Revisiting Statistics and Evidence-Based Medicine: On the Fallacy of the Effect Size Based on Correlation and Misconception of Contingency Tables

Sergey Roussakow, MD, PhD (0000-0002-2548-895X)

85 Great Portland Street

London, W1W 7LT, United Kingdome

roussakow@neogalen.org

+44 20 3885 0302

Affiliation: Galenic Researches International LLP

85 Great Portland Street

London, W1W 7LT, United Kingdome

Word count: 3,190.

ABSTRACT

Evidence-based medicine (EBM) is in crisis, in part due to bad methods, which are understood as misuse of statistics that is considered correct in itself. This article exposes two related common misconceptions in statistics, the effect size (ES) based on correlation (CBES) and a misconception of contingency tables (MCT). CBES is a fallacy based on misunderstanding of correlation and ES and confusion with 2 × 2 tables, which makes no distinction between gross crosstabs (GCTs) and contingency tables (CTs). This leads to misapplication of Pearson's Phi, designed for CTs, to GCTs and confusion of the resulting gross Pearson Phi, or mean-square effect half-size, with the implied Pearson mean square contingency coefficient. Generalizing this binary fallacy to continuous data and the correlation in general (Pearson's r) resulted in flawed equations directly expressing ES in terms of the correlation coefficient, which is impossible without including covariance, so these equations and the whole CBES concept are fundamentally wrong. MCT is a series of related misconceptions due to confusion with 2 × 2 tables and misapplication of related statistics. The misconceptions are threatening because most of the findings from contingency tables, including CBES-based meta-analyses, can be misleading. Problems arising from these fallacies are discussed and the necessary changes to the corpus of statistics are proposed resolving the problem of correlation and ES in paired binary data. Since exposing these fallacies casts doubt on the reliability of the statistical foundations of EBM in general, we urgently need to revise them.

KEW WORDS (6)

Effect size, correlation coefficient, mean square contingency coefficient, Pearson's Phi, 2 × 2 table, contingency table.

KEW WORDS EXTENDED

Effect size, correlation coefficient, association measure, covariance, mean square contingency coefficient, mean square effect half-size, Pearson's Phi, 2 × 2 table, binary crosstab, gross crosstab, contingency table.

KEY FINDINGS

The study:

- exposes for the first time two related common misconceptions in statistics, the fallacy of effect size based on correlation and misconception of contingency tables;
- shows that the misconceptions are threatening and most of the contingency tables findings, including meta-analyses based on correlation, can be misleading;
- resolves the problems arising from these fallacies and proposes the necessary changes to the corpus of statistics;
- clarifies existing and introduces new statistical definitions, including for 2 × 2 tables, creating the basis for further development;
- questions the reliability of the statistical foundations of EBM and for the first time states the need to revise them.

Revisiting Statistics and Evidence-Based Medicine: On the Fallacy of the Effect Size Based on Correlation and Misconception of Contingency Tables

INTRODUCTION

Evidence-based medicine (EBM) is "one of our greatest human creations,"[1, 2] but there is a growing awareness that it is undergoing a crisis,[1, 3, 4, 5, 6] in part due to "bad methods."[1] However, the idea of bad methods comes down to misusing statistics,[1] that is, misusing methods that are correct in themselves.[6] Therefore, believing in the reliability of the statistical foundations is the cornerstone of EBM. Unfortunately, there is cause for concern. This article exposes two common misconceptions in statistics, a fallacy of the effect size based on correlation (CBES), which has been around for over 70 years and remains unnoticed, and a related misconception of contingency tables (MCT).

The concept of CBES is included in all statistical and meta-analysis manuals and is widely used, especially in psychometrics.[7, 8, 9, 10, 11] The basic equation[10]

[1] $r^2 = \frac{d^2}{d^2+\frac{(n_1+n_0-2)(n_1+n_0)}{n_1 n_0}}$, where d is the effect size (ES) known as Cohen's d, r is the coefficient of bivariate correlation commonly known as Pearson's (product-moment) coefficient of correlation,[12]

given equal groups $(n_0 = n_1 = n)$, reduces to

[2] $r^2 = \frac{d^2}{d^2+\frac{4(n-1)}{n}} \cong \frac{d^2}{d^2+4}$,

so

[3] $r \cong \frac{d}{\sqrt{d^2+4}}$;[13]

this is the basic formula used by Cohen.[14] The corresponding equation for the dependence of d from r is

[4] $d = \frac{2r}{\sqrt{1-r^2}}$.[9, 14]

The CBES was the weakest place in the Cohen's effect size concept, since the large ES ($d = 0.8$) corresponded to $r = 0.371$, which is a weak to moderate correlation according to Pearson. To get around this problem, Cohen had to introduce a "biserial" estimate connected

to the raw "point" estimate with a correction factor of 1.253,[15] but even the adjusted "large" r was only 0.465 and still was lower than the Pearson's strong correlation limit of 0.5 (and much lower than the modern limits of 0.7 or 0.6[16]). Motivated by this discrepancy, I investigated ES versus correlation to identify the cause of the discrepancy.

METHODS

A simple model shown in Table 1 was used to analyze ES versus correlation.

Table 1. An example of analytical model for effect size versus correlation [$X \sim N(100, 20)$, $Y \sim N(120, 25)$].

		Group			Group	
Subject		X	Y	Statistic	X	Y
i	1	78.96	118.34	n	10	10
	2	60.71	142.63	m	82.56	101.02
	3	106.92	67.13	s	19.20	30.71
	4	74.67	116.99	s_{XY}	-116.282	
	5	81.37	87.09	r	-0.197	
	6	98.49	96.04	Δm	18.454	
	7	109.00	74.36	S_p	27.788	
	8	61.96	104.82	d	0.664	
	9	95.84	146.72	d_r	-0.402	
	10	57.73	56.07			

Note: n, sample size; m, sample mean; s, sample standard deviation; s_{XY}, sample covariance; r – coefficient of correlation (Pearson's r); Δm, effect magnitude; Sp, pooled standard deviation; d, Cohen's effect size; d_r, correlation-based effect size.

