# Can Social Dependency Induce (Unintended) Robustness in a Meta-Analysis?

Donghyun Kang[1]

The Department of Sociology

The University of Chicago

## Abstract

The validity of scientific knowledge is based on its replicability/reproducibility; however, the concern regarding the reproducibility has been a big issue among research communities. One way to tackle this issue is to conduct a meta-analysis, pooling research outcomes. When randomized control treatments were employed in each study, it is believed that we can obtain a more precise estimate of a treatment effect. In this study, I challenge this assumption. First, we find that about 28% of statistically significant claims ($p < 0.05$) can be insignificant when the Leave-One-Out procedure is applied to meta-analyses on the biomedical claims (studied by more than three but less than 50 randomized controlled trials with dichotomous outcomes) in the Cochrane Database of Systematic Reviews. This suggests that a conclusion of a meta-analysis can be highly sensitive to the authors' selection of studies. Second, we also find that the robustness of the claims is positively associated with the co-authorship density of the researchers. It implies that the authors of original research articles tend to produce similar research results despite their uses of randomized controlled trials. This study shows that we need to be cautious in terms of interpreting conclusions from meta-analyses. Furthermore, we argue that the funding organizations need to diversify their supports in order not to generate an echo-chamber in the biomedical research community.

[1] kangd@uchicago.edu

## Introduction

One of the most important features that distinguish modern society from pre-modern one is how we produce and apply knowledge. As a human activity to discover reliable knowledge in a systematic way, modern science is often related not only to the production of pure knowledge about the world; but it also shapes our daily practices in our life-world. And it is hard to deny that doing randomized experiments is at the core of modern science. Even though it is not the only way of doing science (e.g., paleontologists and astronomers rely on fossil records and radial signals), conducting randomized experiments is naturally and often referred as the best possible (even though it is not perfect either) way of establishing causal relationships. In theory, scientific knowledge produced by scientists using experiments should be replicated by other scientists if scientific knowledge is truly cumulative; but it is not always the case.

One way of tackling this issue is to pool the research results and conduct a meta-analysis to get a better estimate of treatments by increasing the statistical power. This has been especially quite common in the biomedical research community, particularly because health professionals need evidence to give medical practice. In this context, Cochrane Review Group, founded in 1993, has been publishing systematic meta-analyses to help a better decision making for various stakeholders such as doctors, nurses, patients, and policymakers, drawing upon clinical trials conducted worldwide. But in this paper, we will first demonstrate that their efforts to establish evidence by pooling results from (supposedly) independent trials can be quite sensitive to the Leave-One-Out procedure, using the method of *r*-value (Shenhav et al. 2015). Then, we will show evidence suggesting that the robustness of the claims might also be induced by social dependence (i.e., the co-authorship relationship between them) of the authors. This does not necessarily mean that we should not trust the significant

conclusions from the systematic meta-analyses, but we argue that it suggests a possible social force that could lead to a consensus.

## Backgrounds

### *Replication Crisis in Science*

One hot issue in the contemporary scientific enterprise is the high level of concern for reproducibility/replicability[2] of published studies. Of course, it is not unusual to observe some degrees of errors in science and the modern organization of science is believed to handle with the self-correcting institution: we have the peer review process and the open discussion after the publication of the results. The former aims to control the quality of the scientific publication and the latter often leads to open debates regarding certain topics and sometimes results in retraction of papers, if the errors are found to be serious.[3] These practices give scientific knowledge the certain degree of authority.

However, the issue of non-reproducible science is far beyond errors that can be conceived as white noise. One recent survey conducted by *Nature* in 2016 suggests that 90% of respondents comprising of biologists, chemists, physicists, biomedical researchers, etc., expressed their concern about the reproducibility crisis (Baker 2016). In social sciences, the issue was strongly raised from psychology (especially in the field of social psychology), starting from the early 2010s. For example,

[2] In this paper, the two terms are used interchangeably. More nuanced differentiation is possible. For example, replicability can refer to the situation in which the identical results are achieved with the same experimental settings whereas reproducibility is achieved when more diverse sets of conditions produce the consistent result. However, we decided not to differentiate the term given that articles released through Nature and Science do not differentiate the two terms in a strict sense.

[3] In Oct 2018, a database of retracted research papers was released (http://retractiondatabase.org/).

Pashler et al. (2012), the editors in *Perspectives on Psychological Science*, called serious attention to the non-replicability in psychology. The "Reproducibility Project: Psychology", launched by the Center for Open Science in 2011, concluded that only 35 studies out of the 97 studies published in leading psychology journals with significant results were replicated (Open Science Collaboration, 2015). The field of biomedical research is not an exception. For example, examining 49 studies published from 1990 to 2003 and cited more than 1,000 times (45 out of 49 studies had statistically significant outcomes), Ioanndis (2005a) found that only 44% of the claims were replicated by the subsequent studies. Begley (2012) also argued that only 11% of pre-clinical cancer research was replicated.

One tightly related issue with the replicability of research outcomes is the application of statistics in research practice. The standard statistical procedure allows two types of errors - the type 1 error (false-positive) and the type 2 error (false-negative; i.e., statistical power). While experimental physicists impose a stricter standard [4] in terms of reporting research outcomes to prevent false-positive findings, medical and social science research generally accepts the 5% significance level (i.e., the type 1 error) and this practice is often blamed one of the underlying reasons leading to the non-replicability crisis (Ioannids 2005b; 2016).

Several statistical techniques have been of course developed to overcome the limitation of a single study with small sample size and generate more precise ranges of effect sizes. The False Discovery Rate (Benjamini & Yosef 1995) approach and multilevel regression (Hox et al., 2017) are often employed in medical research. The former method attempts to solve the multiple comparison problems by producing the corrected p-value which takes into accounts the fact that repeating the same

[4] The threshold for "evidence of a particle," corresponds to $p = 0.003$, and the standard for "discovery" is $p = 0.0000003$. (citation needed)

randomized trials can increase both type 1 and 2 errors. The latter method, the multilevel framework is also often employed to evaluate the effect of a treatment by pooling the research outcomes from multiple clinical trials. The fixed-effect meta-analysis posits the invariant treatment effect across the different trials and the random-effect model allows variation of a treatment effect.

