From one environment to many: The problem of replicability of statistical inferences

Michael J. Higgins

Work with James J. Higgins and Jinguang Lin

Kansas State University

April 25, 2019

Reproducibility crisis: Examples

- Recent problem that many scientific findings cannot be replicated.

Reproducibility crisis: Examples

- Recent problem that many scientific findings cannot be replicated.
- Problem exists in nearly all fields of science: **Reproducibility crisis**.

Reproducibility crisis: Examples

- Recent problem that many scientific findings cannot be replicated.
- Problem exists in nearly all fields of science: **Reproducibility crisis**.
- A 2016 survey of *Nature* readers found that **about 70%** of scientists have failed to reproduce other researchers' experiments, and **more than 50%** have failed to reproduce their own studies.

Reproducibility crisis: Examples

- Recent problem that many scientific findings cannot be replicated.
- Problem exists in nearly all fields of science: **Reproducibility crisis**.
- A 2016 survey of *Nature* readers found that **about 70%** of scientists have failed to reproduce other researchers' experiments, and **more than 50%** have failed to reproduce their own studies.
- In 2015, a study in Science had researchers conduct replications of 100 experimental and correlational studies published in three psychology journals using original materials.

  Less than 50% of the original findings were replicable.

Reproducibility crisis: Examples

- Prinz et. al (2019) examined the reproducibility of results of 67 medical projects using in-house data.

  They found that **65% of results** were not consistent with the original studies.

Reproducibility crisis: Examples

- Prinz et. al (2019) examined the reproducibility of results of 67 medical projects using in-house data.

  They found that **65% of results** were not consistent with the original studies.

- The US Federal Reserve Board tried to replicate finding from 67 papers published in 13 well-regarded economics journals.

  More than half of the papers were irreproducible despite using author-provided data and code.

Reproducibility crisis: Response

- Concerns about replicability has lead to a large backlash against terms like p-values and statistical significance.

Reproducibility crisis: Response

- Concerns about replicability has lead to a large backlash against terms like *p*-values and statistical significance.
- In 2015, the journal *Basic and Applied Social Psychology* banned the use of *p*-values and claims of statistical significance.

Reproducibility crisis: Response

- Concerns about replicability has lead to a large backlash against terms like *p*-values and statistical significance.
- In 2015, the journal *Basic and Applied Social Psychology* banned the use of *p*-values and claims of statistical significance.
- A 2019 special issue of *The American Statistican* recommended that the use of the phrase "statistically significant" be abandoned completely.

Reproducibility crisis: Definition

- Additionally, the reproducibility crisis has lead to many different different of what it means to replicate a result.

Reproducibility crisis: Definition

- Additionally, the reproducibility crisis has lead to many different different of what it means to replicate a result.
- Use of nearly–synonymous notions of replicability including: reproducibility, reliability, robustness, and generalizability.

Reproducibility crisis: Definition

- Additionally, the reproducibility crisis has lead to many different different of what it means to replicate a result.
- Use of nearly–synonymous notions of replicability including: reproducibility, reliability, robustness, and generalizability.
- We are most concerned with the idea of *replicability*—"the ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected."

Reproducibility crisis: Causes

- What are the major culprits of the reproducibility crisis?

Reproducibility crisis: Causes

- What are the major culprits of the reproducibility crisis?
- Most often cited problem: "Statistically significant" findings are more likely to be published than insignificant ones.

Reproducibility crisis: Causes

- What are the major culprits of the reproducibility crisis?
- Most often cited problem: "Statistically significant" findings are more likely to be published than insignificant ones.
- Large incentive for researchers to "tweak" results until findings are significant.

Reproducibility crisis: Causes

- What are the major culprits of the reproducibility crisis?
- Most often cited problem: "Statistically significant" findings are more likely to be published than insignificant ones.
- Large incentive for researchers to "tweak" results until findings are significant.
- Little consequence if findings are not replicable by other researchers.

Reproducibility crisis: Causes

- *p*-hacking: manipulating and dredging through data until a significant result is found.

  E.g.: No significant average treatment effect, but results suggest that treatment is effective on left-handed, blonde, 38 year olds.

Reproducibility crisis: Causes

- *p*-hacking: manipulating and dredging through data until a significant result is found.

  E.g.: No significant average treatment effect, but results suggest that treatment is effective on left-handed, blonde, 38 year olds.

