Is N-Hacking Ever OK? A simulation-based study

Pamela Reinagel
Section of Neurobiology, Division of Biological Science, University of California, San Diego.

Abstract

After an experiment has been completed and analyzed, a trend may be observed that is "not quite significant". Sometimes in this situation, researchers incrementally grow their sample size N in an effort to achieve statistical significance. This is especially tempting in situations when samples are very costly or time-consuming to collect, such that collecting an entirely new sample larger than N (the statistically sanctioned alternative) would be prohibitive. Such post-hoc sampling or "N-hacking" is condemned, however, because it leads to an excess of false positive results. Here Monte-Carlo simulations are used to show why and how incremental sampling causes false positives, but also to challenge the claim that it necessarily produces alarmingly high false positive rates. In a parameter regime that would be representative of practice in many research fields, simulations show that the inflation of the false positive rate is modest and easily bounded. But the effect on false positive rate is only half the story. What many researchers really want to know is the effect N-hacking would have on the likelihood that a positive result is a real effect that will be replicable. This question has not been considered in the reproducibility literature. The answer depends on the effect size and the prior probability of an effect. Although in practice these values are not known, simulations show that for a wide range of values, the positive predictive value (PPV) of results obtained by N-hacking is in fact *higher* than that of non-incremented experiments of the same sample size and statistical power. This is because the increase in false positives is more than offset by the increase in true positives. Therefore in many situations, adding a few samples to shore up a nearly-significant result is in fact *statistically beneficial*. It is true that uncorrected N-hacking elevates false positives, but in some common situations this does not reduce PPV, which has not been shown previously. In conclusion, if samples are added after an initial hypothesis test this should be disclosed, and if a false positive rate is stated it should be corrected. But, contrary to widespread belief, collecting additional samples to resolve a borderline P value is not invalid, and can confer previously unappreciated advantages for efficiency and positive predictive value.

Background

There has been much concern in recent years concerning the lack of reproducibility of results in some scientific literatures. The call for improved education in statistics and greater transparency in reporting is justified and welcome. But if we apply overly-conservative rules dogmatically, we as a community risk throwing out a lot of babies (good data, promising leads) with the statistical bath water. Experiments in biology and psychology often require substantial financial resources, scientific talent, and use of animal and/or human subjects. There is an ethical imperative to use these resources efficiently. To ensure both reproducibility and efficiency of research, experimentalists need to understand statistical issues rather than blindly apply rules.

The rule in question is a cornerstone of null hypothesis significance testing: sample exactly the predetermined sample size N, and then accept the verdict the hypothesis test, whatever it is. Adding a few samples and retesting after a negative result would be invalid on a standard null hypothesis statistical testing (NHST) framework, and can produce misleading outcomes. But this depends on the parameter regime in which one is operating; what researchers need to know is what can occur in their operating regime.

Empirical scientists are accustomed to looking at data, so simulation is an excellent way to gain intuitions about the implications of statistical methods. Here I simulate the denounced practice of "N-hacking" – incrementally adding more samples after the fact whenever a preliminary result is "almost significant". The simulations demonstrate the known effect that post-hoc sample growth of this kind elevates the false positive rate, and show why this is the case. After exploring a broad range of assumptions bracketing common practice in many fields of research, however, it emerges that the elevation in false positive rate is quite modest, and it becomes apparent that it could readily be corrected for. Moreover, additional simulations show that there is a truth underlying researchers' intuition that growing the sample size is a good idea. The purpose of this article is not to dismiss concerns about sampling procedures, but rather to demonstrate that there are better options than either starting over from scratch or abandoning a hypothesis after obtaining a nearly-significant outcome.

Results

These simulations can be taken to represent a large number of independent studies, each collecting separate samples to test a different hypothesis. I assume that a significance criterion α has been set in advance, and the sample size would be increased *only* for those tests that meet a policy of "P close to α". I further assume that the maximum number of samples the study could or would add is no more than a few times greater than the original sample size, or a few hundred total samples. All simulations were performed in Matlab 2018a.

Effect of incrementally growing sample size on the false positive rate

Experiments were simulated by drawing two independent samples of size N from the same normal distribution. An independent sample t-test was then used to accept or reject the null hypothesis that the samples came from distributions with the same mean. Because the samples always came from the same distribution, any positive result is a false positive. By definition, the t-test produces false positives at a rate of exactly α, the significance threshold, regardless of the mean or standard deviation of the source distribution or the sample size N.

I defined a plausible Asymmetric N-increasing policy as follows: every time a comparison yielded a P value that was "almost significant", additional sample points were added incrementally to the sample, and the t-test repeated. This was iterated until the P value was either significant, or no longer close, or the maximum number of samples was reached. The definition of "almost significant" was: $\alpha \leq P < (1+w)\,\alpha$, where $0 < w \leq 1$. For example if $\alpha = 0.05$ and $w = 0.2$, one would accept a hypothesis if $P < 0.05$, reject if $P > 0.06$, and add samples for P values in between. This would be representative of conditions under which I have seen researchers increment sample size in the fields of biology in which I have worked.

Results of such a policy are shown in Figure 1, assuming an initial sample size of $N_{init} = 12$, incremental sample growth of $N_{incr} = 6$, and maximum sample size of $N_{max} = 24$. For every choice of w and α, $M = 10^6$ independent experiments were simulated. This is meant to represent 10^6 separate studies, each using this policy to test only one hypothesis.

