Scientific Research Publishing

Open Journal of Applied Sciences, 2025, 15(9), 2648-2661
https://www.scirp.org/journal/ojapps
ISSN Online: 2165-3925
ISSN Print: 2165-3917

Commentary on Grice *et al.*, 2020: A Critical Examination of Adjusted Effect Sizes (r_h and PCC_h) and Comparisons across Psychology and Medicine

James W. Grice[1*], Paul T. Barrett[2], Mateo Martin[1]

[1]Department of Psychology, Oklahoma State University, Stillwater, Oklahoma, United States of America
[2]Cognadev UK Ltd., Harrow, United Kingdom
Email: *james.grice@okstate.edu

How to cite this paper: Grice, J.W., Barrett, P.T. and Martin, M. (2025) Commentary on Grice *et al.*, 2020: A Critical Examination of Adjusted Effect Sizes (r_h and PCC_h) and Comparisons across Psychology and Medicine. *Open Journal of Applied Sciences*, **15**, 2648-2661.
https://doi.org/10.4236/ojapps.2025.159178

Received: August 7, 2025
Accepted: September 12, 2025
Published: September 15, 2025

Abstract

In their recent paper, *Persons as Effect Sizes*, Grice and colleagues advocated for a technique that adjusts statistical effect sizes upwards for intervention studies with low base rates like those found in medical and epidemiological research. This technique was developed by Ferguson in part as an aid for comparing medical effect size magnitudes to those reported in psychological studies. Herein we challenge the rationale behind this technique by particularly examining Ferguson's proposed distinction between "hypothesis-relevant" and "hypothesis-irrelevant" cases which lies at the heart of his method. We then advocate against using this technique and instead demonstrate graphical and numerical procedures, many of which are well known, that are rooted in the unadjusted raw data and that are consistent with the person-centered approach toward evaluating effect sizes. Finally, we explore the pitfalls associated with comparing effect sizes from medical studies, which often have miniscule base rates, to those found in psychological studies and conclude that such comparisons should be avoided.

Keywords

Effect Size, Correlation, Percent Correct Classifications, Epidemiology, Relative Risk, Medical Intervention Studies

1. Introduction

In their paper, *Persons as Effect Sizes*, Grice and colleagues [1] introduced a straightforward statistic to complement traditional effect size indices (e.g., d, η^2,

V). This statistic, termed the Percent Correct Classifications Index (PCC), quantifies the percentage of study participants whose behavior or responses align with theoretical expectations. Using examples from published studies, the authors illustrated its application in evaluating hypotheses about group differences (both within and between groups), associations, and relative risk. For relative risk assessment, Grice *et al.* developed a modified statistic, herein labeled PCC_h, by adapting a procedure from Ferguson [2] that adjusts small effect sizes in large clinical trials. This method (detailed below) corrects raw frequencies for low base rates and calculates a correlation, termed r_h, for the variables of interest. Ferguson developed this procedure following research that compared effect sizes between psychological and medical studies [3]. He concluded that psychological effect sizes may or may not compare favorably to medical ones but cautioned that such comparisons, even when using r_h, remain complex, challenging, and potentially unwarranted.

In the current paper, we revisit and build upon this line of prior research, making four key contributions. First, we reexamine the PCC_h, providing a detailed explanation of how Grice and colleagues applied Ferguson's adjustment to their person-centered statistic. This clarification sets the stage for our second point, where we critically evaluate the logic underpinning the computation of Ferguson's r_h. Our analysis reveals that Ferguson's frequency adjustments are based on inconsistent logic, therefore rendering r_h and PCC_h unsound. Third, we propose an alternative approach for utilizing the original PCC without relying on Ferguson's adjustment, offering a more robust framework for its application in studies of relative risk. Finally, we revisit the comparison of psychological and medical effect sizes, endorsing Ferguson's original conclusions but grounding our agreement in an expanded methodological perspective that incorporates diverse effect size indices.

2. PCC, r_h, and PCC_h Statistics

Grice *et al.*'s [1] Percent Correct Classifications (PCC) index has its roots in the earlier work of Cliff [4] and Grissom [5] and can also be seen in the recent work of Speelman and McGann [6]. As an example of relative risk assessment, Grice and colleagues analyzed data from the famous Salk vaccine trial conducted on over four-hundred thousand children. The results from the trial are presented in **Figure 1** which shows proportions of children who contracted or did not contract polio separated by their vaccination status. As can be seen, slightly more children receiving the placebo contracted polio (0.06%) compared to those who received the vaccine (0.02%). The frequency bars are also color-coded according to a naïve Bayesian classifier [7], with green bars indicating correctly classified observations and red bars indicating incorrectly classified observations. Converting the total number of correct classifications to a percentage yields a PCC equal to 49.96%. With only 50% of the children classified correctly, the result appears inconsistent with the known efficacy of the polio vaccine and scientists' understanding of virology [8] [9]. By comparison, Grice and colleagues analyzed data from an exper-

imental study of 307 men that compared hormonal treatment to a placebo. The results of that study yielded a high PCC value equal to 92.51%, indicating a clear difference in testosterone levels (the measured outcome dichotomized via median split) between the two groups. Almost all of the men in the treatment group had higher levels of testosterone when compared to men in the placebo group.

