QUBAHAN JOURNAL OF COACHING AND SPORTS SCIENCES
VOL. 4, NO. 4, Oct 2026
https://doi.org/10.48161/qcsj.v4n4a90

Large-Scale Hypothesis Testing and Confidence Interval Estimation: An Effect-Size and Multiplicity-Aware Framework

Ibrahim Mallam Fali[1*], Eze Claris Nnenna[2]

[1] Accounting, Faculty of Administration and Management Sciences, University of Calabar, Calabar, Nigeria;
[2] Educational Management and Foundational Studies, Faculty of Education, Alex Ekwueme Federal University, Ndufu Alike Ikwo, Ebonyi, Nigeria;

* **Corresponding author:** faliibrahim7@gmail.com

ABSTRACT: This study develops and empirically evaluates an integrated framework for statistical hypothesis testing and confidence interval estimation in large-scale data analysis. Using the 2015 Behavioral Risk Factor Surveillance System (BRFSS) binary diabetes health-indicator file, the analysis examined 253,680 observations and 22 variables, including diabetes status, cardiometabolic risk indicators, lifestyle indicators, healthcare-access measures, and socioeconomic variables. The analysis used descriptive statistics, prevalence estimates, Welch mean-comparison tests, chi-square tests, confidence intervals, standardized effect sizes, odds ratios, risk differences, multiple-comparison correction, and adjusted logistic regression. Diabetes prevalence in the main file was 13.93% (95% CI, 13.80% to 14.07%). The results showed that the large sample size produced very small p-values for most comparisons; however, effect magnitudes varied substantially across variables. Body mass index (BMI) was higher among respondents with diabetes by 4.14 units, with a 95% confidence interval of 4.06 to 4.22 and a moderate standardized effect. High blood pressure showed a strong unadjusted association with diabetes (odds ratio 5.04; 95% CI, 4.91 to 5.17), while physical activity showed a lower odds pattern for diabetes. After adjustment, high blood pressure, high cholesterol, general health, age, and BMI remained important predictors. The study demonstrates that hypothesis testing is most reliable when p-values are interpreted together with confidence intervals, effect sizes, adjusted estimates, and multiplicity-aware decision rules. The framework provides a reproducible model to strengthen statistical interpretation in large empirical datasets.

Keywords: hypothesis testing; confidence intervals; large-scale data; effect size; estimation.

I. INTRODUCTION

Large-scale data analysis has become central to empirical research across health, business, education, economics, and social-science domains. As datasets grow, researchers gain greater numerical precision and a greater capacity to detect small differences among groups. However, the same advantage can create interpretive difficulties: very large samples can produce extremely small p-values even when the difference being tested is too small to be practically meaningful. Therefore, statistical inference in large datasets should not rely on a binary reading of significant or non-significant results. It requires a broader interpretive framework that combines hypothesis testing, confidence intervals, effect-size estimation, multiplicity control, and substantive judgment.

Hypothesis testing remains useful because it provides a formal procedure for assessing whether an observed pattern is compatible with a stated null hypothesis. Yet, a p-value alone does not measure a finding's importance, quantify the size of an association, or show uncertainty around an estimate. Confidence intervals are therefore essential because they present a plausible range for a parameter or effect. A narrow interval indicates greater precision, while its location and width help readers assess both statistical evidence and practical importance.

Large-scale datasets also introduce the multiple-comparisons problem. When many variables are tested against an outcome, the chance of obtaining at least one apparently important result increases even if some relationships are weak or incidental. This problem is especially important in applied datasets that contain several health, behavioral, demographic, or socioeconomic variables. In such cases, inferential quality depends not only on selecting an appropriate test, but also on transparently interpreting p-values and reporting effect sizes.

The present study uses the BRFSS2015 diabetes health-indicator file as a large-scale empirical case. This dataset is appropriate because it contains a large number of observations, a binary health outcome, several binary and ordinal predictors, and a mix of behavioral, clinical, and socioeconomic measures. These characteristics allow the study to demonstrate how different inferential tools can be combined within one coherent framework. The analysis does not aim to build a clinical diagnosis model. Instead, it uses diabetes as an empirical platform to examine how statistical tests, confidence intervals, effect sizes, and multiplicity-aware procedures can improve interpretation in large-scale data analysis.

The study has three objectives. First, it estimates diabetes prevalence and confidence intervals in a large dataset. Second, it evaluates group differences and associations using hypothesis tests and effect sizes. Third, it compares unadjusted, multiplicity-adjusted, and regression-adjusted interpretations to show why large-sample findings require careful evaluation. The main contribution is methodological and empirical: the paper provides a reproducible inference framework and applies it to a large public dataset to show how statistical and practical conclusions can be aligned.

II. LITERATURE REVIEW

The literature on statistical inference increasingly rejects mechanical threshold interpretation and promotes a broader model of evidence. Amrhein, Greenland, and McShane argued that researchers should avoid treating statistical-significance labels as proof of either presence or absence of an effect [1]. Amrhein, Trafimow, and Greenland similarly proposed treating inferential statistics as descriptive tools that support graded interpretation rather than binary decisions [2]. Baker documented broader reproducibility concerns across scientific fields, reinforcing the need for transparent statistical reporting [3].

Threshold reform and false-discovery control have remained central to the debate. Benjamin and colleagues proposed stricter evidence thresholds for new claims, while Calin-Jageman and Cumming emphasized estimation, effect sizes, and interval reporting as more informative inferential practices [4, 5]. Colquhoun showed that conventional p-value thresholds can produce misleading discovery claims when false-discovery rates are ignored [7].

Estimation-based and large-scale perspectives show why p-values should be read together with magnitude and uncertainty. Cumming argued for a new-statistics approach that emphasizes confidence intervals, effect sizes, and cumulative evidence [8]. Fan, Han, and Liu explained that high-dimensional and large-volume datasets introduce analytical challenges that require more careful inferential design [9]. Funder and Ozer argued that researchers should interpret effect sizes through context and concrete consequences rather than universal labels [10].

Several methodological studies clarify the risk of inferential error. Gelman and Carlin showed that low-power or poorly interpreted studies can produce exaggerated magnitude estimates and directional errors [11]. Goeman and Solari explained how multiple-hypothesis testing procedures can control error rates while preserving interpretable conclusions [12]. Greenland and colleagues clarified that people often misinterpret p-values and confidence intervals and should not treat them as direct probabilities about hypotheses [13].