As expected, this Asymmetric N-increasing policy yielded an increase in false positives, which was more severe as w increased (Fig. 1a). Nevertheless the overall elevation in false positives was rather modest. For example with a policy of $\alpha = 0.05$ and $w = 1$, sample size was grown whenever P was between $0.05 - 0.10$, resulting in a realized false positive rate $FP = 0.0625$ instead of the nominal 0.05. Following this policy resulted in a negligible increase in the sample size on average (Fig. 1b). Note that no multiple comparison correction was done within study for the interim retesting on the policy; instead the false positives due to multiple comparisons are included in the reported false positive rates, i.e. these are the *uncorrected* false positive rates.

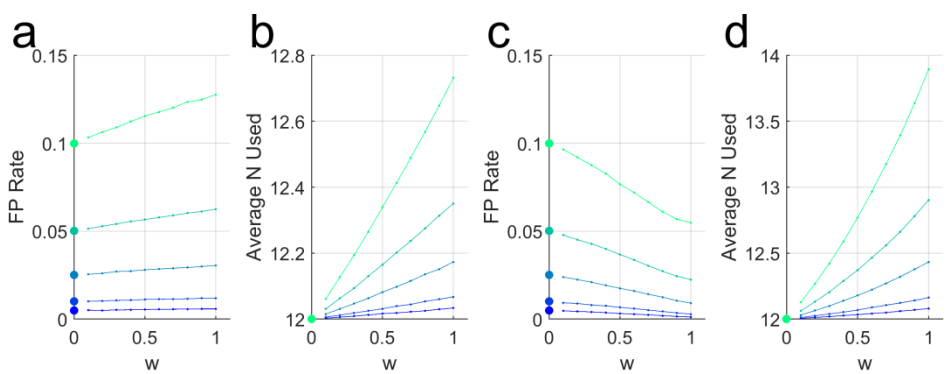

Figure 1. Effect of selective sample-increasing on false positive rate. Results shown are for M=10^6 experiments with initial sample size N_{init}=12, and sample increments N_{incr}=6, and maximum sample size N_{max}=24. **(a)** For any choice of α (0.005, 0.01, 0.025, 0.0500, 0.1, colors), the Asymmetric N-increasing policy yields an increase in false positives (FP Rate) which grows with the decision window w. The case of ($w = 0$) is identical to the standard pre-determined N policy and yields FP at a rate of α (solid symbols). **(b)** Average N in the final sample using the Asymmetric policy, as a function of α and w. **(c)** The Symmetrical N-increasing policy yields a net decrease in false positives, which grows with w. **(d)** Average N in the final sample using the Symmetric policy.

The main reason false positives are elevated by this policy is that the experiments that were incremented and retested were chosen in a biased way. By selectively incrementing only the subset of true negatives in which the difference between experimental and control groups was rather large, and thus nearly significant, even a small difference between groups in the added samples would be sufficient to push the overall group difference over the threshold for significance, purely by chance.

This is a problem because the policy is asymmetric: N was incremented when P was just above threshold, but not when it was just below threshold. To demonstrate this point, I simulated a Symmetric N-increasing policy, in which incremental sample growth occurred whenever a P value was either just below or just above α: $(1 - w)\,\alpha \leq P \;<\; (1 + w)\,\alpha$. It turns out that a symmetric policy more than overcomes the problem – it would convert more false positives to true negatives than it converts true negatives to false positives, resulting in a

net reduction in false positives (Fig. 1c). This is because in addition to the effect noted above, this policy also incremented the sample size in a biased subset of the false positives: ones in which the difference between experimental and control groups was rather small and thus barely significant. The Symmetric policy resulted in a slightly larger final sample size on average (Fig. 1d). In discussions of statistical malpractice, it is often asserted that an experimentalist would never add more samples after obtaining a significant P value, but interestingly there is evidence that they do (*1*). Therefore the consequences of both policies will be explored further below.

Dependence of false positive rate on the parameters α and *w*

For the Asymmetric N-increasing policy, analysis of the simulated data reveals that for any given choice of w, the false positive rate depends linearly on α: $FP = k\alpha$ (Fig 2a). The slopes of these lines are in turn an increasing function of the decision window w (Fig 2b, symbols). On the Symmetric policy, the dependence of FP on α is also linear (Fig 2c) and the slope k declines with w (Fig 2d).

