Karolinska Institutet

A discussion on the use of the p-value

Methodological issues in epidemiology and population research

A LUPOP SEMINAR SERIES

27 May 2021

Laura Pazzagli, Ph. D.
Statistician

Centre for Pharmacoepidemiology,
Department of Medicine, Solna
Karolinska Institutet

About me

- BSc in Statistics and Computer Science, University of Perugia, Italy, 2003
- MSs in Statistics and Computer Science, University of Perugia, Italy, 2007
- Break in my academic education to enter the job market, 2007-2013
- PhD in Mathematical and statistical methods, University of Perugia, Italy, 2015
  Thesis title: “Inverse probability weighting and doubly robust estimators under model misspecification and a study on the socioeconomic determinants of end-stage renal disease in Sweden”
- Visiting PhD student, Research Assistant, and Associate Researcher, Department of Statistics, Umeå University, Sweden, 2013-2016
- Postdoctoral fellow, Centre for Pharmacoepidemiology, Karolinska Institutet, 2016-2020
  Project title: “Methodologies for time-varying exposure in pharmacoepidemiology”
- Statistician, Centre for Pharmacoepidemiology, Karolinska Institutet, 2020-Present working in several studies related to drug safety
- Visiting Professor, University of Milano, Italy, 2021

Outline

- Background
- Introduction to the p-value
- Advantages of the p-value
- Disadvantages of the p-value
- Problems with the use of the p-value in the research environment
- P-mining and conflict of interest
- Three common misuses of p-values
- The fickle P value generates irreproducible results
- Alternatives to the p-value
- Conclusions
- References

Background

- ***Swedish Statistics Promotion*** *is a non-profit association* for friends of statistics - statisticians, statistics users and people with a general interest in statistics.

Statistikfrämjandet

ETT BORTGLÖMT JUBILEUM
SIDORNA 2-4

ATT JOBBA INOM MEDICINSK FORSKNING
SIDORNA 7-8

Qvintensen

NR 2 2019

STATISTIKUTBILDNING OCH SIMULERING
SIDORNA 30-31

TEMA/

P-värde – ett redskap i den statistiska verktygslådan

SIDORNA 11-2

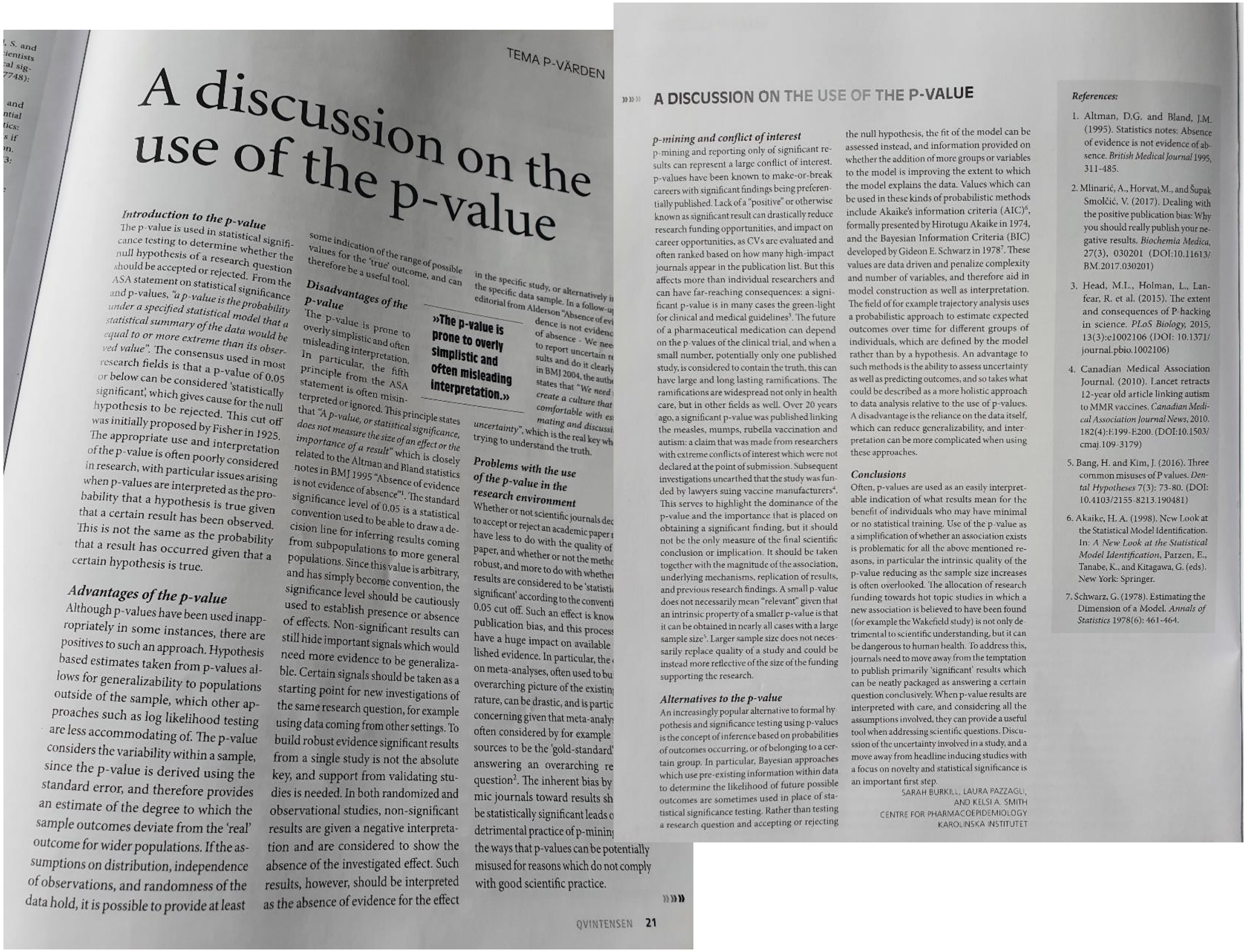
Background

- Sarah Burkill, Kelsi A. Smith, & Laura Pazzagli

- Centre for Pharmacoepidemiology, Department of Medicine Solna, Karolinska Institutet

- Clinical Epidemiology Division, Department of Medicine Solna, Karolinska Institutet

Background



A discussion on the use of the p-value

Introduction to the p-value

The p-value is used in statistical significance testing to determine whether the null hypothesis of a research question should be accepted or rejected. From the ASA statement on statistical significance and p-values, *"a p-value is the probability under a specified statistical model that a statistical summary of the data would be equal to or more extreme than its observed value"*. The consensus used in most research fields is that a p-value of 0.05 or below can be considered 'statistically significant', which gives cause for the null hypothesis to be rejected. This cut off was initially proposed by Fisher in 1925. The appropriate use and interpretation of the p-value is often poorly considered in research, with particular issues arising when p-values are interpreted as the probability that a hypothesis is true given that a certain result has been observed. This is not the same as the probability that a result has occurred given that a certain hypothesis is true.

Advantages of the p-value

Although p-values have been used inappropriately in some instances, there are positives to such an approach. Hypothesis based estimates taken from p-values allows for generalizability to populations outside of the sample, which other approaches such as log likelihood testing are less accommodating of. The p-value considers the variability within a sample, since the p-value is derived using the standard error, and therefore provides an estimate of the degree to which the sample outcomes deviate from the 'real' outcome for wider populations. If the assumptions on distribution, independence of observations, and randomness of the data hold, it is possible to provide at least some indication of the range of possible values for the 'true' outcome, and can therefore be a useful tool.