The instability and presentation of statistical evidence have also received attention. Halsey, Curran-Everett, Vowler, and Drummond showed that p-values can vary substantially from sample to sample [14]. Ho and colleagues introduced estimation graphics as a way to make effect sizes and uncertainty more visible [15]. Hurlbert, Levine, and Utts called for more precise statistical language and warned against misleading significance terminology [16].

A related debate concerns evidence thresholds, equivalence, and sample-size justification. Ioannidis cautioned that lowering p-value thresholds alone cannot fix bias, poor design, or selective reporting [17]. Lakens, Scheel, and Isager showed that equivalence testing can help researchers evaluate whether effects are

small enough to be practically negligible [18]. Lakens later argued that researchers should justify sample size based on inferential goals, desired accuracy, and research constraints [19].

Broader discussions of research validity show that p-values are only one component of evidence. Leek and Peng argued that p-values should be understood within a wider research-validity structure [20]. Makin and Orban de Xivry summarized common statistical mistakes and emphasized the need to align test choice, data structure, and interpretation [21]. Mansournia and colleagues developed the CHAMP statement to support statistical assessment in medical papers, including clarity about design, analysis, and reporting [22].

The move away from dichotomous statistical labels is reinforced by work on p-value interpretation, big data, and confidence intervals. McShane and colleagues recommended treating p-values as continuous evidence alongside design quality, prior knowledge, and practical relevance [23]. Meng showed that large datasets can still contain serious selection, measurement, and representation errors [24]. Morey and colleagues explained that confidence intervals also require careful interpretation and should not be presented as certainty statements [25].

Reproducibility and statistical communication are closely connected to the present framework. Munafò and colleagues proposed a manifesto for reproducible science that emphasized better design, transparent analysis, and stronger reporting [26]. Nuzzo described common statistical errors that occur when researchers overstate p-values [27]. Rafi and Greenland proposed semantic and cognitive tools for improving statistical communication, particularly around compatibility and surprise rather than certainty and significance [28].

Effect-size interpretation, power, and official p-value guidance complete the methodological foundation. Schäfer and Schwarz showed that the meaning of effect sizes varies across fields and should not be reduced to universal cutoffs [29]. Szucs and Ioannidis reported limitations in effect sizes and power across parts of the cognitive neuroscience and psychology literature [30]. Trafimow and Earp criticized routine null-hypothesis testing when the inferential question is poorly matched to the method [32]. Wasserstein and Lazar presented formal ASA guidance on p-value interpretation [33], while Wasserstein, Schirm, and Lazar encouraged moving beyond simplistic $p < .05$ logic [34].

Taken together, the reviewed studies indicate that large datasets should be analyzed through integrated inference rather than isolated tests. The present study follows this direction by combining hypothesis tests, confidence intervals, effect sizes, multiplicity correction, sample-size sensitivity, and adjusted regression within a single empirical design.

Table 1. Comparison of studies on statistical inference, confidence intervals, and large-scale data analysis.

Study / source	Scope and method	Main finding	Relevance to this study
Amrhein, Greenland, & McShane (2019) [1]	Statistical-significance debate.	Called for avoiding binary evidence labels.	Supports graded conclusions.
Amrhein, Trafimow, & Greenland (2019) [2]	Inferential statistics as description.	Recommended graded and descriptive interpretation.	Supports continuous evidence.
Baker (2016) [3]	Reproducibility concerns.	Reported widespread concern about reproducibility.	Supports transparent analysis.
Benjamin et al. (2018) [4]	Threshold reform.	Proposed stricter evidence thresholds for new claims.	Supports p-value calibration.
Calin-Jageman & Cumming (2019) [5]	Estimation inference.	Linked effect sizes and interval estimates.	Supports study design.
Colquhoun (2014) [7]	False discovery and p-value misuse.	Showed discovery claims can be misleading under common thresholds.	Supports multiplicity control.
Cumming (2014) [8]	New-statistics approach.	Promoted effect sizes, confidence intervals, and cumulative evidence.	Supports interval-based reporting.
Fan et al. (2014) [9]	Big-data challenges.	Discussed analytical problems in high-dimensional data.	Frames large-scale inference.

Study / source	Scope and method	Main finding	Relevance to this study
Funder & Ozer (2019) [10]	Effect-size interpretation.	Emphasized concrete consequences and context.	Supports practical meaning.
Gelman & Carlin (2014) [11]	Design analysis and inferential error.	Explained Type S and Type M errors.	Supports effect-size caution.
Goeman & Solari (2014) [12]	Multiple hypothesis testing.	Clarified error-rate control in many-test settings.	Supports adjusted p-values.
Greenland et al. (2016) [13]	Tests, p-values, intervals, and power.	Explained common misinterpretations.	Guides conceptual framing.
Halsey et al. (2015) [14]	Instability of p-values.	Showed p-values can vary greatly across samples.	Supports interval reporting.
Ho et al. (2019) [15]	Estimation graphics.	Visualized effect sizes and uncertainty.	Supports modern figures.
Hurlbert et al. (2019) [16]	Statistical language.	Recommended eliminating misleading labels.	Supports precise wording.
Ioannidis (2018) [17]	Threshold debate.	Warned that threshold changes cannot remove bias.	Supports balanced interpretation.
Lakens et al. (2018) [18]	Equivalence testing.	Presented a framework for negligible-effect testing.	Supports practical significance.
Lakens (2022) [19]	Sample-size justification.	Explained sample-size reasoning and desired accuracy.	Supports large-sample interpretation.
Leek & Peng (2015) [20]	Interpretation of p-values.	Argued p-values are only part of research evidence.	Supports broader inference.
Makin & Orban de Xivry (2019) [21]	Statistical mistakes.	Identified common analysis and interpretation errors.	Supports robustness checks.
Mansournia et al. (2021) [22]	Statistical assessment checklist.	Provided reporting-quality criteria for medical papers.	Supports validity audit.
McShane et al. (2019) [23]	Abandoning threshold logic.	Recommended treating p-values as one evidence component.	Supports integrated framework.
Meng (2018) [24]	Big-data paradoxes.	Showed large data can still contain serious bias.	Supports data-quality audit.
Morey et al. (2016) [25]	Confidence interval interpretation.	Explained confidence-interval misconceptions.	Supports careful interval use.
Munafò et al. (2017) [26]	Reproducible science manifesto.	Recommended stronger design and transparency.	Supports reproducibility.
Nuzzo (2014) [27]	Statistical errors in scientific reporting.	Showed that p-values are often misread and overused.	Supports cautious interpretation.
Rafi & Greenland (2020) [28]	Statistical communication.	Improved semantic tools for statistical interpretation.	Supports clearer reporting.
Schäfer & Schwarz (2019) [29]	Effect-size meaning.	Showed that effect-size interpretation varies by field.	Supports contextual evaluation.
Szucs & Ioannidis (2017) [30]	Power and effect-size assessment.	Showed limitations in common empirical reporting.	Supports effect-size emphasis.
Trafimow & Earp (2017) [32]	NHST critique.	Questioned routine null-hypothesis testing.	Supports alternative interpretation.
Wasserstein & Lazar (2016) [33]	ASA p-value guidance.	Warned against threshold-based misuse.	Supports p-value interpretation.
Wasserstein et al. (2019) [34]	Beyond $p < .05$.	Encouraged abandoning mechanical dichotomies.	Supports combined inference.

