From Sample Size to Effect-Size: Small Study Effect Investigation (SSEi)

F Richy, O Ethgen, O Bruyere, F Deceulaer, J Reginster

Citation

F Richy, O Ethgen, O Bruyere, F Deceulaer, J Reginster. *From Sample Size to Effect-Size: Small Study Effect Investigation (SSEi)*. The Internet Journal of Epidemiology. 2003 Volume 1 Number 2.

Abstract

A small-sized study may report over or under estimated effects of the investigated treatment in randomized controlled trials, which is called small-study effect (SSE). It is commonly related to publication bias. Notwithstanding, the intrinsic, probabilistic, component of SSE has never been assessed yet, which is the purpose of this study.

A stochastic model, simulating the results of a given controlled trial with an increasing number of subjects in each group (from 1 to 200) was used. Predefined sets of input data (expected means and standard deviations) covering a full range of analytical situations were entered in a pseudo-random generator to reflect the variability of the individual's responses in each group. For each set of data, the process was repeated 200 times to take variability into account and therefore to investigate the validity of the model. Effect-size and its standard normal deviate were notably computed for each sample-size in order to determine a threshold for SSE.

The median (25%; 75%; 90%) sample above which SSE was absent was 16.5 (8;30;50) subjects per group and was non-significantly impacted by the different input data sets. A sample size of 50 subjects in each group allowed for no small-study effect in 90% of the simulations.

This study pointed out the fact that SSE is not only linked to selective publication, as well as the rationale to take it into account in addition to power calculation in the design of a RCT as well as in publication bias analysis.

BACKGROUND

The modern era of clinical epidemiology began in the late 1940's with the pioneering work of Bradford Hill. [1] Randomized controlled trials (RCT) are nowadaysbecame the keystone of a medicine progressively based on evidence. RCT both accounts for the confirmation of preliminary studies and for the evidence provided by meta-analyses. Their progressively standardized design and analytical techniques provided the public health specialist more reliable tools to assess and compare therapeutic approaches. As RCT were growing in number and, more recently, included in quantitative systematic reviews, quality assessment strategies were developed, including component approaches, evaluating selected aspects of the trial; checklists, involving lists of items and scales, providing an integrated numeric score of quality such as Jadad Score [2]. Nevertheless, little evidence is supporting their validity due to the lack of empirical research in the field. [3] A seldom considered item is the appropriateness of clinical epidemiology and statistics and their interpretation, also called type III errors. One of the most misunderstood and poorly investigated type III error is called “Small Study Effect” (SSE): the fact that small-sized trials in terms of included subjects are prone to provide over- (or under-) estimated results [4]. SSE is linked to several concepts. First, publication bias was reported in the late eighties and shown to lead to the preferential publication of small trials showing a statistically significant effect of the investigated treatment [5, 6]. Secondly, it was noticed that, when compared to smaller ones, larger studies might give a subtly differential intensity of interventions (i.e. lower doses), or show differences in the underlying confounding factors or symptom patients characteristics at inclusion (milder ones) [7, 8, 9] which provides more conservative estimates than small-sized ones. Thirdly, the probabilistic component of SSE may be due to a high variability of estimates when the number of observations is low. This may provide pseudo-random fluctuations of these parameters and a biased estimation of

the “real” effect. One could think that, even if the effect-size (ES) is potentially biased in small-sized studies, the significance level could not be sufficient to induce a type I error (wrongly positive assumption of difference), and consequently SSE would have a limited impact on decision making in medicine. On the other hand, it must be stated that today, no study directly focused on this effect, nevertheless prone to bias entire research lines: generally, non-significant results associated with high effect-sizes are attributed to a lack of statistical power, and the authors facing such a result often conclude “nevertheless, further studies of higher sample size are needed to clearly assess the efficacy of the treatment (...)”. A meta-analysis of such trials has then good “chances” to produce significant but biased results, and in contrast high chances to be accepted for publication [10]. Recently, statistical methods were developed to deal with publication bias, including funnel plot analysis [[[>7]]] and “trim and fill” methods [11]. Notwithstanding, those methods suffer low statistical power and may not be adequate when the studies are heterogeneous (that is, when they estimate different effects) [12]. Consequently, SSE remains directly and indirectly a serious threat on the validity of the evidence provided by RCTs and meta-analyses, respectively.

