Aggregating Evidence for Informative Hypotheses: the Value of the Unconstrained Hypothesis

Anita Lyubenova

Introduction

Sometimes studies investigating the same research question utilize very different methodologies, such as design, experimental manipulations, and measurement tools. Bayesian evidence synthesis (BES, Kuiper, Buskens, Raub, & Hoijtink, 2013) is a novel synthesis method that allows to aggregate evidence across such studies. It uses Bayes factors and posterior model probabilities (BFs and PMPs, Kass & Raftery, 1995) to identify which of a set of (informative) hypotheses is most likely to be true in the population when taking all studies together (for a detailed introduction, see Klugkist & Volker, 2022). In individual studies testing an informative hypothesis (e.g. H1: $\mu1>\mu2>\mu3$) against its complement Hc: "not H1", has been promoted because it directly addresses the question whether the hypothesis/theory is correct or not, which is often of interest in applied research (van de Schoot, Hoijtink, Hallquist, & Boelen, 2012; van Rossum, van de Schoot, & Hoijtink, 2013). However, if there are multiple studies it may happen that for some studies H1 is true, while for others Hc is true. In such cases BES will indicate which hypothesis is relatively most likely in "the population" when there is, in fact, no common population for all studies. This issue makes the interpretation of the results from BES ambiguous and might lead to erroneous inferences.

The problem has been illustrated in a simulation study by van Wonderen (2022). The hypothesis H1 regarding a regression coefficient $\beta1>0$ was tested against its complement hypothesis Hc: "not H1" when some studies originated from H1 and others from Hc. Across repetitions H1 was considered to be most likely in the population 50% of the time while in the remaining 50% Hc was more supported (van Wonderen, 2022). Thus, the conclusion drawn from BES would be based on chance in such setting. In a different setting, e.g. with different H1 and Hc, one of the hypotheses could be consistently more supported, even if it is not true for all studies (Klugkist & Volker, 2022).

One potential solution would be to test the unconstrained hypothesis Hu along with H1 and Hc (from now on referred to as conjoined testing). Hu does not impose any restrictions on the parameters, and is, thus, is always true. Therefore, conjoined testing ensures that there is at least one common true hypothesis for all studies. If neither H1, nor Hc is true for all studies, Hu would be preferred (van Wonderen, 2022). However, if H1 (or Hc) is true for all studies, it would be most supported (Volker, 2022). This is the case, because BES balances the fit of a hypothesis, i.e., the degree to which the data supports it, and its complexity, i.e., the degree to which it constrains the parameter space (with less constraints meaning higher complexity). Hu has a perfect fit but is the most complex hypothesis. Thus, a hypothesis with a lower complexity would be preferred, as long as its fit is good enough. Because of this, with conjoined testing the question we ask is *"Which is the most parsimonious hypothesis true for all studies?"*.

The answer to this question is expected to be correct if the number of aggregated studies and the individual sample sizes are very large. However, conjoined testing would only be useful in practice if it works well under realistic conditions. Previous simulation studies testing H1 vs. Hu have shown that increasing number of studies is necessary to accumulate evidence for Hu when it is the only true hypothesis (van Wonderen, 2022; Volker, 2022). Because conjoined testing has not been previously investigated, it is yet unclear under which conditions it can correctly identify the most parsimonious common true hypothesis.

The aim of this study was to scrutinize the usability of conjoined testing for BES in case some studies came from H1 and other from Hc. To this end, (1) conjoined testing was compared to testing only H1 vs. Hc; and (2), it was examined how many studies are necessary for Hu be correctly identified as the most parsimonious common true hypothesis in conjoined testing. These questions were answered for different proportions of studies originating from H1 and Hc.

Methods

To address the aims of this study simulations were performed. Data generation and analyses were performed in R [version 4.2.0; R Core Team (2022)].

