PLOS | ONE

RESEARCH ARTICLE

The Dilemma of Heterogeneity Tests in Meta-Analysis: A Challenge from a Simulation Study

Shi-jun Li[1☯], Hua Jiang[1,2,3☯*], Hao Yang[1☯], Wei Chen[1,2☯], Jin Peng[1], Ming-wei Sun[1], Charles Damien Lu[1], Xi Peng[1,3], Jun Zeng[1☯]

1 Department of Computational Mathematics and Bio-statistics, Metabolomics and Multidisciplinary Laboratory for Trauma Research, Sichuan Provincial People's Hospital, Sichuan Academy of Medical Sciences, Chengdu, Sichuan, China, **2** Department of Parenteral and Enteral Nutrition, Peking Union Medical College Hospital, Beijing, China, **3** Institute of Burn Research, Southwest Hospital of the Third Military Medical University, Chongqing, China

☯ These authors contributed equally to this work.
* cdjianghua@gmail.com

Abstract

Introduction

After several decades' development, meta-analysis has become the pillar of evidence-based medicine. However, heterogeneity is still the threat to the validity and quality of such studies. Currently, Q and its descendant I^2 (I square) tests are widely used as the tools for heterogeneity evaluation. The core mission of this kind of test is to identify data sets from similar populations and exclude those are from different populations. Although Q and I^2 are used as the default tool for heterogeneity testing, the work we present here demonstrates that the robustness of these two tools is questionable.

Methods and Findings

We simulated a strictly normalized population S. The simulation successfully represents randomized control trial data sets, which fits perfectly with the theoretical distribution (experimental group: p = 0.37, control group: p = 0.88). And we randomly generate research samples Si that fits the population with tiny distributions. In short, these data sets are perfect and can be seen as completely homogeneous data from the exactly same population. If Q and I^2 are truly robust tools, the Q and I^2 testing results on our simulated data sets should not be positive. We then synthesized these trials by using fixed model. Pooled results indicated that the mean difference (MD) corresponds highly with the true values, and the 95% confidence interval (CI) is narrow. But, when the number of trials and sample size of trials enrolled in the meta-analysis are substantially increased; the Q and I^2 values also increase steadily. This result indicates that I^2 and Q are only suitable for testing heterogeneity amongst small sample size trials, and are not adoptable when the sample sizes and the number of trials increase substantially.

OPEN ACCESS

Citation: Li S-j, Jiang H, Yang H, Chen W, Peng J, Sun M-w, et al. (2015) The Dilemma of Heterogeneity Tests in Meta-Analysis: A Challenge from a Simulation Study. PLoS ONE 10(5): e0127538. doi:10.1371/journal.pone.0127538

Academic Editor: Jian-Xin Gao, Shanghai Jiao Tong University School of Medicine, CHINA

Received: January 19, 2015

Accepted: April 16, 2015

Published: May 29, 2015

Copyright: © 2015 Li et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The work we present in this manuscript is based on a simulation study, all data are generated in R and we provide all details of the simulation process in our supplementary information section. Anyone may repeat our study when the codes are executed in R.

Funding: This work was partially supported by Sichuan Provincial Department of Science and Technology (to Hua Jiang, No. 2014FZ0125 and Jun Zeng, No. 15ZC0656), and Chengdu Municipality Department of Science and Technology (to Hua Jiang, No. 11PPYB099SF) and Sichuan Academy of Medical Sciences (to Hao Yang). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Conclusions

Every day, meta-analysis studies which contain flawed data analysis are emerging and passed on to clinical practitioners as “updated evidence”. Using this kind of evidence that contain heterogeneous data sets leads to wrong conclusion, makes chaos in clinical practice and weakens the foundation of evidence-based medicine. We suggest more strict applications of meta-analysis: it should only be applied to those synthesized trials with small sample sizes. We call upon that the tools of evidence-based medicine should keep up-to-dated with the cutting-edge technologies in data science. Clinical research data should be made available publicly when there is any relevant article published so the research community could conduct in-depth data mining, which is a better alternative for meta-analysis in many instances.

Introduction

Currently, Q and its descendent I^2 tests are widely used, especially the I^2 test, in meta-analysis [1–3]. Established in 2003 by Higgins et al, it is becoming the mainstay for testing heterogeneity [1]. Q and I^2 tests have been integrated into *Review Manager* and almost all other meta-analysis software, and are used as the default tool to determine heterogeneity. In the past decade, along with the emergence of meta-analysis as a core technique for evidence-based approach in almost all branches of bio-medical research, Q and I^2 make up an important methodological component of the enormous number of systematic reviews and clinical guidelines.

