The Journal of Clinical Endocrinology & Metabolism, 2026, **111**, e70–e82
https://doi.org/10.1210/clinem/dgaf332
Advance access publication 9 June 2025
Clinical Research Article

Fragility of Evidence for the Efficacy of Anti-Fracture Medications

Nick Tran,[1,2] Thach S. Tran,[2,3] and Tuan V. Nguyen[2,4]

[1]School of Psychological Sciences, Faculty of Medicine, Health and Human Sciences, Macquarie University, Sydney, NSW 2109, Australia
[2]School of Biomedical Engineering, University of Technology Sydney, Sydney, NSW 2007, Australia
[3]Skeletal Diseases Program, Garvan Institute of Medical Research, Sydney, NSW 2010, Australia
[4]School of Population Health, UNSW Medicine, UNSW Sydney, Sydney, NSW 2052, Australia

Correspondence: Tuan Van Nguyen, PhD, DSc, School of Biomedical Engineering, University of Technology Sydney, Level 10, Building 11, City Campus, Broadway, NSW 2007, Australia. Email: TuanVan.Nguyen@uts.edu.au.

Abstract

Context: A *P* value and statistical significance, conventionally considered for assessing an intervention's effectiveness, are usually misused and misinterpreted.

Objective: To quantify fragility of randomized controlled trial (RCT) evidence for anti-fracture efficacy.

Methods: This retrospective analysis included 27 phase 3/4 RCTs in high-impact medical journals which assessed anti-fracture efficacy, allocated participants in a 1:1 ratio to pharmacological intervention or control, and reported a statistically significant result. Fragility of the results were assessed using the Fragility Index (FI) and Fragility Quotient (FQ). FI is the minimum number of participants in a positive analysis result for whom reversing the reported status would eliminate statistical significance, while FQ is a function of FI to the sample size.

Results: The median FI was 9 (IQR: 4, 19), indicating that adding 9 fracture patients (~0.51% of the study size) to the intervention group would eliminate the documented evidence of anti-fracture efficacy. Notably, the number of participants lost to follow-up exceeded the corresponding FI in 60% of analyses. The most robust evidence for anti-fracture efficacy was documented for romosozumab (FI: 19.5; IQR: 7.0, 31.5); whereas the least found for denosumab (4; 3, 17) and calcium/vitamin D supplementation (7.0; 2.3, 16.8). Anti-fracture efficacy evidence improved among the results that considered fractures the primary endpoint measure (14; 11, 33) or those with *P* value < .001 (26; 18, 42).

Conclusion: The existing RCT evidence of anti-fracture efficacy is highly fragile. The FI, its comparison with loss to follow-up and FQ should be incorporated into clinical guideline development and doctor-patient risk communication.

Key Words: robustness, anti-fracture efficacy, fragility index, fragility quotient

Abbreviations: FI, Fragility Index; FQ, Fragility Quotient; IQR, interquartile range; RCT, randomized controlled trial.

Randomized controlled trials (RCTs) are widely considered the gold standard for evaluating the efficacy and safety of interventions. These trials involve randomly assigning participants to either receive the intervention being studied or a placebo/standard care. The true effect of the intervention is assessed by comparing outcomes between the groups. The factor that helps distinguish between effect and noise is the *P* value, defined as the probability of observing the data (or more extreme data) if the intervention is, in fact, not effective. Traditionally, a *P* value < .05 is considered "statistically significant," suggesting a real effect of the intervention. This threshold is used to inform clinical practice guidelines, ultimately shaping how healthcare professionals manage patients with specific conditions.

However, the current reliance on *P* values, particularly those obtained in analyses examining binary outcomes (eg, fracture vs no fracture), has several limitations because it oversimplifies the complexity of statistical inference and can lead to misinterpretation of results. These limitations include the use of arbitrary thresholds (eg, $P < .05$), emphasis on significance rather than effect size, and sensitivity to sample size. In current practice, an intervention yielding a *P* value of .05 is often deemed effective, while one with a *P* value of .051 is interpreted as lacking sufficient evidence of efficacy. This sharp cutoff illustrates the arbitrary nature of significance thresholds. More importantly, *P* values can be highly sensitive to minor changes in the data, especially when sample sizes are small and event counts are low (1, 2). In such cases, even slight variations in the number of observed events can shift the result from statistically significant to nonsignificant, or vice versa. This volatility highlights the potential fragility of evidence based solely on *P* values.

To address this volatility, the *Fragility Index* (FI) was proposed as an ancillary metric used to quantify the robustness of RCT evidence with statistically significant results (3). The FI is the minimum number of "non-events" in the experimental group that need to be changed to "events" to make the reported statistically significant results statistically nonsignificant (3). Its extension, the *Fragility Quotient* (FQ), as a relative measure of the FI to the sample size, provides an additional means to

Received: 22 February 2025. **Editorial Decision:** 4 June 2025. **Corrected and Typeset:** 19 June 2025

measure the robustness of RCT evidence relative to the sample (4). Smaller FI and FQ values indicate more fragile RCT evidence. Given its intuition and ease in calculation, interpretation and communication, the FI has quickly become popular in multiple specialties in medical research (5), including surgical orthopedics and trauma research (6) and osteoporosis research (7). A recent systematic review including 42 RCTs referenced in the guidelines for the treatment of osteoporosis revealed the RCT evidence for osteoporosis treatment was statistically fragile (7), although evidence for the anti-fracture efficacy of specific pharmacological interventions has never been examined.

We sought to quantify the robustness of evidence for the anti-fracture efficacy of common pharmacological interventions using the FI and FQ. The findings are expected to help clinicians identify robust evidence for the management of osteoporosis.

Methods

Inclusion and Exclusion Criteria

We searched PubMed using the keywords "*Bone Density Conservation Agents/therapeutic use*," "*Calcium supplementation*," "*Vitamin D*," "*Fracture*," and "*Randomized controlled trial*" to identify RCTs that had reported a statistically significant result for at least one dichotomous (ie, fracture vs no fracture) or time-to-fracture outcome published in 11 high-impact general medicine and bone-specific journals between July 31, 1992, and May 15, 2024. The high-impact general medicine journals included the *New England Journal of Medicine (N Engl J Med), Lancet, British Journal of Medicine (BMJ), the Journal of American Medical Association (JAMA), JAMA Internal Medicine (JAMA Intern Med), Annals of Internal Medicine (Ann Intern Med),* and *PLos Medicine (PLos Med)*; whereas the bone-specific journals included the *Journal of Bone and Mineral Research (J Bone Miner Res), Journal of Clinical Endocrinology & Metabolism (J Clin Endocrinol Metab), Bone (Bone)* and *Osteoporosis International (Osteoporos Int)*. These journals were selected under the assumption that the trials published in these top-tier journals are of high quality and provide sufficient information for the calculation of FI/FQ and other measures, including the number of participants lost to follow-up.

The inclusion criteria were: (i) phase 3 or 4 RCTs that assessed the effects of a pharmacological intervention on anti-fracture efficacy; (ii) parallel arm RCTs that allocated participants in a 1:1 ratio to pharmacological intervention and control; and (iii) RCTs that reported a statistically significant result (P value $< .05$) for at least one dichotomous or time-to-event outcome under a null hypothesis that the pharmacological intervention was not effective. The significant results from the extended phases of a RCT were also considered. One RCT might have more than 1 statistically significant result. We excluded studies with other designs, such as noninferiority, crossover, economic, mechanistic, or review designs, non-pharmacological interventions, or those beyond the osteoporosis medications. The secondary analyses for descriptive or predictive analyses were also excluded.

Data Extraction

Two investigators (N.T., T.S.T.) independently reviewed all identified abstracts, and discrepancies were adjudicated by a third investigator (T.V.N.). We used a prespecified data collection form to extract details of the statistically significant results, including fracture site, outcome type (binary or time-to-event), the observed number of participants sustaining a fracture, the number of participants analyzed in each group, and the reported number of participants lost to follow-up. For a time-to-event analysis, the number of fracture patients in each group over the follow-up time was included, whereas the number of participants analyzed was calculated as a sum of the number of fracture patients and the number of participants at risk at the time of fracture assessment.

Calculation of Fragility Index and Fragility Quotient

The FI and FQ were calculated for each of the included analysis results using the *fragility* package (8). Specifically, the FI was calculated from a 2×2 contingency table by iteratively adding a fracture event to the intervention group while concomitantly subtracting a non-fracture event from the same group to keep the total number of participants constant until the first time the calculated two-sided P value on Fisher's exact test $> .05$. The FI for each of the included results was the number of additional fracture events required to obtain a P value $> .05$ (3) (Fig. 1). The FQ was then calculated as the FI divided by the total number of participants analyzed.

Statistical Analysis

We first conducted Fisher's exact test for each of the reported statistically significant results and excluded any results with the re-calculated two-sided P value $> .05$.

