UC DAVIS HEALTH

HARKing:
What is it and why is it bad?

CLINICAL AND TRANSLATIONAL SCIENCE CENTER

Machelle Wilson, PhD

Clinical and Translational Science Center

Division of Biostatistics, Department of Public Health Sciences

School of Medicine

University of California, Davis

The UC Davis CTSC receives support from the NIH National Center for Advancing Translational Sciences (award TR001860).

HARKing: What is it and why is it bad?

- Outline
  - Confirmatory versus Exploratory Research (Schwab and Held)
  - HARKing: When does it hurt (Rubin)
  - Taxonomy of HARKing Behaviors (Murphy and Aguinis)

Confirmatory vs Exploratory Research (Schwab and Held)

- Confirmatory research:
  - Begins with a clear set of hypotheses before data are collected.
  - Tests only *a priori* hypotheses.
  - Goal is high specificity, i.e, eliminating false hypotheses.
  - Suitable for establishing strong evidence and confirming expected.
- Example: Assessing the efficacy of a drug in humans.

Confirmatory vs Exploratory Research

- Exploratory Research:
  - Deals with questions that have not yet been studied at length.
  - May begin with loosely formulated hypotheses.
  - The goal is high sensitivity, i.e., identifying true hypotheses.
  - Suitable for exploring new possibilities and finding the unexpected.
- Example: Testing new compounds in mice.

Dangers of Exploratory Research

- **There is more room for fudging**.
  - An investigator may see interesting results and present them as if they confirm an *a priori* hypothesis.

- **Researcher may only report significant results.**
  - Researcher may test for many associations and only report those that were found to be significant.
- **That is, an investigator may begin exploratory research and then pretend it was confirmatory all along.**

Dangers of Exploratory Research

"…Scientists are on dangerous ground when they merge and confuse confirmatory and exploratory research. For example, researchers might present exploratory results as confirmatory in order to increase the probability of publication. But what they will have done here is engage in the practice of HARKing, and this increases the risk that a false-positive finding will make its way into the scientific literature, and decreases the overall likelihood of the result being reproducible, replicable, or generalizable."

HARKing: When does it hurt? (Rubin)

- **Types of HARKing**:
  - Including *post hoc* hypotheses as if they were *a priori*.
  - Excluding *a priori* hypotheses that were not confirmed.
  - Retrieving hypotheses from a *post hoc* literature search and reporting them as *a priori* hypotheses.

Example of HARKing

- Consider a researcher who hypothesizes that expressions of prejudice increase self-esteem.
- They randomly assign a sample of participants to either describe negative feelings about immigrants or to describe positive feelings about immigrants and then measure the participants' self-esteem.
- They find that participants in the negative feelings condition have significantly *lower* self-esteem scores than those in the positive feelings condition.
- In an effort to accommodate this unexpected finding, they construct a new *post hoc* hypothesis that expressions of prejudice *reduce* self-esteem. They then include this *post hoc* hypothesis in the research report as if it were an *a priori* hypothesis.
- Then, they remove any mention of the original hypothesis from the report.

How Does HARKing Hurt Science?

- Results in hypotheses that are always confirmed and never falsified.
- Hence HARKing harms the progress of science by preventing the research community from identifying already falsified hypotheses.
- HARKing leads to irreproducibility or the ‘Replication Crisis’.
- When hypotheses are uniquely tailored to a given sample, it increases the probability that the findings are not reproducible or generalizable in the population of interest. *This is the key concept*.

A Taxonomy of HARKing Behaviors (Murphy and Aguinis)

- Less problematic (little potential to bias cumulative knowledge)
  - Hypothesis proliferation: An author adds hypotheses to a study after data are collected and analyzed to place added emphasis on a result that was not part of the original conceptual design but was nevertheless going to be reported in the manuscript (e.g., full correlation table or interesting results from Table 1 (demographics and clinical variables).
  - THARKing: An author transparently HARKs in the discussion section of a paper by forming new hypotheses on the basis of results obtained.

A Taxonomy of HARKing Behaviors

- More problematic (great potential to bias cumulative knowledge)
  - Cherry-picking: An author searches through data involving alternative measures or samples to find the results that offer the strongest possible support for a particular hypothesis or research question.
  - Question-trolling: An author searches through data involving several different constructs, measures of those constructs, interventions, or associations to find seemingly notable results worth writing about.

