COBIDAS: Guidelines to Improve Reporting

Thomas Nichols, PhD
Department of Statistics &
Warwick Manufacturing Group
University of Warwick

WIN Annual Conference
24 January 2017
Overview

- Motivation: The Crises of Reproducibility
- COBIDAS guidelines
A careful argument for intense skepticism of modern scientific results

Cited 4332 times (Jan. 2017, Google Scholar)

Study Positive Predictive Value

- Sampling Units
  - Not a set of subjects
  - A set of research hypotheses!
    - E.g. Hypothesis set in cognitive decline in aging:
      - Vitamin D reduces risk of cognitive decline
      - Exercise reduces risk of cognitive decline
      - Fish oil reduces risk of cognitive decline
      - ...

- For a randomly selected study:
  - Given the study is positive, what is the probability the studied hypothesis is true?
  - I.e. what is the study PPV?
### PPV Arithmetic

<table>
<thead>
<tr>
<th></th>
<th>True Hypothesis H+</th>
<th>False Hypothesis H-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive Finding D+</td>
<td>$P(D+</td>
<td>H+)$ <strong>Power</strong> 1-β</td>
</tr>
<tr>
<td>Negative Finding D-</td>
<td>$P(H+)$</td>
<td>$P(H-)$</td>
</tr>
</tbody>
</table>

- **Notation**
  - $P(H+)$: *probability of a true hypothesis*
  - $R = P(H+) / P(H-)$: *odds of a true hypothesis*
  - Odds vs. probability
    - $P(H+) = R / (R+1)$
### PPV Arithmetic

<table>
<thead>
<tr>
<th></th>
<th>True Hypothesis H+</th>
<th>False Hypothesis H-</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Positive Finding D+</strong></td>
<td>P(D+</td>
<td>H+) \textit{Power} 1-β</td>
</tr>
<tr>
<td><strong>Negative Finding D-</strong></td>
<td>P(H+)</td>
<td>P(H-)</td>
</tr>
</tbody>
</table>

- **With Bayes Theorem...**

\[
PPV = P(H+|D+) = \frac{(1-\beta) R}{(1-\beta) R + \alpha}
\]

- That is, PPV depends on
  - Power \((1-\beta)\)
  - Odds of a true hypothesis \((R)\) &
  - False positive rate \((FPR, \alpha)\)
PPV Arithmetic

- When is PPV > ½?

\[
0.5 > PPV = \frac{(1-\beta) R}{(1-\beta) R + \alpha} \Rightarrow (1-\beta)R > \alpha
\]

- This is always true: \((1-\beta) > \alpha\)
- So, if \(R = 1\), PPV > ½
- If \(R < \frac{1}{2}\), then PPV might < ½

- PPV & Power

\[
PPV = \frac{(1-\beta) R}{(1-\beta) R + \alpha} = (1-\beta) \frac{R}{R + \alpha/(1-\beta)} \approx (1-\beta)
\]

- Bigger PPV, better power
- Lower PPV, worse power
PPV Arithmetic

- **PPV & “bias”**
  - Suppose fraction $u$ of all studies shouldn’t have been published but are
    - i.e. won’t have been published if no bias
    - Due to “vibration effects”
      - *Not* the $\alpha$ fraction of chance false positive studies
      - *Not* usual estimation bias per se
  - Then...
    \[ PPV = \frac{(1-\beta) R + u \beta R}{(1-\beta) R + u \beta R + \alpha + u(1-\alpha)} \]
  - As $u$ increases, PPV drops
Exploring study PPV

- PPV depends on $u$ & power
  - Skepticism of a discipline (high ‘bias’ frequency $u$) translates to lower PPV

**PPV vs. R - For different levels of bias $u$**

- Power = 80%
- Power = 20%
Exploring “any” PPV

- Suppose $n$ research teams all study a hypothesis
- Define “D+” as one or more of those teams getting a finding
  - They ‘busier’ the discipline, the lower the PPV

**PPV vs. R - For number of research teams**

- Power = 80%
- Power = 20%
Table 4. PPV of Research Findings for Various Combinations of Power ($1 - \beta$), Ratio of True to Not-True Relationships ($R$), and Bias ($u$)