The primary analysis included statistically significant results from the original RCTs and their extended phases. We presented the overall FIs/FQs and those by trial characteristics as median (interquartile range [IQR]). The trial characteristics included pharmacological interventions, fracture sites, nature of the control group, participant's sex, and journals of publication. The overall FIs/FQs were calculated from all analysis results included in the current project, whereas the pharmacological interventions-specific FIs/FQs were computed for each pharmacological intervention, regardless of fracture sites, nature of the control group, participant's sex, and publication journals. Among the analysis results that reported the loss to follow-up, we compared the calculated FI with the number of participants lost to follow-up to further examine the fragility of evidence for anti-fracture efficacy. We prespecified 2 subgroup analyses and 1 sensitivity analysis. The first subgroup analysis examined trials with fractures being prespecified as a primary endpoint, and the second focused on the highly significant results with a P value $< .001$. The predefined sensitivity analysis excluded the extended phases of the included trials.

As suggested by reviewers, we conducted 4 post hoc exploratory analyses. The first and second exploratory analyses calculated fracture site-specific fragility measures for each specific pharmacological intervention and examined the relationship between FI and years of publication, respectively. The third analysis excluded all results from trials with calcium and/or vitamin D supplementation; whereas the fourth analysis included all relevant RCTs regardless of their publication journals.

Results

Selection of Trials

Among 386 initially identified abstracts, we excluded 285 irrelevant abstracts, leaving 101 abstracts in which full texts

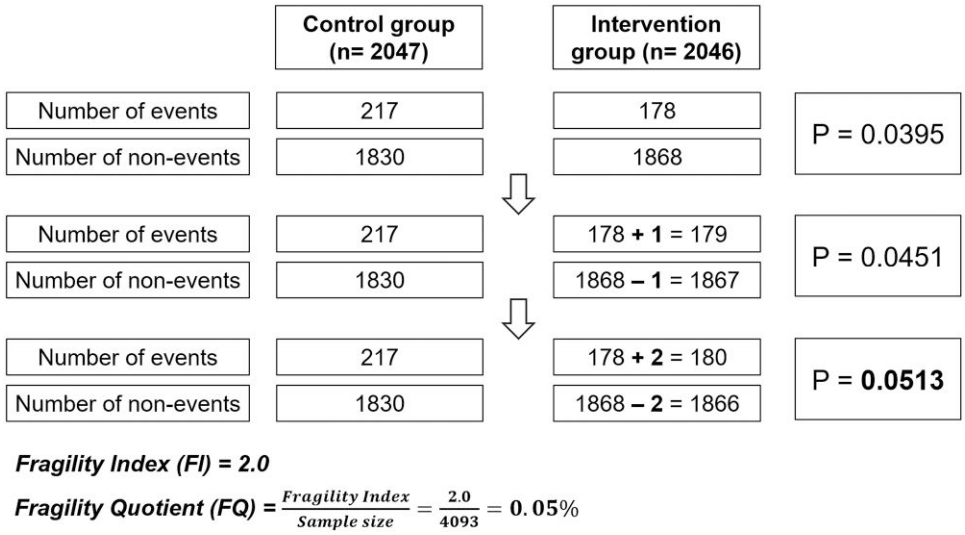

Figure 1. Illustration of calculation of Fragility Index and Fragility Quotient.

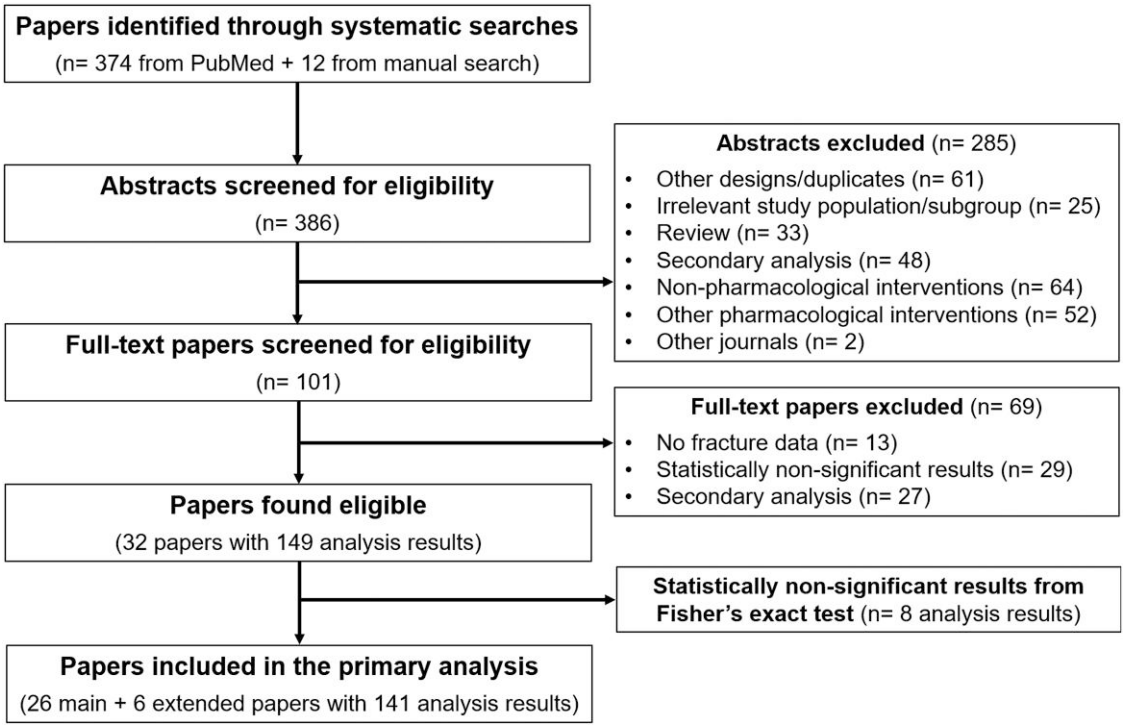

Figure 2. Flowchart of literature search.

were screened for eligibility (Fig. 2). We then excluded 69 papers with no fracture data, statistically nonsignificant results on fracture outcome or secondary analysis of RCT data for mechanistic or prediction research. An additional 8 analysis results that became statistically nonsignificant with a Fisher's exact test used to calculate FI were excluded, leaving 141 analysis results from 32 papers of 26 original RCTs (9-34) and 6 extended phases (35-40) (Supplementary Table S1) (41). The statistically significant results were found over the extended 10-year period (37) of the original 5-year Auckland calcium RCT, which had found no effect of calcium supplements on the 5-year fracture incidence (42). The predefined subgroup analyses included 33 results with fragility fractures being predefined as a primary endpoint and 45 highly statistically significant results.

Trial Characteristics

Approximately three-fourths of the included trials, or 82.3% of the included analysis results, were conducted in postmenopausal women. Bisphosphonates were the most studied agent over the last 30 years (44% of the included RCTs and 34% of the analysis results), followed by the parathyroid hormone (PTH) analogue (18.5% and 21%) (Table 1). Pharmacological

Table 1. Characteristics of the trials included in the primary analysis

Characteristics	Studies (n = 27)	Analysis results (n = 141)
Medication		
Bisphosphonates	12 (44.4%)	48 (34.0%)
Parathyroid hormone analog	5 (18.5%)	30 (21.3%)
Romosozumab	2 (7.4%)	28 (19.9%)
Denosumab	2 (7.4%)	9 (6.4%)
Strontium ranelate	2 (7.4%)	12 (8.5%)
Calcium and/or vitamin D	4 (14.8%)	14 (9.9%)
Journal		
N Engl J Med	12 (44.4%)	79 (56.0%)
Lancet	2 (7.4%)	9 (6.4%)
BMJ	1 (3.7%)	5 (3.5%)
JAMA	3 (11.1%)	10 (7.1%)
Ann Intern Med	1 (3.7%)	1 (0.7%)
J Bone Miner Res	1 (3.7%)	6 (4.3%)
J Clin Endocrinol Metab	2 (7.4%)	12 (8.5%)
Osteoporos Int	5 (18.5%)	19 (13.5%)
Placebo		
Active	3 (11.1%)	29 (20.6%)
Placebo	24 (88.9%)	112 (79.4%)
Sex		
Both	5 (18.5%)	15 (10.6%)
Men	2 (7.4%)	10 (7.1%)
Women	20 (74.1%)	116 (82.3%)
Number of participants at randomization, median (IQR)	1910 (1280, 2980)	1960 (1220, 4090)

Abbreviations: Ann Intern Med, Annals of Internal Medicine; BMJ, British Journal of Medicine; IQR, interquartile range; JAMA, Journal of American Medical Association; J Bone Miner Res, Journal of Bone and Mineral Research; J Clin Endocrinol Metab, Journal of Clinical Endocrinology & Metabolism; N Engl J Med, New England Journal of Medicine; Osteoporos Int, Osteoporosis International.

interventions were compared with placebo in 89% of the included trials and nearly 80% of the results. We found 90% of the analysis results included had reported effect sizes of 0.726 or lower with a median of 0.51 (IQR: 0.35, 0.64) (Supplementary Fig. S1) (41).

