Please review the Supplemental Files folder to review documents not compiled in the PDF.

Understanding unexpected results from randomized clini□cal trials

Journal:	*NEJM Evidence*
Manuscript ID	Draft
Article Type:	Original Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Brophy, James; McGill University Faculty of Medicine and Health Sciences,
Abstract:	Objective: The recent DECAF randomized trial reported a statistically significant (p <0.01) reduction in the recurrence of atrial fibrillation with continued consumption of caffeinated coffee compared to abstinence from coffee and caffeinated products. As coffee has historically be viewed as proarrhythmic, this result may be seen as "unexpected". Evidence Review: Using digitally extracted individual data from this publication, frequentist and Bayesian analyses were performed. Findings: Frequentist analyses showed the trial was underpowered to detect realistic differences and consequently any statistically significant findings are likely subject to a two fold M (magnitude) error. Bayesian analyses, using priors from the DECAF power calulations, show posterior probabilities for a clinically meaningful differences to be inconclusive with the probability of a hazard ratio < 0.9 equal to 88% and the probablity of a risk difference > 2% equal to 82%. Conclusions and Relevance: Supplemental frequentist and Bayesian analyses can test the robustness of "unexpected" results and provide additional insights and nuances into their interpretation. For DECAF which had very positive, but surprising, results in favor of caffeine, this approach shows the probability of a clinically meaningful benefit is inconclusive.

Understanding unexpected results from randomized clini□cal trials

Does coffee reduce atrial fibrillation recurrences?

James M Brophy

2025-11-25

Abstract

Abstract:

Objective: The recent DECAF randomized trial reported a statistically significant (p <0.01) reduction in the recurrence of atrial fibrillation with continued consumption of caffeinated coffee compared to abstinence from coffee and caffeinated products. As coffee has historically be viewed as proarrhythmic, this result may be seen as “unexpected”.

Evidence Review: Using digitally extracted individual data from this publication, frequentist and Bayesian analyses were performed.

Findings: Frequentist analyses showed the trial was underpowered to detect realistic differences and consequently any statistically significant findings are likely subject to a two fold M (magnitude) error. Bayesian analyses, using priors from the DECAF power calulations, show posterior probabilities for a clinically meaningful differences to be inconclusive with the probability of a hazard ratio < 0.9 equal to 88% and the probablity of a risk difference > 2% equal to 82%.

Conclusions and Relevance: Supplemental frequentist and Bayesian analyses can test the robustness of “unexpected” results and provide additional insights and nuances into their interpretation. For DECAF which had very positive, but surprising, results in favor of caffeine, this approach shows the probability of a clinically meaningful benefit is inconclusive.

Backgound

A recent paper, "Caffeinated Coffee Consumption or Abstinence to Reduce Atrial Fibrillation The DECAF Randomized Clinical Trial"(1) published online in JAMA on Nov 9 2025 concluded

> "In this clinical trial of coffee drinkers after successful cardioversion, allocation to consumption of caffeinated coffee averaging 1 cup a day was associated with less recurrence of AF or atrial flutter compared with abstinence from coffee and caffeinated products."

This conclusion is likely surprising to most physicians as the historical belief is that caffeine is associated with increased cardiac ectopy. However the authors make the case that while

> "Caffeinated coffee has traditionally been considered proarrhythmic"

its role in atrial fibrillation is less uncertain. Supporting this uncertainty, the authors refer to a previous randomized trial(2) that was not of atrial fibrillation patients but rather of 100 ambulatory patients studies in a crossover design. That study concluded

> " the consumption of caffeinated coffee did not result in significantly more daily premature atrial contractions than the avoidance of caffeine"

This small trial(2) examined only short-term surrogate endpoints, not atrial fibrillation outcomes, and actually found a non significant trend towards more atrial ectopy in the caffeine group.