Disadvantages of the p-value

The p-value is prone to overly simplistic and often misleading interpretation. In particular, the fifth principle from the ASA statement is often misinterpreted or ignored. This principle states that *"A p-value, or statistical significance, does not measure the size of an effect or the importance of a result"* which is closely related to the Altman and Bland statistics notes in BMJ 1995 "Absence of evidence is not evidence of absence"[1]. The standard significance level of 0.05 is a statistical convention used to be able to draw a decision line for inferring results coming from subpopulations to more general populations. Since this value is arbitrary, and has simply become convention, the significance level should be cautiously used to establish presence or absence of effects. Non-significant results can still hide important signals which would need more evidence to be generalizable. Certain signals should be taken as a starting point for new investigations of the same research question, for example using data coming from other settings. To build robust evidence significant results from a single study is not the absolute key, and support from validating studies is needed. In both randomized and observational studies, non-significant results are given a negative interpretation and are considered to show the absence of the investigated effect. Such results, however, should be interpreted as the absence of evidence for the effect

in the specific study, or alternatively i
the specific data sample. In a follow-u
editorial from Alderson "Absence of evi
dence is not evidenc
of absence - We nee
to report uncertain re
sults and do it clearly
in BMJ 2004, the auth
states that *"We need*
create a culture that
comfortable with es
mating and discussi
uncertainty", which is the real key wh
trying to understand the truth.

»The p-value is prone to overly simplistic and often misleading interpretation.»

Problems with the use of the p-value in the research environment

Whether or not scientific journals dec
to accept or reject an academic paper r
have less to do with the quality of
paper, and whether or not the meth
robust, and more to do with whethe
results are considered to be 'statisti
significant' according to the conventi
0.05 cut off. Such an effect is know
publication bias, and this process
have a huge impact on available
lished evidence. In particular, the
on meta-analyses, often used to bu
overarching picture of the existin
rature, can be drastic, and is partic
concerning given that meta-analy
often considered by for example
sources to be the 'gold-standard'
answering an overarching re
question[2]. The inherent bias by
mic journals toward results sh
be statistically significant leads
detrimental practice of p-minin
the ways that p-values can be potentially misused for reasons which do not comply with good scientific practice.

»»»

p-mining and conflict of interest

p-mining and reporting only of significant results can represent a large conflict of interest. p-values have been known to make-or-break careers with significant findings being preferentially published. Lack of a "positive" or otherwise known as significant result can drastically reduce research funding opportunities, and impact on career opportunities, as CVs are evaluated and often ranked based on how many high-impact journals appear in the publication list. But this affects more than individual researchers and can have far-reaching consequences: a significant p-value is in many cases the green-light for clinical and medical guidelines[3]. The future of a pharmaceutical medication can depend on the p-values of the clinical trial, and when a small number, potentially only one published study, is considered to contain the truth, this can have large and long lasting ramifications. The ramifications are widespread not only in health care, but in other fields as well. Over 20 years ago, a significant p-value was published linking the measles, mumps, rubella vaccination and autism: a claim that was made from researchers with extreme conflicts of interest which were not declared at the point of submission. Subsequent investigations unearthed that the study was funded by lawyers suing vaccine manufacturers[4]. This serves to highlight the dominance of the p-value and the importance that is placed on obtaining a significant finding, but it should not be the only measure of the final scientific conclusion or implication. It should be taken together with the magnitude of the association, underlying mechanisms, replication of results, and previous research findings. A small p-value does not necessarily mean "relevant" given that an intrinsic property of a smaller p-value is that it can be obtained in nearly all cases with a large sample size[5]. Larger sample size does not necessarily replace quality of a study and could be instead more reflective of the size of the funding supporting the research.

Alternatives to the p-value

An increasingly popular alternative to formal hypothesis and significance testing using p-values is the concept of inference based on probabilities of outcomes occurring, or of belonging to a certain group. In particular, Bayesian approaches which use pre-existing information within data to determine the likelihood of future possible outcomes are sometimes used in place of statistical significance testing. Rather than testing a research question and accepting or rejecting the null hypothesis, the fit of the model can be assessed instead, and information provided on whether the addition of more groups or variables to the model is improving the extent to which the model explains the data. Values which can be used in these kinds of probabilistic methods include Akaike's information criteria (AIC)[6], formally presented by Hirotugu Akaike in 1974, and the Bayesian Information Criteria (BIC) developed by Gideon E. Schwarz in 1978[7]. These values are data driven and penalize complexity and number of variables, and therefore aid in model construction as well as interpretation. The field of for example trajectory analysis uses a probabilistic approach to estimate expected outcomes over time for different groups of individuals, which are defined by the model rather than by a hypothesis. An advantage to such methods is the ability to assess uncertainty as well as predicting outcomes, and so takes what could be described as a more holistic approach to data analysis relative to the use of p-values. A disadvantage is the reliance on the data itself, which can reduce generalizability, and interpretation can be more complicated when using these approaches.

Conclusions

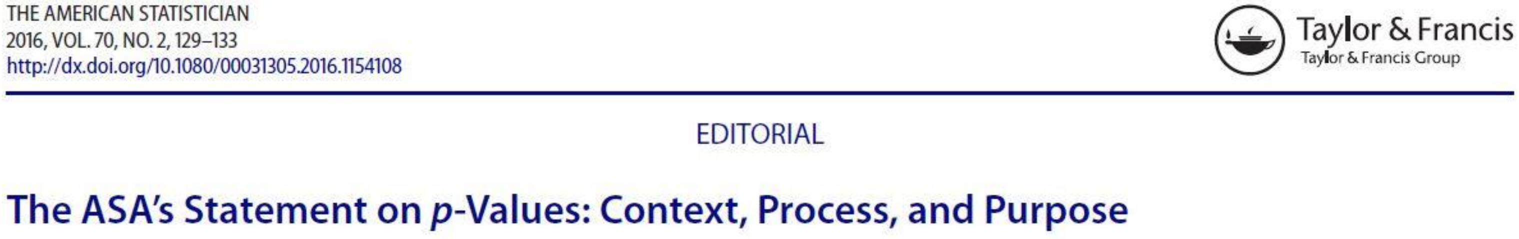
Often, p-values are used as an easily interpretable indication of what results mean for the benefit of individuals who may have minimal or no statistical training. Use of the p-value as a simplification of whether an association exists is problematic for all the above mentioned reasons, in particular the intrinsic quality of the p-value reducing as the sample size increases is often overlooked. The allocation of research funding towards hot topic studies in which a new association is believed to have been found (for example the Wakefield study) is not only detrimental to scientific understanding, but it can be dangerous to human health. To address this, journals need to move away from the temptation to publish primarily 'significant' results which can be neatly packaged as answering a certain question conclusively. When p-value results are interpreted with care, and considering all the assumptions involved, they can provide a useful tool when addressing scientific questions. Discussion of the uncertainty involved in a study, and a move away from headline inducing studies with a focus on novelty and statistical significance is an important first step.

SARAH BURKILL, LAURA PAZZAGLI,
AND KELSI A. SMITH
CENTRE FOR PHARMACOEPIDEMIOLOGY
KAROLINSKA INSTITUTET

References:

1. Altman, D.G. and Bland, J.M. (1995). Statistics notes: Absence of evidence is not evidence of absence. *British Medical Journal* 1995, 311-485.
2. Mlinarić, A., Horvat, M., and Šupak Smolčić, V. (2017). Dealing with the positive publication bias: Why you should really publish your negative results. *Biochemia Medica*, 27(3), 030201 (DOI:10.11613/BM.2017.030201)
3. Head, M.L., Holman, L., Lanfear, R. et al. (2015). The extent and consequences of P-hacking in science. *PLoS Biology*, 2015, 13(3):e1002106 (DOI: 10.1371/journal.pbio.1002106)
4. Canadian Medical Association Journal. (2010). Lancet retracts 12-year old article linking autism to MMR vaccines. *Canadian Medical Association Journal News*, 2010. 182(4):E199-E200. (DOI:10.1503/cmaj.109-3179)
5. Bang, H. and Kim, J. (2016). Three common misuses of P values. *Dental Hypotheses* 7(3): 73-80. (DOI: 10.4103/2155-8213.190481)
6. Akaike, H. A. (1998). New Look at the Statistical Model Identification. In: *A New Look at the Statistical Model Identification*, Parzen, E., Tanabe, K., and Kitagawa, G. (eds). New York: Springer.
7. Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics* 1978(6): 461-464.

