Original paper

New effect size and sample size guidelines in dentistry

Grzegorz Zieliński[1,A–F], Mieszko Więckiewicz[2,E,F]

[1] Department of Sports Medicine, Medical University of Lublin, Poland
[2] Department of Experimental Dentistry, Wroclaw Medical University, Poland

A – research concept and design; B – collection and/or assembly of data; C – data analysis and interpretation;
D – writing the article; E – critical revision of the article; F – final approval of the article

Dental and Medical Problems, ISSN 1644-387X (print), ISSN 2300-9020 (online) *Dent Med Probl*. 2025;62(5):907–917

Address for correspondence
Grzegorz Zieliński
E-mail: grzegorz.zielinski@umlub.pl

Funding sources
None declared

Conflict of interest
None declared

Acknowledgements
None declared

Received on July 2, 2025
Reviewed on July 17, 2025
Accepted on September 8, 2025

Published online on October 28, 2025

Abstract

Background. Cohen has emphasized that the recommended thresholds for effect sizes should only be used in the absence of detailed information about effect size distributions within specific fields.

Objectives. The study aimed to establish updated effect size thresholds (Cohen's *d*, Hedges' *g* and Pearson's *r*) tailored for research in dentistry.

Material and methods. Following methodologies from prior research on effect sizes, the data was extracted from meta-analyses published in the top 10 ranked dentistry journals. The 25th, 50th and 75th percentiles were calculated for Pearson's *r* values, as well as for Cohen's *d* or Hedges' *g*. A total of 4,250 studies were analyzed, with statistical analyses conducted using the R programming language.

Results. The 25th, 50th and 75th percentiles for Pearson's *r* in individual differences research were 0.16, 0.40 and 0.67, respectively. For Hedges' *g*, the percentiles corresponding to small, medium and large effect sizes were 0.10, 0.35 and 0.86, respectively.

Conclusions. In light of these findings, researchers in the field of dentistry are encouraged to adopt the following thresholds: for Pearson's *r*, 0.20 for small effects, 0.40 for medium effects and 0.70 for large effects; and for Cohen's *d* or Hedges' *g*, 0.10 for small effects, 0.40 for medium effects and 0.90 for large effects. These updated thresholds can improve the rigor and quality of dental research, ultimately benefiting patients through enhanced diagnostics and treatment strategies.

Keywords: dentistry, sample size, effect size, stomatology, statistical power

Cite as
Zieliński G, Więckiewicz M. New effect size and sample size guidelines in dentistry. *Dent Med Probl.* 2025;62(5):907–917. doi:10.17219/dmp/210478

DOI
10.17219/dmp/210478

Highlights

- Recommended effect size thresholds for dental research are: Pearson's r = 0.20 (small), 0.40 (medium) and 0.70 (large); and Hedges' g = 0.10 (small), 0.40 (medium) and 0.90 (large).
- Adoption of these thresholds may improve methodological rigor, enhance research quality, and support more accurate diagnostics and treatment in dentistry.
- The study provides guidance on determining appropriate sample sizes in dental research based on desired statistical power and effect size.

Introduction

Scientific research has a tangible impact on health of the population. Over the years, an increase in the number of studies has been observed, and a systematic rise is predicted.[1] Alongside the growth in research quantity, quality should also improve. One factor determining the quality of research is the rigor of the statistical analysis.[2] Contemporary research focuses primarily on reporting p-values for statistical significance, often neglecting the value of the effect size.[2,3]

The significance of research findings is not always adequately represented by statistical significance.[4–7] Results that achieve the predetermined significance level may not be clinically significant, and vice versa.[5] For example, in very large samples, statistical significance is almost always achieved, which may be misinterpreted (without analyzing the effect size) as sample variability.[7] Therefore, regardless of statistical significance, researchers must assess whether the results are clinically meaningful and relevant to their scientific field.[5]

Recommendations for reporting effect sizes are systematically published to enhance the quality of scientific research, thereby improving decision-making in patient treatment.[7,8] Cohen is the most prominent researcher who provided guidelines for effect size analysis. He defined the following thresholds for Cohen's d and Hedges' g: 0.20 (small effect); 0.50 (medium effect); and 0.80 (large effect). For Pearson's r, the established thresholds are 0.10 (small effect), 0.30 (medium effect) and 0.50 (large effect).[9] However, it has been observed that effect sizes may vary across research fields.[10]

For instance, different thresholds have been developed for rehabilitation.[11] To further refine the statistical framework, specific guidelines have been established for physiotherapy.[12] In addition, thresholds have been formulated for gerontology,[6] hearing research,[5] and exercise-based treatments for tendinopathy,[13] as well as for research related to the temporomandibular joint (TMJ) and masticatory muscles.[14]

To date, no guidelines for Cohen's d, Hedges' g or Pearson's r specific to dentistry have been identified. Dentistry, as a branch of medicine, differs significantly from other medical fields. These differences are evident from the outset, including preclinical and clinical education for dentistry students compared to medical students.[15–21] Further distinctions emerge in professional practice, with unique methods of treatment and patient care.[20,22–26] The analysis of the function, pathologies, and treatment of teeth, periodontium, tongue, oral mucosa, and surrounding tissues, as well as TMJ, sets dentistry apart from other medical fields.[14,27–29] Based on these differences, it is rational to investigate whether distinct effect size thresholds exist in dentistry, as observed in other medical fields.[5,6,30]

This issue is of particular concern in the context of public health. The World Health Organization (WHO) has noted a strong relationship between socioeconomic status and the prevalence and severity of oral diseases. This connection has been observed across various populations, ranging from childhood to advanced age.[31]

Dental diseases affect a significant proportion of the population. The global prevalence of dental caries in primary teeth among children is 46%, while the prevalence of caries in permanent teeth among children reaches 54%.[32] Periodontal disease in adults is estimated to impact around 62% of the population, with severe periodontitis occurring in 24%.[33] Approximately 22% of individuals experience edentulism.[34] Sleep bruxism is present in 21% of the population, while daytime bruxism afflicts 23%.[35] Temporomandibular disorders affect 34% of the population, and it is projected that by 2050 this figure will rise to 44%.[36]

Cleft palate has been diagnosed in 33% and cleft lip in 30% of cases involving cleft conditions, with cleft lip and palate occurring approximately once in every 1,000 live births.[37] Cancers of the lip, oral cavity, and pharynx account for about 4% of all cancer cases and 4% of all cancer-related deaths worldwide.[38] In the past decade, noma has been diagnosed in at least 23 countries.[39] Oro-dental trauma affects about 20% of children.[31] These are just a few examples of conditions and disorders associated with dentistry. This highlights the importance of improving research methods, including statistical approaches, within this field.

