Circulation: Cardiovascular Quality and Outcomes

ORIGINAL ARTICLE

Systematic Evaluation of the Robustness of the Evidence Supporting Current Guidelines on Myocardial Revascularization Using the Fragility Index

See Editorial by Seligman et al

Mario Gaudino, MD
Irbaz Hameed, MD
Giuseppe Biondi-Zoccai, MD
Derrick Y. Tam, MD
Stephen Gerry, MSc
Mohamed Rahouma, MD
Faiza M. Khan, MD
Dominick J. Angiolillo, MD
Umberto Benedetto, MD
David P. Taggart, MD, PhD
Leonard N. Girardi, MD
Filippo Crea, MD
Marc Ruel, MD, MPH
Stephen E. Fremes, MD

BACKGROUND: RCTs (randomized controlled trials) are the preferred source of evidence to support professional societies' guidelines. The fragility index (FI), defined as the minimum number of patients whose status would need to switch from nonevent to event to render a statistically significant result nonsignificant, quantitatively estimates the robustness of RCT results. We evaluate RCTs supporting current guidelines on myocardial revascularization using the FI and FI minus number of patients lost to follow-up.

METHODS AND RESULTS: The FI and FI minus number of patients lost to follow-up of RCTs supporting the 2012 American College of Cardiology/American Heart Association Guideline for the Diagnosis and Management of Patients with Stable Ischemic Heart Disease, the 2014 Focused Update of the American College of Cardiology/American Heart Association Guideline for the Diagnosis and Management of Patients with Stable Ischemic Heart Disease, and the 2018 European Society of Cardiology/European Association for Cardio-Thoracic Surgery Guidelines for Myocardial Revascularization were calculated. Of 414 RCTs identified, 160 were eligible for FI calculation. The median FI was 8.0 (95% CI, 5.0–9.0) and the median FI minus number of patients lost to follow-up was 1.0 (95% CI, 0.0–3.0). FI was ≤3, indicating very limited robustness, in 44 (27.5%) RCTs, and was lower than the number LTF, indicating limited robustness, in 68 (42.5%) RCTs. FI was significantly (all $P<0.05$) correlated with the sample size, number of events, statistical power, journal impact factor, use of intention-to-treat analysis, and of composite end points and negatively correlated with the use of percutaneous interventions in the treatment arm and the *P*-value level.

CONCLUSIONS: More than a quarter of RCTs that support current guidelines on myocardial revascularization have a FI of 3 or lower, and over 40% of trials reveal a FI that is lower than the number of patients lost to follow-up. These findings suggest that the robustness of the findings that support current myocardial revascularization guidelines is tenuous and vulnerable to change as new evidence from RCTs appears.

Key Words: intention to treat analysis ■ lost to follow-up ■ myocardial revascularization ■ sample size

https://www.ahajournals.org/journal/circoutcomes

WHAT IS KNOWN

- RCTs (randomized controlled trials) are the preferred source of evidence in guidelines and their statistical significance is evaluated using the *P*-value of the CI.
- This approach has been repeatedly criticized by authors and statistical associations.
- The fragility index (FI), defined as the number of patients needed to switch from nonevent to event to render a trial's result insignificant, can be used to determine the solidity of a trial's results.

WHAT THE STUDY ADDS

- We analyzed the solidity of the randomized trials supporting current guidelines on myocardial revascularization using the FI and the FI minus number of patients lost to follow-up.
- More than a quarter of RCTs supporting current guidelines on myocardial revascularization have a FI of 3 or lower, and over 40% of trials reveal a FI that is lower than the number of patients lost to follow-up

RCTs (randomized controlled trials) are considered the gold standard to compare 2 or more treatments. Randomization aims to minimize (and ideally eliminate) the effect of known, unmeasured, or unknown confounders and is traditionally accepted to be able to provide a reliable estimate of whether or not a treatment has an effect (or a larger effect than another).

RCTs are the preferred source of evidence in guidelines and professional societies or expert position papers. They are the foundations of evidence-based medicine and influence the clinical decision-making of the great majority of physicians and affect the outcomes of most of our patients.

The statistical significance of the results of RCTs is generally evaluated using a fixed threshold of metrics, such as the *P*-value of the CI. However, this approach has been repeatedly criticized by authors and statistical associations in the recent past.[1,2] Indeed, the *P*-value is fraught with all the limitations inherent to a frequentist statistical framework, which include null hypothesis testing, inability to incorporate prior knowledge, risk of being misinterpreted as proof of evidence or direct probability statement, occasional internal logical inconsistency, accurate estimation only based on large samples, and reliance on approximation for many computational models. In addition, the *P*-value does not provide any direct information on the existence of a true treatment effect and, according to many, is a simplistic solution to the complexities of probability theory.[3] *P*-value-based statistical significance is also heavily affected by methodological limitations and can be lost (or gained) by a shift of few events in one group.

To partially overcome these limitations and to provide an objective estimate of the solidity of RCTs, the fragility index (FI) was introduced in 1990.[4] The FI is defined as the minimum number of patients whose status would need to switch from nonevent to event to render a significant result nonsignificant. The lower the value of FI, the lower the solidity and robustness of the results.[4,5] While FI is still dependent on the *P*-value and the problems associated with choosing an arbitrary threshold, it adds additional information as a form of sensitivity analysis which combines the sample size and the precision of the point estimate. Accordingly, FI offers a way to stress-test any study *P*-values, offering a pragmatic and poignant sensitivity measure and it evidently proves less focus of dichomotous testing in itself but actually weighs it in light of sample size and events accrued. The FI may provide an additional perspective informing practitioners on the robustness of the findings of a RCT analyzed according to a frequentist framework and reporting *P*-values. Bayesian approaches may provide a better alternative to counter the problems associated with frequentist analyses of randomized trials.

Cardiovascular disease affects >85 million of people in the United States only.[6] Current guidelines on myocardial revascularization potentially affect the lives of millions of patients worldwide.