Background



THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 2, 129–133
http://dx.doi.org/10.1080/00031305.2016.1154108

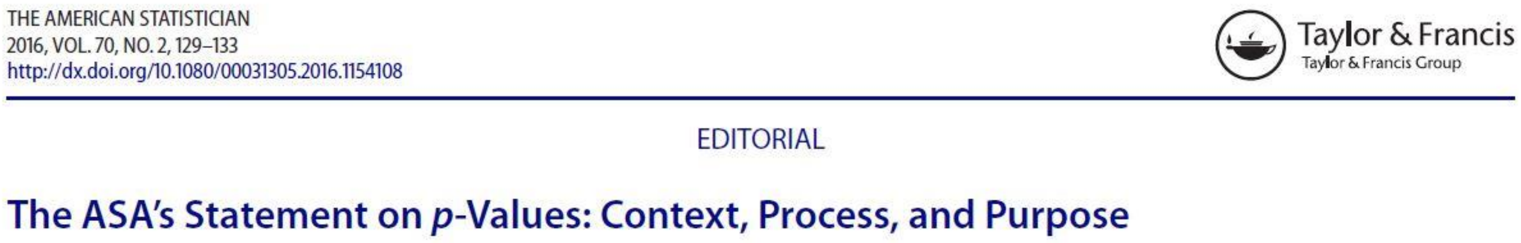
Taylor & Francis
Taylor & Francis Group

EDITORIAL

The ASA's Statement on *p*-Values: Context, Process, and Purpose

"*The Board envisioned that the American Statistical Association (ASA) statement on p-values and statistical significance would shed light on an aspect of our field that is too often misunderstood and misused in the broader research community, and, in the process, provides the community a service.*"

Background



THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 2, 129–133
http://dx.doi.org/10.1080/00031305.2016.1154108

Taylor & Francis
Taylor & Francis Group

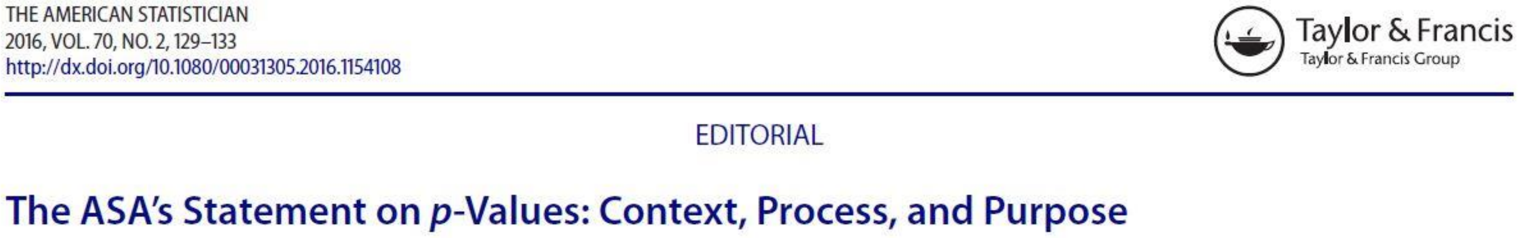
EDITORIAL

The ASA's Statement on *p*-Values: Context, Process, and Purpose

- *The Board tasked Ronald L. Wasserstein with assembling a group of experts representing a wide variety of points of view. On behalf of the Board, he reached out to more than two dozen such people, all of whom said they would be happy to be involved.*

- *The statement development process was lengthier and more controversial than anticipated.*

Background



THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 2, 129–133
http://dx.doi.org/10.1080/00031305.2016.1154108

Taylor & Francis
Taylor & Francis Group

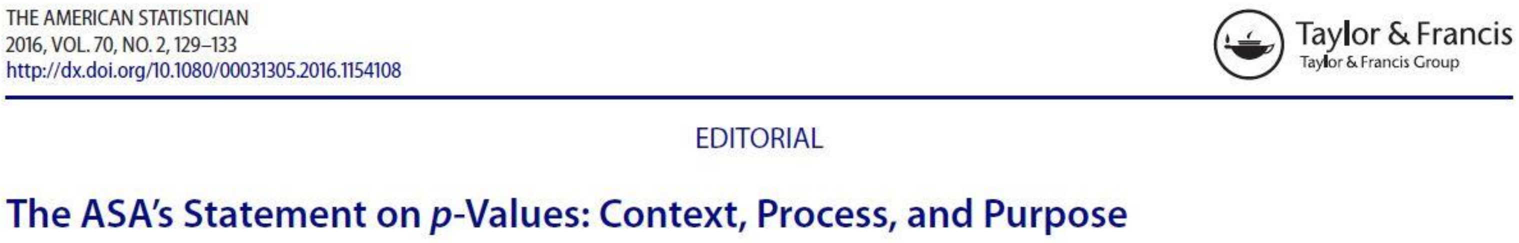
EDITORIAL

The ASA's Statement on *p*-Values: Context, Process, and Purpose

- *A group of discussants was contacted to provide comments on the statement (Naomi Altman, Douglas Altman, Daniel J. Benjamin, Yoav Benjamini, Jim Berger, Don Berry, John Carlin, George Cobb, Andrew Gelman, Steve Goodman, Sander Greenland, John Ioannidis, Joseph Horowitz, Valen Johnson, Michael Lavine, Michael Lew, Rod Little, Deborah Mayo, Michele Millar,Charles Poole, Ken Rothman, Stephen Senn, Dalene Stangl, Philip Stark and Steve Ziliak).*

- *Their statements were published in the online supplement to the ASA statement.*

Principles



THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 2, 129–133
http://dx.doi.org/10.1080/00031305.2016.1154108

Taylor & Francis
Taylor & Francis Group

EDITORIAL

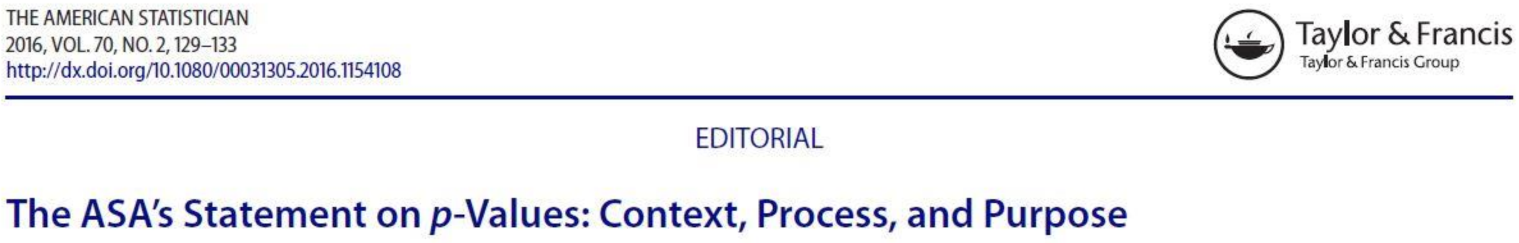
The ASA's Statement on *p*-Values: Context, Process, and Purpose

1. ***P*-values can indicate how incompatible the data are with a specified statistical model.**

2. ***P*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.**

3. **Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.**

Principles



THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 2, 129–133
http://dx.doi.org/10.1080/00031305.2016.1154108

Taylor & Francis
Taylor & Francis Group

EDITORIAL

The ASA’s Statement on *p*-Values: Context, Process, and Purpose

4. **Proper inference requires full reporting and transparency.**

5. **A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.**

6. **By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.**

Introduction to the p-value

- H_0: the difference is 0 vs. H_a : the difference is not 0
- Type I error = α = significance level, is the probability of rejecting H_0 when H_0 is true (false positive)
- Type II error = β, is the probability of not rejecting H_0 when H_a is true (false negative)
- Power = 1- β, is the probability of rejecting H_0 when H_a is true
- p-value is the probability of obtaining a test statistic at least as extreme as the statistic observed under the H_0 (and every other assumption made). P-value depends on the sample size.

Introduction to the p-value

- The p-value is used in statistical significance testing to determine whether the null hypothesis of a research question should be rejected or not.

- From the ASA statement on statistical significance and p-values, "*a p-value is the probability under a specified statistical model that a statistical summary of the data would be equal to or more extreme than its observed value*".

- The consensus used in most research fields is that a p-value of 0.05 or below can be considered 'statistically significant', which gives cause for the null hypothesis to be rejected.