Overall Robustness of Evidence for Anti-Fracture Efficacy

The overall robustness of evidence for anti-fracture efficacy was first assessed for all fractures reported in the analyses included in this project. The median FI and FQ were 9 (IQR: 4, 19) and 0.51% (0.2%, 1.0%), respectively, indicating that the reported statistically significant efficacy of anti-fracture medications would be lost if only 9 individuals without fracture in the intervention group (or 0.51% of the study population) were ultimately found to be fractured (Table 2). Moreover, one-third of these results had FI of 4 or lower and the estimated FI exceeded the reported number of participants lost to follow-up in approximately 60% of the results with the number of loss-to-follow-up reported. The median number of participants recruited in the trials was 1910 (IQR: 1280, 2908), and of fracture events was 86 (52, 210). As expected, the estimated FI was positively associated with the trial size (Fig. 3A) and number of fracture events (Fig. 3B).

Evidence for anti-fracture efficacy became more robust among the analysis results where fractures were predefined as the primary endpoint with a median FI of 14 (IQR: 11, 33) and FQ of 1.07% (0.53%, 2.06%) and those with highly significant results (26 [18, 42] and 1.31% [0.67%, 1.93%]) (Table 3). Moreover, the estimated FI exceeded the reported number of loss-to-follow-ups in half of these results (60.6% and 53.3%, respectively), suggesting evidence for anti-fracture efficacy remains fragile even among the strongest results.

Analysis by Pharmacological Interventions

For bisphosphonates, changing the fracture status of, on average, 8.5 fracture-free participants given bisphosphonates would make the reported anti-fracture efficacy statistically insignificant (Table 2; Fig. 4). Among all fracture types, the most robust evidence for the efficacy of bisphosphonates was observed in the prevention of osteoporotic fractures (Supplementary Table S2) (41). Similar patterns, albeit far more robust, were found in the analysis results with fractures predefined as a primary endpoint or those with P value < .001 (Table 3).

Romosozumab demonstrated the strongest evidence of anti-fracture efficacy, with a median FI of 19.5 (IQR: 7, 31.5) and a corresponding FQ of 0.35% (0.15%, 0.63%). Among studies where fragility fractures were defined as a primary endpoint, the FI and FQ increased to 40 (IQR: 26, 42) and 0.71% (0.60%, 1.15%), respectively. Similarly, in analyses with P values less than .001, the FI and FQ were 36 (IQR: 26, 42) and 0.60% (0.39%, 1.15%). Site-specific analyses showed that romosozumab provided strong evidence for preventing both osteoporotic fractures (FI: 26.0) and vertebral fractures (FI: 25.5) (Supplementary Table S2) (41).

We also found evidence for anti-fracture efficacy of strontium ranelate was the second most robust with the FIs being 18.5 (IQR: 6.3, 45.0), 48.5 (46.8, 50.3) and 45 (21, 48.5) for all results, results with fractures being predefined as the primary endpoints and results with P value < .001, respectively.

In contrast, denosumab and calcium and/or vitamin D supplementation showed the least robust evidence. Changing the fracture status of just 4 participants receiving denosumab or 7 receiving calcium and/or vitamin D would nullify the statistical significance of the results. Of the 9 analyses involving denosumab, 5 focused on vertebral fractures and had a median FI of 4 (IQR: 4, 7). For calcium and/or vitamin D supplementation, 5 analyses targeted hip fractures and yielded a median FI of 8 (IQR: 3, 11) (Supplementary Table S2) (41).

Analysis by Fracture Sites

Vertebral fracture was the most common outcome in RCTs. The median FI and FQ for morphological vertebral fractures were 11 (IQR: 4, 23) and 0.67% (0.31%, 1.68%), respectively, slightly higher than those for any fractures (Table 2; Fig. 5). Importantly, if hip fractures had been additionally identified in only 5 previously fracture-free participants (~0.15% of the sample size), the documented evidence for the benefits of osteoporosis treatment on hip fractures was lost.

Three-fourths of the included results that had predefined vertebral fractures as the primary endpoint had a median FI of 15 (IQR: 11.5, 40.0) or 1.29% (0.69%, 2.58%) of the

Table 2. Fragility of evidence for anti-fracture efficacy

	Number of analysis results	Number of fracture patients, median (IQR)	Fragility Index, median (IQR)	Fragility Quotient, median (IQR)
Overall	141	86 (52, 210)	9.0 (4.0, 19.0)	0.51% (0.24%, 1.04%)
Pharmacological intervention:				
Bisphosphonates	48	72.5 (58.8, 141)	8.5 (4.0, 14.0)	0.54% (0.27%, 0.86%)
Parathyroid hormone analogue	30	44 (34, 59)	6.5 (3.0, 12.5)	0.74% (0.33%, 1.37%)
Romosozumab	28	173 (106, 272)	19.5 (7.0, 31.5)	0.35% (0.15%, 0.63%)
Denosumab	9	69 (29, 121)	4.0 (3.0, 17.0)	0.26% (0.22%, 0.30%)
Strontium ranelate	12	234 (125, 346)	18.5 (6.3, 45.0)	1.31% (0.45%, 2.19%)
Calcium and/or vitamin D	14	183 (128, 345)	7.0 (2.3, 16.8)	0.43% (0.34%, 0.68%)
Fracture site:				
Any fractures	18	210 (76, 401)	9.0 (5.0, 20.3)	0.73% (0.24%, 1.15%)
Osteoporotic fractures	10	163 (67.3, 320)	12.0 (6.0, 29.0)	0.69% (0.23%, 0.75%)
Non-vertebral fractures	19	168 (71, 365)	5.0 (2.0, 17.5)	0.32% (0.15%, 0.48%)
Vertebral fractures	68	72 (38.5, 123)	11.0 (4.0, 23.0)	0.67% (0.31%, 1.68%)
Clinical vertebral fractures	14	63 (51, 95.5)	6.5 (3.25, 11.8)	0.46% (0.31%, 0.53%)
Hip fractures	9	140 (106, 190)	5.0 (2.0, 11.0)	0.15% (0.10%, 0.34%)
Forearm fractures	3	99 (81, 111)	4.0 (3.0, 6.5)	0.34% (0.27%, 0.43%)
Timing of fracture assessment:				
> 6- 12 months	26	65.5 (52.3, 137)	5.0 (3.0, 9.8)	0.24% (0.15%, 0.49%)
> 12- 18 months	18	53 (34.5, 142)	7.5 (4.0, 11.0)	0.66% (0.36%, 1.02%)
> 18- 24 months	48	79.5 (42.8, 189)	10.5 (4.0, 23.0)	0.54% (0.30%, 1.19%)
> 24- 36 months	35	142 (83.5, 341)	15.0 (4.0, 38.5)	0.66% (0.24%, 1.30%)
> 36- 48 months	5	204 (121, 409)	12.0 (7.00, 13.0)	0.53% (0.31%, 1.04-%)
> 48- 60 months	2	96 (83, 109)	2.0 (2.0, 2.0)	0.35% (0.34%, 0.35%)
> 60- 72 months	7	147 (85.5, 281)	10.0 (7.5, 16.5)	0.54% (0.42%, 0.85%)
Journal:				
N Engl J Med	79	117 (54.5, 248)	12.0 (4.5, 26.0)	0.39% (0.23%, 0.83%)
Lancet	9	91 (63, 100)	11.0 (4.0, 15.0)	0.69% (0.26%, 1.43%)
BMJ	5	315 (262, 475)	16.0 (11.0, 22.0)	0.70% (0.62%, 1.09%)
JAMA	10	70.5 (47.5, 112)	10.0 (4.8, 11.0)	0.51% (0.20%, 0.79%)
Ann Intern Med	1	60	7	0.41%
J Bone Miner Res	6	70.5 (64.5, 101)	7.5 (4.5, 12.0)	0.38% (0.23%, 0.62%)
J Clin Endocrinol Metab	12	38 (28.5, 145)	4.0 (2.0, 11.5)	0.86% (0.42%, 1.87%)
Osteoporos Int	19	68 (56, 89.5)	5.0 (3.0, 10.5)	0.66% (0.35%, 1.44%)
Sex:				
Both	15	39 (32.5, 58.5)	4.0 (2.5, 7.0)	0.84% (0.45%, 1.32%)
Men	10	26 (21.5, 34.3)	3.0 (1.3, 4.0)	0.22% (0.12%, 0.30%)
Women	116	107 (65.8, 251)	11.0 (5.0, 23.0)	0.54% (0.24%, 1.10%)
Placebo:				
Active	29	107 (53, 255)	10.0 (5.0, 18.0)	0.44% (0.22%, 0.91%)
Placebo	112	83 (50.3, 194)	8.50 (4.0, 19.3)	0.52% (0.24%, 1.06%)

Abbreviations: Ann Intern Med, Annals of Internal Medicine; BMJ, British Journal of Medicine; IQR, interquartile range; JAMA, Journal of American Medical Association; J Bone Miner Res, Journal of Bone and Mineral Research; J Clin Endocrinol Metab, Journal of Clinical Endocrinology & Metabolism; N Engl J Med, New England Journal of Medicine; Osteoporos Int, Osteoporosis International.

study sample. By contrast, hip fractures were primarily examined in 4 analyses with a median FI of 5.5 and FQ of 0.15% (Table 3). Evidence for anti-fracture efficacy at vertebral fractures (~ 69% of all results) became far more robust among highly statistically significant results with a median FI of 24 (IQR: 17, 43.5), equivalent to 1.85% (0.80%, 2.58%) of the sample size.