In this report, we evaluate the solidity of the evidence supporting the current US and European guidelines on myocardial revascularization using the FI.

METHODS

Selection of Randomized Controlled Trials

We identified all RCTs cited in the 2012 American College of Cardiology (ACC)/American Heart Association (AHA) Guideline for the Diagnosis and Management of Patients with Stable Ischemic Heart Disease,[7] the 2014 Focused Update of the ACC/AHA Guideline for the Diagnosis and Management of Patients with Stable Ischemic Heart Disease,[8] and the 2018 European Society of Cardiology (ESC)/European Association for Cardio-Thoracic Surgery (EACTS) Guidelines for Myocardial Revascularization.[9] All trials were independently reviewed by two reviewers (Drs Hameed and Rahouma) and were included for analysis and data extraction if they reported at least one statistically significant dichotomous primary or secondary outcome ($P<0.05$ or a 95% CI that excluded the null value). For trials reporting multiple significant primary and secondary outcomes, data were extracted for the primary and secondary outcome with the lowest *P*-value. The data that support the findings of this study are available from the corresponding author on reasonable request.

Data Extraction

For each RCT, the following data were recorded: citation in European or American guideline, journal of publication and impact factor (according to Thomson Reuters-Clarivate Analytics), year of publication, use or surgical, medical or

percutaneous intervention in treatment arm, single- or multicenter study, geographic locations of the participating centers, class of recommendation and level of evidence (LOE) supported. Details of the primary or secondary outcome (definition of outcome, composite or noncomposite end points) sample size, number of events, primary analysis (intention-to-treat or alternatives), statistical power, summary statistics, *P*-value for the primary outcome, number of patients lost to follow-up (LTF), and number of crossovers were also collected. Two reviewers (Drs Hameed and Rahouma) independently extracted data from the included trials and the first author (Dr Gaudino) resolved any discrepancy.

Calculation of Fragility Index

FI for the statistically significant primary or secondary outcomes were calculated as described by Walsh et al.[5] The results for each outcome were entered in a 2×2 contingency table following which the *P*-value for each outcome was calculated using the 2-sided Fisher exact test.

Single participants were iteratively shifted one at a time in the lower-incidence treatment group from nonevent to event and the *P*-value for the 2×2 table re-calculated. The FI for an outcome equalled the smallest number of participants required to turn the re-calculated *P*-value nonsignificant (≥0.05).

Calculation of Fragility Index Minus Lost to Follow-Up

FI minus number of patients lost to follow-up (FI-LTF) were calculated following the methods used by Mazzinari et al[10] as the difference between the fragility index and number of patients lost to follow-up.

Statistical Analysis

Continuous variables were reported as medians with their first-third quartile, whereas categorical variables were reported as counts and percentages.

A visual inspection of the data showed that FI and FI-LTF were non-normally distributed; therefore, nonparametric methods were used to compare the groups. The Mann-Whitney *U* test was used to compare 2 groups, and the Kruskal-Wallis test was used to compare 3 or more groups. To assess the interaction between 2 categorical variables with respect to FI/FI-LTF, the Scheirer-Ray-Hare test was used. Categorical variables were compared using Fisher exact test.

Correlations between FI, FI-LTF, and different variables were calculated using Spearman correlation coefficient for continuous variables and rank-biserial correlation coefficient for dichotomous variables. A local regression curve was used to explore the relationship between the variables.

Multivariable linear regression with a quasi-Poisson distribution was used to explore for independent predictors of FI and FL-LTF, with an exploratory and hypothesis-generating scope. Results are reported as regression coefficient (β) and 95% CI. The 95% CIs were calculated using percentile bootstrapping with 1000 samples. Two-sided significance testing was used and a *P*-value<0.05 was considered significant. All analyses were performed using SPSS version 22 (IBM, Chicago, IL) and R (version 3.3.3 R Project for Statistical Computing) within RStudio.

RESULTS

Selection of Randomized Controlled Trials and Data Analysis

A total of 414 RCTs were identified: 171 in the ACC/AHA guidelines, and 243 in the ESC/EACTS guidelines. Two hundred eighty-one RCTs reported dichotomous outcomes, of which 167 (64 in the ACC/AHA guidelines and 103 in the ESC/EACTS guidelines) reported at least 1 statistically significant primary or secondary outcome (125 primary, 35 secondary). Seven RCTs were quoted in both ACC/AHA and ESC/EACTS guidelines and were thus entered only once to avoid duplication, leaving 160 RCTs for the analyses (Figure 1; Table I in the Data Supplement). Thirty-three RCTs (20.6%) were published before the year 2000, 69 (43.1%) from 2000 to 2010, and 58 (36.3%) after 2010.

There were 143 (90.5%) multicenter RCTs, and 78 (48.8%) originated from Europe. The median sample size was 1192 (379–2672). The details of the RCTs and end points are summarized in Table 1.

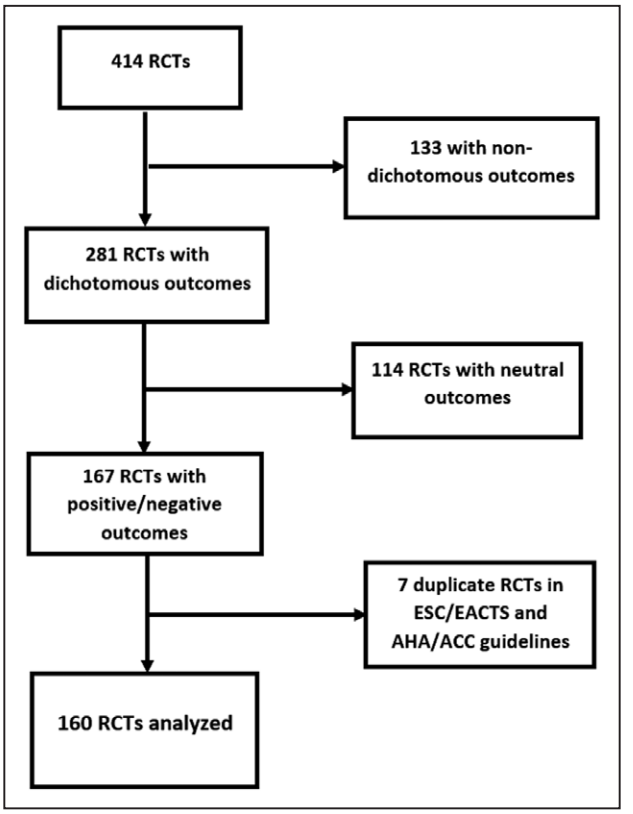

Figure 1. Flowchart of the selection process of randomized controlled trials (RCTs).
ACC indicates American College of Cardiology; AHA, American Heart Association; EACTS, European Association for Cardio-Thoracic Surgery; and ESC, European Society of Cardiology.