It included two samples X and Y of ten subjects each representing two normally distributed populations $\bar{X}$ and $\bar{Y} \sim N(\mu, \sigma)$, where $\bar{X} \sim N(100, 20)$, and $\bar{Y} \sim N(120, 25)$. The samples were randomized using the Excel NORM.INV and RAND functions:

[5] $X = NORM.INV(Pr, \mu, \sigma)$, where μ is the population mean, σ is the population standard deviation, Pr is the random probability density function of $0 \leq Pr \leq 1$ given by Excel RAND function.

The calculated parameters included sample size (n), sample arithmetic mean (m), sample standard deviation (s), effect magnitude (Δm), sample covariance (s_{XY}), Pearson's correlation coefficient (r), pooled standard deviation (S_p)

$$[6] \qquad S_p = \sqrt{\frac{s_X^2 + s_Y^2 - 2rs_x s_y}{2}},$$

actual effect size (Cohen's d)

$$[7] \qquad d = \frac{\Delta \mathrm{m}}{S_p},$$

and the CBES (d_r, equation [4]). A Monte Carlo simulation of 10,000 iterations was performed. Then scatter charts were plotted and the coefficients of linear correlation (R^2) calculated for the CBES versus the actual effect size (AES) (d_r vs. d), covariance versus correlation (s_{XY} vs. r), AES versus Correlation (d vs. r) and covariance versus AES (s_{XY} vs. d).

A working model in MS Excel is available in the Supplement.

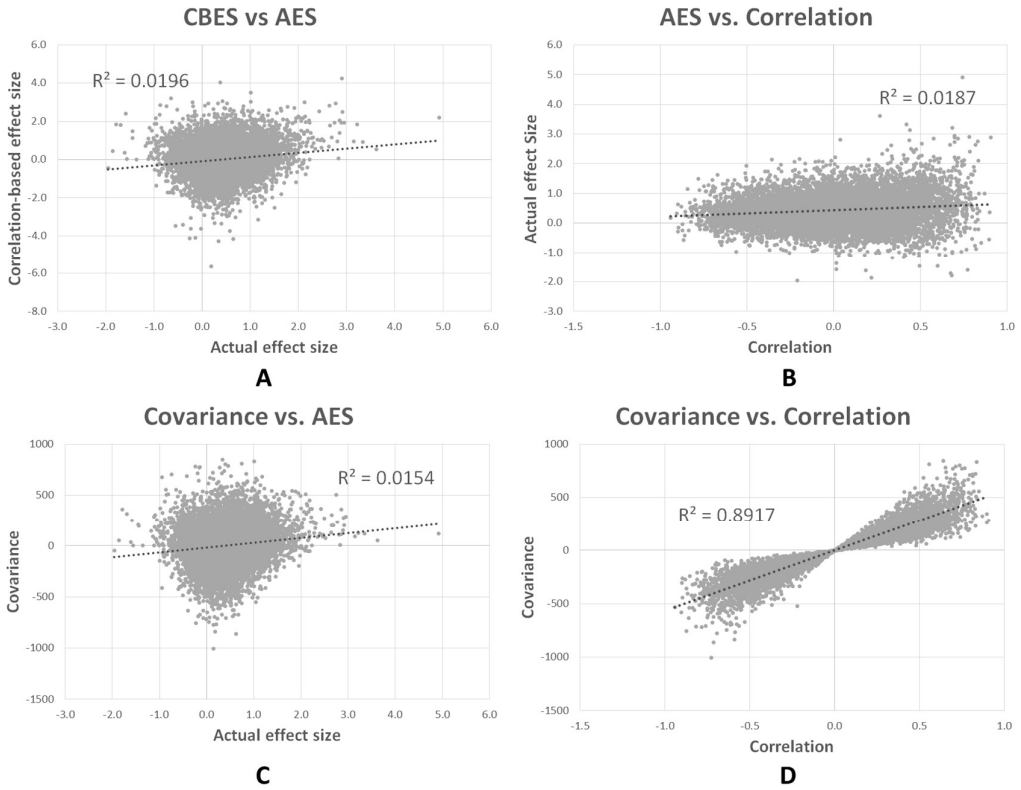

Figure 1. *Effect Size vs. Correlation vs. Covariance (Monte Carlo simulation, 10,000 iterations): A, correlation-based effect size (CBES) versus the actual effect size (AES); B, AES versus correlation; C, covariance versus AES; D, covariance versus correlation.*

RESULTS

In conflict with equations [2]–[4] implying functional relationship between CBES and AES, the simulation showed an extremely weak positive correlation between them (Figure 1A), as well as between AES and correlation (Figure 1B) and AES and covariance (Figure 1C) ($0 < R^2 < 0.02$ in all cases). The result did not depend on the parameters of the samples (μ, σ). Note the striking difference between AES and CBES (Table 1). Thus, CBES and correlation are not related to the AES, suggesting that equations [2]–[4] and the whole CBES concept are flawed.

DISCUSSION

Although after Cohen's milestone monograph,[7] the relationship between correlation and ES seems apparent and even trivial, it is actually a logical fallacy stemming from the trivial notion that, since they are related to between-group differences, they are interrelated and therefore mutually convertible.

