Replication and Evidence

Tyler J. VanderWeele
Departments of Epidemiology and Biostatistics
Harvard T.H. Chan School of Public Health

Prelude to Replication Crisis: Some Bad Examples of P-Value Misuse

Hazard ratio of autism by antidepressant exposure during pregnancy (*JAMA* 2017; 317(15):1544-52)

Without adjustment: 2.16 (95% CI: 1.64-2.86)
Conventional adjustment: 1.59 (95% CI: 1.17-2.17)
IP weighting adjustment: 1.61 (95% CI: 0.997-2.59)

Conclusion: "antidepressant exposure compared with no exposure was not associated with autism spectrum disorder in the child"

Yes, it was!

Prelude to Replication Crisis: Some Bad Examples of P-Value Misuse

Blumenthal et al. (2019; Neurology, 92(3):e212-e223)

Design. 2-by-2 factorial randomized clinical trial of aerobic exercise (AE) and DASH diet

Results. “Participants who engaged in AE (d = 0.32, p = 0.046) but not those who consumed the DASH diet (d = 0.30, p = 0.059) demonstrated significant improvements in the executive function domain.”

Prelude to Replication Crisis: Some Bad Examples of P-Value Misuse

Schmidt et al. (2011; BMJ 2011;343:d3450)

Adjusted RR for current use of newer selective COX-2 inhibitors on atrial fibrillation was 1.20 (95% CI: 1.09, 1.33)

Chao et al. (2013; Int J Cardiol 168:312-316)

Adjusted RR for current use of newer selective COX-2 inhibitors on atrial fibrillation was 1.20 (95% CI: 0.97, 1.48)

"A unique finding of the present study was that the use of selective COX2 inhibitors was not associated with AF risk"

"In contrast to their [Schmidt et al.] findings, selective COX2 inhibitors were not associated with new-onset AF in our study"

Plan of Presentation

(1) Evidence, Truth, and Knowledge

(2) Stability and Evidence within a Study

(3) Replication and the Replication Crisis

(4) Reasons for Lack of Replication

(5) Solutions to Improve Replicability

(6) Humility in Knowledge

(7) Concluding Remarks

Science, Truth and Decision-Making

Goals of science are often taken as (i) discovery of truth and (ii) information needed for decision-making and action

Truth as Correspondence:

Truth (OED): “that which is in accordance with fact or reality”

Truth gives us understanding of the world
If understanding is in accord with reality, decision-making will be better
Decision-making requires more than science; it also requires values, but it does generally require good science and the knowledge that results
Sometimes we must make decisions under uncertainty, without knowing whether we have arrived at truth, but evidence and information is still relevant for decision making (RDS280)

We often do not know whether we have arrived at truth, but when evidence becomes stronger and stronger we may eventually begin to speak of *knowledge*

Knowledge

Classical Definition of Knowledge (Plato): Justified true belief

But…
Gettier Paradox (1963): Smith sees a clock in a train station that reads 1:17 and believes it is 1:17. The clock has in fact been broken for days and its hands not moving, but as it turns out it is in fact 1:17 and it is one of the rare moments the clock is correct. Does Smith "know" that it is 1:17…? (Russell, 1948)

Most would say "no" even the belief is true and justification is offered
"Justified true belief" may not be sufficient for "knowledge"
There have been various proposals to modify the above definition of "knowledge" as "justified true belief" sometimes (roughly) along the lines of:

Revised Definition: "Justified true belief… the evidence for which cannot be overturned"

Knowledge

Knowledge: Justified true belief… the evidence for which cannot be overturned

When do we think that the evidence for a causal relationship between an exposure and an outcome cannot be overturned?