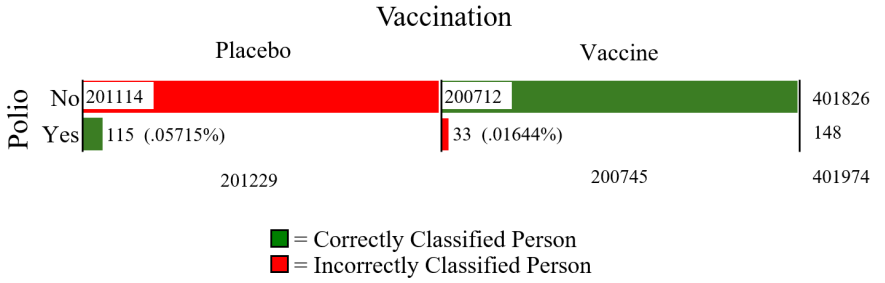

Figure 1. Frequencies of polio incidence by vaccination status. Note. Frequencies are reported in or next to the histogram bars. Total frequencies are reported in the margins of the figure. Percentages of children contracting or failing to contract polio in the vaccinated and non-vaccinated groups are also reported. Cases are classified using a naïve Bayesian classifier.

As an additional analysis of the Salk vaccine data, Grice and colleagues employed Ferguson's [2] procedure that upwardly adjusts small effect size correlations found in large clinical trials. Specifically, his procedure adjusts raw frequencies for low base rates (the exact steps will be described below) and then computes a correlation, dubbed r_h, for the two variables under consideration. Applying these adjustments to the data for the Salk vaccine trial yields the frequencies shown in **Figure 2**. The correlation for the original frequencies is 0.01 while the value for r_h, based on the adjusted frequencies, is 0.74. The PCC similarly increases from 49.96% to 85.65% (referred to herein as PCC_h) once the adjustment is applied. The higher r_h and PCC_h effect magnitudes are more consistent with what researchers might expect given the well-known efficacy of the Salk vaccine. These results therefore appear more valid than the original correlation and PCC… but is the reasoning behind their creation sound?

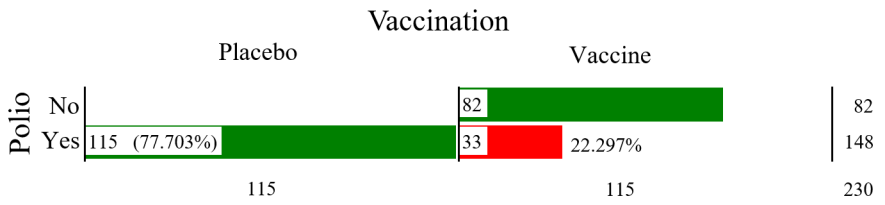

Figure 2. Adjusted frequencies of polio incidence by vaccination status. Note. "Hypothesis Irrelevant" cases are not included in the figure. The remaining cases are classified using a naïve Bayesian classifier.

The Salk vaccine trial data are reported in **Table 1** ([3], p. 226). As can be seen, 401,974 children received either the experimental vaccine (n = 200,745) or a placebo (n = 201,229). Of those receiving the former, 33 contracted polio while 115

of those receiving the placebo contracted the disease. The simplest formula for Pearson's *r* (or *phi*) for the dichotomous variables in Table 1 is as follows (letters denote cells in the table):

Table 1. Salk vaccine contingency table.

	Placebo	Vaccine	Totals
No Polio	201,114 (a)	200,712 (b)	401,826
Polio	115 (c)	33 (d)	148
Totals	201,229	200,745	401,974

$$\begin{aligned}\phi &= \frac{bc-ad}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \\ &= \frac{(200712*115)-(201114*33)}{\sqrt{(201114+200712)(115+33)(201114+115)(200712+33)}} \\ &= 0.01\end{aligned} \quad (1)$$

A value of zero would indicate complete independence between the drug and outcome variables, and the result is thus surprisingly low given the widely held belief in the vaccine's efficacy. Rosnow and Rosenthal [3] presented additional surprisingly small correlations for treatments and outcomes from the medical literature. The point of presenting such small magnitudes was to suggest that psychologists should not be ashamed of nor defensive about the small statistical effect sizes they often find in their studies of the human psyche. After all, if modern medicine has advanced over the past decades with such small statistical effect sizes, so too can psychology.

Ferguson [2] took Rosnow and Rosenthal's [3] work a step further by arguing that phi is not an optimal measure of effect size for judging efficacy because of the impact of low base rates in the context of large samples. He argued the computation of phi includes unnecessary "hypothesis irrelevant" cases, and removing such cases would yield an estimate of effect size that is more ecologically valid. Using the data presented in Table 1, Ferguson reasoned that the hypothesis of interest is "The Salk vaccine is effective in preventing polio in individuals who are exposed to the polio virus" ([2], p. 132). Moreover, he pointed out the researchers used a "wide net" sampling procedure, and by doing so many children in the sample were not relevant to testing the hypothesis. Ferguson's task, then, was to remove these children and recompute the correlation. The steps he took were as follows:

1) Compute the proportion of children in the control (placebo) group who contracted polio: 115/201,229 = 5.71×10^{-4} (see Table 1). It is important to note the 115 children in this group who contracted polio without being vaccinated are considered as hypothesis relevant, whereas the remaining 201,114 children are deemed hypothesis irrelevant and will consequently be deleted from the table.