The table shows that the literature consistently recommends combining p-values with confidence intervals, effect sizes, transparent reporting, and multiplicity-aware interpretation.

III. RESEARCH GAP

The reviewed literature provides strong conceptual guidance on p-values, confidence intervals, effect sizes, and multiplicity. However, many discussions remain either conceptual or focused on isolated statistical problems. Fewer studies present a single applied framework showing how these components work together in one large public dataset. This study addresses this gap by combining hypothesis tests, confidence

intervals, effect sizes, multiplicity correction, sample-size sensitivity, and adjusted regression in one reproducible empirical design.

IV. CONCEPTUAL FRAMEWORK AND HYPOTHESES DEVELOPMENT

The conceptual framework treats valid inference as a sequence of connected decisions. Data quality and variable coding are first audited. Descriptive statistics establish the empirical structure. Hypothesis tests evaluate compatibility with null assumptions. Confidence intervals quantify precision. Effect sizes assess substantive magnitude. Multiple-comparison procedures protect against inflated claims of discovery. Regression adjustment then tests whether the selected relationships remain evident after accounting for other covariates.

- H1: Respondents with diabetes have a higher mean BMI than respondents without diabetes.
- H2: High blood pressure is positively associated with diabetes status.
- H3: High cholesterol is positively associated with diabetes status.
- H4: Physical activity is negatively associated with diabetes status.
- H5: Multiple-comparison adjustment changes the interpretation of some large-scale tests when p-values are considered without effect sizes.
- H6: Some variables remain associated with diabetes after adjustment, but their practical interpretation differs from the unadjusted tests.

V. PROPOSED METHOD

The proposed method is an effect-size and multiplicity-aware inference framework. It begins by defining the outcome and candidate variables, then separates estimation from decision testing. Estimation is used to report prevalence, group means, proportions, odds ratios, and confidence intervals. Hypothesis testing evaluates whether observed differences are unlikely under a null model. Effect-size metrics then assess whether a statistically significant result has sufficient substantive magnitude. Finally, multiplicity correction and adjusted regression are used to reduce overinterpretation of many simultaneous tests.

The procedure is deliberately transparent. Each statistical result is reported with at least three components: a test statistic or p-value, an interval estimate, and a magnitude measure. For continuous or ordinal variables, mean differences and Cohen-type standardized effects are reported. For binary variables, risk differences, odds ratios, confidence intervals, and Cramer-type association measures are reported. For multiple-test interpretation, we apply Bonferroni and Benjamini-Hochberg procedures to distinguish raw statistical significance from multiplicity-aware evidence.

The main estimators and tests used in the framework are summarized below. The equations are included in professional equation format to make the analytical procedure reproducible.

$$\text{prevalence estimate} = x / n \quad (1)$$

$$95\%\ \text{CI for prevalence} = \text{estimate} \pm 1.96 \times \text{standard error} \quad (2)$$

$$\text{Welch } t = \text{mean difference} / \text{standard error of the difference} \quad (3)$$

$$\text{odds ratio} = (a \times d) / (b \times c) \quad (4)$$

$$95\%\ \text{CI for odds ratio} = \exp[\ln(\text{odds ratio}) \pm 1.96 \times SE] \quad (5)$$

$$\text{Cohen } d = \text{mean difference} / \text{pooled standard deviation} \quad (6)$$

$$\text{Benjamini} - \text{Hochberg adjusted } p = \min[(m / \text{rank}) \times \text{ordered } p - \text{value}, 1] \quad (7)$$

$$\text{logit probability} = \text{intercept} + \text{beta1X1} + \text{beta2X2} + \ldots + \text{betakXk} \quad (8)$$

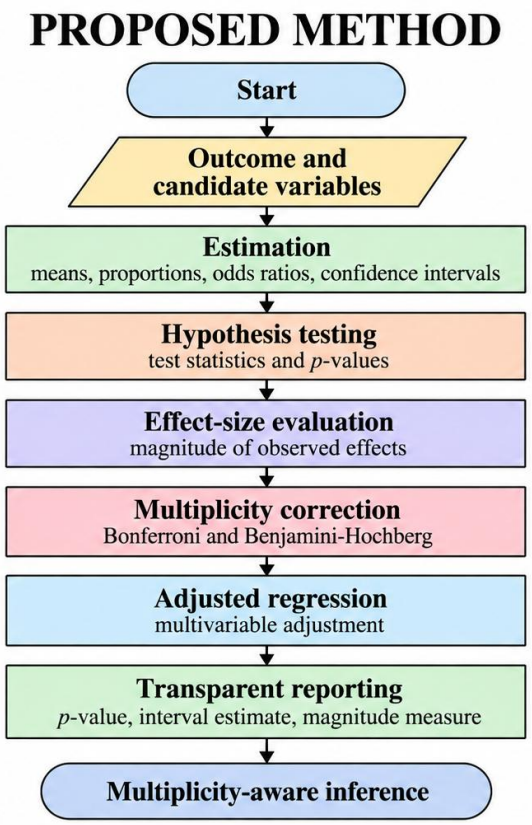


FIGURE 1. Proposed effect-size and multiplicity-aware inference framework.

VI. STUDY CONTRIBUTION

This study makes a methodological and empirical contribution to large-scale statistical inference. Methodologically, it develops a unified interpretation framework in which p-values are not treated as final conclusions, but are evaluated together with confidence intervals, effect sizes, risk differences, odds ratios, multiplicity-adjusted decisions, and adjusted regression estimates. This integrated structure helps reduce the common problem of overstating statistically visible results in very large samples.