Other Trial Characteristics

Approximately one-third of the analyses evaluated anti-fracture efficacy over a follow-up period of 18 to 24 months, yielding a median FI of 10.5 (IQR: 4, 23). However, the most robust evidence was observed in studies assessing fracture outcomes between 24 and 36 months, with a higher median FI of 15 (IQR: 4, 38.5) (Table 2). Moreover, 55% of the results

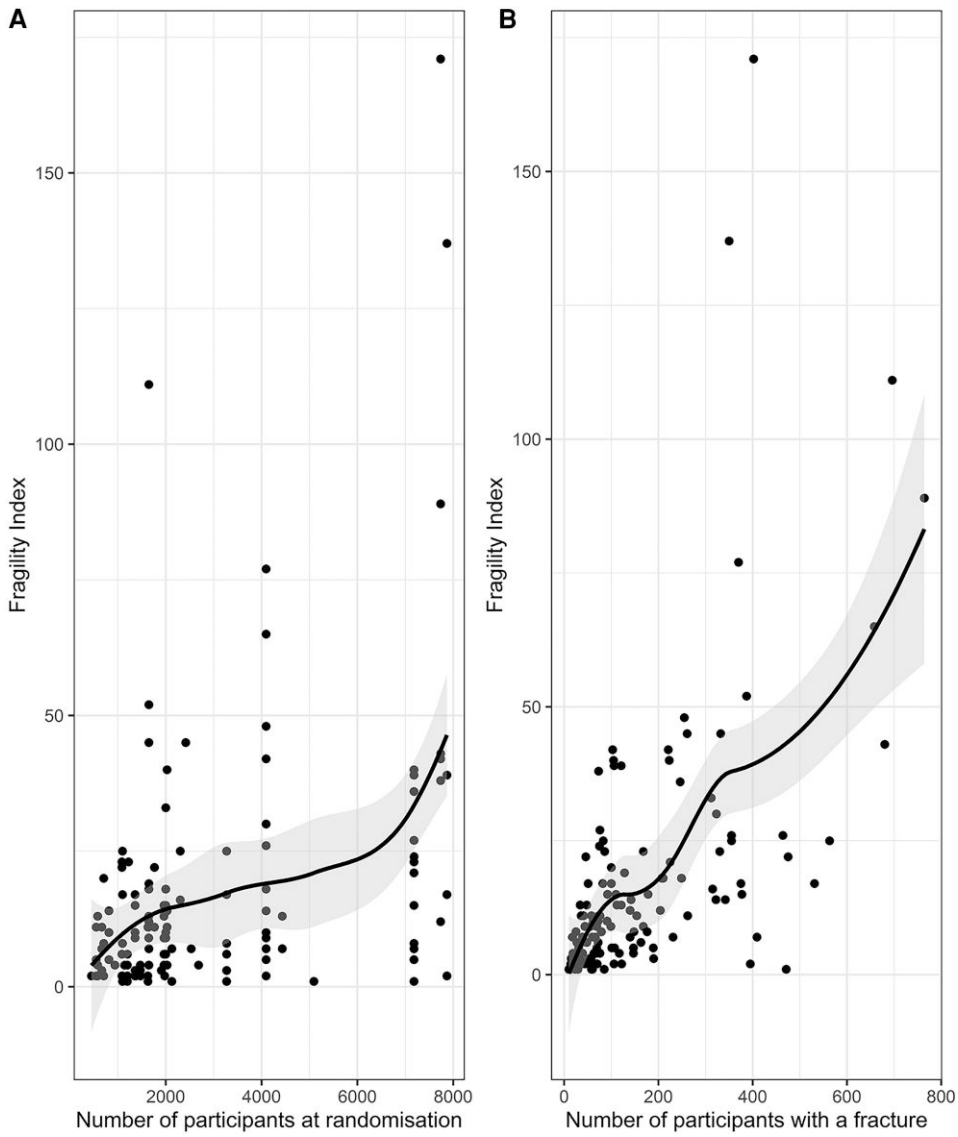

Figure 3. Correlation between Fragility Index and sample size. A, Correlation between Fragility Index and number of participants at randomization. B, Correlation between Fragility Index and number of participants with a fracture.

were published in the *New England Journal of Medicine*, with a median FI of 12 (IQR: 4.5, 26) (Table 2). Notably, no association was found between FI and years of publication (Supplementary Fig. S2) (41).

Our analysis also showed that 82% of the results were derived from studies involving postmenopausal women, with a median FI of 11 (IQR: 5, 23)—higher than that observed in studies involving only men (FI: 3) or mixed-sex populations (FI: 4). In contrast, there was no statistically significant difference in the robustness of evidence between trials using active placebos and those that did not (FI: 10.0 vs 8.5, respectively). As expected, analyses in which fractures were predefined as the primary endpoint and those reporting P values $< .001$ demonstrated stronger evidence for anti-fracture efficacy across all subgroups—regardless of follow-up duration, publication journal, study population, or control group type.

Prespecified Sensitivity Analysis

There were 124 analysis results in the predefined sensitivity analysis, after excluding 17 results from the extended phases of the original trials. The sensitivity analysis reveals similar, though numerically different, results as the primary analysis, confirming the robustness of the primary findings (Supplementary Tables S3 and S4) (41).

Post Hoc Exploratory Analysis Excluding Trials of Calcium and/or Vitamin D Supplementation

We excluded 14 results from 5 trials involving calcium supplementation (37), combined calcium and vitamin D (32, 34, 40), and vitamin D alone (26), resulting in 127 analysis results included in the post hoc exploratory analysis. The overall median FI and FQ were 9 (IQR: 4.0, 19.5) and 0.52% (0.24%, 1.05%), respectively—comparable to the findings of the primary analysis (Supplementary Table S4) (41). Furthermore, this exploratory analysis produced patterns similar to the primary analysis, albeit with some numerical differences, across all trial characteristics (Supplementary Table S5) (41) and subgroup analyses (Supplementary Table S6) (41).

Table 3. Fragility of evidence for anti-fracture efficacy in the subgroup analyses

	Analyses with fragility fractures predefined as the primary endpoint				Analyses with highly significant results ($P < .001$)			
	Number of analysis results	Number of fracture patients, median (IQR)	Fragility Index, median (IQR)	Fragility Quotient, median (IQR)	Number of analysis results	Number of fracture patients, median (IQR)	Fragility Index, median (IQR)	Fragility Quotient, median (IQR)
Overall	33	112 (60, 223)	14 (11, 33)	1.07% (0.53%, 2.06%)	45	128 (82, 330)	26.0 (18.0, 42.0)	1.31% (0.67%, 1.93%)
Pharmacological intervention:								
Bisphosphonates	14	131 (103, 206)	13.0 (8.0, 18.8)	0.93% (0.57%, 1.95%)	10	167 (101, 380)	39.0 (23.3, 42.8)	1.09% (0.69%, 1.97%)
Parathyroid hormone analogue	8	54 (39.8, 83.8)	13.0 (11.0, 17.0)	2.00% (0.89%, 2.65%)	11	48 (42.5, 84.5)	15.0 (12.0, 19.5)	1.93% (1.18%, 2.58%)
Romosozumab	5	221 (105, 370)	40.0 (26.0, 42.0)	0.71% (0.60%, 1.15%)	13	225 (106, 323)	36.0 (26.0, 42.0)	0.60% (0.39%, 1.15%)
Denosumab	1	350	137.0	1.85%	3	121 (102, 236)	39.0 (28.0, 88.0)	0.50% (0.37%, 1.18%)
Strontium ranelate	2	360 (346, 373)	48.5 (46.8, 50.3)	3.36% (3.24%, 3.48%)	7	330 (234, 360)	45.0 (21.0, 48.5)	1.88% (1.62%, 3.36%)
Calcium and/or vitamin D	3	176 (117, 183)	3 (2, 5.5)	0.11% (0.08%, 0.22%)	1	355	25.0	1.05%
Fracture site:								
Any fractures	2	348 (289, 406)	16.5 (11.8, 21.3)	1.09% (0.90%, 1.27%)	2	505 (376, 635)	62.5 (49.3, 75.8)	0.95% (0.74%, 1.15%)
Osteoporotic fractures	1	312	33.0	1.69%	5	323 (312, 355)	30.0 (26.0, 33.0)	0.75% (0.74%, 1.69%)
Non-vertebral fractures	0				4	290 (211, 436)	24.0 (22.5, 29.5)	0.50% (0.34%, 0.76%)
Vertebral fractures	26	103 (51, 204)	15.0 (11.5, 40.0)	1.29% (0.69%, 2.58%)	31	100 (74, 258)	24.0 (17.0, 43.5)	1.85% (0.80%, 2.58%)
Clinical vertebral fractures	0				3	121 (112, 164)	39.0 (28.5, 40.5)	0.65% (0.58%, 0.95%)
Hip fractures	4	158 (120, 180)	5.5 (2.5, 9.0)	0.15% (0.09%, 0.22%)	0			
Timing of fracture assessment:								
> 6- 12 months	1	75	24.0	0.36%	5	76 (75, 128)	24.0 (19.0, 27.0)	0.75% (0.39%, 0.83%)
> 12- 18 months	8	53 (39.8, 89)	9.5 (6.0, 11.5)	0.59% (0.28%, 1.34%)	6	42.5 (37.3, 47)	12.0 (11.0, 13.0)	0.99% (0.81%, 2.20%)
> 18- 24 months	12	96 (72, 222)	24.0 (13.0, 40.0)	1.29% (0.69%, 2.19%)	18	137 (94, 241)	25.5 (21.3, 40.0)	1.53% (0.64%, 2.02%)
> 24- 36 months	10	193 (141, 346)	14.5 (12.3, 50.3)	1.66% (0.78%, 2.77%)	14	296 (113, 378)	42.5 (26.8, 50.3)	1.30% (0.66%, 1.87%)
> 36- 48 months	1	121	13.0	0.31%	1	696	111.0	9.66%
> 48- 60 months	0				0			
> 60- 72 months	1	312	33.0	1.69%	1	312	33	1.69%
Journal:								
N Engl J Med	19	190 (84.5, 341)	25.0 (7.5, 43.5)	1.15% (0.35%, 2.23%)	31	221 (94.5, 353)	36.0 (23.5, 44.0)	1.15% (0.59%, 1.87%)
Lancet	2	158 (125, 190)	27.5 (21.3, 33.8)	1.74% (1.59%, 1.90%)	3	100 (96, 162)	17.0 (16.0, 28.5)	1.62% (1.53%, 1.84%)
JAMA	4	78.5 (35.5, 129)	12.0 (11.0, 13.0)	0.79% (0.66%, 0.83%)	4	40 (35.5, 47.5)	11.0 (10.5, 11.5)	0.80% (0.76%, 0.85%)
Ann Intern Med	1	60	7.0	0.41%	0			
J Bone Miner Res	2	111 (111, 112)	14.0 (13.5, 14.5)	0.72% (0.69%, 0.74%)	1	110	15.0	0.77%
J Clin Endocrinol Metab	2	44.5 (42.8, 46.3)	12.0 (11.5, 12.5)	2.73% (2.66%, 2.81%)	4	155 (46.3, 278)	18.0 (12.5, 28.5)	2.23% (1.87%, 2.66%)
Osteoporos Int	3	100 (69.5, 121)	14.0 (10.5, 17.0)	2.03% (1.55%, 2.85%)	2	398 (249, 547)	65.5 (42.8, 88.3)	6.67% (5.17%, 8.16%)