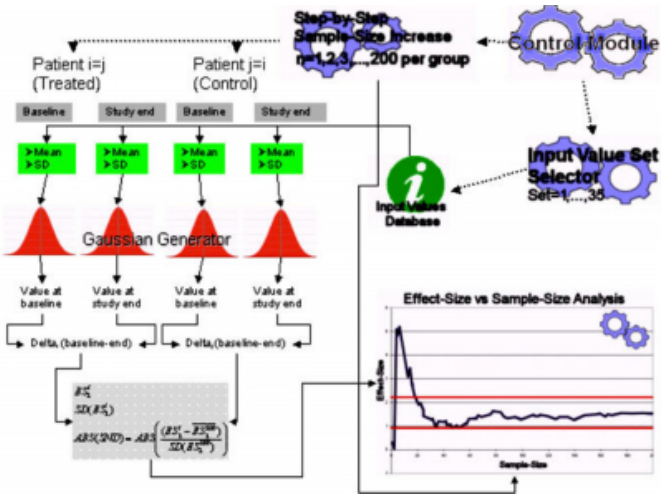
We conducted this study in order to investigate the relationship between sample-size and effect-size. Our secondary objective was to define a cut-off value, in terms of sample-size for SSE. To this end, a stochastic model, reflecting the variability of the individual's response to allocated treatment was specifically developed.

METHODS

The general structure of the SSE model is described in Figure 1. The input data were the expected means and standard deviations (SD) in groups 1 and 2, at the beginning and the end of the simulated trial (G1t0, G1t1, G2t0 and G2t1, respectively). Out of these parameters, pseudo-random, Gaussian values were generated for G1t0, G1t1, G2t0 and G2t1 to reflect the variability of the individual responses in each group. The “before-after” differences (G1t1-G1t0 and G2t1-G2t1) were used for the calculations. The application generated, step by step, 200 trials of the 35 different input sets, of which the sample-size ranged from 1 record per group to 200 patients per group. The whole process was repeated 200 times per input values set, so, a total of 1.400.000 iterations were used for statistical assessment.

Figure 1

Figure 1: SSE Algorithm



Effect-sizes of the group 1 against group 2 were calculated for each iteration by using the pooled standard deviation as denominator [13], as opposed to Glass score [14]. This method has been shown to be more reliable and conservative approach [15]. Standard normal deviates were calculated at each iteration, by:

Figure 2

$$ABS(SND) = ABS\left(\frac{(ES_2^i - ES_2^{200})}{SD(ES_2^{200})}\right)$$

Figure 3

$$SD(ES_2^{200}) = \sqrt{\frac{(n_n - 1).(s_n)^2 + (n_\alpha - 1).(s_\alpha)^2}{n_n + n_\alpha - 2}}$$

The threshold for SSE was 1.96. Any ESi being out of this interval was considered small-study effect, and the corresponding sample-size was highlighted. A threshold for small study effect was determined by establishing the maximal sample-size corresponding to ABS(SND)<1.96 for each set of data. The variability of this threshold was taken into account by simulating 200 times the same trial, while increasing its sample-size, with 35 different sets of data (table 1), so 1.400.000 (200*200*35) iterations were requested to this end. The distribution of the 200 SSE thresholds was plot against the 200 corresponding sample-sizes for each of the 35 input values sets, and the main threshold was set as the percentiles 25, 50, 75 and 90 of the sample-sizes corresponding to those thresholds.

