Use of Open Data R Packages to Detect P-hacking in Scientific Literature

Kroeger CM, Brown AW, Allison DB
Indiana University School of Public Health-Bloomington

abstract

background P-hacking can inflate type 1 error rates and bias published scientific literature [1-6].

objective We aimed to determine whether p-value reporting consistent with p-hacking occurs in abstracts 1) more frequently with atypical statistical analyses (e.g., non-parametric statistics) compared to common analyses (e.g., t-tests, ANOVA), and 2) when atypical analyses are versus not mentioned in abstracts.

methods We used rEntrez to search PubMed and PubMed's Dietary Supplement Subset (DSS) for: 1) p-values within 0.01 interval bins ranging from 0.03 to 0.06, 2) terms consistent with common statistical analyses, and 3) terms consistent with atypical analyses. Associations between abstracts reporting only atypical or common terms and only a single p-value bin within each condition were assessed via Pearson's Chi-squared test.

results When comparing common to atypical terms in PubMed, atypical yielded p-values in the 0.05 bin 10% less often than expected and in the 0.04 bin 2% more often than expected (p=0.024504) – a hallmark of p-hacking. While not statistically significant (p = 0.1068), patterns were similar within the DSS (10% underrepresentation in 0.05 bin, and a 2% overrepresentation in the 0.04 bin). No associations were seen in PubMed or the DSS when abstracts mention versus do not mention atypical terms (p = 0.78125; p=0.54593, respectively).

conclusion Overrepresentation of p-values in the 0.04 range and underrepresentation in the 0.05 range with atypical compared to common terms is consistent with p-hacking.

public health implications Public health research may be biased by investigators preferentially switching to and reporting statistically significant atypical analyses when common analyses are just above the p < 0.05 threshold.

background

figure 1. Example of p-hacking [7]

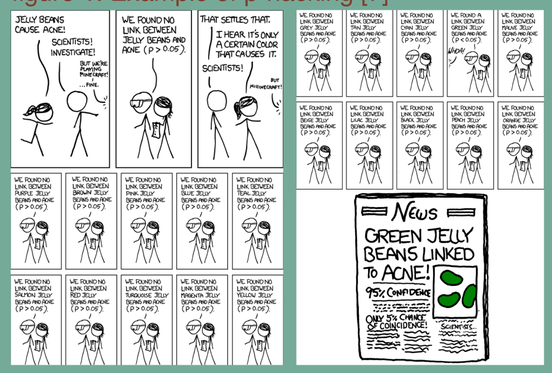


figure 2. The effect of p-hacking on the distribution of p-values in the range of significance [5]

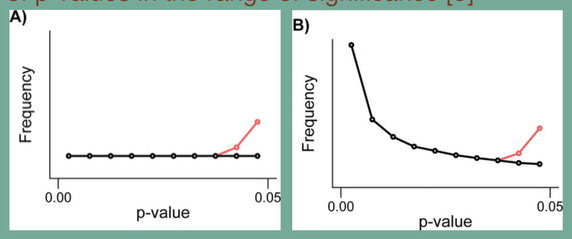


figure 3. Some conditions under which distributions cannot be relied on to indicate p-hacking [8]