Empirically, the study applies the framework to a large public BRFSS2015 diabetes health-indicator dataset containing 253,680 observations and 22 variables. The empirical analysis demonstrates how the same dataset can produce extremely small p-values while still requiring effect-size screening and substantive interpretation. The framework therefore provides a reproducible reporting model for researchers who work with health, social-science, administrative, or business datasets where sample size, test multiplicity, and practical meaning must be interpreted together.

The contribution is also practical because it converts methodological recommendations into a usable analytical sequence: data audit, descriptive profiling, hypothesis testing, interval estimation, effect-size evaluation, multiplicity control, adjusted modeling, and final interpretation. This sequence can support journal reviewers, applied researchers, and policy analysts in distinguishing statistical visibility from practical importance.

VII. METHODOLOGY

1. RESEARCH DESIGN

The study uses a quantitative, observational, cross-sectional design. The purpose is not to infer clinical causality, but to demonstrate a rigorous statistical-inference framework using a large empirical dataset. The design is appropriate because the dataset includes a binary outcome, multiple predictor types, and a large sample size that illustrates both the strengths and risks of large-scale inference.

2. DATA SOURCE / PARTICIPANTS / CORPUS

The main empirical file is diabetes_binary_health_indicators_BRFSS2015.csv, which contains 253,680 observations and 22 variables. It is derived from the Behavioral Risk Factor Surveillance System survey context [6] and distributed as a public Kaggle health-indicator archive [31]. The outcome variable is Diabetes_binary, where 1 represents diabetes, and 0 represents no diabetes. The dataset includes cardiometabolic indicators, lifestyle factors, healthcare-access variables, and demographic or socioeconomic variables.

3. SAMPLE SELECTION AND INCLUSION CRITERIA

We retained all observations in the main binary file because it contained no missing values after inspection. The artificial 50/50 balanced file was not used as the main dataset because it changes the natural prevalence structure and would distort prevalence estimates and population-oriented confidence intervals. The three-class diabetes file was reserved as a possible sensitivity source because it separates no diabetes, prediabetes, and diabetes, but the main analysis uses the binary outcome to maintain a direct hypothesis-testing framework.

4. VARIABLES / MEASURES / CODING FRAMEWORK

Table 2. Variables used in the empirical analysis.

Variable	Type	Coding / scale	Analytical role
Diabetes_binary	Binary outcome	0 = no diabetes, 1 = diabetes	Dependent variable
HighBP	Binary	0 = no high blood pressure, 1 = high blood pressure	Clinical risk indicator
HighChol	Binary	0 = no high cholesterol, 1 = high cholesterol	Clinical risk indicator
CholCheck	Binary	0 = no cholesterol check, 1 = cholesterol check	Healthcare behavior
BMI	Continuous numeric	Body mass index	Anthropometric measure
Smoker	Binary	0 = no, 1 = yes	Lifestyle indicator
Stroke	Binary	0 = no, 1 = yes	Health history
HeartDiseaseorAttack	Binary	0 = no, 1 = yes	Health history
PhysActivity	Binary	0 = no, 1 = yes	Lifestyle indicator
Fruits / Veggies	Binary	0 = no, 1 = yes	Diet-related indicators
HvyAlcoholConsump	Binary	0 = no, 1 = yes	Lifestyle indicator
AnyHealthcare / NoDocbcCost	Binary	Healthcare access / cost barrier	Access indicators
GenHlth	Ordinal	1 = excellent to 5 = poor	Self-rated health
MentHlth / PhysHlth	Count scale	Number of unhealthy days	Health-burden indicators
DiffWalk	Binary	0 = no, 1 = yes	Functional limitation
Sex	Binary	0 = female, 1 = male	Demographic variable
Age	Ordinal category	1 to 13 age categories	Demographic variable
Education / Income	Ordinal categories	Education and income levels	Socioeconomic variables

5. ANALYTICAL PROCEDURE

- Dataset audit was conducted to verify row count, column count, variable types, and missingness.

- Descriptive statistics were calculated for the full sample and by diabetes status.
- Diabetes prevalence and its 95% confidence interval were estimated.
- Welch tests and Mann-Whitney tests were used for continuous or ordinal predictors.
- Chi-square tests, risk differences, odds ratios, and Cramer-type effects were used for binary predictors.
- Bonferroni and Benjamini-Hochberg corrections were applied to the family of variable-level tests.
- Logistic regression was used to estimate adjusted odds ratios and confidence intervals.
- Sample-size sensitivity was examined to show how p-values change as sample size increases.

6. VALIDITY, RELIABILITY, AND ETHICAL CONSIDERATIONS

Validity was addressed by using a large public dataset, retaining the natural prevalence structure, and reporting interval estimates and effect sizes rather than p-values alone. We strengthened reliability by applying reproducible Python-based statistical procedures and saving numerical outputs for verification. Ethical risk is minimal because the dataset is public, de-identified, and used only for aggregate statistical analysis. We did not attempt individual-level identification or sensitive personal tracing.

VIII. RESULTS / FINDINGS

The main dataset included 253,680 observations: 218,334 respondents without diabetes and 35,346 with diabetes. Diabetes prevalence was 13.93%, with a 95% confidence interval from 13.80% to 14.07%. The very narrow interval reflects the large sample size and high precision of the prevalence estimate.

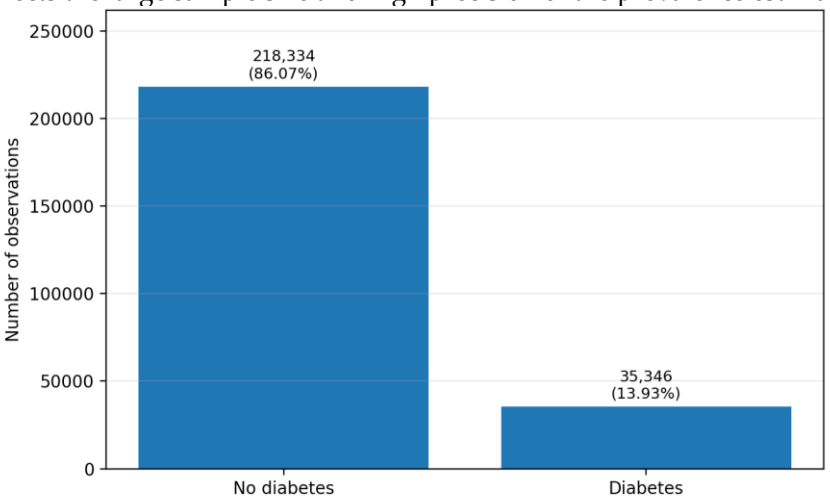


FIGURE 2. Distribution of diabetes status in the main BRFSS2015 binary file.