> "The consumption of caffeinated coffee was associated with 58 daily premature atrial contractions as compared with 53 daily events on days when caffeine was avoided (rate ratio, 1.09; 95% confidence interval [CI], 0.98 to 1.20; P=0.10). The consumption of caffeinated coffee as compared with no caffeine consumption was associated with 154 and 102 daily premature ventricular contractions, respectively (rate ratio, 1.51; 95% CI, 1.18 to 1.94)."(2)

Apart from assuming that atrial ectopy is a good surrogate for atrial fibrillation, this suggests confusion between "absence of evidence" with "evidence of absence".

The DECAF authors(1) also cite a meta-analysis of 12 observational studies(3) that reported no conclusive evidence for or against an association between caffeine intake and incident atrial fibrillation (OR 0.95 (0.84–1.06). The quality of these 12 studies was not reported and as the DECAF authors state

> "observational studies are prone to confounding,and whether these findings are biased by systematic differences between coffee and noncoffee drinkers is unclear"(1)

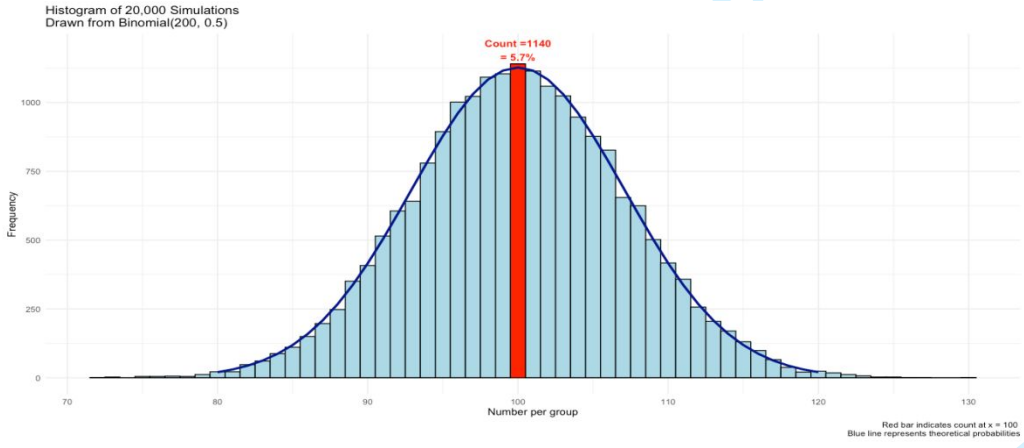
Given this uncertainty, DECAF(1) was designed to investigate the effect of caffeinated coffee consumption versus abstinence on atrial fibrillation recurrence following successful cardioversion.

DECAF Trial Design

DECAF(1) was a randomized controlled trial of 200 regular coffee drinkers with a recent successful cardioversion for atrial fibrillation (AF). Patients were randomized to either continue their usual caffeinated coffee consumption (median 1 cup/day) or to abstain from all caffeine products (decaf group). The primary outcome was recurrence of atrial fibrillation or atrial flutter within 6 months. The sample size calculations were based on assuming

> "… a 50% incidence of AF recurrence within 6 months following cardioversion. A clinically relevant effect size was assumed to approximate the effectiveness of commonly prescribed antiarrhythmic drugs for recurrentAFafter cardioversion. To provide 80% power to detect a minimum 41% reduced relative hazard of AF, we enrolled 200 patients (100 per group) assuming a 1:1 randomization scheme, potential 10% loss to follow-up, and .05 2-tailed α level."(1)