Unfortunately, despite the wide use and acceptance of Q and I^2 tests, the work we present here demonstrates that the robustness of these two tools are questionable; and in many circumstances, relying solely on these tools to measure heterogeneity could lead to the wrong conclusion in meta-analysis, which forms the foundation of evidence-based medicine.

Materials and Methods

Theoretical Analysis and Simulation

Analyzing on the Structure of Q and I2. The structure of the equation of Q is the following:

$$\mathrm{Q} = \sum\nolimits_k \hat{\omega}_k (\mu_k - \hat{\bar{\mu}}_{\hat{\omega}})^2 \tag{1}$$

$$\hat{\omega}_k = n_k / \sigma_k^2 \tag{2}$$

Here, $\bar{\mu}_{\hat{\omega}} = \sum \omega_k \mu_k / \sum_k \omega_k$ and represents the weight of the k-th study, n_k is the sample size of the k–th study. It is assumed that the sample from any trial is independent and the distribution is normalized [3].

Q does not consider the influence from the number of enrolled trials (degree of freedom, *df)*. We can understand this shortcoming of Q from its equation: Q is the weighted sum of the squares of deviations (WSSD) of data sets from the enrolled trials. Along with the increase of the number of trials (n), the non-negative term also increases. Therefore, the number of enrolled trials significantly influences the increase of Q value. Thus the increase of Q value cannot simply be attributed to the variants between enrolled trials. To overcome this shortcoming, Higgins *et al* constructed I^2. It modifies Q and aims to balance the extra variant, which comes from the increase of the number of enrolled trials. Strictly speaking, I^2 is not a test but a descriptive measure.

The equation of I^2 is the following

$$I^2 = \frac{Q - df}{Q} \times 100\% \tag{3}$$

Here *df* is the degree of freedom, *df* = n-1

Although I^2 proposes to overcome quasi-heterogeneity from extra variants, a more serious influence is not considered, which is the sample size n_k (Eq.2). We can easily find that $\hat{\omega}_k$ is in proportion to n_k. Along with the increase of the sample size, the corresponding deviation will also increase. Consequently, the Q value will increase.

Let

$$\mathrm{T} = \frac{\mu_k - \hat{\hat{\mu}}_{\hat{\omega}}}{\frac{\sigma_k}{\sqrt{n_k}}} \tag{4}$$

Remember that the default assumption behind the statistics of the t-test is that the distribution of all enrolled trials met $\mathrm{N}\,(\mu_k, \sigma_k^2)$ and we therefore have $\mathrm{T} \sim \mathrm{T}\,(n_k - 1) \rightarrow N\,(0, 1)$. So Q is indeed the sum of the squares of T_k. Consequently, constructing Q is a process that is made up by the sum of the square of T_k. It is easy to infer that the sample size of each trial cannot be too big, otherwise the T value will surge.

To explore the evolutionary patterns between Q, I^2 and n_k, we herein introduce a simulation process to verify the influence of N and n to Q and I^2.

Simulation Process. We illustrated the research flow and simulation process of the study in Fig 1.

We simulated a population S and its distribution is strictly normalized, which means $S \sim N(\mu, \sigma^2)$ (Table A in S1 File). Now we have samples S_i (i = 1, 2, 3. . .n) where each is a random sample from S (Table B in S1 File). Let $S_i \sim (\mu_i, \sigma_i^2)$. The variation between the samples is only made by random error ε, and $\varepsilon \sim N(0,\sigma_\varepsilon)$.

The distribution parameters of Si can be descripted as following :

$$\mu_i = \mu + E(\boldsymbol{\varepsilon}) \tag{5}$$

$$\sigma_i = \sqrt{\sigma^2 + \sigma_\varepsilon^2} \tag{6}$$

Let $\sigma_\varepsilon << \sigma$, then we have:

$$\mathrm{E}(S_i) = \mu_i = \mu + 0 = \mu = E(S) \tag{7}$$

$$\mathrm{D}(S_i) = \sigma_i^2 = \sqrt{\sigma^2 + \sigma_\varepsilon^2} \approx D(\mathrm{S}) \tag{8}$$

We know Si is a non-skewed sampling of the population. Therefore the simulated data sets are homogenous. We then synthesized these data sets by meta-analysis (fixed model, meta: meta-analysis with R was employed for data aggregating) and we calculated Q and I^2 for each synthesis experiment (Tables C and D in S1 File). To each Si, sampling process will be repeated in 1000 times. Thus we get the distribution of I^2 variations in synthesizing different number of trials (the sample size of each trial is the same). Finally we generated heat map to see the impact of I^2, Q and the number of trials (n) and sample size N (Tables E, F and G in S1 File)

