In defense of correct use of statistical significance

Michal Abrahamowicz *[1] with

James Carpenter * [3,4] , Victor Kipnis *[2], Marie-Eve Beauchamp [1]

*** "P-values" Project Co-Leaders**

[1] Department of Epidemiology & Biostatistics, McGill University, Montreal, Canada
[2] Biometry Research group, National Cancer Institute, USA
[3] Department of Medical Statistics, London School of Hygiene & Tropical Medicine, UK
[4] MRC clinical trials unit at UCL, Holborn, London, UK

STRATOS INITIATIVE

McGill

Background/Rationale

- In March 2019, in the ***Nature*** Comment "***Retire statistical significance***" **V. Amrhein, S. Greenland & B. McShane(AGM)** [1] recommended "*a stop to the use of P values in the conventional dichotomous way – to decide whether a result refutes or supports a scientific hypothesis*" and concluded: *"… **it's time for statistical significance to go**"*

- **The Comment was endorsed by >800 signatories**, mostly end-users of statistical methods, but also a few dozen statisticians, including a few STRATOS members **

  ** **Sampling properties of signatories selection are UNclear ☺**

- This Comment has created a major confusion among both:
  1. Non-statistical researchers, i.e. End-users (including Editors and Reviewers)
  2. Statisticians who Teach Applied Statistics and/or are involved in Collaborative Research

[1] Amrhein *et al. Nature.* March 2019;567:305-307.

Selected Verbatim Citations from AGM's *Nature* Comment

- In the Opening 4 sentences Amrhein *et al* state:

  "When... you heard a... speaker claim there was 'no difference'... because the difference was 'statistically non-significant'? ... We hope that... someone was perplexed if... ***a plot or table showed there actually was a difference***** *. How do statistics so often lead scientists to deny* ***differences that those not educated in statistics can plainly see*** *?**"*

** AGM do NOT explain what is the Empirical Basis to establish that

"there ***was a difference"*** or to ***"plainly see"*** such differences ?

Potential Concerns about AGM's "Black vs. White" recommendations

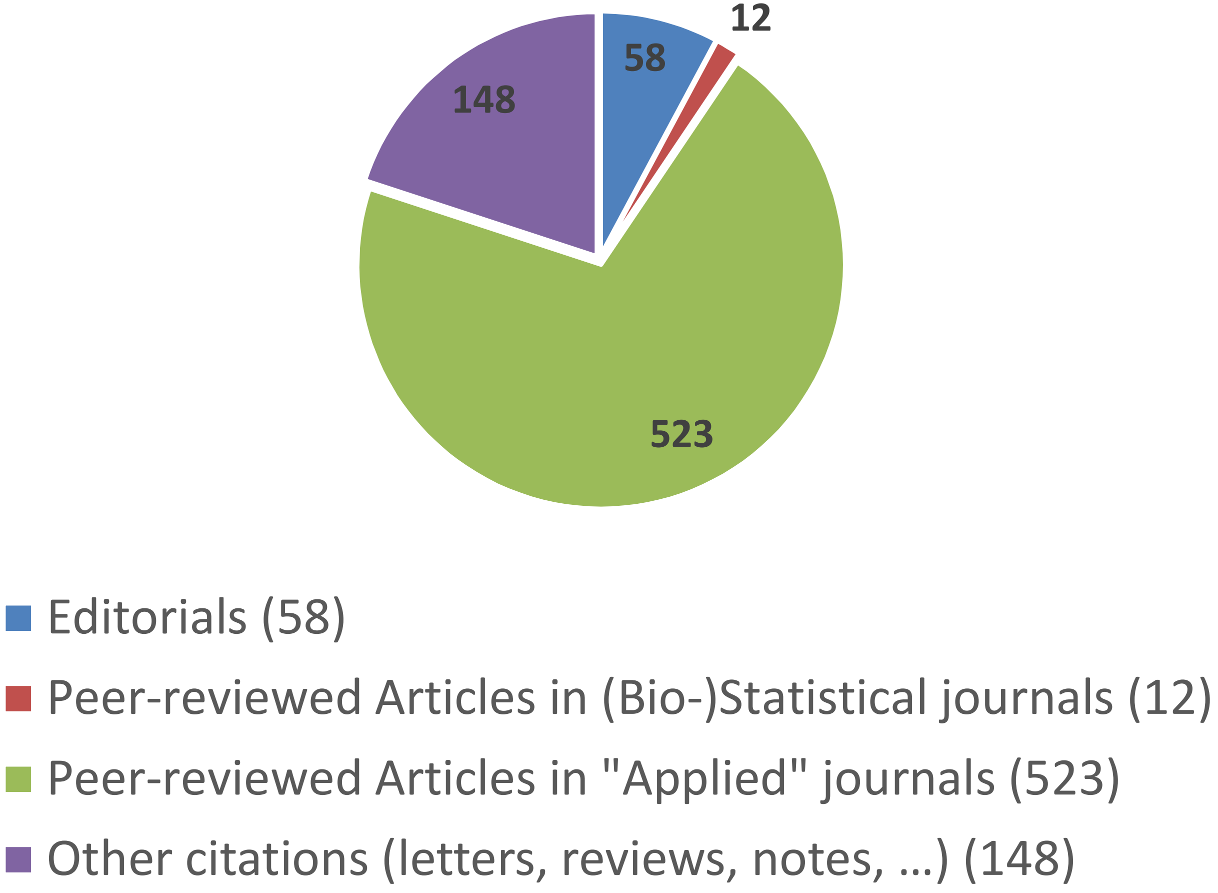
- **Removing the "gatekeeper" of statistical significance may open the floodgates toward an uncontrolled reporting of "associations" that may likely reflect just a combination of (i) sampling errors & (ii) Authors' wishful thinking**

