01 May 2024

White Paper

Bridge
Evidence that matters

Is journal impact factor a reliable indicator of study quality?

A Pilot Study

Through our combined experience as outcomes researchers at Bridge, we have observed many instances of observational studies of low quality being published in prestigious journals, and conversely, high-quality studies in low-impact journals. Data on whether journal quality (as measured by journal Impact Factor) can be used as a surrogate for the quality of studies published in that journal are limited. We present here the summary findings of a pilot study to investigate the relationship between journal Impact Factor and the methodological quality of real-world observational studies.

Background

An integral part of conducting systematic literature reviews [SLRs] is the **quality appraisal** of the underlying studies included for reporting. Based on our experience of evaluating thousands of observational research papers each year using validated quality assessment scales, we have witnessed numerous examples of poorer-quality studies being published in high-impact journals; and, conversely, high-quality studies being published in low-impact journals.

On investigation, we found limited and inconsistent existing literature on whether the journal impact factor (IF) – as a surrogate for journal quality – is a reliable indicator of the methodological quality of research published in that journal.

The existing evidence on the relationship between IF and research quality is primarily based on data from clinical trials [2-4], systematic reviews [1] or a combination of clinical trials and observational studies [5]; with no comprehensive studies specifically exploring this association in the context of real-world observational studies.

While some studies have found journal IF to be a poor indicator of research quality [1, 2], a few have reported a weak-to-moderate positive association between the two [3-5].

We therefore hypothesized that journal IF might not necessarily be a good indicator of the methodological quality of published observational research. Consequently, the **primary objective** of this **pilot study** was to investigate the association between journal IF and research quality **in the context of real-world observational studies**. We also hypothesized that the IF-research quality association might be different across different levels of certain factors such as study design, type of funding and geographic location. Therefore, the **secondary objective** of this study was to explore if the association between journal IF and research quality varies as a function of these factors.

Methods

We perform quality assessment of research papers for all our SLR projects for multiple reasons. First, it helps in determining if a study should be included in the systematic review. Second, it helps us understand the overall strength of conclusions (e.g., are any conclusions based on studies of relatively low-quality?). Third, it ensures that the interpretation is not distorted by low-quality studies (i.e., when studies of varying quality have been included). And finally, it helps in performing sensitivity analysis and meta-regression, especially when the SLR is accompanied by a meta-analysis.

This pilot study included **457** research papers published in **208** unique journals across **11** consecutive SLR projects conducted by one research team within Bridge over the last 5 years. All papers had been assessed using the **Newcastle-Ottawa Scale (NOS)** for observational studies. The NOS contains 3 domains: selection (4 questions), comparability (1 question) and assessment of outcome or exposure (3 questions) [6]. Three different NOS instruments were used, one for cohort studies, one for case-control studies, and an adapted version for cross-sectional studies [7]. The scores for all scales range from a minimum of 0 to a maximum of 9 (with higher ratings indicating better quality). All papers were assessed by one team member, whose results were then verified by a second team member. Any discrepancies were resolved through consensus. Appendix 1 shows how the scoring on each of the three NOS instruments work for their respective study designs.

The latest IFs were obtained directly from the official websites of the respective journals.

The majority of journals provided IFs for the year 2022; in a few instances, for 2021.

Analytical approach

In order to inform the choice between parametric and non-parametric statistical procedures, the normality of the NOS score and IF was assessed using multiple sources of information, such as a histogram (with a superimposed normal curve), Shapiro-Wilk test, skewness and kurtosis z-values, and Normal Q-Q Plot.

The primary objective of the study (i.e., the association between journal IF and NOS score in the overall study sample) was evaluated first using Kendall's tau-b correlation coefficient and secondly using one-way analysis of variance (ANOVA). Each analytic method provides a different measure of effect size, and together, they allow for a more comprehensive evaluation of the relationship.

The secondary objective was also analyzed using the same techniques, except for the fact that the analysis was performed separately within each category of the 3 factors (study design, type of funding and geographic location). All data were analyzed using SPSS version 28.0 (IBM, Armonk, NY, USA). All analyses were two-tailed, and a difference was considered statistically significant if the p value was ≤ 0.05.

A detailed description of the statistical methods is presented in Appendix 2.