Table 2. *Fundamental differences between effect size and correlation.*

	Effect size	**Correlation**
Characteristic	Mean difference normalized to variance	Covariance cleared of variance
Essence	Signal-to-noise ratio	Relationship
Magnitude of the mean difference	Matters	Does not matter
Concordance of specific differences	Does not matter	Matters
Samples	Any (paired and unpaired)	Paired

As the mean difference normalized to variance (Table 2), ES is a kind of signal-to-noise ratio that characterizes the magnitude of the mean difference, regardless of the concordance of specific (paired) differences, and therefore applies to any sample, paired or unpaired. Correlation, which is covariance cleared of variance,

$$[8] \qquad r = \frac{s_{XY}}{s_X s_Y} = \frac{\sum_{i=1}^{n}\left((x_i - m_X)(y_i - m_Y)\right)}{\sqrt{\sum_{i=1}^{n}(x_i - m_x)^2 \sum_{i=1}^{n}\left(y_i - m_y\right)^2}},$$

is a measure of relationship (causation or dependence) that characterizes the concordance of specific (paired) differences, regardless of their magnitude, and therefore applies only to paired samples. Thus, these are fundamentally different parameters that in principle cannot be reduced to each other, which can be proved mathematically.

Correlation coefficient r (equation [8]), given equal variances, comes down to the equation

[9] $$r = \frac{s_{XY}}{s^2},$$

so the ES, following equation [7], is

[10] $$\delta = \Delta\mu\sqrt{\frac{\mathrm{r}}{s_{XY}}}$$

for equal variances and

[11] $$d = \Delta\mathrm{m}\sqrt{\frac{2}{s_X^2+s_Y^2-2s_{XY}}}$$

for unequal variances. Thus, ES cannot be expressed in terms of correlation or covariance, but only in terms of both, so equations [2]–[4] and the CBES concept in general are fundamentally wrong.

Variance is a pure measure of sample variability, correlation is a pure measure of the association (concordance of changes) of samples, and covariance is a complex parameter that combines variability and association. As seen from Table 3, correlation has nothing to do with variance, so the variance-based ES has nothing to do with the correlation (Figure 1B). ES is also not related to covariance (Figure 1C), while correlation and covariance are strongly correlated (e.g., $R^2 \approx 0.9$ in Figure 1D) since both depend on association.

Table 3. *Properties of variance, covariance, and correlation.*

	Variance	**Covariance**	**Correlation**
Variability	Yes	Yes	No
Association	No	Yes	Yes
Direction	No	Yes	Yes

A visual representation of the CBES fallacy is given in Figure 2 (Table S1 in the Supplement) (the lines visually show the association (concordance of changes) between the samples).

Contrary to what was expected according to the CBES, it shows opposite correlations with the same ES (options 1–2), opposite ESs with the same correlation (options 3–6), and possibility of any combinations of ESs and correlations (options 7–8). This is an approximate model that uses the unpaired ES because the paired one gives error if r=1, but it shows the principle well.

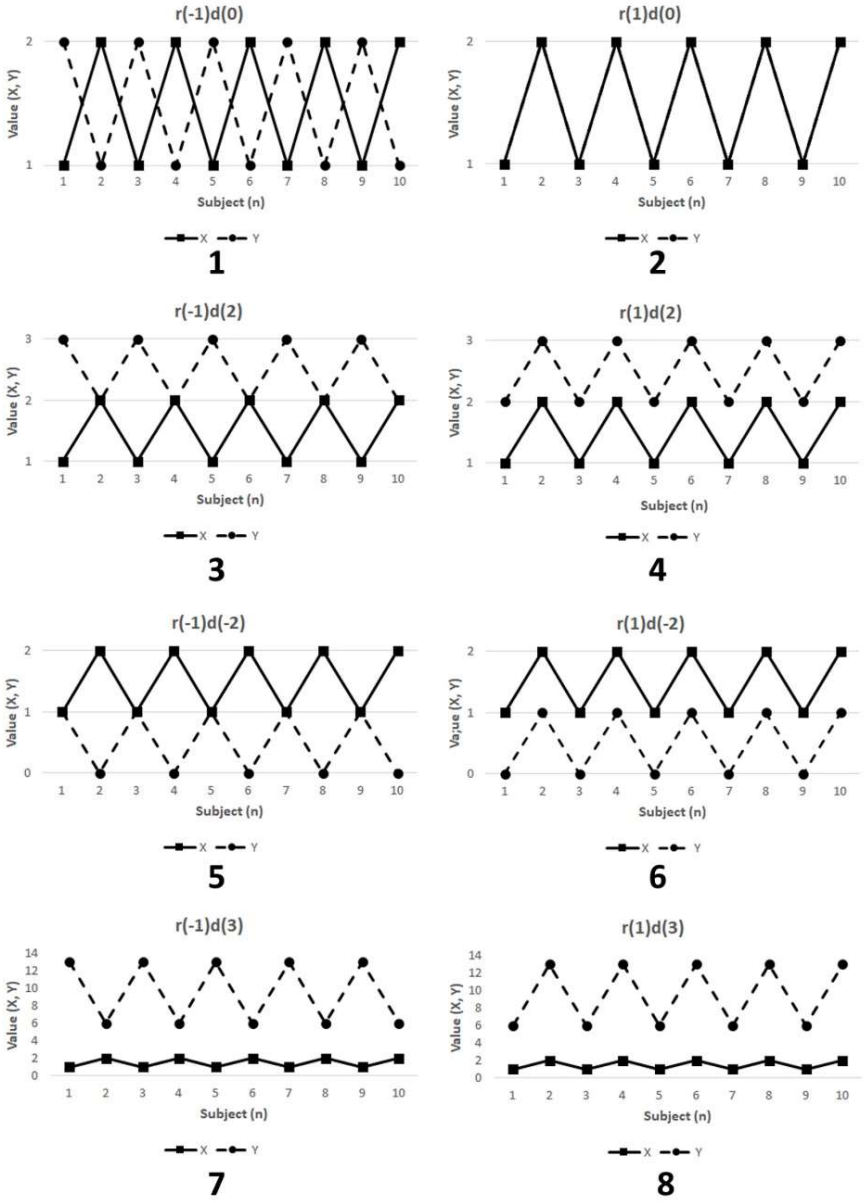

Figure 2. *Visual demonstration of the CBES misconception.*

Left column – discordant cases (r=-1); right column – concordant cases (r=1).