- **Similar concerns** expressed (right after AGM Comment publication) **by other statisticians** [e.g. 2-5]:
  - Julia Haaf: "… *when statistical testing is skipped, ... any differences between observations would be considered meaningful* " [4]
  - **John Ioannidis warns that removal of statistical significance may lead to** "*statistical anarchy*", and "…(reliance) *less on data and evidence and more on subjective opinions and interpretations*" [5]

[2] Ioannidis, *Nature* 2019. [3] Johnson, *Nature* 2019. [4] Haaf et al, *Nature* 2019.
[5] Ioannidis, JAMA 2019. [6] Karl R. Popper, *The Logic of Scientific Discovery* 1959.

Impact of AGM’s *Nature* Comment

741 Citations (Scopus): March 2019 – June 2021



Example of Clinical study citing [AGM]
Thapa *et al*, Cancers (IF=6.7)

- In Methods:

  "Consistent with recommendations..., our analysis focused on effect estimation rather than statistical significance testing [Ref to AGM]'

- Then, in Results, they report effects estimated in different subgroups [10], e.g.: **

  0.149 (95% CI: 0.007, 0.292) for H. Pyl. + *versus*

  0.103 (95% CI: -0.285, +0.490) for H. Pyl. –

  and Naively INTERPRET the 'difference' in Point Estimates:

  "a LARGER increase... was observed for … H. Pyl +…." [10]

- Yet, the observed "DIFFERENCE" may be entirely due to sampling error;

- 0.149 – 0.103 = 0.046 (95% CI: -0.367, 0.459), p = 0.827 !!

- ** Similar issues e.g. in [Ranapurwala *et al*, *Am J Prev Med*] (IF = 4.5) [12]

[1] Amrhein *et al. Nature* 2019. [10] Thapa *et al, Cancers* 2019.
[11] Wasserstein *et al, Am Stat* 2016. [12] Ranapurvala *et al, Am J Prev Med* 2020

NO “symmetry”?: *Significance* reported in many studies that cite AGM

- On the other hand, **many authors who cite the AGM’s *Nature* Comment , explicitly comment on “significant results”**
- 3 Examples from high-ranking journals:

➢ 1/ e.g. Marmor *et al*, ***Cancer*** **(IF = 5.7)** 2020, state:

“… AI/AN women were found to be significantly more likely to have a high-risk (OR=1.28; ***95% CI: 1.01-1.66****)”.*

➢ 2/ Rosoff *et al* - ***JAMA Psychiatry*** **(IF = 21.6)** 2021

➢ 3/ Perez-Cornago *et al* – ***Int J Epidemiology*** **(IF = 7.7)** 2021

AGM's "Flagship example" of Mis-use of (Non-)Significance

- **AGM provide just 1 empirical example of a grossly incorrect interpretation of the results of significance testing** [1], based on **comparing 2 studies of a similar association:**
  - **(i) Larger study 1: 'statistically significant' RR = of 1.2 (95% CI: 1.09 to 1.33, p=0.0003)** [18]
  - **(ii) A later, Smaller study 2: Identical RR=1.2; but association was deemed 'NON-significant':**