Admittedly, all the approaches are invaluable toolkits to establish scientific validity. However, a problematic issue is that they assume that each trial is independent of each other, which might not always be the case considering the domination of team science in modern science.

*The Dominance of Team Science*

A crucial trend in contemporary science is the increase in collaboration. This is somewhat contrasting to a popular image of solitary scientists in the $17^{th}$-$18^{th}$ centuries who might have been struggling with experimental equipment in solitary. Collaboration became an essential part of contemporary science partly due to the amounts of capital - both in terms of experimental instruments and various level of researchers including postdocs and graduate students - were required for doing scientific research (Hagstrom 1964). Especially the wartime science during the second world war and the subsequent period of the cold war are known to have contributed to the emergence of "big science" (de Solla Price 1963, Wolfe 2012) supported by the funding organizations such as the National Science Foundation (NSF) and the Defense Advanced Research Projects Agency (DARPA) in the U.S. Since the "violent transition" (de Solla Price 1963) from "solo investigators" (Leahey 2016) to team scientists in the mid-20th century, the trend has been continuing; for example, the proportion of coauthored papers in JSTOR increased to over 60% in 2011 from 6% in 1900 (West et

al. 2013). And it is reported that the trend holds across almost of all academic disciplines (Leahey, 2016).

Beyond the historical context facilitating collaboration in scientific research, it is pointed out that the development of the communication technology reducing the cost of collaboration (Binz-Scharf et al. 2015) and the intensified competition among scientists is also considered as one of driving forces towards more collaborations. However, the oft-cited fundamental benefit from collaborative research spans to the potential synergies based on the "division of cognitive labors" (i.e., specialization) of scientists with different types of expertise (Leahey 2016). In addition to this, it has been argued that collaborative research activities from a larger team are in a better position for discovery (Esparza & Yamada 2007) and they produce more reliable research in a sense that errors and bad ideas could be filtered out in advance (Clark & Llorens 2012, Bikard et al. 2015). In this context, various policy recommendations and methods to promote collaboration in science have been proposed (Stokols et al., 2008, Croyle 2008, Bennett et al. 2012).

Despite this enthusiasm, recent studies demonstrate some inherent issues of doing team science as well. For example, embedding the co-authorship network of physicists of the 20$^{th}$-century into a 2-dimensional Poincare disk at the institution level, Wu et al. (2018) find that the dense connection among physicists created by the growing collaboration has resulted in less diversity (or contraction) in research agendas. In the same vein, drawing upon scientific publications, patents, and software codes spanning from 1954 to 2014, Wu et al. (2018) argue that large teams tend to conduct research conservatively in a sense of choosing their research topics, whereas small teams tend to produce disruptive work more frequently than larger teams.

The issue we would like to address goes one step further from the mentioned above potential disadvantages of doing team science per se. We further hypothesize that the increased social density among researchers can produce an echo-chamber, leading to the under-dispersion of the research outcomes which stems from the violation of the wisdom of crowds.

*The Violation of the Wisdom of Crowds with Increased Collaboration*

The "wisdom of crowds" refers to a social phenomenon in which the aggregated judgments or decisions from independent individuals often leads to better results than those made by a single individual (Surowiecki, 2005). The wisdom of crowds was first documented by Francis Galton (1907), who was surprised by the fact that the collective guessing about weights of an ox was almost close to the actual weight when "the middlemost estimate" (i.e., the median) was considered. In his report, Galton concluded that this showed the "more creditable to the trust-worthiness of a democratic judgement than might have been expected" (Galton 1907: 451).

One of the assumptions in the original ox weight-guessing that Galton observed is the independence of the judgment which prevents the group thinking or social influence that can result in bias. In addition to this, the diversity of opinions, thoughts, or guessing matters as well; to wit, the wisdom of crowds is better than an individual guess on the condition that independent guesses are aggregated in a macro scale. In other words, the wisdom of crowds works when decentralized individuals with diverse viewpoints make their judgments independently and aggregate them into a collective one.[5] While the debates regarding the relationship between social influence and the quality of

[5] More formally, $\text{group error} = \sum \text{individual error} - \text{diversity}$ (Page, 2008).

collective judgment are proceeding recently (see Lorenz et al. (2010) and Becker et al. (2016) for the negative impact of social influence on collective judgment; Almaatouq et al. (2017) for the conditions that mitigate the negativity), the research regarding the wisdom of crowds basically argues that a group of people can make a good decision when appropriate social conditions are met.

At first glance, the wisdom of crowds does not seem related to the field of science considering that scientists are supposed to have expertise in their fields: they are not mere "crowds". But the empirical evidence suggests that the collective knowledge production system of science is also conceived similarly to the wisdom of crowds. Collaboration such as coauthoring a paper is not the only case; beyond this, modern science constitutes "Invisible colleges" (Crane, 1972) as an autonomous self-regulating system of knowledge production. In other words, the validity of the collective scientific endeavor is based on the "collective wisdom of scientists" (Surowiecki, 2005: 169).

Evidence, however, also suggests that scientists are not free from social influence and bias. For example, studies have shown low inter-rater reliability during the peer review processes in scientific publications (Mahoney, 1977; Emerson et al., 2010, Lee et al. 2013). Moreover, a recent study on the grant reviews process at the National Institute of Health showed the trade-off between the expertise and bias of the grant reviewers (Li, 2017). These empirical studies display different pictures from what Merton (1942) hypothesized about modern science with the normative ethos of universalism, communism, disinterestedness, and organized skepticism.

In sum, the reliability of science can be also affected by the social structure of the field of science. Thus, we need to take into accounts social aspects of science to explain how contemporary science is organized and to evaluate the robustness of science.