Another reason why the inconvertibility of correlation and ES remains obscured is rooted in the misconception of contingency tables (MST). Currently, any crosstab is considered contingency table by default,[17, 18] resulting in the severe fallacy shown in Figure 3.

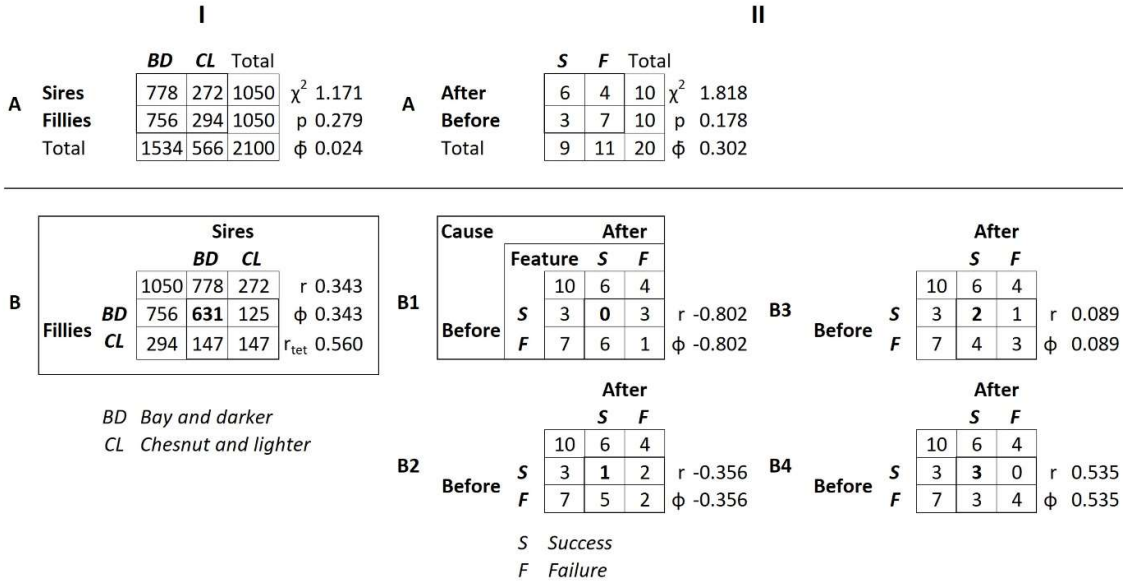

Figure 3. *Example of 2 × 2 tables: I, Pearson's example;[19] II, simulated example; A, gross (parent, master) crosstab; B, contingency (child) tables.* r*, Pearson's* r*;* φ*, Pearson's* φ *;* $\bar{\varphi}$*, gross Pearson's Phi;* χ^2*, Pearson's chi-square statistic; p, p-value.*

Table IIA is a 2 × 2 matrix (crosstab) of two binary variables, intervention (Before–After) and outcome (Success–Failure). It is a gross crosstab that reports that 3 out of 10 subjects were a success before intervention and 6 out of 10 after the intervention. This table gives information on the effect but does not contain the information on the association of the variables. To display the contingency information, it must be converted to a contingency table IIB, where the cells of the gross (parent, master) table become the marginal statistics of the contingency (child) table, and a new 2 × 2 matrix appears that relates to the contingency (association). For example, in the table IIB2, of three before-after pairs who were a success before intervention, one remained a success after the intervention (SS pair) and the other two changed to a failure (SF). Likewise, of seven failure pairs, two remained a failure (FF), and five changed to a success (FS). Therefore, 3 pairs have kept their outcomes, while 7 changed them.

For this table, we can calculate the Pearson mean square contingency coefficient φ (Pearson's Phi)[19]

[12] $\varphi = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$, where $\frac{a \mid b}{c \mid d}$ are the corresponding cells of the 2 × 2 matrix,

as well as any other measures of association,[20, 21, 22, 23, 24, 25] to reveal a negative association (ϕ=-0.356) between the variables in terms of the outcomes (i.e., the groups change their outcomes discordantly).

Thus, we come to new definitions of crosstabs:

- Categorical crosstab is an n × m matrix that displays the mutual frequency distribution of two categorical variables having n and m categories, respectively.
- Binary crosstab (BCT) is a 2 × 2 matrix that displays the mutual frequency distribution of two binary variables.
- Gross crosstab (GCT) is a BCT that displays the mutual frequency distribution of two different binary variables, one of which is paired (i.e., the options are interrelated).
- Contingency table (CT) is a BCT that displays the frequency distribution of the feature pairs against the featured paired cause, so that the marginal statistics of CT match the cells of the parent GCT for these paired binary variables (Figure 3-B1).

The differences between BCT, GCT and CT are summarized in Table 4.

Table 4. *Differences between binary crosstabs, gross crosstabs and contingency tables.*

	Binary crosstab	**Gross crosstab**	**Contingency table**
Number of levels	1		2
Unit	Subject		Pair
Sample	Unpaired	Paired	
Mutual relationship	Unrelated	Master table	Child table
Reduces to contingency tables	Not applicable	Yes	No
Restores to the gross table	Not applicable	No	Yes
Significance of differences	Unpaired tests		Paired tests
Association measures	Inapplicable		Applicable

CT has a double-decker design (Figure 3-IIB1),[26] where the first level is formed by the feature binary variable, and the second by the causal binary variable, and counts pairs. GCT and BCT

are single-decker (Figure 3-IIA) and count subjects. Each parent GBC can be reduced to one or several child CTs with different association, where the number of CTs is equal to the minimum GCT cell value plus one. For example, in the example II (Figure 3), four association options are available with the SS values of 0, 1, 2 and 3 (B1-4). Accordingly, any CT can be restored to the parent GCT (Figure 3-I). BCT refers to unpaired samples, so it cannot be converted. Unpaired significance tests (Pearson's chi-square, Fisher's exact test, etc.) are used with BCTs and GCTs, but give wrong results when applied to CTs, so the CTs-based significance testing requires paired tests (e.g., McNemar's tests). Thus, CTs, GSTs and BCTs are different in nature.

