EVIDENCE-BASED ORTHOPAEDICS

Risk or Uncertainty?

The Anatomy & Physiology of Evidence-Based Orthopaedics

Why the number on the page is never the outcome in the patient.

John Walsh, DO
Prepared for orthopaedic surgeons, patients, and the intellectually curious

THE CENTRAL CLAIM

A trial reports what happened to the average patient. It cannot report what will happen to yours.

01

Anatomy

The static structure of evidence — study designs, the evidence hierarchy, the literature's architecture.

02

Physiology

How EBM actually functions in a real encounter — appraisal, biostatistics, judgment, and the patient in front of you.

03

The gap

Between the two sits risk literacy — and a distinction most training never makes: risk versus uncertainty.

WHAT EBM ACTUALLY IS

Evidence-based medicine was never meant to be evidence alone

> *"Evidence based medicine is the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients."*
>
> — Sackett, Rosenberg, Gray, Haynes & Richardson, BMJ 1996

1

Best research evidence

Systematic reviews, RCTs, cohort and case-control studies — the published literature.

2

Clinical expertise

The surgeon's accumulated skill, pattern recognition, and judgment.

3

Patient values & circumstances

What matters to this specific person — goals, risk tolerance, comorbidity, life context.

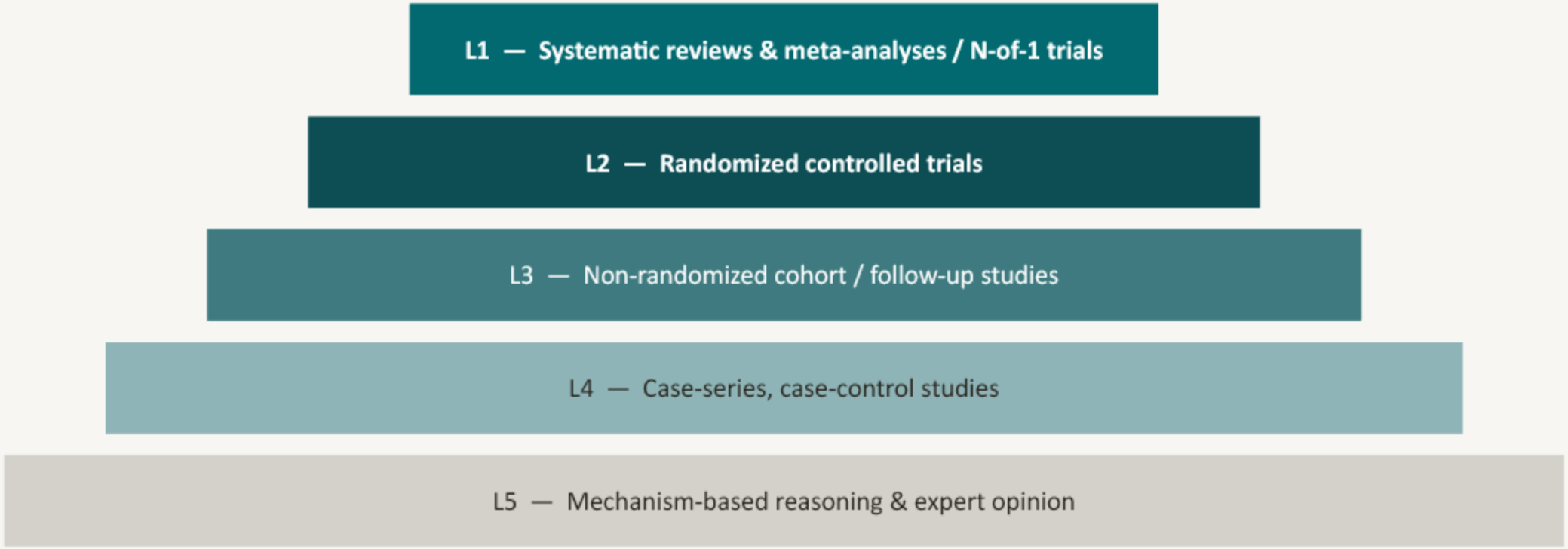
EBM was designed as an integration of all three. In practice, "evidence" often crowds out the other two — which is where this talk begins.