Introduction to the p-value

- The appropriate use and interpretation of the p-value is often poorly considered in research.
- Particular issues arise when p-values are interpreted as the probability that a hypothesis is true given that a certain result has been observed.
- This is not the same as the probability that a result has occurred given that a certain hypothesis is not rejected.

Advantages of the p-value

- Hypothesis based estimates taken from p-values allows for generalizability to populations outside of the sample.

- The p-value considers the variability within a sample, since the p-value is derived using the standard error.

- If the assumptions on distribution, independence of observations, and randomness of the data hold, it is possible to provide at least some indication of the range of possible values for the ‘true’ outcome.

Disadvantages of the p-value

- The p-value is prone to overly simplistic and often misleading interpretation.

- In particular, the fifth principle from the ASA statement is often misinterpreted or ignored.

- This principle states that “*A p-value, or statistical significance, does not measure the size of an effect or the importance of a result*” which is closely related to the Altman and Bland statistics notes in BMJ 1995 “Absence of evidence is not evidence of absence”.

Disadvantages of the p-value

- The standard significance level of 0.05 is a statistical convention used to be able to draw a decision line for inferring results coming from subpopulations to more general populations.
- Since this value is arbitrary, and has simply become convention, the significance level should be cautiously used to establish presence or absence of effects.
- Non-significant results can still hide important signals which would need more evidence to be generalizable.

Disadvantages of the p-value

- Certain signals should be taken as a starting point for new investigations of the same research question, for example using data coming from other settings.

- To build robust evidence significant results from a single study are not the absolute key, and support from validating studies is needed.

- In both randomized and observational studies, non-significant results are given a negative interpretation and are considered to show the absence of the investigated effect, however such results should be interpreted as the absence of evidence for the effect in the specific study, or alternatively in the specific data sample.

Disadvantages of the p-value

- In a follow-up editorial from Alderson “Absence of evidence is not evidence of absence - We need to report uncertain results and do it clearly”, in BMJ 2004, the author states that

“***We need to create a culture that is comfortable with estimating and discussing uncertainty***”

- which is the real key when trying to understand the truth.

Problems with the use of the p-value in the research environment

- Whether or not scientific journals decide to accept or reject an academic paper may have less to do with the quality of the paper, and whether or not the method is robust, and more to do with whether the results are considered to be ‘statistically significant’ according to the conventional 0.05 cut off.

- Such an effect is known as **publication bias**, and this process can have a huge impact on available published evidence.

Problems with the use of the p-value in the research environment

- The effect on particular **meta-analyses**, often used to build an overarching picture of the existing literature, can be drastic, and is particularly concerning given that meta-analyses are often considered by for example media sources to be the ‘gold-standard’ when answering an overarching research question.

- The inherent bias by academic journals toward results shown to be statistically significant leads on to the detrimental practice of p-mining, and to the ways that p-values can be potentially misused for reasons which do not comply with good scientific practice.

P-mining and conflict of interest

- P-mining and reporting only of significant results can represent a large conflict of interest.

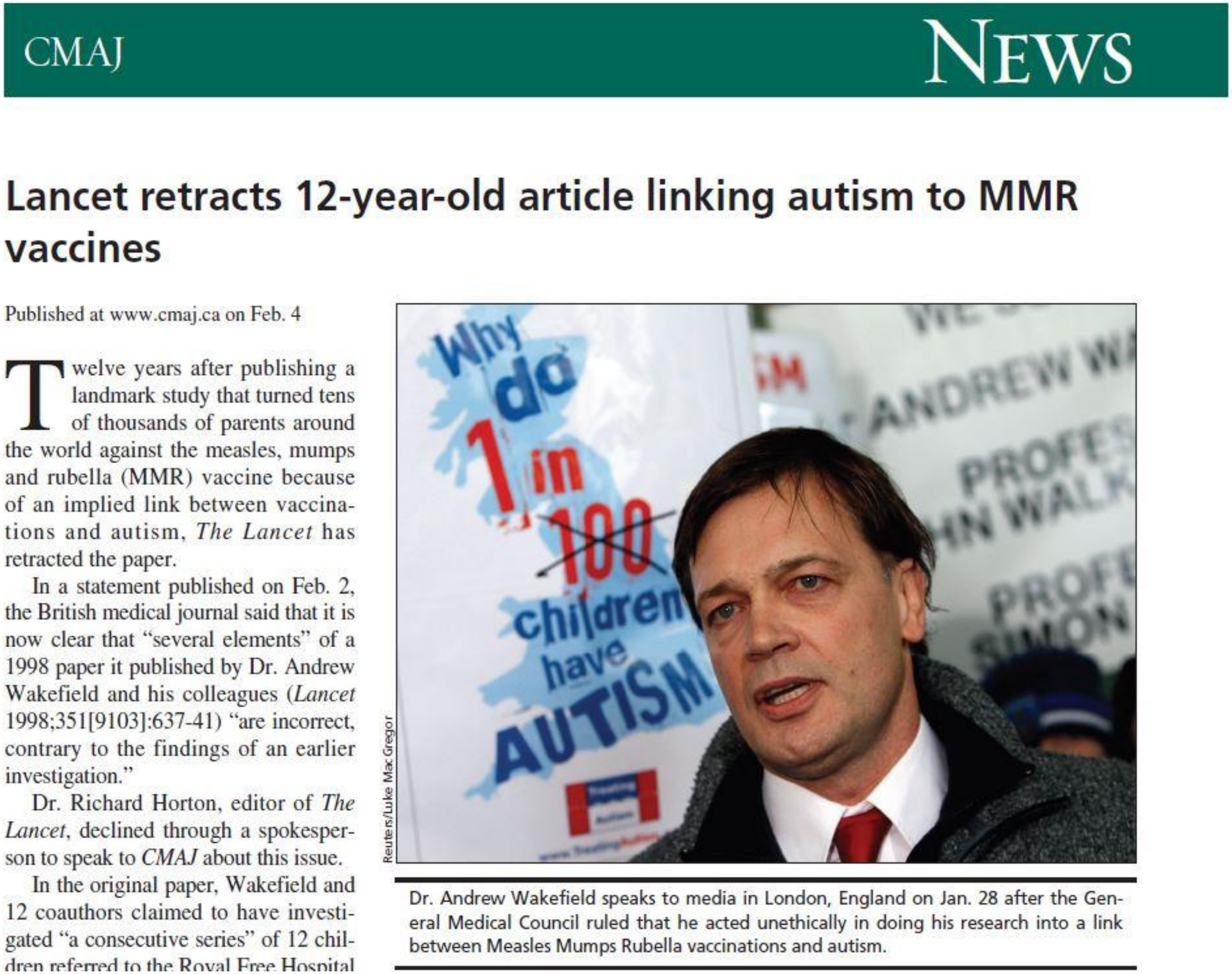
- P-values have been known to make-or-break careers with significant findings being preferentially published.

- Lack of a “positive” or otherwise known as significant result can drastically reduce research funding opportunities, and impact on career opportunities, as CVs are evaluated and often ranked based on how many high-impact journals appear in the publication list.

P-mining and conflict of interest

- The ramifications are widespread not only in health care, but in other fields as well.
- Over 20 years ago, a significant p-value was published linking the measles, mumps, rubella vaccination and autism: a claim that was made from researchers with extreme conflicts of interest which were not declared at the point of submission.
- Subsequent investigations unearthed that the study was funded by lawyers suing vaccine manufacturers.

P-mining and conflict of interest



CMAJ NEWS

Lancet retracts 12-year-old article linking autism to MMR vaccines

Published at www.cmaj.ca on Feb. 4

Twelve years after publishing a landmark study that turned tens of thousands of parents around the world against the measles, mumps and rubella (MMR) vaccine because of an implied link between vaccinations and autism, *The Lancet* has retracted the paper.

In a statement published on Feb. 2, the British medical journal said that it is now clear that "several elements" of a 1998 paper it published by Dr. Andrew Wakefield and his colleagues (*Lancet* 1998;351[9103]:637-41) "are incorrect, contrary to the findings of an earlier investigation."