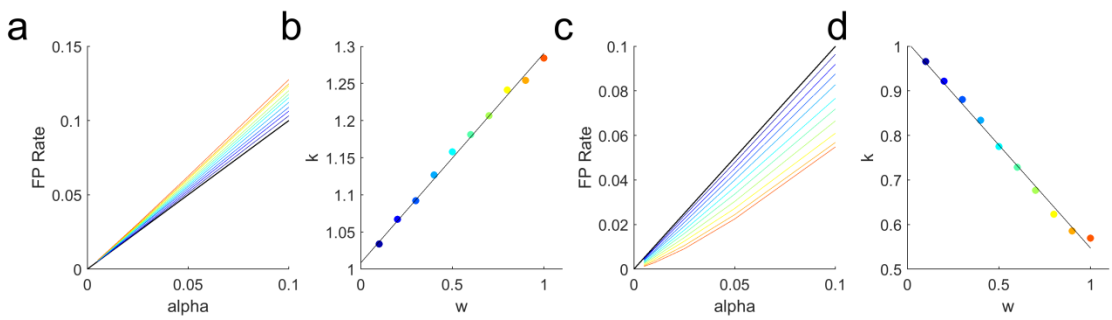

Figure 2. Dependence of FP rate on α and w. Simulations with $N_{init} = 12$, $N_{incr} = 6$, $N_{max} = 24$. **(a)** Asymmetrical N-increasing policy. Color indicates w (cf. panel b). For each value of w, the realized false positive rate is plotted for each value of α and these data points connected (not fit). The identity line (black) is the FP rate of the standard Fixed-N policy. **(b)** The slopes k from linear fits to the data shown in (a), plotted as a function of w (colored symbols). The dependence of the slope k on w is not linear in general, but was approximately linear in this parameter range (linear fit, black). **(c)** Like (a) but for the Symmetrical N-increasing policy. Note that for $w > 0.5$, FP rates are biased upward at larger values of α, due to the imposed cap of $2N$ additional samples. Therefore to determine slopes, lines for each w were fit using only values of α for which this cap did not impact any simulation. **(d)** The slopes k from the linear fits to the data shown in (c), for all values of w for which a slope could be estimated (*i.e.* at least two values of α were not impacted by sampling cap). The dependence of the slope k on w is not linear in general, but was approximately linear in this parameter regime (linear fit, black).

Dependence on N_{init} and N_{incr}

These results depend quantitatively on the choice of N_{init}, N_{incr}, and N_{max}. Additional simulations below explore values of N_{init} ranging from 5–160 initial sample points, incremental sampling N_{incr} ranging from 1 to N_{init} per increment, capping the maximum total sample size at either$N_{max} = 256$ or $N_{max} = 5N_{init}$. These assumptions more than bracket the range of realistic sample sizes and ad-hoc sample growth that would be commonly used in many experimental research fields. These simulations tested $\alpha \leq 0.05$ and $w \leq 0.4$. For each combination of N_{init}, N_{incr}, α and w, $M = 10^6$ experiments were simulated to estimate the fraction yielding false positive results.

The simulations show that initial sample size and incremental sample size are important. Results for the Asymmetric policy with $\alpha = 0.05, w = 0.4$ are shown in Figure 3a. The false positive rate is always elevated compared to α (black line), and this is more severe when the intial sample size is larger (curves slope upward) or the incremental sample growth is smaller (cooler curves are higher). Nevertheless the false positive rate doesn't exceed 0.06 for any condition. In this range of parameters, the dependence of k on w was approximately linear, so one can visualize the results for all combinations of of α and w on the same scale by plotting them as $(\frac{FP}{\alpha} - 1)/w$ (Fig 3b). On this scale, positive values indicate an increase in the FP rate compared to α, and negative values reflect a FP rate less than α. Combining results for all choices of α and w and fitting curves as a function of N_{init} (Fig 3c) will allow us to summarize trends, independent of choice of α and w.

In the case of the Symmetric policy, the false positive rate is always lower than α; this beneficial effect is strongest when N_{incr} is large (warm colors in Fig 3d-f) or N_{init} is small (positive slopes in Fig 3d-f). In summary, these simulations show that the effect of incremental sampling on the false positive rate is real, but modest in size and lawfully related to a handful of parameters.