Data Generating Mechanism

Study sets were simulated such that they included from 1 to 40 studies. The data for individual studies was generated and analyzed according to multivariate ordinary least squares (OLS) regression model using the simulation procedure of (Volker, 2022). In each model the predictor values were generated from a multivariate standard normal distribution with zero mean vector μ and covariance matrix Σ:

$$\mu = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & & \\ 0.3 & 1 & \\ 0.3 & 0.3 & 1 \end{bmatrix}$$

where .3 is the common correlation between the predictor variables. The ratio between the regression coefficients was specified in the vector B and determined from which population the study came from. For the population in which H1 was true B= [3, 2, 1]; for the population in which Hc was true B= [1, 2, 3]. This means that under H1 $\beta 2 = 2\beta 1$, $\beta 1 = 3\beta 3$ and $\beta 1$ was most strongly related to the outcome Y, followed by $\beta 2$ and $\beta 3$. From this, the population level regression coefficients could be computed as follows:

$$\beta = B\sqrt{\frac{\text{Var}(\hat{Y})}{G'((BB')\ \odot\ \Sigma)G}},$$

where $\text{Var}(\hat{Y})) = R^2$, G is a p $\times$ 1 vector of ones, G $'$ is its transpose and$\odot$ indicates the Hadamard (element-wise) product. R^2 was fixed to a medium effect size of 0.13 (Cohen, 1988). Then, continuous outcomes Y were drawn from a normal distribution

$$Y \sim \mathcal{N}(X\beta, 1 - R^2),$$

where X is the matrix with the predictor values.

All studies had the same sample size (n=350) and were highly powered (91%)[1]. In the first condition there was equal number of studies originating from H1 and Hc (ratio H1:Hc = 1:1); in the second condition there was a larger proportion of studies that came from H1 relative to Hc: for every 4th study Hc was true (ratio H1:Hc = 3:1). For each condition 1000 iterations were performed.

Bayesian Evidence Synthesis

The hypotheses of interest concerned the ordering of 3 regression coefficients: H1: $1 > 2 > 3$, Hc: "not H1", and Hu: $\{1, 2, 3\}$. Bayes factors in each individual study (BFs) were computed via the R package bain (Gu, Hoijtink, Mulder, & van Lissa, 2021). They were used to quantify the relative support for each hypothesis of interest Hi against Hu and were denoted by BFiu. BFiu can be computed as a function of the fit (fi) and the complexity (ci) of Hi (Klugkist, Laudy, & Hoijtink, 2005):

$$BF_{\text{iu}} = \frac{f_i}{c_i}$$

Note that the fit and the complexity of Hu are both equal to 1, therefore the BF for Hu is also always 1.

Posterior model probabilities (PMPs) for each hypothesis were used to compare the relative support across multiple hypotheses in each study. In individual studies the PMPs for each Hi can be computed as

$$PMP(H_i) = \frac{\pi_i BF_{iu}}{\sum_{i'=1}^{m} \pi_{i'} BF_{i'u}},$$

where π_i is the prior model probability for hypothesis Hi, which indicates how likely is this hypothesis before looking at the study data; i' stands for 1, c or u for each hypothesis of interest, respectively, and m is the total number of hypotheses. The higher the value of PMP(Hi), the more plausible it is relative to the remaining hypotheses given the data. BES uses the prior model probabilities π_i to inform the PMPs for each hypothesis from one study to the next. The initial prior model probabilities (before looking at the data from the first study) can be denoted by π_i^0 and are commonly fixed to be equal for all hypotheses. The PMPs resulting from the first study can be denoted by π_i^1 and are then used as prior model probabilities in the second study. Then, PMPs resulting from the second study (π_i^2) are used as prior model probabilities in the third study, and so on. The PMPs of Hi after aggregating T studies are denoted by π_i^T and are computed as follows:

$$\pi_i^T = \frac{\prod_{t=1}^{T} BF_{iu}^t}{\sum_{i'=1}^{m} \prod_{t=1}^{T} BF_{i'u}^t},$$

where BF_{iu}^t is the BF of Hi tested against Hu from study t, i' stands for 1, c or u for each hypothesis of interest, respectively, and m is the total number of hypotheses. Because of the multiplicative nature of the aggregation, the order in which the studies are aggregated is not relevant to the final value.