Source: Sackett et al., BMJ 1996

ANATOMY — THE STRUCTURE

The evidence hierarchy is EBM's skeleton

What evidence IS, and how it is classified



Higher = more protection against bias and chance. Lower = faster, cheaper, more available — and far more common in orthopaedic practice than the pyramid suggests.

Source: OCEBM Levels of Evidence, Oxford CEBM

PHYSIOLOGY — THE FUNCTION

The evidence hierarchy is inert until someone works it

1

Ask (PICO)

Population · Intervention · Comparison · Outcome — turn a vague clinical question into an answerable one.

→

2

Acquire

Search the literature efficiently for the highest available level of evidence.

→

3

Appraise

Critically evaluate validity, size of effect, and precision — the biostatistics layer (next section).

→

4

Apply

Integrate the appraisal with clinical judgment and this patient's values — not the trial population's.

This is EBM's physiology: a repeating loop of asking, finding, appraising, and applying — not a one-time lookup.

Source: PICO framework review, Saudi Journal of Anaesthesia (PMC)

The p-value trap

What a p-value actually is

"The probability under a specified statistical model that a statistical summary of the data would be equal to or more extreme than its observed value."

— American Statistical Association, 2016

What it is not

- Not the probability the treatment works
- Not the probability the null hypothesis is true
- Not a measure of effect size or clinical importance

The ASA's own warning

"The p-value is commonly misused and misinterpreted" — to the point that some journals now discourage its use and some statisticians recommend abandoning it outright.

If a "positive" trial can be undone by a handful of patients changing outcome status, was the p-value ever telling you what you thought it was?

Source: Wasserstein & Lazar, The American Statistician, 2016

"Statistically significant" and "matters to the patient" are different claims

Statistical significance

A p-value below an arbitrary threshold (usually 0.05). Tells you the observed difference was unlikely under the null hypothesis — nothing about its size or relevance.

Clinical significance (MCID)

The Minimal Clinically Important Difference: the smallest change in outcome score that a patient would actually perceive as beneficial — enough to justify the risk, cost, or recovery of an intervention.

A trial can report p = 0.001 for an outcome-score improvement well below the MCID. Statistically real. Clinically invisible to the patient.

Source: Slobogean, OTA, Bloom et al., Am J Sports Med 2023

How solid is "significant"? Ask the Fragility Index

The Fragility Index (FI) asks a simple question: how many patients would have to flip from a non-event to an event before a "statistically significant" result stops being significant?