(*continued*)

Table 3. Continued

	Analyses with fragility fractures predefined as the primary endpoint				Analyses with highly significant results ($P < .001$)			
	Number of analysis results	Number of fracture patients, median (IQR)	Fragility Index, median (IQR)	Fragility Quotient, median (IQR)	Number of analysis results	Number of fracture patients, median (IQR)	Fragility Index, median (IQR)	Fragility Quotient, median (IQR)
Sex:								
Both	4	44.5 (40.5, 93.8)	9.0 (7.0, 11.5)	2.02% (1.36%, 2.66%)	2	44.5 (42.8, 46.3)	12.0 (11.5, 12.5)	2.73% (2.66%, 2.81%)
Men	2	33 (31, 35)	3.5 (2.3, 4.8)	0.31% (0.20%, 0.42%)	0			
Women	27	140 (89, 268)	20.0 (12.5, 40.0)	0.93% (0.51%, 2.04%)	43	168 (84.5, 331)	27.0 (19.5, 42.5)	1.25% (0.66%, 1.88%)
Placebo:								
Active	6	157 (50, 333)	20.5 (13.5, 38.0)	1.04% (0.81%, 1.36%)	11	221 (68, 339)	26.0 (14.0, 45.0)	1.15% (0.76%, 1.53%)
Placebo	27	112 (67.5, 207)	13.0 (7.5, 29.0)	1.07% (0.39%, 2.58%)	34	125 (82.3, 326)	26.0 (20.3, 41.5)	1.37% (0.61%, 2.55%)

Abbreviations: Ann Intern Med, Annals of Internal Medicine; BMJ, British Journal of Medicine; IQR, interquartile range; JAMA, Journal of American Medical Association; J Bone Miner Res, Journal of Bone and Mineral Research; J Clin Endocrinol Metab, Journal of Clinical Endocrinology & Metabolism; N Engl J Med, New England Journal of Medicine; Osteoporos Int, Osteoporosis International.

Post Hoc Exploratory Analysis Including Relevant Trials Regardless of Their Publication Journals

We identified 2 main trials (43, 44) and 4 follow-up or extended studies (45-48) with 11 statistically significant analysis results published beyond the predefined high-impact general and bone-related journals (Supplementary Table S7) (41). A single-center, open-label, randomized 1-year trial was conducted to examine the efficacy of risedronate primarily on BMD changes among 316 osteoporotic Japanese (43), and another double-blind, randomized 2-year trial examined the efficacy of alendronate on the incidence of vertebral fracture observed more than 6 months after randomization among 365 Japanese elderly people (44). The original results of these 3 follow-up studies (45-47) were already included in the current analysis; whereas the remaining study (48) is the 3-year follow-up study of Kushida et al (44).

The exploratory analysis, including 152 analysis results from 28 trials regardless of their publication journals, yielded the overall FI of 9 (IQR: 3, 18) and FQ of 0.50% (0.2, 1.0%), almost identical to the primary analysis.

Discussion

The Fragility Index and Fragility Quotient have emerged as innovative measures to quantify the robustness of evidence for an intervention. Our analysis indicated that the evidence for anti-fracture efficacy was highly fragile. Specifically, reversing the fracture status of only 9 participants, or 0.51% of the total sample size, would change the reported results from statistically significant to nonsignificant, making the reported evidence for the anti-fracture efficacy of the pharmacological interventions no longer valid.

Our estimated FI was comparable with the previous studies that examined the fragility of statistically significant findings from 399 RCTs (3), 643 RCTs (6) published in high-impact journals, or 42 RCTs referenced in the guidelines for the treatment of osteoporosis (7). In a retrospective analysis that included both statistically significant and nonsignificant results from RCTs where fracture was the primary endpoint, Huang and colleagues (7) found that changing 10 fracture-free cases to fractures in the intervention group would reverse the documented statistical significance of osteoporosis efficacy. By contrast, our estimated FI was greater than the FI pooled from an umbrella review of 21 reviews that included RCTs that had examined the surgical orthopedic interventions, including hip/knee arthroplasty, spine surgery, or surgical construction (6). Our estimated FIs of anti-fracture efficacy among results with fractures predefined as a primary endpoint or those with P value < .001 (13.5 and 26, respectively) were greater than the median FI from RCTs referenced in the osteoporosis treatment guidelines (7) or those in high-impact journals (3, 6).

The most robust evidence for anti-fracture efficacy was found for romosozumab and strontium ranelate with the estimated median FIs of 19.5 and 18, respectively. The trials examining these pharmacological interventions had relatively large sample sizes (the median sample size was 4090 for romosozumab trials and 1440 for strontium ranelate trials) and number of participants sustaining at least one fracture during the study period (~ 173 and 330, respectively). These relatively high FIs, though potentially indicating the robust findings (49), still reflect the questionable robustness of evidence for anti-fracture efficacy, as the documented positive findings