**Data: The Cochrane Database of Systematic Review**

The Cochrane Database of Systematic Review (CDSR), published by Cochrane Review Group, is one of the leading journals providing systematic reviews on human health intervention and diagnosis to inform diverse sets of stakeholders such as policymakers, patients, health professionals, starting from April 1995[6]. The Cochrane Review Group originally was found in the U.K. as a non-profit organization; now it has 53 different groups[7] with over 30,000 expert volunteers in health science across the world (Cochrane, 2019). The CDSR has been regarded as one of the most reliable medical resources that guide the evidence-based medical practice (Handoll et al. 2008) since its foundation.

[Figure 1 here]

Each review provides the meta-analysis *p*-values and confidence intervals based on the methodology proposed by Higgins et al., (2011), combining the statistical information from the studies that a review attempts to examine. For instance, review CD008792 published in 2013 evaluates the efficacy of the combination of chemotherapy in women with metastatic breast cancer comparing to the sequential application of the same drugs, using 12 studies. Figure 1 shows the overall response rates of the two different therapies from those compiled by the authors of the review.

In total, we initially collected 4,543 Cochrane review articles, which were published before Nov 18, 2017. Each review has its history. Protocols in a preparation stage are first produced by the authors (which sometimes are not published) and they conduct a review after a set of original research articles are identified. This means each Cochrane review article at least has two elements in its history. If a

[6] From 1995 to 2009, CDSR was published quarterly; since then, it has been published in a monthly basis.

[7] The full list of the review groups can be found in https://www.cochranelibrary.com/about/cochrane-review-groups. Note that some topics such as influenza virus can be address by several groups including Acute Respiratory Infections Group, Childhood Cancer Group, etc.

major updated is conducted later, a newer version will replace the last one, but it maintains the same review ID. Minor amendments such as changes in formatting and an addition of a summary in plain English can also take place, but it does not lead to an updated version of a review. For instance, review CD001321 investigating the efficacy of Cranberries in preventing urinary tract infections was first published in 2003 but updated in 2004, 2008, and 2012. Considering this, we developed a web-scraper using Selenium module in Python to collect all the version histories of 4,543 reviews (which led to 8,212 different versions of the 4,543 reviews) and selected the most recent articles before Nov 18, 2017. The mean of the number of authors of the most recent reviews is 4.59 (std = 2.11, min = 1, max = 41).

The reference sections of each Cochrane review disclose both the included and excluded original studies.[8] The metadata of these original studies (e.g., the individual studies in Figure 1) in each review, including the names of the authors, the titles, the journals' names, and the publication year, and the PMID of original research articles (when it is possible) were also collected using a web-scraper. In total, we identified 128,563 original articles included in 4,543 original studies. The PMID is of importance here. Each PMID can be used to identify a medical research article in the MEDLINE database, which is co-maintained by the National Library of Medicine (NLM) and the National Center for Biotechnology Information (NCBI) at the National Institutes of Health (NIH) of the United States. The PMID of each original research article examined by each Cochrane review allowed us to collect the additional data

[8] The authors first collect the relevant published research and sort them with some evaluation criteria to safeguard the scientific validity or to guarantee fair comparisons. The authors decide to exclude studies when they do not meet the criteria or if the study designs are inconsistent with other studies. A reader can find detailed inclusion criteria in each review. In addition to this, note that each review only includes one of the studies duplicated by the same authors. For example, the reference section of the review CD001321 includes the two studies conducted by Salo et al. (2010) and Salo et al. (2011). Since the two studies are conducted by the same group of the authors, the review CD001321 only cites Salo et al. (2010). But when the PubMed IDs were identified, the metadata of every original article in the reference sections was collected regardless of whether it is included or excluded, and whether it is directly cited.

such as the abstracts of the original articles, the references of the original articles, and the list of other original studies published by the authors with the same names, all of which we used to compute the social and intellectual dependencies of the original studies within review articles.

For this study, we decided to focus on the biomedical claims examined by more than 2 but less than 50 clinical trials, using dichotomous outcome variables. Given that we attempt to first test the robustness of the medical claims with the Leave-One-Out procedure, we thought a meta-analysis with many trials would be trivially robust to the method. The final dataset for the following analysis includes 17,891 medical claims from 3,390 reviews. In subsequent analysis, we study the relationship between the robustness of 6,186 statistically significant claims (from 2,134 reviews) and the social/prior knowledge dependency.

**Dependent variable: The r-value as a measure replicability**

In this paper, the *r*-value (Shenhav et al., 2015) is used as a measure of replicability, which can be computed based on a meta-analysis. The computation of the *r*-value basically follows the logic of sensitivity analysis in which whether the significance of a claim holds when samples of studies in a systemic review are excluded. By measuring the impact of each study or sets of studies on the overall conclusion of a review, the *r*-value can provide more rigorous evidence of the replicability of scientific claims by identifying one single study drives the overall significance of a meta-analysis with the increased power. Figure 2 visually demonstrates how the logic can be applied using a forest plot from one review in the CDSR: when Study 7 is excluded, the overall conclusion becomes insignificant.

[Figure 2 here]

Like *p*-value, the significance level of the *r*-value depends on a researcher; but the smaller the *r*-value, it is less likely that the null hypothesis (i.e., no actual effect of treatment based on a meta-analysis) is true. This implies that the smaller the *r*-value, it is more likely that a conclusion from a meta-analysis is not driven by a specific single study. Conceptually speaking, the *r*-value is the largest *p*-value among the *p*-values from the *N* meta-analysis when the basic Leave-Out-Out procedure is applied. Other frameworks such as False Discovery Rate, which controls a multiplicity of endpoints. The computational details as following.

Let $p^L_{i_1,\ldots,i_k}$ and $p^R_{i_1,\ldots,i_k}$ be the left and right *p*-values from a meta-analysis on the subset $S = (i_1, \ldots, i_k) \subset 1, \ldots, N$, where *N* = the full number of studies in the meta-analysis and $k < N$. And let $\prod(k)$ be the set of all possible subsets of size *k*. Each subset from (*N - u + 1*) studies is used to compute the *r*-value of a systematic meta-analysis. The value of *u* is subject to a choice, but we can see that $\binom{N}{u-1}$ subsets of $(N - u + 1)$ studies are used to calculate the *r*-value for a given value of *u*. For example, if *u* = 2, a meta-analysis of each of the *N* subsets of *N* - 1 studies will be employed.

