Assessing replicability across studies: the r-value

Ruth Heller

Tel-Aviv University
www.math.tau.ac.il/∼ruheller

Joint work with Marina Bogomolov and Yoav Benjamini

Quantifying replicability across studies

- The Royal Society of London introduced in the 17th century the convention of counting a discovery as scientific if it was reported by at least two independent studies.
- Whether the effect is present in a study may depend on:
  - the specific cohorts in the study, that are from specific populations exposed to specific environments.
  - the specific experimental protocol in the study.
  - the specific care givers in the study.
  - ...
- We need an objective way to quantify the evidence that findings were replicated across studies: the r-value.

Outline

1. The r-value for the single endpoint design.
2. The FWER/FDR r-values for the primary to follow-up design.
3. More designs, and conclusions.

Replicability claims for the single endpoint design

Let $\theta_1, \ldots, \theta_N$ be the unknown effects in the N studies, where a value of say 1 means that there is no effect.

Is $\theta_i > 1$ in at least two studies?

- This is a minimum requirement for replicability.
- The replicability claim is true if indeed $\theta_i > 1$ in at least two studies.
- The replicability claim is false if $\theta_i \leq 1$ except possibly in one study.

Key difference from meta-analysis

- A meta-analysis p-value tests the intersection null hypothesis that $\theta_i \leq 1$ in all the studies.
- A meta-analysis finding is a claim that $\theta_i > 1$ in at least one study, not a replicability claim.
- This claim may be true even when the replicability claim is false:
  - if $\theta_i > 1$ in exactly one study, and $\theta_i \leq 1$ in all other studies.

The *r*-value for the single endpoint design

The *r*-value is

- the smallest significance level at which we claim replicability.
- the *p*-value for the test of the null hypothesis that $\theta_i \leq 1$ except possibly in one study.

Claiming replicability whenever the *r*-value $\leq \alpha$, guarantees control over a false replicability claim at level α.

The r-value computation for the single endpoint design

For two studies if the finding is significant in both, replicability is also established, in the sense that the union hypothesis $H_{01} \cup H_{02}$ is rejected.

- The r-value is $\max(P_1, P_2)$.

For $N > 2$ studies:

- For every subset of $N - 1$ studies, compute a meta-analysis p-value for testing the global (intersection) null hypothesis.
- The r-value is the largest of the N meta-analysis p-values.

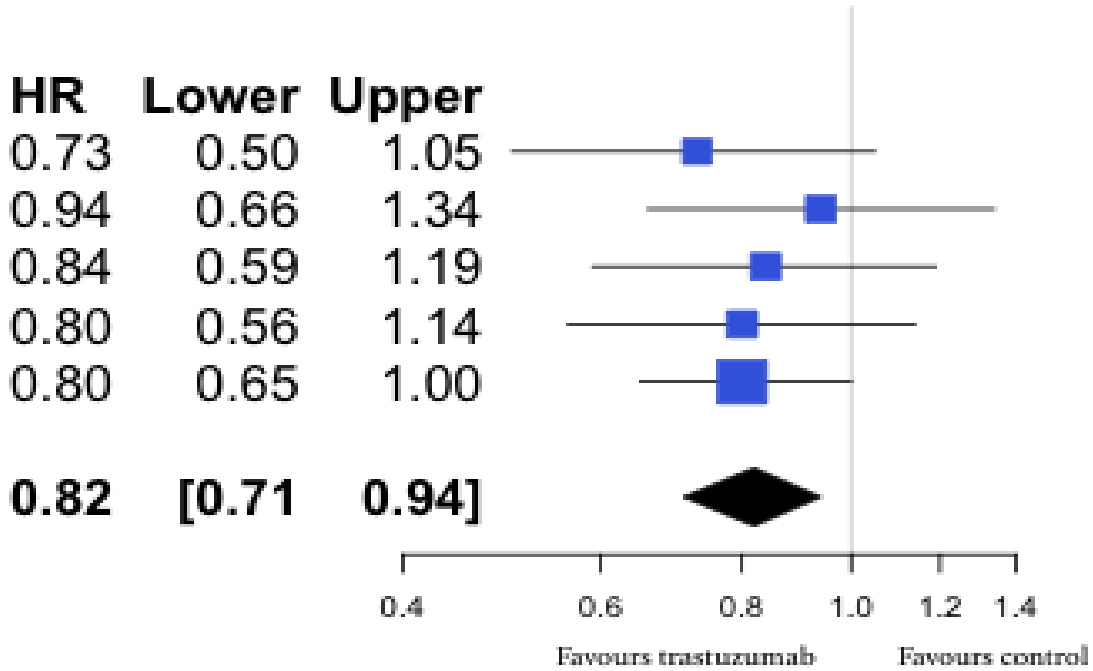
Example application of the single-endpoint design