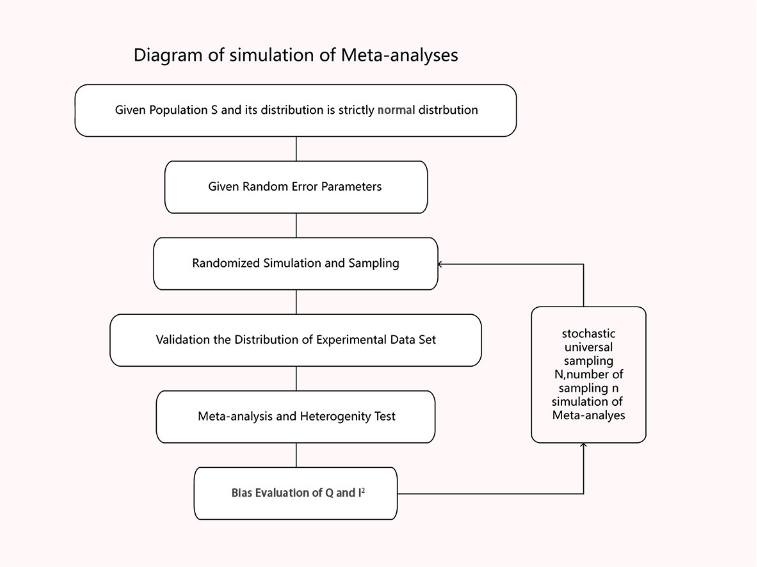


Diagram of simulation of Meta-analyses
Given Population S and its distribution is strictly normal distrbution
Given Random Error Parameters
Randomized Simulation and Sampling
Validation the Distribution of Experimental Data Set
Meta-analysis and Heterogenity Test
Bias Evaluation of Q and I²
stochastic universal sampling N,number of sampling n simulation of Meta-analyes

Fig 1. Diagram of the simulation process.

doi:10.1371/journal.pone.0127538.g001

Distribution Test. We used Kolmogorov-Smirnov Tests to test the distribution of the samples,$\alpha = 0.05$.

Simulation Algorithm. We employed Mersenne-Twister (Matsumoto and Nishimura, 1998) from RNG to simulate data sets [4, 5]. Simulation programming in R see Tables A-G in S1 File.

Environment and Setting of Computation. All computing processes were done using a high performance-computing platform at the Sichuan Academy of Medical Sciences, by using R (version 3.1.1 for win7 64bit) [4].

Results and Discussion

The simulation successfully represents randomized control trial data sets that meet normal distribution and generates *S* (Table 1 and Fig 2), which fits perfectly with the theoretical distribution (experimental group: p = 0.37, control group: p = 0.88). And we randomly generate research samples S_i that fits the population with tiny distributions. In short, these data sets are perfect and can be seen as completely homogenous data from the exactly same population. If Q and I^2 are truly robust tools, the Q and I^2 test results on our simulated data sets here should not be positive. We then synthesized these trials by using fixed model. We exhibit here three meta-analyses that are selected from our simulation experiments (Figs 3–5). Pooled results indicated

Table 1. Distribution of simulated data sets.

Parameters of Distribution	True value of S (population)	Estimation of simulated S	Error	P
Experimental Group				0.37
μe	100	99.99	0.01	
σe	1	0.963	0.037	
Control group				0.88
μc	10	10.01	0.01	
σc	1	1.059	0.059	

doi:10.1371/journal.pone.0127538.t001

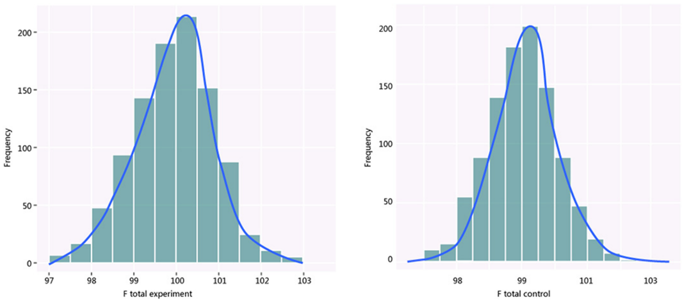


Fig 2. Distribution of simulated S, which is typical normal distribution. (A: Experimental group; B: control group).