Median FI across 399 high-impact RCTs: **8**

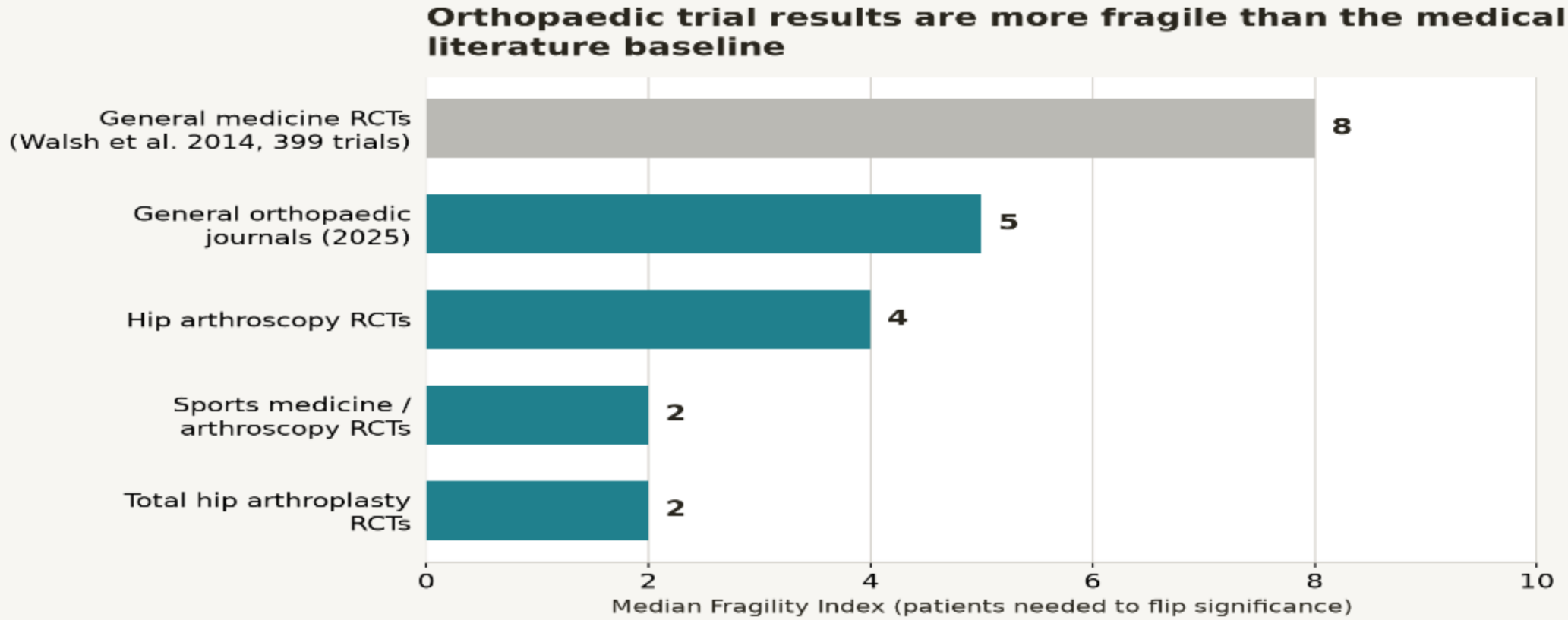
25% of trials had an FI of 3 or less. In 53% of trials, the FI was smaller than the number of patients lost to follow-up — meaning dropout alone could plausibly have erased the "significant" finding.

The method

- Add one event to the smaller-event group, subtract one non-event to keep totals constant
- Recalculate Fisher's exact test
- Repeat until $p \geq 0.05$ — the number of steps taken is the Fragility Index

Source: Walsh et al., J Clin Epidemiol 2014

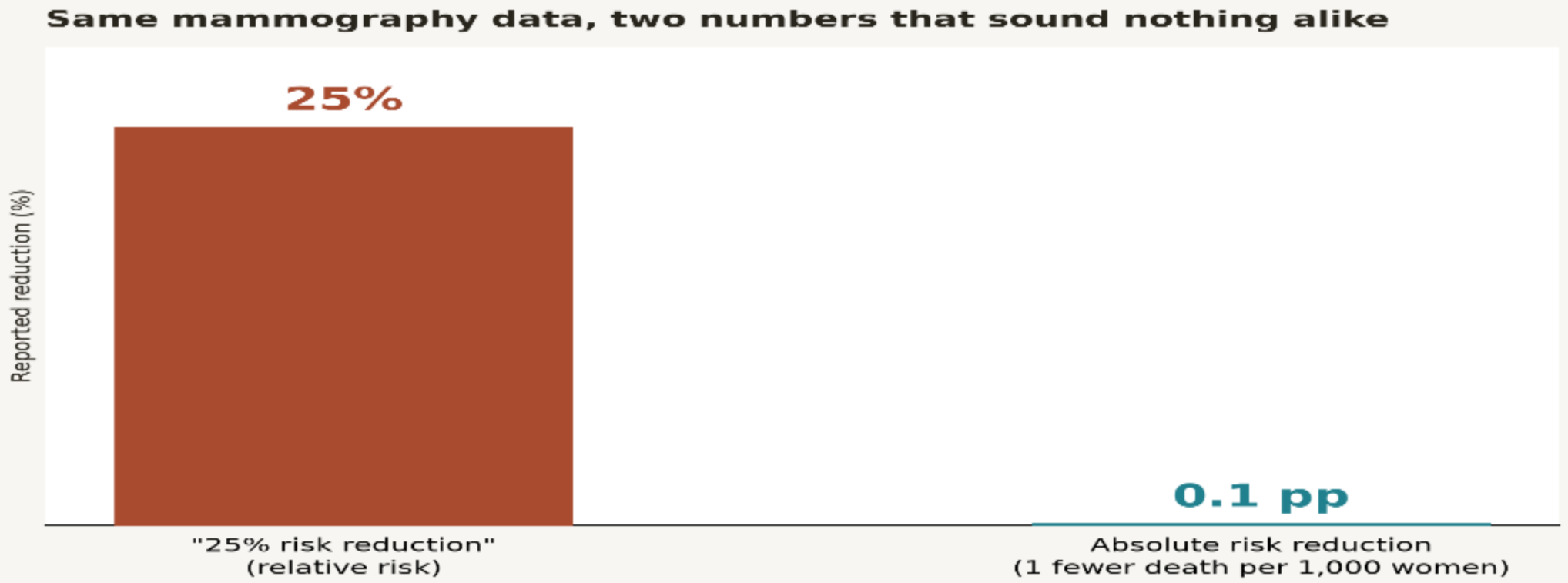
Orthopaedic trials run even more fragile than the medical literature baseline



