The Stata Journal (2020)
20, Number 3, pp. 548–564 DOI: 10.1177/1536867X20953567

emagnification: A tool for estimating effect-size magnification and performing design calculations in epidemiological studies

David J. Miller
Office of Pesticide Programs
U.S. Environmental Protection Agency
Washington, DC
miller.davidj@epa.gov

James T. Nguyen
Office of Pesticide Programs
U.S. Environmental Protection Agency
Washington, DC
nguyen.james@epa.gov

Matteo Bottai
Division of Biostatistics
Institute of Environmental Medicine, Karolinska Institutet
Stockholm, Sweden
matteo.bottai@ki.se

Abstract. Artificial effect-size magnification (ESM) may occur in underpowered studies, where effects are reported only because they or their associated p-values have passed some threshold. Ioannidis (2008, *Epidemiology* 19: 640–648) and Gelman and Carlin (2014, *Perspectives on Psychological Science* 9: 641–651) have suggested that the plausibility of findings for a specific study can be evaluated by computation of ESM, which requires statistical simulation. In this article, we present a new command called `emagnification` that allows straightforward implementation of such simulations in Stata. The commands automate these simulations for epidemiological studies and enable the user to assess ESM routinely for published studies using user-selected, study-specific inputs that are commonly reported in published literature. The intention of the command is to allow a wider community to use ESMs as a tool for evaluating the reliability of reported effect sizes and to put an observed statistically significant effect size into a fuller context with respect to potential implications for study conclusions.

Keywords: st0608, emagnification proportion, emagnification rate, inflation, magnification, p-value, type M error, effect-size magnification, winners curse

1 Introduction

Effect-size magnification (ESM) is a phenomenon by which low-powered studies that detect an effect will characteristically tend to exaggerate the size of that effect when the effect is required to pass some kind of threshold such as the $p < 0.05$ criterion often used for statistical significance. As an example of how ESM may occur, it is useful to imagine a thought experiment in which a trial is run thousands of times, each with different sample sizes. In practice, experiments are generally conducted once and have one fixed sample size. In this thought experiment, a broad distribution of

observed effect sizes over the thousands of runs will be seen. The observed medians of these estimated effect sizes are expected to be close to the true effect size regardless of sample size. However, trials from smaller-sized studies from these simulations will in fact systematically produce wider variation in observed effect sizes than the larger trials. Further, only a small proportion of the observed effects in these small-size, low-power studies will pass any given statistical threshold of significance, popularly $p < 0.05$. Thus, when associations deemed "statistically significant" are found, they are more likely to overestimate the true size of that effect. This is the ESM phenomenon, which can lead to misinterpretation of inflated results from an experiment or observational study as important or "discovered" scientific findings. Stated mathematically: conditional on a result passing some predetermined threshold of statistical significance, test level, or magnitude, the estimated effect size is a biased estimate of the true effect size with the magnitude of this bias inversely related to power of the study.

The remainder of this article introduces the new command `emagnification`, which facilitates performing these ESM simulations. The command is a tool that permits reported statistically significant effect-size estimates from possibly underpowered epidemiological studies to be better evaluated and judged. It is an outgrowth of work done by two of the authors as part of a European Food Safety Agency (EFSA) Panel on Plant Protection Products and their Residues (PPR) and is the result of an expansion and extension of the original EFSA PPR panel work during post-PPR panel collaboration by the authors.[1] The command follows directly from the work of Ioannidis (2008) and, to a lesser extent, the work of Gelman and Carlin (2014). Ioannidis (2008) illustrates by simulation the ESM phenomenon; these ideas were incorporated into `emagnification` so that the calculations similar to those done by Ioannidis to estimate the degree of potential ESM in any given study can be easily performed in Stata. The Stata simulations can be useful when evaluating reported effect sizes in a published epidemiological article, and the new command makes this a simple numerical exercise. As such, it can assist individuals reviewing such studies to put an observed statistically significant effect size into a fuller context that allows better judgments regarding adequacy of sample size vis-à-vis the observed effect size. In doing so, users will gain a better understanding of power and sample-size issues and in interpreting their potential implications with respect to study conclusions.

2 ESM: Illustrating the phenomenon

For illustrative purposes and as an introduction to the issue, we draw on the concrete example from the work of Ioannidis (2008) discussed in the introduction and appearing

1. As part of their work on the PPR, two of the authors (Matteo Bottai and David J. Miller) contributed to the review and writing of "Scientific opinion of the PPR panel on the follow-up of the findings of the external scientific report 'Literature review of epidemiological studies linking exposure to pesticides and health effects'" (EFSA Panel on Plant Protection Products and their Residues [PPR] 2017) and its Annex D, where much of this material originally appeared. The ESM calculations and simulations appearing in Annex D of that EFSA report were generated using a custom-coded SAS program by one of the authors (James T. Nguyen) of this article. The results for the current `emagnification` command introduced here were tested against this earlier SAS script and compared favorably.

in table 2 of his article. This section shows how to replicate this analysis with the new command.