2) Use the proportion from Step 1 to compute the number of children in the vaccine group who are expected to contract polio: $5.71 \times 10^{-4} \times 200{,}745 = 114.72$. This value is rounded to a whole number, 115. This result indicates that, if the

vaccine were perfectly ineffective, we would still expect 115 children in the vaccinated group to contract polio.

3) Adjust the original contingency table by removing the hypothesis irrelevant children and replacing the total children vaccinated (n = 200,745) with the expected number of polio victims under inefficacy (n = 115), as can be seen in Table 2. The vaccinated children who contracted polio (n = 33) are considered hypothesis relevant and kept in the table. With 33 vaccinated children contracting polio, 82 of the 115 expected polio victims remain. This value is entered into the contingency table to replace the 200,712 vaccinated children who did not contract polio. The 200,630 vaccinated healthy children removed from the table are therefore considered as hypothesis irrelevant.

Table 2. Adjusted Salk vaccine contingency table.

	Placebo	Vaccine	Totals
No Polio	0	82	82
Polio	115	33	148
Totals	115	115	230

4) Use Equation (1) above on the adjusted frequencies in Table 2 to compute r_h. The result is equal to 0.74.

A second example included in Rosnow and Rosenthal's [3] original paper entails the relationship between aspirin consumption and incidence of heart attack [10] [11]. The data for the original study of 22,071 men is shown in Table 3 and the correlation between the two variables is equal to only 0.03 (PCC = 50.46%). Following the steps outlined above yields the adjusted frequencies reported in Table 4, for which the value of r_h is quite a bit higher at 0.52. The value for PCC_h computed from the adjusted frequencies is also notably higher at 70.92%.

Table 3. Aspirin study contingency table.

	Placebo	Aspirin	Totals
No Heart Attack	10,795	10,898	21,693
Heart Attack	239	139	378
Totals	11,034	11,037	22,071

Table 4. Adjusted aspirin study contingency table.

	Placebo	Aspirin	Totals
No Heart Attack	0	100	100
Heart Attack	239	139	378
Totals	239	239	478

3. Criticisms of r_h and PCC_h

As described above, Ferguson removed cases from contingency tables because he

reasoned they were "hypothesis irrelevant". Not only is the removal of individuals inconsistent with the person-centered approach advocated by Grice *et al.* [1], but the two examples show this is no small matter. In the Salk vaccine study 401,744 cases (99.94%) were removed and in the aspirin study 21,592 cases (97.82%) were removed. Are such large percentages of cases truly irrelevant to the hypotheses? We think the answer is "no", and the reason for this answer can be found in the computations above. Consider the Salk vaccine study. The baseline proportion of children expected to contract polio without being given a vaccine plays a critical role in the computation of r_h. This baseline is calculated as: $115/201{,}229 = 5.71 \times 10^{-4}$. The computation is possible only because there are 201,114 healthy (no polio), unvaccinated children in the sample; yet these children are deemed as "hypothesis irrelevant". In other words, to compute the necessary proportion of children expected to contract polio without intervention and compute r_h as a test of the hypothesis "The Salk vaccine is effective in preventing polio in individuals who are exposed to the polio virus", some of the hypothesis irrelevant children are needed. It appears, then, these children are not truly hypothesis irrelevant.

A counterargument might suggest the children are irrelevant because the baseline proportion is computed using the number of children who were not vaccinated and contracted polio ($n = 115$) and the total number of unvaccinated children ($n = 201{,}229$). The specific number of unvaccinated children who did not contract polio ($n = 201{,}114$) is not included in the computation. In a sense this argument is technically true, which is to say it is mathematically true, but it obviously glosses over the fact that the latter group is included in the former group of children. In other words, it is detached from the real experiences of individual children represented by the numbers…ironically a reality Grice *et al.* [1] sought to protect in their *persons as effect sizes* approach. Each child in the placebo group is observed by a doctor and judged to be either healthy or sick, and with each judgement something is learned about the prevalence of polio in the sample…a rate which is central to computing r_h. The judgment that unvaccinated Johnny is sick with polio is just as valuable and necessary to computing the proportion above as is the judgment that unvaccinated Susie is not sick. Healthy or sick then, each child is relevant to the hypothesis under consideration, thus undercutting the rationale supporting r_h and, by extension, PCC_h.