- Multiple studies, multiple investigators
- Large effect, large sample size, small standard error
- Robust in sensitivity analysis
- Strong study designs (e.g. RCTs)
- More than one form of evidence (e.g. regression and ITS)
- Understanding of the mechanisms

Knowledge and Consensus

Sometimes we eventually take something as impossible to overturn (e.g. evidence for an effect of smoking on lung cancer); we have knowledge

Often consensus is used to assess what constitutes knowledge (e.g. sometimes this is reflected in what makes it into a textbook)

If everyone agrees (and the evidence has been examined from multiple perspectives) it is likely more difficult to overturn evidence

Consensus is fallible (everyone might be wrong, or consensus may be achieved without examining multiple perspectives) but it is perhaps one helpful guide to whether we have achieved knowledge

Knowledge and Consensus

We might distinguish between *actual knowledge* (dependent upon the evidence, our understanding of it, and its actual relation to truth) and what is *taken as knowledge* (dependent upon consensus)

If science is working well, the evidence is what shapes consensus
In that case, then consensus may end up being a reasonable criterion by which to assess knowledge, but we must be cautious…

For consensus to be a reasonable criterion we need that consensus to be shaped by evidence, not coercion
This points to the importance of academic freedom/free expression; bad ideas should be refuted rather than suppressed; this strengthens our knowledge

Moreover, a focus on evidence, and accurately and objectively assessing it, will generally more easily bring about greater consensus
We can assess evidence *within a study* and we can do this *across studies*

Single Study - Virtues for Objectivity and Consensus

Gelman and Hennig (2017) describe “virtues” for attaining objectivity and hence consensus:

- Transparency: Clear description of protocols, assumptions, etc.
- Common method: Use standard approaches and rules when reasonable
- Impartiality: Consideration of relevant alternatives, evaluation of bias, openness to criticism
- Correspondence to observable reality: making appropriate use of data, reproducibility
- Stability: Robustness to alternative decision/approaches

All of these help attain consensus in knowledge in a community
Of these, if we want “evidence that cannot be overturned” then arguably stability is central (use of sensitivity analysis)

[Gelman and Hennig (2017) also mention “awareness of multiple perspective and context dependence” but arguably what is important is how these are handled]

Virtues for Objectivity and Consensus

We might distinguish between the evidence available *from the data* versus that *from the published study*

Evidence from the published study: How convincing is it that the investigator made right use of the data?

Evidence from the data: Would any reasonable method or approach or way of viewing the data lead to the same conclusion?

Transparency, common method, impartiality, correspondence arguably all pertain principally to the evidence from the published study

Stability is arguably central when we consider the evidence available from the data

The goal of a *published paper* should be to present the evidence so that the evidence from the study corresponds as best as possible to the evidence *from the data*

Stability of Evidence

Stability can be assessed within a study by:

- Robustness to statistical method / modeling
- Robustness to unmeasured confounding / measurement error / missing data / assumptions

This has been much of the focus of the course

We want each of our studies to be as rigorous as possible
We want to be honest about the extent of the evidence
In some cases, the evidence may be substantial
The paper may be transparent, impartial, using established methods and the results stable across perspectives

Can we stop there…?

Replication and Stability

While evidence from a study can be strong, we can also assess the stability of evidence across studies:

- Results should be stable across investigators
- We hope results are at least somewhat stable to the protocol, the time, at least to some extent to the population

This is the topic of replication; this is what we need if we are to have multiple studies contributing to the evidence

Terminology

Reproducibility: The extent to which other investigators can obtain the same results using the data and analysis methods or software used in the original studies (this concerns evidence from a given study and the stability of results)

Replicability: The repetition of a study by applying the same design to the same type of subjects but conducted by a different set of investigators ("replication" if similar results are obtained)

Note: Perfect replication of design is a theoretical construct
Replication is always imperfect; the time is at least different

Consistency (Hill, 1965): a consistent finding is one reported across multiple populations, over time, and using different study designs

"Conceptual Replication": Replication intentionally changing the population, design, intervention or protocol

Replication

Results should be stable across investigators
If we re-run the study we should find something similar

Replication is important both for:

- Assessing the accumulation of evidence
- Assessing whether we can trust the results of a single study

In recent years there have been very large attempts to examine whether results do replicate across numerous studies in several disciplines

