

From the research bench

What is p-hacking?

P-hacking is trying data and analysis choices until a result is significant, then reporting that one. It can leave an excess of p-values just below 0.05.

Author	Published	Reading time	Words
Sinan Isoglu, MBA (Quantic) Commercial growth leader, lecturer and doctoral researcher	18 October 2026	16 min	3,490

Management summary

P-hacking is collecting or selecting data or statistical analyses until a nonsignificant result becomes significant, and reporting the version that crossed the threshold. This concept guide defines it (Head et al.; Munafò et al.), separates it from the file drawer problem, lists the analysis choices it works through, explains how a p-curve can detect it across many studies and what that test misses, reports a text-mining study that found evidence of it in p-values pooled across disciplines but judged its effect on the conclusions of meta-analyses probably weak, shows why an estimate selected for its significance tends to overstate the effect (Gelman and Carlin), and gives a decision record for a commercial analytics team. The sources are one text-mining study with a reanalysis of twelve meta-analyses, one methods article, one analytical model and one perspective.

Contents	§ 1	What is p-hacking?	2
	§ 2	Which analysis choices can turn into p-hacking?	3
	§ 3	How does a p-curve reveal p-hacking, and what does it miss?	4
	§ 4	How widespread did p-hacking look in published research?	5
	Figure 1	P-values text-mined from Results sections, in the two bins just below 0.05	6
	§ 5	Does p-hacking change what a body of evidence concludes?	7
	§ 6	Why does a result chosen for its significance overstate the effect?	8
	§ 7	How can an analytics team spot p-hacking in its own reports?	9
	Figure 2	An analysis decision record	10
	§ 8	What do these sources leave open for business analytics?	11
		References	12

What is p-hacking?

§ 1 of 8

It is adjusting how data are collected or analysed until a result crosses the significance threshold, and then reporting that version. Head et al. (2015)² define it in one sentence: “One type of bias, known as “p-hacking,” occurs when researchers collect or select data or statistical analyses until nonsignificant results become significant.” They also call it inflation bias or selective reporting, and say it “is the misreporting of true effect sizes in published studies”: researchers “try out several statistical analyses and/or data eligibility specifications” and then “selectively report those that produce significant results”.

² Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, 13(3), e1002106.

Munafò et al. (2017)⁴ describe the same thing from the analyst’s side. Choices of models, exclusion rules and covariates can be made “with knowledge of the observed data”, and “Such data-contingent analysis decisions constitute what has become known as P-hacking”. They also call it “data dredging (also known as P-hacking)”, and they note that a researcher who has measured several possible outcomes may, once the results are known, “intentionally or unintentionally select the subset of outcomes that show statistically significant results as the outcomes of interest”.

⁴ Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1(1), 0021.

P-hacking is not the same as the file drawer problem, and the two sources draw that line with different words. Head et al.² name “two widely recognized types of researcher-driven publication bias”: selection, the file drawer effect, in which studies with nonsignificant results are published less often, and inflation, which is p-hacking. Munafò et al.⁴ use the term publication bias for the first type alone and describe the file drawer problem as “the fact that many more studies are conducted than published”. On either usage, the file drawer works on whole studies; p-hacking works inside a study, on which analysis is shown.

In this guide, p-hacking means making data or analysis decisions after seeing how they change the p-value, and reporting the version that crossed the threshold without the versions that did not. It covers when to stop collecting data, which outcome to report, which observations and groups to keep and which covariates to include. With Head et al.²,

it also covers reporting a p-value above 0.05 as below it, which they count as part of p-hacking. It is distinct from the file drawer problem, where whole studies stay unpublished, from multiple testing that is reported in full and corrected, which the [statistical conclusion validity guide](#) covers together with what a p value can and cannot carry, and from exploratory analysis that is labelled as exploratory. Intent is not part of the definition: on Munafò et al.⁴'s account, the selection can be unintentional.

Which analysis choices can turn into p-hacking? § 2 of 8

The examples Head et al.² list are running analyses partway through an experiment to decide whether to keep collecting data, recording many response variables and deciding after the analysis which to report, deciding afterwards whether to keep or drop outliers, excluding, combining or splitting treatment groups after the analysis, adding or removing covariates after the analysis, and stopping the search once an analysis comes out significant. Ioannidis (2005)³, describing a hypothetical genetic association study, names “serendipitous inclusion or exclusion of certain patients or controls, post hoc subgroup analyses”, contrasts that were not specified at the start, changed definitions of the disease or the controls, and selective or distorted reporting. He defines bias as “the combination of various design, data, analysis, and presentation factors that tend to produce research findings when they should not be produced”, and adds: “Selective or distorted reporting is a typical form of such bias.”

³ Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124.

Munafò et al.⁴ add outcome switching, the possibility of changing the outcomes of interest “in the study depending on the observed results”, and they point to analytic flexibility: in a dataset with many variables “there may be hundreds or thousands of reasonable alternative approaches to analysing the same data”. In one study they cite, “applying almost 7,000 analytical pipelines to a single fMRI dataset resulted in over 90% of brain voxels showing significant” activation in at least one analysis. Ioannidis turns flexibility into one of the corollaries of his model: the more flexible the designs, definitions, outcomes and analyses in a field, the less likely its findings are to be true, because “Flexibility increases the potential for transforming what would be “negative” results into “positive” results”. That corollary is derived from the model, not counted in data.

In a commercial analytics team the same choices can go by other names: a readout checked every week and closed when it looks good, one metric picked for the slide from a dashboard of many, a date window or an attribution window chosen afterwards, a large account dropped as an outlier, segments cut until one shows a difference, a control variable added or removed. Each is a legitimate decision when it is fixed in advance or reported with its alternatives. It becomes p-hacking when the choice follows the p-value and the alternatives disappear from the report. The [A/B test guide](#) covers two of these inside online experiments, early stopping and segment cuts made after the test. This mapping is the author's and has not been tested.

How does a p-curve reveal p-hacking, and what does it miss? § 3 of 8

By its shape near 0.05. In Head et al.²'s words, "A p-curve is the distribution of p-values for a set of studies." With no true effect, "every p-value is equally likely to be observed", so the curve is flat; with a true effect, the distribution "is exponential with a right skew", and small p-values are more common than p-values close to 0.05. If researchers turn nonsignificant results into significant ones, the curve changes close to the threshold, and "a p-hacked p-curve will have an overabundance of p-values just below 0.05". That is what separates p-hacking from the file drawer: "Both p-hacking and selective publication bias predict a discontinuity in the p-curve around 0.05, but only p-hacking predicts an overabundance of p-values just below 0.05".

Their test compares two narrow bins of width 0.005, the lower from 0.04 to 0.045 and the upper from 0.045 to 0.05. A one-tailed sign test asks whether the upper bin holds more p-values than the lower one. Without p-hacking, the authors expect the distribution there to be "uniform close to 0.05 (if the true effect sizes are zero), or right skewed", so the upper bin should not hold more. The authors state the limits. Broader tests of the curve's skew "can detect severe p-hacking, but are insensitive to more modest (and arguably more realistic) levels of p-hacking", in particular when true effects are strong, "as the right skew introduced to the p-curve will mask the left skew caused by p-hacking". Their test on the two bins just below 0.05 is the more sensitive one, and they still call it "a simple, and conservative, test for p-hacking". The shape of the curve also depends "on both the true effect (i.e., the p-curve before p-hacking) and the intensity of p-hacking". On this guide's reading, a curve without the excess therefore does not show that no one p-hacked.

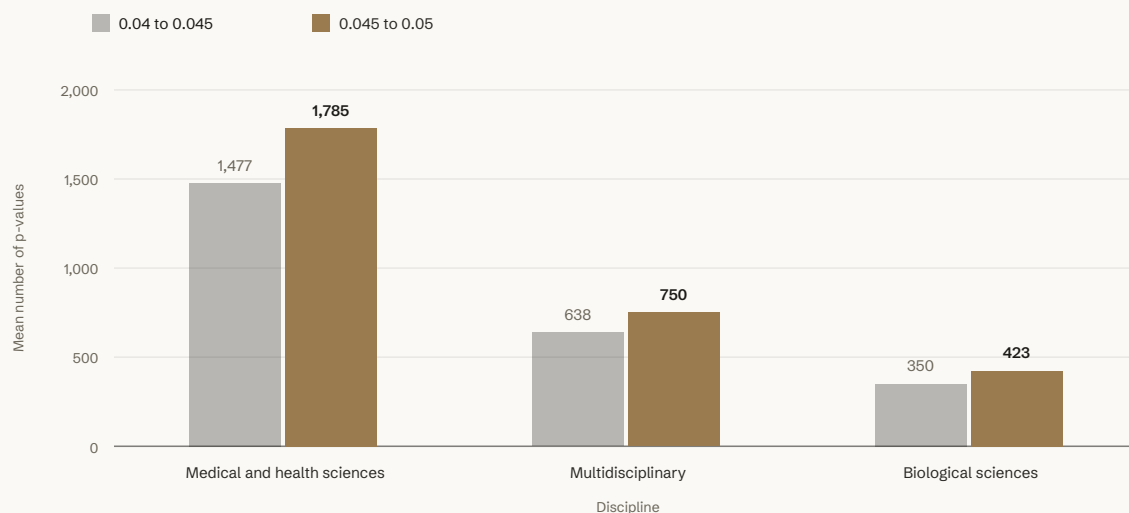
How widespread did p-hacking look in published research?