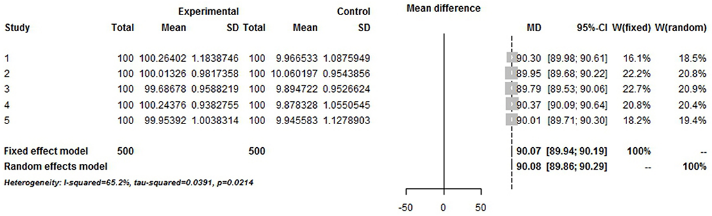
doi:10.1371/journal.pone.0127538.g002

that the mean difference (MD) corresponds highly with the true values, and the 95% is narrow. But, along with the increase of the numbers of trials and sample size, the value of the I^2 steadily increased (Figs 6 and 7A). Relatively, the influence of number of trials is relatively smaller. In terms of Q, we found that the value of Q increases along with the increase of the number of trials synthesized into the meta-analysis, and with the increase of the sample sizes of enrolled trials (Fig 7B).

Forest plots of Simulated Meta-analysis

We demonstrate here that the validity of Q and I^2 test is questionable and unstable to evaluate heterogeneity for meta-analysis. The purpose of the heterogeneity test is to determine whether the included trials are sampled from similar populations. If the samples of included trials are from similar populations, then the expected mean of the samples should equal the mean of the populations (true data). If it is not, then the mean of the samples does not equal the mean of the populations (false data). The core philosophy of meta-analysis is to include those trials from populations that are *de facto* the same. The mission of any heterogeneity test is to detect the trials that are *de facto* not the same. A good heterogeneity-testing tool therefore should not make the mistake to classify a homogenous trial as heterogeneous.

Because all the data sets of the simulated enrolled trials in our study are from the sample population, there could be no heterogeneity between them. When the sample size is small, the bias from sampling will increase with the frequency of sampling. When sampling increases in frequency, the theoretical true bias will decrease, thus heterogeneity should decrease. The mean and variance tend to stabilize when the sampling frequency continues to increase. In this scenario, the I^2 and Q value will increase proportionally along with the sample size n_k, thus causing the quasi-heterogeneity. In summary, both Q and I^2 are sensitive and dependent on sample size n_k (Fig 6 and Fig 7).

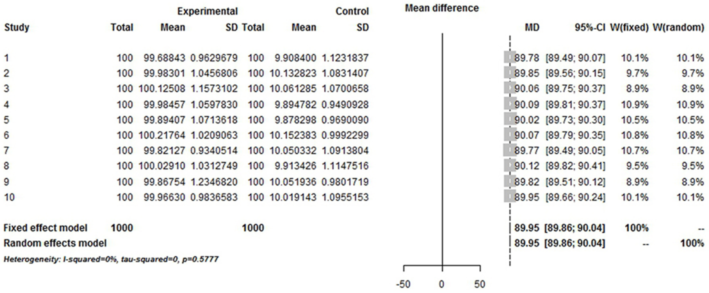


	Experimental			Control						
Study	Total	Mean	SD	Total	Mean	SD	MD	95%-CI	W(fixed)	W(random)
1	100	100.26402	1.1838746	100	9.966533	1.0875949	90.30	[89.98; 90.61]	16.1%	18.5%
2	100	100.01326	0.9817358	100	10.060197	0.9543856	89.95	[89.68; 90.22]	22.2%	20.8%
3	100	99.68678	0.9588219	100	9.894722	0.9526624	89.79	[89.53; 90.06]	22.7%	20.9%
4	100	100.24376	0.9382755	100	9.878328	1.0550545	90.37	[90.09; 90.64]	20.8%	20.4%
5	100	99.95392	1.0038314	100	9.945583	1.1278903	90.01	[89.71; 90.30]	18.2%	19.4%
Fixed effect model	500			500			90.07	[89.94; 90.19]	100%	--
Random effects model							90.08	[89.86; 90.29]	--	100%

Heterogeneity: I-squared=65.2%, tau-squared=0.0391, p=0.0214

Fig 3. Simulated Meta-analysis. Enrolled 5 trials, total number 1000, pooled MD 90, I^2 = 65.2%.

