BMJ Open

# Influence of the statistical significance of results and spin on readers' interpretation of the results in an abstract for a hypothetical clinical trial: a randomised trial

Sofyan Jankowski ,[1,2] Isabelle Boutron,[1] Mike Clarke [2]

**To cite:** Jankowski S, Boutron I, Clarke M. Influence of the statistical significance of results and spin on readers' interpretation of the results in an abstract for a hypothetical clinical trial: a randomised trial. *BMJ Open* 2022;**12**:e056503. doi:10.1136/bmjopen-2021-056503

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2021-056503).

Received 27 August 2021
Accepted 28 February 2022



[1]Université Paris Cité, INSERM, INRAE, CNAM, Centre for Research in Epidemiology and Statistics (CRESS), F-75004, Paris, France
[2]Centre for Public Health, Queen's University Belfast, Belfast, UK

**Correspondence to**
Dr Mike Clarke;
m.clarke@qub.ac.uk

## ABSTRACT

**Objectives** To assess the impact on readers' interpretation of the results reported in an abstract for a hypothetical clinical trial with (1) a statistically significant result (SSR), (2) spin, (3) both an SSR and spin compared with (4) no spin and no SSR.

**Participants** Health students and professionals from universities and health institutions in France and the UK.

**Interventions** Participants completed an online questionnaire using Likert scales and free text, after reading one of the four versions of an abstract about a hypothetical randomised trial evaluating 'Naranex' and 'Bulofil' (two hypothetical drugs) for chronic low back pain. The abstracts differed in (1) reported result of 'mean difference of 1.31 points (95% CI 0.08 to 2.54, p=0.04)' or 'mean difference of 1.31 points (95% CI −0.08 to 2.70, p=0.06)' and (2) presence or absence of spin. The effect size for the trial's primary outcome (pain disability score) was the same in each abstract, slightly in favour of Naranex.

**Primary outcome** The reader's interpretation of the trial's results, based on their answer (1, disagree; 4, neutral; 7, agree) to the following statement: 'About the main findings of the study, what is your opinion about the following statement: 'Naranex is better than Bulofil'?'

**Results** Two hundred and ninety-seven of the 404 people randomised to receive one of the four abstracts completed the study. Respondents were more likely to favour Narenex when the abstract reported an SSR without spin, a statistically significant result with spin, a non-statistically significant result with spin, compared with when it reported a non-SSR without spin.

**Conclusion** Statistical significance appears to have influenced readers' perception whatever the level of spin, while spin influenced readers' perception when the results were not statistically significant but did not appear to have an impact when results were statistically significant.

## Strengths and limitations of this study

- The study design used a 2×2 randomised trial to compare abstracts containing different types of information.
- This randomised trial involved both students and professionals.
- Twenty-six per cent of the randomised participants did not complete the study, but this attrition was balanced between groups.
- The study is based on a single hypothetical abstract and should be reproduced using other samples of abstract and other populations, greater variations in the point estimate for the effect and its 95% CI, and with other amounts of information on the reported trial to see if the findings are generalisable.

## INTRODUCTION

Randomised trials are a key element in evidence-based medicine, particularly when brought together in systematic reviews. Readers need to be able to interpret the results accurately.

The use of p values in the conventional, dichotomous way indicating statistical significance and non-significance and an over-reliance on statistical significance to determine if an intervention is, or is not, effective has been criticised, particularly because statistical significance does not imply scientific importance.[1 2] Further, previous studies have shown that spin, defined as reporting practices so that results are viewed in a more favourable light, is prevalent in published reports.[3–6]

This study assesses the impact on health students' and professionals' interpretation of the results reported in an abstract of a clinical trial with (1) a statistically significant result (SSR), (2) spin, (3) both an SSR and spin compared with an abstract with no spin and no SSR. This question is important, considering the need to ensure that students and professionals are able to critically interpret research results presented in an abstract and in the context of debate on the reporting of p values and CIs.

## METHODS

### Design

We used a web-based 2×2 factorial randomised trial to compare readers' interpretation of abstracts reported with or without spin ('spin' or 'no spin') and with a result that was reported as statistically significant ('mean difference of 1.31 points (95% CI 0.08 to 2.54, p=0.04)') or not ('mean difference of 1.31 points (95% CI −0.08 to 2.70, p=0.06)'). Both abstracts presented the same mean difference. We generated four different abstracts reporting a trial comparing two hypothetical drug treatments ('Naranex' and 'Bulofil') for chronic low back pain, with the direction of benefit appearing to favour Naranex. We report the study in accordance with guidance from the Checklist for Reporting Results of Internet E-Surveys[7] and Consolidated Standards of Reporting Trials (CONSORT),[8] and have tried to ensure that we have not fallen victim to using spin in our own reporting.[9]

### Abstract construction

We created an abstract that contained simple notions, which are likely to be well understood by readers. We prepared the abstract using ideas gleaned from several published trials in patients with low back pain[10–15] and the abstracts we used are shown in table 1. Each of the four abstracts had the following key characteristics:

- Two-group randomised trial: the most common design for randomised trials that is easy to interpret.
- Patients with chronic low back pain: a common condition that is easily understood.
- Comparison of two new drugs (which are described as equivalent to non-steroidal anti-inflammatory drugs (NSAIDs), a typical treatment in chronic low back pain) but noting that these new drugs had not previously been compared against each other.
- Hypothetical drugs (Naranex and Bulofil) to avoid using the names of real drugs that might have led to influence from participants' pre-existing knowledge, preferences or experience about the named drugs and to avoid anyone from regarding the evidence in these abstracts as something that they might use in making decisions about treating chronic low back pain.
- Pain disability as the primary outcome of the trial, measured by the Roland-Morris Disability Questionnaire (0–24).
- Pain intensity measured on a rating scale (0–10) as the trial's secondary outcome.
- Treatment effect estimates for the primary and secondary outcomes were the same in all abstracts, and both had wide 95% CIs.