[1]Power of an individual study was determined as the smaller value from the proportion of times H1 was supported when it was true and the proportion of times Hc was supported when it was true over 10 000 iterations Fu (2022)

Table 1: Simulations set-up

Manipulated factors	Proportion of studies from each population	Tested hypotheses	Number of aggregated studies
Levels	50% from H1-population and 50% from Hc-population	H1 vs Hc	1 through 40
	75% from H1-population and 25% from Hc-population	H1 vs. Hc vs. Hu	

Note:
The simulation had a full factorial design. In H1-population the ratio of the regression coefficients was $\beta 1 : \beta 2 : \beta 3 = 3 : 2 : 1$; in Hc-population $\beta 1 : \beta 2 : \beta 3 = 1 : 2 : 3$. The hypotheses were H1: $\beta 1 > \beta 2 > \beta 3$, Hc: "not H1", and Hu: $\beta 1, \beta 2, \beta 3$.

The aggregated PMPs were computed either for conjoined testing (H1, Hc, and Hu) or only for a test of H1 and Hc. The number of aggregated studies increased by 1 at a time with the maximum of 40 studies. Table 1 provides an overview of the design of the simulation study.

Performance Indices

The performance of BES was evaluated as the proportion of times the PMP for Hu were the highest. It indicates the proportion of times correct inference would be made as Hu is the most parsimonious common true hypothesis in the current simulations. Conjoined testing was compared to testing only H1 vs. Hc by plotting the resulting distributions of PMPs for each hypothesis. The distributions were described by median PMPs and the interval from the 2.5^{th} to the 97.5^{th} percentile of the PMPs for each hypothesis across the 1000 iterations.

Results

The first aim was to compare the distribution of aggregate PMPs for each hypothesis between conjoint testing and H1/Hc testing. The results show that including Hu to the tested hypotheses drastically changed the distribution of the PMPs of H1 or Hc in both H1:Hc ratio conditions. When every second study originated form Hc testing H1 vs. Hc resulted in median PMP for Hc approaching 1 after aggregating only 4 studies (Figure 1A). This is not a desirable result because Hc was true only in half of the studies. In this case Hc was preferred because it had greater complexity than H1, resulting in higher power to support it (Volker, 2022). That is, the same sample size was able to provide stronger support for Hc when Hc was true than for H1 when H1 was true. When Hu was included, the support for Hc quickly decreased over the first 10 studies (Figure 1B). On the contrary, the support for Hu, which was the only common true hypothesis, increased over studies.