4. Persons as Effect Sizes

Where does this conclusion leave Grice *et al.* [1] with their *persons as effect sizes* approach? First, it only applies to the efficacy assessment (denoted as "risk assessment") examples they reported for intervention studies using 2 × 2 experimental designs. Such designs include a treatment vs. control variable (e.g., vaccine vs. placebo) and a dichotomous outcome variable (e.g., polio present vs. polio absent). It has no bearing on the other examples of within-subject designs, between-group comparisons, and variable associations for which they computed the PCC in their paper. Second, when working with two dichotomous variables from a

medical or psychological intervention study, it follows from the criticisms above that raw frequencies in contingency tables should *not* be adjusted. In other words, the adjusted graphs employed by Grice *et al.* are not to be recommended and PCC_h is not to be computed. All participants are to be considered as hypothesis relevant in an intervention study and should be included in contingency tables and figures. This recommendation also avoids the confusion that is likely to result from the deletion of cases. Returning to **Figure 2**, no children are reported for the Polio/Placebo condition. A naïve observer will rightly question why no children are reported for this condition, why 100% of the children in the placebo group contracted polio, and why only 230 children are reported from a total sample of over four-hundred thousand (assuming the actual sample size is known). The graph is unfortunately a radical distortion of reality. If the graphing procedures by Grice *et al.* are to be used, then, we recommend reporting all of the cases for the treatment and control conditions, as shown in **Figure 1**. While low base rates may barely register as frequency bars in a given figure, they should not be hidden from view as a truly person-centered approach requires all persons to be included in any effort to interpret the meaning of a given study's outcome. Reporting frequencies, proportions, or percentages in the histogram bars will moreover facilitate interpretation of the results as well as the computation of the unadjusted PCC and other person-relevant statistics.

These additional statistics (described below) are well known among epidemiologists, and we consider them to be person-centered because they can be understood from the vantage point of the layperson, whose central question regarding an intervention study's outcome is "how necessary and effective is this treatment likely to be for me?" Generally speaking, person-centered approaches treat individuals and their unique response patterns as the primary units of analysis, in contrast to variable-centered approaches, which rely on aggregate statistics (e.g., means, variances, correlations) to examine group differences or variable relationships (see [12]). Consider the Salk vaccine trial data reported in **Figure 1** above. Using this figure and the numbers reported therein, how is a layperson to answer this question? The correlation coefficient is not up to the task as it addresses a variable-centered question; namely, are the two variables associated? As is well known, Pearson's r is a unit-free, generic metric of association as can be seen in its z-score formulation:

$$r_{xy} = \frac{\sum z_x z_y}{n-1} \tag{2}$$

The z-transformation removes mean and variance differences between variables, which can be quite useful when working with a wide array of variables such as ability test scores, self-report inventory sum scores, or clinical scale scores. This diversity of variables is also why psychologists find standardization to be useful for aggregating effects across studies. The downside is that, missing from the z-score-based correlation formula above, are the base rates for dichotomous variables like those in intervention studies; and it is these base rates that are critical for

interpreting the meaning of a given risk-assessment study's results. This fact is well known (see [13]) and is what drove Ferguson [2] to decimate sample sizes in the development of r_h. As noted above, the PCC (reported in isolation from **Figure 1**) similarly fails to make plain the importance of low base rates.

One route forward for the layperson in interpreting binary outcomes like the Salk data reported in **Figure 1** is to compare the percentages of disease in the treatment and control groups as follows:

$$\text{ARR} = 0.05715\% - 0.01644\% = 0.04071\% \tag{3}$$

This value is the well-known Absolute Risk Reduction (or Risk Difference, RD) statistic. It here shows 0.041% fewer vaccinated children contracted polio compared to those who were not vaccinated. The risk of contracting polio without the vaccine is already quite low (0.057%), but the vaccine lowers it by an additional 0.041%. With such a low base rate and small ARR, it is not surprising the Number Needed to Treat (NNT; [14]), another effect size index, is quite high:

$$\text{NNT} = \frac{1}{0.0004071} = 2456.40\left(\text{or } 2456 \text{ persons}\right) \tag{4}$$

This result means that 2456 children would need to be vaccinated to prevent one case of polio. This statistic clearly fits the spirit of the person-centered approach and on the surface suggests the vaccine is not highly efficacious; however, caution is warranted. The statistic cannot be interpreted in isolation from the raw data as the prevalence of polio is extremely low, as clearly shown in **Figure 1**.

Should the low risk of infection without vaccine be taken into account in the computation of a numeric effect size? Consider the Relative Risk Reduction (RRR) as one alternative:

$$\text{RRR} = \frac{0.05715\% - 0.01644\%}{0.05715\%} = 0.71234\left(\sim 71\%\right) \tag{5}$$

This result is interpreted as indicating 71% efficacy and is commonly reported for modern vaccines (e.g., "A two-dose regimen of BNT162b2 conferred 95% protection against Covid-19 in persons 16 years of age or older." [15], p. 2603). It is clear, however, the reported efficacy represents a decrease in the infection rates relative to the extremely low base rate observed in the placebo group.

Yet another method for including the low base rate in a computed effect size is the Odds Ratio (OR):

$$\text{OR} = \frac{0.05715\%}{0.01644\%} = 3.47628\left(\sim 3.48\right) \tag{6}$$

Those receiving the placebo were 3.48 times more likely to contract polio compared to those receiving the Salk vaccine. Like the RRR, this result is also apt to strike the layperson's ear as indicating impressive efficacy.