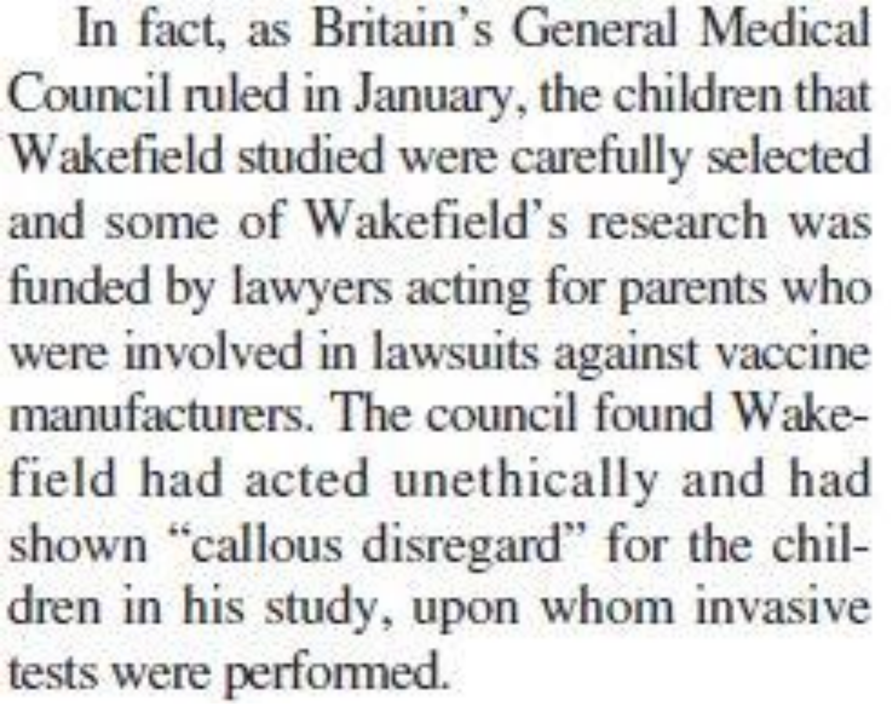
Dr. Richard Horton, editor of *The Lancet*, declined through a spokesperson to speak to *CMAJ* about this issue.

In the original paper, Wakefield and 12 coauthors claimed to have investigated "a consecutive series" of 12 children referred to the Royal Free Hospital

Reuters/Luke MacGregor

Dr. Andrew Wakefield speaks to media in London, England on Jan. 28 after the General Medical Council ruled that he acted unethically in doing his research into a link between Measles Mumps Rubella vaccinations and autism.

P-mining and conflict of interest

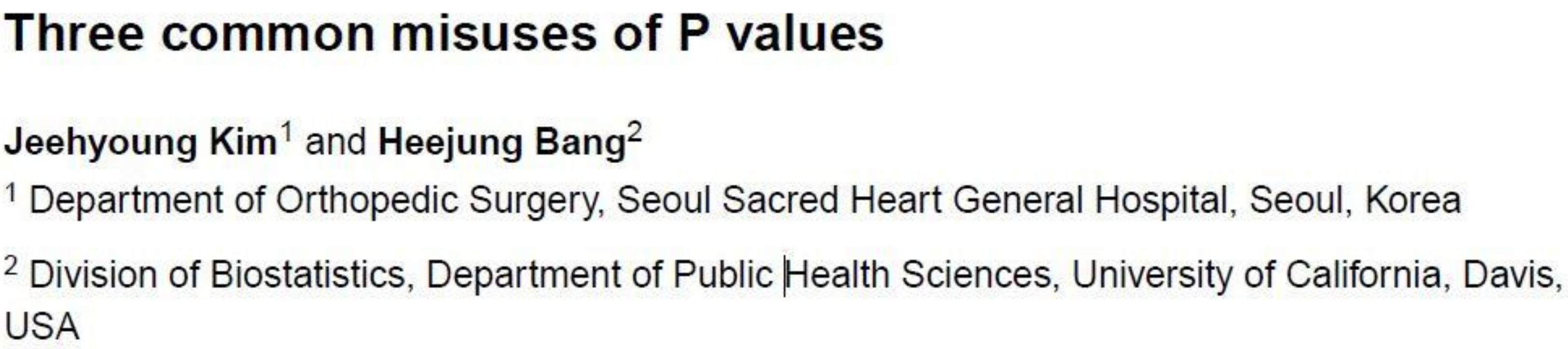


In fact, as Britain's General Medical Council ruled in January, the children that Wakefield studied were carefully selected and some of Wakefield's research was funded by lawyers acting for parents who were involved in lawsuits against vaccine manufacturers. The council found Wakefield had acted unethically and had shown "callous disregard" for the children in his study, upon whom invasive tests were performed.

P-mining and conflict of interest

- This serves to highlight the dominance of the p-value and the importance that is placed on obtaining a significant finding, but it should not be the only measure of the final scientific conclusion or implication.
- It should be taken together with the magnitude of the association, underlying mechanisms, replication of results, and previous research findings.
- A small p-value does not necessarily mean “relevant” given that an intrinsic property of a smaller p-value is that it can be obtained in nearly all cases with a large sample size.
- Larger sample size does not necessarily replace quality of a study and could be instead more reflective of the size of the funding supporting the research.

Three common misuses of p-values



Three common misuses of P values

Jeehyoung Kim[1] and **Heejung Bang**[2]

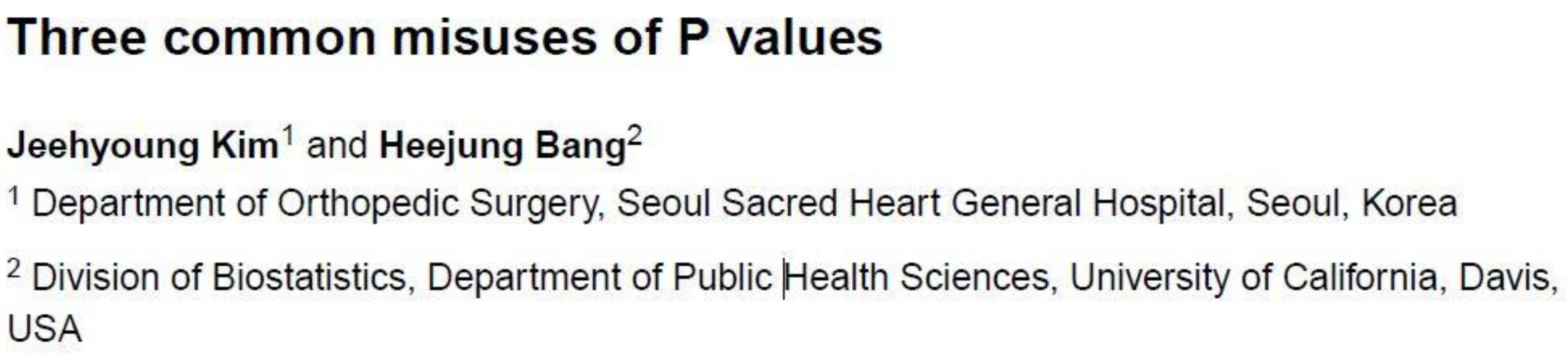
[1] Department of Orthopedic Surgery, Seoul Sacred Heart General Hospital, Seoul, Korea

[2] Division of Biostatistics, Department of Public Health Sciences, University of California, Davis, USA

Large p-value means no difference: Wrong

“One property of the p-value is that it is a function of the sample size (N) (under the Ha). Thus, when N is large, the p-value is destined to be small; this feature can be a reward – acknowledging how hard it is to collect a large sample – but can cause other problems.”

