Simple text-mining methodology to detect comparative p-hacking is sensitive to text searching variations

Kroeger CM, Brown AW, Allison DB

Nutrition Obesity Research Center, University of Alabama at Birmingham, AL, USA

Abstract

Background
P-hacking (altering statistical analyses to obtain a desired p-value) inflates type 1 error rates [1-3]. Simple text-mining methods can help to efficiently extract the numerous p-values needed to assess prevalence of p-hacking from scientific literature [4]. Our previous findings using this technique suggest p-hacking occurs more frequently when investigators use atypical statistical analyses (e.g., Wilcoxon-rank, outlier removal) compared to common analyses (t-tests, ANOVA) [5].

Objective
Here, we aimed to: 1) compare them to a targeted subset of literature, and 3) repeat steps one and two with a refined search method.

Design
We used our original and a refined method to search PubMed and PubMed's Dietary Supplement Subset for: 1) p-values within 0.01 interval bins ranging from 0.03 to 0.06, 2) common statistical analyses and 3) atypical analyses. Associations between analysis choice and p-value bins within each of the four conditions were assessed by chi-square tests.

Results
Findings from our original method were replicated. Atypical analyses yielded p-values in the 0.05* bin 19% less often than expected and in the 0.04* bin 3% more often than expected (p=0.0311) – a hallmark of p-hacking. When using the refined model, analysis type was not associated with p-values in each bin (p=0.2397). Likewise, no associations were seen within the Dietary Supplement PubMed Subset for original and refined models (p=0.6084 and p=0.5955, respectively).

Conclusion
The replication of our findings using our original method but failure to replicate using a conceptually refined method suggests our method may not be robust to small refinements. Neither method provided evidence of p-hacking (differential p-value distributions) between analysis types within the Dietary Supplement subset. Ongoing work aims to investigate factors that can influence robustness, as well as to efficiently extract p-values and statistical models from full-texts, as this method extracts information from abstracts only.

Tables

Table 1. Original Search Query in PubMed

Search	PubMed Query		Items	Time
#22	#14 AND #12	Atypical + 0.06*	181	11:18
#21	#14 AND #11	Atypical + 0.05*	103	11:17
#20	#14 AND #10	Atypical + 0.04*	1011	11:17
#19	#14 AND #9	Atypical + 0.03*	1209	11:17
#18	#13 AND #12	Common + 0.06*	465	11:16
#17	#13 AND #11	Common + 0.05*	364	11:16
#16	#13 AND #10	Common + 0.04*	2480	11:16
#15	#13 AND #9	Common + 0.03*	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

Table 2. Refined Search Query in PubMed

#8	"non parametric" [tiab] OR wilcoxon-mann-whitney [tiab] OR mann-whitney [tiab] OR u-test [tiab] OR wilcoxon [tiab] OR log-transformed [tiab] OR "log transformed" [tiab]
#7	"t test" [tiab] OR t-student [tiab] OR ANOVA [tiab] OR "parametric tests" [tiab]