Ioannidis's table 2 is excerpted here as table 1. Ioannidis generated this from a series of simulations designed to illustrate the ESM phenomenon in which a low study power can be seen to lead to exaggerated effect sizes for those results that are statistically significant. As shown in the first data row of table 1, Ioannidis begins by assuming a true odds ratio (OR) for an association of 1.10 and that the proportion of exposed individuals in the control (or nondiseased) group is 30%. It follows then, mathematically, that the expected proportion of exposed individuals in the case group would be 0.3204.[2] Ioannidis then simulates a set of epidemiological studies in which i) the control group in each simulated study includes 1,000 subjects and the number of exposed subjects within the control group is randomly drawn from a binomial distribution with probability 0.3000 representing the control group proportion; and ii) the case group in each simulated study includes 1,000 subjects and the number of exposed subjects within the case group is randomly drawn from a binomial distribution with probability of 0.3204 representing the case group proportion. The observed OR of each of many simulated studies in which "n" samples are drawn per group is then computed and stored. The median OR of these simulated studies is expected to be equal to the true OR value of 1.10 (as can be calculated by the new `emagnification` command),[3] but we would expect that only a proportion of those observed ORs that happened to have large values would be statistically significant ($p < 0.05$). This is what is illustrated in table 1, focusing on and highlighting the simulation results of the ORs that happened to be found significant at $p < 0.05$. When we look at only those ORs that pass this $p < 0.05$ statistical threshold, the medians among this subset of (statistically significant) ORs are observed to be 1.23, shown in the first row of data in table 1, which is higher by 11% than the true OR of 1.1 used to generate the simulation. In fact, table 1 shows that a considerable fraction ($> 75\%$) of the simulated significant ORs are inflated compared with the true OR of 1.10 because the interquartile range (IQR) of the medians were found in Ioannidis's simulations to be (1.20–1.29).[4] This phenomenon illustrates (via computer simulation) that when a researcher's or data user's focus is on statistically significant results, these results will be systematically biased high (magnified or inflated) for underpowered studies. As the sample size gets smaller (say, from 1,000 in each of the comparison groups in the third line of table 1 to 250 as shown in the fourth line and finally to 50 in the last line) and power thus becomes lower, the magnification becomes greater, going from a relatively small 3% for the case of $n = 1000$ to as much as 118% for the $n = 50$ case!

2. $P1 = (P0 \times \mathrm{OR})/\{(1 - P0) + (P0 \times \mathrm{OR})\}$, where $P1$ = expected proportion of exposed individuals among cases; $P0$ = expected background or control group proportion; and OR = true odds ratio between exposed and control individuals.
3. When we set the `level()` option very close to 1 at `level(0.9999)`, the command `emagnification proportion, p0(0.30) or(1.1) n0(1000) n1(1000) pctile(50) nsim(5000) level(0.9999) onesided seed(123)` produces a `p50` value of 1.100.
4. Simulation with `emagnification` done by the authors suggests the 1.23 listed as `p25` in table 2 of Ioannidis (2008) is a typographical error, and the actual value for the `p25` should be closer to 1.20.

Table 1. Simulations for Effect Sizes Passing the Threshold of Formal Statistical Significance ($P = 0.05$)

True OR	Control Group Rate (%)	Sample n Per Group	Observed OR in Significant Associations	
			Median (IQR)	Median Fold Inflation
1.10	30	1000	1.23 (1.20 – 1.29)	1.11
1.10	30	250	1.51 (1.49 – 1.55)	1.37
1.25	30	1000	1.29 (1.26 – 1.67)	1.03
1.25	30	250	1.60 (1.50 – 1.67)	1.28
1.25	30	50	2.73 (2.60 – 3.16)	2.18

IQR indicates interquartile range.

While table 1 above is useful to illustrate the ESM phenomenon, demonstrate how it arises, and quantify it, the tabulated conditions and numbers in table 1 are necessarily fixed, and it would be useful to be able to generate such a table "on the fly" with inputs that are specific to a study of a researcher's particular interest. This is what the new `emagnification` command accomplishes.

3 The emagnification command

3.1 Syntax

Effect-size magnification for proportions

`emagnification` p`roportion, p0(`*numlist*`) or(`*numlist*`) n0(`*numlist*`) n1(`*numlist*`)`
`[pctile(`*numlist*`) ifactor(`*numlist*`) nsim(`#`) level(`#`) onesided exact`
`seed(`*string*`) log clean format(`*format*`)]`

Effect-size magnification for rates

`emagnification` r`ate, r0(`*numlist*`) rr(`*numlist*`) n0(`*numlist*`) n1(`*numlist*`)`
`[pctile(`*numlist*`) ifactor(`*numlist*`) nsim(`#`) level(`#`) onesided exact`
`seed(`*string*`) log clean format(`*format*`)]`

3.2 Description

The `emagnification` command estimates effect-size magnification of proportions and rates through simulations.