Three common misuses of p-values



Three common misuses of P values

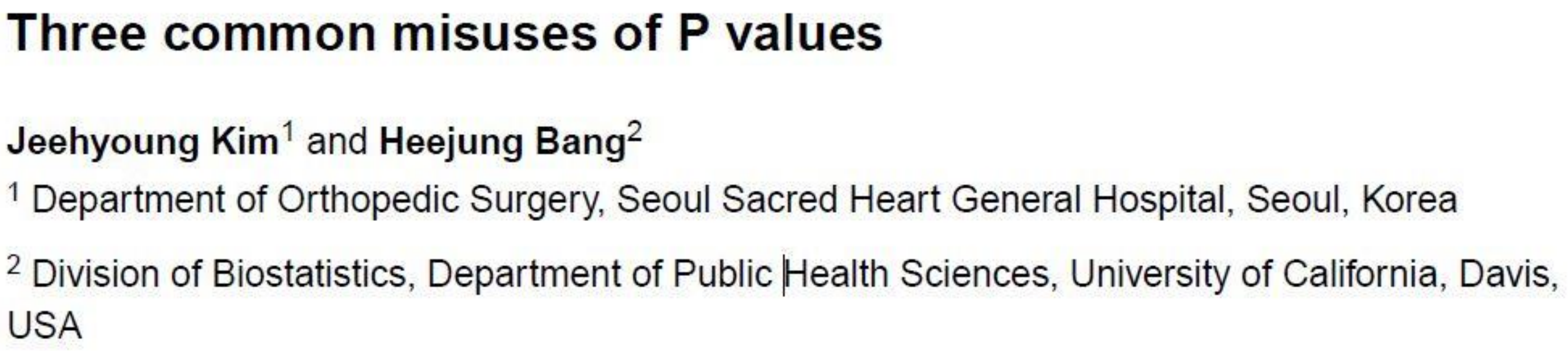
Jeehyoung Kim[1] and **Heejung Bang**[2]

[1] Department of Orthopedic Surgery, Seoul Sacred Heart General Hospital, Seoul, Korea

[2] Division of Biostatistics, Department of Public Health Sciences, University of California, Davis, USA

Multiple testing & 0.05
"The underlying mechanism of multiple testing may be well described as...“If you torture the data enough, nature will always confess.” Acceptable solutions are 1) to designate a single Primary hypothesis (and outcome/parameter/method), while all others are secondary/sensitivity/confirmatory; 2) to reveal all analyses performed (under a given aim in one publication)…”.

Three common misuses of p-values



Three common misuses of P values

Jeehyoung Kim[1] and **Heejung Bang**[2]

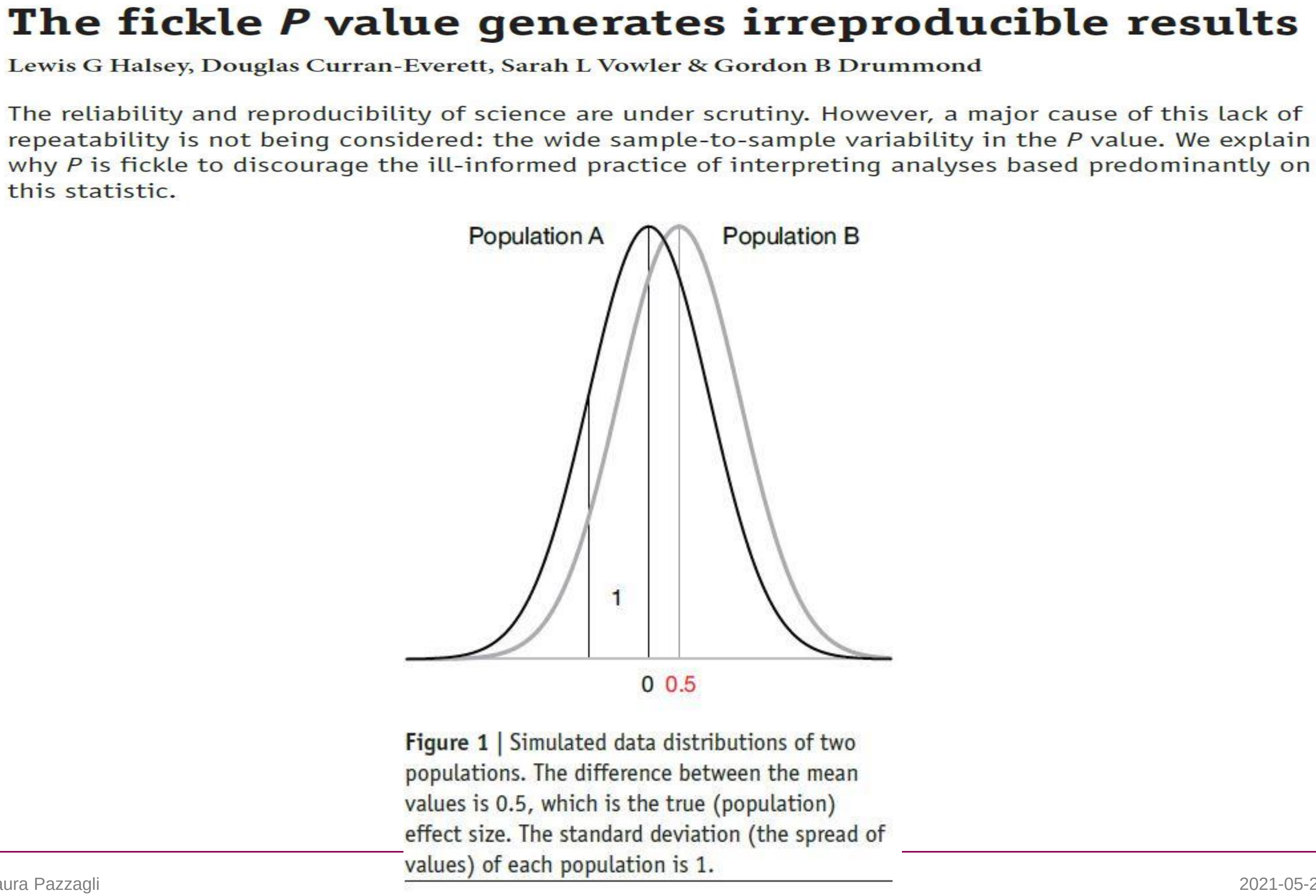
[1] Department of Orthopedic Surgery, Seoul Sacred Heart General Hospital, Seoul, Korea

[2] Division of Biostatistics, Department of Public Health Sciences, University of California, Davis, USA

Smaller p-value is more significant? Not necessarily

“We have discussed the well-known ‘large N→small p’ phenomenon. Below we illustrate that ‘smaller p-value, smaller effect’ can happen, when Ns are different. Another philosophical question may be: Which more strongly supports the effect, ‘a large effect size from a small sample’ vs. ‘a small effect size from a large sample’? The answer can vary and may be not straightforward; yet we are easily convinced that ‘sole reliance on p-values’ can be problematic.”

The fickle *P* value generates irreproducible results (Nature Methods 2015)