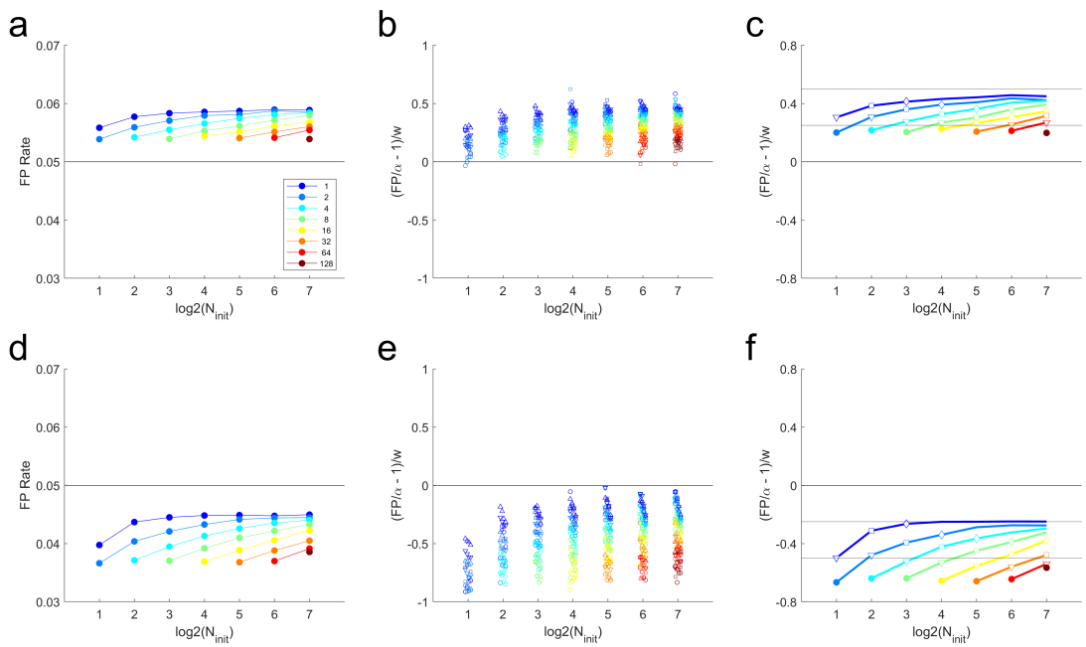

Figure 3. Dependence of FP rate on the initial sample size and sample increment. **a.** The realized false positive rat of the Asymmetric N-increasing policy, as a function of $\log_2 N_{init}$ (horizontal axis) and N_{incr}(colors), for the case $\alpha = 0.05$, $w = 0.4$, $N_{max} = 256$. Each symbol represents the result from 10^6 experiments simulated with a ground truth of no effect. **b.** Results for all choices of α (0.005, 0.010, 0.025, or 0.050; symbol shapes) and w (0.1, 0.2, 0.3 or 0.4, small horizontal shifts), plotted as $(\frac{FP}{\alpha} - 1)/w$ (vertical axis) to reveal regularities. Each symbol represents the results from 10^6 simulated experiments. **c.** Summary of simulations in (b) obtained by fitting the equation $FP = (cw + 1)\alpha$. Symbols indicate simulations in which $N_{incr} = N_{init}$ (closed circles), $N_{incr} = N_{init}/2$ (open triangles), $N_{incr} = N_{init}/4$ (open squares) and $N_{incr} = N_{init}/8$ (open diamonds). Dotted black lines relate to the rules of thumb (see text). Panels **d-f**: as in a-c, for the Symmetrical N-increasing policy. Similar results were found for both policies using $N_{max} = 5\ N_{init}$.

From the simulations one can take away some rules of thumb (dotted lines in Fig 3 c,f):

Assymetric Policy	***Symmetric Policy***
$FP < \alpha(1+\frac{w}{2})$ for $N_{incr} \leq N_{init}$	$FP < \alpha\left(1-\frac{w}{4}\right)$ for $N_{incr} \leq N_{init}$
$FP < \alpha(1+\frac{w}{4})$ for $N_{incr} = N_{init}$	$FP < \alpha\left(1-\frac{w}{2}\right)$ for $N_{incr} = N_{init}$

For example: a study that started with $N = 10$ samples and $\alpha = 0.05$, but was willing to add 10 more samples if it obtained a non-significant result with P<0.07 (up to a cap $N_{max} \leq 50$), would correspond to an Asymmetric policy with $w = 0.4$, $N_{init} = 10$, $N_{incr} = 10$, $N_{max} = 5N_{init}$, and could conservatively estimate that its false positive rate is $FP < 0.0550$ by rule of thumb, compared to the simulation result of $FP = 0.0541 \pm 0.0001$. Additional simulations for $\alpha = 0.05\ or\ 0.10$, $N_{incr} = 1$ (*i.e.* the worst case conditions) were extended to $w = 19.0$ for $N_{init} = 2 - 128$ with $N_{max} = 256$ and still did not exceed this bound (not shown).

These rules of thumb are meant to be helpful guides, but have not been formally proven, which limits generalization to other conditions not tested. The code provided in Supplementary Materials can be used to compute a Monte Carlo estimate of the false positive rate for other parameter combinations. These observed regularities could in principle be used to place a bound on the false positive rate or make a conservative correction (c.f. (*2*)), but there already analytic corrections available to account for incremental sampling – even if the decision to increment the sample size was made after the fact (*3-7*).

Trade-off between statistical power and positive predictive value

Regardless of the amplitude of the effect, an Asymmetrical N-increasing policy certainly increases the false positive rate. This is because some true negative results will, by chance, be converted to false positives when samples are added. But the motivation for doing it is the hope of increasing sensitivity: some "almost-significant" effects are false negatives, which might be converted to true positives with added samples. Conversely the Asymmetric policy reduces false positive rate, but at the risk that some true positives will be converted to false negatives. How these two factors balance depends on what fraction of the tested hypotheses are in fact true (prior probability of effect) and how large the effects are when present (*8*). The simulations presented below will explore this interaction.

First, it is helpful to remember that even in the standard fixed-N policy there is a trade-off between sensitivity and selectivity, which is controlled by the choice of α. The sensitivity, or statistical power, is the fraction of real effects for which the null hypothesis is rejected – the chance that a real effect, if present, will be discovered. The selectivity, or positive predictive value (PPV), is the fraction of the experiments rejecting the null hypothesis that reflect real effects – the chance that a positive result will turn out to be reproducible. In experiments, the probability of a real effect and the true size of the effect are not known, but in simulations these facts are known precisely. Given those facts, for any given sample size N, increasing the arbitrary cutoff for significance α increases sensitivity, at the expense of reduced PPV. By varying α one can define a curve for the sensitivity-selectivity trade-off (e.g., Fig 4a, any red curve). This curve summarizes the options available for interpreting data sets acquired in this

way. The choice of α is up to the investigator, depending on the relative priority one sets on avoiding missing real effects vs. avoiding believing false ones.