Figures

Figure 1. Replication of Original Text-Mining Methods in PubMed

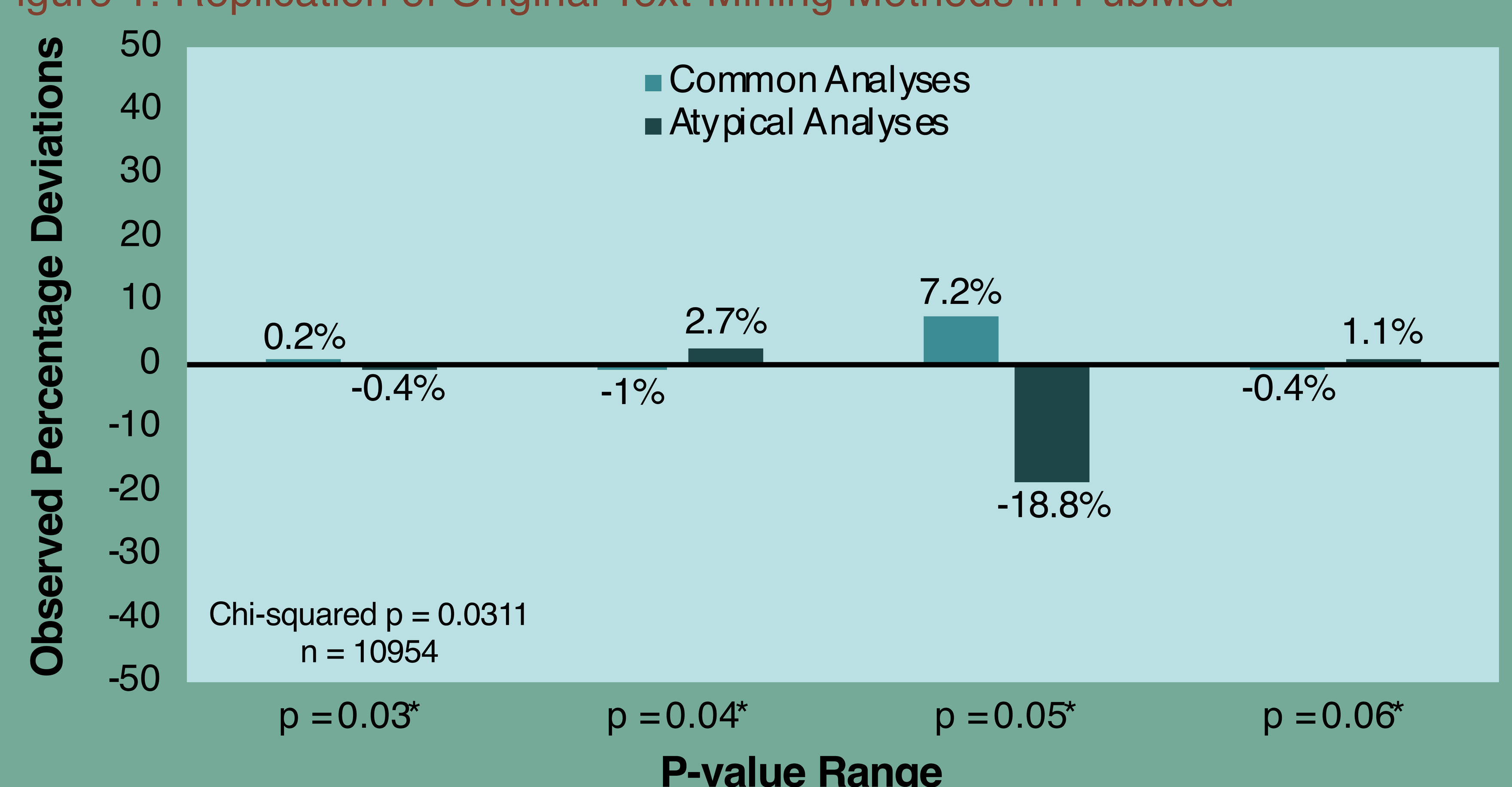


Figure 2. Refined Text-Mining Methods in PubMed

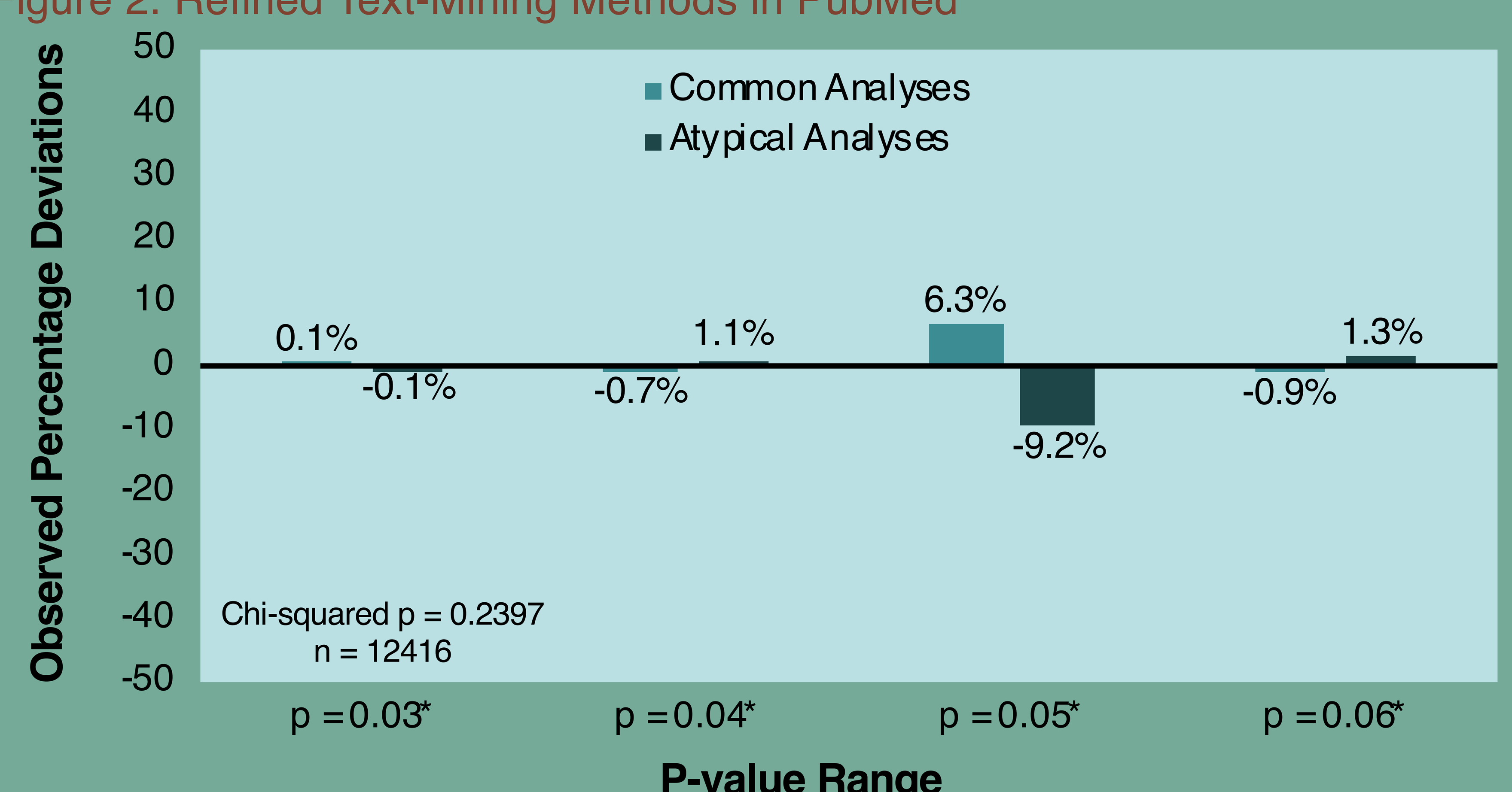