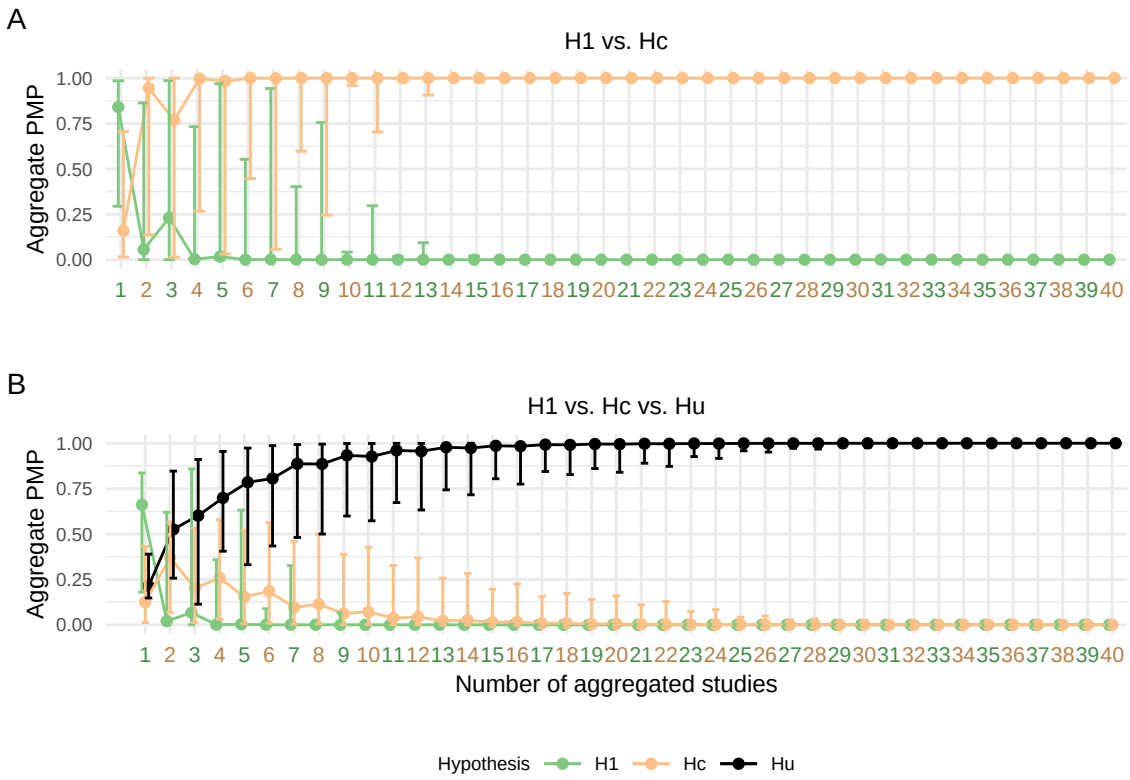

Figure 1: Aggregate PMPs for each hypothesis for increasing number of aggregated studies when the studies originating from H1:Hc=1:1. Points indicate the median PMP per hypothesis;the bars indicate the interval from the 2.5th and 97.5th percentile of the PMPs and contain 95% of the values; the color of the study number indicate the true hypothesis for the corresponding study. The hypotheses were H1: $\beta 1 > \beta 2 > \beta 3$, Hc: "not H1", and Hu: $\beta 1, \beta 2, \beta 3$. Abbreviations: PMP: posterior model probability

Similar patterns were observed when the H1:Hc ratio was increased to 3:1, i.e. when every fourth study originated from Hc (Figure 2). As expected, when testing H1 vs. Hc the support for H1 was higher than for Hc because more studies originated from a H1-population. Again, by adding Hu the support for H1 drops with increasing number of aggregated studies, while the support for Hu increases, as indicated by the median PMPs. However, with larger ratio the spread of the PMPs for H1 and Hu was increased - it covered almost the complete range from 0 to 1.