In a fixed-N policy, a larger sample size N is always better: it increases both sensitivity and selectivity, moving the entire curve up and to the right (Fig 4a, compare red curves). From this observation we suggest a generalization that the statistical quality of any two experimental policies can be compared by relating these curves. A higher curve is better – it means one could choose some α to achieve greater sensitivity for any target selectivity, or to achieve greater selectivity for any target sensitivity, relative to any curve that lies below it.

Dependence on N_{init} and α

To show how N-hacking affects these curves, simulations were done exactly as described above, but now 1% of all experiments were simulated with a real effect of 1σ difference between the population means, such that rejecting the null is the correct conclusion. This was simulated for several choices of N_{init} and α, comparing $w = 0$ (i.e. fixed-N policy) to $w = 0.4$, on either the Asymmetric or Symmetric N-increasing policy.

The curves for the standard fixed-N policy (red curves, Fig 4a) provide the benchmark to which other sampling policies may be compared. For example, an Asymmetric N-increasing policy with $w = 0.4, N_{incr} = N_{init}$ and a sample size cap of $N_{max} = 2N_{init}$ is shown (blue curves, Fig 4a). Because few experiments experience incrementing on this policy, the average final sample size was negligibly greater than the fixed-N policy: $\langle N_{\text{final}} \rangle \leq 1.02\ N_{init}$ for all parameter combinations tested (not shown; cf. Fig.2a). Therefore, the overall sensitivity and selectivity of the policy can be reasonably compared to the fixed-N policy with $N = N_{init}$ (paired curves). For all choices of N_{init} simulated, the curve for the Asymmetric N-increasing policy (blue) fell entirely above and to the right of the corresponding curve for the fixed-N policy (red). Thus on average the Asymmetric N-increasing policy resides entirely on a better frontier than the fixed-N policy: for any point on the fixed-N curve there exists some choice of α for which the Asymmetric policy curve has equal selectivity with higher sensitivity, and another choice of α for which the Asymmetric policy has equal sensitivity with higher selectivity.

Comparing the two policies with the same choice of α (symbols of same shape on the red vs. blue curves) is also informative. For the parameter combinations with lower power ($N_{init} = 6\ or\ 12$ with any α, or $N_{init} = 24$ with $\alpha < 0.01$), both sensitivity and selectivity were higher on the Asymmetric N-increasing policy than the fixed-N policy for the same α, and this was true up to $w = 1$ (not shown). For the parameter combinations with higher power ($N_{init} = 48$ with any α, or N_{init}=24 with $\alpha < 0.01$), using the same α for the Asymmetric N-increasing policy led to a loss in selectivity relative to the fixed-N policy (the matched symbols are above, but to the left, of the fixed-N benchmark). Still, this loss in selectivity was accompanied by a far greater improvement in statistical power than could be achieved by moving along the red curve (changing α) to obtain the same selectivity. Thus the Asymmetric policy represented a superior trade-off even in these cases.

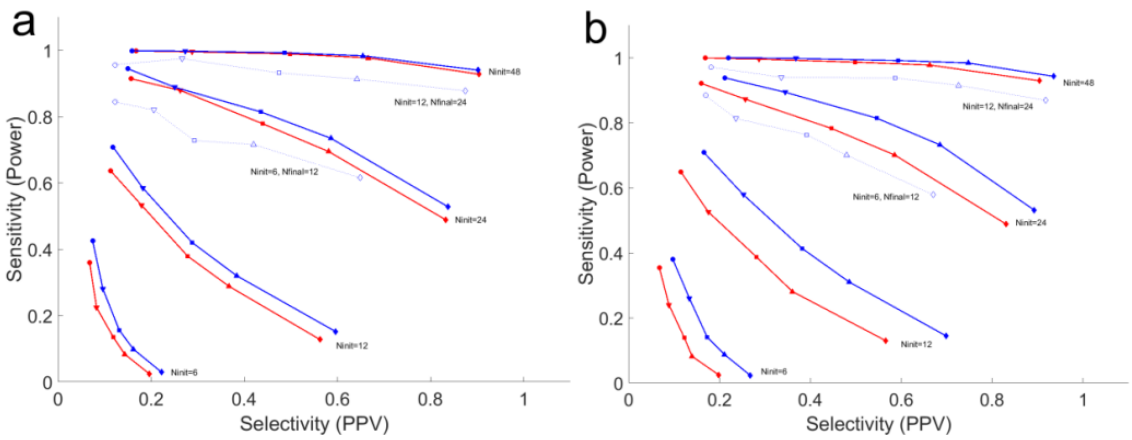

Figure 4. Trade-off between selectivity and sensitivity. (a) Realized selectivity vs. sensitivity in simulations with effect size 1σ, prior effect probability 0.01, and N_{init}=6, 12, 24 or 48 (four curves of each color). Symbols indicate α (○=0.001, ▽=0.005, □=0.01, △=0.025, ◊=0.05). Text labels indicate N_{init} for pairs of curves. For every combination of these parameters a total of $M = 10^4/\alpha$ experiments were simulated to estimate the probabilities of true positives, true negatives, false positives and false negatives caused by the policy. Results for the standard fixed-N policy (red) and for an Asymmetric N-increasing policy using $N_{incr} = N_{init}$, $N_{max} = 2N_{init}$, and $w = 0.4$ (blue) are shown. The small subset of experiments that added samples (to $2N_{init} = N_{final}$) are shown in dotted blue curves and open symbols for $N_{init} = 6$ and $N_{init} = 12$. **(b)** Results for standard fixed-N vs. Symmetric N-increasing policy, details otherwise as in (a).