The `emagnification proportion` syntax estimates ORs with the `logit` command.

The `emagnification rate` syntax estimates relative risks with the `poisson` command.

Iterations that do not converge (for example, zero events are generated that may happen with small counts) are dropped, with the number of valid (completed) iterations shown at the end of the Stata run in the results table in the column labeled `valid`. If all iterations are completed with valid results and no runs are dropped, this will equal the number of iterations requested by the user. The program will continue running even with invalid (and dropped) results.

3.3 Options

`p0(`*numlist*`)` specifies the proportions in the reference group when used with the `emagnification proportion` syntax. This is sometimes estimated in case–control studies using ORs as the number of exposed subjects in the reference (control) group divided by the number of subjects in the reference group. `p0()` is required for `emagnification proportion`.

`or(`*numlist*`)` specifies the ORs of the case (comparison) group versus the reference group. `or()` is required for `emagnification proportion`.

`r0(`*numlist*`)` specifies the rates in the reference group when used with the `emagnification rate` syntax. This is sometimes estimated in cohort studies using rate ratios (or relative risks) as the number of diseased individuals in the reference (unexposed) group divided by the number of subjects in the reference group. `r0()` is required for `emagnification rate`.

`rr(`*numlist*`)` specifies the risk ratios of the exposure (comparison) group versus the reference group. `rr()` is required for `emagnification rate`.

`n0(`*numlist*`)` specifies the sample size in the reference group. `n0()` is required.

`n1(`*numlist*`)` specifies the sample size in the comparison group. `n1()` is required.

`pctile(`*numlist*`)` specifies the percentiles of the distribution of significant effects sizes specified in *numlist* where what is significant is defined in the `level()` option. The default is `pctile(10 50 90)`.

`ifactor(`*numlist*`)` specifies the inflation factors of the percentiles specified in *numlist* where the inflation (exaggeration) factor is equal to the relevant percentile divided by the true odds ratio. The set of percentiles specified in the `ifactor()` option may be different from that specified in the `pctile()` option. For example, both the following two lines are allowed:

```
emagnification proportion, p0(.5) or(2) n0(100) n1(100) pctile(50) ifactor(50)
emagnification proportion, p0(.5) or(2) n0(100) n1(100) pctile(50) ifactor(90)
```

`nsim(#)` specifies the number of simulated datasets. The default is `nsim(10)`.

`level(#)` specifies the significance level of the test. The default is `level(0.05)`.

`onesided` specifies a one-sided test. The default is two sided.

`exact` for `emagnification proportion` specifies Fisher's exact test instead of the default chi-squared test; `exact` for `emagnification rate` specifies the exact Poisson regression instead of the default Poisson regression.

`seed(`*string*`)` specifies the seed for the pseudo-random-number generator. The `emagnification` command is based on simulated pseudo–random data. Therefore, the same command line can produce nonidentical results, when run multiple times.

The pseudo-random-number generator seed is stored in the `r(seed)` macro, regardless of whether the `seed()` option is specified in the `emagnification` command. The `r(seed)`-stored macro can be used to replicate the results of the latest simulation. For example, the following two lines, run consecutively, produce identical estimates:

```
emagnification proportion, p0(.5) or(2) n0(100) n1(100)
emagnification proportion, p0(.5) or(2) n0(100) n1(100) seed(`=r(seed)´)
```

`log` shows the simulation iterations. This is convenient to track Stata's progress on long runs with many iterations.

`clean` shows the results without separator lines.

`format(`*format*`)` specifies the display format for the percentiles.

3.4 Stored results

`emagnification` stores the following in `r()`:

Scalars	
`r(level)`	level of the tests
Macros	
`r(cmdline)`	command as typed
`r(seed)`	seed used by the pseudo-random-number generator
Matrices	
`r(table)`	table of the results

3.5 Example inputs

The following examples illustrate the `emagnification` command:

Estimate the effect-size magnification for a proportion

```
emagnification proportion, p0(.5) or(2) n0(100) n1(100)
```

Estimate the effect-size magnification for a rate

```
emagnification rate, r0(.5) rr(2) n0(100) n1(100)
```

Estimate the effect-size magnification for a proportion in multiple scenarios using 0.1 as the level of significance and showing the inflation factor for only the median statistically significant result

```
emagnification proportion, p0(.5 .9) or(1.5 2) n0(100 200) n1(100 200)   ///
   pctile(50 90) ifactor(50) nsim(100) level(.1) onesided seed(123) log clean
```

Show the stored results of the latest estimation

```
return list
matrix list r(table)
```

4 ESM: Illustrating the emagnification command

To illustrate the use of the `emagnification` command, we revisit here the Ioannidis (2008) example in section 2 and use `emagnification` to replicate the simulation results shown there. To do this, the `emagnification` command requires the following four input values:

1. the number of subjects in the reference group;
2. the number of subjects in the comparison group;
3. the specific proportion or rate of interest in the reference group (here this is the proportion of exposed subjects in the control group because the Ioannidis example uses ORs); and
4. the assumed (true) ORs (here) or rate ratios of interest.