Considering the abovementioned information, a study was conducted to establish novel effect size thresholds (Cohen's d, Hedges' g and Pearson's r) for research in dentistry.

Material and methods

The project was initially registered with the Open Science Framework (OSF).[40]

The search procedure was replicated in accordance with the methodology outlined by Brydges.[6] Ten journals were searched: Journal of Dental Research (ISSN 0022-0345); Journal of Endodontics (ISSN 0099-2399); Dental Materials (ISSN 0109-5641); International Endodontic Journal (ISSN 0143-2885); Journal of Dentistry (ISSN 0300-5712); Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology (ISSN 2212-4403); Journal of the American Dental Association (ISSN 0002-8177); Community Dentistry and Oral Epidemiology (ISSN 0301-5661); Caries Research (ISSN 0008-6568); and Journal of Oral Rehabilitation (ISSN 0305-182X). The identification of these journals was conducted using the Scimago Journal & Country Rank database,[41] with a selection of the "dentistry (miscellaneous)" category and a sorting method based on the highest H-index over the entire period.[2,6,12,14,42] The list of journals was created at the beginning of the project on August 12, 2024.[40]

Considering the continuous development of dentistry, the search period was constrained to the last 20 years, a decision informed by prior studies.[12,14,35,43,44] Articles published between December 31, 2003, and December 31, 2023, were reviewed. The search focused on studies containing the term "meta" in the title during the specified timeframe. The following types of articles were excluded from the analysis: editorials; corrections; correspondence; short communications; conference abstracts; and reviews that did not involve meta-analyses, such as systematic reviews, narrative reviews and scoping reviews. Subsequently, full-text articles were analyzed.

A database similar to the one created by Brydges[6] was constructed, containing the Digital Object Identifier (DOI) numbers of the meta-analyses, along with information on study category, authors, publication year, sample size, and effect size. A total of 4,250 records were screened, and 567 meta-analyses were included for full-text analysis. In 326 publications, none of the studied effect sizes (Cohen's *d*, Hedges' *g*, or Pearson's *r*) were reported. Individual effect sizes were not specified in 89 studies. In 17 studies, data was missing (e.g., sample size explicitly tied to the effect size was not available). Ultimately, 135 meta-analyses were included in the analysis. Comprehensive details regarding the studies, the reasons for exclusion, and the number of included studies per journal are provided in the supplementary materials.

Statistical analysis

In the current study, 2 types of analyses were conducted: studies estimating effects within a group over time (test–retest); and studies evaluating differences between 2 groups. For the within-group analyses, the effect size was measured using the Pearson's *r* correlation coefficient, while for the between-group analyses, the effect size was quantified using Hedges' *g*.

The evaluation of effect sizes was based on Cohen's convention for small, medium and large effects. For the calculation of correlation coefficients, the thresholds were set at 0.10, 0.30 and 0.50, respectively. For between-group differences, the corresponding thresholds were 0.20, 0.50 and 0.80.[9]

The distribution of effect sizes was made by calculating a range of percentiles for both Pearson's *r* and Hedges' *g*. In line with previous literature,[6,30,45] the 25th, 50th (median) and 75th percentiles were interpreted as approximate indicators of small, medium and large effects according to Cohen's guidelines.[9,46] It should be noted, however, that this comparison is conceptual and does not assume that the underlying distribution of effect sizes perfectly aligns with Cohen's benchmarks. This convention does not imply that the distribution of effect sizes in the current data was symmetric.

Additionally, percentiles were determined for 2 subsamples of Hedges' *g* effect sizes, with studies classified into biopsychosocial, diagnosis, health promotion and prevention, and treatment categories according to the research focus of the meta-analysis. Furthermore, to account for the specificity of dental research, an additional division into 7 descriptive subgroups was made: cariology; periodontology; fixed and removable prosthodontics; oral surgery; orthodontics; endodontics; and conservative dentistry. These subgroup analyses were exploratory in nature and aimed to provide a descriptive overview of effect size distributions across research domains. No inferential statistical comparisons were performed between the subgroups; hence, no adjustments were applied for multiple comparisons.

To assess potential inflation bias, one-directional contour-enhanced funnel plots were generated. In these plots, effect sizes are plotted against their corresponding standard errors, with added contour regions representing key levels of statistical significance. Specifically, the orange-shaded region corresponds to the range of $0.10 > p > 0.05$, while the red-shaded region corresponds to $0.05 > p > 0.01$.[6,12,14] An excessive proportion of studies falling within these contours may indicate the presence of inflation bias, suggesting that the reported effect sizes could be overestimates of the true effect sizes. Such inflation may result from factors such as sampling error, publication bias or *p*-hacking. These funnel plots serve as a diagnostic tool to identify potential biases in the reported data.

A series of a priori power analyses were conducted to determine the sample sizes required for future research to achieve various levels of statistical power for both within-group and between-group differences, including biomedical and psychosocial subsamples. For within-group differences, calculations were based on

correlation analyses, while for between-group differences, calculations assumed a two-sample comparison with equal group sizes.

All analyses utilized a two-tailed alpha level of 0.05 and estimated the sample sizes necessary to achieve power levels of 60%, 70%, 80%, and 90% for small, medium and large effect sizes, corresponding to the 25th, 50th and 75th percentiles of the observed effect size distributions, respectively. These calculations provide critical benchmarks for designing adequately powered future studies.[6,12,14]

The analyses were conducted using the R programming language (v. 4.3.3; R Foundation for Statistical Computing, Vienna, Austria) on a Windows 11 Pro 64-bit operating system (build 22631; Microsoft Corporation, Redmond, USA). A comprehensive description of the statistical analysis, including the use of packages in the R language, the estimation of effect sizes for individual studies with group differences, the estimation of variance for Hedges' *g*, and the random-effects model, is provided in the supplementary material 2.