For a left-sided alternative, the *r*-value is computed as

$$r^L = max_{\{i_1,\ldots i_{(N-u+1)}\} \in \prod(N-u+1)} p^L_{i_1,\ldots,i_{(N-u+1)}}$$

For a right-sided alternative, the r-value is computed as:

$$r^R = max_{\{i_1,\ldots i_{(N-u+1)}\} \in \prod(N-u+1)} p^R_{i_1,\ldots,i_{(N-u+1)}}$$

For a two-sided alternative,

$$r = 2\min(r^L, r^R).$$

The sensitivity of the confidence interval from a meta-analysis is also can be calculated using the same underlying logic. The sensitivity interval can be defined as the union of the confidence intervals from the meta-analysis using the $\binom{N}{u-1}$ subsets of $(N - u + 1)$ studies. With the α level significance level, the upper limit of the (1 - α) sensitivity interval equals to the upper limit of the (1- α) confidence interval from the meta-analysis on $L = \{i_1^L, \ldots, i_{(N-u+1)}^L\}$, where $L$ is the subset resulting in the largest *p*-value for the left-side *r*-value. The lower limit of the (1 - α) is computed using the same logic. This implies that a meta-analysis is robust (with a given $u$) if and only if the sensitivity interval does not contain the value proposed in the null hypothesis (e.g., an odds-ratio equals to 1).

Furthermore, when multiple primary endpoints need to be considered, Family-Wise Error Rate (FWER) or False Discovery Rate (FDR) controlling procedures can also be employed to lower the significance threshold. For example, review CD005211, which examines the effectiveness of exercise interventions to prevent/minimize the upper-limb dysfunction for those who receive breast cancer treatment, investigates 4 different types of study designs that can be grouped according to the timing of the introduction of the exercise intervention programs. In this case, we have multiple point estimates of which the relatedness should be controlled. Using FDR framework, we can apply the Benjamini-Hochberg (BH) adjustment[9] (Benjamini & Hochberg, 1995) to achieve this.

**Independent Variables: Co-authorship and Reference Overlap**

[Figure 3 here]

[9] The BH-adjusted *r*-values for a sorted list of *r*-values: $r_{(1)} \leq \ldots \leq r_{(M)}$ for $min_{i \geq j} \frac{M * r_{(i)}}{i}$, $j = 1, \ldots, M$. For instance, if we have $S$= (0.1231, 0.0017, 0.0167, 0.1176), the sorted list is (0.0017, 0.0167, 0.1231, 0.1776) and the adjusted r-values with BH procedure is (0.0068, 0.0334, 0.1641, 0.1776).

We used the Jaccard Coefficient (JC) to measure the degree of overlap for authors and references between the original papers in each Cochrane review. The JC is defined as the proportion of the intersection of two sets to the union of them: $JC\ (A_i, A_j) = \frac{|A_i \cap A_j|}{|A_i| + |A_j|\ |A_i \cap A_j|}$. For example, if paper 1 and paper 2 were written by author A & B, and author B, C, D, and E, the authorship JC between the two papers is computed as 0.2. The resulting values for authors and references were employed to construct the edge weights between pairs of articles in the network layers represented in Figure 3.

*Dependence scores*

The dependence score (*D score)* for authorship and references of each Cochrane review is defined as following: $D = \frac{W}{E_{max}}$, where $E_{max} = n(n-1)\ /2$, and $W$ is the sum of all E, corresponding to the JC between a pair of pair in terms of authorship or reference. Social and knowledge independence are computed as on the JC of authors and reference overlaps, respectively. For example, Claim 1 in Figure 3 has low social and knowledge independence scores whereas Claim 4 has high independence scores for the two dimensions. Claim 2 scores high in knowledge independence but low for social independence whereas Claim 3 has the opposite pattern.

## Results

*Robustness of meta-analyses*

In this study, the Leave-One-Out procedure was chosen to check the robustness of the biomedical claims (which means that $u = 2$). The more trials are excluded, the higher *r*-values will be produced due to the reduced power with fewer samples. Thus, this gives us the smallest *r*-values from the space of

possible *r*-values. In other words, the Leave-One-Out is the most conservative approach among possible choices of the value for *u*, which would make *r*-values as closest as possible to the original *p*-values.

[Figure 4 here]

Figure 4 displays how the *p*-values and *r*-values from 17,891 dichotomous claims are distributed and Table 1 provides the summary using the 5% significance level as a criteria. And Among 17,891 dichotomous claims, 6,186 claims (34.58%) are statistically significant ($p < 0.05$) when meta-analyses with multi-level regressions were conducted[10]. The fact that approximately 65% of total claims were insignificant does not necessarily indicate the results from biomedical research using randomized controlled trials are incredible because the claims examined by the Cochrane reviews included not only beneficial effects of certain types of treatment but also side-effects of drugs or harmful effect of environs. And given that r-values are always greater than or equal to original p-values, we can see in the table that the cell for the cases of originally insignificant with p-values but significant with r-values is empty.

[Table 1 here]

Here, focusing on 6,186 significant claims provides us an interesting result. Table 1 shows that 1,822 claims (29.45%) out of 6,186 claims were identified to be sensitive to the Leave-One-Out procedure. Even though we do not have the ground truths that can be used as a baseline to evaluate whether the 30% is high or low, we can at least see that the proportion is much higher than the 5% false-positive level. This basically suggests that the overall conclusion from the systematic reviews can be highly dependent on the choices of the authors of systematic reviews in terms of how they choose which studies to be included or excluded. Of course, the authors of the Cochrane systematic review discuss

[10] `Meta` package in R was used. For a reference, see Schwarzer et al. (2015).

how they initially identified relevant clinical trials how they evaluate the validity of the trials and release the list of the clinical trials excluded. However, the result still implies a possibility that a small degree of arbitrariness can change the conclusion from the meta-analyses in the Cochrane Reviews.