Figure 4

Table 1: Input data table

		Group 1 (treated)				Group 2 (control)			
Name	Step	G1t0	sd	G1t1	sd	G2t0	sd	G2t1	sd
No difference SD increasing	1	100	1	100	1	100	1	100	1
	2	100	10	100	10	100	10	100	10
	3	100	20	100	20	100	20	100	20
	4	100	50	100	50	100	50	100	50
	5	100	80	100	80	100	80	100	80
Delta G1>G2 Fixed SD	6	100	10	90	10	100	10	100	10
	7	100	10	70	10	100	10	100	10
	8	100	10	50	10	100	10	100	10
	9	100	10	30	10	100	10	100	10
	10	100	10	10	10	100	10	100	10
Delta G1<G2 Fixed SD	11	100	10	100	10	100	10	90	10
	12	100	10	100	10	100	10	70	10
	13	100	10	100	10	100	10	50	10
	14	100	10	100	10	100	10	30	10
	15	100	10	100	10	100	10	10	10
Delta G1>G2 SD G1>G2	16	100	20	70	20	100	10	90	10
	17	100	30	70	30	100	10	90	10
	18	100	50	70	50	100	10	90	10
	19	100	70	70	70	100	10	90	10
	20	100	90	70	90	100	10	90	10
Delta G1>G2 SD G1<G2	21	100	10	70	10	100	20	90	20
	22	100	10	70	10	100	30	90	30
	23	100	10	70	10	100	50	90	50
	24	100	10	70	10	100	70	90	70
	25	100	10	70	10	100	90	90	90
Delta G1<G2 SD G1>G2	26	100	20	90	20	100	10	70	10
	27	100	30	90	30	100	10	70	10
	28	100	50	90	50	100	10	70	10
	29	100	70	90	70	100	10	70	10
	30	100	90	90	90	100	10	70	10
Delta G1<G2 SD G1<G2	31	100	10	90	10	100	20	70	20
	32	100	10	90	10	100	30	70	30
	33	100	10	90	10	100	50	70	50
	34	100	10	90	10	100	70	70	70
	35	100	10	90	10	100	90	70	90

The magnitude of SSE was defined as the ratio of difference highest ES of the set and the ES for 200 records in each group and the ES for 200 records in each group.

The data were generated using Microsoft® Visual Basic 6.0 and filed away in spreadsheets using Microsoft® Excel XPpro, while statistical analyses were performed using Statsoft Statistica 6.0 ®. The application was run on an Intel® Pentium 4 based system at 2.0 Ghz.

RESULTS

Figure 2 displays an example of the primary outputs obtained from the application, corresponding to a given set of input values.

Figure 5

Figure 2: Example of primary SSE outputs illustrated by ES for a single dataset and 200 iterations

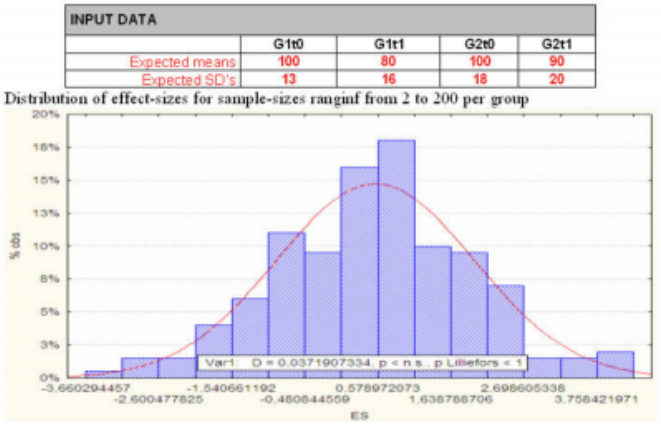
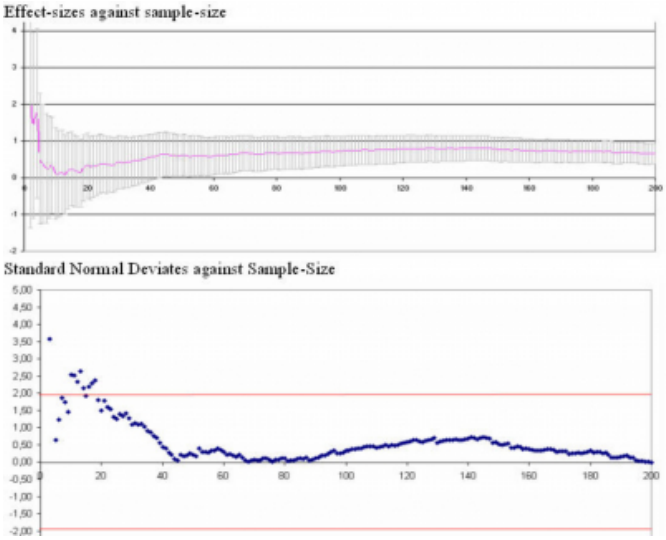