The example of thoroughbred racehorses[19] (Figure 3-IB), which Karl Pearson used when introducing his Phi, is a double-decker CT, from which the parent GCT can be easily restored (Figure 3-IA). When applied to CTs, Pearson's Phi obtains its original meaning of the "mean square contingency coefficient" ϕ, which is another calculation of the Pearson's r for two binary variables (Figure 3-B). The confusion is in the fact that mathematically Pearson's Phi can be applied to any BCT or GCT, but, due to the mentioned difference in nature, this results in a completely different parameter. Pearson's Phi applied to BCT or GCT has nothing to do with Pearson's r and correlation (contingency) at all. This a size-independent derivative of the chi-squared statistic

[13] $\bar{\varphi} = \sqrt{\frac{\chi^2}{N}}$, where N is the sample size,

an effect size-like parameter (Figure 4), which is about half the actual effect size

[14] $\delta \approx 2\bar{\varphi}$,

exactly

[15] $d = \frac{2\bar{\varphi}}{\sqrt{1-\bar{\varphi}^2}}\sqrt{\frac{n-1}{n}}$,[27]

and in large samples matches equation [4], albeit is totally different in nature.

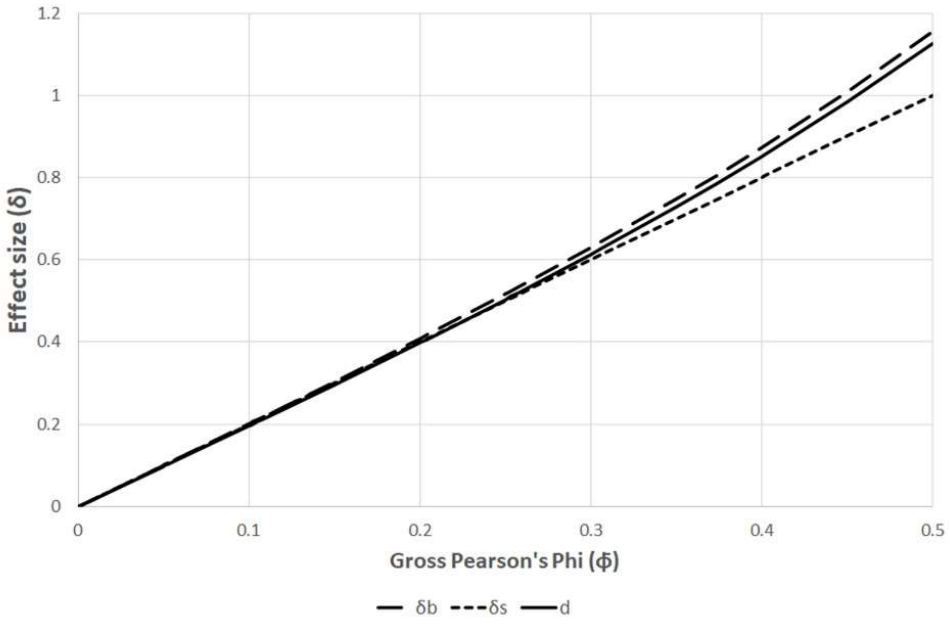

Figure 4. *Binary effect sizes versus the gross Pearson Phi ($\bar{\varphi}$): A Monte Carlo simulation (100 iterations). δ_b, the biased effect size by equation [4]; δ_s, the simplified effect size by equation [14]; d, the binary Cohen's d by equation [15].*

So, the term Pearson's Phi (ϕ) or "mean square contingency coefficient" should only be applied to CTs. The result of applying equations [12] and [13] to GCTs is in fact the "mean-square (or chi-square) effect half-size" and should be denoted as the gross Pearson Phi ($\bar{\varphi}$). Figure 3 shows that $\bar{\varphi}$, which is an effect size parameter (part A), has nothing to do with φ (part B), which is a correlation parameter.

The last source of the CBES fallacy is the equation [6] that establishes a relationship between correlation and S_p. Since S_p determines the ES (equation [7]), which is functionally related to the significance of differences (SOD)

[16] $t = d\sqrt{\frac{n}{2}}$, [10]

it seems logical to conclude that correlation and the ES are mutually related, which is a fallacy, the nature of which is parsed in Figure 5.

As discussed above, ES depends on both variance and covariance (equation [11]). As schematically shown in Figure 5A, the basic ES values (δ_1 and δ_2) depend on the variance and correspond to zero-correlation significances $\chi^2_{1_0}$ and $\chi^2_{2_0}$. With these particular variances, the change in correlation does indeed change the ES and significance.