In the `emagnification proportion` version of the syntax,[5] because we are looking at ORs, we have the following inputs (all derived from the first row of table 1):

- `n0(`*numlist*`)` is the number of subjects in the reference group, here 1,000 control subjects as listed under the "Sample n Per Group" column from table 1;
- `n1(`*numlist*`)` is the number of subjects in the comparison group, here 1,000 case subjects as listed under the "Sample n Per Group" column from table 1;
- `p0(`*numlist*`)` is the proportion of interest in the reference group; in the Ioannidis example, this is the proportion of exposed subjects in the control group = 0.30 as listed under the "Control Group Rate (%)" column of table 1;
- `or(`*numlist*`)` is the assumed true ORs, here 1.10, as listed under the "True OR" column in table 1.

We insert these values into the `emagnification proportion` command with several additional options described in section 3 to simulate the results from the first row of table 1 from Ioannidis:

```
. emagnification proportion, p0(0.30) or(1.1) n0(1000) n1(1000) pctile(25 50 75)
> ifactor(50) nsim(1000) level(0.05) onesided seed(123)

The tests are one-sided with level = .05
```

p0	p1	true_or	n0	n1	valid	power	p25	p50
.3	.3203883	1.1	1000	1000	1000	.274	1.202	1.235

p75	if_p50
1.289	1.123

As can be seen, the values for the `p50` (that is, median) of 1.235 and IQR of (1.202–1.289) approximate well those given in the simulation performed by Ioannidis.

Using similar syntax and taking advantage of the ability of the `emagnification` command to use Stata *numlists*, we can replicate both the second and fourth rows of table 1 with the following single command, adding the `log` option to monitor Stata's progress in real time:

5. For the syntax and detailed example associated with the `emagnification rate` version of the command, see the earlier working paper on ESM available for download from the Karolinska Institute website at http://www.imm.ki.se/biostatistics/emagnification/.

```
. emagnification proportion, p0(0.30) or(1.10 1.25) n0(250) n1(250)
> pctile(25 50 75) ifactor(50) nsim(1000) level(0.05) onesided seed(123) log
Scenario 1: p0 = .3, or = 1.1, n0 = 250, n1 = 250
Completed: 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
Scenario 2: p0 = .3, or = 1.25, n0 = 250, n1 = 250
Completed: 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
The tests are one-sided with level = .05
```

p0	p1	true_or	n0	n1	valid	power	p25	p50
.3	.3203883	1.1	250	250	1000	.114	1.423	1.481

p75	if_p50
1.563	1.346

p0	p1	true_or	n0	n1	valid	power	p25	p50
.3	.3488372	1.25	250	250	1000	.298	1.440	1.519

p75	if_p50
1.623	1.215

Similarly, the values estimated here by Stata (medians) of 1.481 and 1.519 and respective IQRs of (1.423–1.563) and (1.440–1.623) approximate reasonably well those provided by Ioannidis in table 1. The remaining two simulations from Ioannidis (corresponding to the third and fifth rows in table 1) can be re-created using the following two commands:

```
. emagnification proportion, p0(0.30) or(1.25) n0(1000) n1(1000)
> pctile(25 50 75) ifactor(50) nsim(1000) level(0.05) onesided seed(123) log
```

(*output omitted*)

```
. emagnification proportion, p0(0.30) or(1.25) n0(50) n1(50) pctile(25 50 75)
> ifactor(50) nsim(1000) level(0.05) onesided seed(123) log
```

(*output omitted*)

Similarly, these last two commands correspond to the simulation values generated by Ioannidis.[6] Importantly, the above series of simulations illustrate as he did that the more underpowered (and generally smaller) a study is and the smaller the true effect