*Social/knowledge dependence and robustness of claims*

Now, the association between social and knowledge dependence between studies and robustness of claims was tested, using a logistic regression model. Note that the model used 6,186 initially significant claims as observations. Moreover, I decided to introduce the number of clinical trials as a control variable, considering that a claim is more likely to be robust to the Leave-One-Out procedure when it contains more trials, even though I restricted the data to the claims studied by more than or equal to 3 but less than 50 trials. For example, we can expect that a significant conclusion from a meta-analysis with 49 studies is more robust to the Leave-One-Out procedure than a claim drawn from a meta-analysis using only 3 studies. The model can be written down as below.

$$logit\ P(R_i = 1) = \beta_0 + \beta_1 * S_{Dep} + \beta_2 * K_{Dep} + \beta_3 * N \quad \text{--------- (1)}$$

$$where,$$

$$\begin{cases} R_i = 1, if\ p_i < 0.05\ and\ r_i < 0.05 \\ R_i = 0, if\ p_i < 0.05\ but\ r_i \geq 0.05, \end{cases}$$

$$S_{Dep} = social\ dependence,$$
$$K_{Dep} = knowledge\ depdence,$$
$$N = the\ number\ of\ trials.$$

[Table 2 here]

Since the independent variables range from 0 to 1, I decided to multiply 10 times to the independent variables before I estimated the model coefficients more interpretable (one unit increase in

the dependence measurement means the change from 0 to 1, which would be unrealistic.) Table 2 displays the result. As expected, the number of trials is highly correlated with the robustness of the significant claims to the Leave-One-Out test. And we can see that co-authorship dependency is also significantly associated with the robustness of the claims. In the odds ratio term, the result suggests that the increase of 0.1 unit of the dependence score would lead to 21.77% increase in the odds of being robust to the procedure. What this implies is that there would be some under-dispersions in the estimates of the effect size of the treatments when there are overlapping co-authorships between the researchers even though randomized controlled trials were employed. Further implications will be discussed in the conclusion.

**Conclusion and Future Work**

In this paper, we first attempted to quantify the robustness of the biomedical claims examined the Cochrane Database of Systematic Review by applying the framework of *r*-value developed by Shenhav (2015). Specifically, the Leave-One-Out procedure was employed to generate the *r*-values of the 17,891 claims from 3,390 reviews. It turned out that 29.45% of the 6,186 initially significant claims according to the meta-analyses were sensitive. It is higher than the 5% of the false-positives which we often expect as an error rate. Moreover, we identified a possible positive association between the co-authorship overlaps and the degree of agreement in the effect size of the treatments.

The result basically demonstrates that the difficulty of establishing replicability in biomedical studies, using meta-analyses based on randomized controlled trials. First, we showed that the simple Leave-One-Out procedure can undermine the validity of the initial conclusion from the systematic meta-analyses. This does not point to a case in which the authors of meta-analyses always intentionally

include studies favoring their underlying positions and otherwise exclude valid studies. We believe in general that the majority of authors volunteer to take the responsibility of publishing meta-analyses with good faith. However, a recent tumult in the Cochrane Foundation about the evaluation of the HPV vaccine (Jørgensen et al., 2018) raises a concern that the process of selecting legitimate studies for a meta-analysis might be affected by non-scientific consideration. Since the debate regarding the validity of the systematic review on the HPV vaccine is still going on, we will not discuss it further here; but we think the HPV vaccine case and our analysis together indicate that we need to be indeed cautious in evaluating the results from statistical evidence that supports a biomedical claim.

The positive association between the robustness of the meta-analyses and the co-authorship dependence of the researchers suggests a possibility of the under-dispersion of the effect size of treatments which we suspect stemming from the underlying assumptions about the effect of treatment by researchers who have social ties. Of course, we do not think that this is the smoking gun given that the current analysis has rooms for improvement. For example, additional factors such as journal prominence and author's characteristics of the systematic reviews could be included; the latent space approach (Hoff, 2002) might be helpful in terms of reducing unmeasured latent dependency between the authors of original trials. Furthermore, a better method to disambiguate the authors of the original clinical trials might change the result. But we would like to note that the current measure for reference overlap was highly correlated with the authorship overlap and the co-authorship overlap was still significant despite the multicollinearity, which inflates the variance of coefficients.

Our analysis hints us a clue about how the current science funding practice should be reconsidered. If the increase of collaboration in scientific endeavor could lead to dense social and knowledge space like an echo-chamber, which would undermine a condition in which the "wisdom of crowds" can operate appropriately, we might need to more diversify the funding. Theoretically, if this is

the case, it will require us to think about the importance of simultaneous multiple discoveries (Merton, 1961) in terms of making science reliable.

**References**


A. Almaatouq, A. Noriega & A. Pentland. “Wisdom of Dynamic Networks: Selective Social Learning and the Adaptive Wisdom of Crowds.” Collective Intelligence Conference, NYU, New York, USA. June 2017

Becker, Joshua, Devon Brackbill, and Damon Centola. "Network dynamics of social influence in the wisdom of crowds." *Proceedings of the national academy of sciences 114,* no. 26 (2017): E5070-E5076.

Bergstrom, Carl T., Jevin D. West, and Marc A. Wiseman. "The eigenfactor™ metrics." *Journal of Neuroscience* 28, no. 45 (2008): 11433-11434.

Begley, C. Glenn, and Lee M. Ellis. "Raise standards for preclinical cancer research." *Nature* 483, no. 7391 (2012): 531-533.

Benjamini, Yoav, and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal statistical society: series B (Methodological)* 57, no. 1 (1995): 289-300.

Bennett, L. Michelle, and Howard Gadlin. "Collaboration and team science: from theory to practice." Journal of Investigative Medicine 60, no. 5 (2012): 768-775.

Binz-Scharf, Maria C., Yuval Kalish, and Leslie Paik. "Making science: New generations of collaborative knowledge production." *American Behavioral Scientist* 59, no. 5 (2015): 531-547.

Bonacich, Phillip. "Power and centrality: A family of measures." *American journal of sociology* 92, no. 5 (1987): 1170-1182.