Results

Study characteristics

As shown in Table 1 below, the majority of the studies in the sample were cross-sectional, followed by retrospective cohort and prospective cohort. With regard to the distribution of studies across various geographic regions, North America and Europe had the highest representation, followed by Asia Pacific and multi-region studies. Approximately 40% of the studies were industry-funded.

With respect to the underlying disease area, based on our consecutive convenience sample, episodic and chronic migraine studies had the highest representation, followed by anemia in chronic kidney disease and sleep disturbances due to pruritis; while diabetic macular ischemia and diabetic gastroparesis were amongst those with the lowest representation.

Table 1: Study characteristics

Characteristic	Categories	Number (%)
By study design	Cross-sectional	207 (45.3)
	Retrospective cohort	123 (26.9)
	Prospective cohort	122 (26.7)
	Others[1]	5 (1.1)
By geography	North America	174 (38.1)
	Europe	146 (31.9)
	Asia Pacific	74 (16.2)
	Multi-region[2]	54 (11.8)
	Others[3]	9 (2)
By the type of sponsor	Industry	175 (38.3)
	Non-industry[4]	160 (35)
	Unfunded	50 (10.9)
	Funding undisclosed	72 (15.8)

Characteristic	Categories	Number (%)
By disease area	Anemia in chronic kidney disease	56 (12.3)
	Angelman syndrome	37 (8.1)
	Crohn's disease	43 (9.4)
	Diabetic gastroparesis	20 (4.4)
	Diabetic macular edema	39 (8.5)
	Diabetic macular ischemia	23 (5)
	Episodic and chronic migraine	81 (17.7)
	Hemophilia	38 (8.3)
	Sleep disturbances due to pruritis	53 (11.6)
	Treatment-resistant depression	29 (6.3)
	Wet age-related macular degeneration	38 (8.3)
Characteristic	**Mean (SD)**	**Median (range)**
NOS score	6.6 (1.03)	7 (3 - 9)
IF	5.2 (4.5)	3.9 (0.2 - 39)

[1] 3 case-control, 2 ambispective
[2] spanning more than 1 continent
[3] 5 Turkey, 2 Brazil, 1 Egypt, 1 Israel
[4] academia, government, non-governmental organization -(NGO) and not-for-profit organization (NPO)

Key findings

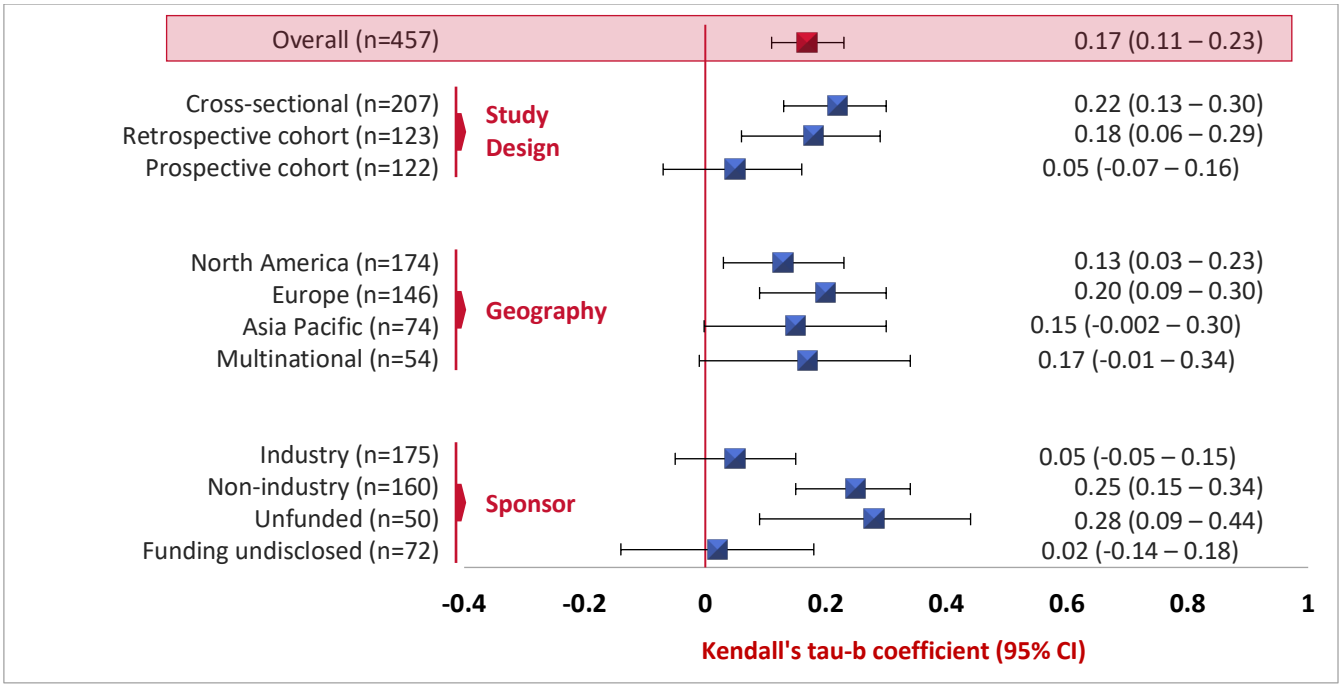
To determine the quantitative relationship between NOS score and journal IF, Kendall's tau-b correlation (a nonparametric correlation) was calculated. As shown in Figure 1 below, overall, there was a **weak positive correlation between NOS score and IF [Kendall's tau-b = 0.17 (95% CI: 0.11 – 0.23), p<0.001]** for the overall sample of 457 studies.