All of these effect sizes and the PCC can be quickly computed from the numbers presented in **Figure 1**, and they are all perfectly legitimate even though they are likely to invoke different interpretations of efficacy regarding the Salk vaccine.

Considering the ARR (or RD) and RRR, for instance, is one to conclude the vaccine decreases the chance of contracting polio by 0.04%, a trivial amount, or by 71%, a large amount? According to Stegenga [16], "The answer is that it does both, because the question is ambiguous." (p. 67). The percentage of individuals contracting polio after vaccination decreases by .04071%; but 0.04071% of something that has a prevalence rate of 0.05715% across many people is ~71%. He therefore concluded both effect sizes should be considered but that "At the very least, they [the laypersons] need the absolute measure [RD] to make an informed treatment decision. Effectiveness of an intervention, from the first-person perspective of a patient, is, roughly, the degree to which the intervention increases the probability that the patient will experience the beneficial outcome in question. This difference-making notion is adequately represented by RD and is not adequately represented by RRR [sic]…From the individual patient's perspective, then, the appropriate outcome measure is RD." ([16], p. 67; see also, [17] [18]). The safest route forward, then, appears to be a "full disclosure" approach to reporting results for these types of intervention studies. Once **Figure 1** is generated, all the effect sizes above can easily be computed and interpreted by researchers and laypersons alike in the context of the specific study's design and results (viz., 2 × 2 design, sample size and control and treatment rates; cf. [19]). Admittedly, conflicting statistical effect size indices may confuse some readers, particularly when interpreting complex phenomena. However, transparency in reporting all relevant metrics remains the superior scientific approach, as it fosters comprehensive understanding and allows for a critical evaluation of a given study's findings. This "full disclosure" practice also supports replication efforts and ensures that the nuances of effect size interpretations are not obscured, thus promoting robust scientific discourse.

5. Effect Size Indices in Psychology and Medicine

Recall Ferguson [2] was inspired by Rosnow and Rosenthal's ([3], p. 227) brief discussion of effect sizes across medical and psychological studies. **Table 5** reports results from most of the studies used in their discussion along with several added examples.[1] The table includes all of the effect size indices discussed above as well as their intercorrelations. Some of the disparities between the various effect sizes in the table are quite remarkable. For the Covid-19 vaccine, for instance, $r = 0.06$, RRR = 95.03%, ARR = 0.84%, NNT = 119, OR = 20.28, and PCC = 50.25. While RRR and OR suggest high efficacy for the vaccine, the other indices suggest low efficacy. By comparison, a study investigating the impact of a cholesterol treatment on coronary heart disease yielded one of the highest ARR (20.99%) and PCC

[1]Necessary frequencies for computing the indices of effect sizes could not be located for a number of studies reported in the original table by Rosnow and Rosenthal. These studies are omitted. Moreover, a number of r and r_h values in **Table 5** differ from the corresponding values reported by Ferguson by 0.03 or less (in absolute value). These differences are likely due to rounding. However, r_h values for the Vietnam Veteran and Testosterone studies differed substantially (0.44 vs. 0.15 and 0.62 vs. 0.36 for Ferguson vs. **Table 5**). The two values reported in **Table 5** are consistent with the high correlation between r_h and RRR. All values in **Table 5** were computed on the basis of the study frequencies using equations programmed into Excel.

(60.49%) values and lowest NNT values (5), indicating impressive efficacy. The Pearson's correlation was still only 0.22, a small value considering a range of 0 (no association) to 1 (perfect association). The RRR (34.69%) and OR (2.34) results were also less impressive than those for the Covid-19 vaccine and other interventions. Which effect size is telling the scientist or layperson the true story of efficacy for these treatments or any others reported in Table 5?

Table 5. Indices of effect size for various intervention studies.