Bonacich, Phillip. "Some unique properties of eigenvector centrality." *Social networks* 29, no. 4 (2007): 555-564.

Büchter, Roland Brian, and Dawid Pieper. "Most overviews of Cochrane reviews neglected potential biases from dual authorship." *Journal of clinical epidemiology* 77 (2016): 91-94.

Clark, Benjamin Y., and Jared J. Llorens. "Investments in scientific research: Examining the funding threshold effects on scientific collaboration and variation by academic discipline." Policy Studies Journal 40, no. 4 (2012): 698-729.

Collins, Francis S., and Lawrence A. Tabak. "Policy: NIH plans to enhance reproducibility." *Nature News* 505, no. 7485 (2014): 612.

Crane, Diana. "Invisible colleges; diffusion of knowledge in scientific communities." (1972).

Croyle, Robert T. "The National Cancer Institute's transdisciplinary centers initiatives and the need for building a science of team science." *American journal of preventive medicine* 35, no. 2 (2008): S90-S93.

Danchev, Valentin, Andrey Rzhetsky, and James A. Evans. "Centralized" big science" communities more likely generate non-replicable results." *arXiv preprint arXiv:1801.05042*(2018).

Fortunato, Santo, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Staša Milojević, Alexander M. Petersen et al. "Science of science." *Science* 359, no. 6379 (2018): eaao0185.

Galton, Francis. "Vox populi (the wisdom of crowds)." *Nature*75, no. 7 (1907): 450-451.

Handoll, Helen H., William J. Gillespie, Lesley D. Gillespie, and Rajan Madhok. "The Cochrane Collaboration: a leading role in producing reliable evidence to inform healthcare decisions in musculoskeletal trauma and disorders." *Indian journal of orthopaedics* 42, no. 3 (2008): 247.

Hagstrom, Warren O. "Traditional and modern forms of scientific teamwork." *Administrative Science Quarterly* (1964): 241-263.

Hicks, Diana, and J. Sylvan Katz. "Science policy for a highly collaborative science system." Science and public policy 23, no. 1 (1996): 39-44.

Higgins, Julian PT, Douglas G. Altman, Peter C. Gøtzsche, Peter Jüni, David Moher, Andrew D. Oxman, Jelena Savović, Kenneth F. Schulz, Laura Weeks, and Jonathan AC Sterne. "The Cochrane Collaboration's tool for assessing risk of bias in randomised trials." *Bmj* 343 (2011): d5928.

Hoff, Peter D., Adrian E. Raftery, and Mark S. Handcock. "Latent space approaches to social network analysis." *Journal of the american Statistical association* 97, no. 460 (2002): 1090-1098.

Hox, Joop J., Mirjam Moerbeek, and Rens Van de Schoot. *Multilevel analysis: Techniques and applications*. Routledge, 2017

Ioannidis, John PA. "Contradicted and initially stronger effects in highly cited clinical research." *JAMA* 294, no. 2 (2005a): 218-228.

Ioannidis, John PA. "Why most published research findings are false." *PLoS medicine* 2, no. 8 (2005b): e124.

Ioannidis, John PA. "Why most clinical research is not useful." *PLoS medicine* 13, no. 6 (2016): e1002049.

Jørgensen, Lars, Peter C. Gøtzsche, and Tom Jefferson. "The Cochrane HPV vaccine review was incomplete and ignored important evidence of bias." *BMJ evidence-based medicine*23, no. 5 (2018): 165-168.

Kivelä, Mikko. "Multilayer networks library for python (pymnet)." (2017).

Leahey, Erin. "From sole investigator to team scientist: Trends in the practice and study of research collaboration." Annual review of sociology 42 (2016): 81-100.

Lee, Carole J., Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. "Bias in peer review." *Journal of the American Society for Information Science and Technology* 64, no. 1 (2013): 2-17.

Latour, Bruno. *Science in action: How to follow scientists and engineers through society*. Harvard university press, 1987.

Li, Danielle. "Expertise versus Bias in Evaluation: Evidence from the NIH." *American Economic Journal: Applied Economics* 9, no. 2 (2017): 60-92.

Lindson, Nicola, Samantha C. Chepkin, Weiyu Ye, Thomas R. Fanshawe, Chris Bullen, and Jamie Hartmann-Boyce. "Different doses, durations and modes of delivery of nicotine replacement therapy for smoking cessation." *Cochrane Database of Systematic Reviews* 4 (2019).

Lorenz, Jan, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. "How social influence can undermine the wisdom of crowd effect." *Proceedings of the national academy of sciences 108*, no. 22 (2011): 9020-9025.

MacCoun, Robert, and Saul Perlmutter. "Blind analysis: hide results to seek the truth." *Nature News* 526, no. 7572 (2015): 187.

Mahoney, Michael J. "Publication prejudices: An experimental study of confirmatory bias in the peer review system." *Cognitive therapy and research* 1, no. 2 (1977): 161-175.

Merton, Robert K. "Singletons and multiples in scientific discovery: A chapter in the sociology of science." *Proceedings of the American Philosophical Society* 105, no. 5 (1961): 470-486.

Merton, Robert K. "The Normative Structure of Science", in Merton, Robert K. (ed.), *The Sociology of Science: Theoretical and Empirical Investigations*, Chicago: University of Chicago Press (1973) [1942]

Nissen, Silas Boye, Tali Magidson, Kevin Gross, and Carl T. Bergstrom. "Publication bias and the canonization of false facts." *Elife* 5 (2016): e21451.

O'Brien, Timothy L. "Change in academic coauthorship, 1953–2003." Science, Technology, & Human Values 37, no. 3 (2012): 210-234.

Olsen, Ole, Philippa Middleton, Jeanette Ezzo, Peter C. Gøtzsche, Victoria Hadhazy, Andrew Herxheimer, Jos Kleijnen, and Heather McIntosh. "Quality of Cochrane reviews: assessment of sample from 1998." *Bmj* 323, no. 7317 (2001): 829-832.

Open Science Collaboration. "Estimating the reproducibility of psychological science." *Science* 349, no. 6251 (2015).