Figure 3. Original Methods in PubMed Dietary Supplement Subset

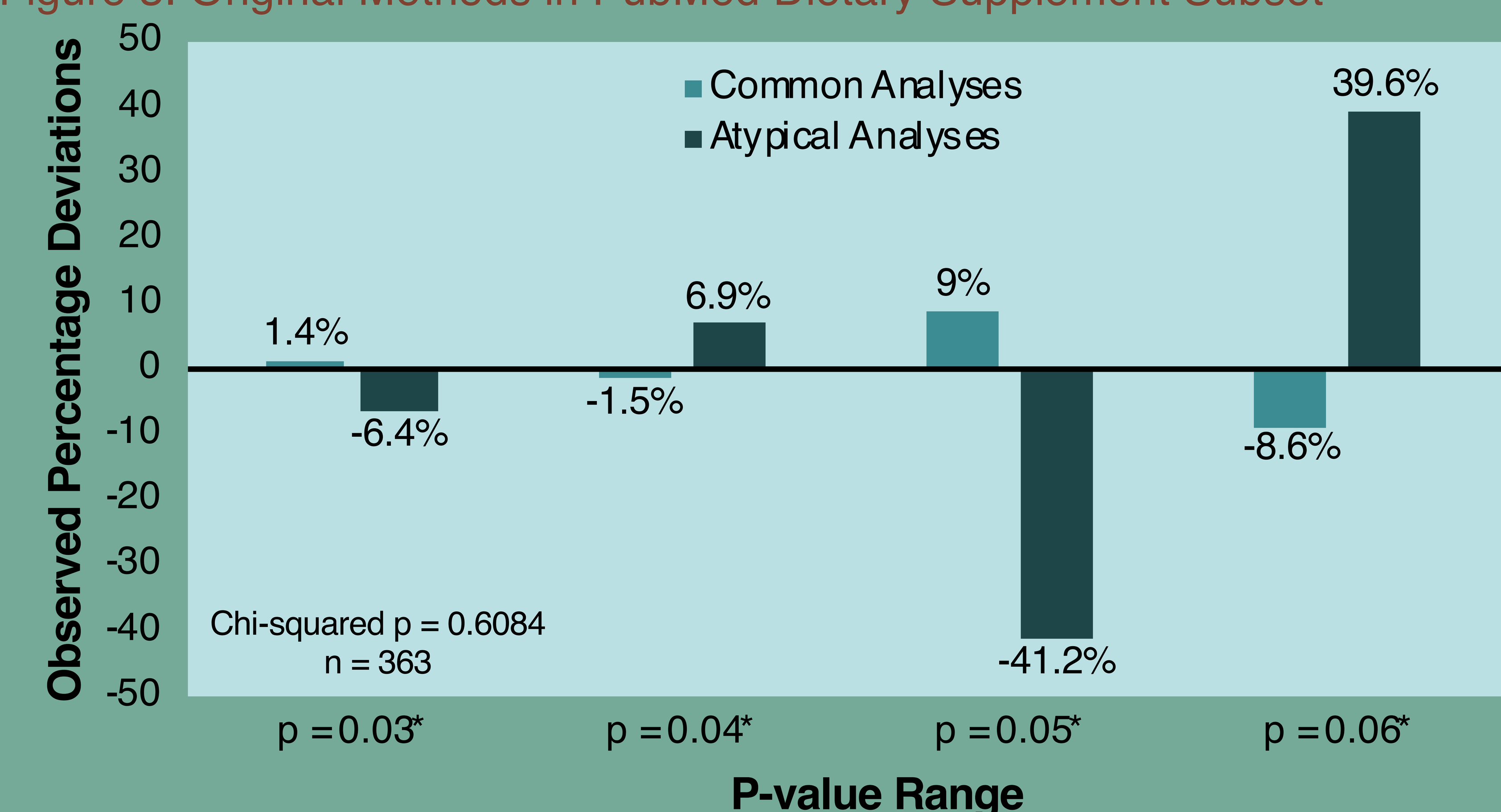
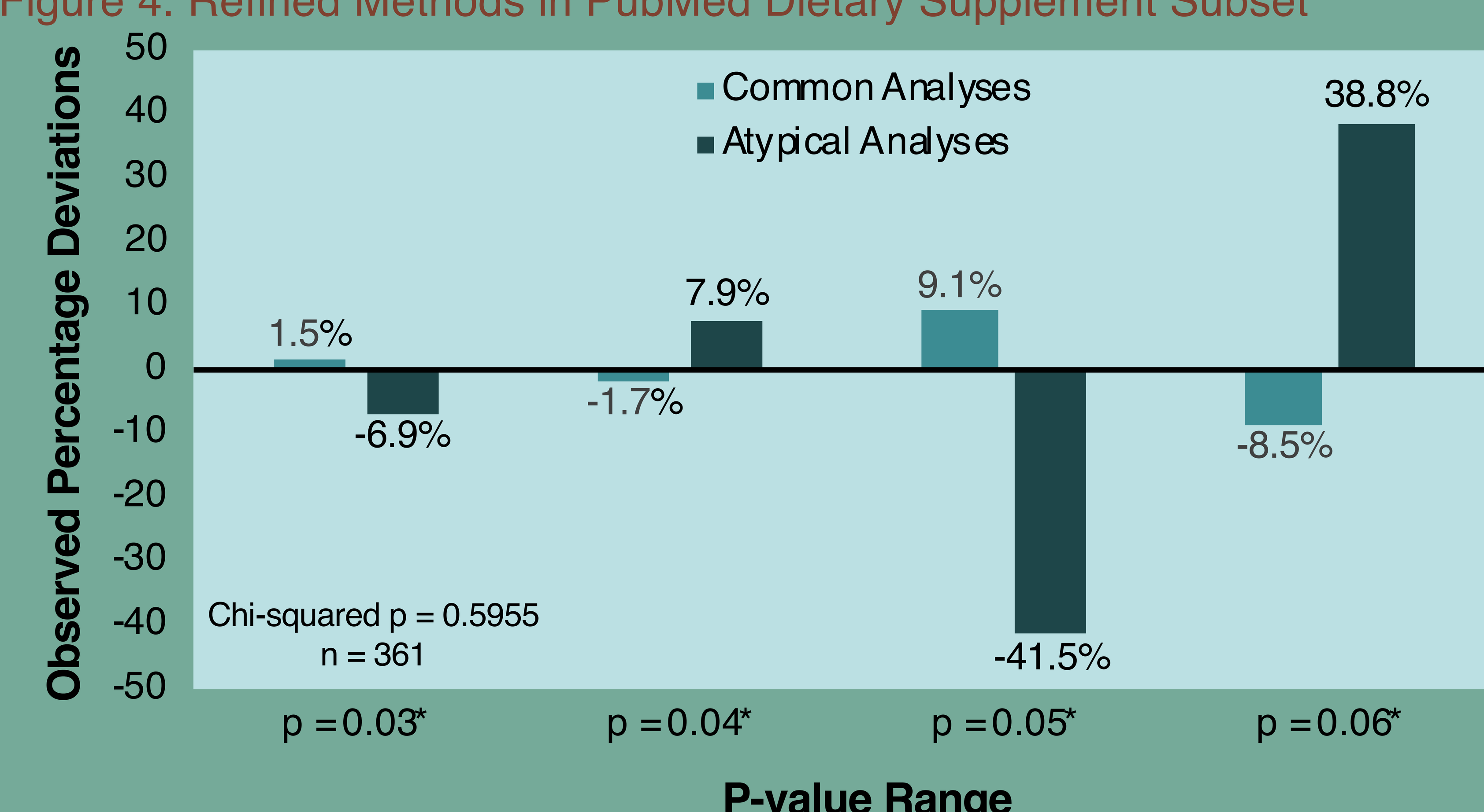


Figure 4. Refined Methods in PubMed Dietary Supplement Subset



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]** 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.

Funding

Provided by: NIH (F32DK107157)