Table 1. Characteristics of Randomized Controlled Trials Cited in the Guidelines

	All Trials*	AHA/ACC Guidelines	ESC/EACTS Guidelines	*P*-Value†
Number	160	64	103	
Year of publication (<2000)	33 (20.6)	31 (49.2)	2 (2.1)	<0.001
Year of publication (2000–2010)	69 (43.1)	29 (46.0)	40 (41.2)	
Year of publication (>2010)	58 (36.3)	3 (4.8)	55 (56.7)	
Primary/secondary end point, n (%)	125/35 (78.1/21.9)	53/11 (82.8/17.2)	79/24 (76.7/23.3)	0.44
Composite primary end point	81 (50.9)	32 (50.8)	54 (52.4)	0.87
Composite secondary end point	9 (5.6)	4 (6.2)	5 (4.9)	0.73
Sample size	1192.0 (378.8–2671.5)	1805.0 (562.5–4033.0)	888.0 (337.0–2157.0)	0.02
Number of intervention patients	594.0 (191.3–1291.3)	905.5 (281.3–2012.0)	442.0 (180.5–1061.5)	0.02
Number of control patients	598.0 (191.0–1275.3)	903.0 (281.3–2021.0)	438.0 (178.5–1081.0)	0.02
% power	80.0 (80.0–90.0)	85.0 (80.0–90.0)	80.0 (80.0–90.0)	0.03
Number of events in intervention patients	53.0 (19.8–158.3)	108.5 (34.3–275.5)	38.0 (16.5–103.0)	<0.01
Number of events in control patients	66.0 (20.0–187.3)	150.0 (25.3–262.5)	52.0 (17.5–112.5)	0.001
Number of patients lost to follow-up	6.5 (0.0–40.3)	15.5 (0.3–63.0)	5.0 (0.0–38.0)	0.07
% crossover	4.2 (1.9–8.0)	5.19 (2.4–8.3)	4.2 (1.8–8.0)	0.75
Surgery trials, n (%)	123 (76.9)	42 (65.6)	83 (80.6)	0.04
Percutaneous intervention trials, n (%)	70 (43.8)	43 (67.2)	28 (27.2)	<0.001
Medical treatment trials, n (%)	83 (51.9)	20 (31.2)	66 (64.1)	<0.001
Location				0.05
Asia	9 (5.6)	1 (1.6)	9 (8.7)	
Europe	78 (48.8)	30 (46.9)	51 (49.5)	
North America	25 (15.6)	14 (21.9)	12 (11.7)	
South America	2 (1.2)	2 (3.1)	0 (0.0)	
Multicontinental	46 (28.7)	17 (26.6)	31 (30.1)	
Single/multicenter trials, n (%)	15/143 (9.5/90.5)	7/56 (11.1/88.9)	8/94 (7.8/92.2)	0.58
Intention to treat analysis, n (%)				0.03
Yes	140 (88.1)	61 (96.8)	86 (83.5)	
No	16 (10.1)	2 (3.2)	14 (13.6)	
Not reported	3 (1.9)	0 (0.0)	3 (2.9)	
P-value, n (%)				0.10
0.01–0.001	38 (23.8)	18 (28.1)	23 (22.3)	
0.05–0.01	78 (48.8)	24 (37.5)	56 (54.4)	
<0.001	44 (27.5)	22 (34.4)	24 (23.3)	

Numbers reported as median (IQR) or total (%). ACC indicates American College of Cardiology; AHA, American Heart Association; EACTS, European Association for Cardio-Thoracic Surgery; ESC, European Society of Cardiology; and IQR, interquartile range.

*Total after exclusion of 7 duplicate trials.

†*P*-value was calculated using Mann-Whitney *U* test and Fisher exact tests for continuous and categorical variables, respectively.

Details of CORs, LOEs, and numbers of RCTs for the individual guidelines are summarized in Table 2. There were a total of 396 recommendations (161 ACC/AHA, 235 ESC/EACTS) supported by 375 RCTs (151 ACC/AHA, 224 ESC/EACTS): 168 Class I recommendations supported by 168 RCTs; 120 Class IIa recommendations supported by 80 RCTs; 66 Class IIb recommendations supported by 63 RCTs; and 42 Class III recommendations supported by 64 RCTs.

Of these recommendations, 75 were designated LOE A with 178 RCTs; 154 LOE B with 172 RCTs; and 167 LOE C with 25 RCTs.

Fragility Index and Patients Lost to Follow-Up

The distribution of FI, losses to follow-up, and FI-LTF are shown in Figure 2.

The median FI for all the 160 trials analyzed was 8.0 (95% CI, 5.0–9.0; IQR, 3.0–15.0): 10.0 (95% CI, 8.0–16.0; IQR, 4.0–21.0) for ACC/AHA guidelines, and 5.0 (95% CI, 4.0–8.0; IQR, 2.0–13.5) for ESC/EACTS guidelines (*P*=0.02). FI was ≤ 3 for 44 RCTs (27.5%): 12 (18.8%) in the ACC/AHA and 34 (33.0%) in the ESC/EACTS guidelines (*P*=0.07; Table 3).