The background, objective and methods in each of the four abstracts were identical, but we varied the content of their title, results and conclusions. These variations related to the two main factors we wished to study, which were

- Statistical significance of the effect estimates for the primary (pain disability) and secondary outcome (pain intensity) (based on the p-value of the statistical test and the 95% CI). The results were reported with either
  p=0.06, mean difference of 1.31 (favouring Naranex) but lower limit of the 95% CI below 0.
  p=0.04, mean difference of 1.31 (favouring Naranex) and lower limit of the 95% CI above 0.
- Presence of spin: in the abstract with spin, we added three known spin strategies[6]: spin in the title ('Naranex improved patient's condition compared with Bulofil…'), a focus on statistically significant secondary outcomes (pain intensity) or subgroup analyses (compliant patients and women) and linguistic spin ('particularly large' and 'much better') to emphasise the benefit of Naranex compared with Bulofil.

### Participants

Participants were any health students or professionals who were willing to participate. They were invited through the following channels from 2 April to 17 June 2017, and our aim was to recruit as many as possible in the time available.

1. We contacted 150 associations of French health students (medicine, dentistry, midwifery and pharmacy), of which 50 agreed to advertise our study on their private Facebook group wall.
2. We posted an invitation to participate on the Facebook group wall of four public Facebook groups of health professionals in France.
3. The Centre for Public Health in the School of Medicine, Dentistry and Biomedical Sciences at Queen's University Belfast, UK, advertised the study on its internal email notification list.

People were informed in the social media post that they would have to answer a dozen questions about a summary of the results of a research study. They were informed that they would receive a randomly selected summary by email with a multiple-choice questionnaire to complete. If they were willing to participate, they had to provide their email address and consent by ticking a box indicating 'I want to receive the participation link, and to answer the study questionnaire'.

### Outcomes

We used a questionnaire that was to be answered using Likert scales (online supplemental appendix 1) and free text. Our primary outcome measure was the reader's interpretation of the results presented in the abstract, which was collected from their answer to the following statement on a 7-point Likert scale (1, disagree; 4, neutral; and 7, agree):

> About the main findings of the study, what is your opinion about the following statement: 'Naranex is better than Bulofil'?

The use of the word 'better' in the primary outcome question was to capture the intuitive feeling of the reader.

**Table 1** Experimental abstracts with the variations regarding p value and spin

| Abstract with no SSR and no spin | Abstract with no SSR and spin | Abstract with SSR and no spin | Abstract with SSR and spin |
|---|---|---|---|
| Title:<br>Naranex vs Bulofil for patients with chronic low back pain: a randomised controlled trial | | Title:<br>Naranex *improved* patient's condition compared with Bulofil in chronic low back pain: a randomised controlled trial | |
| Background<br>Chronic low back pain* is a common condition and causes significant pain, distress and disability across the world. Its causes are multifactorial and it is challenging to manage. NSAIDs† are widely used as first-line therapy, but side effects, such as gastrointestinal toxicity, can occur with long-term use. Naranex and Bulofil have been shown to have similar effects to NSAIDs, but there have been no studies directly comparing these two new drugs.<br>Objective<br>To compare the effects of Bulofil and Naranex for the treatment of chronic low back pain, in a randomised controlled trial‡ (ISRCTN053139911).<br>Methods<br>We did a multicentre, randomised controlled trial in 18 primary care centres in Northern Ireland from 12 November 2014 to 5 March 2016. We randomly allocated patients with chronic low back pain to receive either regular doses of Bulofil (200 mg three times per day) or Naranex (200 mg three times per day). All patients also received best-evidence advice and were followed up for 1 year. Patients and staff at all centres were masked to treatment allocation. The primary efficacy measure was the pain disability (RMDQ)§. Secondary outcomes were pain intensity during the last 24 hours (Numerical Rating Scale from 0 to 10)¶, healthcare resource use, safety, tolerability and compliance.** | | | |
| Results<br>In total, 120 patients were randomised to receive either Naranex (n=63) or Bulofil (n=57).<br>The pain disability (RMDQ score) at 6 months was reduced by 4.87±1.83 in the Naranex group and by 3.56±1.52 in the Bulofil group: between-group difference of 1.31 points **(–0.08 to 2.70, p=0.06),** no statistically significant difference.<br>We found no statistically significant difference in pain intensity: between-group difference of 0.59 points **(–0.09 to 1.27, p=0.07**).<br>Healthcare resource use, safety and tolerability were not statistically different: patients with non-severe adverse events were 5/63 in the Naranex group and 6/57 in the Bulofil group (p=0.89). | Results<br>In total, 120 patients were randomised to receive either Naranex (n=63) or Bulofil (n=57).<br>The pain disability (RMDQ score) at 6 months was reduced by 4.87±1.83 in the Naranex group and by 3.56±1.52 in the Bulofil group. Naranex *improved* the RMDQ score at 6 months compared with Bulofil: between-group difference of 1.31 points **(–0.08 to 2.70, p=0.06)**.<br>Moreover, the RMDQ score was much better for compliant people in the Naranex group compared with those in the Bulofil group: between-group difference of 2.4 points (0.26 to 4.54, p=0.02). For women, pain intensity improvement was much better with Naranex compared with Bulofil: between-group difference of 0.91 points (0.10 to 1.72, p=0.04). Non-severe adverse events were less common in the Naranex group (5/63) than in the Bulofil group (6/57), p=0.89. | Results<br>In total, 120 patients were randomised to receive either Naranex (n=63) or Bulofil (n=57). The pain disability (RMDQ score) at 6 months was reduced by 4.87±1.83 in Naranex group and by 3.56±1.52 in Bulofil group: statistically significant between-group difference of 1.31 points **(0.08 to 2.54, p=0.04)**. We found a statistically significant difference in pain intensity in favour of Naranex: between-group difference of 0.59 points **(0.09 to 1.09, p=0.04)**.<br>Healthcare resource use, safety and tolerability were not statistically different: patients with non-severe adverse events were 5/63 in the Naranex group and 6/57 in the Bulofil group (p=0.89). | Results<br>In total, 120 patients were randomised to receive either Naranex (n=63) or Bulofil (n=57).<br>The pain disability (RMDQ score) at 6 months was reduced 4.87±1.83 in Naranex group and by 3.56±1.52 in the Bulofil group. Naranex *improved* the RMDQ score at 6 months compared with Bulofil: statistically significant between-group difference of 1.31 points **(0.08 to 2.54, p=0.04**). Moreover, the RMDQ score was much better for compliant people in the Naranex group compared with those in the Bulofil group: between-group difference of 2.4 points (0.26 to 4.54, p=0.02). Pain intensity improvement was better with Naranex: between-group difference of 0.59 points **(0.09 to 1.09, p=0.04**). That improvement was much better for women: 0.91 points (0.10 to 1.72, p=0.04). Non-severe adverse events were *less common* in the Naranex group (5/63) than in the Bulofil group (6/57), p=0.89. |