- Overstatement of power.

  E.g.: Performing a cluster randomized trial, but analyzing experiments assuming treatment was given to observational units.

Reproducibility crisis: Causes

- *p*-hacking: manipulating and dredging through data until a significant result is found.

  E.g.: No significant average treatment effect, but results suggest that treatment is effective on left-handed, blonde, 38 year olds.

- Overstatement of power.

  E.g.: Performing a cluster randomized trial, but analyzing experiments assuming treatment was given to observational units.

- Inaccurate protocol/statistical code described to obtain result.

  E.g.: Missing details on how missing data were imputed, whether observations were eliminated, if/how data were transformed, etc.

Environment by treatment interaction

- Most of the aforementioned issues pointed at researcher ignorance, honest mistakes, or deliberate fraud as primary culprits of the reproducibility crisis: **researcher's fault**.

Environment by treatment interaction

- Most of the aforementioned issues pointed at researcher ignorance, honest mistakes, or deliberate fraud as primary culprits of the reproducibility crisis: **researcher's fault**.
- **Our focus**: Environment by treatment interaction.

Environment by treatment interaction

- Most of the aforementioned issues pointed at researcher ignorance, honest mistakes, or deliberate fraud as primary culprits of the reproducibility crisis: **researcher's fault**.
- **Our focus**: Environment by treatment interaction.
- Statistical inferences that are drawn from the analysis of data apply only to the environment in which the study is conducted. Often, researchers want their result to apply more broadly.

Environment by treatment interaction

- Most of the aforementioned issues pointed at researcher ignorance, honest mistakes, or deliberate fraud as primary culprits of the reproducibility crisis: **researcher's fault**.
- **Our focus**: Environment by treatment interaction.
- Statistical inferences that are drawn from the analysis of data apply only to the environment in which the study is conducted. Often, researchers want their result to apply more broadly.
- We show that the presence of differing treatment effects across environments can dramatically decrease the likelihood of replicating results from a study **even if both the initial and follow-up studies follow best statistical practices**.

Environment by treatment interaction

- Most of the aforementioned issues pointed at researcher ignorance, honest mistakes, or deliberate fraud as primary culprits of the reproducibility crisis: **researcher's fault**.
- **Our focus**: Environment by treatment interaction.
- Statistical inferences that are drawn from the analysis of data apply only to the environment in which the study is conducted. Often, researchers want their result to apply more broadly.
- We show that the presence of differing treatment effects across environments can dramatically decrease the likelihood of replicating results from a study **even if both the initial and follow-up studies follow best statistical practices**.
- In the worst case: replicating a study is **equivalent to flipping a fair coin**.

Goal of talk

- How can we assess how environment by treatment interaction can impact replicability of a result? If researchers are aware of this interaction, how can inferences be adjusted?

Goal of talk

- How can we assess how environment by treatment interaction can impact replicability of a result? If researchers are aware of this interaction, how can inferences be adjusted?
- **Measures of replicability**:
  1. Probability of replicability
  2. Adjusted p-value
  3. Adjusted confidence interval

Goal of talk

- How can we assess how environment by treatment interaction can impact replicability of a result? If researchers are aware of this interaction, how can inferences be adjusted?
- **Measures of replicability**:
  1. Probability of replicability
  2. Adjusted p-value
  3. Adjusted confidence interval
- Methods for assessing the magnitude of environmental by treatment interaction.

Model

Observations from the initial study follow **Model 1**:

$$Y_{ij} = \mu_i + \epsilon_{ij}, \quad i = 1, 2, \quad j = 1, ..., n \tag{1}$$

Observations from the follow-up study follow **Model 2**:

$$Y_{ij} = \mu_i + \theta + \delta_i + \epsilon_{ij}, \quad i = 1, 2, \quad j = 1, ..., n \tag{2}$$

- μ_i: mean for treatment i.
- ϵ_{ij}: i.i.d. $N(0, \sigma_e^2)$: Random error terms.
- θ: $N(0, \sigma_\theta^2)$: Random environment effect affecting both treatments.
- δ_i: i.i.d. $N(0, \sigma_B^2)$: Random environment effect affecting only treatment i.