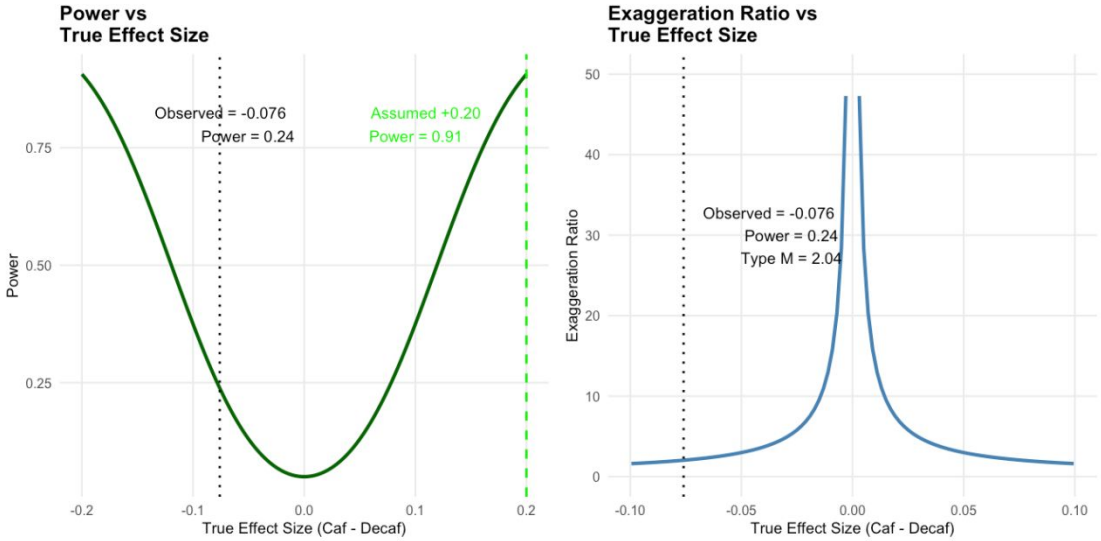
This description is problematic for the following reasons.
1. The implication is, since the methods section is written before any results are available, that 1:1 randomization of 200 subjects will automatically produce two equal groups. Based on binomial probability theory, there is only a 5.7% chance of achieving a perfectly balanced split (see Figure below).



Distribution of group sizes for n=200 randomized 1:1 into two groups. The probability of a perfect 100/100 split is only 5.7%.

2. Given the prevailing prior uncertainty, designing a trial to detect a very large risk reduction is overly optimistic. The 80% power estimate assumed a 41% reduction applied to a 50% baseline rates, meaning the intervention is expected to reduce recurrence by an absolute 20.5%. What would be the power with this sample size of 200 to detect a more modest relative risk reduction, say 15% (i.e. from 50% to 42.5%)? The answer is 24% (see Figure below, left panel), illustrating the study

was severely underpowered to detect more realistic effect sizes and therefore at risk of Type M (magnitude) error(4), meaning any statistically significant result tends to exaggerate the true effect(4). Thus the observed −17% risk difference may be overstating the true benefit, possibly by a factor of two (see right panel of above Figure). The published result should therefore be interpreted cautiously. Further exploration of the probability of a clinically meaningful result awaits the Bayesian analysis provided in later sections.



Power curves for DECAF trial (n=200) to detect various relative risk reductions from a 50% baseline recurrence rate. The design has only 24% power to detect a 15% relative risk reduction(RRR). If the true effect is a 15% RRR, it is likely that any statistically significant result from this underpowered study will overstate the true effect by about two-fold.

3. A lack of specification as to which arm (abstinence or caffeinated) is hypothesized to have the reduced risk. This ambiguity creates interpretive flexibility as whichever arm shows benefit could be claimed as success. Such directional absence hinders inference and inflates the risk of misleading conclusions. This lack of statistical clarity is evident not only in the published article but also in the trial protocol and trial registration documentation at ClinicalTrials.gov. It is disconcerting that these design issues were overlooked not only by the authors but also apparently by peer reviewers, journal editors, and national funding agencies. These design flaws contribute to uncertainty in interpreting the final results.