6. The reader will note that the last command, which uses `n0(50) n1(50)` as sample sizes, produces results that differ somewhat from those of Ioannidis. Specifically, the `emagnification` command produces a `p50` and IQR of 2.344 and $(2.173, 2.786)$, while the Ioannidis results are 2.73 and $(2.60, 3.16)$ and `p50` inflation factors of 1.891 and 2.18, respectively. This is likely because Ioannidis used a test other than that used by `emagnification`. More specifically, as stated in the `emagnification` help file, `emagnification proportion` uses a chi-squared and score test, while Ioannidis uses a Wald test from logistic regression. The chi-squared and score test, Wald test, and likelihood-ratio test are asymptotically equivalent, but the chi-squared and score test and the likelihood-ratio test tend to perform better in small samples. This might explain why the difference is largest in the smallest sample scenario (row 5) with the lowest power (at 15%) in table 1 and negligible in the other rows. Arguments can be advanced for either type of test (`emagnification`'s chi-squared and score versus Ioannidis's Wald test shown in table 1), but we chose the test that tends to work better with smaller samples. Regardless of the numerical differences, the substantive issues are unaffected: both Ioannidis's results and the results from `emagnification` convey the same message: that the effect size for "discovered" statistically significant results may be substantially inflated by about twofold.

size that the study is investigating, the greater the degree that observed effect sizes that pass some preestablished statistical threshold or are by other means "discovered" will be inflated. Here we see the median inflations vary from 3% with a (moderate) true OR of 1.25 and a large sample of 1,000 to a near doubling of the true OR with a smaller sample size of only 50.

5 Discussion

The above examples have demonstrated that ESM has the potential to be considerable when the power of a study is low. From a practical perspective, these simulation results demonstrate that ESM should be of interest to those evaluating statistically significant results from low-powered studies and that any large effect sizes observed from such studies should be interpreted cautiously.

One question the reader may ask is how these estimated e-magnification intervals differ from or relate to the typical confidence intervals around point estimates that populate much of the literature. In addition, the reader may ask what advantages there are to considering both the (classic) 95% confidence interval around the effect size typically reported in any literature study and any estimated effect-size magnification interval as derived through `emagnification`. Further, the reader may wonder about the extent to which the (classic) confidence interval and the ESM interval are expected to be similar and how these two intervals should best be interpreted by the practitioner with respect to specific study results. Finally, the reader may be interested in determining what assumed true effect sizes should be used in the `emagnification` simulations and how these should be selected.

First, the (classic) confidence interval around an estimated effect size (such as a mean or mean difference, or an OR, a rate ratio, or a hazard ratio) is an interval that is expected to contain the true parameter (or effect size) over an infinite number of repetitions of the study with a frequency no less than the confidence level, if the underlying statistical model is correct and there is no bias (Rothman, Greenland, and Lash 2008). This confidence interval can be interpreted as a plausible range of estimated ORs if the observational study were repeated many times in the exact same way and if all differences in results in those study replications could be ascribed entirely to the random nature of the Bernoulli and binomial data-generation process in the case of an OR or a rate ratio.

What `emagnification` does is address a separate question from that addressed by the classic confidence interval: if the study were repeated thousands of times and all errors were random, what is a plausible range for other statistically significant results given the size or power of the actual study, the exposed proportion among the control reference group, and any true effect size? This discrete question was originally suggested by Ioannidis and later more directly for the case of continuous data by Gelman and Carlin (2014): "What would be expected to happen if the study were repeated many times given a range of user-assumed (true) effect sizes if one were interested in and focused on only statistically significant results?" This may be particularly useful to the

user who is studying expected small effects using noisy measurements with small sample sizes because it is this user that is most likely to experience (or be bitten by) ESM. This is partly because the low power of the study to detect small effect sizes leads to unstable p-values.[7]

Consider as an example the last row of table 1 simulated here with Stata:

```
. emagnification proportion, p0(0.30) or(1.25) n0(50) n1(50) pctile(25 50 75)
> ifactor(50) nsim(1000) level(0.05) onesided seed(123) log
Scenario 1: p0 = .3, or = 1.25, n0 = 50, n1 = 50
Completed: 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
The tests are one-sided with level = .05
```

p0	p1	true_or	n0	n1	valid	power	p25	p50	p75
.3	.3488372	1.25	50	50	1000	.146	2.173	2.364	2.786

if_p50
1.891

The `emagnification` command allows one to say that even if the true OR were only 1.25, the median estimated OR for statistically significant results is 2.364, and half the statistically significant results if the study were repeated many times would be between 2.173 (`p25`) and 2.786 (`p75`), considerably greater than the true OR of 1.25, which is unknown to the researcher. What does this mean for the data user? The data user should understand that the study might not be adequately powered (here power is 14.6% to detect a true OR of 1.25), and a true OR of 1.25 can be expected to produce a median "discovered" statistically significant effect size of 2.364. In short, consideration of ESM allows the user to view the data from a different perspective from that of the classic 95% confidence interval, one that allows "what if" scenarios to be played out to examine the effects that low-powered studies with imprecise effect sizes might have on statistically significant results.