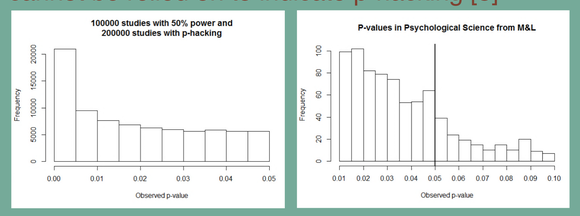


table 1. Manual search query in PubMed

Search	PubMed Query	Items	Time
#22	#14 AND #12	181	11:18
#21	#14 AND #11	103	11:17
#20	#14 AND #10	1011	11:17
#19	#14 AND #9	1209	11:17
#18	#13 AND #12	465	11:16
#17	#13 AND #11	364	11:16
#16	#13 AND #10	2480	11:16
#15	#13 AND #9	3134	11:15
#14	#8 NOT #7	162910	11:11
#13	#7 NOT #8	83881	11:11
#12	#6 NOT (#1 OR #3 OR #5)	17059	11:10
#11	#5 NOT (#1 OR #3 OR #6)	10323	11:08
#10	#3 NOT (#1 OR #5 OR #6)	90871	11:07
#9	#1 NOT (#3 OR #5 OR #6)	111569	11:06
#8	nonparametric [tiab] OR non-parametric [tiab] OR wlicoxon-rank [tiab] OR "wilcoxon rank" [tiab] OR kruskal-wallis [tiab] OR "kruskal wallis" [tiab] OR transformation [tiab] OR (outlier* [tiab] AND remov* [tiab])	166384	11:04
#7	t-test [tiab] OR anova [tiab] OR ancova [tiab] OR "mixed model" [tiab]	87355	11:02
#6	p=.06* [tiab] OR p=0.06* [tiab]	25343	11:01
#5	p=.05* [tiab] OR p=0.05* [tiab]	16409	10:59
#3	p=.04* [tiab] OR p=0.04* [tiab]	130032	10:55
#1	p=.03* [tiab] OR p=0.03* [tiab]	151564	10:53

Atypical + 0.06*
Atypical + 0.05*
Atypical + 0.04*
Atypical + 0.03*
Common + 0.06*
Common + 0.05*
Common + 0.04*
Common + 0.03*

methods

figure 4. Programmatically reproducible rEntrez search query in PubMed

```
# Create objects for individual p-value bin search strings
three <- "(p=.03*[TIAB] OR p=0.03*[TIAB])"
four <- "(p=.04*[TIAB] OR p=0.04*[TIAB])"
five <- "((p 050[TIAB] OR p 0500[TIAB] OR p 05006[TIAB] OR p 0501[TIAB]...
six <- "(p=.06*[TIAB] OR p=0.06*[TIAB])"

# Create objects for exclusive p-value bin search strings
three_ex <- paste(three, "NOT", four, "NOT", five, "NOT", six)
four_ex <- paste(four, "NOT", three, "NOT", five, "NOT", six)
five_ex <- paste(five, "NOT", three, "NOT", four, "NOT", six)
six_ex <- paste(six, "NOT", three, "NOT", four, "NOT", five)
# Create object for common tests search string for Method d
com_d <- "(t-test[TIAB] OR t test[TIAB] OR t-student[TIAB] OR t student[TIAB] OR
          anova[TIAB] OR ancova[TIAB])"

# Create object for atypical tests search string for Method d
atyp_d <- "(nonparametric[TIAB] OR non-parametric[TIAB] OR non parametric[TIAB]
           OR kruskal-wallis[TIAB] OR rank-sum test[TIAB] OR rank sum test[TIAB]
           OR spearman rank correlation coefficient[TIAB] OR
           spearman correlation[TIAB] OR wilcox test[TIAB] OR
           kolmogorov-smirnov test[TIAB] OR kolmogorov smirnov test[TIAB] OR
           u-test[TIAB] OR mann whitney[TIAB] OR mann-whitney[TIAB] OR
           wilcoxon-mann-whitney[TIAB] OR wilcoxon[TIAB] OR wilcoxon-rank[TIAB]
           OR kruskal wallis[TIAB] OR sign test[TIAB] OR signed-rank[TIAB] OR
           friedman test[TIAB] OR mood's median test[TIAB] OR bootstrapping[TIAB]
           OR permutation test[TIAB] OR log-transform*[TIAB] OR
           log transform*[TIAB] OR log transformed[TIAB] OR
           (outlier*[TIAB] AND remov*[TIAB]))"
```

```
# Create object for exclusive common tests search string
com_d_ex <- paste(com_d, "NOT", atyp_d)

# Create object for exclusive atypical tests search string
atyp_d_ex <- paste(atyp_d, "NOT", com_d)

# Creat object for each exclusive p-value bin and test type combination search
three_com_d_ex <- paste(com_d_ex, "AND", three_ex)
four_com_d_ex <- paste(com_d_ex, "AND", four_ex)
five_com_d_ex <- paste(com_d_ex, "AND", five_ex)
six_com_d_ex <- paste(com_d_ex, "AND", six_ex)
three_atyp_d_ex <- paste(atyp_d_ex, "AND", three_ex)
four_atyp_d_ex <- paste(atyp_d_ex, "AND", four_ex)
five_atyp_d_ex <- paste(atyp_d_ex, "AND", five_ex)
six_atyp_d_ex <- paste(atyp_d_ex, "AND", six_ex)
# Create counts for each exclusive p-value bin and test type combination
three_com_d_ex_ct <- entrez_search(db = "pubmed",
                                   term = three_com_d_ex,
                                   retmax = 0)
```