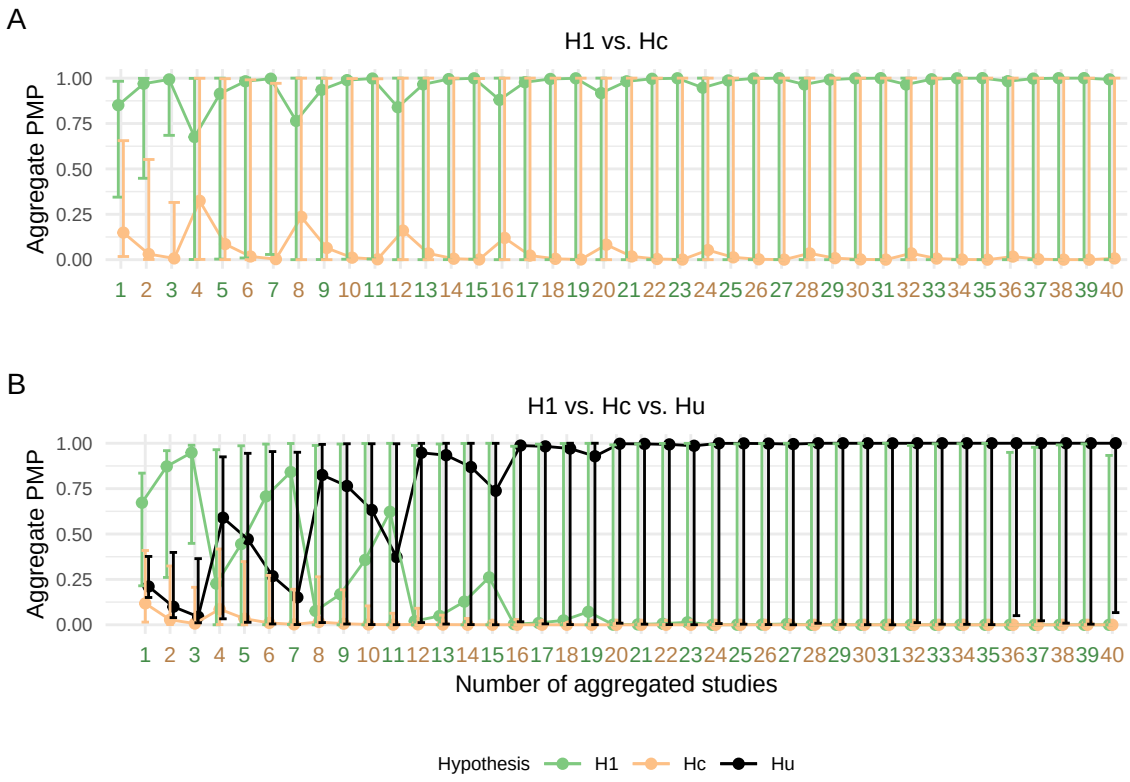

Figure 2: Aggregate PMPs for each hypothesis for increasing number of aggregated studies when the studies originating from H1:Hc=3:1. Points indicate the median PMP per hypothesis;the bars indicate the interval from the 2.5th and 97.5th percentile of the PMPs and contain 95% of the values; the color of the study number indicate the true hypothesis for the corresponding study. The hypotheses were H1: $\beta 1 > \beta 2 > \beta 3$, Hc: "not H1", and Hu: $\beta 1, \beta 2, \beta 3$. Abbreviations: PMP: posterior model probability

The second research question was about the number of studies necessary to obtain the highest support for Hu on aggregate level. When the ratio is 1:1, the proportion of times Hu receives the most support approached 1 for less than 10 studies (Figure 2). In particular, after aggregating 7 studies the proportion was 96.5. Then, the probability of making a wrong conclusion would be less than 5%. However, when the ratio was 3:1 testing Hu had less power because many studies came from H1 and relatively few studies disproved that H1 is a common true hypothesis. This was reflected in the proportion of times Hu received most support, as they did not increase as quickly and monotonously as in the former condition. The "peaks" were at the inclusion of a study that comes from Hc, after which there was a steady decrease while H1-studies were being included. From this, we could see which ratios would be more or less probable to result in high support for Hu. For instance, when aggregating over 7 studies, while only 1 comes from Hc, the probability of concluding that Hu is the only common true hypothesis is 30%, i.e., the probability of making "erroneous" conclusion is 70%. While adding more studies generally reduced the probability of not choosing Hu, it only reached 5% at the 40th study.

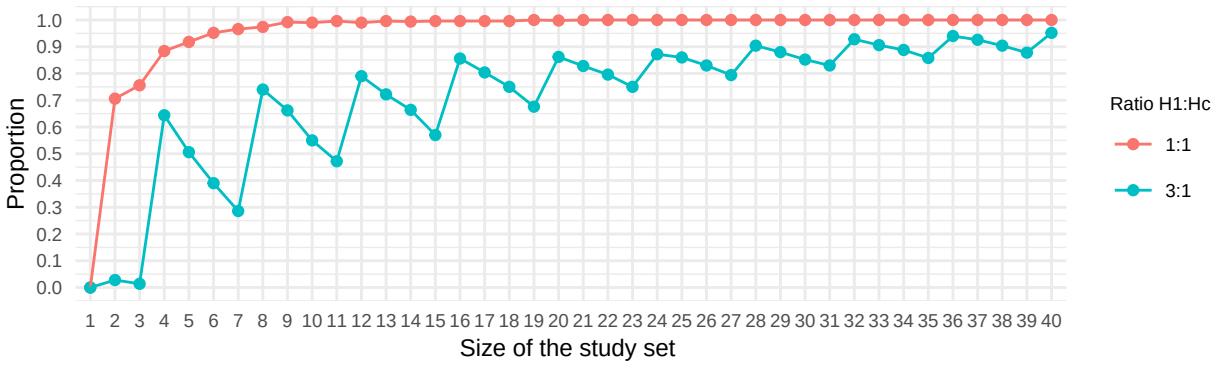

Figure 3: Proportion of times the PMP of Hu was the highest for each condition across number of aggregated studies