Some readers may question these ESM calculations that focus on and emphasize the power of a study and consider them to be simply a variant of (discredited) post hoc power calculations. They are not.[8] Instead, ESM calculations can be considered calculations related to the "design calculations" or "post-data design analysis" advocated

7. See, for example, Geoff Cumming's the "Dance of the p values" video at https: // youtu.be / ez4DgdurRPg, which illustrates how the p-value—particularly for low-powered studies—can be very imprecise.

8. The `emagnification` calculations are not post hoc power calculations, because they do not use the effect-size estimates estimated by the study but instead estimate the observed effect-size distributions for statistically significant observed effects found assuming different, user-selected potential effect sizes.

by Gelman and Carlin (2014)[9,10] and discussed further and more recently in greater mathematical detail in Lu, Qiu, and Deng (2019). Gelman and Carlin (2014) advocate using power calculations—reemphasized and named "design calculations" to focus on errors in magnitude and sign instead of declarations of statistical significance—after the data have been collected to help inform a statistical data summary.[11,12] Although Gelman and Carlin (2014) focus on continuous outcomes rather than the categorical and contingency table outcomes focused on here with `emagnification`, they state that such design calculations are intended to address the relevant post-data collection question of not "What is the power of a test?" but instead the more relevant post hoc question of "What might be expected to happen in studies of this size?" This is what was done here with `emagnification`: a variety of plausible ORs were selected to cover a broad range of plausible underlying true effect sizes, and the question "What might be expected to happen in studies of this size if the researcher focuses on discovered, statistically significant effect sizes?" was addressed. The Stata ESM calculations and simulations discussed can be considered "sister" calculations specific for categorical data to the post hoc design calculations for continuous data advocated by Gelman and Carlin (2014) because they derive from the same principles and address the same issues.

Finally, it is important when conducting simulations to use realistic hypothesized true effect sizes based on information that is external to the study under review. Specifically, Gelman and Carlin (2014) indicate that ranges of plausible effect sizes can be developed from auxiliary data, from direct literature, from meta-analyses derived from a systematic review, or from general subject matter expertise or knowledge. However,

9. The article includes an R program on design calculations for experiments whose outcomes are continuous and that would be more typical in research in psychology (for example, effect sizes measured as standardized mean differences) than epidemiology (for example, effect sizes measured as odds, rate, or hazard ratios). Although the authors discuss the program in terms of the design calculations they advocate, the underlying concepts and their implications are the same as applied here with `emagnification`.
10. The R code published as part of the Gelman and Carlin (2014) article has been recently translated to Stata in the community-contributed command `rdesigni`, written and recently updated by Klein (2017) at the University of Kassel and available for download from the Statistical Software Components archive. This command implements the design analysis approach discussed in Gelman and Carlin (2014) and—as is true for that article—approaches the issue from a design analysis and calculation perspective for continuous outcome data that is not necessarily easily adapted to the odds and rate ratios considered more typical in epidemiology and discussed here in the context of `emagnification`. A more recent community-contributed command that is similar to Klein's (2017) `rdesigni` is `retrodesign`, written by Linden (2019) and also available for download from the Statistical Software Components archive. Like Klein's (2017) `rdesigni`, `retrodesign` also is specific for continuous outcomes rather than the categorical ones considered here.
11. See "Yes, it makes sense to do design analysis (power calculations) after the data have been collected" at http: // andrewgelman.com / 2017 / 03 / 03 / yes-makes-sense-design-analysis-power-calculations-data-collected / .
12. Gelman and Carlin (2014) specifically recommend that such ESM-like post hoc design calculations be done when strong statistically significant evidence for nonnull effects has been found because "a [discovered statistically] significant result is often surprisingly likely to be in the wrong direction and to greatly overestimate an effect when researchers study small effects using noisy measurements and small sample sizes." They state that such post hoc design calculations may be even more relevant for findings that are found to be statistically significant because the interpretation of a statistically significant result can change quite substantially depending on the researcher's belief in a plausible size of the underlying effect.

they acknowledge that some fields may have very unclear effect sizes and that in other cases, the investigational area may be brand new or novel, and estimates of effect size may not be readily available; in these cases, they recommend that researchers consider a broad range of possible effect sizes and perform calculations such as those illustrated here with `emagnification` as a form of sensitivity analysis. With respect to specific values that might be used for true effect sizes in the simulations, Ioannidis (2008) says in most fields reasonable guesses about the effect sizes can be made; he suggests, for example, that for genetic associations of common variants with common disease, effect sizes tend to be small or even very small (most ORs $= 1.1 - 1.4$; a few $1.4 - 2$). Similarly, he says relative risk decreases of up to 10–30% are the best that are seen or hoped for with most medical interventions with hard clinical outcomes. In environmental epidemiology, ORs or risk ratios of 2 or 3 are considered by some to be the border between small and larger effects; effect sizes of 1.1 or 1.2 can be difficult to distinguish from real null effects and are sometimes considered "null", "near null", or "very small". Taubes (1995) and Wynder (1996) discuss, characterize, and give examples of odds or rate ratios that have been observed in the past in environmental epidemiology studies. Other authors provide additional general thoughts on these matters (Monson 1990; Rosenthal 1996; Chen, Cohen, and Chen 2010; Grimes and Schulz 2012; Olivier, May, and Bell 2017). What is important in ESM calculations and has been emphasized by both Ioannidis (2008) and Gelman and Carlin (2014) is that a range of plausible true effect sizes be estimated and applied and that these plausible true effect sizes be generated external to the study of interest. The bottom line is that true effect sizes will never be known, so in most cases it is likely best to regard selection of these values as a kind of sensitivity analysis. That is why we chose here to use effect sizes of 1.2, 1.5, 2.0, and 3.0 for illustrative purposes.