Results

Characteristics of the sample

The analysis encompassed a total of 4,250 dentistry studies, which were categorized into 4 research domains: biopsychosocial (n = 127, 2.99%); diagnosis (n = 796, 18.73%); health promotion and prevention (n = 271, 6.38%); and treatment (n = 3,056, 71.91%). Two types of effect sizes were utilized in the studies: those measuring between-group effects (Hedges' g, n = 4,038, 95.01%), and those measuring within-group effects (Pearson's r, n = 212, 4.99%). The median group sizes ranged from 20 to 24, with an interquartile range of 12–45. The complete database of publications used in the analyses is available in the supplementary material 3.

Within-group differences

The first (25%), second (50%) and third (75%) percentiles for within-group differences research corresponded to Pearson's r values of 0.16, 0.40 and 0.67, respectively (Table 1,2). This finding indicates that, in dentistry research focused on individual differences, the median effect size is Pearson's r = 0.40.

The observed effect sizes were noticeably higher than those in Cohen's guidelines for small, medium and large effects, with differences ranging from 0.06 for small effects to 0.17 for effects classified as large (Table 2). Compared to Cohen's benchmarks, only 64.2% of the observed correlations would qualify as medium effects or stronger ($r \geq 0.30$), and just 34.0% would be classified as strong effects ($r \geq 0.50$).

Table 1. Percentiles associated with observed within-group correlations (Pearson's r) and between-group differences (Hedges' g)

Percentile	Pearson's r	Hedges' g
5th	0.02	0.01
10th	0.05	0.03
15th	0.08	0.05
20th	0.12	0.08
25th	0.16	0.10
30th	0.21	0.14
35th	0.28	0.20
40th	0.33	0.25
45th	0.35	0.32
50th	0.40	0.35
55th	0.44	0.48
60th	0.46	0.58
65th	0.50	0.69
70th	0.53	0.84
75th	0.67	0.86
80th	0.83	1.35
85th	0.89	1.80
90th	0.92	2.64
95th	0.95	4.36

The distributions of effect sizes for within-group and between-group differences in Fig. 1A and Fig. 1B are reported with 25th, 50th and 75th percentiles corresponding to small (r = 0.16, g = 0.10), medium (r = 0.40, g = 0.35) and large (r = 0.67, g = 0.86) effects, respectively. This indicates that the majority of observed relationships in dentistry research on individual differences are of medium to large size, suggesting clinically meaningful associations in this domain. Additionally, it is important to emphasize the differences observed in the domains of dentistry: in oral surgery, the small effect was 0.08, the medium effect was 0.27 and the large effect was 0.66; in orthodontics, the respective values were 0.40, 0.93 and 1.87; in periodontology, the values were 0.11, 0.29 and 0.63; in cariology – 0.10, 0.40 and 1.01; in conservative dentistry – 0.10,

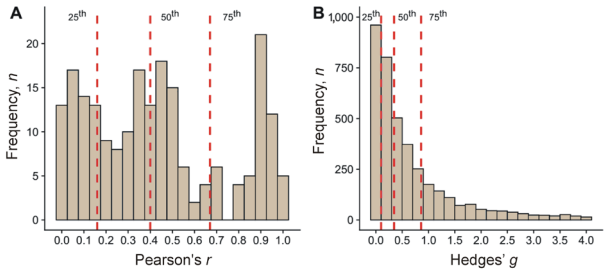

Fig. 1. Distribution of Pearson's r (A) and Hedges' g (B) effect sizes for within-group and between-group differences

Dashed red lines represent the 25th, 50th and 75th percentiles corresponding to small (r = 0.16, g = 0.10), medium (r = 0.40, g = 0.35) and large (r = 0.67, g = 0.86) effect sizes.

Table 2. Comparison of Cohen's guidelines with quantitatively derived estimates of effect sizes

Characteristic			Studies, *n*	Effect size		
				small	medium	large
Individual differences (Pearson's *r*)	Cohen[9]		–	0.10	0.30	0.50
	current study	obtained values	212	0.16	0.40	0.67
		rounded values		0.20	0.40	0.70
	category	diagnosis	44	0.29	0.50	0.85
		health promotion and prevention	87	0.06	0.17	0.41
		treatment	81	0.35	0.47	0.89
		oral surgery	81	0.35	0.47	0.89
		cariology	87	0.06	0.17	0.41
		conservative dentistry	31	0.23	0.40	0.52
Group differences (Hedges' *g*)	Cohen[9]		–	0.20	0.50	0.80
	current study	obtained values	4,038	0.10	0.35	0.86
		rounded values		0.10	0.40	0.90
	category	biopsychosocial	127	0.05	0.14	0.36
		diagnosis	752	0.05	0.18	0.51
		health promotion and prevention	184	0.09	0.27	0.80
		treatment	2975	0.15	0.50	1.29
		oral surgery	1274	0.08	0.27	0.66
		orthodontics	199	0.40	0.93	1.87
		periodontology	474	0.11	0.29	0.63
		cariology	480	0.10	0.40	1.01
		conservative dentistry	282	0.10	0.31	0.73
		endodontics	176	0.04	0.19	0.77
		fixed and removable prosthodontics	517	0.56	1.54	3.35
		temporomandibular joint and masticatory muscle research*	456	0.10	0.30	0.70

* data obtained from the study by Zieliński and Gawda.[14]

0.31 and 0.73; in endodontics – 0.04, 0.19 and 0.77; and in fixed and removable prosthodontics – 0.56, 1.54 and 3.35 (Table 2). It should be noted that most of the effects for between-group differences were below the thresholds recommended by Cohen.

The median sample size for within-group differences was 117 participants. This sample size is large enough to detect a medium (r = 0.40; power = 1.00) or large effect (r = 0.67; power = 1.00), but not to detect a small effect (r = 0.16; power = 0.41).

A visual assessment of the distribution of effect sizes was prepared to evaluate potential publication bias and the symmetry of the effect size distributions within each domain. Contour lines indicate the regions of statistical significance (supplementary material 2: Fig. 2).