doi:10.1371/journal.pone.0127538.g003



	Experimental			Control						
Study	Total	Mean	SD	Total	Mean	SD	MD	95%-CI	W(fixed)	W(random)
1	100	99.68843	0.9629679	100	9.908400	1.1231837	89.78	[89.49; 90.07]	10.1%	10.1%
2	100	99.98301	1.0456806	100	10.132823	1.0831407	89.85	[89.56; 90.15]	9.7%	9.7%
3	100	100.12508	1.1573102	100	10.061285	1.0700658	90.06	[89.75; 90.37]	8.9%	8.9%
4	100	99.98457	1.0597830	100	9.894782	0.9490928	90.09	[89.81; 90.37]	10.9%	10.9%
5	100	99.89407	1.0713618	100	9.878298	0.9690090	90.02	[89.73; 90.30]	10.5%	10.5%
6	100	100.21764	1.0209063	100	10.152383	0.9992299	90.07	[89.79; 90.35]	10.8%	10.8%
7	100	99.82127	0.9340514	100	10.050332	1.0913804	89.77	[89.49; 90.05]	10.7%	10.7%
8	100	100.02910	1.0312749	100	9.913426	1.1147516	90.12	[89.82; 90.41]	9.5%	9.5%
9	100	99.86754	1.2346820	100	10.051936	0.9801719	89.82	[89.51; 90.12]	8.9%	8.9%
10	100	99.96630	0.9836583	100	10.019143	1.0955153	89.95	[89.66; 90.24]	10.1%	10.1%
Fixed effect model	1000			1000			89.95	[89.86; 90.04]	100%	--
Random effects model							89.95	[89.86; 90.04]	--	100%

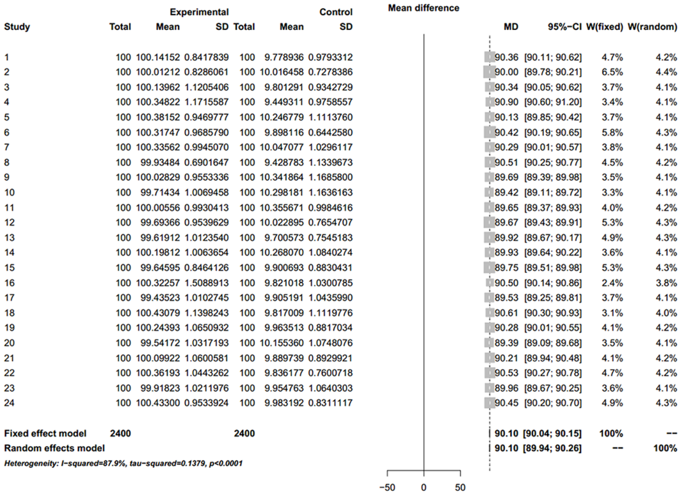
Heterogeneity: I-squared=0%, tau-squared=0, p=0.5777

Fig 4. Simulated Meta-analysis. Enrolled 10 trials, total number 2000, pooled MD 89.95, $I^2 = 0\%$.

doi:10.1371/journal.pone.0127538.g004

Gerta Rücker et al have published an article in 2008 also tried to address the I^2 problem [6]. The result of Rücker's study seemed similar to ours: we both reached the conclusion that the I^2 will increase to 100% along with the sample size increasing in a meta-analysis. But, there was a major methodological flaw in Rücker's study, which it was the fact that they did not test the homogeneity and distribution of the data sets included in their simulation. As is well known, most people performing meta-analysis do not conduct distribution tests on their data set from the original trials, and heterogeneity is quite real in most circumstances. Because the data sets of Rücker's study are from real meta-analysis which quite possibly contains high heterogeneous trials, it is impossible to get rid of the heterogeneity risk by directly and randomly sampling from these data sets. In other words, when the sample size is large enough and the heterogeneity is de facto existent, the increase of I^2 is most likely expected. But, such a simulation cannot be seen as a strict mathematic proof. What we did in our study was to give the complete proof in full generality. In short, we simulated a pure homogenous population S and strictly normalized its distribution, and then we repeated the sampling in 1000 times and proved the I^2 was unstable in any case when sample size increased. To our best knowledge, the study we presented here is the very first one that generally proved that using I^2 test can lead erroneous results in any case when sample size of a meta-analysis is large.

After several decades' development, meta-analysis has become a pillar of evidence-based medicine. But heterogeneity is still the threat to the validity and quality of meta-analysis. The