    95% CI: 0.97 to 1.48, p=0.091] because the 95% CI included 1 [16]

- The authors of study 2 then **concluded that** [16]**:**

  their ("Non-significant') results "*stood in contrast"* with ("significant") results of study 1

- Obviously, **we agree with AGM that this "conclusion" is entirely unjustifiable and reflect a glaring misinterpretation of the results of statistical significance testing!**

- **However, we do NOT think that the 'main culprit' was the use of significance testing!**

[1] Amrhein *et al. Nature* 2019. [16] Chao et al. *Int J Cardiology* 2013.
[17] Schmidt & Rothman, *Int J Cardiol* 2014. [18] Schmidt *et al, BMJ* 2011.

Revisiting the "Flagship example" with a Proper use of a Significance test

- The paradoxical "conclusion" about the "*contrast*" between the results of the two studies is **due to mixing up (i) 2 independent formal tests with (ii) informal and incorrect comparison of their dichotomized p-values**

- **Formal statistical test of the Significance of the Difference between the 2 estimates yields p=1.0 as the two point estimates are *identical* (RR=1.2)**

- The **95% CI for the difference** of the log(RR)'s is **(-0.23 to +0.23)**

- Thus, **formal statistical inference, whether based on significance test or on the 95% CI for the difference, clearly indicates NO evidence of the Difference between the results of the two studies and, thus, will permit avoiding the totally erroneous conclusion**

Conclusions

- AGM's *Nature* Comment leads to "loose" interpretations of apparent effects/ differences/ associations that may likely reflect just sampling error in Empirical studies (as predicted e.g. by Haaf [4], Ioannidis [5], and others)
- Many problems pointed out by AGM could be avoided by a Correct Rigorous use of statistical inference combined with better Education of End-Users

[4] Haaf et al, *Nature* 2019. [5] Ioannidis, JAMA 2019.

Proposed STRATOS approach

- Members of the STRATOS Initiative recently decided to **propose a more Balanced Perspective on the role and use of Significance Testing** (and statistical inference in general) **in Applied Research**
- **Writing group of 17 statisticians** with different expertise/opinions (*8 countries on 3 continents, All 9 STRATOS Topic Groups*) will **discuss the pros & cons of different approaches and will aim at ‘partial consensus’ while recognizing potential divergent opinions**
- We’ll focus on better Education of End-Users about Correct use of Significance Tests through both (i) theoretical arguments & (ii) empirical examples
- The draft document will be circulated to all > 100 STRATOS members for further comments/revisions and/or endorsements

Current Members of the Writing Group

- **Anne-Laure Boulesteix**, Germany
- **Daniela Dunkler,** Austria
- **Mitch Gail**, USA
- **Els Goetghebeur,** Belgium
- **Marianne Huebner**, USA
- **Saskia Le Cessie**, the Netherlands
- **Kate Lee**, Australia
- **Roderick Little**, USA
- **Willi Sauerbrei**, Germany
- **Ewout Steyerberg**, the Netherlands
- **Ben van Calster**, the Netherlands
- **Michael Wallace**, Canada
- **Mark Woodward**, Australia
- **Laure Wynants**, Belgium

Project co-Leaders:

Michal Abrahamowicz, Canada

James Carpenter, UK

Victor Kipnis, USA

References

1. Amrhein V, Greenland S, McShane B. Retire statistical significance. *Nature* 2019;567:305-307.
2. Ioannidis JPA. Retiring significance: a free pass to bias. *Nature* 2019;567:461.
3. Johnson VE. Retiring significance: raise the bar. *Nature* 2019:567:461.
4. Haaf JM, Ly A, Wagenmakers EJ. Retire significance, but still test hypotheses. *Nature* 2019;567:461.
5. Ioannidis JPA. The importance of predefined rules and prespecified statistical analyses: do not abandon significance. *JAMA* 2019;321(21):2067-2068.
6. Popper K.R. The Logic of Scientific Discovery. Hutchinson & Co. 1959.
7. Panikkar B, Lemmond B, Allen L, Dipirro C, Kasper S. Making the invisible visible: Results of a community-led health survey following PFAS contamination of drinking water in Merrimack, New Hampshire. *Environmental Health: A Global Access Science Source* 2019;18(1):79.
8. He Y, Theodoratou E, Li X, Din FVN, Vaughan-Shaw P, Svinti V, Farrington SM, Campbell H, Dunlop MG, Timofeeva M. Effects of common genetic variants associated with colorectal cancer risk on survival outcomes after diagnosis: A large population-based cohort study. International Journal of Cancer 2019;145(9):2427-2432.
9. Fischer R. Statistical Methods for Research Workers. Edinburgh: Oliver & Boyd. 1925.
10. Thapa S, Fischbach LA, Delongchamp R, Faramawi MF, Orloff M. The association between salt and potential mediators of the gastric precancerous process. *Cancers* 2019; 11(4):535.
11. Wasserstein RL, Lazar NA. The ASA's Statement on p-values: context, process, and purpose. American Statistician 2016;70:129–133.
12. Ranapurwala SI, Ringwalt CL, Pence BW, Schiro S, Fulcher N, McCort A. DiPrete BL, Marshall, SW. State Medical Board Policy and Opioid Prescribing: A Controlled Interrupted Time Series. *American Journal of Preventive Medicine* 2021;60(3):343-351.
13. Marmor S, Longacre CF, Altman AM, Hui JYC, Jensen EH, Tuttle TM. Genomic expression assay testing among American Indian and Alaska Native women with breast cancer. *Cancer* 2020;126(24):5222-5229.
14. Rosoff DB, Smith GD, Mehta N, Clarke TK, Lohoff FW. Evaluating the relationship between alcohol consumption, tobacco use, and cardiovascular disease: A multivariable Mendelian randomization study. *PLoS Medicine* 2020;17(12):e1003410.
15. Perez-Cornago A, Crowe FL, Appleby PN, *et al*. Plant foods, dietary fibre and risk of ischaemic heart disease in the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort. *International Journal of Epidemiology* 2021;50(1):212-222.
16. Chao TF, Liu CJ, Chen SJ, Wang KL, Lin TJ, Chang SL, *et al*. The association between the use of non-steroidal anti-inflammatory drugs and atrial fibrillation: a nationwide case-control study. *International Journal of Cardiology* 2013;168(1):312-316.
17. Schmidt M, Rothman KJ. Mistaken inference caused by reliance on and misinterpretation of a significance test. *International Journal of Cardiology* 2014;177(3):1089-1090.
18. Schmidt M, Christiansen CF, Mehnert F, Rothman KJ, Sørensen HT. Non-steroidal anti-inflammatory drug use and risk of atrial fibrillation or flutter: population based case-control study. *BMJ* 2011; 343:d3450.

Our and Other Statisticians' Concerns about AGM's "Black vs. White" recommendations

- **Removing the "gatekeeper" of statistical significance may open the floodgates toward an uncontrolled reporting of "associations" that may likely reflect just a combination of (i) sampling errors & (ii) Authors' wishful thinking**

- **Similar concerns** expressed (right after AGM Comment publication) **by other statisticians** [e.g. 2-5]:
  - E.g., **Julia Haaf *et al* state:** "… *when statistical testing is skipped, ... any differences between observations would be considered meaningful*" [3]
  - **John Ioannidis warns that removal of statistical significance**, a necessary "gatekeeper" to ensure **falsifiability** of the postulated scientific hypotheses [6], **may lead to** "*statistical anarchy"*, and concludes *"Without clear rules for analyses, science and policy may rely less on data and evidence and more on subjective opinions and interpretations*" [5]

[2] Ioannidis, *Nature* 2019. [3] Johnson, *Nature* 2019. [4] Haaf et al, *Nature* 2019.
[5] Ioannidis, JAMA 2019. [6] Karl R. Popper, *The Logic of Scientific Discovery* 1959.

NO “symmetry”?: *Significance* reported in many studies that cite AGM

- On the other hand, **IF the 95% CI for effects of interest excluded the null, or equivalently p<0.05, many authors reported the “significant associations” or “effects” in a conventional way, in spite of having cited the AGM’s Comment,** e.g.:

➢ Marmor *et al* - ***Cancer*** **(IF = 5.7)** 2020

“… AI/AN women were found to be significantly more likely to have a high-risk (OR=1.28; 95% CI: 1.01-1.66)”.

➢ Rosoff *et al* - ***JAMA Psychiatry*** **(IF = 21.6)** 2021

“… we used a stringent selection threshold (P < 5 × 10−6) for the pain medication use and ASRD risk instruments to compensate for lack of SNVs with effect P values less than conventional genome-wide significance (P < 5 × 10−8)”.