Although the curve for the Asymmetric N-increasing policy fell above the benchmark curve for N_{init}, the small subset of experiments for which sample size was increased had $2N_{init}$ final samples. Is the whole effect due to the fact that un-incremented experiments lie on the fixed-N curve for $N = N_{init}$ and the incremented subset lie on the curve for $N = 2N_{init}$? The answer is no. Considering the incremented subset of experiments separately (dotted blue curves) reveals that they live on a frontier *above* the curve for fixed-N experiments with a sample size $N = 2N_{init}$. The subset of experiments that were not incremented (which had a sample size of exactly N_{init}) lay on a curve that was either slightly above or indistinguishable from the fixed-N benchmark in all cases examined (not shown).

These simulations demonstrate that for an effect size of 1σ and effect probability of 0.01, the Asymmetric N-increasing is a clear win-win scenario: for any initial sample size, whatever selectivity (PPV) one can achieve on the fixed-N policy, that same selectivity can be achieved with higher statistical power on the Asymmetric N-increasing policy for some choice of α. Moreover, in situations where statistical power is limited, using the same choice of α in an Asymmetric N-increasing policy – even without any correction for the false positive rate or multiple comparisons – yields improvements in *both* statistical power *and* PPV relative to fixed-N. Additional simulations verified that the Asymmetric policy curves remained above the curves of the fixed-N policy as either the prior probability or effect size approached 0 (although PPV approaches 0 in both cases), for a range of N_{init} and $N_{incr} = N_{init}$ (not shown).

To be clear, ordinarily the prior probability of a true effect and the effect size are not known to the investigator, so the PPV of any given empirical result is unknown. Nevertheless under certain conditions, whatever the PPV would be using fixed-N, the PPV after Asymmetric N-increasing would be greater. This can be shown in simulations, in which the prior probability

and effect size *are* known. Under other conditions, the PPV after Asymmetric N-increasing is lower than that of a fixed-N experiment, but the simulations demonstrate that there must be some choice of α that would provide the same PPV as fixed-N with higher power, and some other choice that would provide the same power with higher PPV. How to find these values of α is not addressed.

The Symmetric N-increasing policy was superior to the fixed-N policy (Fig 4b, compare red to blue as described for Fig 4a), as well as beating the Asymmetric policy (compare blue curves in 5a to 5b). Comparing results using the same choice of α the Symmetric policy increased both selectivity and sensitivity relative to fixed-N for all conditions tested. The Symmetric policy had about the same sensitivity as the Asymmetric policy with the same α, but selectivity was much higher. The subset of experiments on the Symmetric policy that had added samples to a final $2N$ fell on a curve well above the fixed-N experiments $2N$ samples, and the subset of experiments that reached a verdict with N samples fell on a curve either above or indistinguishable from the fixed-N curve with N samples.

Dependence on the decision window *w*

To demonstrate the impact of the decision window w on these conclusions, I further simulated results for a range of N_{init} for up to $w = 10$, for the Asymmetric case with $N_{incr} = 1$ and $N_{max} = 50$, (Fig 5). In the case of $= 0.05$, a policy with $w = 10$ means adding samples if $P < 0.5$, an extremely liberal policy (Fig. 5). While this will result in a rather high false positive rate, how does it affect statistical inference?

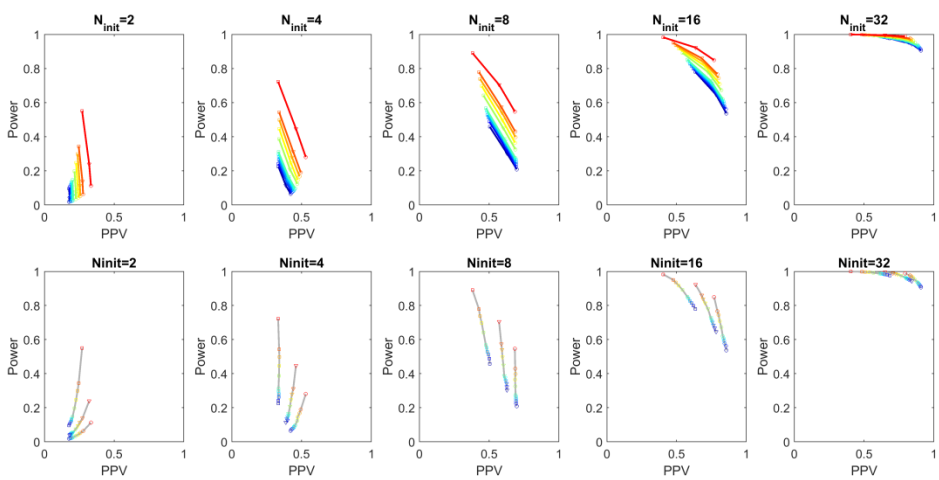

Figure 5. Asymmetric N-increasing policies improve statistical inference even when *w* is large. Realized selectivity vs. sensitivity in simulations with effect size 1σ, prior effect probability 0.1, N_{init} as indicated on panel title, $N_{incr} = 1$, $N_{max} = 50$. Each symbol represents the uncorrected result from $M = 10^6$ simulated experiments. Symbols indicate α (○=0.01, ▽=0.02, □=0.05). Colors indicate w (blue→red= 0, 0.2, 0.4, 0.6, 0.8, 1, 2, 3, 4, 5, 10). Note that dark blue ($w =$ 0) is the standard fixed-N policy. **Top row:** simulations with the same w and different α connected with curves, color indicates w. **Bottom row:** the same data, but simulations with the same α and different w connected.