Cherry-picking and Question-trolling

- Because both cherry-picking and question-trolling systematically capitalize on chance fluctuations in the effect size estimates (i.e., differences between groups or associations with risk factors) produced by different samples or measures, the effects of HARKing in a field of study are likely to be systematically related to:
  - (1) sample size,
  - (2) the size of the pool of sample results the researcher has to choose from,
  - (3) the heterogeneity of the population effects that underlie that pool of sample statistics, and
  - (4) the prevalence of the forms of HARKing in the field.

Study Features: Sample Size

- The smaller the sample, the more variability one expects in sample statistics, and therefore, the larger the opportunity for a biased sampling procedure to lead to results that deviate sharply from the population effect.

Study Features: Number of Tests

- All other things being equal, a researcher who scans a set of ten sample statistics before selecting one as the basis for their *post hoc* hypothesis will have more opportunities to seriously overestimate population effects than another researcher who scans a set of just three sample statistics.

Study Features: Heterogeneity

- Cherry-picking: Involves choosing the strongest association among the many tested for a given outcome.
- For example: A researcher has 3 measures of toxicity and 3 measures of self-reported pain. This would yield 9 estimated correlations that are all measuring *the same thing*.
- Cherry-picking would involve reporting only the correlation that was strongest and most significant.
- Hence, cherry picking involves selection among *homogeneous* effects.

Study Features: heterogeneity

- Question-trolling involves studying a large variety of associations across a large variety of measures.
- For example, a researcher might study all pair-wise associations between toxicity, pain, patient satisfaction, 1 year survival, several adverse events, and tumor regression and then chose to report only the associations that are significant.
- Hence, question-trolling involves selection among *heterogenous* effects.

Effects of Cherry-Picking and Question-Trolling

- Both cherry-picking and question-trolling result in bias in the scientific literature.
- Question-trolling is worse.
- Small sample sizes, large numbers of tests, and heterogenous selection result in higher bias than large sample sizes, small numbers of tests, and homogeneous selection.
- THARKing is unlikely to lead to bias.

THARKing: Being honest and Upfront

- **For example, say you were interested in the effects of BMI on the risk of heart failure in women**.
  - You test the association using a time-to-event proportional hazards model.
  - BMI is not significant, though many of the potential confounders you controlled for in the model are significant.
  - One of the confounders is hypertension. You hypothesize that BMI is closely associated with hypertension and that the relationship moderates the effect of BMI on heart failure.
  - Hence, you add an interaction term between BMI and hypertension to see if the effects of BMI on heart failure differ for those with hypertension compared to those without. The results turn out to be significant.
  - You then add this analysis and its results to the Discussion section, while reporting the analysis and results for BMI alone in the main body of your article.

Who's Responsible for HARKing?

- Note that to progress the body of knowledge in a scientific field there is a need for reducing the degree of HARKing, which requires increasing the dissemination of negative results.
  - Journals should be willing to accept negative results from well-designed studies.

    There is an increasing trend among scientific journals to allow the publication of interesting negative results.
  - Investigators should always perform well-designed studies.

    Negative results from under-powered studies are not interesting. It's important for investigators to power their studies at a clinically interesting effect size.

  - *To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of.* ~ Sir Ronald Aylmer Fisher
  - *The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.* ~ John Tukey

Take Home Message

- Design your study so that all your hypotheses are well-characterized and justified by theory and preliminary evidence.
- Make sure your study is adequately powered.
- Do not look at the data or results and then come up with hypotheses, except in the Discussion section.
- Do not test hypotheses on the same data that generated them.
- Report the results of the planned analyses as observed, even if unfavorable.
- If any new and interesting hypotheses are generated after seeing the results, be transparent about the process in your publication.

References

- Schwab, Simon; Held, Leonard (2020) “Different Worlds: Confirmatory vs Exploratory Research”, *Significance, April 2020. The Royal Statistical Society.*
- Rubin, Mark (2017) “When Does HARKing Hurt? Identifying When Different Types of Undisclosed Post Hoc Hypothesizing Harm Scientific Progress”, *Review of General Psychology, 21,* 308-320.
- Murphy, Kevin R.; Aguinis, Herman (2019) “HARKing: How Badly Can Cherry-Picking and Question Trolling Produce Bias in Published Results?”, *Journal of Business and Psychology (2019) 34:1–17.*