	Experimental			Control						
Study	Total	Mean	SD	Total	Mean	SD	MD	95%−CI	W(fixed)	W(random)
1	100	100.14152	0.8417839	100	9.778936	0.9793312	90.36	[90.11; 90.62]	4.7%	4.2%
2	100	100.01212	0.8286061	100	10.016458	0.7278386	90.00	[89.78; 90.21]	6.5%	4.4%
3	100	100.13962	1.1205406	100	9.801291	0.9342729	90.34	[90.05; 90.62]	3.7%	4.1%
4	100	100.34822	1.1715587	100	9.449311	0.9758557	90.90	[90.60; 91.20]	3.4%	4.1%
5	100	100.38152	0.9469777	100	10.246779	1.1113760	90.13	[89.85; 90.42]	3.7%	4.1%
6	100	100.31747	0.9685790	100	9.898116	0.6442580	90.42	[90.19; 90.65]	5.8%	4.3%
7	100	100.33562	0.9945070	100	10.047077	1.0296117	90.29	[90.01; 90.57]	3.8%	4.1%
8	100	99.93484	0.6901647	100	9.428783	1.1339673	90.51	[90.25; 90.77]	4.5%	4.2%
9	100	100.02829	0.9553336	100	10.341864	1.1685800	89.69	[89.39; 89.98]	3.5%	4.1%
10	100	99.71434	1.0069458	100	10.298181	1.1636163	89.42	[89.11; 89.72]	3.3%	4.1%
11	100	100.00556	0.9930413	100	10.355671	0.9984616	89.65	[89.37; 89.93]	4.0%	4.2%
12	100	99.69366	0.9539629	100	10.022895	0.7654707	89.67	[89.43; 89.91]	5.3%	4.3%
13	100	99.61912	1.0123540	100	9.700573	0.7545183	89.92	[89.67; 90.17]	4.9%	4.3%
14	100	100.19812	1.0063654	100	10.268070	1.0840274	89.93	[89.64; 90.22]	3.6%	4.1%
15	100	99.64595	0.8464126	100	9.900693	0.8830431	89.75	[89.51; 89.98]	5.3%	4.3%
16	100	100.32257	1.5088913	100	9.821018	1.0300785	90.50	[90.14; 90.86]	2.4%	3.8%
17	100	99.43523	1.0102745	100	9.905191	1.0435990	89.53	[89.25; 89.81]	3.7%	4.1%
18	100	100.43079	1.1398243	100	9.817009	1.1119776	90.61	[90.30; 90.93]	3.1%	4.0%
19	100	100.24393	1.0650932	100	9.963513	0.8817034	90.28	[90.01; 90.55]	4.1%	4.2%
20	100	99.54172	1.0317193	100	10.155360	1.0748076	89.39	[89.09; 89.68]	3.5%	4.1%
21	100	100.09922	1.0600581	100	9.889739	0.8929921	90.21	[89.94; 90.48]	4.1%	4.2%
22	100	100.36193	1.0443262	100	9.836177	0.7600718	90.53	[90.27; 90.78]	4.7%	4.2%
23	100	99.91823	1.0211976	100	9.954763	1.0640303	89.96	[89.67; 90.25]	3.6%	4.1%
24	100	100.43300	0.9533924	100	9.983192	0.8311117	90.45	[90.20; 90.70]	4.9%	4.3%
Fixed effect model	2400			2400			90.10	[90.04; 90.15]	100%	--
Random effects model							90.10	[89.94; 90.26]	--	100%

Heterogeneity: I−squared=87.9%, tau−squared=0.1379, p<0.0001

Fig 5. Simulated Meta-analysis. Enrolled 24 trials, total number 2400, pooled MD 90.1, $I^2 = 87.9\%$.