A total of 70.95% of the studies demonstrated sufficient power to detect a medium effect, as indicated by their distribution within the gray region of the contour-enhanced funnel plot (supplementary material 2: Fig. 2A,3) (Table 3), corresponding to $p < 0.01$. Furthermore, the funnel plot did not exhibit an overrepresentation of just-significant results (p-values: 0.05–0.01, represented by the red region) or marginally significant results (p-values: 0.10–0.05, represented by the orange region). This pattern indicates that inflation bias, including potential concerns such as publication bias or p-hacking, is unlikely to pose a significant issue in dentistry studies investigating individual differences.

The sample size calculations presented in Table 4 provide critical benchmarks for designing future studies in individual differences research. Achieving adequate statistical power necessitates the determination of the required sample size, which varies substantially depending on the effect size and the desired power level. For small effects (r = 0.16), achieving 80% power requires a sample size of 304, which increases to 406 for 90% power, indicating the need for larger samples to reliably detect subtle effects. For medium effects (r = 0.40), a sample size of 46 is sufficient for 80% power, while for large effects (r = 0.67), only 15 participants are needed to achieve the same power level.

Only 62% of the analyzed studies were adequately powered to detect a medium effect, and nearly 90% were powered to identify a large effect.

Table 3. Distribution of studies across funnel plot color regions based on the research domain and type of comparison

Comparison	Category	Color region [%]			
		white ($p > 0.10$)	orange ($0.10 \geq p > 0.05$)	red ($0.05 \geq p > 0.01$)	gray ($p \leq 0.01$)
Within-group differences	overall	18.09	3.33	7.62	70.95
	diagnosis	4.65	0.00	9.30	86.00
	health promotion and prevention	18.60	3.49	3.49	74.42
	treatment	24.69	4.94	11.11	59.26
	oral surgery	24.69	4.94	11.11	59.26
	cariology	18.60	3.49	3.49	74.42
	conservative dentistry	6.67	0.00	13.33	80.00
Between-group differences	overall	56.71	5.19	8.00	30.09
	treatment	50.45	4.90	7.61	37.04
	health promotion and prevention	55.49	6.59	6.59	31.32
	diagnosis	61.82	4.62	7.34	26.22
	biopsychosocial	68.03	3.28	8.20	20.49
	oral surgery	28.08	4.93	11.33	55.66
	orthodontics	64.76	5.01	7.83	22.39
	periodontology	62.74	5.51	4.94	26.81
	cariology	46.77	4.92	8.00	40.31
	conservative dentistry	61.64	6.51	6.16	25.69
	endodontics	62.23	3.12	12.45	22.23
	fixed and removable prosthodontics	28.33	3.87	6.19	61.61

Between-group differences

In the between-group differences sample, the 25th, 50th and 75th percentiles corresponded to Hedges' g values of 0.10, 0.35 and 0.86, respectively (Table 1,2). For small and medium effects, these values are lower than Cohen's benchmarks of 0.20 and 0.50,[9,46] while for large effects, the 75th percentile exceeds Cohen's guideline of 0.80. A comparison of these results with Cohen's recommendations reveals that only 40.4% of the observed effect sizes in this sample would qualify as medium or stronger effects ($g \geq 0.50$), and just 27% would be considered large ($g \geq 0.80$). This finding indicates that a substantial proportion of the observed group differences reflects smaller-than-expected effects, based on established guidelines.

An examination of specific research domains revealed significant variability. In biopsychosocial studies, the derived thresholds for small ($g = 0.05$), medium ($g = 0.14$) and large ($g = 0.36$) effects are substantially smaller than those reported in Cohen's guidelines, indicating that even the modest effects within this domain hold practical significance. Similarly, diagnostic studies show lower thresholds for small ($g = 0.05$) and medium ($g = 0.18$) effects, with large effects ($g = 0.51$) aligning more closely with Cohen's recommendations.

Health promotion and prevention studies have demonstrated thresholds for small, medium and large effects

Table 4. Distribution of sample sizes required to achieve various levels of statistical power in research on within-group differences

Category	Effect size	Statistical power			
		60%	70%	80%	90%
All studies ($N = 212$)	small ($r = 0.16$)	191	240	304	406
	medium ($r = 0.40$)	30	37	46	61
	large ($r = 0.67$)	10	12	15	19
Diagnosis ($n = 44$)	small ($r = 0.29$)	56	70	89	118
	medium ($r = 0.50$)	18	22	27	36
	large ($r = 0.85$)	6	6	7	9
Health promotion and prevention ($n = 87$)	small ($r = 0.06$)	1,360	1,712	2,177	2,914
	medium ($r = 0.17$)	168	212	269	359
	large ($r = 0.41$)	28	35	44	58
Treatment ($n = 81$)	small ($r = 0.35$)	39	48	61	81
	medium ($r = 0.47$)	21	26	32	42
	large ($r = 0.89$)	5	6	7	8
Oral surgery ($n = 81$)	small ($r = 0.35$)	39	48	61	81
	medium ($r = 0.47$)	21	26	32	42
	large ($r = 0.89$)	5	6	7	8
Cariology ($n = 87$)	small ($r = 0.06$)	1,360	1,712	2,177	2,914
	medium ($r = 0.17$)	168	212	269	359
	large ($r = 0.41$)	28	35	44	58
Conservative dentistry ($n = 31$)	small ($r = 0.23$)	88	110	139	185
	medium ($r = 0.40$)	30	37	46	61
	large ($r = 0.52$)	17	21	26	34

Table 5. Distribution of sample sizes required to achieve various levels of statistical power in research on between-group differences