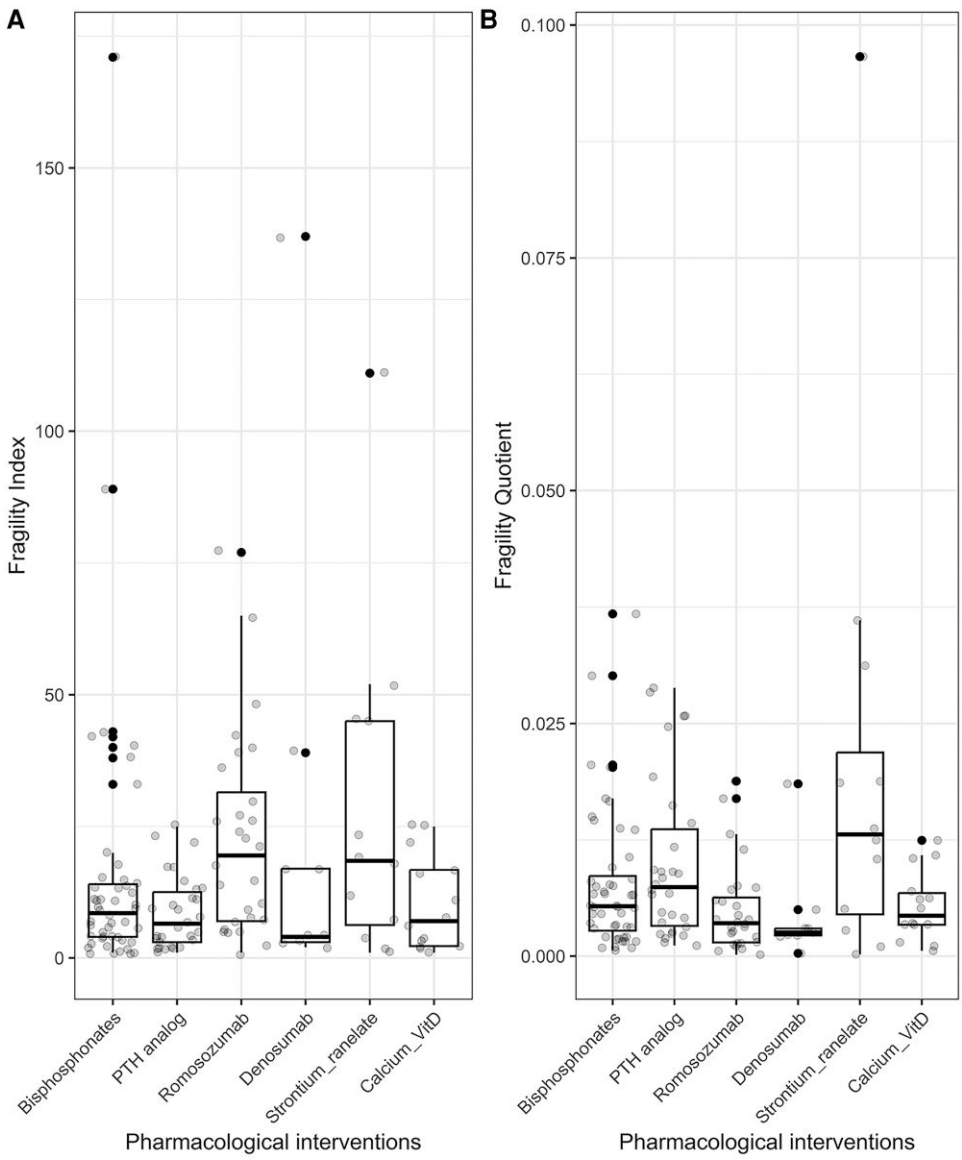

Figure 4. Fragility of evidence for anti-fracture efficacy by pharmacological interventions. A, Fragility Index. B, Fragility Quotient.

were no longer significant if fewer than 20 fracture-free participants treated with these medications or 1% of the sample size, were ultimately found to be fractured.

While a *P* value and statistical significance are usually misused and misinterpreted (50), the interpretation of effect size using conventional metrics such as relative or absolute risks is challenging. A relative risk has been shown to often give an exaggerated impression (51) and an absolute risk, as a probability-based metric, is much harder to understand (52). The FI and FQ have been proposed as a natural frequency measure, making the interpretation of the robustness of evidence more intuitive (3) and potentially more effective than the conventional metrics, as a natural frequency metric is known to be more friendly compared to probability-based metrics (53-55). The FI, not intentionally designed to replace a *P* value, serves as a supplementary metric for assessing the robustness of statistical significance, making a *P* value and the concept of statistical significance more clinically meaningful. However, when using the FI to assess the robustness of evidence, it is important to consider factors that are inherently associated with a low FI. In particular, small RCTs or studies with few outcome events tend to produce lower FIs. To account for the influence of sample size, the FQ, calculated by dividing the FI by the total sample size, has been proposed as a complementary measure (4). Additionally, incorporating the outcome event rate into the interpretation of FI can enhance its clinical relevance in real-world settings. Secondly, sample sizes in RCTs are typically determined using power calculations aimed at detecting a minimally significant difference, which can result in relatively low FIs. To enhance the robustness of trial findings and reduce reliance on a small number of outcome events, it has been proposed that the FI be integrated into sample size calculations during the study design phase (56). This FI-based approach incorporates a target FI threshold alongside traditional power calculations, ensuring that trials are both statistically powered and more resilient to minor changes in the outcome events. Thirdly, FI was not designed to assess a time-to-event or a continuous outcome, which are also common in RCTs examining intervention efficacy. Recently, the survival-inferred FI, operationally defined

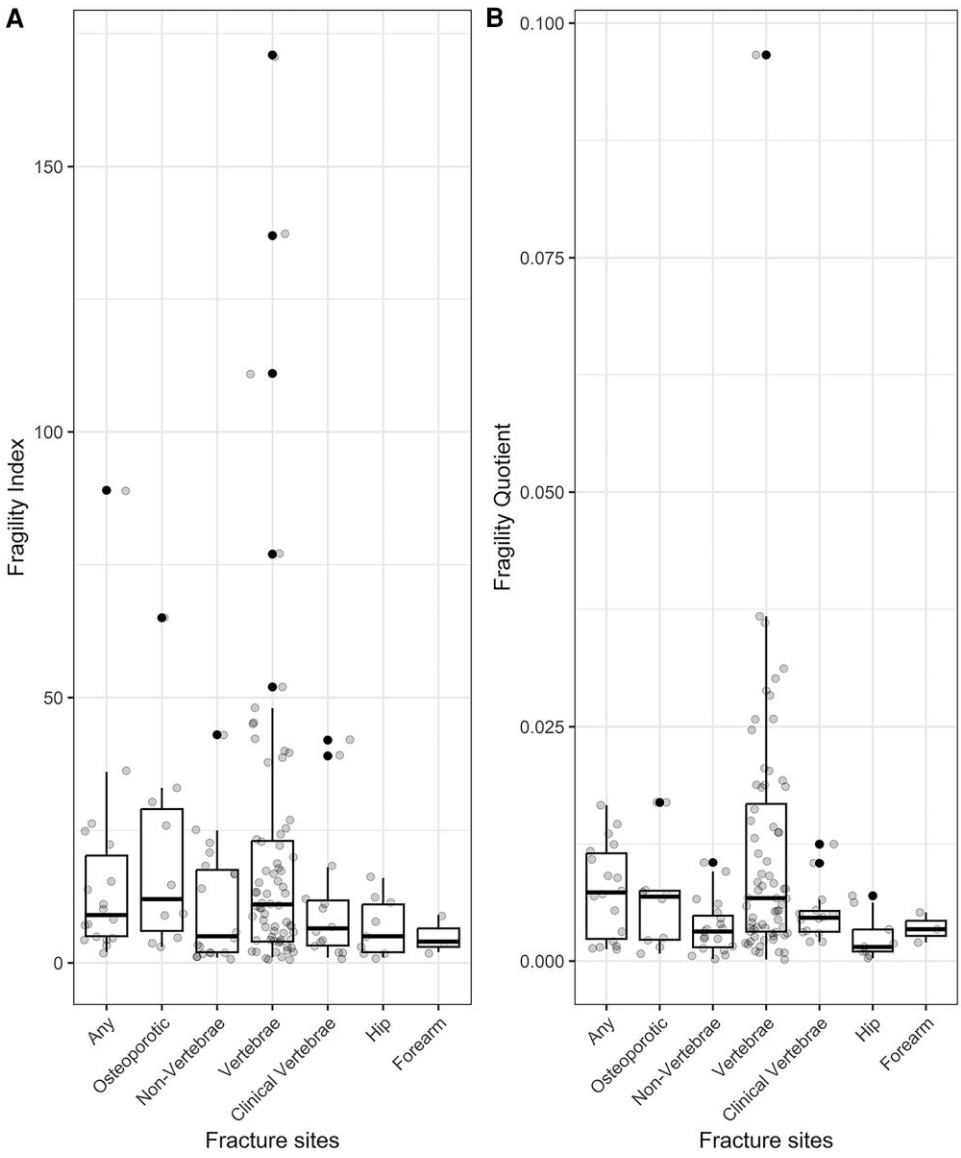

Figure 5. Fragility of evidence for anti-fracture efficacy by fracture sites. A, Fragility Index. B, Fragility Quotient.

as the minimum number of reassignments of the participants with the longest follow-up time from the intervention group to the control group, leading to a loss of statistical significance using a log-rank test (57), has been proposed for a time-to-event outcome. This survival-inferred FI, nevertheless, does not share the exact principles of the original FI as it reassigns the participants between the study arms but not the outcome of interest (57). Similarly, a Continuous Fragility Index (CFI) has been proposed for continuous outcomes (58). In contrast to the FI, the CFI requires more complex calculations, including iterative simulations and adjustments to individual data points (58). Many RCTs have used improvements in bone mineral density as a surrogate for evaluating the anti-fracture efficacy of pharmacological interventions. Future research applying the CFI to assess the robustness of evidence based on bone mineral density changes would complement our findings and enhance the assessment of fragility of evidence for anti-fracture efficacy in real-world clinical scenarios.

The current findings have potential implications for research, guideline development, and clinical practice. Researchers might consider lowering the *P* value threshold, most likely to .001 as demonstrated by our findings, to enhance the strength and reliability of RCT evidence for anti-fracture efficacy. It has recently been proposed to lower the threshold for statistical significance to .005 (59-61) to minimize the risk of false-positive conclusions (59, 60), *P* hacking (1), and underpowered trials (1). This proposed threshold might also encourage researchers to rely more on effect size rather than *P* values, though it would make future studies larger, more costly, and less feasible. Lowering a *P* value threshold might also be associated with increased false negatives, as clinically important findings with a relatively large *P* value might not be captured. Strong RCT evidence from well-designed and sufficiently powered trials is practically preferred as the implementation of an intervention with fragile evidence might lead to more harm than good to patients. Future trials examining anti-fracture efficacy are thus expected to have better study design, better pharmacological efficacy, and longer duration of follow-up to improve the robustness of RCT evidence. Finally, the FI, its comparison to the number of study participants lost to follow-up, the average

rate of the outcome event, and FQ should be incorporated into not only the development of clinical guidelines but also in the decision-making process. The incorporation of these measures that quantify the fragility of the evidence with other statistical measures such as a *P* value and effect sizes help clinicians and readers identify the most robust evidence, ultimately improving the quality of health care interventions.