Increasing w always increases sensitivity. As w increases, at some point depending on N_{init} the uncorrected Asymmetric policy switches from increasing the PPV to eroding it, compared to the fixed-N policy using the same α (Fig 5, bottom row). Nevertheless as w

increases, the curves of PPV vs. power continue to move up and to the right (Fig 5, top row). Thus, one may have to adjust α to obtain the same PPV as fixed-N, but the Asymmetric policy produces statistically better options: for the same PPV, the power is higher. It is noteworthy that the most extreme case, $N_{init} = 2$ and retesting after every $N_{incr} = 1$ sample, Asymmetric N-increasing is strictly beneficial, increasing both power and PPV without even adjusting α. The significance of this will be discussed below.

Discussion

Main conclusions

These simulations demonstrate that false positive rates are increased if the sample size is grown incrementally whenever a result is non-significant but close. I demonstrate that the problem arises from the fact that equally close significant results are not similarly challenged. Most writers warn that this practice will lead to extremely high or even 100% false positive rates (*6, 9-12*). But those projections are based on assumptions that are not representative of typical practice in some scientific fields, such as that an experimentalist would add more samples after obtaining a non-significant result no matter how far from α the P value was, or would continue adding samples indefinitely until achieving a significant outcome.

If instead one considers circumstances in which the P value would have to be rather close to α for one to add samples (e.g., no more than twice α), and a limited number of total samples could be added before giving up (e.g., no more than five times the initial sample), the effects on the false positive rate are modest and clearly bounded.

The increase in false positive rate depends quantitatively on the initial sample size N_{init}, significance criterion α, closeness criterion w, increment size N_{incr}, and total sample cap N_{max}. Simulations demonstrate in which direction and how steeply the false positive rate depends on these factors. Some rules of thumb emerge for how bad the effect could possibly be, given those parameters. These empirically observed bounds are not yet analytically proven. The take home lesson is that the increase false positive rate is a lawful function of known parameters, and therefore one can correct for it rigorously – even if it was not planned in advance (*3-6*).

Further simulations demonstrate that under many conditions this type of N-hacking is superior to a fixed-N policy in the sense that it increases the statistical power achievable for any given positive predictive value (PPV), compared to studies that strictly adhere to the initially planned N. In particular, for the great many experimental studies that use a small initial sample size ($N \leq 12$) and $\alpha = 0.05$, if one would only add samples post-hoc when $P < 0.1$ and would always quit after accumulating $\sim N = 50$ samples, it is simply not true that N-hacking leads to an elevated risk of unreproducible results. A verdict of "statistical significance" reached in this manner, far from being dubious, is more likely to be reproducible than results reached by fixed-N experiments with the same initial or final sample size – even if no correction were applied for sequential sampling or multiple comparisons. This result may only be true in a narrow range of parameters, and thus is not of mathematical interest; but for experimentalists working within that range of parameters, the result is of considerable practical value.

Scientists in exactly this situation are currently being told (by teachers, advisors, reviewers, editors, and even staff biostatisticians) that if they have obtained a non-significant finding with a P value just above α, they cannot validly add more samples to their data set to improve statistical power; they must either run a completely independent replication, or accept the null hypothesis. The results shown here imply that this is bad advice. It is true that adding samples after the test violates the basic premise of null hypothesis significance testing (NHST). But that is not the same as being *invalid*. Adding more samples *with disclosure* is never invalid, and there are methods for rigorous correction of the P value within the NHST framework. Moreover, these simulations show that there are statistical benefits of incremental sampling that are often overlooked.

Extensions and limitations

These simulations used a normal distribution for the ground truth source distribution and an independent sample t-test as our basic hypothesis test. But the analysis of the false positive rate only depends on the assumption that the statistical test used generates P values that are uniformly distributed between 0-1 on the null hypothesis. In other words, as long as the statistical test being used is valid for the distribution being sampled and the structure of the experiment, the dependence of false positive rate on parameters in these simulations should generalize to any source distribution and statistical test. Power analysis might be affected by the shape of the source distribution, however, so generality of those results to other distributions should not be assumed.

These simulations only considered experiments in which a single hypothesis is tested on each sample. Multiple tests on a single sample (such as a gene chip array experiment) is a very different situation, because in that case incrementing N and retesting would lead to re-testing of all the hypotheses, regardless of their original P values. This case has been discussed by others.

Numerical simulations and graphs are easy for experimental scientists to understand, because they present the expectations of the null hypothesis in terms directly comparable to data. These simulations covered a broad range of parameters to provide concrete intuitions about the directions and orders of magnitudes of effects in relation to experimental parameters. But analytic treatments would be necessary to determine exactly under what conditions these results will apply, to provide rigorous proofs and precise bounds or corrections.

