Statistical power in RCTs Why does CISMAC insist on so high statistical power? -Relevance for Getting Research into Policy and Practice (GRIPP)

Halvor Sommerfelt and José C. Martines

CISMAC Webinar 2\textsuperscript{nd} of May 2018
Confidence interval of a Relative risk

<table>
<thead>
<tr>
<th>Exposed</th>
<th>Disease</th>
<th>No disease</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>a</td>
<td>b</td>
<td>NE</td>
</tr>
<tr>
<td>No</td>
<td>c</td>
<td>d</td>
<td>NU</td>
</tr>
</tbody>
</table>

$$RR = \frac{a/(a+b)}{c/(c+d)}$$

$$SE_{\ln[RR]} = \sqrt{\frac{1}{a} - \frac{1}{NE} + \frac{1}{c} - \frac{1}{NU}}$$

95% CI of OR = $e^{\ln[RR] \pm 1.96(SE_{\ln[RR]})}$
High statistical power will enable you to detect a smaller effect size and/or detect the expected effect size despite the outcome occurring more rarely (Hawthorne).

Example: In a recent application to the RCN, we stated the following hypothesis:

- 10 mg of daily zinc sulphate given orally as adjunct therapy to Ugandan infants aged 3 to 59 days hospitalized with clinical signs of sepsis will reduce the risk of treatment failure from 20% to ≤13%, i.e. an efficacy of ≥35%.

- Stata: `power twoprop .2 .13, power(0.9)` → Sample size=589 per arm
- ×1.05 for 5% attrition: 589/0.95=620
- Zinc and placebo arms: Enroll: 620*2=1,240
- `power twoprop .2 .13, power(0.8)` → Sample size=441 per arm and lower precision
In addition to presenting a simplistic picture of a set power, effect size and outcome risk, present a matrix or a graph:

Stata: `power twoprop (.1 (.01) .2), n(1240) rrisk(.6 .65 .7)`

Estimated power for a two-sample proportions test

- Pearson's $\chi^2$ test
- $H_0: p_2 = p_1$ versus $H_a: p_2 \neq p_1$

Parameters: $\alpha = .05$, $N = 1178$, $N_1 = 589$, $N_2 = 589$
With a given effect size, what drives “genuine” statistical precision on RCTs/Cohort studies?

• Dichotomous outcomes: Event risk/rate
• Continuous outcomes: SDs
How high quality research supports transition from results to policy and practice
To create and support a sustainable global network of institutions and individuals who carry out high-quality research to develop and test the delivery of interventions to improve maternal, newborn and child health in low- and middle-income countries, and translate the results into policy and practice.
The quality of the evidence affects the confidence on whether estimates of effect are adequate to support a recommendation

- Policy recommendations are based on a systematic and comprehensive assessment: the balance of a recommendation’s potential benefits/harms (and explicit consideration of other relevant factors).
- WHO uses the GRADE (Grading of Recommendations Assessment, Development and Evaluation)\(^1\) approach to assess the quality of evidence and to develop and report recommendations.

\(^1\)http://www.gradeworkinggroup.org/
Guideline Development follows strict process

- Systematic searches, synthesis and quality assessment of the evidence;
- Expert groups - including content experts, methodologists, target users, policy makers, with gender and geographical balance;
- Appropriate collection and management of experts’ conflicts of interest;
- Standards for transparent decision-making process, considering potential harms and benefits, end users values and preferences.
GRADE evidence profiles

Based on systematic reviews for each important outcome and each key question (in PICO format)

• Assessment of the quality of the evidence
• Summary of findings across studies
One example: Multiple micronutrient supplements

For pregnant women (P), do micronutrient supplements (I) compared with iron and folic acid supplements (C) improve maternal and perinatal outcomes (O)?