Continued

Table 1 Continued

| Abstract with no SSR and no spin | Abstract with no SSR and spin | Abstract with SSR and no spin | Abstract with SSR and spin |
|---|---|---|---|
| Conclusions<br>We did not find a statistically significant difference between Naranex and Bulofil on pain disability improvement in patients with chronic low back pain. No statistically significant differences were found for pain intensity, and safety. | Conclusions<br>Naranex *improved* the pain disability compared with Bulofil. That improvement is *particularly large* for patients who take their medication. Pain intensity was also *significantly much better* for women who took Naranex. *Our results support the effectiveness and the safety of Naranex for chronic low back pain and suggest that it might be effective for patients with other types of chronic pain*. | Conclusions<br>Naranex was significantly better than Bulofil on pain disability improvement in patients with chronic low back pain. Naranex was also statistically significantly better for pain intensity. No statistically significant differences were found in terms of healthcare resource use and safety. | Conclusions<br>Naranex *improved* statistically significantly the pain disability compared with Bulofil. That improvement is *particularly large* for patients who take their medication. Pain intensity was also *significantly much better* in the Naranex group, with an *important effect* for women. *Our results support the effectiveness and the safety of Naranex for chronic low back pain and suggest that it might be effective for patients with other types of chronic pain*. |

To highlight the differences in this table, p values and 95% CI for primary and secondary outcome are shown in bold; primary and secondary outcome subgroup analysis on compliant people and women are underlined; and linguistic spin is italicised. No such highlighting was used in the abstracts used in the study.
*Chronic low back pain: low back pain that has lasted for more than 3 months.
†NSAIDs: a class of drugs that includes aspirin and ibuprofen.
‡Randomised controlled trial: participants are randomly selected to receive one of the treatments being assessed.
§RMDQ (*pain disability*): 24-item self-report questionnaire to assess the effects of low back pain on functional activities, ranging from 0 (no disability) to 24 (severe disability).
¶Numerical Rating Scale (*pain intensity*): self-report scale for assessing pain, ranging from 0 (no pain) to 10 (extreme pain).
**Compliance: in this case, this means taking the allocated drug.
NSAID, non-steroidal anti-inflammatory drug; RMDQ, Roland-Morris Disability Questionnaire; SSR, statistically significant result.

We also used Likert scales to gather each reader's opinions on whether they would (1) use either drug if they had chronic low back pain and (2) fund future trials to test either drug. They were also asked to record their level of trust in the findings, whether they felt that the abstract could have come from a publication in an important, high-impact medical journal and whether it was similar to other scientific summaries they have read. The full questionnaire is in the online supplemental appendix.

### Participant and public involvement

The online questionnaire was revised following pilot testing and input from 30 health students (whose data are not included in the analyses presented here). Participants who wished to do so could register an e-mail address to which the published results will be sent.

### Randomisation and blinding

One researcher (SJ) followed these steps:

1. Extracting the email addresses from the social media opt-in participation form.
2. Randomising, using a 1.1.1.1 ratio in blocks of 4 using Microsoft Excel V. 2016 when each group of four volunteers was available.
3. Sending the abstract and study link to each randomised participant in accordance with their random allocation (online supplemental appendix 2).
4. Reminders were sent every 10 days, if there had been no reply. A maximum of three reminders was sent to each person.

This allocation method is unlikely to lead to bias because participants were recruited through social media and had no prior contact with the investigator. Baseline characteristics were assessed after randomisation from the questions in the study questionnaire that was emailed to volunteers along with their randomly allocated abstract. The investigator had no direct interactions with any of the participants, and reminders were sent regardless of the allocated abstract.

Volunteers were blinded to the study hypothesis. They were informed that the study aimed to 'evaluate the influence of the results presentation format (table, plots, abstract…) on reader's comprehension'. This dummy objective was stated as the objective in all study-related messages and online information about the study. In this way, the volunteers were not aware that we were focussing on spin and statistical significance of the results.

### Statistical analysis

We did not calculate an a priori sample size because we did not have data to allow an accurate calculation.[16] Instead, our aim was to recruit as many participants as possible to increase the precision of the analyses.

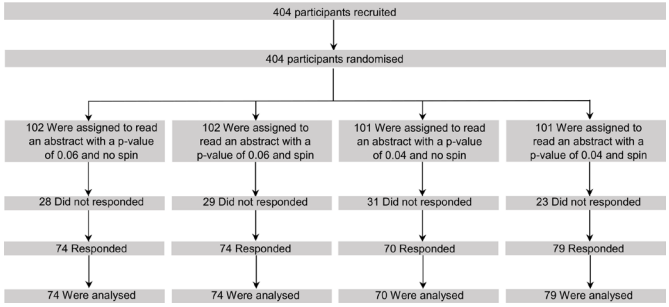


**Figure 1** Study profile.

Since the primary and secondary outcomes were ordinal and not normally distributed, we used non-parametric tests for all analyses. The Scheirer-Ray-Hare test was used to assess the impact of spin, 'SSR' and the interaction of both on the outcomes. The Mann-Whitney U test was used as a post-hoc test for four prespecified pairwise comparisons and a Bonferroni correction was applied to account for multiplicity in analysis.