Cochrane Reviews:

- Systematic reviews of all clinical studies on a topic that meet certain criteria, in order to establish whether or not there is conclusive evidence about a specific treatment.
- They are published online in The Cochrane Library.
- They are designed to facilitate the choices that practitioners, consumers, policy-makers and others face in health care.

A Cochrane Review with evidence of replicability

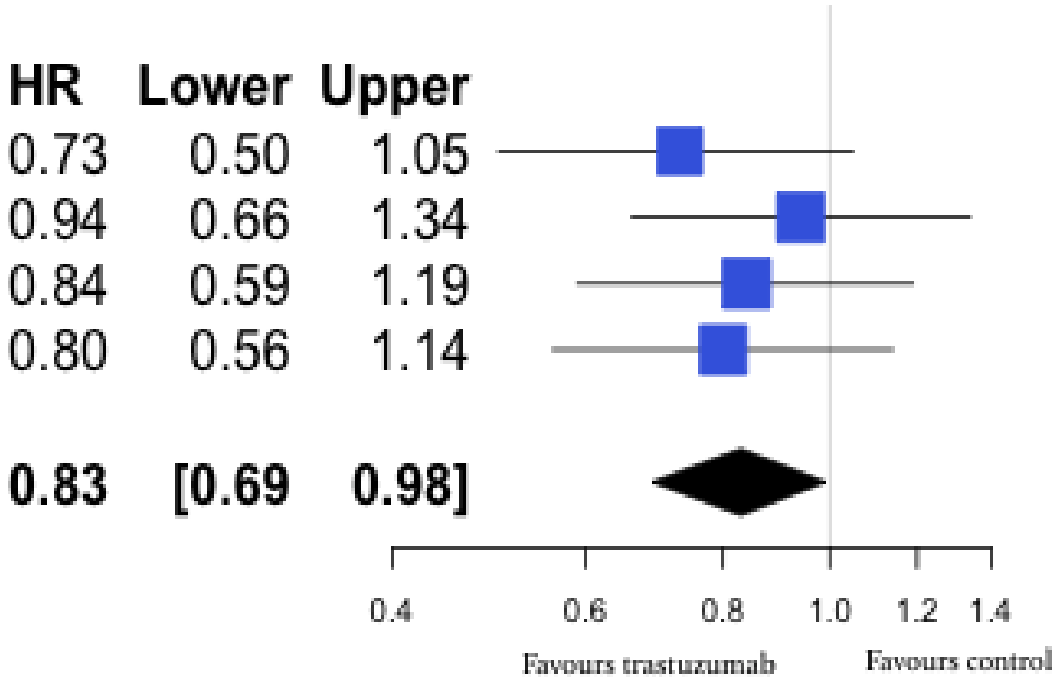
- Review objective: to assess the efficacy of therapy with trastuzumab in women with HER2-positive metastatic breast cancer.
- Significant effect in meta-analysis:



- Note that only one single study is (barely) significant.

