REPLY TO GOEMAN ET AL.:

Trade-offs in model averaging using multilevel tests

Daniel J. Wilson[a,1]

There were 2 errors in Wilson (1) which I have announced, but I do not accept the 4 claimed by Goeman et al. (2); I rebut them point by point on a figshare page (https://figshare.com/articles/Trade-offs_in_model_averaging_using_multilevel_tests_Appendix_/9699740). However, their letter highlights 2 limitations of the harmonic mean p-value (HMP) procedure that I discuss below with possible countermeasures.

First, they report a model of p-value dependence (herein called "GRN") with parameter $\rho = 0.2$. Unlike the dependence simulated in figure S4 of ref. 1, GRN dependence makes the asymptotically exact HMP anticonservative, producing a type I error rate of 0.09 when all null hypotheses are true, above the theoretical target of $\alpha = 0.05$. This limitation is important but it does not, as claimed, imply an error. The paper states that "the assumptions of equal weights, independence, and identical degrees of freedom can be relaxed." A fair criticism would be that the paper did not qualify that statement sufficiently.

Equation 2.7 of Davis and Resnick (3) implies that the result that $p_{\mathring{p}} \to \mathring{p}$ as $\mathring{p} \to 0$ (equation 5 of ref. 1) holds despite dependence when

$$\Pr\left(p_j < x | p_i < x\right) \to 0, \quad \text{as} \quad x \to 0 \qquad \textbf{[1]}$$

for all p-values $i \neq j$. This condition appears satisfied by GRN (Fig. 1*A*). Simulations confirm convergence of the asymptotically exact test (equation 4 of ref. 1) as α becomes small (Fig. 1*B*). Thus, Eq. **1** formalizes robustness of the HMP to dependence as $\alpha \to 0$, but not necessarily at $\alpha = 0.05$.

Second, Goeman et al. (2) mention that the significance threshold at which the HMP rejects an individual null hypothesis should be more stringent than the Bonferroni threshold, contrary to the paper. This is a special case of the error in which the criterion for declaring set $\mathcal{R}$ significant should be $\mathring{p}_{\mathcal{R}} \leq \alpha_L \, w_{\mathcal{R}}$, rather than $\mathring{p}_{\mathcal{R}} \leq \alpha_{|\mathcal{R}|} \, w_{\mathcal{R}}$, where L is the total number of tests and $\alpha_L \leq \alpha_{|\mathcal{R}|} < \alpha$. Thus the power of the HMP to detect significant groups of hypotheses comes at the cost of reduced power to detect individual hypotheses.

One response is to seek a test that shares some benefits of the HMP while avoiding these issues. Multilevel versions of Bonferroni and Simes' (4) procedures are candidates, as both are robust to GRN dependence (Fig. 1*C*). Their combined p-values for set $\mathcal{R}$ approximate the HMP by bounding it from above (*SI Appendix*, equations 36 and 39 of ref. 1):

$$\begin{aligned} p_{\mathcal{R}}^{\text{Bonf}} &= \frac{w_{\mathcal{R}}}{\max_{i \in \mathcal{R}} \left\{ w_i / p_i \right\}} \\ p_{\mathcal{R}}^{\text{Simes}} &= \frac{w_{\mathcal{R}}}{\max_{i \in \mathcal{R}} \left\{ r_{i\mathcal{R}} \; w_i / p_i \right\}} \\ \mathring{p}_{\mathcal{R}} &= \frac{w_{\mathcal{R}}}{\sum_{i \in \mathcal{R}} w_i / p_i}. \end{aligned} \qquad \textbf{[2]}$$

(Here $w_{\mathcal{R}} = \sum_{i \in \mathcal{R}} w_i$, $\sum_{i=1}^{L} w_i = 1$, and $r_{i\mathcal{R}}$ ranks w_i / p_i within $\mathcal{R}$ from largest, 1, to smallest, $|\mathcal{R}|$.)

The multilevel Bonferroni and Simes methods can therefore be interpreted as approximating the HMP's model averaging approach. These multilevel tests control the strong-sense familywise error rate because a superset of any significant subset must also be significant at threshold $\alpha \; w_{\mathcal{R}}$. This allows the most significant groups of p-values to be identified, so that conclusions are made at the finest resolution permitted by the data, as in the HMP procedure.