Model

Observations from the initial study follow **Model 1**:

$$Y_{ij} = \mu_i + \epsilon_{ij}, \quad i = 1, 2, \quad j = 1, ..., n \tag{1}$$

Observations from the follow-up study follow **Model 2**:

$$Y_{ij} = \mu_i + \theta + \delta_i + \epsilon_{ij}, \quad i = 1, 2, \quad j = 1, ..., n \tag{2}$$

- μ_i: mean for treatment i.
- ϵ_{ij}: i.i.d. $N(0, \sigma_e^2)$: Random error terms.
- θ: $N(0, \sigma_\theta^2)$: Random environment effect affecting both treatments.
- δ_i: i.i.d. $N(0, \sigma_B^2)$: Random environment effect affecting only treatment i.
- Environment: Includes factors such as location/facility, time, personnel, equipment. Certain aspects may be uncontrollable.

Model

Observations from the follow-up study follow **Model 2**:

$$Y_{ij} = \mu_i + \theta + \delta_i + \epsilon_{ij}, \quad i = 1, 2, \quad j = 1, ..., n$$

Under Model 2, the difference of sample means can be expressed as

$$\overline{Y}_1 - \overline{Y}_2 = (\mu_1 - \mu_2) + (\overline{\epsilon}_1 - \overline{\epsilon}_2) + (\delta_1 - \delta_2)$$

Thus, $\overline{Y}_1 - \overline{Y}_2$ has a normal distribution with mean $\mu_1 - \mu_2$ and variance $2\sigma_e^2/n + 2\sigma_B^2$.

Two-sample t-test

We assume that inferences are performed on $\mu_1 - \mu_2$ using a two-sample t-test.

$$T = \frac{\overline{Y}_1 - \overline{Y}_2}{S_p\sqrt{2/n}} = \frac{(\overline{Y}_1 - \overline{Y}_2)}{\sqrt{2\sigma_e^2/n + 2\sigma_B^2}} \frac{\sqrt{2\sigma_e^2/n + 2\sigma_B^2}}{S_p\sqrt{2/n}}$$
$$= NCT\sqrt{1 + n\sigma_B^2/\sigma_e^2}$$

- S_p: the "pooled" sample standard deviation
- NCT: non-central t-distribution with $df = 2(n-1)$ and non-centrality parameter

$$u = \frac{\mu_1 - \mu_2}{\sqrt{2\sigma_e^2/n + 2\sigma_B^2}} = \frac{\sqrt{n}\Delta}{\sqrt{1 + n\sigma_B^2/\sigma_e^2}}$$

Effect Size, EER

- Δ is the **effect size**:

$$\Delta = (u_1 - u_2)/(\sigma_e\sqrt{2}).$$

- σ_B/σ_e is the **environmental effect ratio (EER)**: The ratio of the standard deviation between environments within treatment (or interaction) and the standard deviation of experimental error.

Effect Size, EER

- Δ is the **effect size**:

$$\Delta = (u_1 - u_2)/(\sigma_e\sqrt{2}).$$

- σ_B/σ_e is the **environmental effect ratio (EER)**: The ratio of the standard deviation between environments within treatment (or interaction) and the standard deviation of experimental error.
- Δ and the EER are two critical factors in assessing the replicability of a study.

Probability of replicability

- **Probability of replicability**: The probability that the follow-up study yields a significant result, assuming the initial study significant result.

Probability of replicability

- **Probability of replicability**: The probability that the follow-up study yields a significant result, assuming the initial study significant result.
- If $T \geq t_{\alpha/2,df}$ in the initial study, then we have a replicable result if $T \geq t_{\alpha/2,df}$ in the follow-up experiment.

  ($t_{\alpha,df}$ is the $1-\alpha$ quantile of the t-distribution).

Probability of replicability

- **Probability of replicability**: The probability that the follow-up study yields a significant result, assuming the initial study significant result.
- If $T \geq t_{\alpha/2,df}$ in the initial study, then we have a replicable result if $T \geq t_{\alpha/2,df}$ in the follow-up experiment.

  ($t_{\alpha,df}$ is the $1-\alpha$ quantile of the t-distribution).
- If the initial study is significant in the wrong direction—that is, $T \leq -t_{\alpha/2,df}$ in the initial study—then replicability occurs if $T \leq -t_{\alpha/2,df}$ in the follow-up study.

  This is confirmation of an incorrect result.