A Cochrane Review with evidence of replicability

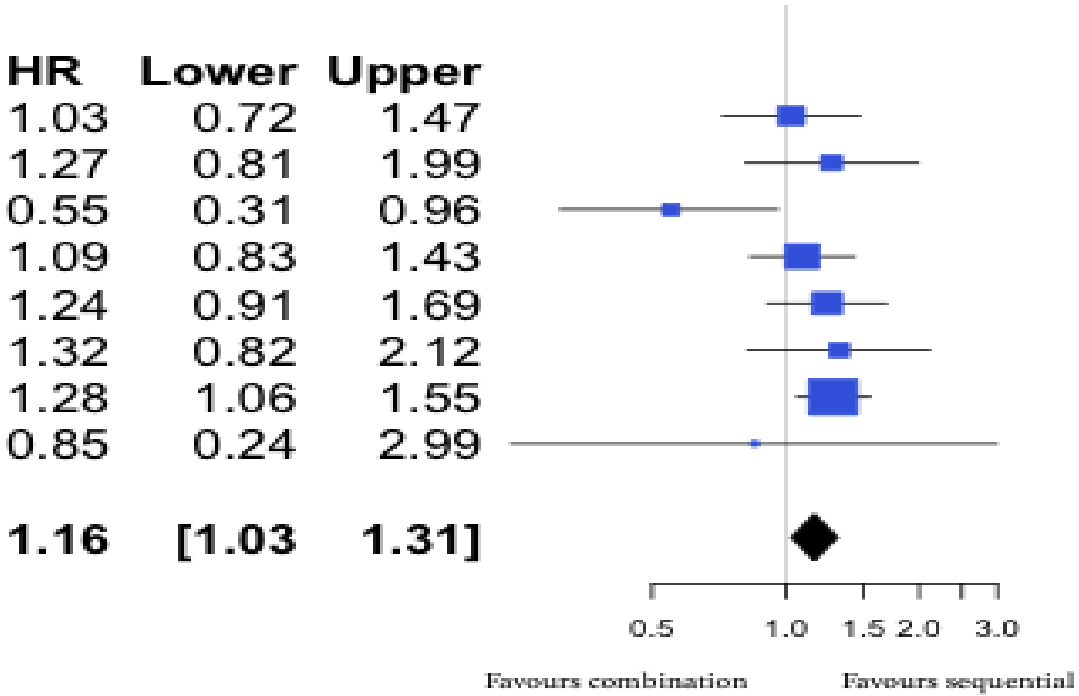
- Review result: Trastuzumab improved overall survival in HER2-positive women with metastatic breast cancer.
- Even when removing each single study, there is a significant effect in the meta-analysis.
- Removing the single significant study:



- The r-value is 0.03549.

A Cochrane Review with no evidence of replicability

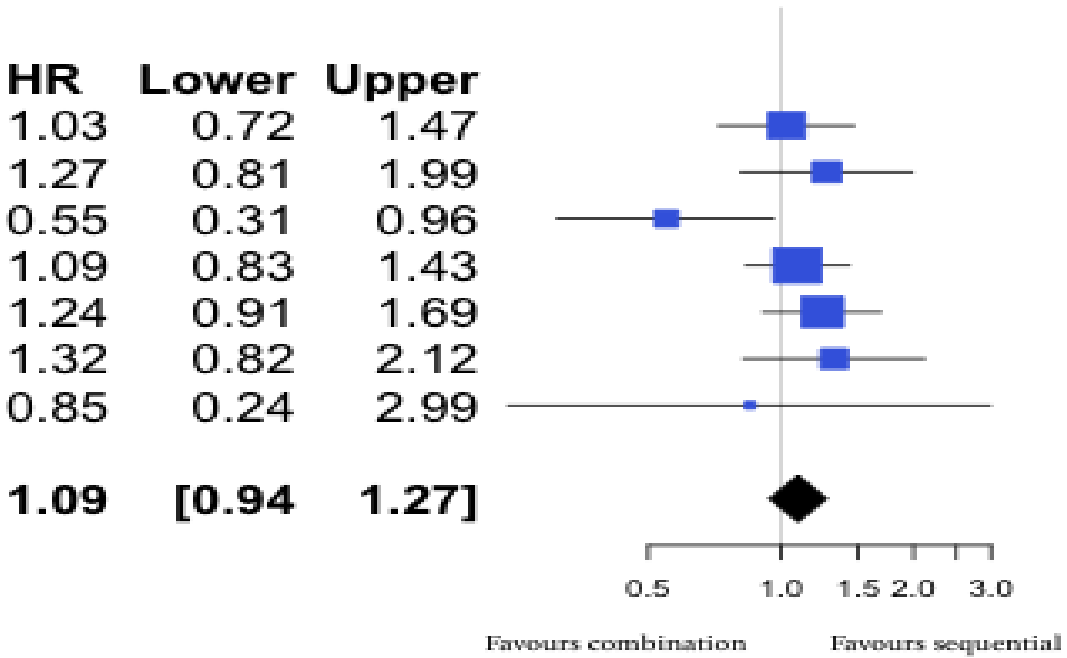
- Review objective: to assess the effect of combination chemotherapy compared to the same drugs given sequentially in women with metastatic breast cancer.
- Significant effect in meta-analysis:



- Note that different studies seem to have different effects!

A Cochrane Review with no evidence of replicability

- Review result: the combination arm had a higher risk of progression than the sequential arm.
- But removing study number 7, there is no longer a significant effect in the meta-analysis:



- The *r*-value was 0.24.

The usefulness of the r-value in Cochrane Reviews

- The meta-analysis cannot serve as evidence towards replicability of effects across studies: one study may drive the entire conclusion.
- The scientific evidence supported by the r-value is stronger than that offered by the meta-analysis p-value.
- A straightforward extension: the $r_{u/N}$-value, that quantifies the evidence that the finding has been replicated in at least u out of N studies

The primary to follow-up design

1. Examine $m \gg 1$ features in the primary study.
2. Follow-up on a fraction of the m features in an independent study.
3. The decision of which features to follow-up on is based on the results from the primary study.

Example application for the primary to follow-up design

Genome-wide association studies (GWAS):

- In the primary study, hundreds of thousands of single nucleotide polymorphisms (SNPs) are tested for association with the phenotype.
- In the follow-up study, a handful of promising SNPs are tested for association with the phenotype.
- If the p-value is fairly small in the follow-up study it is informally regarded as a replicated finding.

Table 1 of Bis et al., 2012: GWAS of hippocampal volume

- Bis et al. (2012) examined $2.5 * 10^6$ SNPs in a primary study, and five SNPs in four loci with primary study p-value$<= 1/(2.5 * 10^6)$ were followed-up.

Locus	Gene	Primary p-values	Follow-up p-values	Meta-analysis p-values	$2.5 \times 10^6 \times$ Meta-analysis p-values
2q24	DPP4	5.2×10^{-8}	0.7	2.9×10^{-7}	0.725
9q33	ASTN2	1.0×10^{-7}	0.2	1.0×10^{-7}	0.25
12q14	MSRB3	5.5×10^{-9}	0.002	5.3×10^{-11}	0.00013
	WIF1	2.2×10^{-8}	0.0007	7.1×10^{-11}	0.00018
12q24	HRK	4.8×10^{-8}	5.8×10^{-5}	2.9×10^{-11}	0.00007

How to compute r-values for this design?

Notation

- p_{1j} are the primary study p-value for SNP $j \in \{1, \ldots, m\}$.
- $\mathcal{R}_1$ is the set of SNPs selected for follow-up, and it is determined by $\{p_{1j} : j = 1, \ldots, m\}$.
- p_{2j} is the follow-up study p-value for SNP $j \in \mathcal{R}_1$.
- FWER: the probability of at least one false replicability claim.
- FDR: the expected fraction of false replicability claims among the replicability claims.
- $l_{00} \in [0, 1)$ is a lower bound on the fraction of features, out of the m features examined in the primary study, that are not associated with the phenotype.
  - $l_{00} = 0.8$ is a conservative choice for GWAS on the whole genome.
- The r-value is the smallest level (of FDR/FWER) at which we can say that the finding has been replicated.