<table>
<thead>
<tr>
<th>$1 - \beta$</th>
<th>$R$</th>
<th>$u$</th>
<th>Practical Example</th>
<th>PPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.80</td>
<td>1:1</td>
<td>0.10</td>
<td>Adequately powered RCT with little bias and 1:1 pre-study odds</td>
<td>0.85</td>
</tr>
<tr>
<td>0.95</td>
<td>2:1</td>
<td>0.30</td>
<td>Confirmatory meta-analysis of good-quality RCTs</td>
<td>0.85</td>
</tr>
<tr>
<td>0.80</td>
<td>1:3</td>
<td>0.40</td>
<td>Meta-analysis of small inconclusive studies</td>
<td>0.41</td>
</tr>
<tr>
<td>0.20</td>
<td>1:5</td>
<td>0.20</td>
<td>Underpowered, but well-performed phase I/II RCT</td>
<td>0.23</td>
</tr>
<tr>
<td>0.20</td>
<td>1:5</td>
<td>0.80</td>
<td>Underpowered, poorly performed phase I/II RCT</td>
<td>0.17</td>
</tr>
<tr>
<td>0.80</td>
<td>1:10</td>
<td>0.30</td>
<td>Adequately powered exploratory epidemiological study</td>
<td>0.20</td>
</tr>
<tr>
<td>0.20</td>
<td>1:10</td>
<td>0.30</td>
<td>Underpowered exploratory epidemiological study</td>
<td>0.12</td>
</tr>
<tr>
<td>0.20</td>
<td>1:1,000</td>
<td>0.80</td>
<td>Discovery-oriented exploratory research with massive testing</td>
<td>0.0010</td>
</tr>
<tr>
<td>0.20</td>
<td>1:1,000</td>
<td>0.20</td>
<td>As in previous example, but with more limited bias (more standardized)</td>
<td>0.0015</td>
</tr>
</tbody>
</table>
OK, but what’s the evidence?

- This is a thought experiment
  - Sampling frame “Research hypotheses”
  - Many studies experience “bias”, but this may take P-values from 0.0001 when then should be 0.005

- Is there really a problem here?
  - Canary in the coal mine, or
  - Chicken Little?
Exhibit A: Law of Small numbers

- Or “Winner’s Curse”
  - Small studies over-estimate effect size

- 256 meta analyses for a dichotomous effect (odds ratio) from Cochrane database
- Studies with smallest N have biggest effect size!
  - Low N studies have low power
  - Low-power studies rarely succeed, but when they do, is result of randomly high effect or randomly small variance, biasing effect size
- Explains difficulty with replication

Two Problems

- Suppressed studies & Biased effects
  - P>0.05 not published
  - Biases that afflict small studies more than large studies

File drawer problem
(Unpublished non-significant studies)

Bias
(Fishing or Vibration Effects)
Vibration Effects

- Sloppy or nonexistent analysis protocols
  
  “Try voxel-wise whole brain, then cluster-wise, then if not getting good results, look for subjects with bad movement, if still nothing, maybe try a global signal regressor; if still nothing do SVC for frontal lobe, if not, then try DLPFC (probably only right side), if still nothing, will look in literature for xyz coordinates near my activation, use spherical SVC… surely that’ll work!”

- You stop when you get the result you expect
- These “vibrations” can only lead to inflated false positives

- Afflicts well-intended researchers
  
  - Modern, “big data” scientific tools have multitude of preprocessing options, modeling choices
    - Pre-modelling normalisation options
    - Even more choices of options, covariates, interactions
Exhibit B: Studies chronically under powered

- Review of 730 neuroscience studies
  - Extracted from 48 meta analyses
  - Power of each of 730 studies calculated

- Median power 20%
  - For 50% of studies, fewer than 1 in 5 replications will succeed!