While we emphasize that low-powered studies tend to produce greater degrees of ESM in results that are found to be statistically significant (or pass other threshold criteria) than higher-powered studies in the context of ORs or rate ratios typically found in epidemiology, we note that the ESM phenomenon is a principle applicable to discovery science in general and is not a specific affliction or malady of epidemiology (Ioannidis 2005, 2008; Yarkoni 2009; Lehrer 2010; Button et al. 2013; Button 2013; Reinhart 2015); hence, it is applicable to any science in which studies tend to be underpowered and emphasize the use of p-values to "discover" an effect. It is often seen in studies in pharmacology, in gene studies, in psychological studies, and in oft-cited medical literature. In short, any discovered associations from an underpowered study that are highlighted or focused upon on the basis of passing a statistical or other similar threshold will be systematically biased away from the null. The potential degree of this inflation or bias away from the null will depend on a number of issues, including the background rate of the outcome of interest, the sample size of the study, and the effect size of interest. It follows that low-powered epidemiological studies investigating small or weak effects in populations that have a low background rate of the (health) outcome of interest will tend toward the greatest degree of ESM. Note that this is an issue related to how studies are interpreted by users and not one that is intrinsic to or the fault of the study design; nor is it an issue related to good scientific principles or practices.

6 Summary and conclusion

In sum, the ESM phenomenon is real, is important for more appropriately interpreting underpowered studies, and in many ways is underrecognized and underappreciated in the research community and among regulators and decision makers. The phenomenon is not specific to epidemiology and is applicable to any science in which studies tend to be underpowered and emphasize the use of p-values to "discover" an effect, and it is important that users of statistical study results recognize this issue and its potential interpretational consequences. The new `emagnification` command introduced here is a tool that permits reported statistically significant effect-size estimates from possibly underpowered epidemiological studies to be better evaluated and judged. Thus, it can assist individuals reviewing such studies to put an observed statistically significant effect size into a fuller context that allows better judgments regarding adequacy of sample size vis-à-vis the observed effect size. In doing so, users will gain a better understanding of power and sample size issues and in interpreting their potential implications with respect to study conclusions.

7 Additional notes

Some material presented here was originally generated by two of the authors who served in various capacities on an EFSA panel on PPR that, in turn, followed up on findings of the external scientific report "Literature review of epidemiological studies linking exposure to pesticides and health effects" (Ntzani et al. 2013) (University of Ioannina Medical School, 2013) (EFSA-Q-2014-00481). As part of their work on the PPR, the authors contributed to the review and writing of "Scientific opinion of the PPR panel on the follow-up of the findings of the external scientific report 'Literature review of epidemiological studies linking exposure to pesticides and health effects'" and its Annex D where much of this material originally appeared. The PPR panel report is published in the *EFSA Journal* (EFSA Panel on Plant Protection Products and their Residues [PPR] 2017), an official publication of EFSA. This *Stata Journal* article introduces the new command `emagnification` and is the result of an expansion and extension of the original EFSA PPR panel work as part of a post-PPR panel collaboration by the authors.

The analysis described in this article has been reviewed by the U.S. Environmental Protection Agency's Office of Chemical Safety and Pollution Prevention and approved for publication. Approval does not signify that the contents necessarily reflect the views, policies, or determinations of the Agency, nor does the mention of trade names of commercial products constitute endorsement or recommendation for use.

8 Programs and supplemental materials

To install a snapshot of the corresponding software files as they existed at the time of publication of this article, type

```
. net sj 20-3
. net install st0608     (to install program files, if available)
. net get st0608         (to install ancillary files, if available)
```

9 References

Button, K. 2013. Unreliable neuroscience? Why power matters. https://www.theguardian.com/science/sifting-the-evidence/2013/apr/10/unreliable-neuroscience-power-matters.

Button, K. S., J. P. A. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. J. Robinson, and M. R. Munafò. 2013. Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14: 365–376. https://doi.org/10.1038/nrn3475.