Figure 6

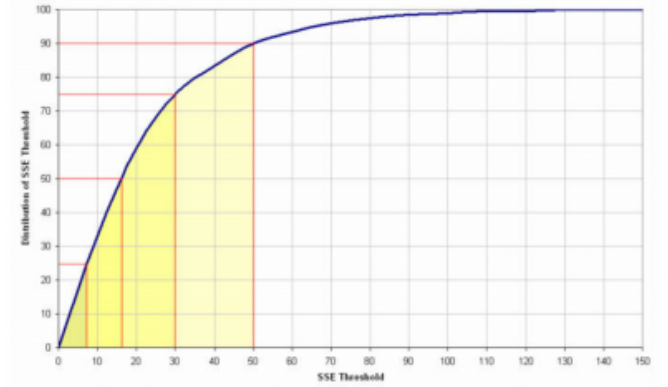


SMALL STUDY EFFECT INVESTIGATION

We found that the input values had no incidence on the obtained SSE threshold (chi-square goodness of fit test, p>0,10). At the end of the 35 simulations of 200 steps each, the median (25%; 75%) sample above which the SSE was considered non-statistically significant was 16,5 (8,30) subjects in each group (figure 3). The distribution of all SSE thresholds fit an exponential distribution (Kolmogorov-Smirnov d=0,026, p=ns). A sample size of 50 subjects in each group allowed for no small-study effect in 90% of the simulations. The application also provided an estimation of the maximal SSE size (%) in function of the input parameters. SSE size was found to be inversely proportional to the effect-size and a power regression model furnished: $SSEsize = 69,76 \times ES^{-1,0443}$, $R^2=0,998$ (figure 4).

Figure 7

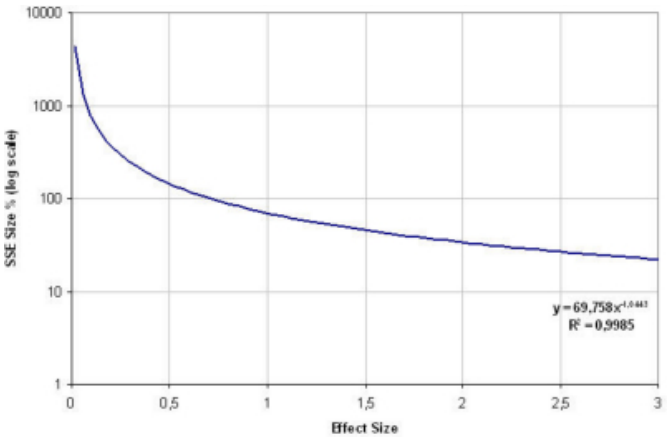
Figure 3: Cumulative distribution of SSE Threshold



Percentile	25	50	75	90
Sample-size	8	16,5	30	50

Figure 8

Figure 4: SSE Size in function of effect-size



DISCUSSION AND CONCLUSION

The results of this study provided a clear rationale to consider small study effect in the methodological design of a trial, and not only when discussing the results of a quantitative systematic review. Hence, a restricted sample-size, even allowing for sufficient statistical power in case of a projected high treatment efficacy, led to a potential over or under estimation of the effect-size that can not be detected in posterior peer reviewing processes. As a matter of fact, the scientific literature remains extremely poor on this phenomenon, and SSE has great chances to be disregarded: using a publication bias detection tool (funnel plot, precision against standard normal deviate) leading to significant results, the authors would certainly conclude that more small sized studies favoring the investigated treatment were published compared to non-significant or negative ones. A part of this bias can be explained by SSE, since the effect-sizes of these studies are prone to be under- or over-estimated. We logically found that SSE size was inversely proportional to effect-size and provided an estimate for it.