Fragility Index = number of patients who would need to switch outcome status to flip a trial's result from statistically significant to non-significant. Lower = more fragile. Orthopaedic subspecialty medians (teal) sit below the general-medicine benchmark (grey).

Source: Khan 2017; Maldonado 2021; Go 2022; Poursalehian 2025

What you should know about risk (but usually aren't told)



Mammography screening is often reported as cutting breast-cancer mortality risk by 25% (relative risk reduction). The absolute risk reduction behind that headline: roughly 1 fewer death per 1,000 women screened.

Source: Gigerenzer et al., Psychological Science in the Public Interest, 2007

The fix: natural frequencies, not percentages

Gigerenzer's central recommendation — tested across doctors, patients, and journalists alike: replace conditional probabilities and percentages with countable, natural frequencies.

Confusing

"A test is 90% sensitive and 91% specific; the disease has a 1% base rate. What is the probability a positive result is a true positive?"

Even trained physicians routinely get this wrong.

Clear

"Out of 1,000 people, 10 have the disease. Of those 10, 9 test positive. Of the 990 without it, about 89 also test positive."

Same information. Now it is obvious that most positives are false positives.

Source: Gigerenzer et al., Psychological Science in the Public Interest, 2007

RISK VS. UNCERTAINTY

Frank Knight drew a line most of medicine ignores

"Risk is measurable and uncertainty is not... The term 'risk' is used for measurable uncertainty, as contrasted with unmeasurable uncertainty, to which the term 'uncertainty' is applied."

— Frank H. Knight, Risk, Uncertainty, and Profit, 1921

Certainty	→	Risk	→	Uncertainty	→	Ignorance
The outcome is known in advance.		Unknown outcome, known probability distribution from a stable reference class.		Unknown outcome, unknown or unreliable probability — the situation is too unique to support a stable number.		Not even the range of possible outcomes is known.

Most orthopaedic RCT evidence sits at "risk." Most orthopaedic clinical encounters — unique patients, comorbidities, anatomy — sit at "uncertainty." EBM trains surgeons almost exclusively for the first column.

Source: Knight, Risk, Uncertainty, and Profit, 1921 (Econlib)

There is no such thing as an average patient

"Determining the best treatment for an individual (the task of a doctor) is fundamentally different from determining the average effect of treatment in a population (the purpose of a trial)... Person-level treatment effects cannot be identified in randomized clinical trials, and population average treatment effects may not apply to all patients."