The Bonferroni r-value

For every feature $j \in \mathcal{R}_1$:

- Compute the e-value

$$e_j = \max\left(2[1 - l_{00}(1 - q/2)]p_{1j},\ 2\frac{|\mathcal{R}_1|}{m}p_{2j}\right), j \in \mathcal{R}_1.$$

- Let $r_j = e_j m$.
- The Bonferroni r-value is the smallest level q s.t. $r_j \leq q$.

The replicability claims at level α are all SNPs with r-values$\leq \alpha$.

The FWER on false replicability claims is controlled at level α .

The FDR r-value

1. For every feature $j \in \mathcal{R}_1$ compute the following e-values

$$e_j = \max\left(2[1 - l_{00}(1 - q/2)]p_{1j},\ 2\frac{|\mathcal{R}_1|}{m}p_{2j}\right), j \in \mathcal{R}_1.$$

2. Sort the e-values $e_{(1)} \leq \cdots \leq e_{(R_1)}$.
3. The ith largest r is

$$r_{(i)} = \min_{j \geq i, j \in \mathcal{R}_1} \frac{e_{(j)}m}{j}.$$

4. The FDR r-value is the smallest level q s.t. $r_j \leq q$.

We provide a web applet for computing the r-values.[1]

[1] `http://www.math.tau.ac.il/~ruheller/App.html`.

Table 1 of Bis et al., 2012

Input for r-value computation: $l_{00} = 0.8$.

Gene	Primary p-value	Follow-up p-value	$2.5 \times 10^6 \times$ meta-analysis p	Bonferroni r-value	FDR r-value
DPP4	5.2×10^{-8}	0.7	0.725	1.0	1.0
ASTN2	1.0×10^{-7}	0.2	0.25	1.0	0.5000
MSRB3	5.5×10^{-9}	0.002	0.00013	0.020	0.011
WIF1	2.2×10^{-8}	0.0007	0.00018	0.023	0.011
HRK	4.8×10^{-8}	5.8×10^{-5}	0.00007	0.053	0.017

Validity of FDR replicability analysis

For the procedure that declares as replicated all findings with FDR r-values $\leq q$, we have a theorem[2] that shows that:

- If the p-values in the primary study are independent, and the p-values from the follow-up study are jointly independent or are positive regression dependent on the subset of null hypotheses, then the FDR on false replicability claims is controlled at level q.
- For arbitrary dependence among the p-values in the primary study,
  - Replacing m by $m^* = m \sum_{i=1}^{m} 1/i$ in the r-value computation, the FDR on false replicability claims is controlled at level q.
  - If the features selected for follow-up are at most a fixed threshold $t \in (0, 1)$, the price paid is smaller than $\sum_{i=1}^{m} 1/i$.

[2] Theorem 2.1 in `http://arxiv.org/pdf/1310.0606v2.pdf`

Robustness of FDR replicability analysis

Modifying the r-value computation may be unnecessary for GWAS:

- Our simulated GWAS examples suggest that the procedure is valid for the type of dependency that occurs in GWAS.

We conjecture that our proposal is robust to deviations from independence incurred in practice:

- It is similar to the (very robust) Benjamini-Hochberg procedure, with the important difference of using e-values instead of p-values.

Example of GWAS of Crohn's disease (CD)

- Barret et al. (2008) examined 635546 SNPs in a primary study.
- SNP j was followed if $p_{1j} \leq 5 \times 10^{-5}$ and it was one of the two SNPs with smallest primary study p-values in a region.
- 126 SNPs were followed up, in 63 distinct regions.
- Barret et al. listed the 30 SNPs that had follow-up study p-values below $0.05/126$ as convincingly replicated CD risk loci.
- Setting $l_{00} = 0.8$ for the FDR r-values computation:
  - We decide that 52 SNP findings are replicated at r-values ≤ 0.05.
  - Taking the conservative approach to dependency within primary study, we decide that 34 SNP findings are replicated at r-values ≤ 0.05.