DECAF Trial Results

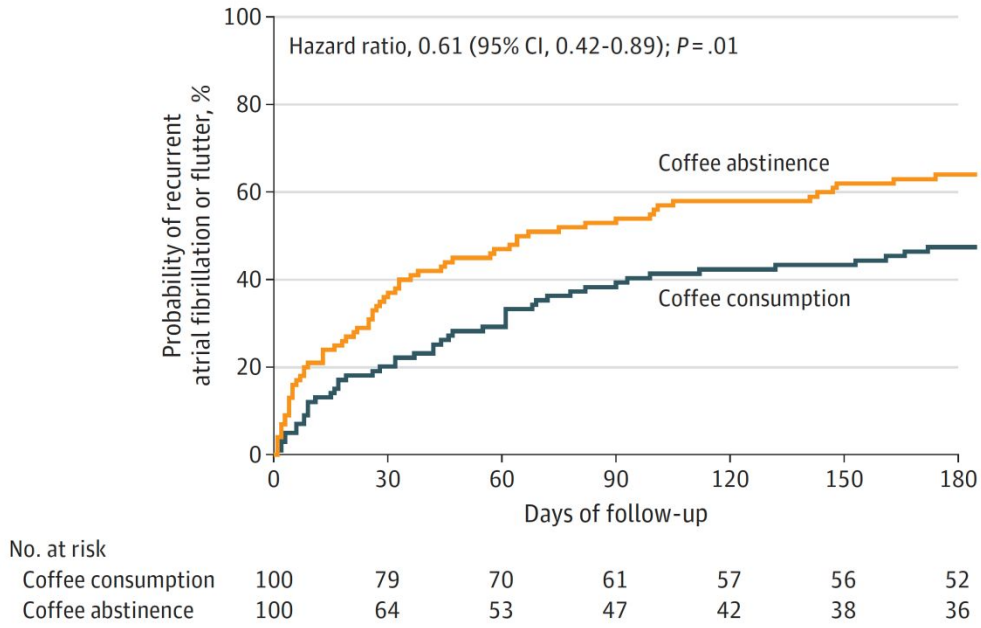
In the caffeinated coffee group, 47% of patients had an AF or atrial flutter recurrence compared to 64% recurrence in the decaf group (hazard ratio, 0.61 [95%CI, 0.42-0.89],). The authors concluded

> "…allocation to consumption of caffeinated coffee averaging 1 cup a day was **associated** with less recurrence of AF or atrial flutter compared with abstinence from coffee and caffeinated products."(1)

Given the randomized design and the statistically significant result (p = 0.01), it is surprising that the conclusion wasn't

> …allocation to consumption of caffeinated coffee averaging 1 cup a day **caused** less recurrence of AF or atrial flutter compared with abstinence from coffee and caffeinated products.

The results were summarized in the Figure below.



Cumulative incidence curve from DECAF trial(1) showing higher recurrence rates in the decaf group (blue) versus caffeinated group (orange).

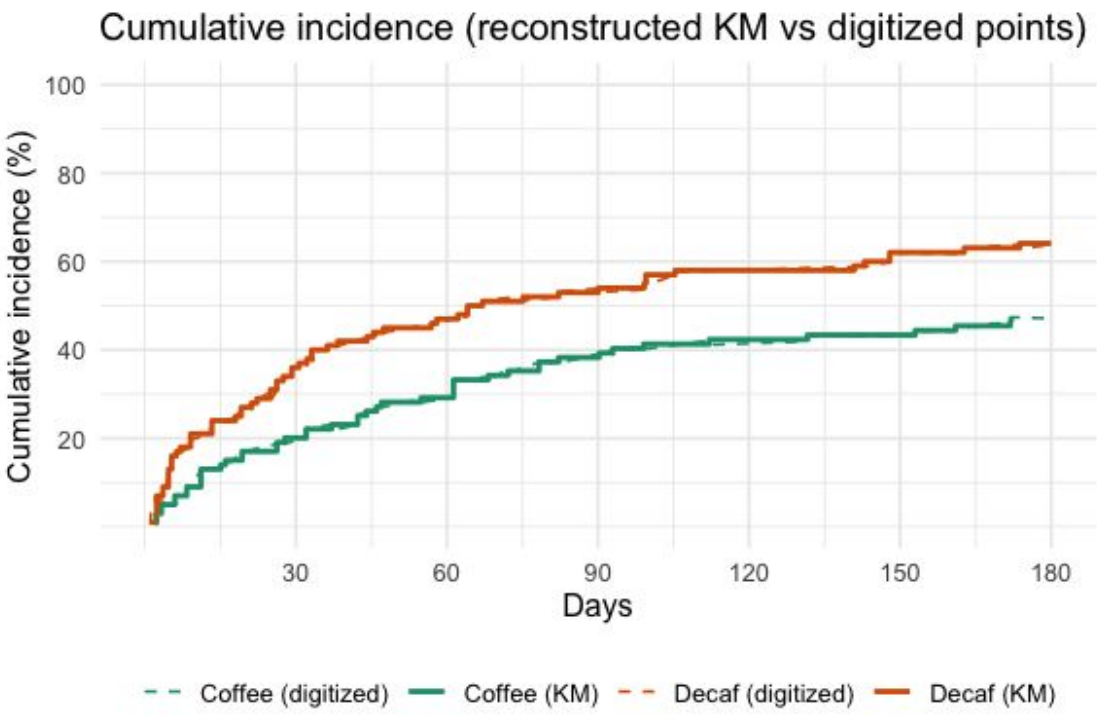
Bayesian reanalysis of DECAF (survival modeling)