Table 2. Number of Randomized Trials Supporting Different Classes of Recommendations and Levels of Evidence

	Overall		ACC/AHA Guidelines		ESC/EACTS Guidelines	
Total number of CORs/RCTs	CORs=396	RCTs=375	CORs=161	RCTs=151	CORs=235	RCTs=224
Class I	168 (42.4)	168 (44.8)	58 (36.0)	44 (29.1)	110 (46.8)	124 (55.4)
Class IIa	120 (30.3)	80 (21.3)	47 (29.2)	39 (25.8)	73 (31.1)	41 (18.3)
Class IIb	66 (26.7)	63 (16.8)	29 (18.0)	36 (23.9)	37 (15.7)	27 (12.1)
Class III	42 (10.6)	64 (17.1)	27 (16.8)	32 (21.2)	15 (6.4)	32 (14.3)
LOEs/RCTs	LOEs=396	RCTs=375	LOEs=161	RCTs=151	LOEs=235	RCTs=224
LOE A	75 (18.9)	178 (47.5)	14 (8.7)	30 (19.9)	61 (26.0)	148 (66.1)
LOE B	154 (38.9)	172 (45.9)	78 (48.4)	102 (67.5)	76 (32.3)	70(31.3)
LOE C	167 (42.2)	25 (6.6)	69 (42.9)	19 (12.6)	98 (41.7)	6 (2.6)

ACC indicates American College of Cardiology; AHA, American Heart Association; COR, Class of recommendation; EACTS, European Association for Cardio-Thoracic Surgery; ESC, European Society of Cardiology; LOE, level of evidence; and RCT, randomized controlled trial.

For primary end points, the overall median FI was 8.5 (IQR, 4.0–16.3): 12.0 (IQR, 5.0–21.0) for ACC/AHA and 7.0 (IQR, 3.0–15.0) for ESC/EACTS guidelines (*P*=0.04). For secondary end points, the median FI was 4.0 (IQR, 2.0–9.0): 7.0 (IQR, 4.0–22.0) for ACC/AHA and 3.5 (IQR, 1.0–7.0) for ESC/EACTS guidelines (*P*=0.07; Table 3).

There was no statistically significant difference in FI for era of publication (*P*=0.66; Table 4), although the median FI was progressively lower for RCTs published in the later study years when considering the eras before 2000, from 2000 to 2010, and after 2010.

There was no difference in the median FI of RCTs used to support different CORs (*P*=0.25) and LOEs (*P*=0.16; Table 4).

The FI of RCTs involving surgery in the treatment arm was not significantly different from that of nonsurgical RCTs (*P*=0.59). RCTs involving percutaneous intervention in the treatment arm had a significantly lower FI than nonpercutaneous intervention RCTs (*P*<0.01).

RCTs with composite end points had a significantly higher FI than RCTs with single end points (*P*=0.03; Table 5).

Details of FI-LTF are given in Table 3. The overall median FI-LTF was 1.0 (95% CI, 0.0–3.0; IQR, 0.0–9.0): 3.0 (95% CI, 0.0–7.0; IQR, 0.0–13.0) for ACC/AHA and 1.0 (95% CI, 0.0–3.0; IQR, 0.0–7.5) for ESC/EACTS guidelines (*P*=0.45).

FI was lower than the number of patients LTF in 68 (42.5%) RCTs: 28 (43.8%) in ACC/AHA and 45 (43.7%) RCTs in ESC/EACTs guidelines (*P*=1.00; Table 3; Figure 2).

Correlation Between Fragility Index, Fragility Index Minus Number Lost to Follow-Up, and Trial Characteristics

The FI was significantly correlated with the sample size (Spearman correlation [R]=0.35; *P*<0.001), number of events (R=0.60; *P*<0.001), statistical power (R=0.21; *P*=0.02), impact factor of journal of publication (R=0.34; *P*<0.001), use of intention to treat analysis (R=0.20; *P*=0.01), and of composite end points (R=0.17; *P*=0.03). FI was negatively correlated with use of percutaneous intervention in the treatment arm (R=–0.25; *P*=0.001), the *P*-value level (R=–0.63; *P*<0.001), and was lower for primary end points (R=–0.18; *P*=0.02; Figure 3; Table II in the Data Supplement).

The FI-LTF was negatively correlated with the *P*-value level (R=–0.42; *P*<0.001) and the number of patients lost to follow-up (R=–0.65; *P*<0.001; Table II in the Data Supplement).

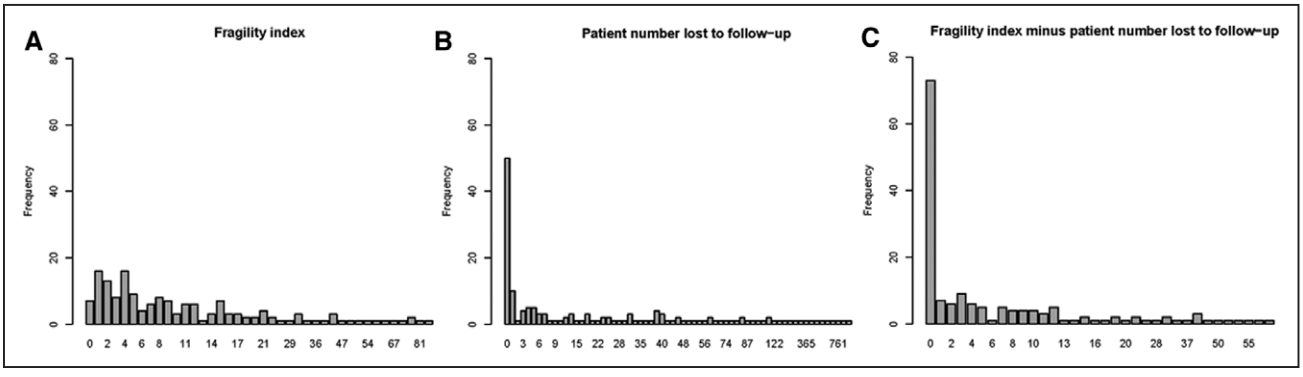

Figure 2. Frequencies of different outcomes.
Frequencies of (**A**) fragility indices, (**B**) patient number lost to follow-up in all trials and (**C**) fragility indices minus patient number lost to follow-up.