The figure shows that the main file keeps a realistic imbalance between respondents without diabetes and respondents with diabetes, which is important for valid prevalence and interval interpretation.

Table 3. Sample profile and group differences for selected continuous or ordinal variables.

Variable	No diabetes mean	Diabetes mean	Mean diff.	95% CI	p-value	Cohen d
BMI	27.81	31.94	4.14	4.06 to 4.22	<0.001	0.64
GenHlth	2.39	3.29	0.91	0.89 to 0.92	<0.001	0.89
MentHlth	2.98	4.46	1.48	1.39 to 1.58	<0.001	0.20
PhysHlth	3.64	7.95	4.31	4.19 to 4.44	<0.001	0.50
Age	7.81	9.38	1.56	1.54 to 1.59	<0.001	0.52
Education	5.10	4.75	-0.35	-0.37 to -0.34	<0.001	-0.36

Variable	No diabetes mean	Diabetes mean	Mean diff.	95% CI	p-value	Cohen d
Income	6.19	5.21	-0.98	-1.00 to -0.96	<0.001	-0.48

The table shows statistically visible differences across all selected variables, but the standardized effects differ in magnitude; general health and BMI show stronger practical contrasts than some smaller differences.

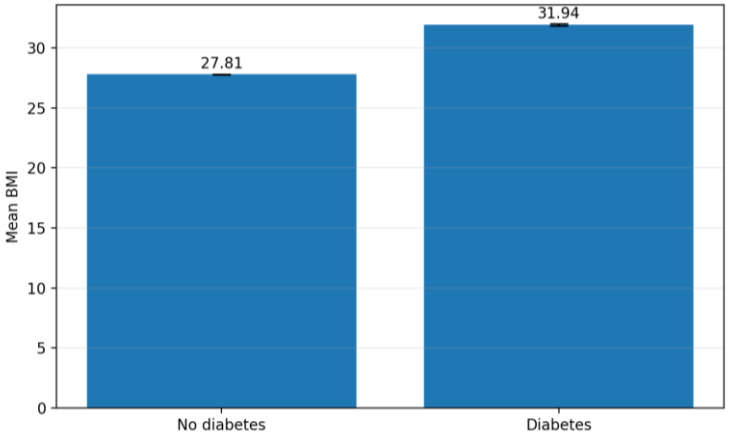


FIGURE 3. Mean BMI by diabetes status with 95% confidence intervals.

The figure shows a clear BMI difference between groups. The narrow intervals reflect the precision created by the large sample size.

Table 4. Main binary-variable hypothesis tests, odds ratios, confidence intervals, and effect sizes.

Variable	Diabetes % if 1	Diabetes % if 0	Risk diff. pp	Odds ratio	95% CI	Cramer V
HighBP	24.45	6.04	18.41	5.04	4.91 to 5.17	0.263
HighChol	22.01	7.98	14.03	3.25	3.18 to 3.33	0.200
HeartDiseaseorAttack	32.97	11.95	21.02	3.62	3.52 to 3.73	0.177
Stroke	31.75	13.18	18.57	3.06	2.94 to 3.20	0.106
DiffWalk	30.75	10.53	20.21	3.77	3.68 to 3.87	0.218
PhysActivity	11.61	21.14	-9.53	0.49	0.48 to 0.50	0.118
HvyAlcoholConsump	5.84	14.42	-8.58	0.37	0.34 to 0.40	0.057
Smoker	16.29	12.06	4.24	1.42	1.39 to 1.45	0.061
Fruits	12.86	15.79	-2.93	0.79	0.77 to 0.81	0.041
Veggies	12.99	18.00	-5.01	0.68	0.66 to 0.70	0.057
Sex	15.16	12.97	2.19	1.20	1.17 to 1.23	0.031
NoDocbcCost	17.52	13.60	3.92	1.35	1.30 to 1.40	0.031

The table shows that several binary variables have extremely small p-values, but interpretation should focus on the odds ratios, risk differences, and Cramer-type effect magnitudes.

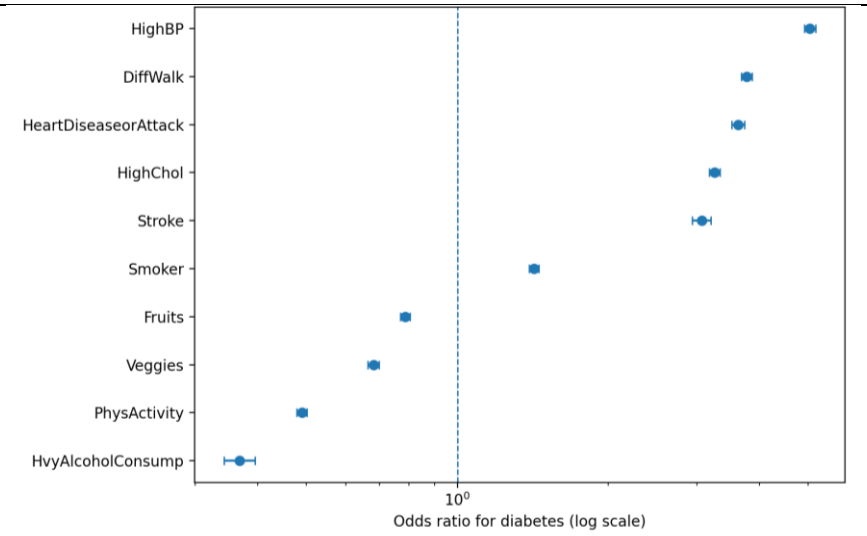


FIGURE 4. Unadjusted odds ratios with 95% confidence intervals.

The figure shows that high blood pressure, high cholesterol, difficulty walking, and cardiovascular-history variables have stronger unadjusted associations with diabetes than lifestyle indicators, which have smaller effects.

Table 5. Multiple-comparison adjustment for variable-level tests.