➢ Perez-Cornago *et al* – ***Int J Epidemiology*** **(IF = 7.7)** 2021

- *“… only the intake of fruit was significantly associated with a lower risk”.*
- *“… and borderline significant inverse association between legume intake and IHD risk based on 10 prospective studies (RR…: 0.91, 95% CI 0.84-0.99)”*

[13] Marmor *et al, Cancer* 2020. [14] Rosoff *et al, JAMA Psych* 2021.
[15] Perez-Cornago *et al, Int J Epidemiol* 2021.

Examples of Impact in Empirical Studies that cite AGM's Comment

Panikkar *et al* [7], *Environmental Health* 2019 (IF = 4.7), state in Methods:

> *"To avoid placing too much emphasis on statistical significance, we emphasize the strength of associations in our results as well* [1]*."*

(Similar statements in Methods of several other papers that cite AGM)

Then, in Results:

> *"Participants who had water filtration were also* ***close to 3 times more likely*** *to report developmental disorders (****OR = 2.960*** *(95%) CI: 0.7–12.8). … Residents who lived in Merrimack for 18–30 years (****OR = 4.966*** *95% CI: 0.6–42.9) and over 30 years (****OR = 5.456*** *95% CI: 0.3–90.6) were* ***5 times as likely*** *to report developmental problems."* [7]

- Interpretating the point estimates as indicating "*close to 3 times*" or "*5 times*" risk increases illustrates the hazards of ignoring statistical (NON-)significance, and statistical inference in general
    i. All the three ORs would have a reasonable chance (>13% or >23%) of being observed even if there were no associations at all, with all **p-values >0.10 (0.14, 0.14 & 0.24)**
    ii. Furthermore, **the 95% CIs indicate that the point estimates are extremely imprecise**, and that the ranges of **ORs consistent with the observed results include even *important (up to 70%) risk reductions***!

[1] Amrhein *et al. Nature* 2019. [7] Panikkar *et al,* Environ Health 2019.

Thapa *et al*, Cancers (IF=6.7)

- In Methods:

  "Consistent with recommendations..., our analysis focused on effect estimation rather than statistical significance testing [1,11]'

- Then, in Results, they discuss "***Differences***" between effects in different subgroups [10] **which very likely reflect just the sampling error**, e.g.: **

  0.149 (95% CI: 0.007, 0.292) for H. Pyl. + *versus*

  0.103 (95% CI: -0.285, +0.490) for H. Pyl. –

  "a LARGER increase... was observed for … H. Pyl +…." [10]

- Yet, for the DIFFERENCE = 0.046 (95% CI: -0.367, 0.459), p = 0.827 !!

- ** Similar issues e.g. in [Ranapurwala *et al*, *Am J Prev Med*] (IF = 4.5) [12]

[1] Amrhein *et al. Nature* 2019. [10] Thapa *et al, Cancers* 2019.
[11] Wasserstein *et al, Am Stat* 2016. [12] Ranapurvala *et al, Am J Prev Med* 2020

He et al, *Int J Cancer* (IF=5.1) [8]

- **He *et al* [8] state: "***We additionally looked into direction of effects to overcome limitations of statistical significance.*"

  And then conclude: "*Though not reaching suggested significance level (p≤0.05), these results are consistent with directions of effects observed in previous studies.*"

- **Yet, if p>0.05, i.e. the 95% CIs include the null effect, the direction of the association cannot be firmly established****, and results are compatible with all: (i) risk increases, (ii) risk decreases, and (iii) H_0 of no association!

** As pointed out by Ronald Fischer, > 90 years ago [9]:

Statistical significance tests are necessary to "... *test if there is anything to justify estimation at all*"

[8] He *et al*, *Int J Cancer 2019.* [9] Fischer 1925

Re-analyses of “Flagship example”: do NOT “mix” Formal Statistical Inference with IN-formal argumentation !