As a post-hoc sensitivity analysis, we also show the results stratified on participants' background (student vs professional) to examine if these subgroup results were consistent with the main analysis. Analyses were performed with R software 3.0.0.

## RESULTS

### Participants

The flow of participants is shown in figure 1. Participants were recruited from 19 April to 13 June 2017. In brief, 404 people volunteered to join the study and were randomised to one of the four experimental groups: 'no SSR and no spin' (n=102), 'no SSR and spin' (n=102), 'SSR and no spin' (n=101) and 'SSR and spin' (n=101). We analysed the responses of the 297 participants who returned a completed questionnaire: 'no SSR and no spin' (n=74), 'no SSR and spin' (n=74), 'SSR and no spin' (n=70), and 'SSR and spin' (n=79). Table 2 shows the baseline characteristics of these participants. They were based in the UK (15%, n=45) or France (85%, n=252). The mean age was 26 (SD 8) years. There were 240 (81%) students and 57 (19%) healthcare or research professionals.

### Outcomes

Effects of spin, SSR and interaction of both on the primary outcome were significant (p<0.0001) indicating that the effect of SSR is not the same regardless of whether a spin was present, and vice versa. Consequently, we analysed the impact of SSRs with and without spin by considering the four groups separately. Figure 2 shows that participants were more likely to favour Narenex when abstracts reported a SSR without spin [median (IQR): 7 (6;7)]; a SSR with spin [7 (6;7)], or a non-SSR with spin [6 (5;7)] compared with a non-SSR without spin [1 (1;1)] (p<0.0001). The two other possible pairwise comparisons were not related to our primary objectives but showed no statistically significant difference between the abstracts with an SSR reported without spin or with spin.

A post hoc analysis for the primary outcome stratified on participants' background (students vs professionals) shows that the findings in each subgroup are visually consistent with each other and the main analyses (figure 3).

Secondary outcomes are consistent with primary outcome analysis, although the interaction was not significant for each outcome.

Full results for the other questions are provided in online supplemental appendix 3. Data are available in an open access repository.[17]

## DISCUSSION

### Summary of finding

Our results show that readers' interpretation is influenced by the presence of SSRs regardless of the presence or absence of spin. Our results also confirm the influence of spin on readers' perception when the results were not statistically significant. However, spin did not appear to have an impact when the results were statistically significant, but this result may have been affected by a ceiling effect because participants gave the maximum score on the Likert scale for abstracts with a statistically significant difference without spin and abstracts with a statistically significant difference with spin.

### Comparison with previous studies

In the wake of the work on clinical trial misreporting,[18] Boutron *et al* proposed a definition of spin as the 'use of specific reporting strategies, from whatever motive, to highlight that the experimental treatment is beneficial, despite a statistically non-significant difference for the primary outcome, or to distract the reader from statistically nonsignificant results'. They developed a classification scheme to standardise strategies used for spin,[6] and up to 70% of the biomedical research literature has been found to contain spin.[3–5] In a study assessing the impact of spin on readers, Boutron *et al* studied 300 oncologists who were experienced in clinical research. They were randomly allocated to read either a spin or non-spin abstract and asked to answer the question 'Based on this abstract, do you think treatment A would be beneficial to patients?' on a scale from 0 (very unlikely) to 10 (very likely). The presence of spin favouring treatment A produced a statistically significant higher score on this Likert scale.[19] Subsequently, Shinohara *et al* asked the same question to primary care physicians randomly allocated to read an abstract with or without overstatements in the conclusion. They concluded that when sufficient information is provided and standardised in other sections, there is no effect of overstatements in the abstract's conclusion.[20]

Table 2 Baseline characteristics of participants

| Abstract | No SSR, no spin | No SSR, spin | SSR, no spin | SSR, spin |
|---|---|---|---|---|
| Number analysed | 74 | 74 | 70 | 79 |
| Age (years), mean (SD) | 26 (8) | 26 (8) | 26 (9) | 26 (8) |
| Gender | | | | |
| Female | 57 (77) | 56 (76) | 56 (80) | 57 (72) |
| Male | 17 (23) | 18 (24) | 14 (20) | 22 (28) |
| Location | | | | |
| France | 65 (88) | 61 (82) | 60 (86) | 66 (84) |
| UK | 9 (12) | 13 (18) | 10 (14) | 13 (16) |
| Profession | | | | |
| Students | | | | |
| Medicine | 38 (51) | 34 (46) | 32 (46) | 41 (52) |
| Midwifery | 9 (12) | 14 (19) | 7 (10) | 12 (15) |
| Dentistry | 4 (5) | 2 (3) | 3 (4) | 1 (1) |
| Pharmacists | 4 (5) | 4 (5) | 7 (10) | 7 (9) |
| Others | 4 (5) | 5 (7) | 5 (7) | 7 (9) |
| Professionals | | | | |
| Medicine | 8 (11) | 8 (11) | 6 (9) | 4 (5) |
| Research | 5 (7) | 4 (5) | 8 (11) | 6 (8) |
| Others | 2 (3) | 3 (4) | 2 (3) | 1 (1) |
| Articles/abstracts read per year (relating to general scientific topics) | | | | |
| <1 | 7 (11) | 11 (16) | 5 (8) | 8 (11) |
| 1–5 | 22 (34) | 16 (24) | 18 (29) | 26 (35) |
| 6–10 | 11 (17) | 11 (16) | 16 (26) | 11 (15) |
| 11–20 | 9 (14) | 11 (16) | 8 (13) | 5 (7) |
| >20 | 16 (25) | 18 (27) | 15 (24) | 24 (32) |
| Clinical articles/abstracts read per year (relating to clinical studies with human participants) | | | | |
| <1 | 17 (23) | 20 (27) | 17 (24) | 19 (24) |
| 1–5 | 24 (32) | 16 (22) | 21 (30) | 28 (35) |
| 6–10 | 13 (18) | 9 (12) | 14 (20) | 8 (10) |
| 11–20 | 6 (8) | 11 (15) | 8 (11) | 11 (14) |
| >20 | 14 (19) | 18 (24) | 10 (14) | 13 (16) |
| Do you feel able to define 'RCT'?: 1 (not able) to 7 (able), median, IQR) | 7 (6–7) | 7 (5–7) | 7 (6–7) | 7 (5.5–7.0) |
| Do you feel able to define 'statistical significance': 1 (not able) to 7 (able), median (IQR) | 7 (6–7) | 7 (5–7) | 7 (5–7) | 7 (5–7) |

Data are shown as means (SD), medians (IQR) or numbers (%).
RCT, randomised controlled trial; SSR, statistically significant result.