results

figure 5. Common vs. atypical tests in abstracts by p-value bin

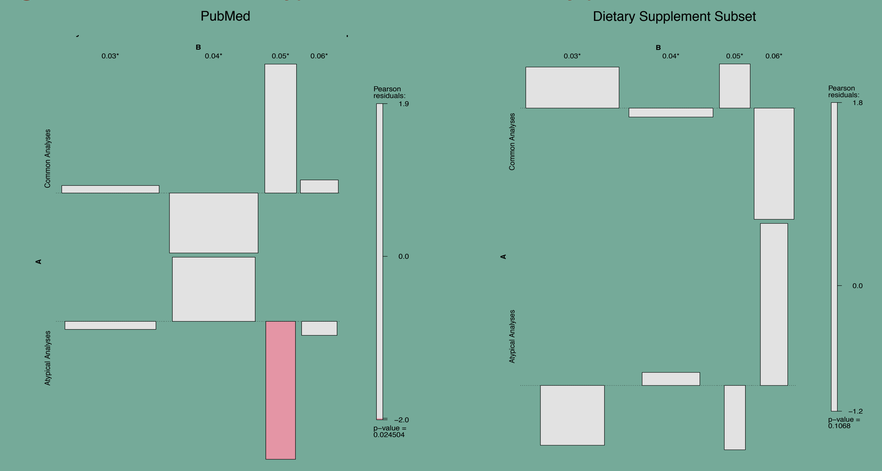
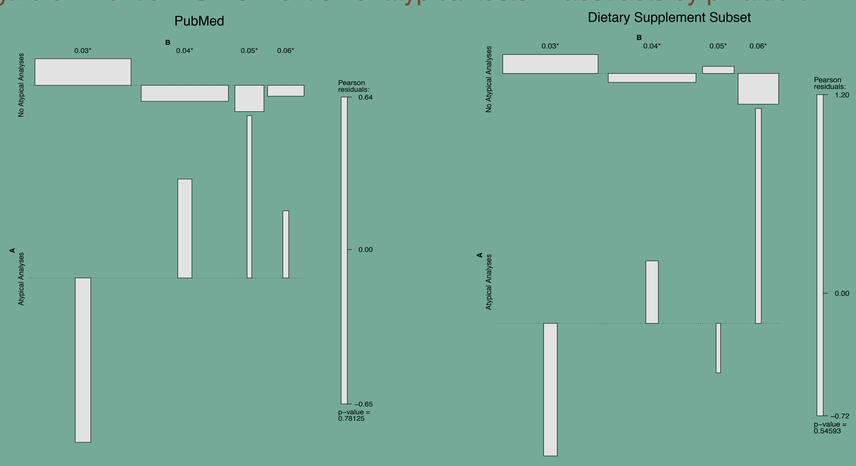


figure 6. Mention vs. no mention of atypical tests in abstracts by p-value bin



references

[1] Motulsky HJ. Common misconceptions about data analysis and statistics. *Naunyn-Schmiedeberg's Arch Pharmacol* (2014) 387:1017-1023. **[2]** Simonsohn U, et al. P-curve: a key to the file drawer. *J Exp Psychol Gen* (2014) 143 (2):534-47. **[3]** Masicampo EJ, et al. A peculiar prevalence of p values just below .05. *Q J Exp Psychol (Hove)* (2012) 65(11):2271-9. **[4]** Gadbury & Allison. Inappropriate fiddling with statistical analyses to obtain a desirable p-value: Tests to detect its presence in published literature. *PLOS One*. 2012. **[5]** Head ML et al. The extent and consequences of P-Hacking in Science. *PLOS Biology* (2015) 13(3). **[6]** Kroeger, CM et al. Evidence of p-value fiddling using a rapid, high-volume, systematic method. Advances and Controversies in Clinical Nutrition. ASN Annual Scientific Meeting, 2014. **[7]** explainxkcd.com/wiki/index.php/882:_significant **[8]** Lakens D. What p-hacking really looks like: A comment on Masicampo and LaLande. *The Quarterly Journal of Experimental Psychology* (2015) 68(4): 829-832.

funding

Provided by: NIH (F32DK107157)