# Towards Development of a Checklist for Scientific Studies

Alison L Gibbs

November 2019

In the most recent issue of *Teaching Statistics*, George Cobb expresses concern that we fail “to teach the difference between uncertainty and ambiguity.” Ambiguity includes questions about how the data were obtained, how the analysis to carry out was chosen, and the meaning and generalizability of the conclusions. We currently do not have models for teaching ambiguity. He suggests that one place to start might be the development of checklists.

Proposed student exercise: *Develop a checklist.*

* Have the students read a study, in the news and scientific reports.
* Have the students brainstorm questions to ask.
* Repeat several times throughout the course, and every time a new study is introduced.

## Prompting questions:

1. Is the statistical analysis appropriate?
2. What might stop the study from being generalizable to a larger group?
3. How might the study results be wrong?
4. Is the result interesting?
5. Given what was observed, what else might be possible?

## Another question:

* If the study is convincing, did you need P-values to be convinced?

## A resource in the clinical trial setting

The series of articles *Endgames: Statistical Question* in the **BMJ** by Philip Sedgwick

*An example:*

<https://www.bmj.com/content/343/bmj.d4176.full>

*Researchers assessed the effects of full length 5 degree lateral wedge insoles on improving symptoms and slowing structural disease progression in medial knee osteoarthritis. A randomised double-blind, controlled trial was performed. The control treatment was flat insoles. Both types of insole were worn inside the shoes daily for 12 months.*

*Participants were recruited from the community if they were aged 50 years or over and had a clinical and radiographic diagnosis of mild to moderately severe medial knee osteoarthritis. In total, 103 individuals were randomised to lateral wedge insoles and 97 to flat insoles. The primary symptomatic outcome was self-rated overall knee pain in the past week. The primary structural outcome was volume of medial tibial cartilage from magnetic resonance imaging scans. For both outcomes, the change at 12 months from baseline was recorded. When compared with flat insoles, lateral wedge insoles provided no symptomatic or structural benefits when worn for 12 months.*

*Which of the following types of bias, if any, would have been minimised by the study design?*

*a) Allocation bias  
b) Ascertainment bias*

*c) Detection bias  
d) Assessor bias  
e) Response bias*

*Another example:*

<https://www.bmj.com/content/343/bmj.d7839.full>

*The association between adherence to lifestyle recommendations and risk of colorectal cancer was investigated using a prospective cohort study. Adherence to lifestyle recommendations was measured by a lifestyle index based on five factors—physical activity, waist circumference, smoking, alcohol intake, and diet.*

*Between 1993 and 1997, all Danish men and women aged 50-64 years who lived in Copenhagen and Aarhus were invited to participate. In total, 160 725 potential participants were identified and 57 053 (35%) accepted the invitation. Of the respondents, 569 were excluded because they had recently been diagnosed with cancer. A further 997 respondents were excluded because of missing data, resulting in a cohort size of 55 487. Adherence to lifestyle recommendations was measured at recruitment. Cohort participants were followed for a median of 9.9 years and the date of diagnosis of any cancer (except non-melanoma skin cancer) was recorded. The researchers concluded that adherence to recommendations for lifestyle factors may considerably reduce the risk of colorectal cancer.*

*Which of the following, if any, might the above cohort study have been prone to?*

* 1. *Allocation bias*
  2. *Healthy entrant effect*
  3. *Non-response bias*
  4. *Selection bias*

*[Answer: Only a is false]*

## Possible ideas for the checklist:

1. Is the statistical analysis appropriate?
   1. Does the method make sense for how the data were collected?
   2. Was the analysis chosen in advance?
   3. Are the conditions satisfied for the statistical method used?
   4. How was the choice of variables to include made?
   5. Was sensitivity analysis carried out to see how important decisions made were?
2. What might stop the study from being generalizable to a larger group?
   1. How were the study participants chosen?
   2. Why might or might not it be possible to extrapolate the conclusions beyond the study setting?
3. How might the study results be wrong?
   1. Are there plausible alternative explanations (e.g., confounders)?
   2. Are there identifiable sources of bias?
   3. Does a laboratory experiment reflect clinical practice?
4. Is the result interesting?
   1. Can causal conclusions be made?
   2. Is the observed effect size meaningful?
   3. Are the results consistent with a meaningful theory?
5. Given what was observed, what else might be possible?

## Other questions:

* Is the study exploratory or confirmatory?
* Does the analysis answer the research question?
* What decisions did the research make in the data analysis? (adjustment of unusual values, imputation of missing values, transforming variables, which variables to include in model, interactions included?, degree of polynomial?, use of graphs to identify patterns to explore, remedial measures from residual plots, which data to include / exclude)
* Was randomization used?
* Are the results practically meaningful?
* Any concerns about the quality of the data?
* Is what is observed plausible?
* Does the study claim that absence of evidence implies evidence of absence?
* Is there a control?
* Do the study conditions reflect the issue under consideration?
* What am I not being told? What else did they look for? Could this result be cherry-picked?
* Might the study participants behave differently because they are in the study (Hawthorne effect)?
* Could results be explained by regression to the mean?
* Are there implications due to the placebo affect?

## Sources of bias (systematic distortion):

* Allocation bias – the method by which experimental units are assigned to treatments causes systematic differences in subjects between treatments
* Ascertainment bias (detection bias) – systematic distortion of measurements because of awareness of treatment allocation
  + Assessor bias – ascertainment bias on the part of the researchers
  + Response bias – ascertainment bias on the part of the research participants, subjects report results in a socially acceptable way
* Selection bias
  + Non-response bias – people who choose not to participate (turn down invitation) are different than those who do
  + Volunteer bias – people who volunteer to participate are not representative of the population
  + Survivorship bias – only some can be studied
* Co-intervention bias – subjects in a study go out and seek other treatment
* Attrition bias – patients in the control group may be more likely to drop out than patients in a treatment group; resentful demoralization is the phenomenon where participants who do not get their preferred treatment do not comply or withdraw; lack of compliance can lead to dilution bias
* Outcome reporting bias – statistically significant outcomes are what get reported in publications
* Publication bias – studies with significant results get published more than those without significant results
* Publication bias
* Healthy worker bias -- those in the workplace are healthier than the general population
* Recall bias – systematic difference in recall between treatment gro

Kari Lock Morgan’s evaluation of evidence.

Comparing A vs B. Observe that A has better outcomes than B in the data.

Possible explanations:

1. A causes better outcomes than B
2. The groups differ at baseline
3. Just random chance