Independent Variable	Dependent Variable	n	B Rate	r	r_h	RRR	ARR	NNT	OR	PCC	PCC$_h$
Salk Vaccine	Polio	401,974	<0.01	0.01	0.74	71.24	0.04	2456	3.48	49.96	85.65
Beta carotene	Death	29,133	0.13	0.01	0.20	7.33	0.93	107	1.09	50.48	53.65
Aspirin	Heart Attack	22,071	0.02	0.03	0.51	41.86	0.91	110	1.74	50.46	70.92
Streptokinase	Death	11,712	0.13	0.03	0.31	17.26	2.24	45	1.24	51.14	58.60
Alendronate	Hip Fracture	2027	0.02	0.04	0.58	50.83	1.11	90	2.06	50.96	75.00
Covid-19 Vaccine	Covid	36,523	0.01	0.06	0.95	95.03	0.84	119	20.28	50.25	97.52
Vietnam veteran status	Alcohol Problems	4462	0.91	0.07	0.15	4.97	4.52	22	1.57	47.78	46.96
Garlic	Death	432	0.10	0.09	0.55	47.97	4.57	22	2.02	53.47	73.17
Testosterone	Adult delinquency	4462	0.90	0.12	0.36	14.04	12.64	8	2.63	83.26	91.41
Cyclosporine	Death	209	0.10	0.15	0.76	71.93	7.46	13	3.86	53.11	86.36
Warfarin	Blood Clots	508	0.15	0.15	0.67	62.46	9.13	11	2.95	54.72	81.08
Hosp. vs. Tx Choice	Alcohol Problems	144	0.38	0.16	0.49	38.76	14.74	7	2.02	57.64	69.09
AZT for neonates	HIV	364	0.22	0.21	0.71	67.50	14.84	7	3.66	57.42	83.75
Cholesterol treatment	Coronary status	162	0.60	0.21	0.46	34.69	20.99	5	2.34	60.49	67.35
AZT	Death	282	0.12	0.23	0.94	94.09	10.99	9	19.04	56.74	96.97
Tx. Choice vs. AA	Alcohol problems	154	0.63	0.25	0.50	39.30	24.62	4	2.73	62.34	71.88
Hosp. vs. AA	Alcohol problems	156	0.63	0.40	0.69	62.83	39.36	3	5.53	69.23	82.65
	Rank Correlations										
	Base Rate	−0.525									
	r	−0.876	0.575								
	r_h	−0.048	−0.560	0.233							
	RRR	−0.044	−0.578	0.218	0.998						
	ARR	−0.884	0.757	0.924	−0.074	−0.088					
	NNT	0.884	−0.757	−0.924	0.074	0.088	1.000				
	OR	−0.238	−0.141	0.520	0.873	0.855	0.265	−0.265			
	PCC	−0.753	0.609	0.797	−0.044	−0.069	0.907	−0.907	0.270		
	PCC$_h$	0.037	−0.331	0.233	0.836	0.806	0.010	−0.010	0.873	0.174	

Note. Hosp. = compulsory hospitalization; tx. = treatment; AA = Alcoholics Anonymous. Sources: Salk Vaccine [20]; Aspirin [10]; Alendronate [21]; Covid Vaccine [15]; Beta carotene [22]; Streptokinase [23]; Vietnam veteran status [24]; Garlic [25]; Testosterone [26]; Hosp. vs. Tx Choice [27]; Cyclosporine [28]; Warfarin [29]; AZT for neonates [30]; Cholesterol treatment [31]; AZT [32]; Tx. Choice vs. AA [27]; Hosp. vs. AA [27].

Examination of the Spearman correlations in Table 5 also indicates that, as noted above, base rates are playing an important role in the variability between the effect size indices. Base rates for the control conditions are correlated positively with r (0.573) and ARR (0.757) and negatively with RRR (−0.578) and OR (−0.141). As can also be seen, sample size is negatively correlated with r (−0.885) and ARR (−0.884) and is nearly orthogonal to RRR (−0.044). As an aside, the correlation between r_h and sample size is also nearly zero (−0.048, as noted by Ferguson [2]) and the correlation between r_h and RRR is near unity (rho = 0.998). This latter result is not surprising, however, as the base rate adjustment used in Ferguson's decimation of sample sizes is the denominator of the RRR (see Equation (5)).

The general lesson learned from the contents of Table 5 is that interpreting effect sizes from either the researcher's or layperson's point of view is no simple matter for intervention studies. Moreover, attempts at cross-discipline comparisons between medical and psychological studies may be poorly motivated. Medical and epidemiological effect sizes from 2 × 2 contingency tables like those above are focused on explicit-intervention efficacy (or risk) estimation, whereas many effect sizes in psychology are focused on generic associations or mean differences between variables ([33] [34]). Ferguson [2] also notes issues of reliability and validity that make it difficult to compare effect sizes across disciplines, and Rosnow and Rosenthal [3] note the importance of context and the nature of the dependent variable. It seems psychologists may nonetheless be tempted to leverage epidemiological arguments from the risk literature to justify interpreting small effect sizes as consequential (e.g. [35]-[37]). This strategy is problematic because the base rate for a psychological phenomenon of interest is meaningless when there is no actual intervention being deployed or assessed as a population outcome or effect. Psychological test scores, for example, do not constitute an "intervention" as might be instantiated when establishing the risk vs. benefit of say an aortic stent surgical operation for patients with a certain kind of heart failure, or the impact of a specific kind of hygiene intervention within a target population. This distinction is why many psychological research reports of effect size magnitude do not engage with absolute or relative risk estimates of outcomes, because no explicit phenomenal interventions are being investigated (unlike in medical research). A more recent exposition from the psychiatric/mental health area is more realistic when it extrapolates small effects for mental health issues as having important epidemiological consequences [38]. Without such careful extrapolation, it's prudent to view small effect sizes as indicating trivial real-world differences, thereby preventing overstated claims about their influence on outcomes or behavior, avoiding misrepresentation of relationship strength, and emphasizing the need for replication or stronger effect sizes in subsequent studies.

6. Conclusion

In summary, attempts to compare psychological effects with medical research outcomes should be abandoned—a conclusion also reached by Ferguson ([2], p. 135).