§ 4 of 8

Widespread, on Head et al.²'s evidence, though their test can only see part of it. They “used text-mining to search for p-values in all Open Access papers available in the PubMed database” and built curves by discipline, separately for p-values in Results sections and in Abstracts. Pooled across disciplines, both sets had a strong right skew, which they read as evidence that researchers mostly study effects that are not zero. The pooled data also showed “strong evidence for p-hacking”: in Results sections the estimated share of p-values in the upper bin was 0.546 (lower confidence limit 0.536) across 14 disciplines, where equal counts in the two bins would give 0.5, and in Abstracts 0.537 (0.518) across 10. “In most disciplines, there were more p-values in the upper than the lower bin.” In the Results sections the difference was significant in every discipline where the authors had good statistical power, which they name as medical and health sciences, biological sciences and the multidisciplinary category; Figure 1 shows those three. In the Abstracts only the multidisciplinary and the information and computing sciences categories were significant.

Figure 1 P-values text-mined from Results sections, in the two bins just below 0.05

In each of the three disciplines with the most p-values, the bin just below 0.05 held more p-values than the bin below it, which the authors do not expect without p-hacking.



The one-tailed binomial test for p-hacking gave $p < 0.001$ for medical and health sciences, 0.001 for the multidisciplinary category and 0.005 for biological sciences. Of the 14 disciplines in Table 1, nine had more p-values in the upper bin than in the lower one; the test was also significant at 5 per cent for psychology and cognitive sciences (29 against 50, $p = 0.012$), which is not drawn.

Source Head et al. (2015), p. 7, Table 1: mean number of p-values from Results sections of open access papers in PubMed, based on 1,000 bootstraps of one p-value per Results section and rounded to whole numbers, in the bins 0.04 to 0.045 and 0.045 to 0.05, for the three disciplines the authors name as having good statistical power (p. 8). One-tailed binomial test for p-hacking from the same table.

Their reading is direct: “Our text-mining suggests that p-hacking is widespread.” They add two qualifications. There was “still a strong right skew in all the p-curves we examined”, which they find reassuring given the concern that many published results are false, and because text-mining pools “many different types of questions”, it remains unclear whether some fields or questions within these disciplines study effects that average zero. They also report that an earlier text-mining study of five medical journals found no evidence of p-hacking, and that it was criticised for using p-values from Abstracts, which are more likely to contain only the strongest results.

Does p-hacking change what a body of evidence concludes?

§ 5 of 8

Probably not by much in the one area Head et al.² tested, though it may inflate the estimates. They repeated the test on p-values from studies used in twelve meta-analyses “by evolutionary biologists studying sexual selection”. Nine of the twelve curves showed significant evidential value, that is, evidence that the effect studied is not zero. Counting misreported p-values in the upper bin (16 values given as below 0.05 that were in fact larger), seven of the twelve curves had more p-values there than in the lower bin, and the difference was significant in one, the dataset with the largest sample. “However, the evidence for p-hacking disappeared when we excluded misreported p-values from our analyses.” Across all twelve, the share in the upper bin was 0.615 (lower limit 0.513, $p = 0.033$) with the misreported values and 0.489 (0.375, $p = 0.443$) without. The authors keep the misreported values in, because reporting nonsignificant results as below 0.05 “is a component of p-hacking that should not be ignored”.