Outcomes:
• Maternal illness (anaemia, pre-eclampsia/eclampsia [PE/E])
• Side-effects
• Maternal death
• Fetal/newborn illness (congenital anomaly, SGA, low birth weight, preterm birth)
• Fetal/newborn death
## A.6. Multiple micronutrient (MMN) supplements

### EB Table A.6: Multiple micronutrient supplements versus iron (with or without folic acid)\(^1\)


<table>
<thead>
<tr>
<th>Quality assessment</th>
<th>No. of women(^1)</th>
<th>Effect</th>
<th>Certainty</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>13-15 micronutrients</td>
<td>Iron (with or without folic acid)</td>
<td>Relative (95% CI)</td>
</tr>
<tr>
<td><strong>Maternal anaemia (third trimester Hb &lt; 110 g/L)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 randomized trials</td>
<td>no serious risk of bias</td>
<td>no serious inconsistency</td>
<td>no serious indirectness</td>
</tr>
<tr>
<td><strong>Caesarean section</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 randomized trials</td>
<td>no serious risk of bias</td>
<td>no serious inconsistency</td>
<td>no serious indirectness</td>
</tr>
<tr>
<td><strong>Maternal mortality</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 randomized trials</td>
<td>serious(^1)</td>
<td>no serious inconsistency</td>
<td>no serious indirectness</td>
</tr>
<tr>
<td><strong>Small for gestational age</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13 randomized trials</td>
<td>no serious risk of bias</td>
<td>serious inconsistency</td>
<td>no serious indirectness</td>
</tr>
<tr>
<td><strong>Low birth weight</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14 randomized trials</td>
<td>no serious risk of bias</td>
<td>no serious inconsistency</td>
<td>no serious indirectness</td>
</tr>
<tr>
<td><strong>Preterm birth</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14 randomized trials</td>
<td>no serious risk of bias</td>
<td>no serious inconsistency</td>
<td>no serious indirectness</td>
</tr>
</tbody>
</table>
Quality of evidence

Quality ratings apply to the body of evidence for each outcome assessed for each key question and not to individual studies.

Judgement on the risk of bias of each individual study to assess the quality domain of study limitations

The starting point for rating the quality of the evidence is study design:

• randomized controlled trials (RCTs); and
• nonrandomized trials and observational studies
Five factors can lower the quality of the evidence for each outcome

Initial ratings can be adjusted in light of five factors:

1. limitations in study design and execution (risk of bias)
2. indirectness
3. imprecision (reflected in 95%CI)
4. inconsistency and
5. publication bias
Risk of Bias
assessing limitations in study design and execution

The following reduce risk of bias:
• random sequence generation;
• concealment of allocation to treatment group;
• blinding of participants and investigators;
• reporting of data on all study participants (incl. attrition and exclusions);
  and
• complete reporting of all study outcomes specified a priori.
Indirectness/Directness

- Refers to generalizability, external validity, transferability and applicability
- Indirectness arises when the evidence differs in terms of population, intervention, comparator or outcome (PICO), from that of the question of interest
- If the measured effect applies to a population likely to differ from the target population, this rates down the quality of the evidence.
Inconsistency

The results for a given outcome are not similar across studies.

• Severity of inconsistency depends on:
  • magnitude of the differences in the direction and size of the effect
  • the significance of such differences, and
  • whether any of these differences can be explained (e.g., differences in populations, interventions, comparators or outcomes)

Important inconsistency is present if:
• the point estimates vary widely across studies;
• confidence intervals show minimal overlap or none; or
• in the meta-analysis, the test for heterogeneity yields a statistically significant result and the $I^2$ value is high
Imprecision

• Results are imprecise when the data include relatively few participants and few events. This leads to large uncertainty: wide confidence intervals around the estimate of effect.

• One can use the 95% confidence interval for the pooled estimate of effect as the primary way to examine for imprecision.

Alternatively, one can look at the data in relation to the sample size needed for a single adequately powered trial to detect the minimum effect of interest. If the total number of patients included in a systematic review is less than the sample size calculated, one should consider downgrading the quality of the evidence for imprecision.
What are the consequences for reduced reproducibility?

Can we compensate by “just” increasing the sample size?

– For continuous outcomes: Yes (BUT: You must pay dearly for any (unnecessary lack of measurement reproducibility) by having to increase your sample size (and budget/work).

– For dichotomous outcomes: WORSE!