The findings should be interpreted in the context of their strengths and limitations. This is, to our knowledge, the first study to systematically assess the robustness of evidence for anti-fracture efficacy of specific pharmacological interventions on specific fracture sites. A comprehensive search was conducted to capture all RCTs that examined the anti-fracture efficacy of different pharmacological interventions in high-impact journals over the last 30 years. The predefined subgroup analyses of analysis results with fragility fractures predefined as a primary endpoint and those with $P < .001$ provided insights into the fragility of evidence for anti-fracture efficacy.

However, FI and FQ were originally developed for assessing a binary outcome, whereas two-thirds of the included analysis results were derived from a time-to-event analysis. As the assumption of proportional hazards was met for these time-to-event analyses, it is unlikely that a statistically significant result from a time-to-event analysis turns out to be nonsignificant when the outcome is treated as binary. To ensure the validity of the FI calculation, we excluded 8 results with statistically nonsignificant results with a Fisher's exact test. Second, the threshold of fragility has not been established, making it practically hard to conclude the robustness of RCT evidence (3, 62). However, it is at least alarming that reversing the fracture status of only 9 participants or 0.51% of the sample size, which might have occurred among those lost to follow-up, would make the reported significance statistically nonsignificant, thus losing the evidence and altering the clinical guidelines. An FI threshold between 19 and 22 has been suggested from a recent meta-analysis in cardiology (49). Nevertheless, this rule of thumb is preliminary, should be specific to medical fields and needs to be considered on a case-by-case basis. Third, our study did not include certain pharmacological interventions, such as hormone replacement therapy (HRT), selective estrogen receptor modulators (SERMs), and odanacatib—nor did it consider trials published outside our predefined list of high-impact general medicine and bone-specific journals. These interventions were excluded either because they are not primarily indicated for fracture prevention (HRT and SERMs) or were never approved (odanacatib) (63). Trials published in these high-impact journals were assumed to be of high quality and to provide the necessary data for our analysis. Our additional search of studies published outside of these high-impact journals found only 2 relevant trials and 4 follow-up studies that met all other criteria. More importantly, the exploratory analysis, including all relevant trials regardless of their publication journals, yielded almost identical results with the primary analysis, confirming the robustness of our findings.

In conclusion, existing RCT evidence of anti-fracture efficacy is highly fragile. The FI and FQ provide an additional easy-to-interpret means of assessing the strength of RCT evidence along with the conventional metrics of *P* value and effect size. Therefore, the FI, its comparison with loss to follow-up, and its extension of FQ should be incorporated into clinical guideline development and doctor-patient risk communication to help identify robust RCT evidence and improve healthcare quality.

Disclosures

N.T. and T.T. have no conflict of interest. T.V.N. has received grants from the Australian National Health and Medical Research Council and Amgen Competitive Grant Program; fees for lectures from Amgen, Bridge Health Care (Vietnam), DKSH Pharma, MSD and VT Health Care (Vietnam); and support from Amgen for attending the annual meeting of APCO and from VT Health Care (Vietnam) for attending the Vietnam Osteoporosis Society Annual Scientific Meeting. Dr. Nguyen is also a Senior Advisor for the Vietnam Osteoporosis Society, and an Executive Member of the Asia Pacific Consortium on Osteoporosis.

Data Availability

The statistical analysis plan, dataset and R Markdown file are publicly accessible (https://github.com/NT-max/Fragility-of-evidence-for-the-efficacy-of-anti-fracture-medications).

References

1. Ioannidis JP. The importance of predefined rules and prespecified statistical analyses: do not abandon significance. *JAMA*. 2019;321(21):2067-2068.
2. Sterne JA, Davey Smith G. Sifting the evidence- what's wrong with significance tests? *BMJ*. 2001;322(7280):226-231.
3. Walsh M, Srinathan SK, McAuley DF, *et al*. The statistical significance of randomized controlled trial results is frequently fragile: a case for a fragility Index. *J Clin Epidemiol*. 2014;67(6):622-628.
4. Ahmed W, Fowler RA, McCredie VA. Does sample size matter when interpreting the fragility Index? *Crit Care Med*. 2016;44(11):e1142-e1143.
5. Lin L, Xing A, Chu H, *et al*. Assessing the robustness of results from clinical trials and meta-analyses with the fragility index. *Am J Obstet Gynecol*. 2023;228(3):276-282.
6. Kampman JM, Turgman O, Sperna Weiland NH, *et al*. Statistical robustness of randomized controlled trials in high-impact journals has improved but was low across medical specialties. *J Clin Epidemiol*. 2022;150:165-170.
7. Huang X, Chen B, Thabane L, Adachi JD, Li G. Fragility of results from randomized controlled trials supporting the guidelines for the treatment of osteoporosis: a retrospective analysis. *Osteoporos Int*. 2021;32(9):1713-1723.
8. Lin L, Chu H. Assessing and visualizing fragility of clinical results with binary outcomes in R using the fragility package. *PLoS One*. 2022;17(6):e0268754.
9. Kendler DL, Marin F, Zerbini CAF, *et al*. Effects of teriparatide and risedronate on new fractures in post-menopausal women with severe osteoporosis (VERO): a multicentre, double-blind, double-dummy, randomised controlled trial. *Lancet*. 2018;391(10117):230-240.
10. Reid IR, Horne AM, Mihov B, *et al*. Fracture prevention with zoledronate in older women with osteopenia. *N Engl J Med*. 2018;379(25):2407-2416.
11. Nakamura T, Fukunaga M, Nakano T, *et al*. Efficacy and safety of once-yearly zoledronic acid in Japanese patients with primary osteoporosis: two-year results from a randomized placebo-controlled double-blind study (ZOledroNate treatment in efficacy to osteoporosis; ZONE study). *Osteoporos Int*. 2017;28(1):389-398.
12. Saag KG, Petersen J, Brandi ML, *et al*. Romosozumab or alendronate for fracture prevention in women with osteoporosis. *N Engl J Med*. 2017;377(15):1417-1427.
13. Cosman F, Crittenden DB, Adachi JD, *et al*. Romosozumab treatment in postmenopausal women with osteoporosis. *N Engl J Med*. 2016;375(16):1532-1543.
14. Miller PD, Hattersley G, Riis BJ, *et al*. Effect of abaloparatide vs placebo on new vertebral fractures in postmenopausal women