Multilevel HMP, Simes, and Bonferroni procedures all have lower power (higher type II error rates) for combining small proportions of p-values, the HMP slightly more so. All procedures have higher power for combining large proportions of p-values, the HMP considerably more so (Fig. 1*D*). Thus, multilevel Bonferroni, Simes, and HMP procedures all offer some benefits of model averaging with different trade-offs in terms of the power of their combined tests and their robustness to dependence.

Acknowledgments

D.J.W. is a Sir Henry Dale Fellow, jointly funded by the Wellcome Trust and the Royal Society (Grant 101237/Z/13/B), and is supported by a Big Data Institute Robertson Fellowship. I thank Jacob Armstrong for comments.

[a] Big Data Institute, Nuffield Department of Population Health, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford OX3 7LF, United Kingdom

Author contributions: D.J.W. wrote the paper.

The author declares no competing interest.

Data deposition: A point-by-point rebuttal to Goeman et al. is available at https://figshare.com/articles/Trade-offs_in_model_averaging_using_multilevel_tests_Appendix_/9699740. R code for figures is available at https://figshare.com/articles/Trade-offs_in_model_averaging_using_multilevel_tests/9699743.

[1] Email: daniel.wilson@bdi.ox.ac.uk.

First published October 29, 2019.

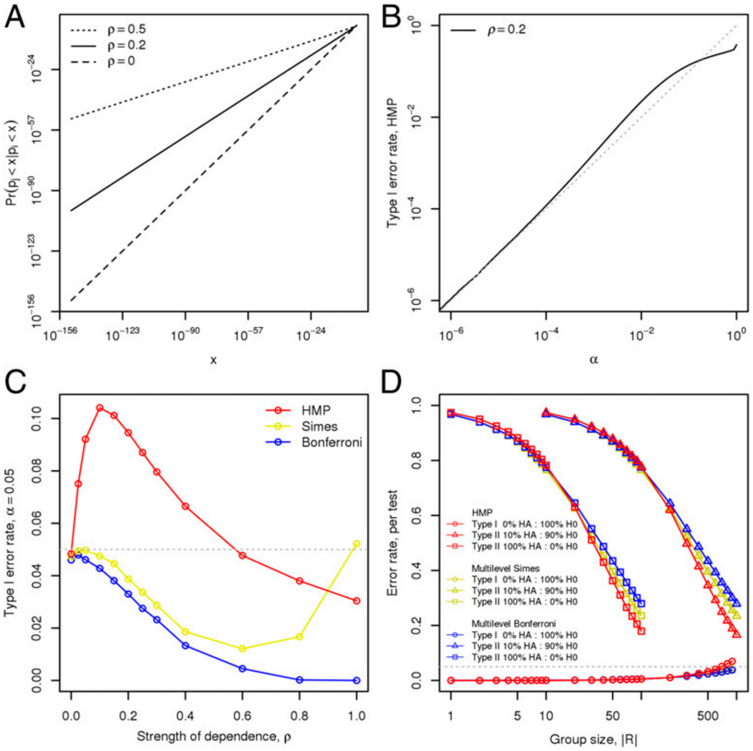

Fig. 1. **Properties of the GRN model. (*A*) GRN satisfies Davis and Resnick's (3) condition. (*B*) For small α, the asymptotically exact HMP procedure converges to the correct type I error rate (10^8 simulations, $L = 10^5$). (*C*) Simes and Bonferroni are robust to GRN dependence (10^4 simulations, $L = 10^5$). (*D*) Error rates for multilevel Bonferroni, Simes, and HMP procedures as a function of the number of *p*-values being combined. The $L = 1{,}000$ normal random variables have means −2.0 and 0.0 under HA and H0, respectively, in proportion 100:900 (10^4 simulations). R code for figures is available at https://figshare.com/articles/Trade-offs_in_model_averaging_using_multilevel_tests/9699743.**

1 D. J. Wilson, The harmonic mean *p*-value for combining dependent tests. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 1195–1200, and correction (2019), **116**, 21948.

2 J. J. Goeman, J. D. Rosenblatt, T. E. Nichols, The harmonic mean *p*-value: Strong versus weak control, and the assumption of independence. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 23382–23383 (2019).

3 R. A. Davis, S. I. Resnick, Limit theory for bilinear processes with heavy-tailed noise. *Ann. Appl. Probab.* **6**, 1191–1210 (1996).

4 R. J. Simes, An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73**, 751–754 (1986).