Category	Effect size	Statistical power			
		60%	70%	80%	90%
All studies (N = 4,038)	small (g = 0.10)	1,017	1,290	1,628	2,178
	medium (g = 0.35)	84	105	133	177
	large (g = 0.86)	15	18	23	30
Biopsychosocial (n = 752)	small (g = 0.05)	3,927	4,947	6,292	8,422
	medium (g = 0.14)	507	638	811	1,086
	large (g = 0.36)	79	99	126	168
Diagnosis (n = 184)	small (g = 0.05)	3,927	4,947	6,292	8,422
	medium (g = 0.18)	294	370	471	629
	large (g = 0.51)	39	49	61	82
Health promotion and prevention (n = 184)	small (g = 0.09)	1,354	1,705	2,168	2,902
	medium (g = 0.27)	139	175	222	297
	large (g = 0.80)	17	21	26	34
Treatment (n = 2795)	small (g = 0.15)	454	572	727	973
	medium (g = 0.50)	41	51	64	85
	large (g = 1.29)	7	9	11	14
Oral surgery (n = 1274)	small (g = 0.08)	62	78	99	133
	medium (g = 0.27)	12	15	19	25
	large (g = 0.66)	4	5	6	7
Orthodontics (n = 199)	small (g = 0.40)	1,655	2,085	2,652	3,550
	medium (g = 0.93)	138	173	220	294
	large (g = 1.87)	24	30	37	50
Periodontology (n = 474)	small (g = 0.11)	824	1,039	1,320	1,767
	medium (g = 0.93)	117	147	187	250
	large (g = 0.63)	26	32	41	55
Cariology (n = 480)	small (g = 0.10)	904	1,139	1,448	1,938
	medium (g = 0.40)	62	78	98	131
	large (g = 1.01)	11	13	16	22
Conservative dentistry (n = 282)	small (g = 0.10)	965	1,215	1,545	2,068
	medium (g = 0.31)	106	133	169	226
	large (g = 0.73)	20	25	31	41
Endodontics (n = 176)	small (g = 0.04)	7,151	9,013	11,462	15,345
	medium (g = 0.19)	275	346	439	589
	large (g = 0.77)	18	22	28	37
Fixed and removable prosthodontics (n = 517)	small (g = 0.56)	23	40	51	68
	medium (g = 1.54)	5	6	8	9
	large (g = 3.35)	2	3	3	3
Temporomandibular joint and masticatory muscle research* (n = 456)	small (g = 0.10)	1,020	1,280	1,630	2,180
	medium (g = 0.30)	80	100	130	180
	large (g = 0.70)	14	17	20	30

* data obtained from the study by Zieliński and Gawda.[14]

(g = 0.09, g = 0.27 and g = 0.80, respectively) that are closer to Cohen's benchmarks, particularly for large effects. Treatment studies have revealed thresholds (g = 0.15, g = 0.50 and g = 1.29, respectively) that closely align with or exceed Cohen's benchmarks, especially for large effects.

A visual representation of the variation in effect sizes within each category highlighted the differences in the distribution of study outcomes (supplementary material 2: Fig. 4,5). The treatment category demonstrates a wide distribution of effect sizes, with a peak around moderate values of Hedges' g and a noticeable tail extending into higher effect sizes. The health promotion and prevention category shows a narrower distribution, with the majority of effect sizes clustering around smaller to moderate values. The diagnosis domain exhibits a sharply peaked distribution, concentrated around smaller effect sizes, with a steep decline as the values increase. The biopsychosocial category has a similarly narrow distribution, with most studies reporting smaller effect sizes and a small proportion extending to moderate values.

The median sample sizes for the case and control groups were 24 and 20 participants, respectively. These sample sizes are insufficient to reliably detect a large (g = 0.86; power = 0.79), medium (g = 0.35; power = 0.20), or small effect (g = 0.10; power = 0.06). Notably, only 6% of the studies included in the analysis were adequately powered to detect a medium effect (g = 0.35) with a statistical power of 0.80. This finding highlights a critical limitation in the statistical power of most studies, emphasizing the need for larger sample sizes in future research to ensure the robustness and reliability of findings.

The data presented in Table 3 further supports the conclusion that inflation bias is unlikely to have a significant impact on dentistry studies that investigate group differences. Across all studies, only 5.19% of results fall within the orange region (marginally significant results: $0.10 \geq p > 0.05$), and 8.00% fall within the red region (just-significant results: $0.05 \geq p > 0.01$). A similar pattern is observed across specific research categories. Treatment studies indicated 4.90% of results in the orange region and 7.61% in the red region. Health promotion and prevention studies showed 6.59% in both regions. Diagnosis studies demonstrated 4.62% in the orange region and 7.34% in the red region. Finally, biopsychosocial studies indicated 3.28% in the orange region and 8.20% in the red region. The relatively low proportion of results in these regions, combined with the high percentage of robustly significant findings in the gray region ($p < 0.01$), suggests that inflation bias, including publication bias or p-hacking, is unlikely to be a major concern in these studies.

The sample size requirements presented in Table 5 provide insights into the feasibility of achieving adequate statistical power in between-group differences research across various dentistry domains.

For all studies combined, detecting small effects (g = 0.10) with 80% power requires substantial sample sizes (n = 1,628), while medium (g = 0.35) and large (g = 0.86) effects require significantly fewer participants (n = 133 and n = 23, respectively). This underscores the challenge of reliably detecting small effects, which necessitate much larger sample sizes compared to medium or large effects.

When examining specific dentistry domains, considerable variability in sample size requirements is evident. In the context of biopsychosocial studies, detecting small effects (g = 0.05) with 80% power demands an extremely large sample size (n = 6,292), while medium (g = 0.14) and large (g = 0.36) effects require 811 and 126 participants, respectively. Similarly, diagnosis studies require large samples to detect small effects (g = 0.05; n = 6,292), with moderate reductions for medium (g = 0.18; n = 471) and large effects (g = 0.51; n = 61). These results highlight the difficulty of achieving sufficient power in studies that target small effects within these domains.

In health promotion and prevention studies, the sample size requirements are comparatively moderate. The detection of small effects (g = 0.09) necessitates a sample size of 2,168 individuals to achieve 80% power, while medium (g = 0.27; n = 222) and large (g = 0.80; n = 26) effects are more easily achievable. Treatment studies, in contrast, have demonstrated the most favorable sample size requirements. For small effects (g = 0.15), a sample of 727 participants is required to attain 80% power, while medium (g = 0.50; n = 64) and large (g = 1.29; n = 11) effects require considerably smaller samples.