Probability of replicability

- Thus, for a two-sided test at level of significance α, the probability of reproducibility is approxmiately

$$1 - G_{df,u}\left(t_{\alpha/2,df}/\sqrt{1 + n\sigma_B^2/\sigma_e^2}\right)$$

- $G_{df,u}(t)$: the c.d.f. of the non-central t-distribution, with non-centrality parameter u.

Probability of replicability

- Thus, for a two-sided test at level of significance α, the probability of reproducibility is approxmiately

$$1 - G_{df,u}\left(t_{\alpha/2,df}/\sqrt{1 + n\sigma_B^2/\sigma_e^2}\right)$$

- $G_{df,u}(t)$: the c.d.f. of the non-central t-distribution, with non-centrality parameter u.
- This probability depends on:
  1. n, α, as usual.
  2. Δ: Larger $\Delta \rightarrow$ Larger probability of reproducibility.
  3. EER: Smaller EER $\rightarrow$ Larger probability of reproducibility (if original study has high power).

Example: Probability of replicability

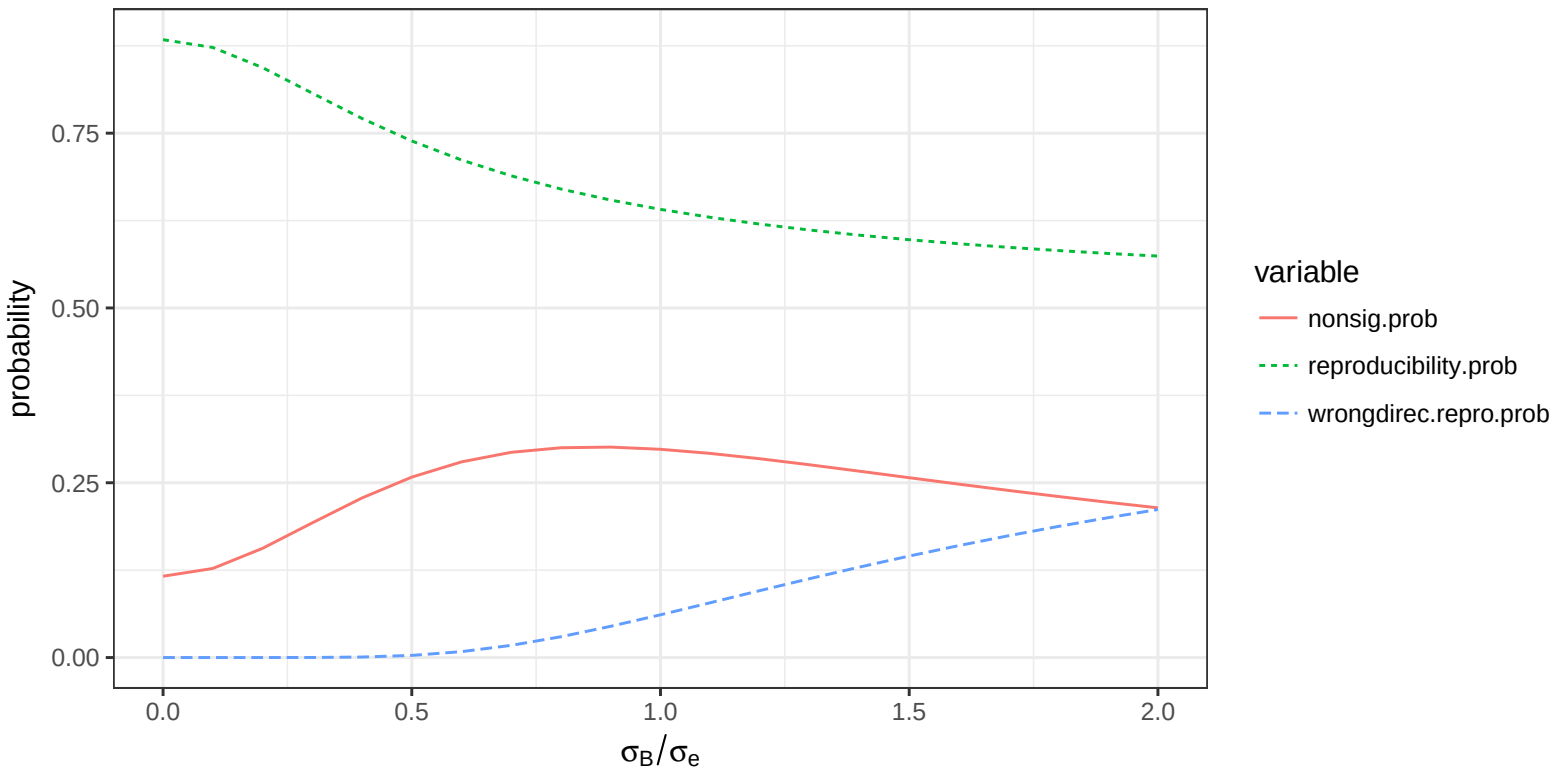

Figure: Probability of replicability, probability of significance in the wrong direction, non-significance, vs. σ_B/σ_e, $n = 11$, Δ=1.02, $\alpha = .05$, initial power = .88

Adjusted p-value

- Consider a hypothesis test of $H_0 : \mu_1 - \mu_2 - 0$ against $H_1 : \mu_1 - \mu_2 \neq 0$.
- The observed effect size is: $\Delta^* = (\overline{y}_1 - \overline{y}_2)/(\sqrt{2}S_p)$.