The small p-value (p(data|null hypothesis) informs us that observing this data or more extreme data, if there was no true effect is unlikely. The inverse probability (p(effect|data) is arguably of greater interest and provides additional insights, most importantly the ability to include prior knowledge and to provide probability estimates of benefit or harm. Inverse probability, or Bayesian analyses, of survival data requires individual patient data (IPD) and well justified priors, which summarize the state of knowledge before seeing the data. Priors are often deemed the Achilles' heel of Bayesian analyses, but well justified, transparent and robust testing of different priors can mitigate subjectiveness concerns. In Bayesian reanalyses of completed studies, prior beliefs can sometimes be extracted from the original power calculations. For example, the DECAF(1) authors assumed a baseline recurrence risk of 50% with a 41% relative risk reduction and therefore a 29.5% risk in the intervention group (0.50 -

0.50*0.41 = 0.295). However we are now faced with the previous identification problem of the direction in the projected improvement. Did DECAF(1) authors believe that the risk reduction would be with decaffeinated or caffeinated exposure? Given the prevailing belief of caffeine as a proarrhythmic agent and given that the study population was comprised initially of coffee drinkers, with the intervention being abstinence we assume the proposed relative risk reduction is in favor of the decaffeinated group.

The priors for both treatment effect and baseline risk can be specified as normal distributions on the log-odds scale. Based on the assumptions used by the DECAF invetigators, the distributional parameters for the baseline control prior will be mean centered at their belief of a 50% recurrence risk (0 on the logit scale) with a standard deviation of 1.5, indicating a weakly informative prior. Similarly, the prior for the treatment effect will be centered at 29.5% (-0.871 on the logit scale) with an assumed standard deviation of 0.5, reflecting slightly higher confidence in our ability to estimate effect size variability. The reasonableness of these priors can be assessed through prior predictive checks which demonstrates that these priors adequately reflect the DECAF position of considerable uncertainty with a wide range of possible effect sizes, principally with expected benefit (93.5% probability), but a small probability of harm, for the intervention versus control.

In addition to the priors, the Bayesian survival analysis requires the individual patient data (IPD). Cumulative incidence data available in graphical format (DECAF(1) Figure3), were extracted with WebPlotDigitizer(5) and transformed into survival format to generate a Kaplan–Meier (KM) plot using the Guyot et al(6) algorithm that incorporates numbers at risk and event counts. This method was implemented via the IPDfromK R package(7) providing a robust and validated approach for reconstructing IPD from published survival curves for secondary analyses. The success in extracting the IPD is evident in the complete overlap between the reconstructed KM curves and the original published curves (see Figure below).