Page, Scott E. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies-New Edition*. Princeton University Press, 2008.

Pashler, Harold, and Eric–Jan Wagenmakers. "Editors' introduction to the special section on replicability in psychological science: A crisis of confidence?." *Perspectives on Psychological Science* 7, no. 6 (2012): 528-530.

Price, Derek J. de Solla (Derek John de Solla). *Little Science, Big Science-- and Beyond*. New York: Columbia University Press, 1986 [1963].

Prinz, Florian, Thomas Schlange, and Khusru Asadullah. "Believe it or not: how much can we rely on published data on potential drug targets?." *Nature reviews Drug discovery* 10, no. 9 (2011): 712.

Rosenthal, Robert. "The file drawer problem and tolerance for null results." *Psychological bulletin* 86, no. 3 (1979): 638.

Schwarzer, Guido, James R. Carpenter, and Gerta Rücker. *Meta-analysis with R*. Vol. 4724. New York, NY: Springer, 2015.

Stokols, Daniel, Kara L. Hall, Brandie K. Taylor, and Richard P. Moser. "The science of team science: overview of the field and introduction to the supplement." American journal of preventive medicine 35, no. 2 (2008): S77-S89.

Shenhav, Liat, Ruth Heller, and Yoav Benjamini. "Quantifying replicability in systematic reviews: the r-value." *arXiv preprint arXiv:1502.00088* (2015).

Surowiecki, James. *The wisdom of crowds*. Anchor, 2005.

Wolfe, Audra J. *Competing with the Soviets: Science, technology, and the state in Cold War America*. JHU Press, 2012.

Wu, Lingfei, Dashun Wang, and James A. Evans. "Large teams develop and small teams disrupt science and technology." *Nature* 566, no. 7744 (2019): 378.

Wu, Lingfei, Linzhuo Li, and James Evans. "Social connection induces cultural contraction: Evidence from hyperbolic embeddings of social and semantic networks." *arXiv preprint arXiv:1807.10216* (2018).

Wuchty, Stefan, Benjamin F. Jones, and Brian Uzzi. "The increasing dominance of teams in production of knowledge." *Science 316, no. 5827* (2007): 1036-1039.