- The observed t-stat is $T = \dfrac{\overline{y}_1 - \overline{y}_2}{S_p\sqrt{2/n}} = \Delta^*\sqrt{n}$

Adjusted p-value

- Consider a hypothesis test of $H_0 : \mu_1 - \mu_2 - 0$ against $H_1 : \mu_1 - \mu_2 \neq 0$.
- The observed effect size is: $\Delta^* = (\overline{y}_1 - \overline{y}_2)/(\sqrt{2}S_p)$.
- The observed t-stat is $T = \dfrac{\overline{y}_1 - \overline{y}_2}{S_p\sqrt{2/n}} = \Delta^*\sqrt{n}$
- Under the null hypothesis, the non-centrality parameter $u = 0$. The two-sided p-value for the observed t-stat under Model 2 is:

$$2(1 - G_{df,u=0}\left(\Delta^*\sqrt{n}/\sqrt{1 + n\sigma_B^2/\sigma_e^2}\right)$$

- This is called the **adjusted p-value**.

Adjusted p-value

Properties of the adjusted p-value:

- The adjusted p-value is smaller for larger effect sizes Δ and smaller EER.

Adjusted p-value

Properties of the adjusted p-value:

- The adjusted p-value is smaller for larger effect sizes Δ and smaller EER.
- For given values of Δ and EER, it may be impossible to achieve an adjusted p-value < 0.05, regardless of the sample size.

Adjusted p-value

Properties of the adjusted p-value:

- The adjusted p-value is smaller for larger effect sizes Δ and smaller EER.
- For given values of Δ and EER, it may be impossible to achieve an adjusted p-value < 0.05, regardless of the sample size.
- **Large effect size is a better indicator of reproducibility than small p-value.**