– Also true when a continuous outcome is dichotomized (e.g. weight-for-length transformed into wasting when <-2Z).
Sample mean

Null: green; Alternative: red
We wish to capture

• The true (biologic) variability of study participants

We wish to reduce

• The variability superimposed on the biologic variability by low reproducibility
True
Suboptimal reproducibility

True
Adequate reproducibility

Suboptimal reproducibility

True
Reducing measurement variability

=Increasing reproducibility:

Will reduce superimposed measurement variability and leave us with more “pure” biologic variability
Consequence of poor reproducibility:

1. Impossible to assess *achieved* validity (=accuracy) [no one is really interested in “average accuracy”].

2. Measurement variability is superimposed on the true (biologic) variability → reduced ability to identify differences (effects) which are actually present = reduced power.

3. What if you wish to use a continuous variable to make a dichotomous variable, e.g. WLZ to wasting (i.e. < -2 WLZ)*
Adequate reproducibility

Wasted

Prevalence of wasting = 75%
Wasted

Adequate reproducibility

Prevalence of wasting = 75%

Suboptimal reproducibility

Prevalence of wasting = 60%!

Artificially reduced reproducibility reduces sensitivity to identify wasting, i.e. reduces its prevalence
Adequate reproducibility

Wasted

Suboptimal reproducibility

Prevalence ratio = 2.5%

Prevalence ratio = 30%

Artificially reduced reproducibility reduces specificity to identify wasting, i.e. inflates its prevalence.
Adequate reproducibility

Wasted

Prevalence ratio = $\frac{2.5\%}{75\%}=0.033$

Effect = $1-0.033 = 0.97 = 97\%$
Suboptimal reproducibility

Adequate reproducibility

Prevalence ratio $= \frac{2.5\%}{75\%} = 0.033$

Effect $= 1 - 0.033 = 0.97 = 97\%$

Prevalence ratio $= \frac{30\%}{60\%} = 0.5!$

Effect $= 1 - 0.5 = 0.5 = 50\%$

Biased towards no effect by low reproducibility
Compensation by increasing sample size?

• For comparing means: Yes

• For comparing proportions, i.e. calculating Relative risk (RR) or Prevalence ratio (PR)?: No!

• in fact, RR and PR will appear artificially precise (narrow 95%CIs)
Objective: To determine if routine daily zinc (1 RDA) supplementation reduces incidence of PNEUMONIA

Possible definitions of pneumonia:

• Cough or difficult & fast breathing (WHO criteria) for ALRI
  Easy to use even by field workers
  Misses very few real cases of pneumonia (high sensitivity, 95%)
  Low specificity (50%)

Crepitations on auscultation by physician
  Difficult to use in field conditions
  Low sensitivity (50%)
  High specificity (95%)
If the diagnostic test were 100% sensitive and 100% specific and the intervention was 50% efficacious

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zinc</td>
<td>50</td>
<td>950</td>
<td>1000</td>
</tr>
<tr>
<td>Placebo</td>
<td>100</td>
<td>900</td>
<td>1000</td>
</tr>
</tbody>
</table>

\[
RR = \frac{50/1000}{100/1000} = 0.5
\]
If the diagnostic test were 100% sensitive and 95% specific and the intervention was 50% efficacious

Pneumonia

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zinc</td>
<td>50+48</td>
<td>950-48</td>
<td>1000</td>
</tr>
<tr>
<td>Placebo</td>
<td>100+45</td>
<td>900-45</td>
<td>1000</td>
</tr>
</tbody>
</table>

\[
RR = \frac{98/1000}{145/1000} = 0.68
\]

\[
SE_{\ln(\text{RR})} = \sqrt{\frac{1}{a} - \frac{1}{N_U} + \frac{1}{c} - \frac{1}{N_E}}
\]
What if the diagnostic test were 50% sensitive and 100% specific and the intervention was 50% efficacious?