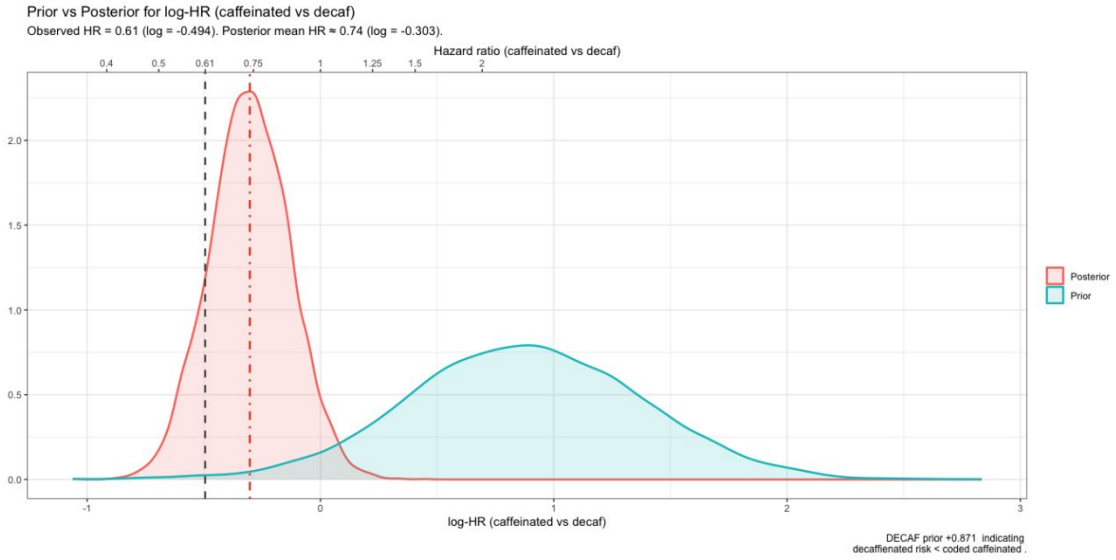


Cumulative incidence plots from digitalized DECAF curves and reconstructed KM plots. Note excellent comparison with Figure 3.

The IPD calculated frequentist survival analysis results, HR = 0.62 (95% CI 0.43–0.91),; p = 0.014 closely matches the published DECAF(1) results, confirming the accuracy of the IPD reconstruction process.

Once the priors and IPD are specified, we can proceed with the Bayesian survival regression modelling. As Cox regression uses the partial likelihood, which cancels out the baseline hazard when estimating the regression coefficients (i.e., log-HRs), in a Bayesian Cox survival model we only need a prior for the the treatment effect.

The Bayesian posterior hazard ratio is 0.74 (95% CrI 0.53–1.04), suggesting that caffeinated coffee consumption may be associated with a lower risk of AF recurrence compared to decaffeinated coffee but the strength of the conclusion has been tempered by the prior belief that decaffeinated and not caffeinated, coffee would be beneficial. Graphically this shift of the observed result by the prior belief is shown in the Figure below. Notice also the width of the posterior probability has narrowed, underscoring that our new data has decreased our uncertainty about the effect of coffee.