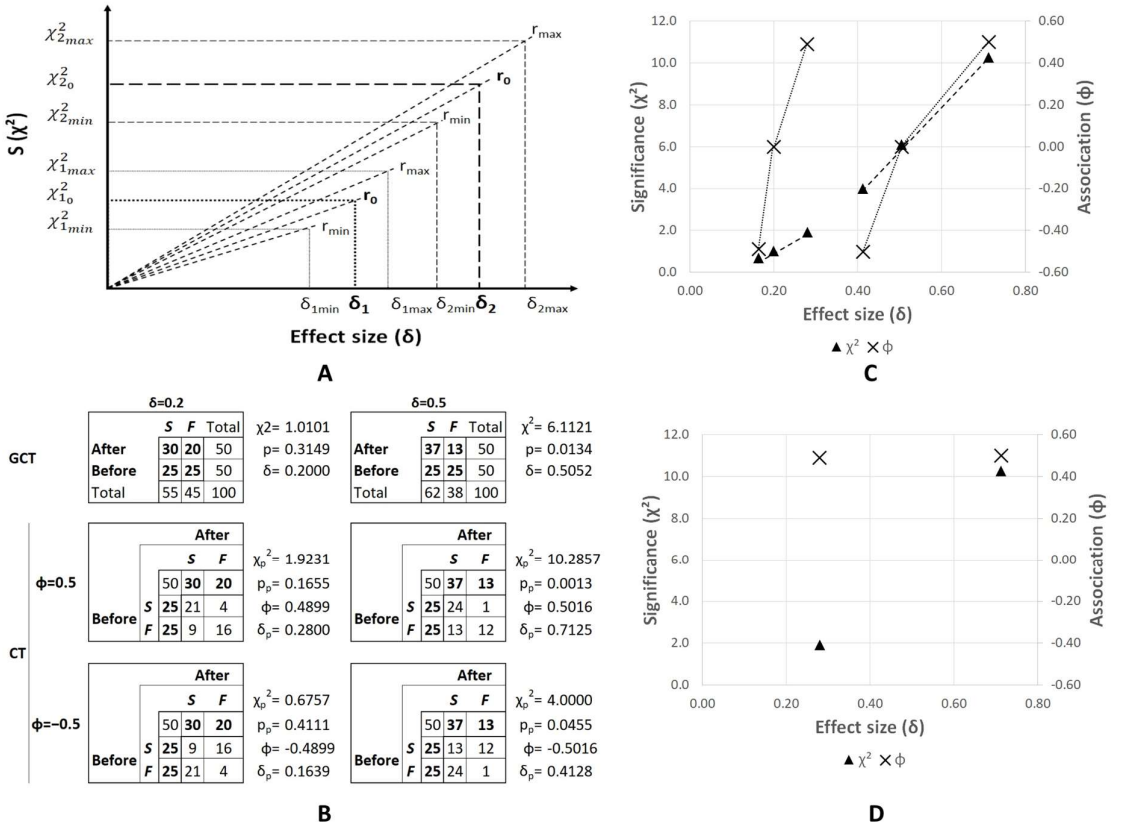

Figure 5. *Explanation of the fallacy of the relationship between correlation, on the one hand, and effect size and significance of differences, on the other: A, schematic representation of the fallacy; B-D, an example: B, original tables; C-D, significance and association vs. AES: C, within-GCT (multiple associations); D, between-GCT (single association per GCT).*

Therefore, the dependence of ES from correlation (equation [6]) is only valid for constant variance, that is, for the set of CTs within each GCT (Figure 5C), which doesn't matter since in real life we are dealing with the only association for each GCT (Figure 5D). An example case is shown in Figure 5B-D. Two equal-sized GCTs with the zero-correlation ESs of 0.2 and 0.5 allow a number of associations (Figure 5B), and within each GCT, the relationship between the correlation and ES / significance is functional (Figure 5C). However, when considering two GCTs, the relationship is blurred, since the same correlations (here, $\varphi \approx 0.5$ and $\varphi \approx -0.5$) correspond to different ESs / significances. In real life, when there is the only association per each GCT, the relationship actually disappears ($0 < R^2 < 0.02$) (Figure 5D, Figure 1B), revealing itself only by the fact that the correlation is always positive (Figure 1). The main

misconception stemming from this fallacy is a widespread tendency to draw conclusions about association based on the SOD and vice versa.

Thus, the CBES fallacy is that ES has nothing to do with correlation, so the concept is fundamentally wrong. However, in the case of binary variables, the misapplication of Pearson's Phi to GCTs results in the mean-square effect half-size. The fallacy is that this parameter is still misleadingly considered a measure of association. Generalization of this binary fallacy to the correlation in general (Pearson's r) and the entire range of binary and continuous data led to the erroneous equations [2]–[4] that, in turn, lead to erroneous meta-analysis,[10, 11] misunderstanding of the nature of correlation,[7] and erroneous conversions and transformations based on the CBES.[28, 29]

The emergence of the CBES fallacy was associated with the introduction of meta-analysis[8, 30] and the idea of effect size[7] in early 80s. The confusion with 2 × 2 tables seems to have started even earlier. At least, Cramer in 1946[31] still correctly applied Pearson's Phi to CTs, while Cohen in 1988[7] was already in the misconception. In fairness, this misconception seems to go back to Karl Pearson himself, who (or someone of his team) boldly calculated φ for all 2 × 2 tables, including GCTs.[32] In a sense, the confusion is due to the term "2 × 2 table," which makes no distinction between GCTs and CTs and facilitates the misuse of Pearson's Phi. Despite the apparent inadequacy, the CBES concept has never been questioned, is included in all guidelines,[9-11] equations [2]–[4] are commonly used for calculating ES and related conversions,[28, 29] and even for the correlation-based definition of the ES,[10] so it is a common misconception. The introduction begins with a flawed equation [1],[10] in which correlation related to paired samples is combined with unequal groups, that is, with unpaired statistics.

The same applies to the MCT, which is a series of related misconceptions. Typically, it looks like treating GCTs, or even unpaired BCTs, as CTs, misleadingly attributing to them the ability to assess the association of the variables, leading to misapplication of association statistics (e. g., applying association statistics to GCTs) and independence statistics (e. g., applying Pearson's chi-square to CTs). Fundamentally, this misconception stems from three fallacies: confusion of effect and association, as discussed above for CBES; misunderstanding of the mutual relationship between GCTs and CTs as parent and child tables, leading to the belief that CTs simply arise in a single specific form;[26] and lack of understanding of the pairwise nature of association (if samples are not paired (i.e., the pairs are not intrinsically bound),

they can be resorted in any order and any association is random (corresponds to a certain random order), therefore meaningless). Finally, Pearson's Phi can be calculated using equation [13] for the gross Pearson Phi,[33] which is non-directional, so it is not a measure of association that requires equation [12].