The real reason you should not N-hack

Some may be suspicious of the claim that N-hacking, in the sense simulated here, provides (slightly) better statistical inference than fixed-N experiments, in terms of the power achieved for any given PPV, as well as the number of samples required to achieve this power. But this result is not at odds with established statistical theory. The N-incrementing policies described here are closely related to other well-described sequential sampling methods – particularly in the limit of $N_{init} = 2, N_{incr} = 1$ (Fig 5, left panels). For example, in Wald's Sequential Probability Ratio Test (*13*), one sets a threshold α to accept and another threshold β to reject a hypothesis. Then one computes a test statistic S after each new sample is added. If $S < \alpha$ the hypothesis is accepted, if $S > \beta$ the hypothesis us rejected, and if $\alpha < S < \beta$ one

continues sampling. Similarly, a Bayesian sequential sampling method sets a criterion c, and then sequentially computes the Bayes Factor for the hypothesis vs. null hypothesis. If $BF > c$ the hypothesis is accepted, $if\ BF < \frac{1}{c}$ the hypothesis is rejected, and otherwise one keeps sampling (*14*). The drift diffusion model (DDM), which is widely used to model decision-making, is closely related (*15*). Fully sequential sampling methods are known to be statistically powerful and efficient. The kind of N-hacking commonly practiced is merely a weak version, conferring minor benefits compared to fixed-N methods. So the real reason not to advocate N-hacking as an intentional method is that fully sequential sampling methods are even better (*16, 17*).

Finally it is worth noting that this entire problem arises because of the currently standard practice of setting arbitrary cutoffs for "statistical significance" and reducing analog P values to binary hypothesis tests. It is not at all clear that experimental science is well served by this overall approach (*18-20*). Converting a P value into a significance verdict necessarily discards information. Maintaining an analog estimate of the evidence for a hypothesis as data are accumulated would be better (*21*), and would eliminate the need for most of this discussion.

Acknowledgements

PR acknowledges Hal Pashler, Casper Albers and Daniel Lakens for valuable discussions and helpful comments on earlier drafts of the manuscript.

References

1. E. C. Yu, A. M. Sprenger, R. P. Thomas, M. R. Dougherty, When decision heuristics and science collide. *Psychon Bull Rev* **21**, 268-282 (2014).
2. R. W. Frick, A better stopping rule for conventional statistical tests. *Behav Res Meth Ins C* **30**, 690-697 (1998).
3. D. L. DeMets, G. Lan, The alpha spending function approach to interim data analyses. *Cancer Treat Res* **75**, 1-27 (1995).
4. D. Lakens, Performing high-powered studies efficiently with sequential analyses. *Eur J Soc Psychol* **44**, 701-710 (2014).
5. B. J. Sagarin, J. K. Ambler, E. M. Lee, An Ethical Approach to Peeking at Data. *Perspect Psychol Sci* **9**, 293-304 (2014).
6. E. Schott, M. Rhemtulla, K. Byers-Heinlein, Should I test more babies? Solutions for transparent data peeking. *Infant Behav Dev* **54**, 166-176 (2019).
7. D. Lakens, E. R. Evers, Sailing From the Seas of Chaos Into the Corridor of Stability: Practical Recommendations to Increase the Informational Value of Studies. *Perspect Psychol Sci* **9**, 278-292 (2014).
8. D. Colquhoun, An investigation of the false discovery rate and the misinterpretation of p-values. *R Soc Open Sci* **1**, 140216 (2014).
9. C. Albers, The problem with unadjusted multiple and sequential statistical testing. *Nat Commun* **10**, 1921 (2019).
10. D. Szucs, A Tutorial on Hunting Statistical Significance by Chasing N. *Front Psychol* **7**, 1444 (2016).
11. J. P. Simmons, L. D. Nelson, U. Simonsohn, False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci* **22**, 1359-1366 (2011).

12. H. J. Motulsky, Common misconceptions about data analysis and statistics. *Naunyn Schmiedebergs Arch Pharmacol* **387**, 1017-1023 (2014).
13. A. Wald, *Sequential analysis*. Wiley mathematical statistics series (J. Wiley & sons, New York, 1947), pp. xii, 212 p.
14. S. N. Goodman, Toward evidence-based medical statistics. 2: The Bayes factor. *Ann Intern Med* **130**, 1005-1013 (1999).
15. J. I. Gold, M. N. Shadlen, Neural computations that underlie decisions about sensory stimuli. *Trends Cogn Sci* **5**, 10-16 (2001).
16. J. Bartroff, T. L. Lai, M.-C. Shih, *Sequential experimentation in clinical trials : design and analysis*. Springer series in statistics, (Springer, New York, 2013), pp. xv, 237 pages.
17. D. Siegmund, *Sequential analysis : tests and confidence intervals*. Springer series in statistics (Springer-Verlag, New York, 1985), pp. xi, 272 p.
18. R. L. Wasserstein, A. L. Schirm, N. A. Lazar, Moving to a World Beyond "p < 0.05". *Am Stat* **73**, 1-19 (2019).
19. R. S. Nickerson, Null hypothesis significance testing: a review of an old and continuing controversy. *Psychol Methods* **5**, 241-301 (2000).
20. S. N. Goodman, Toward evidence-based medical statistics. 1: The P value fallacy. *Ann Intern Med* **130**, 995-1004 (1999).
21. S. N. Goodman, Of P-values and Bayes: a modest proposal. *Epidemiology* **12**, 295-297 (2001).