Unfortunately, replication has not always been as substantial as might be hoped…

Replication Crisis

Pharmaceutical companies often conduct in-house validation studies
Prinz et al. (2011) surveyed 23 heads of laboratories
They reported only 20-25% of validation studies of published results related to drug target development replicated the original

Preclinical Cancer Research (Begley and Ellis, 2012)
Amgen could replicate only 6 of 53 (11%) “landmark” pre-clinical studies in cancer biology

Preclinical cancer studies often have small sample sizes
But arguably then there is even more need for care

Methods in these two papers are only vaguely described

Replication Crisis

Open Science Collaboration (2015) carried out replications of 100 studies in psychology in 3 leading psychology journals

- Only 47% of replication CI's contained the original effect size
- 97% of original studies had "statistically significant" results; only 36% of replications did
- Effect sizes in replication were 1/2 of that of original

- "Replication" was defined a statistically significant result in the same direction of the original (we will return below to this definition of replication)
- Replication was poorer for social psychology topics (25% p-value) than cognitive psychology topics (50% p-values)
- Replication was poorer for interactions (22%) than main effects (47%)
- Replication was poorer for $0.02<p<0.04$ in the original (18%) than for $p<0.001$ in the original (63%)
- Original studies with larger effects were more likely to have $p<0.05$ in replication but also more likely to have larger difference in effect size

Conclusion: <u>we should not place too much trust in an individual study</u>

Critiques

Gilbert et al. (2016) comment…

(1) Often there were different populations in the replication or design
   - ✧ Attitude toward African Americans in US (vs. Italy in replication)
   - ✧ Imaging consequences of military service (vs. honeymoon)

(2) When multiple replication studies are done, results look much better
   - ✧ With 36 replications per study, the original study CI contains combined estimates 85% of the time

(3) When restricting to studies for which original authors approved replication design, replication goes up to 59.7% (vs. 15.4% without approval)

Perhaps the picture is not quite so bleak

But even with author-approved replication, p-value “replication rates” are at 59.7% so we should not draw definitive conclusions in any individual study

Replication Crisis

Camerer et al. (2016) report results from replication studies of all 18 studies of laboratory experiments in economics between 2011-2014 in American Economic Review and Quarterly Journal of Economics

All replications studies powered at 90% (*if* original effect sizes are true)

- Significant effects in same direction in 11 of 18 (61%)
- 95% CI of replication includes original result in 12 of 18 (67%)
- And 15 of 18 (83%) lie within a 95% prediction interval
- Effect sizes were 66% in replication as original
- Original p-value is correlated with "replication success"
- Original sample size is even more strongly correlated

Looks somewhat better than psychology but…

- restricted to "statistically significant" experiments
- only main effects, no interactions
- using similar inclusion criteria gives more similar results, though economics still performs a bit better

Should we be Surprised?

We should not expect all results to “replicate” in the sense of “statistically significant” p-values with effects the same direction as the original

- There is of course statistical variability in estimates; we can have $p<0.05$ when there is no effect, or $p>0.05$ when there is a true effect
- One can demonstrate that even in studies with perfect repetition of the original protocol we would not expect 95% p-value replication (e.g. only 63% in economics and 78% in psychology projects)
- Many existing metrics are flawed if used naively
- Only for the prediction intervals would one expect 95% (and actual proportions are 83% in economics and 76% in psychology)
- This is certainly still not perfect but not quite as bad
- With numerous replications of the original, one can use yet better metrics e.g. quantile in the replication effect size distribution of the original effect size, or estimate proportion of replication effects in the same direction as the original (Mathur and VanderWeele, 2020)

Evidence and Knowledge

Furthermore, various quantitative analyses suggest that, in the published literature, there will be considerably more than 5% false positives

- The use of the $p<0.05$ criteria for publication will create selective reporting in results
- The problem is compounded by cognitive biases of investigators, by decisions about methods ("investigator degrees of freedom", Simmons et al., 2011) by coding errors etc.

Much of the emphasis has been on whether the results of the original study "replicate"

But is assessing the compatibility of the original with subsequent replications really what is most important…?