Additionally, it is important to observe how sample size requirements vary across different categories of dentistry. For example, to detect small effects with 60% power (g = 0.08) in oral surgery, a sample size of 62 participants is needed. Within the same category, detecting large effects (g = 0.66) would require only 6 individuals. However, under the same assumptions (small effect and 60% power), cariology would require a sample of 904 patients, while detecting large effects in this category would necessitate a study sample of 11 participants. In each of the presented categories of dentistry, the results highlight the difficulty of achieving sufficient statistical power in studies targeting small effects in these areas (Table 5).

For large effects with 90% power, a distinct picture emerges. In oral surgery (g = 0.66), a sample size of 7 individuals is needed; in orthodontics (g = 1.87) – 50; in periodontology (g = 0.63) – 55; in cariology (g = 1.01) – 22; in conservative dentistry (g = 0.73) – 41; in endodontics (g = 0.77) – 37; and in fixed and removable prosthodontics (g = 3.35) – 3.

These findings demonstrate that conducting studies focused on detecting large effects is highly feasible for researchers within each category of dentistry.

Discussion

The growing significance of dental diseases and the increasing proportion of affected individuals is evident. Beyond the percentage-based data, this trend is also reflected in the rising number of scientific publications focused on dental research, as well as in the specific nature of the discipline itself.

The aim of the study was to establish novel, data-driven thresholds for effect sizes (Cohen's d, Hedges' g and Pearson's r) relevant to dental research, rather than relying on general, arbitrary benchmarks that may not adequately reflect the specific characteristics of the field. Additionally, the study offers guidance on the minimum required sample size, contingent upon statistical power. The inclusion of information regarding sample size and effect size calculations in standardized sections of research papers constitutes a key component of transparent reporting.[2]

It is important to acknowledge that, while Cohen's benchmarks serve as a useful comparative tool, their application should not be done without careful consideration of the clinical context.[12,47,48] Cohen's thresholds are arbitrary and fail to account for clinical relevance, domain-specific nuances or individual patient needs.

For this reason, researchers are encouraged to explore alternative approaches and to consider effect size as part of a broader clinical evaluation process, rather than as a definitive indicator of an intervention's value. The clinical significance of a change is not solely determined by its effect size. As Sullivan and Feinn have observed, p-values indicate statistical significance, whereas effect sizes convey the magnitude of the difference.[7] However, it is only within a clinical context that one can assess whether a change holds real value for the patient.[7]

Therefore, when interpreting results, it is essential to consider p-values, effect sizes, patient-reported outcomes, functional performance, and clinical judgment collectively. It is crucial not to rely solely on numerical indicators. Clinical relevance should emerge from a comprehensive analysis that accounts for individual needs, therapeutic decisions, treatment conditions, and the patient's quality of life. From this perspective, the new effect size thresholds do not replace clinical judgment but are intended to serve as a tool to facilitate the interpretation of findings.[7,12,47,48]

The results of this analysis indicate that the majority of observed effect sizes in dental research deviate substantially from the thresholds proposed by Cohen. In particular, the majority of the effects were smaller than Cohen's benchmarks, which calls into question the validity of using Cohen's thresholds as reference points in the field of dentistry.

In the present study, it was observed that for Pearson's r, values of 0.16 (≈ 0.20) represented small effects, 0.40 indicated medium effects and 0.67 (≈ 0.70) corresponded to large effects. For Hedges' g, the established thresholds were 0.10, 0.35 (≈ 0.40) and 0.86 (≈ 0.90). Calculations were also performed separately for individual domains within dentistry, such as oral surgery, orthodontics, periodontology, cariology, conservative dentistry, endodontics, and both fixed and removable prosthodontics (Table 3).

With regard to within-group differences (Pearson's r), Cohen's original thresholds are inadequate for research in dentistry. Our results also exceed those reported by

Gignac and Szodorai for psychological studies,[49] Brydges' estimates in gerontology research[6] and Zieliński for physiotherapy.[12] When comparing the effect sizes obtained in the present study for Hedges' *g* (0.10, 0.40 and 0.90), it can be observed that the thresholds for small effects are consistent with those established for TMJ and masticatory muscle research.[14] However, a discrepancy in the values for medium and large effects is evident. In the broader field of dentistry, medium and large effect size thresholds are elevated by 0.10 and 0.20, respectively. This highlights the specificity of the discipline under investigation.

A significant observation presented in Tables 4 and 5 highlights their value as a framework for planning future studies in individual differences research. The minimum sample size requirements to ensure adequate statistical power vary considerably depending on effect size and the desired power levels. For small effects (r = 0.16), achieving 80% power requires a sample size of 304, increasing to 406 for 90% power. This underscores the need for larger samples to ensure reliable detection of small effects. In contrast, medium effects (r = 0.40) require 46 participants for 80% power, while large effects (r = 0.67) require just 15 participants to achieve the same power level. Tables 4 and 5 provide practical guidelines on the appropriate sample size needed for dental studies across the aforementioned fields of dentistry, based on specific assumptions regarding statistical power and effect size.

Limitations

This study has several limitations that should be acknowledged. First, the investigation was restricted to meta-analyses that were published over a 20-year period. Although this temporal constraint may limit the study's scope, it aligns with the dynamic nature of dental and medical research and reflects current developments in the field.[12,14,49] A key limitation is the potential for systematic biases, such as publication bias, sampling error, and questionable research practices (e.g., *p*-hacking), which may distort the distribution and interpretation of effect sizes.[6,50,51] These risks have been extensively documented in meta-research and are acknowledged in similar studies.[5,6,12,14,49] The study relied solely on published data, assuming that the original authors applied appropriate statistical methods. While this is considered standard practice, there is a risk that the included studies may have failed to meet methodological standards.[5,6,12,14,49] On the other hand, the relatively large sample size strengthens the robustness and generalizability of the findings in comparison to prior studies.[6,14]

In conclusion, the present study proposes updated, empirically-based effect size thresholds for dental research, grounded in discipline-specific data rather than arbitrary general values. These thresholds are not intended to replace clinical evaluation; rather, they are designed to serve as a tool that enhances the interpretation of the results, reporting transparency, and the planning of future studies. The clinical relevance of findings should be assessed by integrating statistical data with patient impact, expert judgment, and the broader healthcare context.