Based on **study design**, there was a weak positive correlation between NOS score and IF for cross-sectional and retrospective cohort studies whereas there was no correlation between NOS score and IF for prospective cohort studies.

Based on **geography**, there was a weak positive correlation between NOS score and IF for all major regions, although only the findings from North America and Europe were statistically significant perhaps because of their relatively large sample sizes.

Finally, based on **sponsor**, there was a weak positive correlation between NOS score and IF for non-industry funded and unfunded studies whereas there was no correlation between NOS and IF for industry-funded studies and studies with undisclosed funding.

Figure 1: Correlation between IF and NOS score: overall and as a function of study characteristics

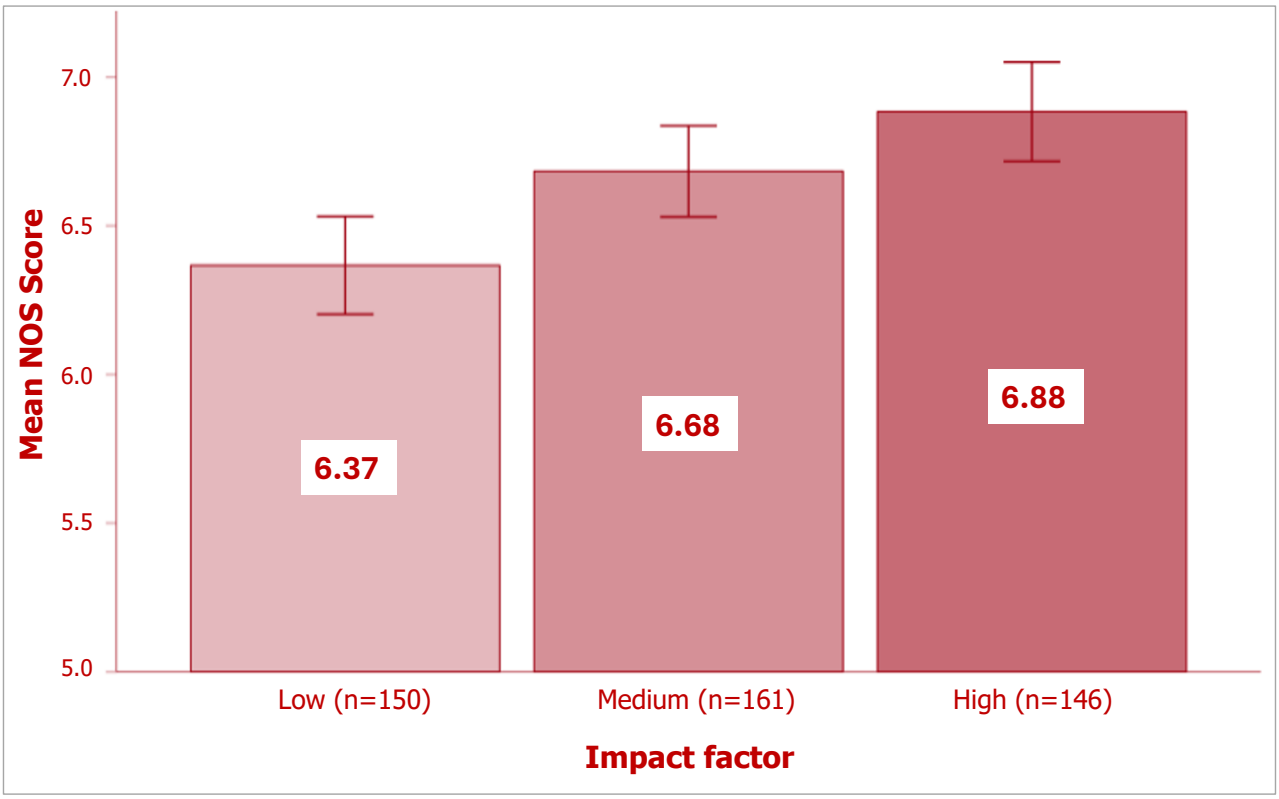


One-way ANOVA was conducted to examine the differences in mean NOS scores across the 3 categories of IF based on tertiles: low (≤3.2), medium (3.3-4.9), and high (≥5) impact. As shown in Figure 2, there was a statistically significant difference in the mean NOS score between the 3 IF groups as determined by one-way ANOVA [F(2,454) = 9.94, p <0.001].

The effect size, eta squared (η^2), was **0.04 (95% CI: 0.01-0.08)**, indicating a small effect (such that only 4% of variation in the NOS score was accounted for by the journal IF).

A post-hoc Bonferroni test showed that the mean NOS score was significantly higher in both high IF and medium IF groups compared to the low IF group (p<0.001 and p=0.02 respectively); however, there was no statistically significant difference in the mean NOS score between high and medium IF groups (p=0.25).

Figure 2: Distribution of NOS score by IF categories



The error bars represent 95% CIs of the mean NOS score

Discussion

The overall relationship between IF and NOS score is positive but weak. Consequently, the IF of a journal is not always a reliable measure of the quality of an individual paper, and cannot replace a careful critical appraisal of underlying research. It is possible that some journals may prioritize novelty over methodological rigor, leading to discrepancies in research quality even among journals with similar IFs. Moreover, journals may be more inclined to publish studies with statistically significant