**Figure 1. An Example of a Forest Plot from the CDSR (CD008792)**

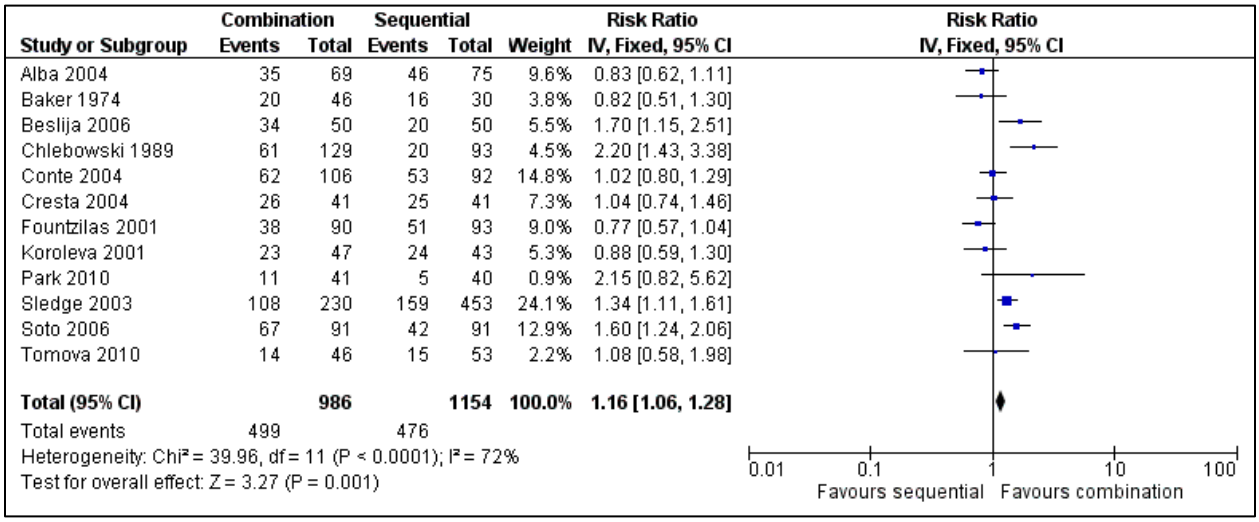


**Figure 2. The Forest Plot from Review CD008792 (Left) and Excluding Study 7 (Right)**

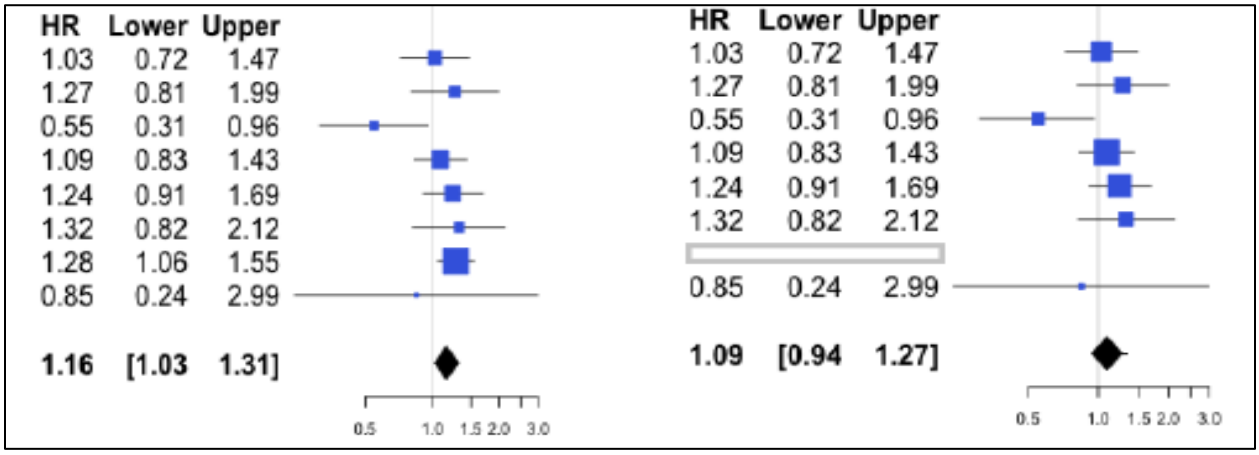


**Figure 3. The Schematic Multi-Layer Presentation**

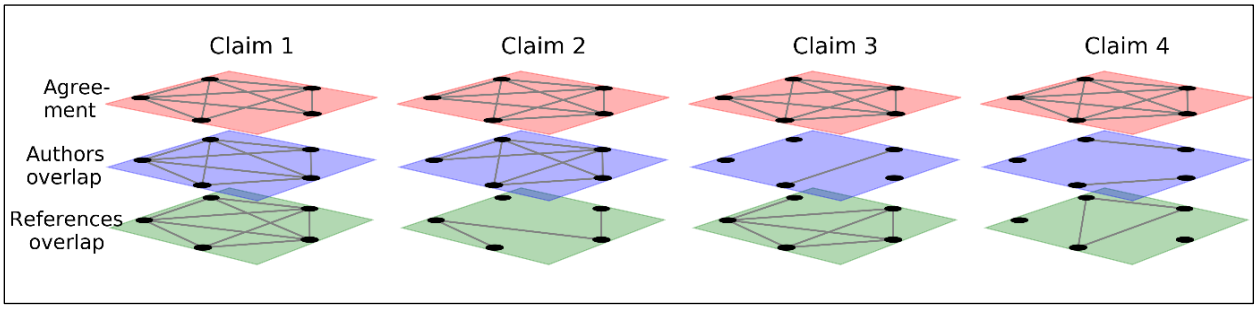

**Figure 4. *p*-values versus *r*-values**

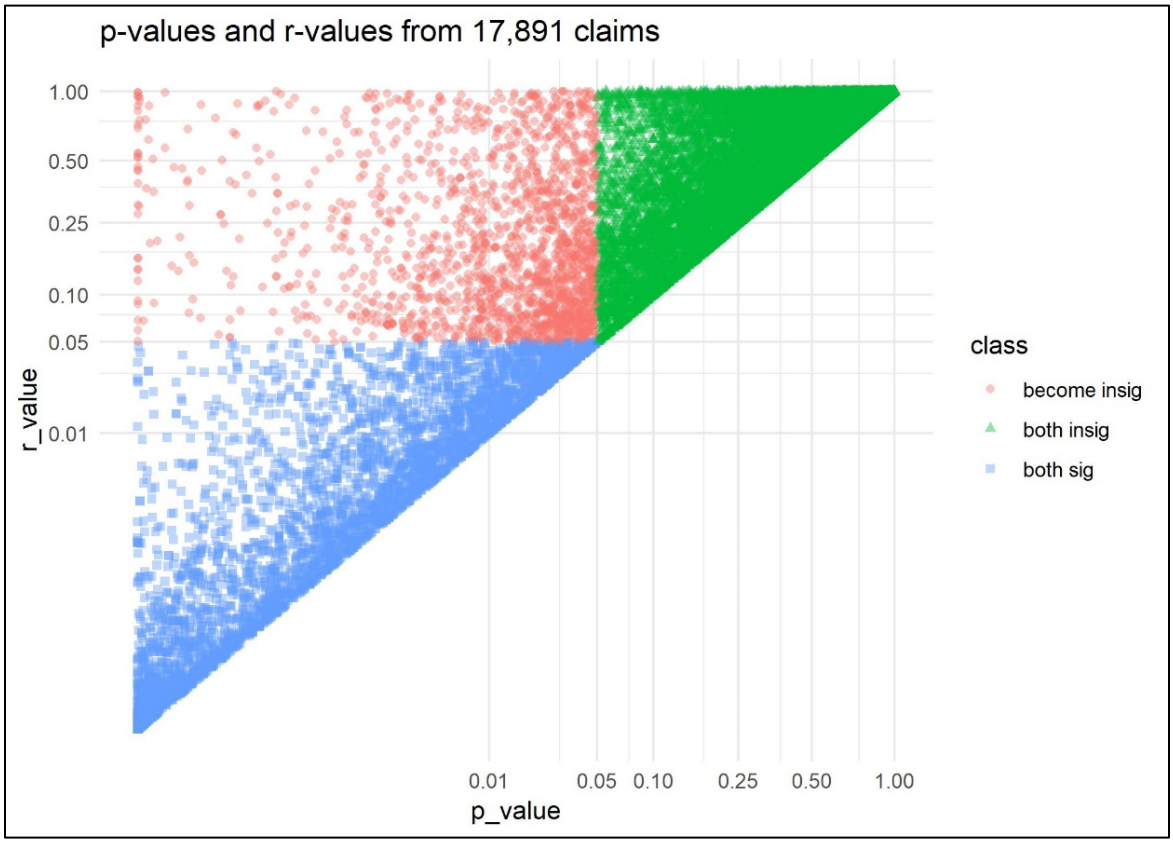

**Table 1**

| | Significant ($p < 0.05$) | Insignificant ($p \geq 0.05$) | Total |
|---|---|---|---|
| Significant ($r < 0.05$) | 4,364 | - | 4,364 |
| Insignificant ($r \geq 0.05$) | 1,822 | 11,705 | 13,527 |
| Total | 6,186 | 11,705 | 17,891 |

**Table 2**

| | ***Logistic Regression*** |
|---|---|
| | **R = 1 if $p$ < 0.05 and $r$ < 0.05** |
| Num_of_Trials | 0.074*** |
| | (0.004) |
| **Co_authorship_Dep*10** | **0.197*** |
| | (0.089) |
| Ref_Dep*10 | 0.206 |
| | (0.359) |
| Intercept | -0.005 |
| | (0.051) |
| Observations | 6,069 |
| Log Likelihood | -3,438.6985 |
| *Note:* | *p<0.05; **p<0.01; ***p<0.001 |