Perhaps more useful would be the positive predictive value of a study

Why Most Published Research Findings are False

Table 4. PPV of Research Findings for Various Combinations of Power (1 – β), Ratio of True to Not-True Relationships *(R)*, and Bias *(u)*

1 – β	*R*	*u*	Practical Example	PPV
0.80	1:1	0.10	Adequately powered RCT with little bias and 1:1 pre-study odds	0.85
0.95	2:1	0.30	Confirmatory meta-analysis of good-quality RCTs	0.85
0.80	1:3	0.40	Meta-analysis of small inconclusive studies	0.41
0.20	1:5	0.20	Underpowered, but well-performed phase I/II RCT	0.23
0.20	1:5	0.80	Underpowered, poorly performed phase I/II RCT	0.17
0.80	1:10	0.30	Adequately powered exploratory epidemiological study	0.20
0.20	1:10	0.30	Underpowered exploratory epidemiological study	0.12
0.20	1:1,000	0.80	Discovery-oriented exploratory research with massive testing	0.0010
0.20	1:1,000	0.20	As in previous example, but with more limited bias (more standardized)	0.0015

u = proportion of results that are “research findings” but would not have been published if it had not been because of bias (Ioannidis, 2005)

Evidence and Knowledge

Positive Predictive Values

- The conclusion that "most published research findings are false" is based on relatively strong assumptions
- Other analyses and/or assumptions suggest the PPV of at least RCTs may be closer to 75% or 80% or higher
- But still this is not great…
- It is hardly "evidence that cannot be overturned"

Implications of Replication Crisis:

- We should not place too much trust in an individual study
- We want to improve PPV, and the quality of each study
- The replication crisis makes clear we have a ways to go in that regard
- But for "evidence that cannot be overturned" we will usually need evidence from multiple strong studies
- The "replication crisis" is (somewhat) less problematic if we do not place too much trust in a single study, and accept that evidence accumulate across studies

Evidence and Knowledge

Arguably the focus should instead be on **progression towards knowledge**

(1) If results differ, it is not necessarily the case the original study is wrong; replication may be flawed
(2) If results do replicate, it does not guarantee the conclusion is correct; both may be flawed or subject to the same biases
(3) If results are replicated repeatedly across many studies this continues to add to evidence, but we don't just want quantity but quality also!
(4) Even with multiple replications, we still must take design, and potential biases, into account
(5) Even if biases can be ruled out, we still have only accumulated evidence for a single population

Replication is important but only part of the process of the accumulation of evidence and knowledge

Ideally, we want to improve the quality of each individual study, avoid publication biases, avoid other biases, **and also** run multiple studies

Failure to replicate challenges the quality and accumulation of evidence

Reasons for Failure to Replicate

(1) Statistical Variation
(2) Selective Reporting
    a) Selective reporting from statistical models
    b) Selective reporting from results
    c) Selective reporting in author publication choice
    d) Selective reporting in journal/reviewer/editor publication choice
(3) Unintentional Errors
    a) Errors of coding or collection or entry of data
    b) Errors in the analysis of data
(4) Unintentional Cognitive Biases
    a) Confirmatory biases in data collection / study administration
    b) Confirmatory biases in analysis
(5) Design and Population Considerations
    a) Different population from the original study
    b) Different design than the original study
(6) Fabrication
    a) Fabrication of data
    b) Fabrication of results

Solutions

(1) Reproducible Research

(2) Change in publication practice in journals

- Move away from hypothesis testing framework
- Blinding to results in review
- Requiring replication
- Requiring reproducibility

(3) Pre-registration of trials (observational studies?)