It seems there is nothing to be gained from these efforts except confusion and misrepresentation of outcomes. The coefficients used by many psychologists and medical diagnostics share little in common, especially the language each employs to address the interpretation of the results. The majority of psychologists do not employ interventions and assess binary outcomes and are therefore unconcerned with base rates, focusing instead on associations or mean differences between variables. Likewise, the base rates of a phenomenon of interest in epidemiological/risk estimation are usually miniscule compared to those found in psychology research. Epidemiological extrapolation of small psychological effects may unfortunately be little more than "armchair speculation", invariably unsupported by clear empirical evidence of the projected outcomes over time. While also recommending psychologists avoid conventions for interpreting effect sizes (e.g., "$d = 0.2$ is a small effect"), Götz, Gosling, and Rentfrow [19] recently concurred that such speculation be avoided when interpreting the importance of a given effect.

Author Contributions

JWG conceptualized and wrote a majority of the manuscript. PTB contributed to writing the manuscript and to data analysis. MM contributed to data collection and analysis.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Grice, J.W., Medellin, E., Jones, I., Horvath, S., McDaniel, H., O'lansen, C., *et al.* (2020) Persons as Effect Sizes. *Advances in Methods and Practices in Psychological Science*, **3**, 443-455. https://doi.org/10.1177/2515245920922982

[2] Ferguson, C.J. (2009) Is Psychological Research Really as Good as Medical Research? Effect Size Comparisons between Psychology and Medicine. *Review of General Psychology*, **13**, 130-136. https://doi.org/10.1037/a0015103

[3] Rosnow, R.L. and Rosenthal, R. (2003) Effect Sizes for Experimenting Psychologists. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, **57**, 221-237. https://doi.org/10.1037/h0087427

[4] Cliff, N. (1993) Dominance Statistics: Ordinal Analyses to Answer Ordinal Questions. *Psychological Bulletin*, **114**, 494-509. https://doi.org/10.1037//0033-2909.114.3.494

[5] Grissom, R.J. (1994) Statistical Analysis of Ordinal Categorical Status after Therapies. *Journal of Consulting and Clinical Psychology*, **62**, 281-284. https://doi.org/10.1037/0022-006x.62.2.281

[6] Speelman, C.P. and McGann, M. (2020) Statements about the Pervasiveness of Behavior Require Data about the Pervasiveness of Behavior. *Frontiers in Psychology*, **11**, Article 594675. https://doi.org/10.3389/fpsyg.2020.594675

[7] Sauer, S. (2017) Observation Oriented Modeling Revised from a Statistical Point of View. *Behavior Research Methods*, **50**, 1749-1761. https://doi.org/10.3758/s13428-017-0949-8

[8] Plotkin, S.A. (2011) History of Vaccine Development. Springer. https://doi.org/10.1007/978-1-4419-1339-5

[9] Giese, M. (2016) Introduction to Molecular Vaccinology. Springer. https://doi.org/10.1007/978-3-319-25832-4

[10] Steering Committee of the Physicians' Health Study Research Group (1988) Preliminary Report: Findings from the Aspirin Component of the Ongoing Physicians' Health Study. *New England Journal of Medicine*, **318**, 262-264. https://doi.org/10.1056/nejm198801283180431

[11] Steering Committee of the Physicians' Health Study Research Group (1989) Final Report on the Aspirin Component of the Ongoing Physicians' Health Study. *New England Journal of Medicine*, **321**, 129-135. https://doi.org/10.1056/nejm198907203210301

[12] Grice, J.W. (2015) From Means and Variances to Persons and Patterns. *Frontiers in Psychology*, **6**, Article 1007. https://doi.org/10.3389/fpsyg.2015.01007

[13] Sackett, D.L. (2001) Why Randomized Controlled Trials Fail but Needn't: 2. Failure to Employ Physiological Statistics, or the Only Formula a Clinician-Trialist Is Ever Likely to Need (or Understand!). *Canadian Medical Association Journal*, **65**, 1226-1237. https://pmc.ncbi.nlm.nih.gov/articles/PMC81587/

[14] McAlister, F.A. (2008) The "Number Needed to Treat" Turns 20—And Continues to Be Used and Misused. *Canadian Medical Association Journal*, **179**, 549-553. https://doi.org/10.1503/cmaj.080484

[15] Polack, F.P., Thomas, S.J., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., *et al.* (2020) Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. *New England Journal of Medicine*, **383**, 2603-2615. https://doi.org/10.1056/nejmoa2034577

[16] Stegenga, J. (2015) Measuring Effectiveness. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, **54**, 62-71. https://doi.org/10.1016/j.shpsc.2015.06.003

[17] Stegenga, J. (2022) Red Herrings about Relative Measures: A Response to Hoefer and Krauss. *Studies in History and Philosophy of Science*, **92**, 56-59. https://doi.org/10.1016/j.shpsa.2022.01.004

[18] Hoefer, C. and Krauss, A. (2021) Measures of Effectiveness in Medical Research: Reporting Both Absolute and Relative Measures. *Studies in History and Philosophy of Science*, **88**, 280-283. https://doi.org/10.1016/j.shpsa.2021.06.012

[19] Götz, F.M., Gosling, S.D. and Rentfrow, P.J. (2024) Effect Sizes and What to Make of Them. *Nature Human Behaviour*, **8**, 798-800. https://doi.org/10.1038/s41562-024-01858-z

[20] Francis, T., Korns, R., Voight, R., Boisen, M., Hemphill, F., Napier, J. and Tolchinsky, E. (1955) An Evaluation of the 1954 Poliomyelitis Vaccine Trials—Summary Report. *American Journal of Public Health*, **45**, 1-63.