## Implication of the results

The study reported here is particularly important because it is the first that we are aware of which shows a possible influence of statistical significance of a study's results and a possible interaction with the presence of spin in the report of a clinical trial. Our findings support the importance of careful interpretation of statistical significance when someone is considering the results of a clinical trial and reinforce concerns about the misuse of statistical testing and interpretation, and about the simple use of statistical significance to determine whether one treatment is better than another.[21] Other researchers have highlighted how the p value and statistical significance are misunderstood and misinterpreted,[22 23] but concerns have been raised about how the absence of the p value might lead authors to make claims for an important signal to fit a pre-existing narrative.[24] An alternative to abandoning the p value might be to lower the threshold from the conventional 0.05[25] or to adopt a wider CI than 95%, even if this leads to a need for larger sample sizes and the

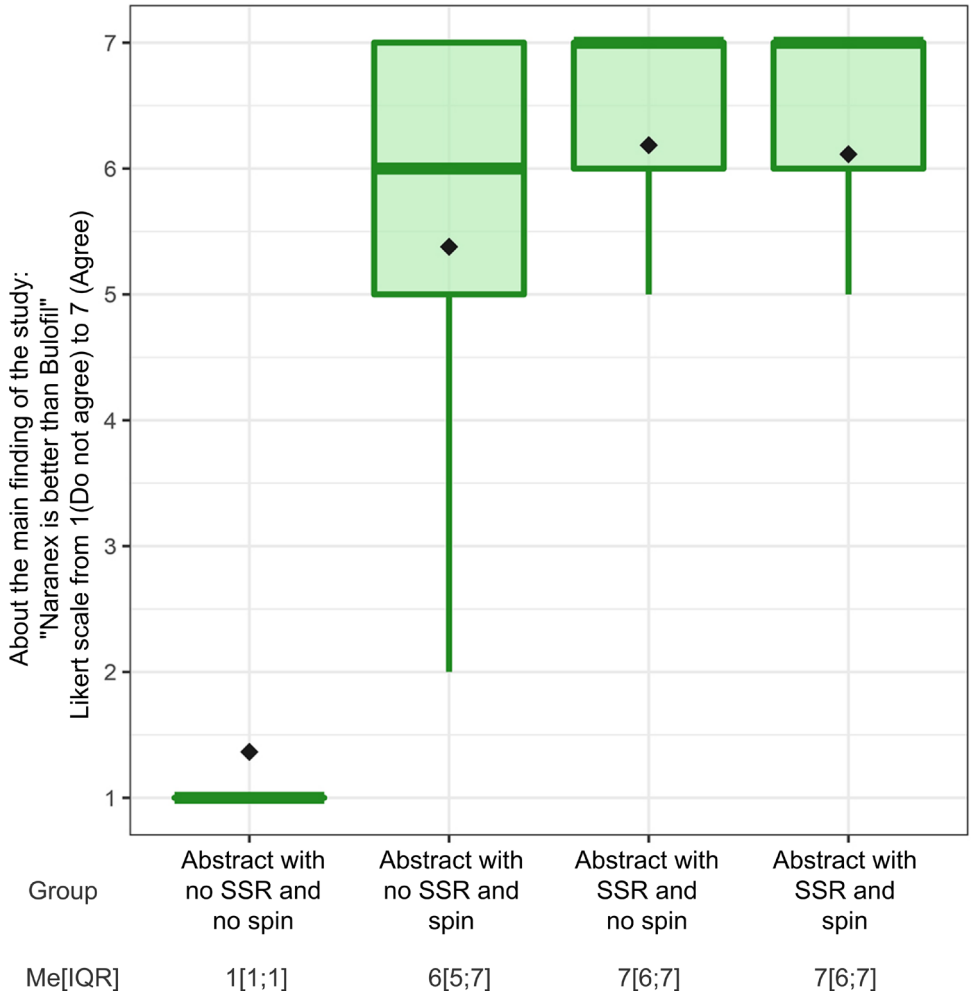


**Figure 2** Readers' assessment of the superiority of 'Naranex' compared with 'Bulofil' readers' assessment of the superiority of Naranex compared with Bulofilafter reading their allocated abstract of a randomised controlled trial reported with or without SSRs and with or without spin. Scores are based on a Likert scale, ranging from 0 (do not agree) to 7 (agree). Boxes represent median observations (horizontal rule) with 25th and 75th percentiles of observed data (top and bottom of the box). The diamonds represent the mean. The end of the vertical line represents the minimum values. IQR, considering first and third quartiles. Me, median; SSR statistically significant result.

consequent adjustments to funding strategies and trial design. This might also provide a solution to mitigate the effects of p-hacking,[26] selective outcome reporting and other forms of bias.