An example of the consequences of these misconceptions is shown in Figure 6. Section I presents example CTs (Table 1–3) taken from a credible source.[26] Sections II and III include the corresponding GCTs and CTs, respectively, obtained by adjusting the original tables. Only Table 1 was indeed a two-decker CT, obtained by applying two binary variables (opinions on death penalty and gun registration) to the same subject (paired samples). However, the confusion of association and effect led to the misleading conclusion that "P value is 0.0232 … suggests that there is an association … ." In fact, there was no association (ϕ=-0.061, Table 1-III), and the conclusion is a statistical error caused by MCT.

I

Table 1

	Death penalty		
Gun registration	*Favor*	*Oppose*	Total
Favor	784	236	1020
Oppose	311	66	377
Total	1095	302	1397

χ^2= 5.1503
p= 0.0232

Table 2

	Lung cancer		
Smoker	*Yes*	*No*	Total
Yes	647	622	1269
No	2	27	29
Total	649	649	1298

χ^2= 22.044
p= 2.7E-06

Table 3

	Depression improved?		
Treatment	*Yes*	*No*	Total
Pramipexole	8	4	12
Placebo	2	8	10
Total	10	12	22

II

	Favor	*Oppose*	Total
Death penalty	1095	302	1397
Gun registration	1020	377	1397
Total	2115	679	2794

χ^2= 10.944
p= 0.0009
δ= 0.146

	LC	*NLC*	Total
SM	647	622	1269
NSM	2	27	29
Total	649	649	1298

χ^2= 22.044
p= 2.7E-06
δ= 0.263

III

		Death penalty		
			Favor	*Oppose*
		1397	1095	302
Gun regi	*Favor*	1020	784	236
stration	*Oppose*	377	311	66

χ^2_p= 10.283
p_p= 0.0013
ϕ= -0.061

Figure 6. *An example of the misconception of 2 × 2 tables:* [26] *I, original tables; II, corrected gross tables; III, corrected contingency tables. LC, Lung Cancer; NLC, No Lung Cancer; SM, Smoker, NSM, Non-Smoker.*

Other examples are not CTs since count subjects, not pairs. A pseudo-two-decker design of Table 2-I is misleading because when the count is not based on pairs, the Yes–No options, misclassified as "a single binary variable," actually represent different variables "Lung cancer" (Lung cancer – No lung cancer) and "Smoking" (Smoker – Non-Smoker). It is also not a GCT since the samples, albeit equal-sized, are unpaired.[34] Thus, this is a BCT (Table 2-II) that cannot be reduced to CT since any association in the BCT is random, therefore misleading. Finally, Table 3-I is a typical BCT with unequal groups that technically cannot be converted to

CT, so its pseudo-two-decker design is simply anecdotal. All examples misuse the significance-based association inference.

In addition, in the Table 1 example, Pearson's chi-square was misapplied to the CT, so that the reported p-value of 0.0232 (χ^2=5.15) is incorrect, and the actual p-value is 0.0013 (χ^2=10.283). Table 5 presents a more illustrative example of the degree of possible error caused by misapplication of significance tests based on the example in Figure 3II. For example, misapplication of Pearson's chi-square to the CT (B1) would show a significant difference (p=0.011), while the actual difference (McNemar's test) is not significant (p=0.317). The same applies to association: a strong negative association in Table B1 (ϕ=-0.802) would be misjudged as a weak positive association (ϕ=0.302), if based on the GCT (A).

Table 5. *Significance and association error due to misconception of contingency tables (Figure 3II).*

		Pearson's test		**McNemar's test**	
Table	**ϕ**	**χ^2**	**p**	**χ^2**	**p**
A	0.302	1.818	0.178	0.143	0.705
B1	-0.802	6.429	0.011	1.000	0.317
B2	-0.356	1.270	0.260	1.286	0.257
B3	0.089	0.079	0.778	1.800	0.180
B4	0.535	2.857	0.091	3.000	0.083

The example in Figure 6 shows that the MCT is threatening, because much of the findings obtained from CTs can be misleading. The misconception is widespread: the idea of CT is typically explained using GCTs;[11, 26, 35] BCTs are misleadingly referred to as CTs;[17, 26] CTs are often (mostly?) pseudo CTs;[26] and even true CTs are still misused in terms of significance testing[26] and association measure.[33] With all of the above, there seems to be no publication on CTs unaffected by the misconception and therefore not misleading.

Given the above, the following changes should be made to the corpus of statistics:

- The CBES concept should be abolished as misleading.
- New definitions for BCTs, GCTs and CTs should be adopted, the term "2 × 2 table" should be avoided as confusing.

- Pearson's mean square contingency coefficient (φ) should only be applied to CTs; this is an association measure that is not applicable to meta-analysis.
- Pearson's Phi applied to GCTs should be referred to as the gross Pearson Phi or "mean-square effect half-size" ($\bar{\varphi}$); it should not be used for estimating association.
- Equations [2]–[4] and all related equations linking effect size to correlation / association are false and should be invalidated.
- Unpaired significance tests should only apply only to BCTs and GCTs; significance testing for CTs should only use paired tests.
- In no case should the results of significance tests be used to assess the association between variables and vice versa.
- All conclusions and inferences based on the CBES, as well as all conversions and transformations based on CBES, should be invalidated.
- All meta-analyses based on CBES should be revised.
- All findings and conclusions based on CTs should be revised.
- The relevant chapters in statistical and meta-analysis manuals should be revised.

CONCLUSIONS

This article exposes two common misconceptions in statistics, CBES that has been around for over 70 years and remains unnoticed, and MCT, that casts doubt on the reliability of the statistical foundations of EBM in general. If the statistical foundations are corrupted, then the problems of EBM are deeper than it is believed, because they are not limited to the misuse of statistics but extends to the bad statistics itself. However, this can be a problem and a solution at the same time, as many of the EBM problems can actually be caused by incorrect statistics and resolved by fixing these flaws. That is why we urgently need to revise the statistical foundations of EBM. This article completely revisits the correlation and effect size problems in binary data and corrects all their shortcomings.