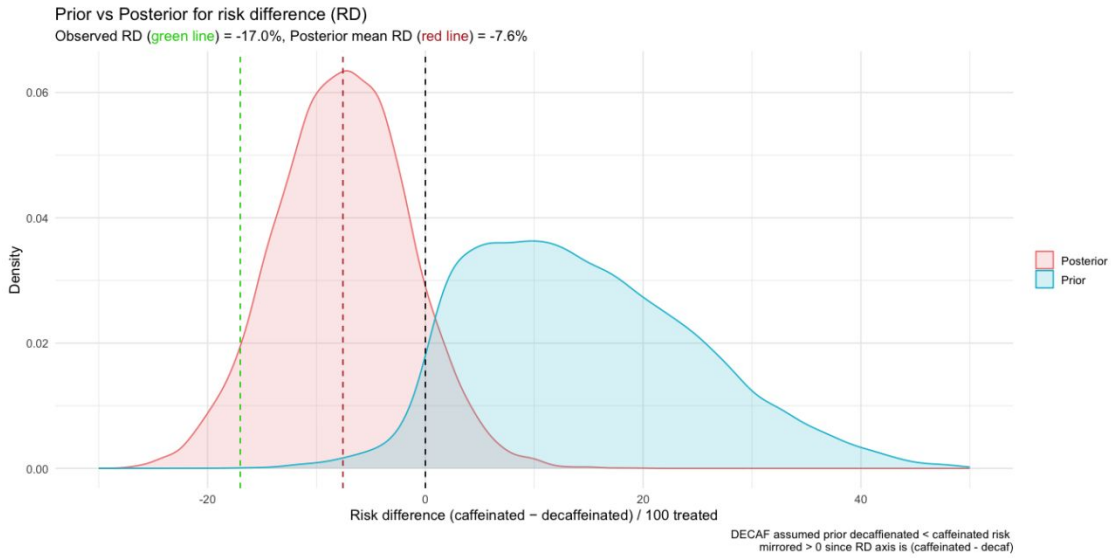


Prior and posterior probability density functions of log HR

Despite this shift, the posterior probability remains high that caffeinated coffee has a beneficial effect (HR < 1) is 96.0. However, the probability that caffeinated coffee has a clinically meaningful benefit, for example a HR < 0.9, falls to 87.9%. Increasing the threshold of what represents a clinically meaningful decrease, will lead to a corresponding reduction in the probability of a beneficial caffeine effect.

Bayesian reanalysis of DECAF (risk difference)

The DECAF(1) result can also be expressed as a risk difference, -17% [95% CI; -.03, -.31, p = 0.01]. Again a Bayesian risk difference analysis may facilitate clinical interpretations. The Bayesian model will be fit to the aggregated data of events and total patients in each group. The model formula specifies a binomial likelihood with a logit link function, appropriate for binary outcome data. The previously developped priors based on the DECAF power calculation can be transformed to the risk difference scale. The Bayesian posterior risk difference, based on these prior beliefs and the observed data is 7.6% (95% CrI: -19.5% to 4.4%), in favor of the caffeinated group. This posterior risk difference showed a shift towards the null compared to the frequentist result, due to the prior's influence even thought it was left purposely vague due to reflect the lack of previous consensus concerning the intervention's value in this population (see Figure below). The prior favored a posterior probability, (89.3%), for decaffeinated benefit but did allocate a smaller probability (10.7%) for a possible caffeinated benefit. While the final result suggests a caffeinated coffee benefit, the strength of the evidence in favor of this benefit is seen as modest at best, and certainly much less than implied by the original analysis. For example, the probability of a clinically meaningful benefit (e.g., say arbitrarily RD < -0.02 or number needed to treat = 50) drops to 82.0%.



Prior and posterior probability density functions of risk difference

Discussion

DECAF(1) was a well performed clinical trial with successful randomization and no lost to follow-up. The trial was unblinded but events were adjudicated by treating physicians and not study coordinators so this risk of bias appears low. No other obvious biases were identified yet the results were highly "unexpected" or "surprising". The published analysis apparently provides a definitive answer in favor of the caffeinated group (p = .01). Notwithstanding DECAF's(1) positive attributes and the emergence of "dataism" - the belief that empirical data should guide decisions as characterized by the pithy aphorism "In God we trust. All others have to bring data.", many clinicians will prefer to trust their intuition or "gut instinct" and not encourage coffee consumption or maintence to prevent episodes of recurrent atrial fibrillation. However, reliance on intuition over evidence remains prone to cognitive error(8).

Given the increasing influence of data-driven decision-making, it is essential that statistical interpretations be both rigorous and nuanced, whether frequentist or Bayesian in nature. From a frequentist standpoint, DECAF's(1) sample size was insufficient to detect realistic effect sizes. Power calculations revealed that the trial had only ~24% power to detect a 15% relative risk reduction, and any statistically significant result from such an underpowered study is likely to exaggerate the true effect — potentially by a factor of two.