However, sustainable solutions are needed to improve readers' understanding and interpretation of the meaning of study findings. First, we may need to revisit guidance that supports reporting exact p values but fails to issue direction on specifying the a priori significance level.[27] Statistical analysis plans are rarely specified to this level of detail, even for randomised trials for which protocols might have been preregistered,[24] but if precise observed p values are to be reported (rather than whether it is larger or smaller than an arbitrary cut-off), the researcher's a priori significance level should be stated. Moreover, guidance on the interpretation of precise p values and their practical importance for clinical practice is needed. Some authors have suggested complementing the use of p values with Bayesian alternatives, such as the delta likelihood incorporating the p value and study power.[28 29] Others have proposed the reporting of parameters such as number needed to treat and effect sizes, or minimal clinically important difference.[30 31] Lastly, our study suggests a need to focus on educating healthcare students and practitioners on (1) interpreting point estimates while acknowledging the uncertainty around these, (2) recognising and avoiding false declarations of 'no difference', and (3) recognising and avoiding overconfident claims.[32]

## Strength and limitations

Some strengths of our study are the large number of participants and their diversity. We included health students from their first year to doctorate level and a variety of professionals in both France and the UK. Our study thereby complements the findings on a population of experienced oncologists by Boutron *et al*[19] and on primary care physicians by Shinohara *et al*.[20]

Our study has some limitations. First, 107 (26%) of the people who were randomised did not return a completed questionnaire and, because the questionnaire was used to

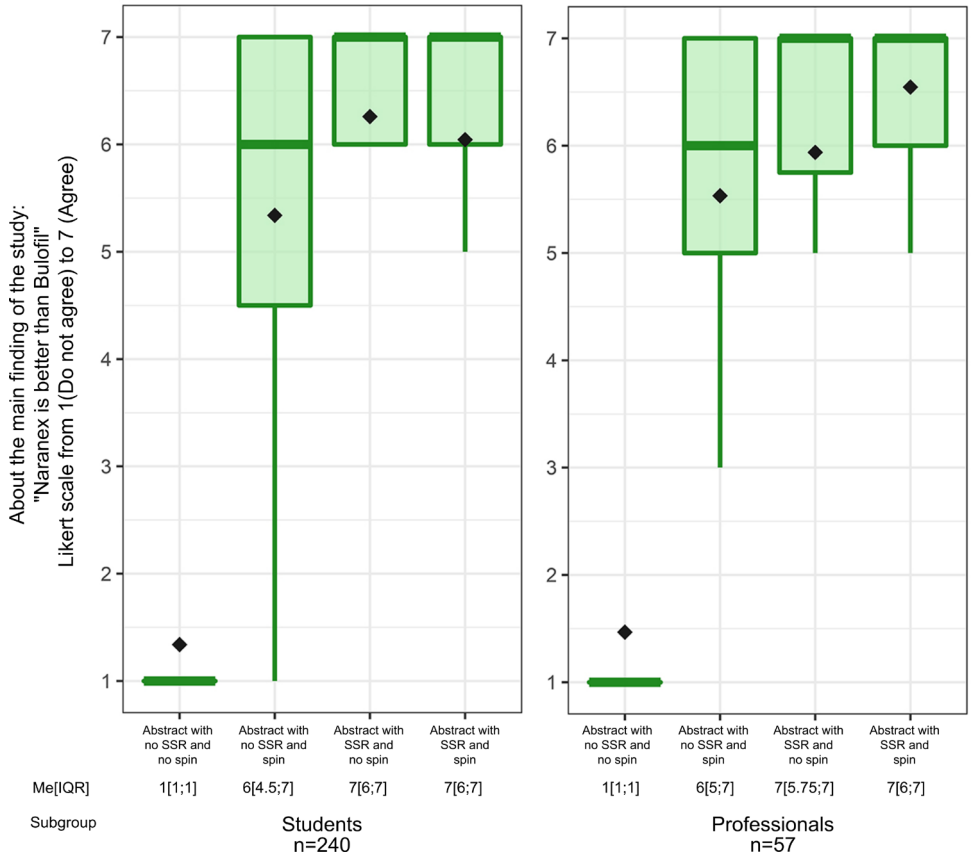


**Figure 3** Post hoc analyses for the primary outcome stratified on participant's background (students vs professionals) readers' assessment of the superiority of 'Naranex' compared with 'Bulofil' after reading their allocated sbstract of a randomised controlled trial reported with or without SSRs and with or without spin. Scores are based on a Likert scale, ranging from 0 (do not agree) to 7 (agree). Boxes represent median observations (horizontal rule) with 25th and 75th percentiles of observed data (top and bottom of the box). The diamonds represent the mean. The end of the vertical line represents the minimum values. IQR, considering first and third quartiles. Data are separated into two subgroups: students (n=240), defined as students from first year of university to PhD and residency, and professionals (n=57), defined as graduated professionals. Me, median; SSR statistically significant result.

collect information on their characteristics as well as their answers to the questions about their allocated abstract, we have no information that would allow us to compare them with the people who did respond and are in the analysis. Nevertheless, attrition is well balanced between the randomised groups, and it is unlikely that it is related to the person's allocated abstract. Second, although all participants were sent their randomly allocated abstract and the investigators could not influence this, the investigators were not blinded once the abstract had been allocated. Blinding can help to minimise allocation biases, conscious or unconscious selection of statistical tests and reporting but, in the context of this study, it is unlikely that a lack of blinding would result in important biases because participants were recruited through social media and had no contact with the investigators, and the intervention and assessment were provided online with no direct involvement of the investigator. Third, although this study was not a clinical trial comparing healthcare interventions and was thus not suitable for registration in a prospective clinical trial registry, the protocol was also not published in, for example, the Open Science Framework. Fourth, we used abstracts based on a single hypothetical trial in order to allow us to control the parameters precisely, which we based on existing abstracts,[10–15] but this means that, unlike previous experiments on spin, we did not use actual published abstracts.[19 20] However, it is also important to note that many published abstracts continue to fail to meet the reporting standards in, for example, the CONSORT statement.[33] We also used a large amount of spin, which we put into the abstract's title, results (with two subgroup analyses) and conclusions. This amount of spin was comparable to the real abstracts used by Boutron *et al*,[19] which showed a clear effect of spin on readers' interpretation. However, it is not possible to know whether one or more of these spin elements were most influential, or if all three were required to influence the readers. For example, when focusing on a subtype of spin, such as 'overstatement' in the conclusion of an abstract (defined as inconsistency between the results of primary outcomes in full text and those deduced from the abstract conclusion), Shinohara *et al* did not find a difference regarding how much the physicians found the experimental treatment *beneficial* in both 'overstated' and