- Erroneous “paradoxical’ conclusion (b) that smaller study 2 results *“stood in contrast”* with “significant” study 1 results is **due to mixing up (i) 2 independent formal tests with (ii) informal and incorrect comparison of their dichotomized p-values**

- **Formal statistical test of the “significance” of the difference between the 2 estimates yields p=1.0 because the point estimates are *identical* (RR=1.2)**

- The 95% CI for the difference between the corresponding log(RR) is (-0.23 to +0.23), implying **the 95% CI (0.63 to 1.59) for the Ratio (RR1/RR2) of the 2 effects**

- Thus, **formal statistical inference, whether based on significance test or on the 95% CI for the difference, clearly indicates NO evidence of the Difference between results of the two studies and, thus, will permit avoiding the totally erroneous conclusion (b)**

- Yet, the **95% CI for the difference** indicates that the **results are still compatible with a moderate yet clinically meaningful difference, with one RR being possibly more than 50% higher than the other.** Thus, the **Equality of the 2 RR point estimates does *NOT* imply that the corresponding (unknown) true effects are exactly the same!**

AGM's "Flagship example" of Mis-use of (Non-)Significance

- **AGM provide just 1 empirical example of a grossly incorrect interpretation of the results of significance testing** [1], discussed earlier by Schmidt & Rothman [17]:

  They compare results of 2 studies of potential atrial fibrillation (AF) risks associated with an anti-inflammatory drug:

  - ➢ **An earlier, larger study 1 reported a 'statistically significant' association with relative risks (RR) of 1.2 (95% CI: 1.09 to 1.33, p=0.0003)** [18]
  - ➢ **In a later, smaller study 2, the point estimate of RR was identical to study 1 (RR=1.2; 95% CI: 0.97 to 1.48, p=0.091] but association was deemed 'non-significant' because the 95% CI included 1** [16]

- The authors of study 2 **concluded that** [16]**:**

  (a) The use of drugs under study was *"not associated"* with AF risks, and

  (b) Their results *"stood in contrast"* with ("significant") results of study 1

- Obviously, **we agree with AGM that both conclusions (a) and (b) are entirely unjustifiable and reflect a glaring misinterpretation of the results of statistical significance testing!**

- **However, we do NOT think that the 'main culprit' was the use of significance testing!**

[1] Amrhein *et al. Nature* 2019. [16] Chao et al. *Int J Cardiology* 2013.
[17] Schmidt & Rothman, *Int J Cardiol* 2014. [18] Schmidt *et al, BMJ* 2011.

"Flagship example": **How to Interpret the results of the smaller study 2 ?**

- AMG's statement *"it is **ludicrous** to conclude ... '**no association' when the interval estimate includes serious risk increases**..."* [1]
- Logically implies that, by symmetry, **we should also take into account the lower range of RR value in the 95% CI** (0.97 to 1.48), which **does include the null effect of RR=1.0**
- **Thus, when considered independently of study 1, study 2 does *not* provide a strong evidence of risk increase**: the point estimate of RR=1.2 or higher would be reasonably likely (probability ~ 0.09) to be observed by chance alone even if there is no true association in the source population, with the true RR=1.0
- **So Interpretation of the results from the smaller study 2 [16] depends on whether they are assessed:**
  - **i.** **INDEPENDENTLY of** earlier results of the larger study 1 [18], **OR**
  - **ii.** **Taking into Account** these Earlier Results

[1] Amrhein *et al, Nature* 2019. [16] Chao *et al, Int J Cardiol* 2013.
[18] Schmidt *et al, BMJ* 2011.

Further comments on the "Flagship example": **Difficulties in avoiding "Dichotomy"**

- NOTE: AGM say: *"It is ludicrous to conclude that the statistically non-significant result showed 'no association' when the INTERVAL ESTIMATE Includes a serious risk increase."* (Earlier they say: "*The 95% CI… included a considerable risk increase of 48%*" (the UPPER Bound of the 95% CI!)

- However, much depends on the Confidence level used for the "*Interval*". E.g. **the 80% CI** for RR **(0.73, 1.38) will Exclude risk increases of 40% or more.**

- **Yet, choosing the confidence level – which determines if the "interval" does or does not include a specific strength of the effect - requires Necessary DICHOTOMIZATION which Amrhein *et al* [1] seem to strongly oppose…**

[1] Amrhein *et al, Nature* 2019.

Outline of Joint Presentations

- Background: Overview of *Nature* 2019 **Amrhein, Greenland & McShane's (AGM)** Comment (MA)
- Examples of the Comment's Impact on Applied research (MA)
- Re-analysis of the AGM "Flagship example" (MA)
- Outline of the proposed STRATOS approach (MA)
- Back to the origins: historical perspective on Significance tests *vs.* Hypothesis testing (VK)
- Some common mistakes/pitfalls to avoid (VK)