The Bayesian framework adds further depth by incorporating prior beliefs and quantifies uncertainty in a way that frequentist methods cannot. It enables clinicians to assess not only the most recently observed data but also how to contextualize it by incorporating existing knowledge and beliefs. This is particularly important in scenarios where prior evidence or expert opinion may conflict with new findings. Using priors derived from the DECAF authors' own assumptions(1), the Bayesian survival and risk difference models suggest that the strength of evidence favoring caffeinated coffee is more modest than the frequentist interpretation implies. For example, while the posterior probability that

caffeinated coffee reduces AF recurrence is high (~96%), the probability of a clinically meaningful benefit (e.g., HR < 0.9) drops to ~88%. On the risk difference scale, the posterior probability of at least absolute 2% reduction in recurrences by avoiding abstinence is 82.0%. This distinction between statistical and clinical significance is crucial for informed decision-making.

Some may dismiss this reanalysis as statistical alchemy or the “haze of Bayes”(9), but it offers a principled approach to interpreting unexpected findings. In the case of DECAF(1), it tempers over-interpretation by including past beliefs, distinguishes between statistical and clinical significance, and provides the statistical justification for a replication study. This reanalysis can help clinicians make more informed decisions by providing a clearer understanding of the evidence at hand. This is especially important given that over 120 media outlets reported the findings within a week of publication, potentially influencing clinical practice and public behavior based on a single, underpowered study with unexpected, and naively interpreted, results.

Supplementary Material

All code for the calculations and figures presented in this article are provided in the supplementary file

References

1. Wong CX, Cheung CC, Montenegro G, Oo HH, Pena IJ, Tang JJ, et al. Caffeinated coffee consumption or abstinence to reduce atrial fibrillation: The DECAF randomized clinical trial. JAMA [Internet]. 2025; Available from: https://www.ncbi.nlm.nih.gov/pubmed/41206802

2. Marcus GM, Rosenthal DG, Nah G, Vittinghoff E, Fang C, Ogomori K, et al. Acute effects of coffee consumption on health among ambulatory adults. N Engl J Med [Internet]. 2023;388(12):1092–100. Available from: https://www.ncbi.nlm.nih.gov/pubmed/36947466

3. Krittanawong C, Tunhasiriwet A, Wang Z, Farrell AM, Chirapongsathorn S, Zhang H, et al. Is caffeine or coffee consumption a risk for new-onset atrial fibrillation? A systematic review and meta-analysis. Eur J Prev Cardiol [Internet]. 2021;28(12):e13–5. Available from: https://www.ncbi.nlm.nih.gov/pubmed/34647581

4. Gelman A, Carlin J. Beyond power calculations: Assessing type s (sign) and type m (magnitude) errors. Perspect Psychol Sci [Internet]. 2014;9(6):641–51. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26186114

5. Rohatgi A. WebPlotDigitizer [computer software] [Internet]. 2025. Available from: https://automeris.io/WebPlotDigitizer cite turn16search49

6. Guyot P, Ades AE, Ouwens MJ, Welton NJ. Enhanced secondary analysis of survival data: Reconstructing the data from published kaplan-meier survival curves. BMC Med Res Methodol [Internet]. 2012;12:9. Available from: https://www.ncbi.nlm.nih.gov/pubmed/22297116

7. Liu N, Lee JJ. IPDfromKM: Map digitized survival curves back to individual patient data [Internet]. 2020. Available from: https://CRAN.R-project.org/package=IPDfromKM

8. Kahneman D. Thinking, fast and slow. 1st ed. New York: Farrar, Straus; Giroux; 2011.

9. Feinstein AR. Clinical biostatistics. XXXIX. The haze of bayes, the aerial palaces of decision analysis, and the computerized ouija board. Clin Pharmacol Ther [Internet]. 1977;21(4):482–96. Available from: https://www.ncbi.nlm.nih.gov/pubmed/403045