General use of multiple testing corrections in life sciences could boost replicability

Otília Menyhárt [1,2,3] , **Balázs Győrffy** [1,2,3,4]

[1] *Cancer Biomarker Research Group, Institute of Molecular Life Sciences, Hungarian Research Network, Budapest, 1117, Hungary*
2 National Laboratory for Drug Research and Development, Budapest, 1117, Hungary
3 Department of Bioinformatics, Semmelweis University, Budapest, 1094, Hungary
4 Department of Biophysics, Medical School, University of Pecs, Pecs, 7624, Hungary

Many studies fail to adjust for multiple hypothesis testing, resulting in irreproducible findings. We created a tool (www.multipletesting.com) to automate these corrections.

BACKGROUND

- Most published research findings may be false.
- Testing multiple hypotheses increases false positives, and uncorrected p-values can obscure real biological insights.
- Many studies, including clinical trials and epidemiology, fail to adjust for multiple testing.

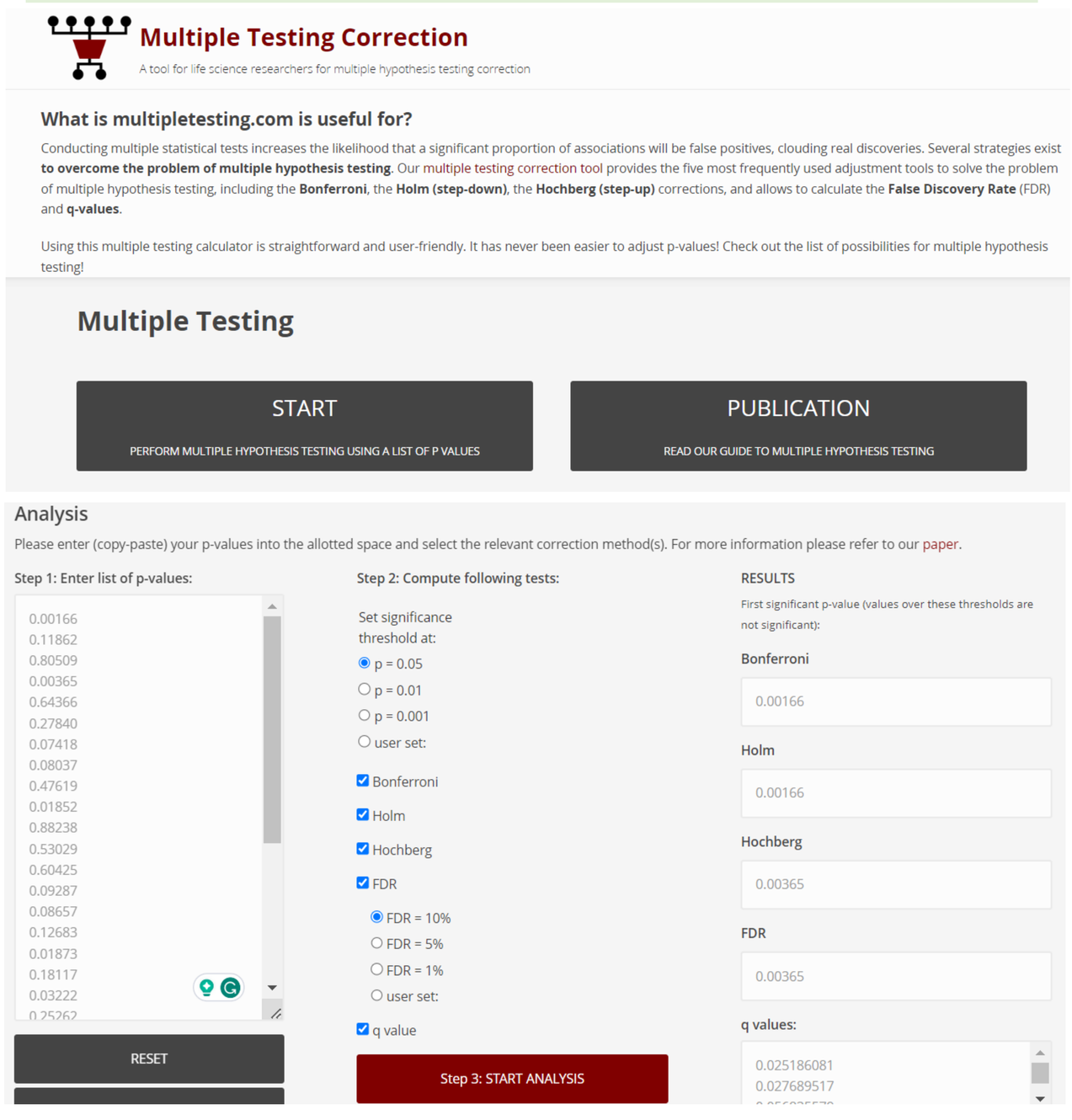
Prevalence of adjustments (reviews)

Study	Year published	Number of studies with multiple comparisons	Proportion of studies with adjustments
Cohen	2010		22%
Tyler et al.	2011	>20	5.80%
Stacey et al.	2012	538	14%
Baron et al.	2013		40%
Wason et al.	2014		51%
Gewandter	2014	33	45%
Chalkidou	2014	15	1 study
Vickerstaff et al.	2015	60	25%
Kirkham et al.	2015	140	10%
Dworkin et al.	2016	29	21%
Benjamini and Cohen	2017		20%
Brand	2021	89	2 studies altogether
Nevin	2022	38	11% final reports, 7% protocols
Pike	2022	28	48%

GOALS

to provide an automated interface for scientists to apply corrections for multiple hypothesis testing

multipletesting.com



BONFERRONI vs. FALSE DISCOVERY RATE?

Bonferroni ⟺ FDR

The original p value

$$\textit{Bonferroni-corrected p value} = \frac{\alpha}{n}$$

The number of tests performed

- strong control of type I error
- effective when a small number of hypotheses are tested
- high risk of false negatives
- drastically lowers statistical power

For 1000 variables, assume that all H0 are true, i.e. no real difference between control and experimental groups

p - values without correction:
p = 0.05 results in 50 false positive
p = 0.01 results in 10 false positives

		actual	
		positive	negative
predicted	positive	true positive	false positive
	negative	false negative	true negative

- controls the expected proportion of false positives
- higher power: more suitable for large datasets
- useful for exploratory research where some true effects are expected among many hypotheses.

when testing 20 variables

adjusted p 0.05/20 = 0.0025 - too stringent threshold!

FDR= 0.05 allows 5% of the discovered biomarkers to be false positives, which is more forgiving than the Bonferroni!

CONCLUSIONS

- our tool allows the **immediate application** of the most commonly used multiple testing correction methods
- **easy data upload:** copy and paste *p*-values
- allows **comparisons across adjustment methods**
- suitable for **all scientific disciplines**

Scan to see the site!

Acknowledgments
This project was supported by the National Research, Development, and Innovation Office, Hungary (PharmaLab, Grant No.: RRF-2.3.1-21-2022-00015). Otília Menyhart was supported by the Janos Bolyai Scholarship of the Hungarian Academy of Sciences and the Hungarian Scientific Research Fund (Grant No.: OTKA FK147194). The support of ELIXIR Hungary (www.bioinformatics.hu) is acknowledged. The authors declare no conflict of interest.

WCE WORLD CONGRESS OF EPIDEMIOLOGY 2024