Their judgement on the consequences is hedged in both directions. P-hacking is “probably common”, but “its effect seems to be weak relative to the real effect sizes being measured”, and this result, they write, “suggests that p-hacking probably does not drastically alter scientific consensuses drawn from meta-analyses”. That one meta-analysis with a significant indication of p-hacking also showed strong evidential value, and its p-values between 0.045 and 0.05 “were only a small proportion of published significant p-values”; of it they write: “It is therefore unlikely that p-hacking would change the qualitative conclusions made in this meta-analysis, although p-hacking might have inflated the estimated mean effect size.” They give two reasons why a meta-analysis may withstand it: all else being equal, the small studies most open to p-hacking receive less weight, and at least in some fields meta-analyses often use data not directly related to the original paper’s main question, and “The p-values associated with secondary questions are less likely to be p-hacked.” Their summary keeps both halves: the evidence supports mean effects that are not zero, but the “estimated mean effect size has probably been inflated by p-hacking”.

Ioannidis (2005)³ puts the same mechanism into a model. In it, bias lowers the probability that a claimed finding is true: “Thus, with increasing bias, the chances that a research finding is true diminish considerably.” That is a result of the model under assumed inputs, not a count of false findings. He also notes that a reader cannot see from

one report how much testing came before it: “usually it is impossible to decipher how much data dredging by the reporting authors or other research teams has preceded a reported research finding”.

Why does a result chosen for its significance overstate the effect?

§ 6 of 8

Because when a study is small relative to the true effect, only large estimates cross the threshold. Gelman and Carlin (2014)¹ give this a number, the exaggeration ratio: “the expectation of the absolute value of the estimate divided by the effect size, if statistically significantly different from zero”. Next to it they put the Type S error rate, the chance that a significant estimate has the wrong sign. For unbiased, normally distributed estimates, both are small at 80 per cent power, with “an expected exaggeration factor of 1.12”. For such estimates, problems with the exaggeration ratio “start to arise when power is” below 0.5, and “problems with the Type S error rate start to arise when power is less than 0.1”.

¹ Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science*, 9(6), 641–651.

Their first example is a published finding from a sample of 2,972 respondents to a longitudinal survey, in which “Of the first-born children of the parents in the most attractive category, 56% were girls, compared with 48% in the other groups”, a difference of 8 percentage points with $p = 0.015$. From the literature on sex ratios at birth, they hypothesise true differences of 0.1, 0.3 and 1.0 percentage points. At 0.1 points a significant result would have “a 46% chance it will have the wrong sign”, and “in expectation the estimated effect will be 77 times too high”; at 0.3 points the chance of the wrong sign is roughly 40 per cent and the overestimate a factor of 25; at 1.0 point, “A statistically significant finding in this case has a 19% chance of appearing with the wrong sign, and the magnitude of the true effect would be overestimated by an expected factor of 8.” In their second example, a reported 17-point difference in voting preferences between women at different points of the menstrual cycle, they set the true effect at 2 points, which they consider the upper end of what is plausible; that gives power of 0.06, a 24 per cent chance of the wrong sign and an expected exaggeration of 9.7.

Their general point is that “Using statistical significance as a screener can lead researchers to drastically overestimate the magnitude of an effect”. In the first example they “ignore multiple comparisons issues and take the published claim of statistical significance at face value”,

and in the second they take “the published comparison and p value at face value”, so the exaggeration they compute comes from the significance filter alone. On this guide’s reading, p-hacking adds to it, because every extra version of the analysis gives a noisy estimate another chance to cross the threshold; Gelman and Carlin¹ make a related point when they write that multiple comparisons or researcher degrees of freedom “make it easy for researchers to find large and statistically significant effects that could arise from noise alone”.

Their remedy is a design analysis based on an effect size “determined from literature review or other information external to the data at hand”. Earlier estimates are a weak guide, because they “will themselves tend to be overestimates”, to the extent that the published literature “selects on statistical significance”; “using a published point estimate is not enough”. And “A design calculation requires an assumed effect size and adds nothing to an existing data analysis if the postulated effect size is estimated from the very same data.” They are open that the input is a judgement: “We recognize that any assumption of effect sizes is just that, an assumption.”

How can an analytics team spot p-hacking in its own reports?