Table 3. Fragility Index

	All RCTs	ACC/AHA Guidelines	ESC/EACTS Guidelines	*P*-Value*
Fragility index	8.0 (3.0–15.0)	10.0 (4.0–21.0)	5.0 (2.0–13.5)	0.02
FI-LTF	1.0 (0.0–9.0)	3.0 (0.0–13.0)	1.0 (0.0–7.5)	0.45
Fragility index ≤3, n (%)	44 (27.5)	12 (18.8)	34 (33.0)	0.05
Fragility index <LTF, n (%)	68 (42.5)	28 (43.8)	45 (43.7)	1.00
Fragility index by year of publication				0.22†
<2000	10.0 (4.0–20.0)	10.0 (3.5–20.5)	7.5 (5.8–9.3)	0.73
2000–2010	8.0 (3.8–16.0)	9.5 (5.0–21.0)	4.5 (2.8–12.3)	0.04
>2010	7.0 (2.0–14.8)	33.0 (33.0–43.5)	6.0 (2.0–12.0)	0.01
Primary end points				
Number of end points	125	52	73	
Fragility index	8.5 (4.0–16.3)	12.0 (5.0–21.0)	7.0 (3.0–15.0)	0.04
FI-LTF	1.0 (0.0–10.0)	0.0 (0.0–13.0)	1.0 (0.0–8.0)	0.72
Secondary end points				
Number of end points	35	11	24	
Fragility index	4.0 (2.0–9.0)	7.0 (4.0–22.0)	3.5 (1.0–7.0)	0.07
FI-LTF	2.0 (0.0–8.0)	4.0 (1.5–18.5)	1.0 (0.0–5.5)	0.19

P values were calculated using Mann-Whitney *U* test and Fisher exact tests for continuous and categorical variables respectively. The Scheirer-Ray-Hare test was used to determine *P*-values for nonparametric interactions. Numbers reported as median (IQR) or total (%). ACC indicates American College of Cardiology; AHA, American Heart Association; EACTS, European Association for Cardio-Thoracic Surgery; ESC, European Society of Cardiology; FI-LTF, fragility index minus number lost to follow-up; IQR, interquartile range; LOE, level of evidence; LTF, lost to follow-up; and RCT, randomized controlled trial.

**P* values calculated using Mann-Whitney *U* test and Fisher exact tests for continuous and categorical variables, respectively.

†*P* value calculated using Scheirer-Ray-Hare test.

Multiple regression revealed that sample size, number of events, and *P*-value level were independent predictors of FI (*P*<0.01), while sample size, number of events, number lost to follow-up, and *P*-value level were independent predictors of FI-LTF (*P*<0.05; Table III in the Data Supplement).

DISCUSSION

We have found that the RCTs supporting the current guidelines on myocardial revascularization are generally fragile. The median FI was 8, meaning that a shift of 8 patients from nonevent to event would have significantly changed the results of the original analysis. Of note, 27.5% of the RCTs had a FI≤3, and even the median FI of RCTs used to support class I LOE A recommendations (considered the most solid in clinical medicine) was <10. Even more concerning, the median FI-LTF was 1 and 42.5% of RCTs had a negative FI-LTF value, indicating that their FI was smaller than their number of patients who were lost to follow-up.

Despite the common critiques of surgical RCTs, trials evaluating surgical interventions had FI similar to nonsurgical trials. In contrast, trials aimed at evaluating percutaneous treatment had significantly lower FI.

Confirming previous reports,[10] the sample size and the impact factor of the journal of publication were significantly correlated with the FI. The number of events, statistical power of the trial, *P*-value level, use of composite end points and intention to treat analysis were also correlated with the FI. Similar to previous findings in other fields,[11,12] the treatment effect size was not correlated with the FI.

Interestingly, the FI decreased (although not significantly) over time, with most recent RCTs having lower FI. This may be related to the increasing difficulty in finding support for large RCTs and may elicit concerns regarding enacted policy changes that may have been based on fragile evidence.

Traditionally, the statistical significance of an RCT is judged based on a fixed threshold of *P*-value (usually 0.05). The *P*-value level is influenced by methodological factors and may vary substantially with shifting of only few events from one group to the other. In fact, the use of the *P*-value approach has been heavily criticized in recent years. The American Statistical Association, in a statement of statistical significance and *P*-values, has summarized the many issues related to the use of boundary *P*-values.[2] Reliance on a fixed *P*-value level has also been identified as one of the possible cause of the low level of replication rate in current scientific research.

The FI was introduced in 1990 with the aim of complementing the *P*-value and providing an intuitive measure of the solidity of the results of an RCT.[4] The FI is the number of participants that need to switch from nonevent to

Table 4. Fragility Index of Randomized Controlled Trials by Classes of Recommendations, Levels of Evidence, and Year of Publication

	Number	FI (Median [IQR])	FI-LTF (Median [IQR])
Class of recommendation			
I	43	9.0 (4.0–26.0)	3.0 (0.0–12.5)
IIa	16	6.0 (3.8–12.0)	3.0 (0.0–8.5)
IIb	12	9.0 (1.8–15.3)	0.0 (0.0–6.5)
III	7	5.0 (3.5–6.5)	0.0 (0.0–0.0)
P value		0.25	0.14
Level of evidence			
A	54	8.5 (4.0–23.0)	2.0 (0.0–12.0)
B	22	5.0 (3.0–10.8)	1.0 (0.0–6.8)
C	2	9.0 (5.0–13.0)	0.0 (0.0–0.0)
P value		0.16	0.28
Year of publication			
<2000	33	10.0 (4.0–20.0)	3.00 (0.00–11.00)
2000–2010	69	8.0 (3.8–16.0)	3.00 (0.00–9.25)
>2010	58	7.0 (2.0–14.8)	0.00 (0.00–6.75)
P value		0.66	0.18

P values were calculated using Kruskal-Wallis test. IQR indicates interquartile range; FI, fragility index; and FI-LTF, fragility index minus number lost to follow-up.