Variable	Test	Raw p	BH-FDR p	Bonferroni p	Effect metric	Effect value
HighBP	Chi-square	<0.001	<0.001	<0.001	Odds ratio	5.04
HighChol	Chi-square	<0.001	<0.001	<0.001	Odds ratio	3.25
CholCheck	Chi-square	<0.001	<0.001	<0.001	Odds ratio	6.43
BMI	Welch t	<0.001	<0.001	<0.001	Mean difference	4.14
Smoker	Chi-square	<0.001	<0.001	<0.001	Odds ratio	1.42
Stroke	Chi-square	<0.001	<0.001	<0.001	Odds ratio	3.06
HeartDiseaseorAttack	Chi-square	<0.001	<0.001	<0.001	Odds ratio	3.62
PhysActivity	Chi-square	<0.001	<0.001	<0.001	Odds ratio	0.490
Fruits	Chi-square	<0.001	<0.001	<0.001	Odds ratio	0.787
Veggies	Chi-square	<0.001	<0.001	<0.001	Odds ratio	0.680
HvyAlcoholConsump	Chi-square	<0.001	<0.001	<0.001	Odds ratio	0.368
AnyHealthcare	Chi-square	<0.001	<0.001	<0.001	Odds ratio	1.27
NoDocbcCost	Chi-square	<0.001	<0.001	<0.001	Odds ratio	1.35
GenHlth	Welch t	<0.001	<0.001	<0.001	Mean difference	0.906
MentHlth	Welch t	<0.001	<0.001	<0.001	Mean difference	1.48
PhysHlth	Welch t	<0.001	<0.001	<0.001	Mean difference	4.31
DiffWalk	Chi-square	<0.001	<0.001	<0.001	Odds ratio	3.77
Sex	Chi-square	<0.001	<0.001	<0.001	Odds ratio	1.20
Age	Welch t	<0.001	<0.001	<0.001	Mean difference	1.56
Education	Welch t	<0.001	<0.001	<0.001	Mean difference	-0.354
Income	Welch t	<0.001	<0.001	<0.001	Mean difference	-0.980

The table indicates that most tests remain statistically significant after adjustment, which is expected in a very large dataset; therefore, the magnitude and meaning of each effect remain essential for interpretation.

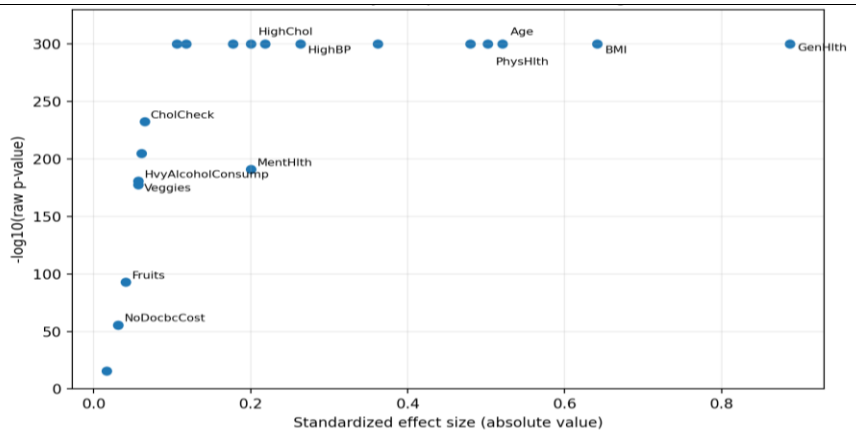


FIGURE 5. Statistical visibility compared with effect magnitude.

The figure shows that variables can have highly visible p-values while differing substantially in effect magnitude, which supports the need for effect-size interpretation.

Table 6. Adjusted logistic regression odds ratios and confidence intervals.

Variable	Adjusted OR	95% CI	p-value	Interpretation
HighBP	2.17	2.11 to 2.23	<0.001	Higher adjusted odds
HighChol	1.80	1.75 to 1.85	<0.001	Higher adjusted odds
BMI	1.06	1.06 to 1.06	<0.001	Higher odds per BMI unit
GenHlth	1.71	1.68 to 1.74	<0.001	Worse general health increases odds
Age	1.14	1.13 to 1.14	<0.001	Higher odds across age categories
HeartDiseaseorAttack	1.25	1.21 to 1.30	<0.001	Higher adjusted odds
Stroke	1.15	1.09 to 1.20	<0.001	Higher adjusted odds
DiffWalk	1.14	1.10 to 1.17	<0.001	Higher adjusted odds
PhysActivity	0.96	0.93 to 0.98	0.001	Lower adjusted odds
Fruits	0.95	0.93 to 0.98	<0.001	Lower adjusted odds
Veggies	0.97	0.94 to 1.00	0.036	Slightly lower adjusted odds
HvyAlcoholConsump	0.46	0.43 to 0.50	<0.001	Lower adjusted odds in this observational file
Sex	1.29	1.25 to 1.32	<0.001	Higher odds for coded male category
Education	0.97	0.96 to 0.98	<0.001	Lower odds at higher levels
Income	0.95	0.95 to 0.96	<0.001	Lower odds at higher levels

Variable	Adjusted OR	95% CI	p-value	Interpretation
Smoker	0.99	0.96 to 1.01	0.285	Not important after adjustment
NoDocbcCost	0.98	0.94 to 1.03	0.419	Not important after adjustment

The adjusted model shows that several unadjusted associations remain visible after covariate adjustment, but some variables with strong raw p-values become weak after accounting for other predictors.

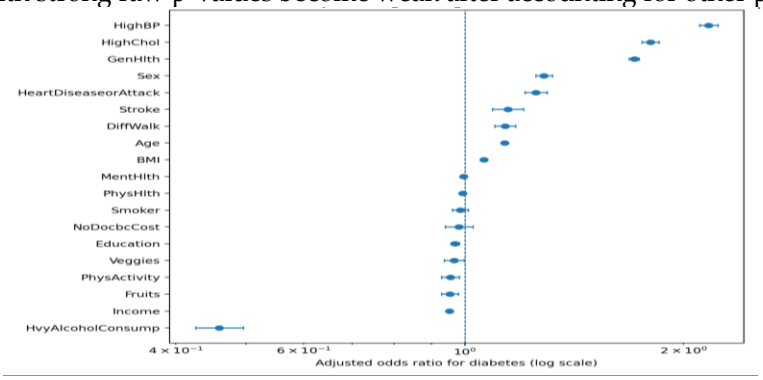


FIGURE 6. Adjusted logistic regression odds ratios.

The figure shows adjusted odds ratios on a log scale, making it easier to compare variables with effects that differ in direction and magnitude.

Table 7. Sample-size sensitivity of selected p-values.