'not overstated' abstracts.[20] However, that study controlled the amount of information in the abstract, standardising it in the Results and Methods sections, for example, by removing subgroup data from all experimental abstracts. Spin is sometimes used to give extra information inappropriate to emphasise the non-significant results, which is why our abstracts with and without spin differed regarding the amount of information they contained. The abstracts with spin included information on women and compliant patients for the subgroup analyses, which was not mentioned in the abstracts without spin. Fifth, we used Likert scales that present several limitations such as the ceiling effect noted previously, the need to analyse the ordinal data with non-parametric methods,[34] and the fact that Likert scaling is a bipolar scaling method, measuring either a positive or negative response to a statement and may be subject to distortion such as central tendency bias and acquiescence bias.[35] Finally, although online trials may help in the recruitment of participants, it can be difficult to determine the impact of any 'volunteer effect', a selection bias that might arise and not be assessable because of a lack of information on the non-participants.[36]

In looking to future research, this should reproduce our experiment using other samples of abstract and other populations, greater variations in the point estimate for the effect and its 95% CI, and other amounts of information on the reported trial to see if the findings are generalisable. First, we were not able to determine whether the respondents' interpretations of the results in the abstracts were influenced by their perception of the clinical relevance of either the point estimate for the effect (which was the same in both the statistically significant and non-statistically significant abstract) or the ends of the 95% CI, which needed to be different in the two abstracts in order to have the CI cross the line of no difference for the non-statistically significant abstracts. This may be important in the context of the debate about statistical versus scientific importance or clinical significance[1 2] and would be worthy of testing in future studies, which would compare a variety of point estimates and 95% CIs that are very different. Second, to identify the type of spin that is the most influential, further research could also use an approach that would test individual features such as the inclusion of a statistically significant subgroup analysis and avoid this difference in the information presented by assessing spin in research summaries that are longer than the typical abstracts that we tested (and therefore could include information on more analyses). Third, future research should also assess whether readers with access to the full text of articles and, perhaps, the protocols for the clinical trials, would be affected differently by the contents of the abstract given that they would be able, for example, to make a more thorough consideration of subgroup analyses. Lastly, the importance of being able to assess the full-text article before making a decision between the two drugs was raised by some participants. To them, statistical significance was not enough to make a decision between the two hypothetical drugs Naranex and Bulofil, but we did make a neutral option available in the Likert scale. Only 37 participants provided answers in the free text box of the questionnaire; therefore, we do not present a qualitative data analysis. However, some participants wrote of the need for additional information (such as the full text report) for other aspects of their answers, and we recognise that an abstract contains insufficient information to reach fully informed conclusions about treatment superiority. However, many clinicians and health students might limit their reading to abstracts without accessing the full text when making initial or time-constrained judgements,[37 38] making our study setting relevant.

## CONCLUSION

We have shown how statistical significance and spin can influence readers' interpretation of the summary results of a clinical trial and lead them to reach different conclusions about the effects of a treatment. Peer reviewers, editors and readers of journal articles or conference abstracts need to be aware of this. Critical thinking should be an important part of the teaching of health students and reinforced to professionals. It might be helpful to train them to recognise spin and to develop the skills needed to form their own conclusions from the results presented in scientific articles, without undue influence from the way that those results are presented.

**Acknowledgements** We thank all the participants, especially the PhD students at the Centre for Public Health, Queen's University Belfast, for their comments and participation during the pilot study; and Dr Gabriel Baron for his clarification on non-parametric statistics. This study is based on a thesis prepared as part of the MSc in Public Health 'Comparative Effectiveness Research' of Université de Paris, France, and the research was conducted in the Centre for Public Health, Queen's University Belfast, UK.

**Contributors** SJ worked on this study as part of his thesis for the MSc in Public Health 'Comparative Effectiveness Research' of Université de Paris, France. He is accountable for all aspects of the work and for ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. IB, an expert in spin, was responsible for evaluation of the thesis, advising during the conception and design of the work, revising the manuscript critically for intellectual content and interpretation of data, and approved this version for publication. MC was thesis supervisor and daily collaborator on the work from conception and design, drafting and piloting, analysis, revising the manuscript critically for intellectual content interpretation of data, and approved this version for publication. SJ is guarantor for this study.

**Funding** The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

**Competing interests** None declared.

**Patient and public involvement** Patients and/or the public were involved in the design, conduct, reporting or dissemination plans of this research. Refer to the Methods section for further details.

**Patient consent for publication** Not applicable.

**Ethics approval** This study involves human participants and was approved by Queen's University Belfast Research Ethics (ref 15/35 v3). The participants gave informed consent to participate in the study before taking part. Additionally, the protocol for the pilot study was not made available, other than to the ethics committee, and the revisions that we made following the pilot study were not incorporated into an updated protocol but were simply adopted in the design of, for example, the revised questionnaire and the modified abstracts.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** Data are available in a public, open access repository. Extra data can be accessed via the Dryad data repository at http://datadryad.org/ with the doi: 10.5061/dryad.0cfxpnw2z.