Table 5. Comparisons of FIs of Different Categories of RCTs

Comparison	FI	*P*-Value
Surgery vs nonsurgery RCTs	7.0 (3.0–20.0) vs 8.0 (3.0–15.0)	0.59
Surgery vs percutaneous intervention RCTs (n=102)*	7.0 (3.0–20.0) vs 4.0 (2.0–10.0)	0.06
Surgery vs medical treatment RCTs (n=110)†	7.0 (3.0–20.0) vs 9.0 (4.0–19.5)	0.59
Percutaneous intervention vs nonpercutaneous intervention RCTs	5.0 (2.0–12.0) vs 11.0 (5.0–21.0)	<0.01
Percutaneous intervention vs medical treatment RCTs (n=147)‡	5.0 (2.0–11.8) vs 11.50 (5.8–21.0)	0.001
Composite vs noncomposite end points	8.0 (4.0–16.0) vs 4.0 (2.0–12.0)	0.03

P values were calculated using Mann-Whitney *U* test. FI indicates fragility index; and RCT, randomized controlled trial.

*Trials with percutaneous intervention and surgery as their treatment arms were categorized as surgery trials.

†Trials with surgery and medical treatment as their treatment arms were categorized as surgery trials.

‡Trials with percutaneous intervention and medical treatment as their treatment arms were categorized as percutaneous intervention trials.

event in the lower incidence treatment group for a trial to lose statistical significance.[5] In a rather misnomer fashion, a lower FI value indicates lower solidity of the results. An additional important measure is the FI-LTF, as one can make the case that the outcome of patients LTF could have changed the statistical results of the trial. This is particularly relevant when the likelihood of occurrence of the outcome may be the reason for the LTF.[13] So far, no defined FI boundaries exist for the definition of frail RCTs. Intuitively, the FI must be correlated with the sample size, treatment effect, power and existing evidence, and the FI threshold may vary from case to case.

Walsh et al[5] found that among 399 RCTs published in high-impact medical journals between 2004 and 2010, the median FI was 8 and that 25% of them had a FI ≤3. Notably, in 53% of the trials, the FI was lower than the number of patients LTF. Docherty et al,[11] in a review of the RCTs supporting the guidelines for the management of patients with chronic heart failure, had more reassuring results with a median FI of 26, 35% of the trials with a FI≤10 and 20% with FI lower than the number of LTF.

In general, the strength of the published RCTs in different fields has been reported to be low. In a review of the RCTs quoted in the 2016 Chest Guidelines and Expert Recommendations for Venous Thrombo-embolism, the median FI was 5 (median sample size, 400), while a similar review of the 2017 guidelines for the treatment of diabetes mellitus reported a median FI of 16 (median sample size, 2548). Low or very low FI have been reported in trauma (median FI, 3; median sample size, 207), critical care (median FI, 2; median sample size, 126.5), nephrology (median FI, 3; median sample size, 134), spine and sport surgery (median FI, 2 for both, median sample size, 132 and 64, respectively), and anesthesiology (median FI, 4; median sample size, 150).[12,14–20]

It is important to note that there are ethical reasons to design RCT sample sizes to produce the required level of evidence using the minimum number of participants. In fact, while enrolling a larger number of participants may produce stronger evidence against the null, this implies randomizing patients when some level of evidence could be already generated from the available data, and may be seen as violating the equipoise principle. On the other hand, fragile results that are contradicted by subsequent studies or require confirmation in other trials may also be potentially harmful to patients and elicit equally important ethical questions. The delicate balance between the number of patients whose treatment is based on randomization and the solidity of the achieved results is the basis of RCT sample size calculation.

Our findings and similar ones from other medical and surgical specialties highlight the need, whenever using an RCT to inform practice guidelines, to carefully consider, on top of clinical and statistical significance, the actual robustness of the results, and the consequent role of play of chance in guiding its conclusions. We suggest that future practice guidelines provide detailed reports of RCT fragility, and that fragility should be considered when planning pivotal and pragmatic RCTs, and its measurements consistently provided in the main publications to inform patients, clinicians, and stakeholders.

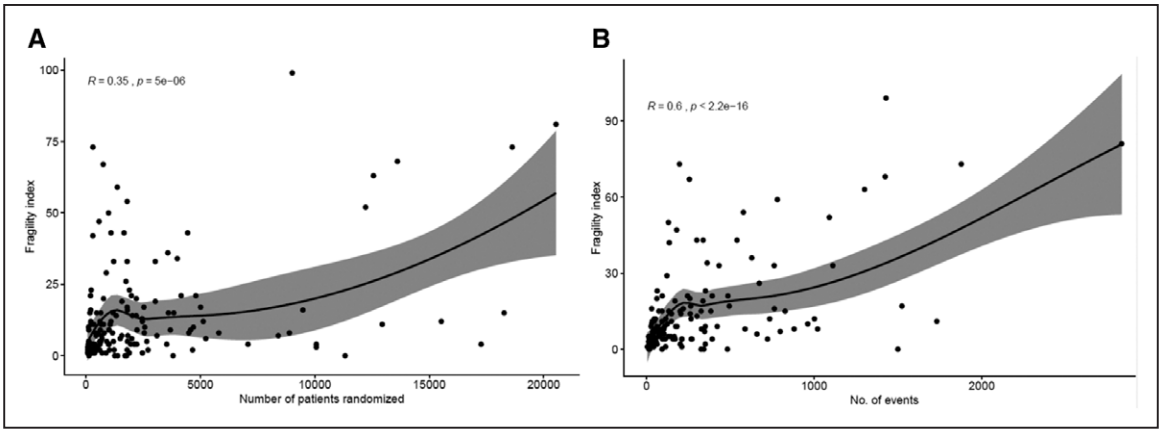

Figure 3. Correlation of fragility index with trial characteristics.
A, Correlation between fragility index and total sample size (Spearman correlation [R]=0.35; *P*<0.001); and (**B**) correlation between fragility index and number of outcomes (Spearman R=0.60; *P*<0.001). Black line represents fitted local regression curve, dark gray area represents 1 SD.