Variable	n = 500	n = 5,000	n = 50,000	Full sample
Fruits	0.002	0.024	<0.001	<0.001
Veggies	0.031	<0.001	<0.001	<0.001
Smoker	0.480	<0.001	<0.001	<0.001

The table illustrates how the same variable can move from weak or moderate evidence in small subsamples to extremely small p-values in the full dataset.

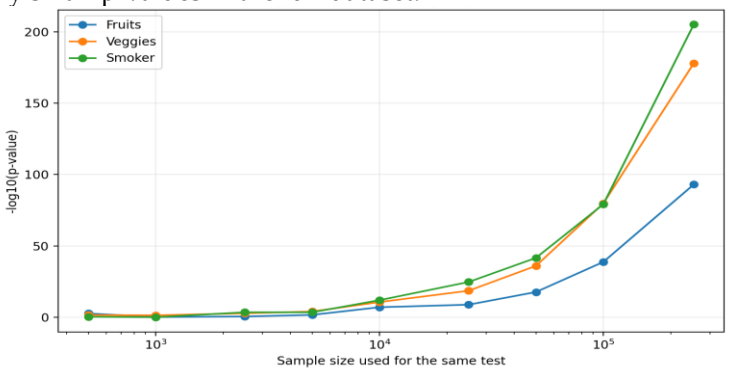


FIGURE 7. Large-sample sensitivity of p-values.

The figure demonstrates that p-values tend to become more extreme as sample size increases, even when the underlying effect magnitude remains modest.

IX. DISCUSSION

The findings support the central argument of the study: hypothesis testing in large-scale datasets should be interpreted jointly with confidence intervals, effect sizes, and multiplicity-aware procedures. The full dataset produced very small p-values for most comparisons. If the analysis stopped here, the results could create the impression that every variable is equally important. The confidence intervals and effect-size estimates show a more informative pattern. Some variables, such as high blood pressure, high cholesterol, general health, BMI, and age, showed stronger and more interpretable associations with diabetes status. Other variables remained statistically visible but had smaller magnitudes or became weaker after adjustment.

The prevalence estimate demonstrates the value of confidence intervals in large samples. The estimated diabetes prevalence was precise, with a narrow interval. This precision is useful, but it should not be confused with causal strength. The interval reflects uncertainty around the observed prevalence estimate in the dataset, while interpreting association still depends on variable measurement, coding, selection, confounding, and study design. This distinction is central to responsible large-scale data analysis.

The BMI result illustrates how a statistical test and an interval estimate can be interpreted together. The mean BMI difference between groups was approximately 4.14 units, with a narrow confidence interval and a moderate standardized effect. This is not merely a p-value result; it is practically interpretable because the magnitude matters in the context of health indicators. In contrast, variables such as fruits and vegetables showed statistically visible differences but comparatively smaller adjusted effects. These findings demonstrate why large-scale inference requires effect-size screening.

The unadjusted odds-ratio results also show why statistical interpretation must distinguish between association and explanation. High blood pressure and high cholesterol had strong unadjusted associations with diabetes. However, adjusted regression reduced the magnitude of several associations because variables share information with age, BMI, general health, and other predictors. This does not invalidate the unadjusted tests; instead, it clarifies that unadjusted tests describe associations, while adjusted models provide conditional estimates.

The multiple-comparison results confirm that p-value correction alone is not enough. Because the dataset is large, many adjusted p-values remained below conventional thresholds. Therefore, pair multiplicity adjustment with effect-size reporting. In smaller datasets, correction may change which results appear statistically significant. In very large datasets, correction may still leave many results visible, so practical interpretation becomes even more important.

The sample-size sensitivity analysis directly demonstrates the large-sample problem. For selected variables, p-values became more extreme as sample size increased. This pattern confirms that p-values are not pure measures of practical importance. They depend on effect magnitude, sample size, variation, and model assumptions. Therefore, the proposed framework recommends reading every p-value alongside confidence intervals, effect metrics, and substantive judgment.

X. PRACTICAL AND POLICY IMPLICATIONS

For applied researchers, the study provides a reporting template for analyzing large datasets. Results should include p-values, confidence intervals, and effect sizes, while tables should identify whether conclusions survive multiplicity correction. This approach improves transparency and prevents exaggerated claims based on p-values alone.

For health-data analysts and policy researchers, the results demonstrate the importance of distinguishing statistical visibility from practical importance. In large public-health files, many associations may be detectable, but policy interpretation should prioritize effects that are large enough, interpretable enough, and consistent enough to support decision-making. Confidence intervals and adjusted estimates help identify which findings deserve closer attention.

For journal reviewers and editors, the framework offers a checklist for assessing large-scale empirical papers. A strong manuscript should report the data source, coding decisions, missingness, test selection, confidence intervals, effect sizes, correction for multiple comparisons, and limitations of causal interpretation.

XI. LIMITATIONS AND FUTURE RESEARCH

The study has several limitations. First, the dataset is observational and cross-sectional; therefore, the results describe associations rather than causal effects. Second, the variables are coded in simplified binary or ordinal form, which may reduce measurement detail. Third, the analysis does not apply survey weights, so the estimates should be interpreted as file-based empirical estimates rather than weighted national estimates. Fourth, self-reporting and measurement error may affect some variables. Fifth, the logistic regression model is used for adjusted interpretation, not for clinical prediction or causal modeling.

Future research can extend the framework by applying survey weights, comparing binary and three-class diabetes outcomes, using bootstrap interval estimation, adding interaction terms, and evaluating how missing-data mechanisms affect confidence intervals and p-values. Additional studies can apply the same framework to financial, educational, environmental, or administrative datasets to test whether the same interpretation principles hold across domains.

XII. CONCLUSION

This study developed and applied a structured framework for statistical hypothesis testing and confidence interval estimation in large-scale data analysis. Using a large BRFSS2015 diabetes health-indicator file, the analysis showed that p-values alone are insufficient for reliable interpretation. The dataset produced many statistically significant results, but their substantive meaning varied across variables once we considered confidence intervals, effect sizes, risk differences, odds ratios, multiplicity correction, and adjusted regression. The empirical findings showed a precise diabetes prevalence estimate, meaningful BMI and general-health differences, strong unadjusted associations for high blood pressure and high cholesterol, and more nuanced results after adjustment. The sample-size sensitivity analysis further demonstrated that p-values become more extreme as sample size grows, even when effects are modest. The main conclusion is that large-scale statistical inference should move beyond threshold-based testing and adopt integrated reporting. A reliable empirical conclusion should state the test result, the confidence interval, the effect magnitude, the multiplicity-aware interpretation, and the practical meaning. This approach strengthens transparency, improves reproducibility, and supports more responsible conclusions in large-scale empirical research.