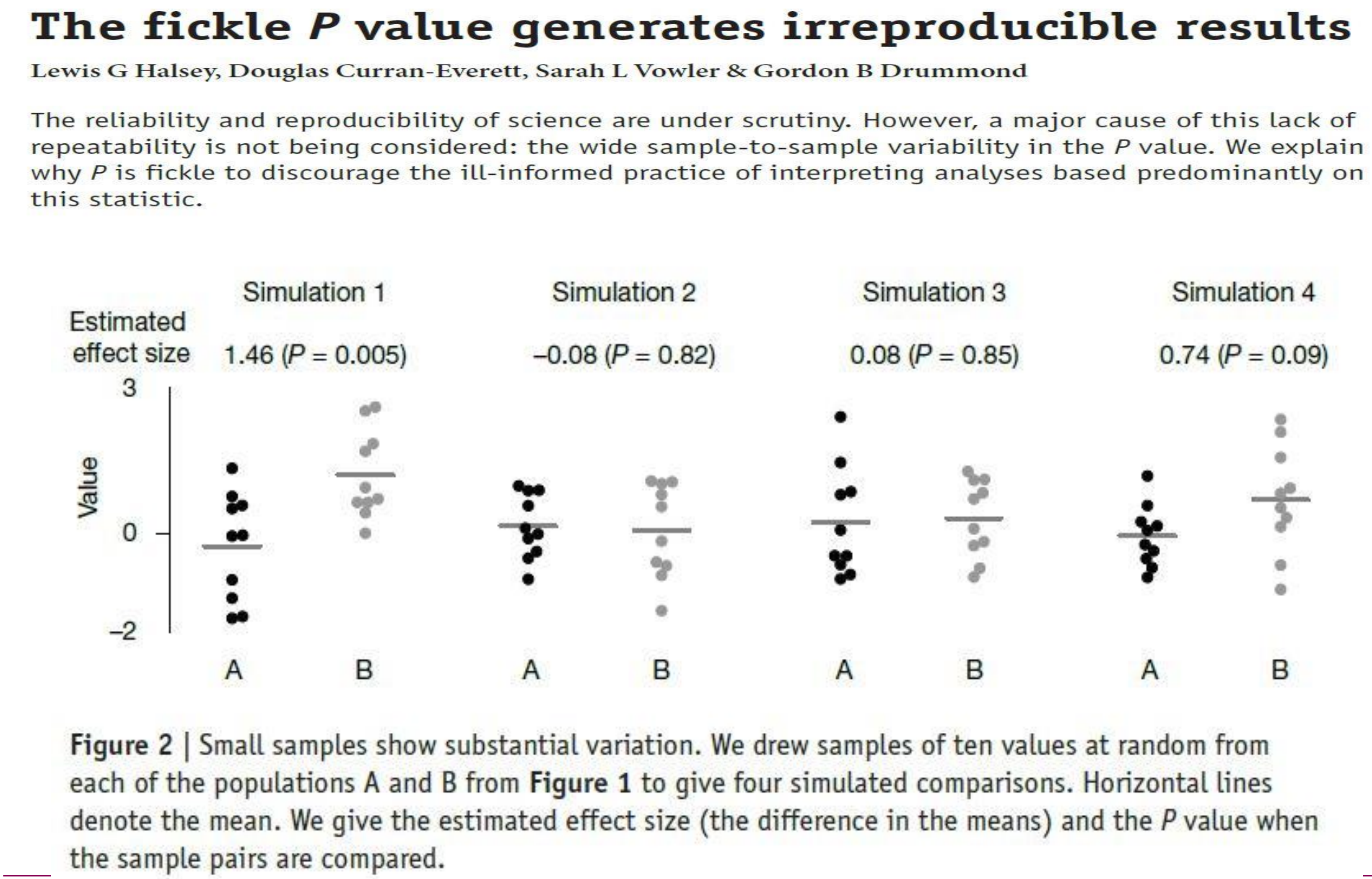
The fickle *P* value generates irreproducible results

Lewis G Halsey, Douglas Curran-Everett, Sarah L Vowler & Gordon B Drummond

The reliability and reproducibility of science are under scrutiny. However, a major cause of this lack of repeatability is not being considered: the wide sample-to-sample variability in the *P* value. We explain why *P* is fickle to discourage the ill-informed practice of interpreting analyses based predominantly on this statistic.

Figure 1 | Simulated data distributions of two populations. The difference between the mean values is 0.5, which is the true (population) effect size. The standard deviation (the spread of values) of each population is 1.

The fickle *P* value generates irreproducible results (Nature Methods 2015)



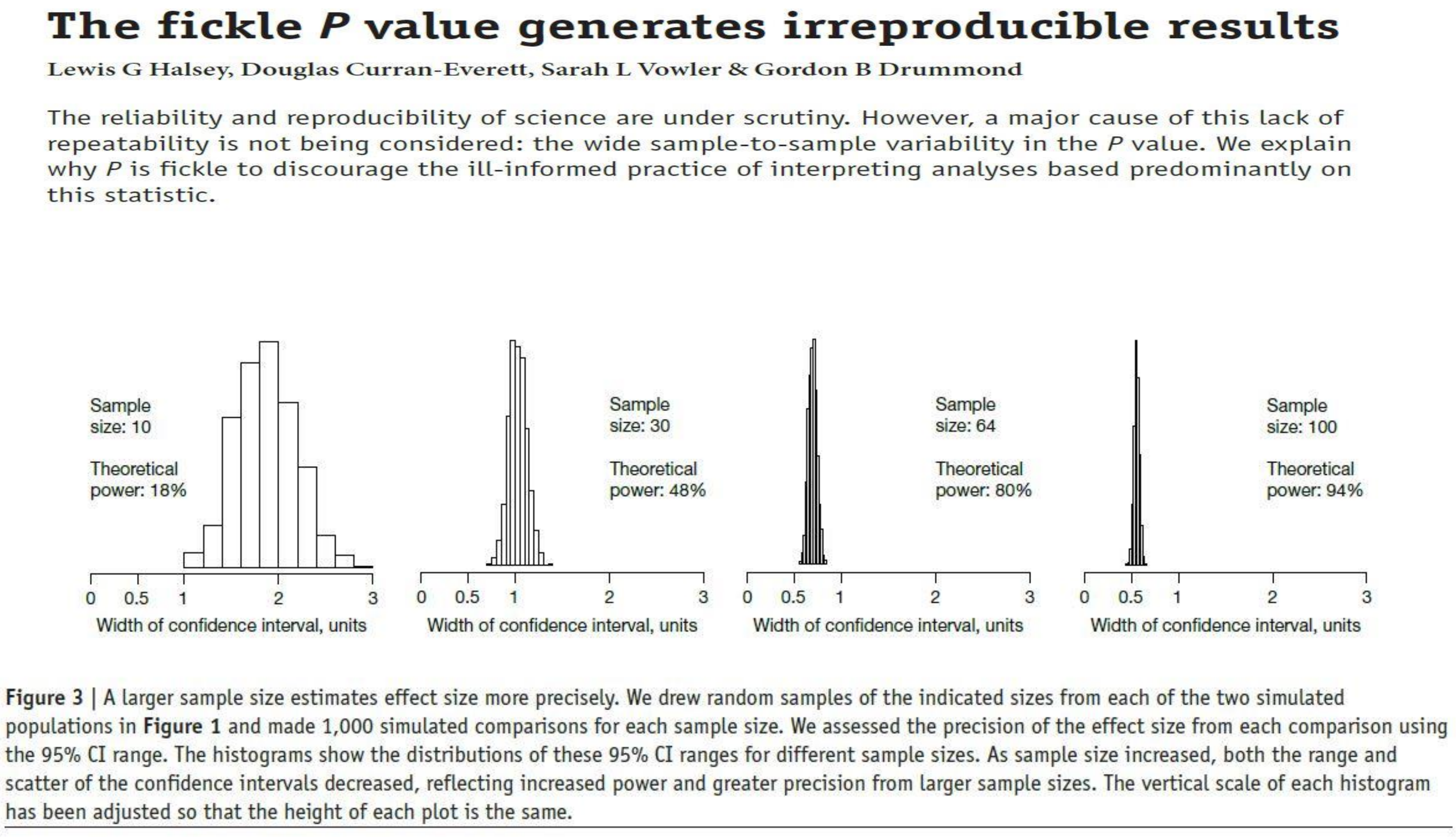
The fickle *P* value generates irreproducible results

Lewis G Halsey, Douglas Curran-Everett, Sarah L Vowler & Gordon B Drummond

The reliability and reproducibility of science are under scrutiny. However, a major cause of this lack of repeatability is not being considered: the wide sample-to-sample variability in the *P* value. We explain why *P* is fickle to discourage the ill-informed practice of interpreting analyses based predominantly on this statistic.

Figure 2 | Small samples show substantial variation. We drew samples of ten values at random from each of the populations A and B from **Figure 1** to give four simulated comparisons. Horizontal lines denote the mean. We give the estimated effect size (the difference in the means) and the *P* value when the sample pairs are compared.