doi:10.1371/journal.pone.0127538.g005

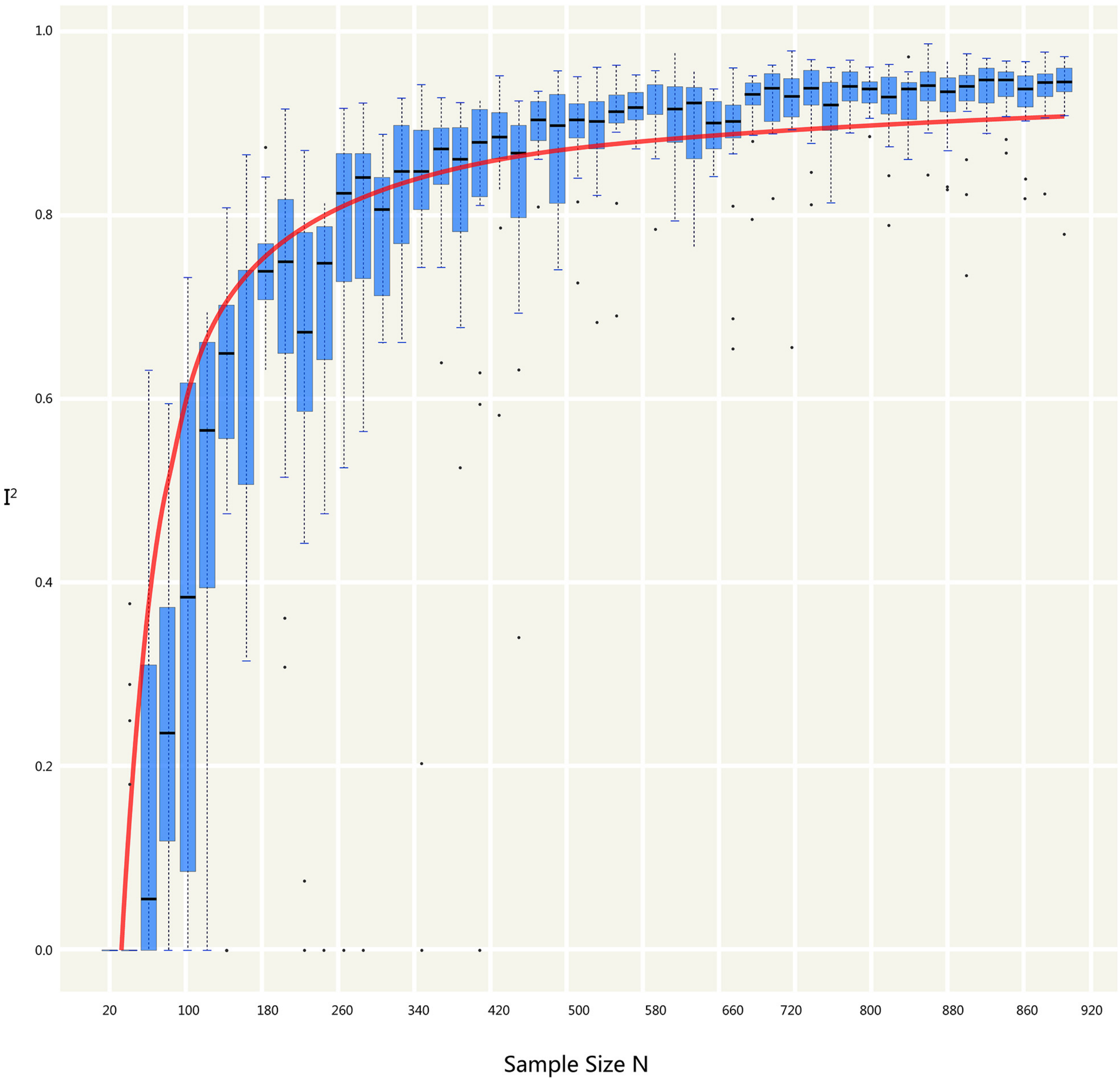


Fig 6. Impact of I^2 and the sample size. Lateral axis represents the sample size; vertical axis represents the I^2 value. Boxes represent the distribution of I^2 variations in synthesizing different number of trials (the sample size of each trial is the same). To each Si, sampling process will be repeated in 1000 times.

doi:10.1371/journal.pone.0127538.g006

core issue is to distinguish data sets from similar populations and exclude the others. First of all, currently meta-analysis researchers accept the data expressed as mean±sd by default as normal distribution, without any further analysis to test whether this distribution hypothesis is correct or not. Thus the heterogeneity challenge is quite real.

Secondly, almost none of the clinical researchers are aware that Q and I^2 are tools that can only be applicable to test heterogeneity between small sample size trials, and will lost their robustness when the sample sizes and the number of trials are substantially increased (as demonstrated by our study presented here).

This represents a dilemma: the purpose of meta-analysis is to enlarge the sample size, in order to expand and validate the implication of the result. New meta-analysis researches