Conclusions

Based on these findings, researchers in the field of dentistry are encouraged to adopt the following thresholds: for Pearson's r, 0.20 for small effects, 0.40 for medium effects and 0.70 for large effects; and for Cohen's d or Hedges' g, 0.10 for small effects, 0.40 for medium effects and 0.90 for large effects. These updated thresholds have the potential to improve the rigor and quality of dental research, ultimately benefiting patients through enhanced diagnostics and treatment strategies.

Ethics approval and consent to participate

Not applicable.

Data availability

The data related to this article, including supplementary materials, can be accessed in the Open Science Framework (OSF) database via the following link: https://osf.io/9fghx/files. The script used in the analysis is available from the corresponding author upon reasonable request.

Consent for publication

Not applicable.

Use of AI and AI-assisted technologies

Not applicable.

ORCID iDs

Grzegorz Zieliński https://orcid.org/0000-0002-2849-0641
Mieszko Więckiewicz https://orcid.org/0000-0003-4953-7143

References

1. Taşkın Z. Forecasting the future of library and information science and its sub-fields. *Scientometrics*. 2021;126(2):1527–1551. doi:10.1007/s11192-020-03800-2
2. Zieliński G, Gawda P. Analysis of the use of sample size and effect size calculations in a temporomandibular disorders randomised controlled trial – short narrative review. *J Pers Med*. 2024;14(6):655. doi:10.3390/jpm14060655
3. Chu B, Liu M, Leas EC, Althouse BM, Ayers JW. Effect size reporting among prominent health journals: A case study of odds ratios. *BMJ Evid Based Med*. 2020;26(4):184. doi:10.1136/bmjebm-2020-111569
4. Bothe AK, Richardson JD. Statistical, practical, clinical, and personal significance: Definitions and applications in speech-language pathology. *Am J Speech Lang Pathol*. 2011;20(3):233–242. doi:10.1044/1058-0360(2011/10-0034)
5. Gaeta L, Brydges CR. An examination of effect sizes and statistical power in speech, language, and hearing research. *J Speech Lang Hear Res*. 2020;63(5):1572–1580. doi:10.1044/2020_JSLHR-19-00299