The fickle *P* value generates irreproducible results (Nature Methods 2015)



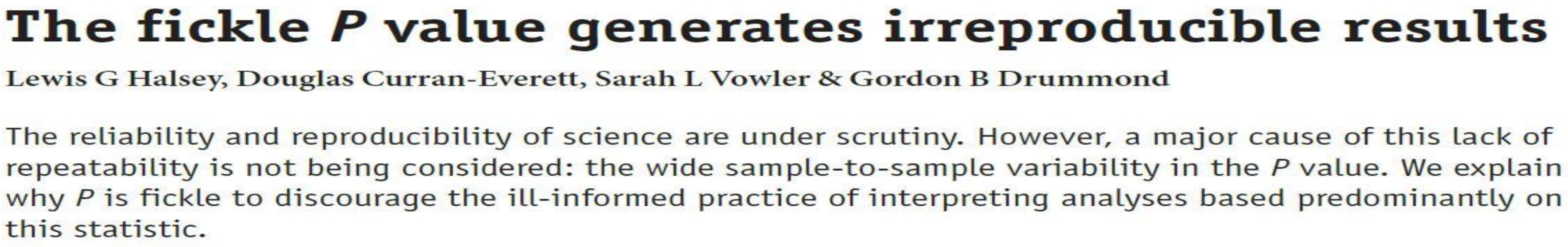
The fickle *P* value generates irreproducible results

Lewis G Halsey, Douglas Curran-Everett, Sarah L Vowler & Gordon B Drummond

The reliability and reproducibility of science are under scrutiny. However, a major cause of this lack of repeatability is not being considered: the wide sample-to-sample variability in the *P* value. We explain why *P* is fickle to discourage the ill-informed practice of interpreting analyses based predominantly on this statistic.

Figure 3 | A larger sample size estimates effect size more precisely. We drew random samples of the indicated sizes from each of the two simulated populations in **Figure 1** and made 1,000 simulated comparisons for each sample size. We assessed the precision of the effect size from each comparison using the 95% CI range. The histograms show the distributions of these 95% CI ranges for different sample sizes. As sample size increased, both the range and scatter of the confidence intervals decreased, reflecting increased power and greater precision from larger sample sizes. The vertical scale of each histogram has been adjusted so that the height of each plot is the same.

The fickle *P* value generates irreproducible results (Nature Methods 2015)



The fickle *P* value generates irreproducible results

Lewis G Halsey, Douglas Curran-Everett, Sarah L Vowler & Gordon B Drummond

The reliability and reproducibility of science are under scrutiny. However, a major cause of this lack of repeatability is not being considered: the wide sample-to-sample variability in the *P* value. We explain why *P* is fickle to discourage the ill-informed practice of interpreting analyses based predominantly on this statistic.

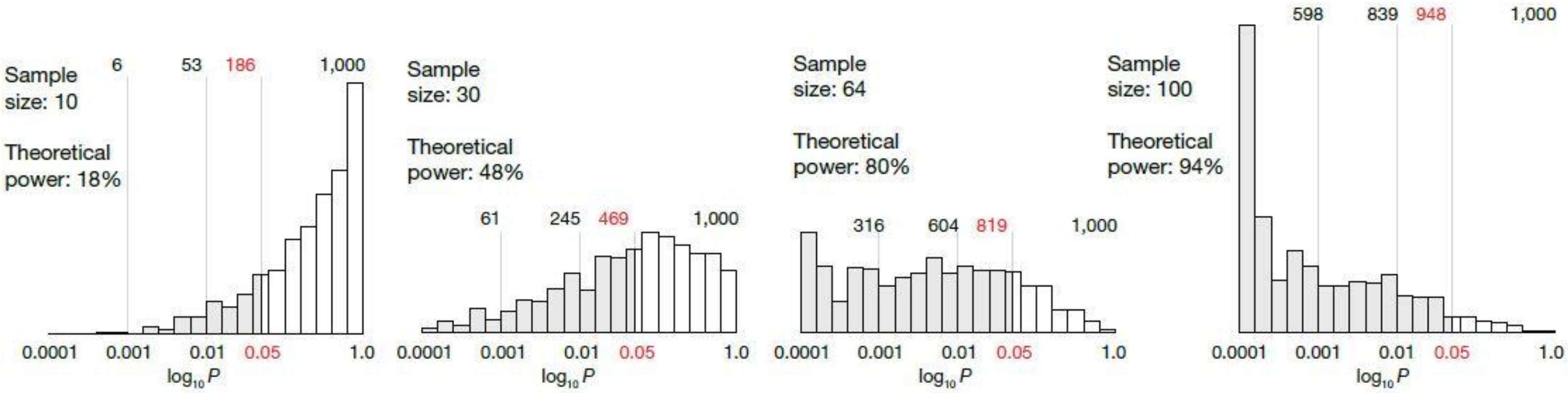


Figure 4 | Sample size affects the distribution of *P* values. We drew random samples of the indicated sizes from each of the two simulated populations in **Figure 1** and made 1,000 simulated comparisons with a two-sample *t*-test for each sample size. The distribution of *P* values is shown; it varies substantially depending on the sample size. Above each histogram we show the number of *P* values at or below 0.001, 0.01, 0.05 (red) and 1. The empirical power is the percentage of simulations in which the true difference of 0.5 is detected using a cutoff of $P < 0.05$. These broadly agree with the theoretical power.

Alternatives to the p-value

- An increasingly popular alternative to formal hypothesis and significance testing using p-values is the concept of inference based on probabilities of outcomes occurring, or of belonging to a certain group.
- In particular, Bayesian approaches which use pre-existing information within data to determine the likelihood of future possible outcomes are sometimes used in place of statistical significance testing.
- Rather than testing a research question and rejecting or not the null hypothesis, the fit of the model can be assessed instead, and information provided on whether the addition of more groups or variables to the model is improving the extent to which the model explains the data.

Alternatives to the p-value

- Values which can be used in these kinds of probabilistic methods include Akaike’s information criteria (AIC), formally presented by Hirotugu Akaike in 1974, and the Bayesian Information Criteria (BIC) developed by Gideon E. Schwarz in 1978.

- These values are data driven and penalize complexity and number of variables, and therefore aid in model construction as well as interpretation.

- The field of for example trajectory analysis uses a probabilistic approach to estimate expected outcomes over time for different groups of individuals, which are defined by the model rather than by a hypothesis.

Alternatives to the p-value

- An advantage to such methods is the ability to assess uncertainty as well as predicting outcomes, and so takes what could be described as a more holistic approach to data analysis relative to the use of p-values.

- A disadvantage is the reliance on the data itself, which can reduce generalizability, and interpretation can be more complicated when using these approaches.

- Recently the e-value became another popular approach to be used as sensitivity analysis.

- The e-value measures the amount of association between an unmeasured confounder and the exposure and outcome that would be necessary to completely cancel away the findings.

Conclusions

- Often, p-values are used as an easily interpretable indication of what results mean for the benefit of individuals who may have minimal or no statistical training.

- Use of the p-value as a simplification of whether an association exists is problematic for all the above mentioned reasons, in particular the intrinsic quality of the p-value reducing as the sample size increases is often overlooked.

- The allocation of research funding towards hot topic studies in which a new association is believed to have been found (for example the Wakefield study) is not only detrimental to scientific understanding, but it can be dangerous to human health.

Conclusions

- To address this, journals need to move away from the temptation to publish primarily ‘significant’ results which can be neatly packaged as answering a certain question conclusively.
- When p-value results are interpreted with care, and considering all the assumptions involved, they can provide a useful tool when addressing scientific questions.
- Discussion of the uncertainty involved in a study, and a move away from headline inducing studies with a focus on novelty and statistical significance is an important first step.

Conclusions

- What about all the other methodological issues related to a specific research question?
- Do statistically significant results equal unbiased results?
- More thinking about potential sources of bias is necessary when interpreting statistically significant results.
- Statistical significance does not correspond to clinical relevance.

Main references

1. Wasserstein, RL, and NA. Lazar. "The ASA statement on p-values: context, process, and purpose“. *The Amrican Statistician* (2016): 70: 129-133.

2. Altman, DG, and JM Bland. "Statistics notes: Absence of evidence is not evidence of absence“. *Bmj* (1995) 311: 485.

3. Mlinarić, A, M Horvat, and V Šupak Smolčić. “Dealing with the positive publication bias: Why you should really publish your negative results”. *Biochemia medica* (2017) 27(3): 447-452.

4. Eggertson, L. "Lancet retracts 12-year-old article linking autism to MMR vaccines." *Canadian Medical Association* (2010) 182(4):199-200.

5. Kim, J, and B Heejung. "Three common misuses of P values“. *Dental hypotheses* (2016) 7(3): 73.