Several limitations of this analysis need to be considered. The FI, like the *P*-value, should not be interpreted as a measure of the effect, but only of the fragility of the results of a trial. In addition, the FI does not per se overcome the limitation of a frequentist framework, but simply provides an additional perspective on a study weaknesses. There are no established boundaries to define an RCT outcome as robust or fragile and the same FI can have different meanings in different clinical contexts. It is also plausible that the FI of primary outcomes consisting of hard end points, such as mortality, may warrant a different interpretation than those involving functional measures. As trials are usually powered for the primary outcome, calculation of the FI for secondary outcomes, which should only be considered hypothesis generating, must be viewed with skepticism. FI, *P*-values, events, and sample are mathematically related and thus their multivariable analysis is limited by collinearity and clustering features, with eventual results mainly exploratory and hypothesis generating. Also, the FI can be applied only to trials with a positive result and a dichotomous outcomes and its calculation convert time-to-event into dichotomous outcomes. Finally, only 160 RCTs were eligible for analysis and the generalizability of our results may be limited.

In conclusion, we have found that the solidity of the RCTs used to support the current guidelines on myocardial revascularization is low and seems to be decreasing in the most recent years. Our data support the need for large RCTs addressing important clinical questions with adequate trial design, power, and sample size. Our findings also suggest that metrics related to FI of the supporting evidence should be incorporated in the recommendations proposed by professional guidelines.

ARTICLE INFORMATION

Received August 13, 2019; accepted October 21, 2019.

The Data Supplement is available at https://www.ahajournals.org/doi/suppl/10.1161/CIRCOUTCOMES.119.006017.

Correspondence

Mario Gaudino, MD, Department of Cardiothoracic Surgery, Weill Cornell Medicine, 525 E 68th St, New York, NY 10065. Email mfg9004@med.cornell.edu

Affiliations

Department of Cardiothoracic Surgery, Weill Cornell Medicine, New York (M.G., I.H., M. Rahouma, F.M.K., L.N.G.). Department of Medico-Surgical Sciences and Biotechnologies, Sapienza University of Rome, Italy (G.B.-Z.). Mediterranea Cardiocentro, Napoli, Italy (G.B.-Z.). Schulich Heart Centre Sunnybrook Health Sciences Centre, University of Toronto, Canada (D.Y.T., S.E.F.). Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology & Musculoskeletal Sciences, University of Oxford, United Kingdom (S.G.). Department of Cardiology, University of Florida, Jacksonville (D.J.A.). Bristol Heart Institute, University of Bristol, School of Clinical Sciences, United Kingdom (U.B.). Nuffield Department of Surgical Sciences, John Radcliffe Hospital, University of Oxford, United Kingdom (D.P.T.). Fondazione Policlinico Universitario A. Gemelli IRCCS, Roma, Italy (F.C.). Department of Cardiovascular and Thoracic Sciences, Università Cattolica de Sacro Cuore, Roma, Italy (F.C.). Division of Cardiac Surgery, University of Ottawa Heart Institute, ON, Canada (M. Ruel).

Acknowledgments

Drs Gaudino, Hameed, Biondi-Zoccai, Tam, Rahouma, Khan, Angiolillo, Benedetto, Taggart, Girardi, Crea, MR, and S. Gerry are experts in clinical research on myocardial revascularization. Dr Biondi-Zoccai and S. Gerry are experts in statistical analysis. Drs Gaudino, Hameed, Biondi-Zoccai, Tam, S. Gerry, MR and Drs Khan, Angiolillo, Taggart, Girardi, Crea, MR, and S. Gerry have expertise in clinical research methodology.

Disclosures

None.

REFERENCES

1. Ioannidis JPA. Why most published research findings are false. *PLoS Med*. 2005;2:e124. doi:10.1371/journal.pmed.0020124
2. Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. *The American Statistician*. 2016;70:129–133. doi:10.1080/00031305.2016.1154108
3. Lee JJ. Demystify statistical significance–time to move on from the p value to bayesian analysis. *J Natl Cancer Inst*. 2011;103:2–3. doi: 10.1093/jnci/djq493
4. Feinstein AR. The unit fragility index: an additional appraisal of "statistical significance" for a contrast of two proportions. *J Clin Epidemiol*. 1990;43:201–209. doi: 10.1016/0895-4356(90)90186-s
5. Walsh M, Srinathan SK, McAuley DF, Mrkobrada M, Levine O, Ribic C, Molnar AO, Dattani ND, Burke A, Guyatt G, Thabane L, Walter SD, Pogue J, Devereaux PJ. The statistical significance of randomized controlled trial results is frequently fragile: a case for a Fragility Index. *J Clin Epidemiol*. 2014;67:622–628. doi: 10.1016/j.jclinepi.2013.10.019