Exhibit C: Mass replication

- Open Science Collaboration: Psychology
  - Replicated 100 new & classic studies
  - Effort of 270 scientists
- Each replication ‘registered’
  - Carefully powered (1-β ≈ 90%)
  - Extensive peer review (usually with original authors contributing) in preparing study
  - Complete details of study protocol & analysis publically recorded and fixed

Exhibit C: Mass non-replication

- **p-value**
  - Red: Not Significant
  - Blue: Significant

- **Replication Power**
  - 0.6
  - 0.7
  - 0.8
  - 0.9

- **Graph**: Scatter plot showing the relationship between original effect size and replication effect size, with different colors and sizes indicating significance and replication power.
Exhibit C: Mass non-replication

- Mean replication effect size half of original
  - In correlation units: Orig. 0.403  Repl. 0.197
- Most replications not significant
  - P<0.05 significant: Orig. 97%  Repl. 36%
- Joint analysis of Orig. & Repl.
  - 68% significant
What can be done?

- **TOP – Transparency Openness Promotion**
  - Advancing open science goals in service of reproducibly
  - Articulated by
  - Provides 8 areas, 4 levels of success

---

### Summary of the eight standards and three levels of the TOP guidelines

Levels 1 to 3 are increasingly stringent for each standard. Level 0 offers a comparison that does not meet the standard.

<table>
<thead>
<tr>
<th></th>
<th>LEVEL 0</th>
<th>LEVEL 1</th>
<th>LEVEL 2</th>
<th>LEVEL 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Citation standards</strong></td>
<td>Journal encourages citation of data, code, and materials—or says nothing.</td>
<td>Journal describes citation of data in guidelines to authors with clear rules and examples.</td>
<td>Article provides appropriate citation for data and materials used, consistent with journal’s author guidelines.</td>
<td>Article is not published until appropriate citation for data and materials is provided that follows journal’s author guidelines.</td>
</tr>
<tr>
<td><strong>Data transparency</strong></td>
<td>Journal encourages data sharing—or says nothing.</td>
<td>Article states whether data are available and, if so, where to access them.</td>
<td>Data must be posted to a trusted repository. Exceptions must be identified at article submission.</td>
<td>Data must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.</td>
</tr>
</tbody>
</table>
Elements of TOP

- Citation standards
- Data transparency
- Analytic methods (code) transparency
- Research materials transparency
- Design and analysis transparency
- Preregistration of studies
- Preregistration of analysis plans
- Replication
TOP Update (1/2)

- Citation standards
  - Citation of data, code and materials
  - Level 3: Complete citation of all data, code and materials
    - e.g. New Science standard

- Data/Code/Materials transparency
  - Availability of data/code/materials
  - Level 3: Before pub., data, code & materials posted to trusted repository; reported analyses independently reproduced
    - e.g. “R” kite-mark in Biostatistics
TOP Uptake (2/2)

- Design and analysis transparency
  - Completely described design, following best practice
  - Level 3: Journal requires and enforces adherence to design standards for review and publication
    - Small steps: *Nature / Nature Neuroscience* check lists

- Preregistration of Study/Analysis Plan
  - Level 3: Required

- Replication
  - Facilitation of replication studies
  - Level 3: Registered report article type
Yes, the sky is falling.

- Many reasons to worry about validity of scientific literature
- Researchers need to...
  - Do power calculations
  - Disclose methods & findings transparently
  - Pre-register your study protocol and analysis plan
  - Make study materials and data available
  - Work collaboratively to increase power and replicate findings
    - Meta-Analyses
OHBM Committee On Best Practice In Data Analysis & Sharing (COBIDAS)

- White paper with checklists of practice & reporting, for all variants of MRI of the brain
- Emphasis on comprehensive *reporting*
  - Practice too varied to be prescriptive, except
- Best practice give for 3 areas
  - Statistical modeling, data sharing & reproducibility
- Published bioRxiv doi:10.1101/054262 20 May 2016
  - Commentary forthcoming in *Nature Neuroscience*

http://www.humanbrainmapping.org/cobidas
COBIDAS Structure: 7 Key Areas

Experimental Design Reporting

Acquisition Reporting
Subject preparation. MRI system description. MRI acquisition. Preliminary quality control.

Preprocessing Reporting

Statistical Modeling & Inference

Results Reporting
Mass univariate analysis. Functional connectivity.

Data Sharing
Define data sharing plan early. Database for organized data.