§ 7 of 8

Not from the p-values. A p-curve needs many independent results, and a team with a handful of analyses on one question cannot draw one, so its check is the record of decisions, not the shape of a distribution. Head et al.² recommend that researchers “Clearly label research as prespecified” or exploratory “so that readers can treat results with appropriate caution” and “Perform data analysis blind wherever possible”, and that journals, in their reporting guidelines, require authors to be explicit “about the entire analysis process (not just the final tests used to generate reported p-values)”. Munafò et al.⁴ describe one form of blinding, in which, during data preparation and cleaning, the identity of conditions or the variable labels “can be masked so that the output is not interpretable in terms of the research hypothesis”, and they argue that writing analysis decisions down “prior to observing the data” keeps them “data-independent”; the [reproducibility guide](#) covers preregistration. Figure 2 turns this into one record per reported result and adds Gelman and Carlin’s¹ check on the plausible size of the effect.

Figure 2 An analysis decision record

Write down the result, the decisions behind it, when they were fixed, every version run, the plausible effect and the label before the result goes into a report.

Result	Decisions	Fixed	Versions	Outside size	Label
Which estimate goes in the report?	Stop, outcome, exclusions, groups, covariates?	Before or after seeing results?	How many run, and are all shown?	Plausible effect from other data?	Prespecified or exploratory?

If a decision was fixed after seeing results, show every version run and label the result exploratory.

One row per reported result. The outside size comes from other data, not from the estimate itself. An estimate far above it, from a small analysis, is likely to overstate the effect.

Source Author's decision worksheet. The decisions field follows the practices listed in Head et al. (2015), p. 1, and Munafò et al. (2017), p. 3; the fixed field follows Munafò et al. (2017), p. 3; the versions and label fields follow Head et al. (2015), p. 12; the outside size field follows Gelman and Carlin (2014), pp. 642 and 648. The fields are reader inputs; no company data is supplied.

An illustration, with no company data: a quarterly review asks whether win rates differed after a change to the discount policy. The planned comparison, the win rate across all deals before and after the change, shows no significant difference. The analyst then drops deals below a size threshold, splits the rest by segment and switches the outcome from win rate to average deal size; one of twelve versions, deal size in the mid-market segment, comes out at $p = 0.03$. The record shows the planned result with its p -value, names the three decisions made after the first result, counts the twelve versions and lists them all, states as the outside size the change in deal size seen after an earlier policy change, if the team has one that was not itself picked for its significance, and labels the mid-market figure exploratory: a question for a planned comparison, not a finding. The record does not make a before-and-after comparison causal; it shows how far the reported

figure was chosen. These are the author's examples and have not been tested.

What do these sources leave open for business analytics? § 8 of 8

How often it happens there, and whether a record like Figure 2 reduces it. Head et al.²'s evidence comes from p-values in open access papers indexed in PubMed and from twelve meta-analyses in one area of evolutionary biology; their test is conservative and pools many questions; in the meta-analyses, whether p-hacking shows at all depends on whether misreported p-values count, and their judgement that its consequences are weak rests on those twelve. Gelman and Carlin's¹ numbers are design calculations under assumed true effects for two published studies, whose published significance they take at face value. Ioannidis's³ essay is a model with assumed inputs, not a measurement. Munafò et al.⁴ propose measures and write that "There is some evidence from both simulations and empirical studies supporting the likely effectiveness of these measures", and that their broad adoption "will require iterative evaluation and improvement". None of the four measured p-hacking in company analytics.

On this guide's reading, two things carry over to business analytics with some confidence: when an analysis is small relative to the plausible effect, an estimate kept because it crossed the threshold is likely to overstate the effect, and a single report cannot be checked for p-hacking from its own p-values, so the check is the record of decisions. Whether such a record changes what analysts report, the sources do not test.

 End of essay

References

- 1 Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science*, 9(6), 641-651. <https://doi.org/10.1177/1745691614551642>
- 2 Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, 13(3), e1002106. <https://doi.org/10.1371/journal.pbio.1002106>
- 3 Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124. <https://doi.org/10.1371/journal.pmed.0020124>
- 4 Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1(1), 0021. <https://doi.org/10.1038/s41562-016-0021>

How to cite this essay

Isoglu, S. (2026, October 18). What is p-hacking? Sinan Isoglu.
<https://isoglu.com/insights/journal/what-is-p-hacking/>
Online version: <https://isoglu.com/insights/journal/what-is-p-hacking/>

Sinan Isoglu, MBA (Quantic)

Commercial growth leader, lecturer and doctoral researcher
© 2026 Sinan Isoglu · isoglu.com · All rights reserved