6. Mozaffarian D, Benjamin EJ, Go AS, Arnett DK, Blaha MJ, Cushman M, de Ferranti S, Després JP, Fullerton HJ, Howard VJ, Huffman MD, Judd SE, Kissela BM, Lackland DT, Lichtman JH, Lisabeth LD, Liu S, Mackey RH, Matchar DB, McGuire DK, Mohler ER 3rd, Moy CS, Muntner P, Mussolino ME, Nasir K, Neumar RW, Nichol G, Palaniappan L, Pandey DK, Reeves MJ, Rodriguez CJ, Sorlie PD, Stein J, Towfighi A, Turan TN, Virani SS, Willey JZ, Woo D, Yeh RW, Turner MB; American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke statistics–2015 update: a report from the American Heart Association. *Circulation*. 2015;131:e29–322. doi: 10.1161/CIR.0000000000000152
7. Fihn SD, Gardin JM, Abrams J, Berra K, Blankenship JC, Dallas AP, Douglas PS, Foody JM, Gerber TC, Hinderliter AL, King SB 3rd, Kligfield PD, Krumholz HM, Kwong RY, Lim MJ, Linderbaum JA, Mack MJ, Munger MA, Prager RL, Sabik JF, Shaw LJ, Sikkema JD, Smith CR Jr, Smith SC Jr, Spertus JA, Williams SV, Anderson JL; American College of Cardiology Foundation/American Heart Association Task Force. 2012 ACCF/AHA/ACP/AATS/PCNA/SCAI/STS guideline for the diagnosis and management of patients with stable ischemic heart disease: a report of the American College of Cardiology Foundation/American Heart Association task force on practice guidelines, and the American College of Physicians, American Association for Thoracic Surgery, Preventive Cardiovascular Nurses Association, society for cardiovascular angiography and interventions, and society of thoracic surgeons. *Circulation*. 2012;126:e354–e471. doi: 10.1161/CIR.0b013e318277d6a0
8. Fihn SD, Blankenship JC, Alexander KP, Bittl JA, Byrne JG, Fletcher BJ, Fonarow GC, Lange RA, Levine GN, Maddox TM, Naidu SS, Ohman EM, Smith PK. 2014 ACC/AHA/AATS/PCNA/SCAI/STS focused update of the guideline for the diagnosis and management of patients with stable ischemic heart disease: a report of the American College of Cardiology/American Heart Association task force on practice guidelines, and the American Association for thoracic surgery, preventive cardiovascular nurses association, society for cardiovascular angiography and interventions, and society of thoracic surgeons. *Circulation*. 2014;130:1749–1767. doi: 10.1161/CIR.0000000000000095
9. Neumann FJ, Sousa-Uva M, Ahlsson A, Alfonso F, Banning AP, Benedetto U, Byrne RA, Collet JP, Falk V, Head SJ, Jüni P, Kastrati A, Koller A, Kristensen SD, Niebauer J, Richter DJ, Seferović PM, Sibbing D, Stefanini GG, Windecker S, Yadav R, Zembala MO. 2018 ESC/EACTS guidelines on myocardial revascularization. *Eur J Cardiothorac Surg*. 2019;55:4–90. doi: 10.1093/ejcts/ezy289
10. Mazzinari G, Ball L, Serpa Neto A, Errando CL, Dondorp AM, Bos LD, Gama de Abreu M, Pelosi P, Schultz MJ. The fragility of statistically significant findings in randomised controlled anaesthesiology trials: systematic review of the medical literature. *Br J Anaesth*. 2018;120:935–941. doi: 10.1016/j.bja.2018.01.012
11. Docherty KF, Campbell RT, Jhund PS, Petrie MC, McMurray JJV. How robust are clinical trials in heart failure? *Eur Heart J*. 2017;38:338–345. doi: 10.1093/eurheartj/ehw427
12. Tignanelli CJ, Napolitano LM. The fragility index-P values reimagined, flaws and all-reply. *JAMA Surg*. 2019;154:674–675. doi: 10.1001/jamasurg.2019.0568
13. Akl EA, Briel M, You JJ, Sun X, Johnston BC, Busse JW, Mulla S, Lamontagne F, Bassler D, Vera C, Alshurafa M, Katsios CM, Zhou Q, Cukierman-Yaffe T, Gangji A, Mills EJ, Walter SD, Cook DJ, Schünemann HJ, Altman DG, Guyatt GH. Potential impact on estimated treatment effects of information lost to follow-up in randomised controlled trials (LOST-IT): systematic review. *BMJ*. 2012;344:e2809. doi: 10.1136/bmj.e2809
14. Edwards E, Wayant C, Besas J, Chronister J, Vassar M. How fragile are clinical trial outcomes that support the CHEST clinical practice guidelines for VTE? *Chest*. 2018;154:512–520. doi: 10.1016/j.chest.2018.01.031
15. Chase Kruse B, Matt Vassar B. Unbreakable? An analysis of the fragility of randomized trials that support diabetes treatment guidelines. *Diabetes Res Clin Pract*. 2017;134:91–105. doi: 10.1016/j.diabres.2017.10.007
16. Khan M, Evaniew N, Gichuru M, Habib A, Ayeni OR, Bedi A, Walsh M, Devereaux PJ, Bhandari M. The fragility of statistically significant findings from randomized trials in sports surgery: a systematic survey. *Am J Sports Med*. 2017;45:2164–2170. doi: 10.1177/0363546516674469
17. Evaniew N, Files C, Smith C, Bhandari M, Ghert M, Walsh M, Devereaux PJ, Guyatt G. The fragility of statistically significant findings from randomized trials in spine surgery: a systematic survey. *Spine J*. 2015;15:2188–2197. doi: 10.1016/j.spinee.2015.06.004
18. Ridgeon EE, Young PJ, Bellomo R, Mucchetti M, Lembo R, Landoni G. The fragility index in multicenter randomized controlled critical care trials. *Crit Care Med*. 2016;44:1278–1284. doi: 10.1097/CCM.0000000000001670
19. Shochet LR, Kerr PG, Polkinghorne KR. The fragility of significant results underscores the need of larger randomized controlled trials in nephrology. *Kidney Int*. 2017;92:1469–1475. doi: 10.1016/j.kint.2017.05.011
20. Grolleau F, Collins GS, Smarandache A, Pirracchio R, Gakuba C, Boutron I, Busse JW, Devereaux PJ, Le Manach Y. The fragility and reliability of conclusions of anesthesia and critical care randomized trials with statistically significant findings: a systematic review. *Crit Care Med*. 2019;47:456–462. doi: 10.1097/CCM.0000000000003527