Notwithstanding, this study had clear limitations. The pseudo-random generator was only able to produce Gaussian values, while certain biomedical outcomes (i.e. platelets rates) have asymmetric distributions. In other words, in our model, all patients had the same probability of response depending on the allocated treatment. In reality, a population is never as homogeneous and subgroups of “good” and “bad” responders are often observed even if the randomization was properly applied. More distributions should be tested, depending on the investigated variables. Secondly, the paper is only concerned with the difference between two population means. Many other types of effect-sizes exist: for differences between proportions, for comparing more than two means, for correlations of several types, etc. Thirdly, we did not investigate whether a differential inter group withdrawal rate of patients could impact the relationshop between sample-size and effect-size. In particular, the hypothesis about which intention-to-treat analysis is always more conservative than the per-protocol approach should be carefully tested in the light of stochastic models, since its handling might be often inadequate [16]. Finally, due to limited informatics resources we had to restrict the maximal sample size to 200 per group. Indeed such algorithms require high resources and therefore any increase in its capabilities involves a geometric rising of the needed computational time.

In conclusion, this empirical study gave a rationale to point out the fact that small study effect is not only linked to selective publication policies. Further assessment is needed to be able to provide a more comprehensive integration of SSE to publication bias analysis and power calculation.

CORRESPONDENCE TO

Florent RICHY University of Liège Public Health and Epidemiology Unit CHU – Bât B23 B-4000 SART-TILMAN Belgium EUROPE Tel. –32-4-3662581 Fax –32-4-3662812 E-Mail: florent.richy@ulg.ac.be

References

1. Silverman WA, Chalmers I. Sir Austin Bradford Hill: an appreciation. Control Clin Trials 1992; 13: 100-105
2. Jadad AR, Moore RA, Carroll D, Jenkinson C. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials 1996; 17: 1-12
3. Moher D, Jadad AR, Tugwell P. Assessing the quality of randomized controlled trials. Current issues and future directions. International J Technol Assess Health Care 1996; 12: 195-208

4. Borenstein M. Hypothesis testing and effect-size estimation in clinical trials. Ann Allergy Asthma Immunol 1997 ; 78 : 5-16
5. Begg CB, Berlin JA. Publication bias ; a problem in interpreting medical data. J R Statist Soc A 1988; 151: 445-63
6. Dickersin K, Chan S, Chalmers TC, Sacks HS, Smith H. Publication bias and clinical trials. Controll Clin Trials 1987; 8: 343-53
7. Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. BMJ 1998; 317: 1185-90
8. Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997; 315: 629-34
9. Richy F, Bruyère O, Ethgen O, Cucherat M, Henrotin Y, Reginster JY. Structural and symptomatic efficacy of glucosamine and chondroitine sulfate in knee osteoarthritis: a comprehensive meta-analysis. Arch Int Med 2003, in press
10. Anonymous. Half of meta-analyses may contain publication bias. BMJ 2000 ; 320 :E
11. Duval S, Tweedie R. Trim and fill : A simple funnel-plot-based method of testing and adjusting for publication bias. Biometrics 2000 ; 56 : 241-9
12. Terrin N, Schmid CH, Lau J, Olkin I. Adjusting for publication bias in the presence of heterogeneity. Stat Med 2003 ; 22 : 2113-26
13. Cohen J (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ; Lawrence Earlbaum Associates:P44
14. Glass G, McGaw B,Smith ML (1981). Meta-analysis in Social Research. Beverly Hills: Sage.
15. Rosnow RL, Rosenthal R. Computing, contrasts, effect sizes, and countermulls on other people's published data : General procedures for reaserach consumers. Psychological Methods, I, 331-340
16. Wright CC, Sim J. Intention-to-treat analysis approach to data from randomized controlled trials : a sensitivity analysis. J Clin Epidemiol 2003 ; 56 : 833-42

Author Information

Florent Richy
Department of Public Health, Public Health and Epidemiology Unit, Faculty of Medicine, University of Liège

Olivier Ethgen
Department of Public Health, Public Health and Epidemiology Unit, Faculty of Medicine, University of Liège

Olivier Bruyere
Department of Public Health, Public Health and Epidemiology Unit, Faculty of Medicine, University of Liège

Frédéric Deceulaer
Department of Public Health, Public Health and Epidemiology Unit, Faculty of Medicine, University of Liège

Jean-Yves Reginster
Department of Public Health, Public Health and Epidemiology Unit, Faculty of Medicine, University of Liège