Reproducibility
Documentation. Archiving. Citation.
Table D.1. Experimental Design Reporting

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Notes</th>
<th>Mandatory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of subjects</td>
<td>Elaborate each by group if have more than one group.</td>
<td></td>
</tr>
<tr>
<td>Subjects approached</td>
<td></td>
<td>N</td>
</tr>
<tr>
<td>Subjects consented</td>
<td></td>
<td>N</td>
</tr>
<tr>
<td>Subjects refused to participate</td>
<td>Provide reasons.</td>
<td>N</td>
</tr>
<tr>
<td>Subjects excluded</td>
<td>Subjects excluded after consenting but before data acquisition: provide reasons.</td>
<td>N</td>
</tr>
<tr>
<td>Subjects participated and analyzed</td>
<td>Provide the number of subjects scanned, number excluded after acquisition, and the number included in the data analysis. If they differ, note the number of subjects in each particular analysis.</td>
<td>Y</td>
</tr>
<tr>
<td>Inclusion criteria and descriptive statistics</td>
<td>Elaborate each by group if have more than one group.</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>Mean, standard deviation and range.</td>
<td>Y</td>
</tr>
<tr>
<td>Sex</td>
<td>Absolute counts or relative frequencies.</td>
<td>Y</td>
</tr>
<tr>
<td>Race &amp; ethnicity</td>
<td>Per guidelines of NIH or other relevant agency.</td>
<td>N</td>
</tr>
<tr>
<td>Education, SES</td>
<td>Education is essential for studies comparing patient and control groups; complete SES reporting less important for single-group studies, but still useful. Specify measurement instrument used; may be parental SES and education if study has minors.</td>
<td>Y</td>
</tr>
<tr>
<td>IQ</td>
<td>Specify measurement instrument used.</td>
<td>N</td>
</tr>
<tr>
<td>Reporting items: Experimental Design</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Handedness</strong></td>
<td>Absolute or relative frequencies; basis of handedness-attribution (self-report, EHI, other tests). (Important for fMRI, may be less important for structural studies.)</td>
<td>Y</td>
</tr>
<tr>
<td><strong>Exclusion criteria</strong></td>
<td>Describe any screening criteria, including those applied to “normal” sample such as MRI exclusion criteria.</td>
<td>Y</td>
</tr>
<tr>
<td><strong>Clinical criteria</strong></td>
<td>Detail the area of recruitment (in- vs. outpatient setting, community hospital vs. tertiary referral center etc.) as well as whether patients were currently in treatment.</td>
<td>Y</td>
</tr>
<tr>
<td><strong>Clinical instruments</strong></td>
<td>Describe the instruments used to obtain the diagnosis and provide tests of intra- or inter-rater reliability. Clarify whether a “clinical diagnosis” or “inventory diagnosis” was used (if applicable). State the diagnostic system (ICD, DSM etc) that was used.</td>
<td>Y</td>
</tr>
<tr>
<td><strong>Matching strategy</strong></td>
<td>If applicable.</td>
<td>Y</td>
</tr>
<tr>
<td><strong>Population &amp; recruitment strategy</strong></td>
<td>Population from which subjects were drawn, and how and where recruitment took place, e.g., schools, clinics, etc. If possible, specify if subjects in research study have participated in other studies before.</td>
<td>Y</td>
</tr>
<tr>
<td><strong>Subject scanning order</strong></td>
<td>With multiple groups, information on ordering and/or scanning sequence; especially report relative to scanner changes/upgrades. (Ideally, if possible, random or interleaved order to avoid bias due to scanner changes/upgrades.)</td>
<td>Y</td>
</tr>
<tr>
<td><strong>Neurocognitive measures</strong></td>
<td>All measures collected on subjects should be described and reported.</td>
<td>Y</td>
</tr>
<tr>
<td><strong>Ethical considerations</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Ethical approval</strong></td>
<td>Describe approval given, including the particular institutional review board, medical ethics committee or equivalent that granted the approval. When data is shared, describe the ethics/institutional approvals required from either the author (source) or recipient.</td>
<td>Y</td>
</tr>
<tr>
<td><strong>Informed consent</strong></td>
<td>Record whether subjects provided informed consent or, if applicable, informed assent.</td>
<td>Y</td>
</tr>
<tr>
<td>Reporting items: Experimental Design</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Design specifications</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Design type</td>
<td>Task or resting state. Event-related or block design. (See body text for usage of ‘block design’ terminology.)</td>
<td>Y</td>
</tr>
<tr>
<td>Condition &amp; stimuli</td>
<td>Clearly describe each condition and the stimuli used. Be sure to completely describe baseline (e.g. blank white/black screen, presence of fixation cross, or any other text), especially for resting-state studies. When possible provide images or screen snapshots of the stimuli.</td>
<td>Y</td>
</tr>
<tr>
<td>Number of blocks, trials or experimental units</td>
<td>Specify per session, and if differing by subject, summary statistics (mean, range and/or standard deviation) of such counts.</td>
<td>Y</td>
</tr>
<tr>
<td>Timing and duration</td>
<td>Length of each trial or block (both, if trials are blocked), and interval between trials. Provide the timing structure of the events in the task, whether a random/jittered pattern or a regular arrangement; any jittering of block onsets.</td>
<td>Y</td>
</tr>
<tr>
<td>Length of the experiment</td>
<td>Describe the total length of the scanning session, as well as the duration of each run. (Important to assess subject fatigue.)</td>
<td>Y</td>
</tr>
<tr>
<td>Design optimization</td>
<td>Whether design was optimized for efficiency, and how.</td>
<td>Y</td>
</tr>
<tr>
<td>Presentation software</td>
<td>Name software, version and operating system on which the stimulus presentation was run. When possible, provide code used to drive experiment.</td>
<td>Y</td>
</tr>
<tr>
<td><strong>Task specification</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Condition</td>
<td>Enumerate the conditions and fully describe and reference each. Consider using a shorthand name, e.g. AUDSTIM, VISSTIM, to refer to each condition, to clarify the distinction between a specific modeled effect and a psychological construct. Naming should reflect the distinction between instruction periods and actual stimuli, and between single parameters and contrasts of parameters.</td>
<td>Y</td>
</tr>
<tr>
<td>Instructions</td>
<td>Specify the instructions given to subjects for each condition (ideally the exact text in supplement or appendix). For resting-state, be sure to indicate eyes-closed,</td>
<td>Y</td>
</tr>
</tbody>
</table>
## Reporting items: Experimental Design