(4) Outcome-wide studies

(5) Greater incentives for replication

(6) Abandon p=0.05 hypothesis testing

(1) Reproducibility

Reproducibility: The extent to which other investigators can obtain the same results using the data and analysis methods or software used in the original studies

Requiring posting of data and code ("transparency") helps…

- detect errors in the original data or its management
- identify mistake in analysis
- identifying instances of unintentional coding error
- sometimes helps detect fabrication
- makes clear when analysis choice drives results
- ensures greater care

This will in general be helpful

But similar problems can also arise in the reproduced research

e.g. searching for the one choice for which results do differ

Different teams with competing hypotheses could analyze same data

Encouraging Reproducibility

Some journals now request or even require that authors submit both data and software analysis

As above, this helps with many aspects of research

But…

- extra burden for researchers
- disincentivizes data collection
- concerns about data privacy
- concerns about intellectual property in statistical computing

Open Science Framework provides a way to make this easy
Center for Open Science has been providing resources and platforms

Reproducibility helps with…

(1) Statistical Variation
(2) Selective Reporting
    a) Selective reporting from statistical models
    b) Selective reporting from results
    c) Selective reporting in author publication choice
    d) Selective reporting in journal/reviewer/editor publication choice
(3) Unintentional Errors
    a) Errors of coding or collection or entry of data
    b) Errors in the analysis of data
(4) Unintentional Cognitive Biases
    a) Confirmatory biases in data collection / study administration
    b) Confirmatory biases in analysis
(5) Design and Population Considerations
    a) Different population from the original study
    b) Different design than the original study
(6) Fabrication
    a) Fabrication of data
    b) Fabrication of results

(2) Journal Practices

Make decisions on publication based on design (blinded to results)
- a large well designed study is informative even if the results are null
- would eliminate reviewer significance-testing selection

Even when unblinded to results, reviewers, and editors could attempt to apply similar principles

Some disciplines (e.g. genetics) require replication for top journals

Some journals now require reproducibility (posting of code/data)

Journals could provide incentives e.g. allow greater word count for studies with replication and/or reproducibility

Journals could help with...

(1) Statistical Variation
(2) Selective Reporting
    a) Selective reporting from statistical models
    b) Selective reporting from results
    c) Selective reporting in author publication choice
    d) Selective reporting in journal/reviewer/editor publication choice
(3) Unintentional Errors
    a) Errors of coding or collection or entry of data
    b) Errors in the analysis of data
(4) Unintentional Cognitive Biases
    a) Confirmatory biases in data collection / study administration
    b) Confirmatory biases in analysis
(5) Design and Population Considerations
    a) Different population from the original study
    b) Different design than the original study
(6) Fabrication
    a) Fabrication of data
    b) Fabrication of results

Plus the advantages of reproducibility (and replication)

(3) Pre-Registration of Clinical Trials

2004: The International Committee of Medical Journal Editors announced a policy requiring the pre-registration of clinical trial protocols in a publicly available repository as a mandatory prerequisite for publication of the trial's results in their journals

- Counter the suppression of results
- Ensure more honest reporting of results
- Avoid potential for new trial to randomize treatment already shown to be ineffective

Now relatively standard practice in clinical trials

Pre-Registration in Observational Studies

2009: European Center for Ecotoxicology and Toxicology of Chemicals proposed same pre-registration for non-randomized studies

- principally as a safe guard against “fishing” and selective reporting of results

Vigorous debate followed

- questions as whether pre-specified hypotheses are more plausible
- quality of design/analysis may be high without pre-specification
- could discourage creativity and discovery
- may inappropriately down-weight strong evidence simply because it did not conform to “regulations”

Perhaps don’t need to require but could be encouraged (e.g. offer incentives)

- additional advantage of more careful study planning
- some of the advantages depend on if data collection is prospective vs. secondary

Pre-Registration in Observational Studies

Intermediate Position:
In 2014, a self-appointed committee promulgated new guidelines called the “Transparency and Openness Promotion Guidelines,” which proposed eight standards that could be implemented:

1) Citation Standards
2) Data Transparency
3) Analytic Methods (Code) Transparency
4) Research Materials Transparency
5) Design and Analysis Transparency
6) Study Preregistration
7) Analysis Plan Preregistration
8) Replication

Each assessed as Levels 0, I, II, or III
Reasonable but arguably still should not be imposed (Lash, 2022)

Pre-registration helps with...