results, leading to publication bias. This can result in high-impact journals publishing studies that are not necessarily of higher quality but are more likely to attract attention and citations. The common theme across these arguments is that the methodological quality of a research paper might be just one of the many factors that a journal considers in its editorial decision-making. More research is needed to understand what those other factors might be.

Another key finding of our study that warrants some discussion is the lack of correlation between NOS score and IF in industry-funded studies. So, how does an industry-funded study with low quality find its way into a high-impact journal, and how does an industry-funded study with high quality end up in a low-impact journal? While it was beyond the scope of this paper to investigate this further, there are several ideas that are worthy of future investigation. For example, are industry-funded studies more likely to report novel and statistically significant findings compared to non-industry-funded studies? How are industry-funded studies perceived by journal editors and by peer-reviewers?

We also found a lack of correlation between NOS score and IF for prospective cohort studies. Although this finding needs further evaluation, it is possible that prospective cohort studies, despite being of low-to-moderate quality, are likely to attract journals' attention simply by virtue of their design, as a prospective design is inherently associated with a lower risk of bias compared to a case-control or a cross-sectional design.

The **strengths** of our pilot study include a large sample size of 457 (this is important since, as we have stated, studies of this nature are done infrequently) and that it covers a diverse range of disease areas, making the results more generalizable to the observational literature. Further, a consecutive series of 11 SLR projects was chosen, reducing selection bias in the identification of research papers. The same research team was used across all 11 projects, reducing variability in assessment of quality (members of the team participated in the same intensive training on critical appraisal using the NOS). Finally, 2 independent researchers scored each research paper during critical appraisal, potentially reducing subjectivity in the assessment.