DATA SHARING

An Excel model described in Methods is available in the Supplement. Extra models are available by emailing SR or at http://en.neogalen.org/projects?mode=view&ret_mode=folder&post_id=4160304&folder_id=220280806

LIST OF ABBREVIATIONS

a	upper left cell of the 2 × 2 matrix
b	upper right cell of the 2 × 2 matrix
BCT	binary cross-tabulation
c	bottom left cell of the 2 × 2 matrix
CBES	effect size based on correlation; correlation-based effect size
d	bottom right cell of the 2 × 2 matrix
d	Cohen's effect size
ES	effect size
m	number of contingency tables for a given gross table
m	sample mean (arithmetic)
n	group size
N	sample size
r	Pearson's product-moment correlation coefficient
s	sample standard deviation
SOD	significance of differences
S_p	pooled standard deviation
s_{XY}	sample covariance
μ	population mean (arithmetic)
σ	population standard deviation
ϕ	mean square contingency coefficient, Pearson's Phi
$\bar{\phi}$	mean-square effect half-size, gross Pearson's Phi
χ^2	Pearson's chi-square statistic

DECLARATIONS

Conflict of interest

The author declares no conflicts of interest.

Funding

No funding.

Contributorship statement

The sole author is the only contributor.

REFERENCES

[1] Horton R. What is medicine's 5 sigma? Lancet 2015;385(9976):P1380.

[2] Kamerow D. Milestones, tombstones, and sex education. BMJ 2007;334:0-a.

[3] Ioannidis JPA. Why most published research findings are false? PLOS Medicine 2005;2(8):e124.

[4] Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. The Lancet 2009;374(9683):86–9.

[5] Macleod MB, Michie S, Roberts I, et al. Biomedical research: increasing value, reducing waste. Lancet 2014;383(9912):101-4.

[6] Ioannidis JPA. Evidence-based medicine has been hijacked: a report to David Sackett. J Clin Epidemiol 2016;73:82-6.

[7] Cohen J. Statistical Power Analysis for the Behavioral Sciences. Routledge 1988:75-144.

[8] Rosenthal R. Meta-analysis procedures for social research. Sage: Beverly Hills 1984.

[9] Cooper H, Hedges LV. The Handbook of Research Synthesis, Volume 236. Russell Sage Foundation 1994:238.

[10] Hedges LV, Olkin I. Statistical methods for meta-analysis. Orlando: Academic Press Inc 1985:77.

[11] Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to Meta-Analysis. John Wiley and Sons 2009:41-43.

[12] Pearson K. Notes on regression and inheritance in the case of two parents. Proc of the Royal Society of London 1895;58:240–2.

[13] Borenstein (2009), p.48, formula 7.7, 7.8.

[14] Cohen (1988), p. 23, formula 2.2.6.

[15] Cohen (1988), p. 82, formula 3.2.1.

[16] Evans JD. Straightforward Statistics for the Behavioral Sciences. Brooks/Cole Publishing, Pacific Grove 1996.

[17] Lauritzen SL. Lectures on Contingency Tables. Electronic edition, 1979-2002. [cited 2021 Jul 14] http://www.stats.ox.ac.uk/~steffen/papers/cont.pdf

[18] Everett BS, Skrondal A. The Cambridge dictionary of statistics, 4th ed. Cambridge University Press 2010.

[19] Pearson K. Mathematical Contributions to the Theory of Evolution. VII. On the Correlation of Characters not Quantitatively Measurable. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 1900;195:1-47.

[20] Yule GU. On the Association of Attributes in Statistics: With Illustrations from the Material of the Childhood Society, &c. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 1900;194:257–319.

[21] Cohen J. A coefficient of agreement for nominal scales. Educat Psychol Measurement 1960;20(1):37–46.

[22] Goodman LA, Kruskal WH. Measures of Association for Cross Classifications. J Amer Stat Assoc 1954;49(268):732–64.

[23] Goodman LA, Kruskal WH. Measures of Association for Cross Classifications. II: Further Discussion and References. J Amer Stat Assoc 1959;54(285):123–63.

[24] Goodman LA, Kruskal WH. Measures of Association for Cross Classifications III: Approximate Sampling Theory. J Amer Stat Assoc 1963;58(302):310–64.

[25] Goodman LA, Kruskal WH. Measures of Association for Cross Classifications, IV: Simplification of Asymptotic Variances. J Amer Stat Assoc 1972;67(338):415–21.

[26] Liu L, Berger VW. Two by Two Contingency Tables. In: Encyclopedia of Statistics in Behavioral Science (Eds: BS Everitt & DC Howell). John Wiley & Sons, Ltd, Chichester 2005;4:2076–81.

[27] Fleiss JL. Measures of effect size for categorical data. In: Cooper H, Hedges LV (Eds.). The handbook of research synthesis. New York: Russell Sage Foundation 1994:245–60.

[28] Olivier J, May WL, Bell ML. Relative effect sizes for measures of risk. Commun Stat Theory Methods 2017;46(14):6774-8.

[29] Olivier J, Bell ML. Effect Sizes for 2×2 Contingency Tables. PLoS One 2013;8(3):e58777.

[30] Glass GV, McGraw B, Smith ML. Meta-analysis in social research. Sage: Beverly Hills 1981.

[31] Cramér H. Mathematical Methods of Statistics. Princeton: Princeton University Press 1946:282.

[32] Ekström J. The Phi-coefficient, the Tetrachoric Correlation Coefficient, and the Pearson-Yule Debate. UCLA 2011 [cited 2021 Jun 15] https://escholarship.org/uc/item/7qp4604r..

[33] Everitt BS, Skrondal A. The Cambridge dictionary of statistics, 4th ed. Cambridge University Press 2010:325.

[34] Doll R, Hill AB. Smoking and carcinoma of the lung; preliminary report. BMJ 1950;2(4682),739–48.

[35] Contingency table. Wikipedia [cited 2021 Jun 14] https://en.wikipedia.org/wiki/Contingency_table