value of the RR parameter (contrary to what many will tell you, 'null' does not necessarily mean zero). Thus if we are consistent in our statistical approach, an RR interval estimate (CI) that does not include 1 should correspond to an RD interval estimate (CI) that does not include 0, and both of these imply a point estimate that if tested against the null would not be statistically significant. Sadly, a consistent approach to RD's and RR's is absent from most textbooks.

7 i CI's for RR are calculated by first constructing symmetric CI's on log scale and then converting them to asymmetric RR scale. (ratios of statistics tend to be skewed)

ii RD... using symmetric CI's based on $ME = 1.96 \sqrt{p_1[1-p_1]/n_1 + p_2[1-p_2]/n_2}$ from difference of two independent sample proportions, each subject to binomial variation.
[Graph is from Excel spreadsheet of the type used to display 4 items 'volume-high-low- closing' for stock market activity. using volumes of zero and deleting the corresponding left vertical axis]



8i If Ho true and if carry out many tests, increase chance of a false positive result. Stricter alpha level for individual tests reduces risk of this.

ii Prob[at least 1 FP] = 1–Prob[all 14 negative] = $1–0.95^{14}$ = 1–0.49 = 0.51.

iii Prob[at least 1 FP] = $1–(1–0.05/14)^{14}$ = 1–0.951 = 0.049.

iv The two outcomes Low Birth Weight and Preterm Delivery are biologically linked and so are not independent.

v If results were in same direction in neighbouring countries with same setup, might be inclined to think that "chance does not strike the same region twice". But this isn't the case [RR 0.67 in Portugal but 1.84 in Spain]. So no consistency, maybe 1.84 is chance.

vi I did use geographic contiguity to try to interpret them; had the pattern in the two been similar I might have been more willing to think it was real. So I am using 'outside' information to read something into the p-values: ie even if p-values are same blinded and unblinded, I do not interpret them in exactly same way. [Then again, I know very little about the medical setup and the possibility of such outcomes in these countries-- it may well be that the two neighbouring countries are as different as Ireland and Hungary in so far as what one would expect of the interventions. Other factors, such as integrity of the randomization, compliance, etc etc come into the interpretations too]

vii The same p-values, should, if combined with unequal prior beliefs, lead to unequal posterior beliefs. We must not interpret p-values in a vacuum.

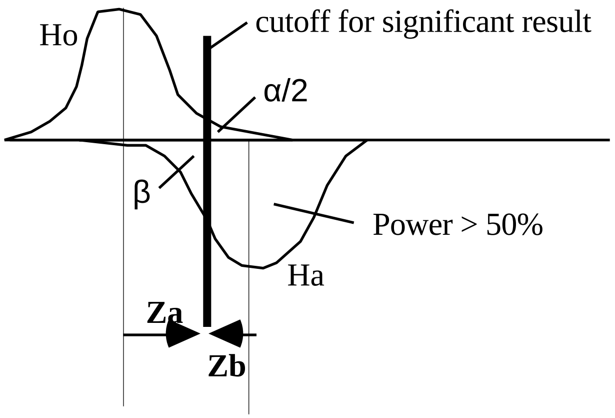
9i Data in form of 7 Binomial proportions, or counts in a 2x7 table. So can use a chi-square test of equality of proportions (or lack of associoation between rows and columns in table).

ii π=proportion of Low Birth Weight.
H_0: $\pi_{Belgium} = \pi_{Denmark} = ... = \pi_{Spain}$
Ha: Some variation among the 7 π's.

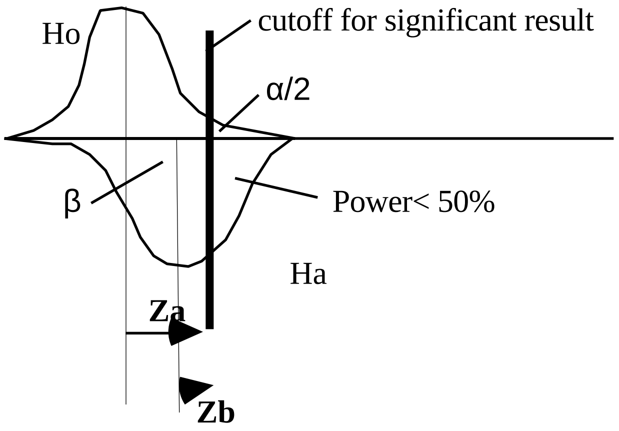
iii That have evidence against H_0. Not clear where source of variation is.

10i $2750 = 2\{z_{\alpha/2} + z_\beta\}^2 \frac{\bar{\pi}[1-\bar{\pi}]}{\Delta^2}$. Here $\bar{\pi}$ = 0.045 and Δ = 0.01.
Thus $\{z_{\alpha/2} + z_\beta\}$ = 1.79.
The correct formula involves $\{z_{\alpha/2} – [z_\beta]\}$ When we have a power of > 50%, z_β will be a <u>negative</u> quantity so the difference comes to more than 1.96. Many, as I do, will write the formula as $\{z_{\alpha/2} + z_\beta\}$ with the understanding that $\{z_{\alpha/2} + z_\beta\}$ will be a <u>positive</u> quantity. Here the amalgam of the two z's comes to less than $z_{\alpha/2}$ alone, so in fact we have less than 50% power [Less than half of the Ha distribution is beyond the cutoff for statistical significance]. Strictly speaking the 1.79 is {1.96 – [+0.17]}. It is best to draw a diagram to illustrate.

Usual Situation (Power > 50%)



Low Power Situation (Power < 50%)



The power is the amount of the Ha distribution that is beyond the cutoff for significance. Here it is the proportion of the Z distribution that is to the right of Z=+0.17. This is approximately 43%.

ii The comment was to illustrate that if use CI of .89 to 1.29, we are 'ruling out' RR's below 0.85; and before the study the authors agreed that the RR had to be below 0.8 to be clinically important. So the study provides a 'definitive negative' result rather than an inconclusive one.

iii The probability that if a given difference exists, the study will produce a positive [ie 'statistically significant'] result.

iv See the text. A CI is an interval produced using a procedure that yields a 'correct' answer (within the margin of error specified) with a given confidence.

v We should not interpret the results in isolation but use them to revise what we believed before the study.

11i Cost, acceptability, displacement of resources, gravity of problem, etc... Clinically significant Δ = the value of Δ at which we should switch from old to new or "the Δ that makes a Δ" [à la W. Spitzer]

ii $95\%CI(RD) = 0.077–0.066 \pm 1.95\sqrt{\frac{0.077x0.923}{2683} + \frac{0.066x0.934}{2688}}$

= 0.011±0.014 or –0.003 to +0.025. These are all on the proportion scale. In the percentage scale the interval becomes –0.3% to +2.5%. [Note also that, unlike for a test, the CI does not have to use a common p in the variance p{1–p}. In actual practice, the use of a common p for both will barely change the margin of error]

The 2.5% means that if one intervenes on 100, one would prevent an average of 2.5 ie NRT=100/2.5 or 40. The reverse sign on the –0.3% means that one might cause 1 for every 100 intervened on, i.e. no benefit. This translates to 100/0 or infinity. So the 95% CI for NRT is (40,infinity).