**ORCID iDs**
Sofyan Jankowski http://orcid.org/0000-0003-0866-3290
Mike Clarke http://orcid.org/0000-0002-2926-7257

## REFERENCES

1 Pike H. Statistical significance should be abandoned, say scientists. *BMJ* 2019;364:l1374.
2 Wasserstein RL, Schirm AL, Lazar NA. Moving to a World Beyond "p. *Am Stat* 2019;73:1–19.
3 Steegmans PAJ, Di Girolamo N, Meursinge Reynders RA. Spin in the reporting, interpretation, and extrapolation of adverse effects of orthodontic interventions: protocol for a cross-sectional study of systematic reviews. *Res Integr Peer Rev* 2019;4:27.
4 Chiu K, Grundy Q, Bero L. 'Spin' in published biomedical literature: a methodological systematic review. *PLoS Biol* 2017;15:e2002173.
5 Ghannad M, Olsen M, Boutron I, *et al*. A systematic review finds that spin or interpretation bias is abundant in evaluations of ovarian cancer biomarkers. *J Clin Epidemiol* 2019;116:9–17.
6 Boutron I, Dutton S, Ravaud P, *et al*. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA* 2010;303:2058–64.
7 Eysenbach G. Improving the quality of web surveys: the checklist for reporting results of Internet E-Surveys (cherries). *J Med Internet Res* 2004;6:e34.
8 Schulz KF, Altman DG, Moher D, *et al*. Consort 2010 statement: updated guidelines for reporting parallel group randomised trials. *PLoS Med* 2010;7:e1000251.
9 Bero L, Chiu K, Grundy Q. The SSSPIN study-spin in studies of spin: meta-research analysis. *BMJ* 2019;367:l6202.
10 Bagg MK, Hübscher M, Rabey M, *et al*. The resolve trial for people with chronic low back pain: protocol for a randomised clinical trial. *J Physiother* 2017;63:47–8.
11 Katz N, Borenstein DG, Birbara C, *et al*. Efficacy and safety of tanezumab in the treatment of chronic low back pain. *Pain* 2011;152:2248–58.
12 Williams CM, Maher CG, Latimer J, *et al*. Efficacy of paracetamol for acute low-back pain: a double-blind, randomised controlled trial. *Lancet* 2014;384:1586–96.
13 Konno S, Oda N, Ochiai T, *et al*. Randomized, double-blind, placebo-controlled phase III trial of duloxetine monotherapy in Japanese patients with chronic low back pain. *Spine* 2016;41:1709–17.
14 Shirado O, Doi T, Akai M, *et al*. Multicenter randomized controlled trial to evaluate the effect of home-based exercise on patients with chronic low back pain: the Japan low back pain exercise therapy study. *Spine* 2010;35:E811–9.
15 Shanthanna H, Gilron I, Thabane L, *et al*. Gabapentinoids for chronic low back pain: a protocol for systematic review and meta-analysis of randomised controlled trials. *BMJ Open* 2016;6:e013200.
16 Charles P, Giraudeau B, Dechartres A, *et al*. Reporting of sample size calculation in randomised controlled trials: review. *BMJ* 2009;338:b1732.
17 Jankowski S, Boutron I, Clarke M. The influence of statistical significance and spin on readers' perception of clinical trial abstracts: a randomized trial 2021.
18 Marco CA, Larkin GL. Research ethics: ethical issues of data reporting and the quest for authenticity. *Acad Emerg Med* 2000;7:691–4.
19 Boutron I, Altman DG, Hopewell S, *et al*. Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of cancer: the SPIIN randomized controlled trial. *J Clin Oncol* 2014;32:4120–6.
20 Shinohara K, Aoki T, So R, *et al*. Influence of overstated abstract conclusions on clinicians: a web-based randomised controlled trial. *BMJ Open* 2017;7:e018355.
21 Gewandter JS, McKeown A, McDermott MP, *et al*. Data interpretation in analgesic clinical trials with statistically nonsignificant primary analyses: an ACTTION systematic review. *J Pain* 2015;16:3–10.
22 Wood J, Freemantle N, King M, *et al*. Trap of trends to statistical significance: likelihood of near significant P value becoming more significant with extra data. *BMJ* 2014;348:g2215.
23 Fisher A, Anderson GB, Peng R, *et al*. A randomized trial in a massive online open course shows people don't know what a statistically significant relationship looks like, but they can learn. *PeerJ* 2014;2:e589.
24 Ioannidis JPA. The importance of predefined rules and Prespecified statistical analyses: do not abandon significance. *JAMA* 2019;321:2067–8.
25 Adibi A, Sin D, Sadatsafavi M. Lowering the P value threshold. *JAMA* 2019;321:1532–3.
26 Head ML, Holman L, Lanfear R, *et al*. The extent and consequences of p-hacking in science. *PLoS Biol* 2015;13:e1002106.
27 Aguinis H, Vassar M, Wayant C. On reporting and interpreting statistical significance and p values in medical research. *BMJ Evid Based Med* 2021;26:39–42.
28 Ruiz-Ruano García AM, López Puga J. Deciding on null hypotheses using P-values or Bayesian alternatives: a simulation study. *Psicothema* 2018;30:110–5.
29 Adams NG, O'Reilly G. A likelihood-based approach to P-value interpretation provided a novel, plausible, and clinically useful research study metric. *J Clin Epidemiol* 2017;92:111–5.
30 Kraemer HC, Neri E, Spiegel D. Wrangling with p-values versus effect sizes to improve medical decision-making: a tutorial. *Int J Eat Disord* 2020;53:302–8.
31 Consonni D, Bertazzi PA. Health significance and statistical uncertainty. The value of P-value. *Med Lav* 2017;108:327–31.
32 Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature* 2019;567:305–7.
33 Chhapola V, Tiwari S, Brar R, *et al*. Reporting quality of trial abstracts-improved yet suboptimal: a systematic review and meta-analysis. *J Evid Based Med* 2018;11:89–94.
34 Allen IE, Seaman CA. Likert scales and data analyses. *Qual Prog* 2007;40:64–5.
35 Pimentel JL. A note on the usage of Likert scaling for research data analysis. *Usm Rd J* 2010;18:109–12.
36 MSDP N, Narayan KA. Strengths and weaknesses of online surveys. *Technology* 2019;6:7.
37 Barry HC, Ebell MH, Shaughnessy AF, *et al*. Family physicians' use of medical abstracts to guide decision making: style or substance? *J Am Board Fam Pract* 2001;14:437–42.
38 Smith R. What clinical information do doctors need? *BMJ* 1996;313:1062–8.