(1) Statistical Variation
(2) Selective Reporting
    a) Selective reporting from statistical models
    b) Selective reporting from results
    c) Selective reporting in author publication choice
    d) Selective reporting in journal/reviewer/editor publication choice
(3) Unintentional Errors
    a) Errors of coding or collection or entry of data
    b) Errors in the analysis of data
(4) Unintentional Cognitive Biases
    a) Confirmatory biases in data collection / study administration
    b) Confirmatory biases in analysis
(5) Design and Population Considerations
    a) Different population from the original study
    b) Different design than the original study
(6) Fabrication
    a) Fabrication of data
    b) Fabrication of results

(4) Outcome-Wide Analyses

For assessing the public health importance of an exposure it would be good to examine many outcomes simultaneously not just one

Move towards “outcome-wide epidemiology” (VanderWeele, 2017, 2020)

- Allows for more rapid expansion of knowledge
- Helps identify exposures with some positive and some negative effects
- Allows for easier publication of null results

This would also assist with:

- Selective reporting of results
- Selective reporting from statistical models
  - With many outcomes, one uses standardized procedures; and it is more difficult to select models across all outcomes
- Authors, reviewers, editors more likely to publish the null results too

We can go on a “fishing expedition”, and then report it
Such practices are arguably not unreasonable and are effectively what is done over time in the literature, over multiple papers, anyway
Such studies could report both regular p-values and p-values corrected for multiple tests (note: results over multiple papers generally are not corrected)

Outcome-Wide Studies helps with…

(1) Statistical Variation
(2) Selective Reporting
    a) Selective reporting from statistical models
    b) Selective reporting from results
    c) Selective reporting in author publication choice
    d) Selective reporting in journal/reviewer/editor publication choice
(3) Unintentional Errors
    a) Errors of coding or collection or entry of data
    b) Errors in the analysis of data
(4) Unintentional Cognitive Biases
    a) Confirmatory biases in data collection / study administration
    b) Confirmatory biases in analysis
(5) Design and Population Considerations
    a) Different population from the original study
    b) Different design than the original study
(6) Fabrication
    a) Fabrication of data
    b) Fabrication of results

(5) Incentives to Replicate

There have been few incentives to replicate studies:

- Public research funding / journals have emphasized innovation
- Investigators thus pursue new research hypotheses over replication of studies with previously studied hypotheses

With increasing recognition of the importance of replication, this is beginning to change

- Media prominence of failure to replicate studies
- Some journals now require replication of research results before publication (genetics)
- New journals focusing on replication
- Funding still lags behind

Replication of course can be very costly

Decisions to fund replication need to be based on the <u>importance of the results</u>

Replication helps with...

(1) Statistical Variation
(2) Selective Reporting
    a) Selective reporting from statistical models
    b) Selective reporting from results
    c) Selective reporting in author publication choice
    d) Selective reporting in journal/reviewer/editor publication choice
(3) Unintentional Errors
    a) Errors of coding or collection or entry of data
    b) Errors in the analysis of data
(4) Unintentional Cognitive Biases
    a) Confirmatory biases in data collection / study administration
    b) Confirmatory biases in analysis
(5) Design and Population Considerations
    a) Different population from the original study
    b) Different design than the original study
(6) Fabrication
    a) Fabrication of data
    b) Fabrication of results

(6) Abandon p=0.05 Hypothesis Testing

Using p=0.05 as a cut-off can lead to poor reasoning
There is nothing wrong with the p-value, only the 0.05 cut-off

There are numerous major problems with using this as the cut-off

- Considering p>0.05 as “no difference”
- Suppression of null results
- Publication bias in the literature (problematic in meta-analyses)
- Encourages placing too much trust in a single study
- Conditional on p<0.05, the effect estimate is likely exaggerated

We should carefully use the p-value and other statistics, and aspects of design and robustness to evaluate evidence
But using the p=0.05 cut-off as absolute is problematic and should be abandoned (Amrhein, Nature 2019)