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
<th>Yes/No</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Stimuli</strong></td>
<td>specifics of stimuli used in each run. For example, the unique number of stimuli used, and whether/how stimuli were repeated over trials or conditions.</td>
<td>Y</td>
</tr>
<tr>
<td><strong>Randomization</strong></td>
<td>describe block or event ordering as deterministic, or report manner of randomization, in terms of order and timing. If pseudo-randomized, i.e. under constraints, describe how and the criteria used to constrain the orders/timings.</td>
<td>Y</td>
</tr>
<tr>
<td><strong>Stimulus presentation &amp; response collection.</strong></td>
<td>specify the presentation hardware (e.g. back projection, in-room display, goggles, etc), and the response systems (e.g. button boxes, eye tracking, physiology). Note how equipment was synched to the scanner (e.g. scanner TTL, or manual sync.)</td>
<td>Y</td>
</tr>
<tr>
<td><strong>Run order</strong></td>
<td>order in which tasks runs are conducted in the scanner.</td>
<td>Y</td>
</tr>
<tr>
<td><strong>Power analysis</strong></td>
<td>specify the type of outcome used as the basis of power computations, e.g. signal in a pre-specified ROI, or whole image voxelwise (or cluster-wise, peak-wise, etc.).</td>
<td>Y</td>
</tr>
</tbody>
</table>
| **Power parameters**                          | specify  
  - Effect size (or effect magnitude and standard deviation separately).  
  - Source of predicted effect size (previous literature with citation; pilot data with description, etc).  
  - Significance level (e.g. uncorrected alpha 0.05 for an ROI, or FWE-corrected significance)  
  - Target power (typically 80%).  
  - Any other parameters set (e.g., for spatial methods a brain volume and smoothness may be needed to be specified).                                                                                      | Y      |
Before starting your next trial, head to EQUATOR Network — Likely has relevant best-practice guidelines for you!
THANK YOU!
Open Discussion

- Has reliability/reproducibility of findings become an issue in your discipline? If so, how has the discipline reacted?
- What practices can you follow to ensure that someone else, given your data, could obtain your same results?
- What practices can you follow to ensure that someone else, starting from scratch, collecting new data, could obtain the same results that you have obtained?