XIII. DATA AND CODE AVAILABILITY

The analysis used the public Diabetes Health Indicators / BRFSS2015 Kaggle archive uploaded for this study. We used Python scripts to audit the file, calculate tests and confidence intervals, generate figures, and prepare reproducible summary tables. We did not use the artificial balanced file as the main empirical source because it changes the natural outcome distribution.

Author Contributions

The author conducted the conceptualization, methodology, data analysis, investigation, writing, review, editing, and final approval of the manuscript.

Funding

This research received no external funding.

Data Availability

The dataset will be available from the author upon reasonable request.

Conflicts of Interest

The author declares no conflict of interest.

REFERENCES

1. Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. Nature, 567(7748), 305–307. https://doi.org/10.1038/d41586-019-00857-9
2. Amrhein, V., Trafimow, D., & Greenland, S. (2019). Inferential statistics as descriptive statistics: There is no replication crisis if we do not expect replication. The American Statistician, 73(sup1), 262–270. https://doi.org/10.1080/00031305.2018.1543137
3. Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. Nature, 533(7604), 452–454. https://doi.org/10.1038/533452a
4. Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., ... Johnson, V. E. (2018). Redefine statistical significance. Nature Human Behaviour, 2(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z
5. Calin-Jageman, R. J., & Cumming, G. (2019). Estimation for better inference in neuroscience. eNeuro, 6(4), ENEURO.0205-19.2019. https://doi.org/10.1523/ENEURO.0205-19.2019
6. Centers for Disease Control and Prevention. (2015). Behavioral Risk Factor Surveillance System survey data. U.S. Department of Health and Human Services.
7. Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. Royal Society Open Science, 1(3), 140216. https://doi.org/10.1098/rsos.140216
8. Cumming, G. (2014). The new statistics: Why and how. Psychological Science, 25(1), 7–29. https://doi.org/10.1177/0956797613504966
9. Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. National Science Review, 1(2), 293–314. https://doi.org/10.1093/nsr/nwt032
10. Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. Advances in Methods and Practices in Psychological Science, 2(2), 156–168. https://doi.org/10.1177/2515245919847202
11. Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing Type S and Type M errors. Perspectives on Psychological Science, 9(6), 641–651. https://doi.org/10.1177/1745691614551642
12. Goeman, J. J., & Solari, A. (2014). Multiple hypothesis testing in genomics. Statistics in Medicine, 33(11), 1946–1978. https://doi.org/10.1002/sim.6082
13. Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, p values, confidence intervals, and power: A guide to misinterpretations. European Journal of Epidemiology, 31(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3
14. Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. (2015). The fickle p-value generates irreproducible results. Nature Methods, 12(3), 179–185. https://doi.org/10.1038/nmeth.3288
15. Ho, J., Tumkaya, T., Aryal, S., Choi, H., & Claridge-Chang, A. (2019). Moving beyond p values: Data analysis with estimation graphics. Nature Methods, 16(7), 565–566. https://doi.org/10.1038/s41592-019-0470-3
16. Hurlbert, S. H., Levine, R. A., & Utts, J. (2019). Coup de grâce for a tough old bull: Statistically significant expires. The American Statistician, 73(sup1), 352–357. https://doi.org/10.1080/00031305.2018.1543616
17. Ioannidis, J. P. A. (2018). Proposal to lower p-value thresholds to .005. JAMA, 319(14), 1429–1430. https://doi.org/10.1001/jama.2018.1536
18. Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. Advances in Methods and Practices in Psychological Science, 1(2), 259–269. https://doi.org/10.1177/2515245918770963
19. Lakens, D. (2022). Sample size justification. Collabra: Psychology, 8(1), 33267. https://doi.org/10.1525/collabra.33267
20. Leek, J. T., & Peng, R. D. (2015). Statistics: P values are just the tip of the iceberg. Nature, 520(7549), 612. https://doi.org/10.1038/520612a
21. Makin, T. R., & Orban de Xivry, J. J. (2019). Ten common statistical mistakes to watch out for when writing or reviewing a manuscript. eLife, 8, e48175. https://doi.org/10.7554/eLife.48175
22. Mansournia, M. A., Collins, G. S., Nielsen, R. O., Nazemipour, M., Jewell, N. P., Altman, D. G., & Campbell, M. J. (2021). Checklist for Statistical Assessment of Medical Papers: The CHAMP statement. British Journal of Sports Medicine, 55(18), 1002–1003. https://doi.org/10.1136/bjsports-2020-103651
23. McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. The American Statistician, 73(sup1), 235–245. https://doi.org/10.1080/00031305.2018.1527253
24. Meng, X. L. (2018). Statistical paradises and paradoxes in big data: Law of large populations, big data paradox, and the 2016 US presidential election. The Annals of Applied Statistics, 12(2), 685–726. https://doi.org/10.1214/18-AOAS1161SF

25. Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E. J. (2016). The fallacy of placing confidence in confidence intervals. Psychonomic Bulletin & Review, 23(1), 103–123. https://doi.org/10.3758/s13423-015-0947-8
26. Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., du Sert, N. P., Simonsohn, U., Wagenmakers, E. J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. Nature Human Behaviour, 1, 0021. https://doi.org/10.1038/s41562-016-0021
27. Nuzzo, R. (2014). Scientific method: Statistical errors. Nature, 506(7487), 150–152. https://doi.org/10.1038/506150a
28. Rafi, Z., & Greenland, S. (2020). Semantic and cognitive tools to aid statistical science: Replace confidence and significance by compatibility and surprise. BMC Medical Research Methodology, 20, 244. https://doi.org/10.1186/s12874-020-01105-9
29. Schäfer, T., & 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, 813. https://doi.org/10.3389/fpsyg.2019.00813
30. Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. PLOS Biology, 15(3), e2000797. https://doi.org/10.1371/journal.pbio.2000797
31. Teboul, A. (2021). Diabetes Health Indicators Dataset. Kaggle. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset
32. Trafimow, D., & Earp, B. D. (2017). Null hypothesis significance testing and Type I error: The domain problem. Frontiers in Psychology, 8, 621. https://doi.org/10.3389/fpsyg.2017.00621
33. Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. The American Statistician, 70(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108
34. Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond p < .05. The American Statistician, 73(sup1), 1–19. https://doi.org/10.1080/00031305.2019.1583913