[21] Black, D.M., Cummings, S.R., Karpf, D.B., Cauley, J.A., Thompson, D.E., Nevitt, M.C., *et al.* (1996) Randomised Trial of Effect of Alendronate on Risk of Fracture in Women with Existing Vertebral Fractures. *The Lancet*, **348**, 1535-1541. https://doi.org/10.1016/s0140-6736(96)07088-2

[22] The α-Tocopherol, Beta Carotene Cancer Prevention Study Group (1994) The Effect of Vitamin E and β Carotene on the Incidence of Lung Cancer and Other Cancers in Male Smokers. *New England Journal of Medicine*, **330**, 1029-1035. https://doi.org/10.1056/NEJM199404143301501

[23] Gruppo Italiano per lo Studio della Streptochinasi Nell'Infarto Miocardico (1986) Ef-

fectiveness of Intravenous Thrombolytic Treatment in Acute Myocardial Infarction. *Lancet*, **1**, 397-402.

[24] Centers for Disease Control Vietnam Experience Study (1988) Health Status of Vietnam Veterans: 1. Psychosocial Characteristics. *Journal of the American Medical Association*, **259**, 2701-2707. https://doi.org/10.1001/jama.259.18.2701

[25] Goldfinger, S.E. (1991) Garlic: Good for What Ails You. *Harvard Health Letter*, **16**, 1-2.

[26] Dabbs, J.M. and Morris, R. (1990) Testosterone, Social Class, and Antisocial Behavior in a Sample of 4,462 Men. *Psychological Science*, **1**, 209-211. https://doi.org/10.1111/j.1467-9280.1990.tb00200.x

[27] Walsh, D.C., Hingson, R.W., Merrigan, D.M., Levenson, S.M., Cupples, L.A., Heeren, T., *et al.* (1991) A Randomized Trial of Treatment Options for Alcohol-Abusing Workers. *New England Journal of Medicine*, **325**, 775-782. https://doi.org/10.1056/nejm199109123251105

[28] Canadian Multicentre Transplant Study Group (1983) A Randomized Clinical Trial of Cyclosporine in Cadaveric Renal Transplantation. *New England Journal of Medicine*, **309**, 809-815. https://doi.org/10.1056/nejm198310063091401

[29] Grady, D. (2003) Safe Therapy Is Found for High Blood-Clot Risk. New York Times, A1-A22.

[30] Altman, L.K. (1994) In Major Finding, Drug Limits HIV Infection in Newborns. New York Times, A1-A13.

[31] Roberts, L. (1987) Study Bolsters Case against Cholesterol. *Science*, **237**, 28-29. https://doi.org/10.1126/science.3603009

[32] Barnes, D.M. (1986) Promising Results Halt Trial of Anti-Aids Drug. *Science*, **234**, 15-16. https://doi.org/10.1126/science.3529393

[33] Fritz, C.O., Morris, P.E. and Richler, J.J. (2012) "Effect Size Estimates: Current Use, Calculations, and Interpretation": Correction to Fritz *et al.* (2011). *Journal of Experimental Psychology*: *General*, **141**, 30-30. https://doi.org/10.1037/a0026092

[34] Schäfer, T. and Schwarz, M.A. (2019) The Meaningfulness of Effect Sizes in Psychological Research: Differences between Sub-Disciplines and the Impact of Potential Biases. *Frontiers in Psychology*, **10**, Article 813. https://doi.org/10.3389/fpsyg.2019.00813

[35] Gignac, G.E. and Szodorai, E.T. (2016) Effect Size Guidelines for Individual Differences Researchers. *Personality and Individual Differences*, **102**, 74-78. https://doi.org/10.1016/j.paid.2016.06.069

[36] Funder, D.C. and Ozer, D.J. (2019) Evaluating Effect Size in Psychological Research: Sense and Nonsense. *Advances in Methods and Practices in Psychological Science*, **2**, 156-168. https://doi.org/10.1177/2515245919847202

[37] Götz, F.M., Gosling, S.D. and Rentfrow, P.J. (2021) Small Effects: The Indispensable Foundation for a Cumulative Psychological Science. *Perspectives on Psychological Science*, **17**, 205-215. https://doi.org/10.1177/1745691620984483

[38] Carey, E.G., Ridler, I., Ford, T.J. and Stringaris, A. (2023) Editorial Perspective: When Is a 'Small Effect' Actually Large and Impactful? *Journal of Child Psychology and Psychiatry*, **64**, 1643-1647. https://doi.org/10.1111/jcpp.13817