Chen, H., P. Cohen, and S. Chen. 2010. How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistics—Simulation and Computation* 39: 860–864. https://doi.org/10.1080/03610911003650383.

EFSA Panel on Plant Protection Products and their Residues (PPR), C. Ockleford, P. Adriaanse, P. Berny, T. Brock, S. Duquesne, S. Grilli, S. Hougaard, M. Klein, T. Kuhl, R. Laskowski, K. Machera, O. Pelkonen, S. Pieper, R. Smith, M. Stemmer, I. Sundh, I. Teodorovic, A. Tiktak, C. J. Topping, G. Wolterink, M. Bottai, T. Halldorsson, P. Hamey, M.-O. Rambourg, I. Tzoulaki, D. C. Marques, F. Crivellente, H. Deluyker, and A. F. Hernandez-Jerez. 2017. Scientific opinion of the PPR panel on the follow-up of the findings of the external scientific report 'Literature review of epidemiological studies linking exposure to pesticides and health effects'. *EFSA Journal* 15: e05007. https://doi.org/10.2903/j.efsa.2017.5007.

Gelman, A., and J. Carlin. 2014. Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science* 9: 641–651. https://doi.org/10.1177/1745691614551642.

Grimes, D. A., and K. F. Schulz. 2012. False alarms and pseudo-epidemics: The limitations of observational epidemiology. *Obstetrics & Gynecology* 120: 920–927. https://doi.org/10.1097/AOG.0b013e31826af61a.

Ioannidis, J. P. A. 2005. Why most published research findings are false. *PLOS Medicine* 2: e124. https://doi.org/10.1371/journal.pmed.0020124.

———. 2008. Why most discovered true associations are inflated. *Epidemiology* 19: 640–648. https://doi.org/10.1097/EDE.0b013e31818131e7.

Klein, D. 2017. rdesigni: Stata module to perform design analysis. Statistical Software Components S458423, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s458423.html.

Lehrer, J. 2010. The truth wears off: Is there something wrong with the scientific method? https://www.newyorker.com/magazine/2010/12/13/the-truth-wears-off.

Linden, A. 2019. retrodesign: Stata module to compute type-S (sign) and type-M (magnitude) errors. Statistical Software Components S458631, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s458631.html.

Lu, J., Y. Qiu, and A. Deng. 2019. A note on type S/M errors in hypothesis testing. *British Journal of Mathematical and Statistical Psychology* 72: 1–17. https://doi.org/10.1111/bmsp.12132.

Monson, R. R. 1990. *Occupational Epidemiology*. 2nd ed. Boca Raton, FL: CRC Press.

Ntzani, E. E., M. Chondrogiorgi, G. Ntritsos, E. Evangelou, and I. Tzoulaki. 2013. Literature review on epidemiological studies linking exposure to pesticides and health effects. *EFSA Supporting Publications* 10(10). https://doi.org/10.2903/sp.efsa.2013.EN-497.

Olivier, J., W. L. May, and M. L. Bell. 2017. Relative effect sizes for measures of risk. *Communications in Statistics—Theory and Methods* 46: 6774–6781. https://doi.org/10.1080/03610926.2015.1134575.

Reinhart, A. 2015. *Statistics Done Wrong: The Woefully Complete Guide*. San Francisco: No Starch Press.

Rosenthal, J. A. 1996. Qualitative descriptors of strength of association and effect size. *Journal of Social Service Research* 21: 37–59. https://doi.org/10.1300/J079v21n04_02.

Rothman, K. J., S. Greenland, and T. L. Lash. 2008. *Modern Epidemiology*. 3rd ed. Philadelphia: Lippincott Williams & Wilkins.

Taubes, G. 1995. Epidemiology faces its limits. *Science* 269: 164–169. https://doi.org/10.1126/science.7618077.

Wynder, E. L. 1996. Invited commentary: Response to science article, "Epidemiology faces its limits". *American Journal of Epidemiology* 143: 747–748. https://doi.org/10.1093/oxfordjournals.aje.a008811.

Yarkoni, T. 2009. Ioannidis on effect size inflation, with guest appearance by Bozo the Clown. https://www.talyarkoni.org/blog/2009/11/21/ioannidis-on-effect-size-inflation-with-guest-appearance-by-bozo-the-clown/.

About the authors

David J. Miller is formerly a senior statistician and currently a supervisory chemist and branch chief in the Health Effects Division of the U.S. Environmental Protection Agency's Office of Pesticide Programs in Washington, DC.

James T. Nguyen is a mathematical statistician in the Health Effects Division of the U.S. Environmental Protection Agency's Office of Pesticide Programs in Washington, DC.

Matteo Bottai is a professor of biostatistics in the Division of Biostatistics at the Institute of Environmental Medicine at the Karolinska Institutet in Stockholm, Sweden.