Some **caveats** of this pilot study require acknowledgment. Only one tool (the NOS) was used to assess research quality. Whilst being acknowledged for its ease of use and a convenient scoring system, the NOS has also been criticized for low inter-rater reliability, and its use as a "quantitative" rating scale is not well-established [8-9]. While IF can provide some insights into the visibility and influence of a journal within its field, it is only one indicator of journal quality. IFs of journals are field-dependent and not comparable across different disease/therapeutic areas. As an example, a "top" journal publishing research on rare diseases (i.e., with very narrow scope) can have an IF which might be lower than the IF of an "average" journal publishing research on a common disease area (i.e., with a broad scope). IFs change differently over time for different journals; however, their rate of change is low. Further, the applicability of these findings to the clinical trial literature cannot be assumed. This study does not allow for causal inferences to be drawn on the relationship between NOS score and IF. Finally, the interpretation of effect sizes (Kendall's tau-b and eta-squared) is context-dependent, and Cohen's guidelines should not be used as strict thresholds but rather as a general reference to help interpret the practical significance of findings.

Despite these limitations, the findings of this study have several important **implications** for various stakeholders. Clinicians, researchers and policy makers [and indeed AI models, an additional focus of ours] must be trained to critically appraise the methodological quality of an original research paper to make informed decisions based on the best available evidence – they should not rely on the perceived 'prestige' of the journal. There may be a need for educational initiatives to help researchers, clinicians, and other stakeholders understand the limitations of journal-based metrics, and critically evaluate the quality of research. Journals should consider conducting, and subsequently publishing, a formal quality assessment of research papers using a validated tool as part of their peer-review process.

There are several avenues for **future research** in this area, in addition to the ones suggested earlier. Future studies should attempt to confirm our findings by using other tools or checklists for quality assessment of observational research as well as other indicators of journal quality such as the type of peer review, reputation within the field, editorial policies, and other metrics such as Eigenfactor score. Investigating the extent to which publication bias influences the relationship between journal quality and research quality is also an important research area. Finally, qualitative research methods, such as interviews and surveys with researchers, editors, and peer reviewers, can also provide insights into the perceived importance of journal quality and its impact on research practices.

In summary, while there is some correlation between journal quality and observational research quality, it is essential to recognize that they are not synonymous. High-quality research can be found in journals of varying IFs, and assessing research quality requires careful consideration of factors such as study design, methodology, analysis, interpretation, and significance of findings.

Authors: Dr. Digant Gupta, MBBS, MPH, Dr. Amandeep Kaur, MS, PhD, Mansi Malik, MSc

References

1. Nascimento, D.P., et al., Journal impact factor is associated with PRISMA endorsement, but not with the methodological quality of low back pain systematic reviews: a methodological review. Eur Spine J, 2020. **29**(3): p. 462-479.
2. Saginur, M., et al., Journal impact factor, trial effect size, and methodological quality appear scantly related: a systematic review and meta-analysis. Syst Rev, 2020. **9**(1): p. 53.
3. Ahmed Ali, U., et al., Journal impact factor and methodological quality of surgical randomized controlled trials: an empirical study. Langenbecks Arch Surg, 2017. **402**(7): p. 1015-1022.
4. Gluud, L.L., et al., The journal impact factor as a predictor of trial quality and outcomes: cohort study of hepatobiliary randomized clinical trials. Am J Gastroenterol, 2005. **100**(11): p. 2431-5.
5. Lee, K.P., et al., Association of journal quality indicators with methodological quality of clinical research articles. Jama, 2002. **287**(21): p. 2805-8.
6. Wells, G., et al., The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Ottawa: Ottawa Hospital Research Institute, 2012.
7. Herzog R, Álvarez-Pasquin MJ, Díaz C, et al. Are healthcare workers' intentions to vaccinate related to their knowledge, beliefs and attitudes? A systematic review. BMC Public Health 2013;13:154. doi: 10.1186/1471-2458-13-154.
8. Hartling L, Milne A, Hamm MP, et al. Testing the Newcastle Ottawa Scale showed low reliability between individual reviewers. J Clin Epidemiol 2013;66(9):982-93. doi: 10.1016/j.jclinepi.2013.03.003 [published Online First: 2013/05/21]
9. Stang A. Critical evaluation of the Newcastle-Ottawa scale for the assessment of the quality of nonrandomized studies in meta-analyses. Eur J Epidemiol 2010;25(9):603-5. doi: 10.1007/s10654-010-9491-z [published Online First: 2010/07/24]
10. Cohen, J., A power primer. Psychol Bull, 1992. 112(1): p. 155-9.
11. Bishara, A.J. and J.B. Hittner, Testing the significance of a correlation with nonnormal data: comparison of Pearson, Spearman, transformation, and resampling approaches. Psychol Methods, 2012. 17(3): p. 399-417.
12. Cohen, J., Statistical power analysis for the behavioral sciences. 2013: Academic press.