If we abandon p=0.05 we are more likely not to view replication as a crisis but as an opportunity to synthesize evidence

Examples of Appropriate Reporting

Estimate Large, Wide CI Covering the Null: “The estimate was substantial (HR=1.6, p=0.12), but there is considerable uncertainty (95% CI: 0.88, 2.88) not allowing for definitive conclusions”

Estimate Large, Wide CI Not Covering the Null: “There was some evidence for an association with a substantial effect estimate (HR=1.6, p=0.04), but with considerable uncertainty (95% CI: 1.01, 2.60)”

Estimate Large, Narrow CI Not Covering the Null: “There was evidence for an association with a substantial effect estimate (HR=1.6, p<0.001); the entire confidence interval suggested fairly substantial associations (95% CI: 1.45, 1.76)”

Estimate Small, Wide CI Covering the Null: “The estimate was small (OR=1.1, p=0.63), and there was considerable uncertainty (95% CI: 0.63, 1.62), not allowing for definitive conclusions”

Estimate Small, Narrow CI Covering the Null: “The estimate was small (OR=1.1, p=0.11), and there was some uncertainty in the estimate and even its direction, but the entire confidence intervals indicates relatively small associations (95% CI: 0.97, 1.23)”

Estimate Small, Narrow CI Not Covering the Null: “There was some evidence for an association (OR=1.1, p<0.001), but the entire confidence intervals indicates relatively small associations (95% CI: 1.07, 1.13)

Different Designs and Populations

None of the “solutions” really directly addresses:

(5) Design and Population Considerations

a) Different population from the original study
b) Different design than the original study

If we change the population, the results may be different
If we change the intervention, the results may be different

Arguably, however, we want interventions that are applicable fairly broadly (not just some carefully selected sub-population) and that are not overly sensitive to slight changes in the intervention itself

We often want <u>some replication across populations and designs</u>

Though even this will be relative to subject-matter (cf. Klein et al,. 2014)
e.g. universal patterns of perception and cognition vs. culturally specific responses (racism, immigration, nationalism, etc.)

Humility with Individual Studies

We need replication **and** better research; with better practices, we can also attain higher replication but if we want…

Knowledge: Justified true belief… the evidence for which cannot be overturned

Any single study is unlikely to provide this

- Evidence can be quite strong, but certainly data collection/entry/coding errors are even then always possible
- Selective reporting is almost certainly always a major issue with results, authors, and journals
- We should always expect some results not to replicate
- Evidence and knowledge accumulate gradually; we need multiple studies, systematic review, meta-analysis etc.
- We have been too ready in the academy and in the media to give too much weight to a single study

The "replication crisis" should make clear that we need to be more cautious with individual studies

Discussions should move **from research to knowledge**

The Way Forward

If our aim is knowledge, we can…

- Make each study and analysis as rigorous as possible
- Whenever possible practice transparency and reproducibility
- Carry out sensitivity analysis and be honest about the results
- Don't over-interpret the results of a single study
- Certainly don't take p=0.05 as the indication of truth
- Don't despair if a particular study "fails to replicate"
- Think about the entire body of evidence across papers
- Focus on meta-analyses and systematic review and restrict these to more rigorous studies
- Examine multiple forms of evidence
- Clarify the context(s) to which the evidence pertains e.g. which populations, ages, intervention types, etc.

Doing so will bring us closer to knowledge
Doing so will make it clearer what further evidence may be needed

Conclusions

(1) Scientific knowledge requires evidence that cannot be overturned

(2) Sensitivity analysis is arguably central in this task, as is replication

(3) Many study results have not replicated; this should lead to greater humility in interpreting any given single study

(4) The ideal of knowledge from a single study is generally unrealistic and problematic

(5) Reproducibility, changes in journal practices, pre-registration, outcome-wide studies, and more replication studies would all help more quickly obtain reliable evidence

(6) Focus on knowledge, and good scientific practices, is essential for correct decision-making in public health

(7) Our aim should be knowledge and contributions to health, with publication principally as a means to those ends