Multiple studies each examining all the features

Examine $m \gg 1$ features in each of N studies.
Relevant in many fields, e.g. for large "omics" consortia, where data on the same features is available from multiple centers.
Relevant methodologies:

- Heller and Yekutieli (2013) estimate the Bayes FDR on replicability claims, for $N \geq 2$.
- For $N = 2$:
  - Li et al. (2011) give a method based on relative ranking to control the "irreproducibility discovery rate", which is used in the ENCODE project.
  - Bogomolov and Heller (2013) suggest methods for FWER and FDR control for replicability analysis; work in progress.

Summary

- We suggested an easy, rigorously motivated way, to quantify the evidence of replicability: the r-value.
- Reporting the r-values gives credibility to replicability claims.
- FDR replicability analysis may be more powerful than current common practices for making replicability claims,
  while at the same time it has the theoretical guarantees of control over false replicability claims.
- The r-value computation depends on the design of the replicability problem, as well as on the choice of error measure to control.

Paper in Nature Genetics

"Common Variants at 12q14 and 12q24 are associated with hippocampal volume", Nature Genetics 44, 2012.

Joshua C Bis, Charles DeCarli, Albert Vernon Smith, Fedde van der Lijn, Fabrice Crivello, Myriam Fornage, Stephanie Debette, Joshua M Shulman, Helena Schmidt, Velandai Srikanth, Maaike Schuur, Lei Yu, Seung-Hoan Choi, Sigurdur Sigurdsson, Benjamin F J Verhaaren, Anita L DeStefano, Jean-Charles Lambert, Clifford R Jack Jr, Maksim Struchalin, Jim Stankovich, Carla A Ibrahim-Verbaas, Debra Fleischman, Alex Zijdenbos, Tom den Heijer, Bernard Mazoyer, Laura H Coker, Christian Enzinger, Patrick Danoy, Najaf Amin, Konstantinos Arfanakis, Mark A van Buchem, Rene F A G de Bruijn, Alexa Beiser, Carole Dufouil, Juebin Huang, Margherita Cavalieri, Russell Thomson, Wiro J Niessen, Lori B Chibnik, Gauti K Gislason, Albert Hofman, Aleksandra Pikula, Philippe Amouyel, Kevin B Freeman, Thanh G Phan, Ben A Oostra, Jason L Stein, Sarah E Medland, Alejandro Arias Vasquez, Derrek P Hibar, Margaret J Wright, Barbara Franke, Nicholas G Martin & Paul M Thompson for Enhancing Neuro Imaging Genetics through Meta-Analysis (ENIGMA) Consortium, Michael A Nalls, Andre G Uitterlinden, Rhoda Au, Alexis Elbaz, Richard J Beare, John C van Swieten, Oscar L Lopez, Tamara B Harris, Vincent Chouraki, Monique M B Breteler, Philip L De Jager, James T Becker, Meike W Vernooij, David Knopman, Franz Fazekas, Philip A Wolf, Aad van der Lugt, Vilmundur Gudnason, W T Longstreth Jr, Matthew A Brown, David A Bennett, Cornelia M van Duijn, Thomas H Mosley, Reinhold Schmidt, Christophe Tzourio, Lenore J Launer, M Arfan Ikram & Sudha Seshadri for the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium

References I

Bogomolov, M. and Heller, R. (2013).

Discovering findings that replicate from a primary study of high dimension to a follow-up study.

Journal of the American Statistical Association, 108 (504): 1480–1492, .

Barret et al. (2008).

Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease.

Nature Genetics, 40: 955–962.

Heller, R. and Yekutieli, D. (2014).

Replicability analysis for genome-wide association studies.

The Annals of Applied Statistics, 8 (1): 481–498.

Li, Q. and Brown, J. and Huang, H. and Bickel, P. (2011)

Measuring reproducibility of high-throughput experiments.

The Annals of Applied Statistics, 5(3): 1752–1779.