Appendix 1: Newcastle-Ottawa scales by study designs

	Cohort	Case-Control	Cross-sectional
Number of questions/items	8 items	8 items	7 items
Domains (number of items)	**3 domains:** • Selection (4) • Comparability (1) • Outcome (3)	**3 domains:** • Selection (4) • Comparability (1) • Exposure (3)	**3 domains:** • Selection (4) • Comparability (1) • Outcome (2)
Questions (and possible stars)	**Selection domain** 1) Representativeness of the exposed cohort (*) 2) Selection of the non-exposed cohort (*) 3) Ascertainment of exposure (*) 4) Demonstration that outcome of interest was not present at start of study (*) **Comparability domain** 5) Comparability of cohorts based on the design or analysis (**) **Outcome domain** 6) Assessment of outcome (*) 7) Follow-up long enough for outcomes to occur (*) 8) Adequacy of follow-up of cohorts (*)	**Selection domain** 1) Is the case definition adequate? (*) 2) Representativeness of the cases (*) 3) Selection of controls (*) 4) Definition of controls (*) **Comparability domain** 5) Comparability of cases and controls on the basis of the design or analysis (**) **Exposure domain** 6) Ascertainment of exposure (*) 7) Same method of ascertainment for cases and controls (*) 8) Non-response rate (*)	**Selection domain** 1) Representativeness of the sample (*) 2) Sample size (*) 3) Non-respondents (*) 4) Ascertainment of exposure (**) **Comparability domain** 5) Comparability of different outcome groups based on study design or analysis (**) **Outcome domain** 6) Assessment of outcome (*) 7) Statistical test (*)
Total stars	9	9	9

Appendix 2

The association between IF and NOS score was examined in two ways

1 Kendall's tau-b correlation (a nonparametric alternative to the Pearson's correlation) along with their 95% confidence intervals were calculated to determine the quantitative relationship between NOS score and journal IF, both for the overall sample (n=457) as well as for different subgroups based on selected stratifying variables (study design, geography, and type of sponsor). Using Cohen's guidelines, **r = 0.10, r = 0.30,** and **r = 0.50** were recommended to be considered as cut-offs for **small, medium**, and **large** effect sizes, respectively [10]. As part of sensitivity analysis, to assess the possible influence of sampling bias on the results, bootstrap estimation based on 1000 random samples with replacement was used to generate bias-corrected and accelerated (BCa) confidence intervals for the correlation coefficient [11]. Bootstrapping estimation technique does not assume any level of normally distributed data and therefore tends to be more robust with skewed data.

2 One-way ANOVA was used to examine the mean NOS scores across the 3 categories of IF based on tertiles: **low (<=3.2), medium (3.3-4.9),** and **high (>=5)** impact. The assumption of homogeneity of variance (i.e., variances of NOS scores are equal across IF groups) was assessed using Levene's test. Brown-Forsythe test and Welch test were used as robust ANOVA procedures if the homogeneity of variance assumption was not met. Bonferroni post-hoc test (assuming equal variances) or Tamhane's T2 test (assuming unequal variances) was used to explore pairwise differences in mean NOS scores across different IF groups. Eta-squared (η^2) was calculated as the measure of effect size which indicates the proportion of variation in the NOS score accounted for by the journal IF. Using Cohen's guidelines, the following benchmarks for judging effect size based on η^2 were used: **small (0.01 - 0.059), medium (0.06 - 0.139),** and **large (>=0.14)** [12].