Adjusted p-value: Example

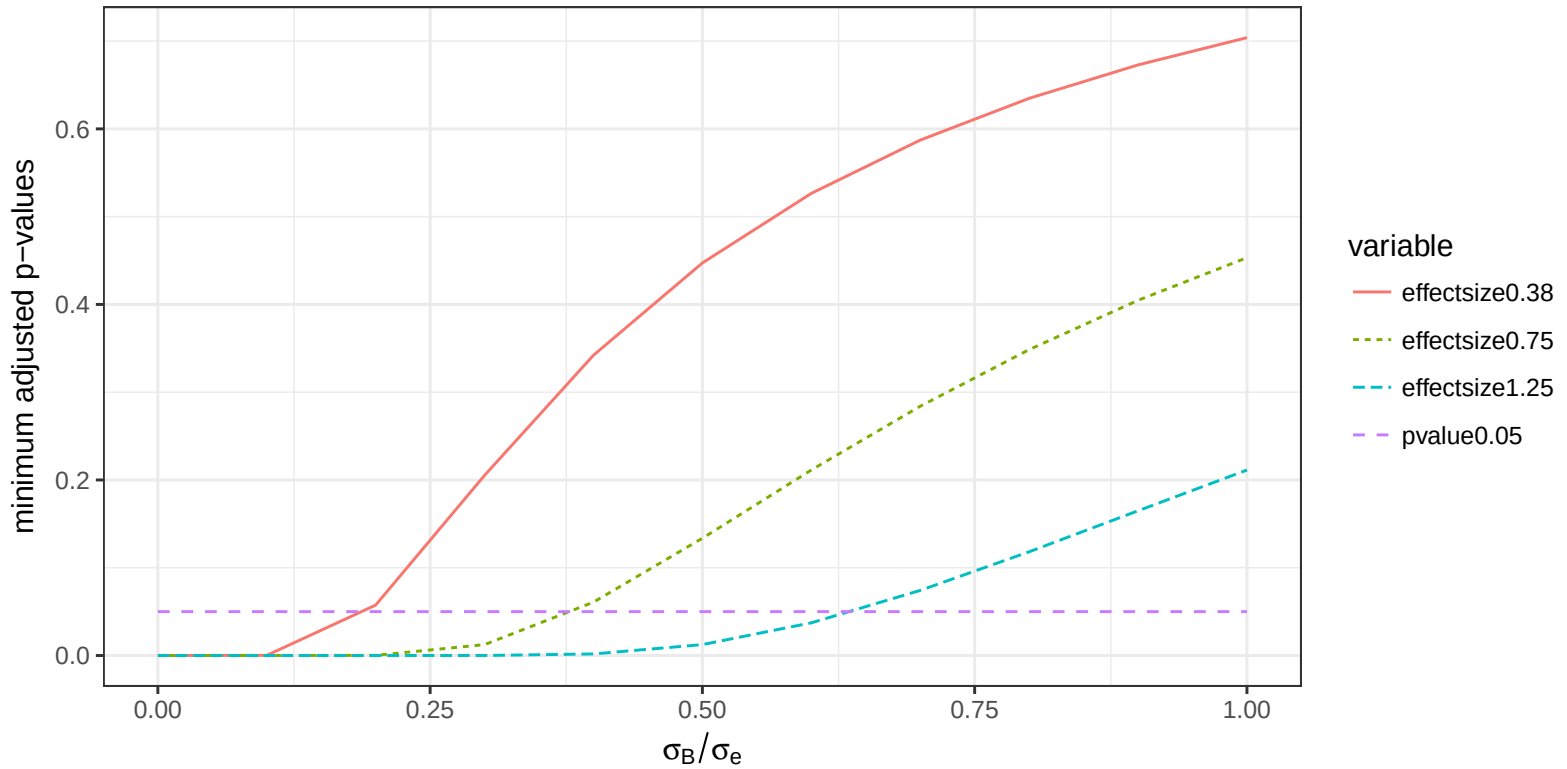

Figure: Adjusted p-values vs. EER for various effect sizes Δ, $n \to \infty$.

Adjusted confidence interval

- The previous hypothesis test can be inverted to form an adjusted confidence interval:

$$(\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2} S_p \sqrt{2/n + 2\sigma_B^2/\sigma_e^2}.$$

Adjusted confidence interval

- The previous hypothesis test can be inverted to form an adjusted confidence interval:

$$(\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2} S_p \sqrt{2/n + 2\sigma_B^2/\sigma_e^2}.$$

- Of note, as $n \to \infty$, the confidence interval length is approximately proportional to the EER $\sigma_B/\sigma_e > 0$.

Adjusted confidence interval

- The previous hypothesis test can be inverted to form an adjusted confidence interval:

$$(\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2} S_p \sqrt{2/n + 2\sigma_B^2/\sigma_e^2}.$$

- Of note, as $n \to \infty$, the confidence interval length is approximately proportional to the EER $\sigma_B/\sigma_e > 0$.
- The larger the EER, the larger the confidence interval, regardless of the sample size.

How to determine EER

- Perform a study at separate labs simultaneously. Can get an estimate of EER directly (though can be quite costly).

How to determine EER

- Perform a study at separate labs simultaneously. Can get an estimate of EER directly (though can be quite costly).
- Historical: A meta-analysis may allow a good estimate of EER for a given type of study.

How to determine EER

- Perform a study at separate labs simultaneously. Can get an estimate of EER directly (though can be quite costly).
- Historical: A meta-analysis may allow a good estimate of EER for a given type of study.
- Sensitivity analysis: Statistical analysis may include values of EER for which the given result will replicate.

Thank you

Thank you!

Paper available at:

https://arxiv.org/abs/1904.10036