<table>
<thead>
<tr>
<th>Pneumonia</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zinc</td>
<td>?</td>
<td>?</td>
<td>1000</td>
</tr>
<tr>
<td>Placebo</td>
<td>?</td>
<td>?</td>
<td>1000</td>
</tr>
</tbody>
</table>

$RR = ?$
If the diagnostic test were 50% sensitive and 100% specific and the intervention was 50% efficacious

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zinc</td>
<td>50-25</td>
<td>950+25</td>
<td>1000</td>
</tr>
<tr>
<td>Placebo</td>
<td>100-50</td>
<td>900+50</td>
<td>1000</td>
</tr>
</tbody>
</table>

\[ RR = \frac{25/1000}{50/1000} = 0.5 \]

\[ SE_{\text{ln}(RR)} = \sqrt{\frac{1}{a} - \frac{1}{N_U} + \frac{1}{c} - \frac{1}{N_E}} \]
Underpowered studies yield statistically imprecise effect measures

- (High impact) journals tend to publish papers that report on “positive findings”, while “negative studies” are much harder to publish.

Well-powered studies yield statistically precise effect measures. If they show no effect, they may never be published.
Daily zinc in addition to antibiotics for childhood (2-35 months) pneumonia

Efficacy against treatment failure (after 2 days):
• in community acquired (mainly non-severe) pneumonia:
  5% (95%CI -20%, 22%)

• in hospitalized children with severe pneumonia:
  12% (95%CI -10%, 29%)
  [Pediatrics 2012; 129: 701-8]

Moderate evidence of no (clinically relevant) effect

NEOVITA trial in Tanzania: Neonatal vitamin A supplementation efficacy on risk of death in the first 6 months of life:
Efficacy: -10% (95% CI -26%, 5%).
Cochrane analysis of neonatal vitA supplementation 11 trials: Efficacy: 2% (95%CI -7%, 11%).
<table>
<thead>
<tr>
<th>Efficacy</th>
<th>95%CI of %Efficacy</th>
<th>RR</th>
<th>95%CI of RR</th>
<th>ULCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zinc pneumonia trials in Nepal:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Community-acquired pneumonia</td>
<td>5 %</td>
<td>-20 %</td>
<td>22 %</td>
<td>0.95</td>
</tr>
<tr>
<td>Children hospitalized with severe pneumonia</td>
<td>12 %</td>
<td>-10 %</td>
<td>29 %</td>
<td>0.88</td>
</tr>
<tr>
<td>Neonatal vitA supplementation and infant survival</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NEOVITA Tanzania</td>
<td>-10 %</td>
<td>-26 %</td>
<td>5 %</td>
<td>1.10</td>
</tr>
<tr>
<td>Pooled analysis (Cochrane)</td>
<td>2 %</td>
<td>-7 %</td>
<td>11 %</td>
<td>0.98</td>
</tr>
</tbody>
</table>
Provide >6 month old children with 20 mg per day of dispersible zinc tablets for 2 weeks
Publication bias

• Publication bias is the systematic underestimate or overestimate of the effect of an intervention resulting from the selective publication of studies based on the study results.
• Studies in which no effect is found are less likely to be published!
• The existence or nonexistence of publication bias cannot be confirmed – it can only be suspected.
• Searches of trial registries and the grey literature to identify unpublished studies and thus minimize the risk of bias.
• The risk of publication bias may be assessed using funnel plots and appropriate statistical tests.
• When publication bias is suspected, downgrade the quality of the evidence by one level.
Evidence limitations:

• No serious limitations – the majority of the studies meet all the minimum quality criteria

• Serious limitations – one of the minimum criteria for quality is not met by the majority of studies in the review.
  Result: lowering of the overall quality rating by one level (e.g. “high” becomes “moderate” for RCTs or “low” becomes “very low” for observational studies).

• Very serious limitations – the risk of bias may strongly influence the estimate of effect; study limitations present in the majority of studies
  Result: lowering of the quality by two levels.
Quality level Definition in GRADE

- **High** We are very confident that the true effect lies close to that of the estimate of the effect.

- **Moderate** We are moderately confident in the effect estimate: the true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different.

- **Low** Our confidence in the effect estimate is limited: the true effect may be substantially different from the estimate of the effect.

- **Very low** We have very little confidence in the effect estimate: the true effect is likely to be substantially different from the estimate of effect.
Criteria to consider when moving from evidence to recommendations

• Quality of the evidence
• Balance of benefits and harms
• Values and preferences
• Resource implications
• Priority of the problem
• Equity and human rights
• Acceptability
• Feasibility