6. Brydges CR. Effect size guidelines, sample size calculations, and statistical power in gerontology. *Innov Aging*. 2019;3(4):igz036. doi:10.1093/geroni/igz036
7. Sullivan GM, Feinn R. Using effect size—or why the *p* value is not enough. *J Grad Med Educ*. 2012;4(3):279–282. doi:10.4300/JGME-D-12-00156.1
8. Wilkinson L. Statistical methods in psychology journals: Guidelines and explanations. *Am Psychol*. 1999;54(8):594–604. doi:10.1037/0003-066X.54.8.594
9. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. New York, NY: Routledge; 1988. doi:10.4324/9780203771587
10. Hemphill JF. Interpreting the magnitudes of correlation coefficients. *Am Psychol*. 2003;58(1):78–79. doi:10.1037/0003-066X.58.1.78
11. Kinney AR, Eakman AM, Graham JE. Novel effect size interpretation guidelines and an evaluation of statistical power in rehabilitation research. *Arch Phys Med Rehabil*. 2020;101(12):2219–2226. doi:10.1016/j.apmr.2020.02.017
12. Zieliński G. Effect size guidelines for individual and group differences in physiotherapy. *Arch Phys Med Rehabil*. 2025:S0003-9993(25)00717-8. doi:10.1016/j.apmr.2025.05.013
13. Swinton PA, Shim JSC, Pavlova AV, et al. What are small, medium and large effect sizes for exercise treatments of tendinopathy? A systematic review and meta-analysis. *BMJ Open Sport Exerc Med*. 2023;9(1):e001389. doi:10.1136/bmjsem-2022-001389
14. Zieliński G, Gawda P. Defining effect size standards in temporomandibular joint and masticatory muscle research. *Med Sci Monit*. 2025;31:e948365. doi:10.12659/MSM.948365
15. Hackenberg B, Schlich MN, Gouveris H, et al. Medical and dental students' perception of interdisciplinary knowledge, teaching content, and interprofessional status at a German University: A cross-sectional study. *Int J Environ Res Public Health*. 2022;20(1):428. doi:10.3390/ijerph20010428
16. Erdilek D, Gümüştaş B, Güray Efes B. Digitalization era of dental education: A systematic review. *Dent Med Probl*. 2023;60(3):513–525. doi:10.17219/dmp/156804
17. Sedky RAF, Ben Dor B, Mustafa DS, et al. Self-assessment skills of undergraduate students in operative dentistry: Preclinical performance and gender. *Dent Med Probl*. 2024. doi:10.17219/dmp/175276
18. Stulginskiene S, Abalikstaite J, Gendviliene I, et al. Importance of education on infection control and on the hand skin health of dental personnel. *Dent Med Probl*. 2022;59(3):373–379. doi:10.17219/dmp/142563
19. Kachabian S, Seyedmajidi S, Tahani B, Naghibi Sistani MM. Effectiveness of educational strategies to teach evidence-based dentistry to undergraduate dental students: A systematic review. *Evid Based Dent*. 2024;25(1):53–54. doi:10.1038/s41432-023-00958-5
20. Spielman AI. Dental education and practice: Past, present, and future trends. *Front Oral Health*. 2024;5:1368121. doi:10.3389/froh.2024.1368121
21. MacNeil RLM, Hilario H. Input from practice: Reshaping dental education for integrated patient care. *Front Oral Health*. 2021;2:659030. doi:10.3389/froh.2021.659030
22. Lobbezoo F, Aarab G. Medicine and dentistry working side by side to improve global health equity. *J Dent Res*. 2022;101(10):1133–1134. doi:10.1177/00220345221088237
23. Almulhim KS, Rehman SU, Ali S, Ahmad S, Khan AS. Bibliometric analysis of the current status and trends in dental applications of glass fiber-reinforced composites from 1998 to 2022. *Dent Med Probl*. 2024;61(5):783–795. doi:10.17219/dmp/171803
24. Jurado CA, Villalobos-Tinoco J, Alshabib A, Afrashtehfar KI. Advanced restorative management of focal microdontia: A brief review and case report. *Dent Med Probl*. 2024;61(3):457–464. doi:10.17219/dmp/158834
25. Ansari G, Toomarian L, Masoum T, Shayeghi S, Eftekhar L. Evaluation of the sedative effect of intranasal versus intramuscular ketamine in 2–6-year-old uncooperative dental patients. *Dent Med Probl*. 2024;61(1):35–41. doi:10.17219/dmp/144364
26. Woźniak-Budych MJ, Staszak M, Staszak K. A critical review of dental biomaterials with an emphasis on biocompatibility. *Dent Med Probl*. 2023;60(4):709–739. doi:10.17219/dmp/172732
27. Gombra V, Kaur M, Hasan S, Mansoori S. Smokeless tobacco- and quid-associated localized lesions of the oral cavity: A cross-sectional study from a dental institute. *Dent Med Probl*. 2024;61(5):687–696. doi:10.17219/dmp/152439
28. Murad M, Al-Maslamani L, Yates J. Removal of mandibular third molars: An overview of risks, a proposal for international community and guidance. *Dent Med Probl*. 2024;61(4):481–488. doi:10.17219/dmp/166156
29. Wadhwa J, Sethi S, Gupta A, Batra P, Lalfakawmi S. Is prevalence of dental anomalies site-specific in cleft lip and palate patients? A systematic review and meta-analysis. *Dent Med Probl*. 2025;62(1):125–133. doi:10.17219/dmp/170879
30. Lovakov A, Agadullina ER. Empirically derived guidelines for effect size interpretation in social psychology. *Eur J Soc Psychol*. 2021;51(3):485–504. doi:10.1002/ejsp.2752
31. World Health Organization (WHO). Oral health. https://www.who.int/news-room/fact-sheets/detail/oral-health. Accessed October 23, 2024.
32. Kazeminia M, Abdi A, Shohaimi S, et al. Dental caries in primary and permanent teeth in children's worldwide, 1995 to 2019: A systematic review and meta-analysis. *Head Face Med*. 2020;16(1):22. doi:10.1186/s13005-020-00237-z
33. Trindade D, Carvalho R, Machado V, Chambrone L, Mendes JJ, Botelho J. Prevalence of periodontitis in dentate people between 2011 and 2020: A systematic review and meta-analysis of epidemiological studies. *J Clin Periodontol*. 2023;50(5):604–626. doi:10.1111/jcpe.13769
34. Borg-Bartolo R, Roccuzzo A, Molinero-Mourelle P, et al. Global prevalence of edentulism and dental caries in middle-aged and elderly persons: A systematic review and meta-analysis. *J Dent*. 2022;127:104335. doi:10.1016/j.jdent.2022.104335
35. Zieliński G, Pająk A, Wójcicki M. Global prevalence of sleep bruxism and awake bruxism in pediatric and adult populations: A systematic review and meta-analysis. *J Clin Med*. 2024;13(14):4259. doi:10.3390/jcm13144259
36. Zieliński G. Quo vadis temporomandibular disorders? By 2050, the global prevalence of TMD may approach 44%. *J Clin Med*. 2025;14(13):4414. doi:10.3390/jcm14134414
37. Salari N, Darvishi N, Heydari M, Bokaee S, Darvishi F, Mohammadi M. Global prevalence of cleft palate, cleft lip and cleft palate and lip: A comprehensive systematic review and meta-analysis. *J Stomatol Oral Maxillofac Surg*. 2022;123(2):110–120. doi:10.1016/j.jormas.2021.05.008
38. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2021;71(3):209–249. doi:10.3322/caac.21660
39. Galli A, Brugger C, Fürst T, Monnier N, Winkler MS, Steinmann P. Prevalence, incidence, and reported global distribution of noma: A systematic literature review. *Lancet Infect Dis*. 2022;22(8):e221–e230. doi:10.1016/S1473-3099(21)00698-8
40. Zieliński G. Effect size guidelines, sample size calculations, and statistical power in dentistry. Published online August 12, 2024. doi:10.17605/OSF.IO/E9ZJU
41. SJR: Scientific Journal Rankings. https://www.scimagojr.com/journalrank.php. Accessed October 23, 2024.
42. Mondal H, Deepak KK, Gupta M, Kumar R. The h-Index: Understanding its predictors, significance, and criticism. *J Family Med Prim Care*. 2023;12(11):2531–2537. doi:10.4103/jfmpc.jfmpc_1613_23
43. Delli K, Livas C, Dijkstra PU. How has the dental literature evolved over time? Analyzing 20 years of journal self-citation rates and impact factors. *Acta Odontol Scand*. 2020;78(3):223–228. doi:10.1080/00016357.2019.1685681
44. Pitts NB, Banerjee A, Mazevet ME, Goffin G, Martignon S. From "ICDAS" to "CariesCare International": The 20-year journey building international consensus to take caries evidence into clinical practice. *Br Dent J*. 2021;231(12):769–774. doi:10.1038/s41415-021-3732-2
45. Quintana DS. Statistical considerations for reporting and planning heart rate variability case–control studies. *Psychophysiology*. 2017;54(3):344–349. doi:10.1111/psyp.12798

46. Cohen J. A power primer. *Psychol Bull*. 1992;112(1):155–159. doi:10.1037/0033-2909.112.1.155
47. Tagliaferri SD, Belavy DL, Fitzgibbon BM, et al. How to interpret effect sizes for biopsychosocial outcomes and implications for current research. *J Pain*. 2024;25(4):857–861. doi:10.1016/j.jpain.2023.10.014
48. Bogduk N. Calibrating effect-size for studies of pain treatment. *Interv Pain Med*. 2022;1(Suppl 2):100123. doi:10.1016/j.inpm.2022.100123
49. Gignac GE, Szodorai ET. Effect size guidelines for individual differences researchers. *Pers Individ Differ*. 2016;102:74–78. doi:10.1016/j.paid.2016.06.069
50. Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. The extent and consequences of *p*-hacking in science. *PLoS Biol*. 2015;13(3):e1002106. doi:10.1371/journal.pbio.1002106
51. Di Amari P, Banks G, Bourque L, Holladay H, O'Boyle E. Effect size benchmarks: Time for a causal renaissance. *Leadership Quart*. 2025;36(1):101855. doi:10.1016/j.leaqua.2024.101855