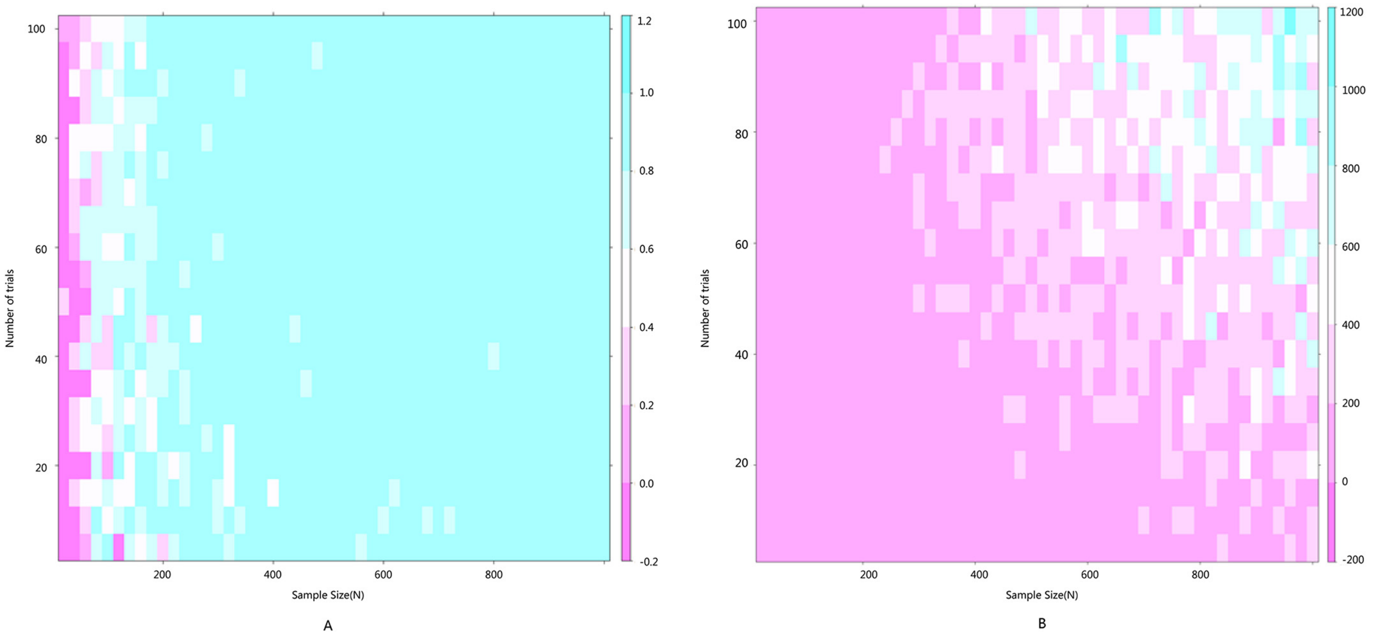


Fig 7. Heat maps of the impact of two heterogeneity test tools. A: Heat map of I^2 and the number of trials (n) and sample size N; B: Heat map of Q and the number of trials (n) and sample size N;.

doi:10.1371/journal.pone.0127538.g007

including these flaws are emerging and passed on to clinical practitioners as "updated evidence", but they are actually not strong as they assumed.

Conclusions

In summary, the validity of widely used Q and I^2 test in current meta-analysis is questionable and unstable on heterogeneity evaluation. Before new heterogeneity evaluation tool which is developed and its robustness are demonstrated, we will suggest more strict applications of meta-analysis. The meta-analysis may only be applied to those synthesized trials with small sample sizes. We call upon that the tools of evidence-based medicine should keep up-to-dated with the cutting-edge technologies in data science. Clinical research data should be made available publically when there is any relevant article published so the research community could conduct in-depth data mining, which is a better alternative for meta-analysis in many instances.

Supporting Information

S1 File. Tables A to G. Code of simulation algorithm and graphics plot in R.
(DOCX)

Acknowledgments

The authors would thank Mr. Lu He of University of Waikato, for his generous help on proofreading of the manuscript. We appreciate Dr. Kai Wang of Metabolomics and Multidisciplinary Laboratory of Sichuan Provincial People's Hospital, Sichuan Academy of Medical Sciences, for his help on manuscript editing.

Author Contributions

Conceived and designed the experiments: HJ HY. Performed the experiments: SL HY WC MS JZ. Analyzed the data: HJ HY SL CDL JP XP MS JZ. Wrote the paper: HJ HY CDL. Reviewed, commented on, and approved the manuscript: SL HJ HY WC JP MS CDL XP JZ.

References

1. Higgins J, Thompson S, Deeks J, Altman D. Measuring inconsistency in meta-analysis. BMJ. 2003; 327: 557–560. PMID: 12958120
2. Borenstein M. Fixed-Effect versus Random-Effects Models. In: Borenstein M, Hedges L, Higgins J, Rothstein H, editors. Introduction to Meta-Analysis. U.S.: John Wiley & Sons, Ltd; 2009. pp. 79–94.
3. Kulinskaya E. Evidence in Cochran's Q for heterogeneity of effects. In: Kulinskaya E, Morgenthaler S, Staudte R, editors. Meta-Analysis: A Guide to Calibrating and Combining Statistical Evidence. U.S.:, John Wiley & Sons, Ltd; 2008. pp. 209–220.
4. Schwarzer G. meta: Meta-Analysis with R. R package version 3.2–1. 2014; Available: http://CRAN.R-project.org/package = meta
5. Kabacoff R. Generalized linear models. In: Kabacoff R, editors. R in Action: Data Analysis and Graphics with R. U.S., Manning Publications; 2011.pp. 158–197.
6. Rücker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I^2 in assessing heterogeneity may mislead. BMC Med Res Methodol. 2008; 8:79 doi: 10.1186/1471-2288-8-79 PMID: 19036172