Conclusion

This investigation (1) illustrated that testing an informative hypothesis H1 only against its complement Hc in BES can lead to inaccurate conclusions about the truth in the population, and (2) showed that using conjoined testing of H1, Hc and Hu has a potential to solve the problem by indicating the most parsimonious common true hypothesis. The median number of studies in meta-analyses was found to be 6 in a review of the Cochrane Database (Davey, Turner, Clarke, & Higgins, 2011). Conjoint testing seems to work well for such number of studies in the extreme situation when half the studies originated from H1 and the other half from Hc. However, when the studies coming from Hc were few, it took more studies for BES to choose Hu. The performance was then particularly "poor" with small number of aggregated studies. Of course, in real research if, for instance 6 out of 7 studies support H1, a reasonable conclusion would be that H1 holds in the population. The single study that is not in line with it may be due to sampling error or due to characteristics of the study. Thus, such behavior of BES in not that undesirable.

The current results are based only a limited set of conditions with all studies being highly powered. Therefore, it is not yet entirely convincing that conjoined testing is going to be useful in realistic settings with underpowered studies and/or different effect sizes for H1 and Hc. Nevertheless, at this stage it is safe to conclude that more rigorous investigations across more diverse settings are worthwhile.

References

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge. https://doi.org/10.4324/9780203771587

Davey, J., Turner, R. M., Clarke, M. J., & Higgins, J. P. T. (2011). Characteristics of meta-analyses and their component studies in the cochrane database of systematic reviews: A cross-sectional, descriptive analysis. *BMC Medical Research Methodology*, *11*, 160. https://doi.org/10.1186/1471-2288-11-160

Fu, Q. (2022). *Sample size determination for bayesian informative hypothesis testing*. https://doi.org/10.33540/1221

Gu, X., Hoijtink, H., Mulder, J., & van Lissa, C. (2021). *Bain: Bayes factors for informative hypotheses: (Version 0.2.8) [r package]*. Retrieved from https://CRAN.R-project.org/package=bain

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A bayesian approach. *Psychological Methods*, *10*(4), 477–493. https://doi.org/10.1037/1082-989x.10.4.477

Klugkist, I., & Volker, T. B. (2022). *Bayesian evidence synthesis for informative hypotheses: An introduction*. Manuscript submitted for publication.

Kuiper, R. M., Buskens, V., Raub, W., & Hoijtink, H. (2013). Combining statistical evidence from several studies. *Sociological Methods & Research*, *42*(1), 60–81. https://doi.org/10.1177/0049124112464867

R Core Team. (2022). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

van de Schoot, R., Hoijtink, H., Hallquist, M. N., & Boelen, P. A. (2012). Bayesian evaluation of inequality-constrained hypotheses in SEM models using mplus. *Structural Equation Modeling : A Multidisciplinary Journal*, *19*(4). https://doi.org/10.1080/10705511.2012.713267

van Rossum, M., van de Schoot, R., & Hoijtink, H. (2013). "Is the hypothesis correct" or "is it not." *Methodology*, *9*(1), 13–22. https://doi.org/10.1027/1614-2241/a000050

van Wonderen, E. (2022). *Comparing bayesian evidence synthesis to meta-analysis: A simulation study and empirical application*. Manuscript submitted for publication.

Volker, T. B. (2022). *Combining support for hypotheses over heterogeneous studies with bayesian evidence synthesis: A simulation study*. Manuscript submitted for publication.