with osteoporosis: a randomized clinical trial. *JAMA*. 2016; 316(7):722-733.
15. Boonen S, Reginster JY, Kaufman JM, *et al.* Fracture risk and zoledronic acid therapy in men with osteoporosis. *N Engl J Med*. 2012;367(18):1714-1723.
16. Nakamura T, Sugimoto T, Nakano T, *et al.* Randomized Teriparatide [human parathyroid hormone (PTH) 1-34] Once-Weekly Efficacy Research (TOWER) trial for examining the reduction in new vertebral fractures in subjects with primary osteoporosis and high fracture risk. *J Clin Endocrinol Metab*. 2012; 97(9):3097-3106.
17. Cummings SR, San Martin J, McClung MR, *et al.* Denosumab for prevention of fractures in postmenopausal women with osteoporosis. *N Engl J Med*. 2009;361(8):756-765.
18. Smith MR, Egerdie B, Hernandez Toriz N, *et al.* Denosumab in men receiving androgen-deprivation therapy for prostate cancer. *N Engl J Med*. 2009;361(8):745-755.
19. Matsumoto T, Hagino H, Shiraki M, *et al.* Effect of daily oral minodronate on vertebral fractures in Japanese postmenopausal women with established osteoporosis: a randomized placebo-controlled double-blind study. *Osteoporos Int*. 2009;20(8):1429-1437.
20. Greenspan SL, Bone HG, Ettinger MP, *et al.* Effect of recombinant human parathyroid hormone (1-84) on vertebral fracture and bone mineral density in postmenopausal women with osteoporosis: a randomized trial. *Ann Intern Med*. 2007;146(5):326-339.
21. Black DM, Delmas PD, Eastell R, *et al.* Once-yearly zoledronic acid for treatment of postmenopausal osteoporosis. *N Engl J Med*. 2007;356(18):1809-1822.
22. Lyles KW, Colon-Emeric CS, Magaziner JS, *et al.* Zoledronic acid and clinical fractures and mortality after hip fracture. *N Engl J Med*. 2007;357(18):1799-1809.
23. Reginster JY, Seeman E, De Vernejoul MC, *et al.* Strontium ranelate reduces the risk of nonvertebral fractures in postmenopausal women with osteoporosis: Treatment of Peripheral Osteoporosis (TROPOS) study. *J Clin Endocrinol Metab*. 2005;90(5):2816-2822.
24. Chesnut CH, 3rd, Skag A, Christiansen C, *et al.* Effects of oral ibandronate administered daily or intermittently on fracture risk in postmenopausal osteoporosis. *J Bone Miner Res*. 2004;19(8): 1241-1249.
25. Meunier PJ, Roux C, Seeman E, *et al.* The effects of strontium ranelate on the risk of vertebral fracture in women with postmenopausal osteoporosis. *N Engl J Med*. 2004;350(5):459-468.
26. Trivedi DP, Doll R, Khaw KT. Effect of four monthly oral vitamin D3 (cholecalciferol) supplementation on fractures and mortality in men and women living in the community: randomised double blind controlled trial. *BMJ*. 2003;326(7387):469.
27. Neer RM, Arnaud CD, Zanchetta JR, *et al.* Effect of parathyroid hormone (1-34) on fractures and bone mineral density in postmenopausal women with osteoporosis. *N Engl J Med*. 2001;344(19): 1434-1441.
28. Reginster J, Minne HW, Sorensen OH, *et al.* Randomized trial of the effects of risedronate on vertebral fractures in women with established postmenopausal osteoporosis. Vertebral Efficacy with Risedronate Therapy (VERT) Study Group. *Osteoporos Int*. 2000; 11(1):83-91.
29. Pols HA, Felsenberg D, Hanley DA, *et al.* Multinational, placebo-controlled, randomized trial of the effects of alendronate on bone density and fracture risk in postmenopausal women with low bone mass: results of the FOSIT study. Fosamax International Trial Study Group. *Osteoporos Int*. 1999;9(5):461-468.
30. Harris ST, Watts NB, Genant HK, *et al.* Effects of risedronate treatment on vertebral and nonvertebral fractures in women with postmenopausal osteoporosis: a randomized controlled trial. Vertebral Efficacy With Risedronate Therapy (VERT) Study Group. *JAMA*. 1999;282(14):1344-1352.
31. Cummings SR, Black DM, Thompson DE, *et al.* Effect of alendronate on risk of fracture in women with low bone density but without vertebral fractures: results from the fracture intervention trial. *JAMA*. 1998;280(24):2077-2082.
32. Dawson-Hughes B, Harris SS, Krall EA, Dallal GE. Effect of calcium and vitamin D supplementation on bone density in men and women 65 years of age or older. *N Engl J Med*. 1997;337(10): 670-676.
33. Black DM, Cummings SR, Karpf DB, *et al.* Randomised trial of effect of alendronate on risk of fracture in women with existing vertebral fractures. Fracture Intervention Trial Research Group. *Lancet*. 1996;348(9041):1535-1541.
34. Chapuy MC, Arlot ME, Duboeuf F, *et al.* Vitamin D3 and calcium to prevent hip fractures in elderly women. *N Engl J Med*. 1992; 327(23):1637-1642.
35. Body JJ, Marin F, Kendler DL, *et al.* Efficacy of teriparatide compared with risedronate on FRAX((R))-defined major osteoporotic fractures: results of the VERO clinical trial. *Osteoporos Int*. 2020;31(10):1935-1942.
36. Leder BZ, Mitlak B, Hu MY, Hattersley G, Bockman RS. Effect of abaloparatide vs alendronate on fracture risk reduction in postmenopausal women with osteoporosis. *J Clin Endocrinol Metab*. 2020;105(3):938-943.
37. Radford LT, Bolland MJ, Mason B, *et al.* The Auckland calcium study: 5-year post-trial follow-up. *Osteoporos Int*. 2014;25(1): 297-304.
38. Meunier PJ, Roux C, Ortolani S, *et al.* Effects of long-term strontium ranelate treatment on vertebral fracture risk in postmenopausal women with osteoporosis. *Osteoporos Int*. 2009;20(10): 1663-1673.
39. Watts NB, Chines A, Olszynski WP, *et al.* Fracture risk remains reduced one year after discontinuation of risedronate. *Osteoporos Int*. 2008;19(3):365-372.
40. Chapuy MC, Arlot ME, Delmas PD, Meunier PJ. Effect of calcium and cholecalciferol treatment for three years on hip fractures in elderly women. *BMJ*. 1994;308(6936):1081-1082.
41. Tran N, Tran T, Nguyen TV. Fragility of Evidence for the Efficacy of Anti-fracture Medications. *figshare*. https://doi.org/10.6084/m9.figshare.28803986. Published April 16 2025.
42. Reid IR, Mason B, Horne A, *et al.* Randomized controlled trial of calcium in healthy older women. *Am J Med*. 2006;119(9):777-785.
43. Ringe JD, Faber H, Farahmand P, Dorst A. Efficacy of risedronate in men with primary and secondary osteoporosis: results of a 1-year study. *Rheumatol Int*. 2006;26(5):427-431.
44. Kushida K, Shiraki M, Nakamura T, *et al.* The efficacy of alendronate in reducing the risk for vertebral fracture in Japanese patients with osteoporosis: a randomized, double-blind, active-controlled, double-dummy trial. *Curr Ther Res Clin Exp*. 2002;63(9):606-620.
45. Cosman F, Miller PD, Williams GC, *et al.* Eighteen months of treatment with subcutaneous abaloparatide followed by 6 months of treatment with alendronate in postmenopausal women with osteoporosis: results of the ACTIVExtend trial. *Mayo Clin Proc*. 2017;92(2):200-210.
46. Sugimoto T, Shiraki M, Nakano T, *et al.* Vertebral fracture risk after once-weekly teriparatide injections: follow-up study of Teriparatide Once-Weekly Efficacy Research (TOWER) trial. *CMRO*. 2013;29(3):195-203.
47. Reginster JY, Felsenberg D, Boonen S, *et al.* Effects of long-term strontium ranelate treatment on the risk of nonvertebral and vertebral fractures in postmenopausal osteoporosis: results of a five-year, randomized, placebo-controlled trial. *Arthritis Rheum*. 2008;58(6):1687-1695.
48. Kushida K, Shiraki M, Nakamura T, *et al.* Alendronate reduced vertebral fracture risk in postmenopausal Japanese women with osteoporosis: a 3-year follow-up study. *J Bone Miner Metab*. 2004;22(5):462-468.
49. Murad MH, Kara Balla A, Khan MS, Shaikh A, Saadi S, Wang Z. Thresholds for interpreting the fragility index derived from sample of randomised controlled trials in cardiology: a meta-epidemiologic study. *BMJ Evid Based Med*. 2023;28(2):133-136.
50. Wasserstein RL, Lazae NA. The ASA's statement on p-values: context, process, and purpose. *Am Stat*. 2016;70(2):129-133.

51. Trevena LJ, Zikmund-Fisher BJ, Edwards A, *et al*. Presenting quantitative information about decision outcomes: a risk communication primer for patient decision aid developers. *BMC Med Inform Decis Mak*. 2013;13 Suppl 2(S2):S7.
52. Zipkin DA, Umscheid CA, Keating NL, *et al*. Evidence-based risk communication: a systematic review. *Ann Intern Med*. 2014; 161(4):270-280.
53. Ahmed H, Naik G, Willoughby H, Edwards AGK. Communicating risk. *BMJ*. 2012;344(jun18 1):e3996.
54. Akl EA, Oxman AD, Herrin J, *et al*. Using alternative statistical formats for presenting risks and risk reductions. *Cochrane Database Syst Rev*. 2011;2011:CD006776.
55. Gigerenzer G, Gaissmaier W, Kurz-Milcke E, Schwartz LM, Woloshin S. Helping doctors and patients make sense of health statistics. *Psychol Sci Public Interest*. 2007;8(2):53-96.
56. Baer BR, Gaudino M, Fremes SE, Charlson M, Wells MT. The fragility index can be used for sample size calculations in clinical trials. *J Clin Epidemiol*. 2021;139:199-209.
57. Bomze D, Asher N, Hasan Ali O, *et al*. Survival-inferred fragility Index of phase 3 clinical trials evaluating immune checkpoint inhibitors. *JAMA Netw Open*. 2020;3(10): e2017675.
58. Caldwell JE, Youssefzadeh K, Limpisvasti O. A method for calculating the fragility index of continuous outcomes. *J Clin Epidemiol*. 2021;136:20-25.
59. Benjamin DJ, Berger JO, Johnson VE. Redefine statistical significance. *Nat Hum Behav*. 2018;2(1):6-10.
60. Ioannidis JPA. The proposal to lower P value thresholds to .005. *JAMA*. 2018;319(14):1429-1430.
61. Wayant C, Scott J, Vassar M. Evaluation of lowering the P value threshold for statistical significance from .05 to .005 in previously published randomized clinical trials in major medical journals. *JAMA*. 2018;320(17):1813-1815.
62. Niforatos JD, Zheutlin AR, Chaitoff A, Pescatore RM. The fragility index of practice changing clinical trials is low and highly correlated with P-values. *J Clin Epidemiol*. 2020;119:140-142.
63. US Preventive Services Task Force; Mangione CM, Barry MJ, *et al*. Hormone therapy for the primary prevention of chronic conditions in postmenopausal persons: US preventive services task force recommendation statement. *JAMA*. 2022;328(17):1740-1746.