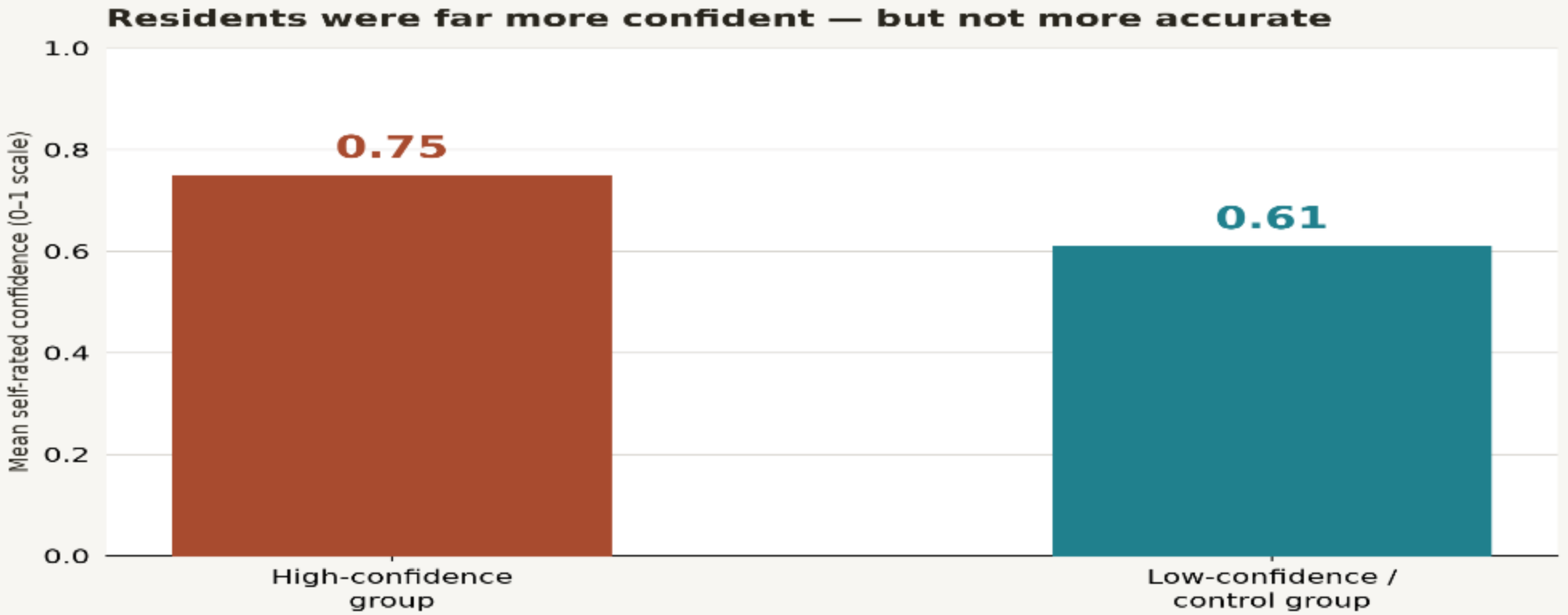
— Dahabreh, Hayward & Kent, International Journal of Epidemiology, 2016

The trial answers a population question. Your patient is asking an individual one.

A mean improvement score, an absolute risk reduction, an NNT — each is a statement about a distribution of outcomes across the trial's enrolled population. None of them is a prediction about which outcome this particular patient, with their particular biology and circumstances, will draw.

Source: Dahabreh, Hayward & Kent, Int J Epidemiol, 2016

Confidence is not a proxy for correctness



In a controlled study of internal medicine residents, the high-confidence group rated itself far more confident than the control group, yet diagnostic accuracy did not differ significantly between the two groups — direct evidence of confidence–accuracy miscalibration.

Source: Overconfidence, Time-on-Task, and Medical Errors, 2024

Uncertainty is not the enemy of good decisions

"Just because absolute certainty is unachievable does not mean that rational and optimal decisions cannot be made... rational decisions, even in the absence of guaranteed absolute certainty, are not only possible but, on average, beneficial."

— Djulbegovic, Ethics of Uncertainty, Patient Education and Counseling, 2021

"We want to reduce uncertainty, but we do not want to eliminate it totally. Only because we do not know what the future holds can we have our hope and choices... uncertainty should not be looked on as the enemy but rather as a friend — two cheers for uncertainty."

— Djulbegovic, BMJ letter, 2004

Source: Djulbegovic, Patient Education and Counseling, 2021, Djulbegovic, BMJ, 2004

PRESSURE-TESTING THE ARGUMENT

Isn't ‘uncertainty’ just an excuse for bad judgment?

THE OBJECTION

“If uncertainty is real and, by definition, unmeasurable — what stops any surgeon from labeling an inconvenient trial result ‘uncertain’ and reverting to gut feeling? Knight’s distinction becomes permission to override evidence, exactly the practice EBM exists to replace.”

The risk: uncertainty-talk as a loophole, not a discipline.

THE ANSWER

Sackett’s 1996 definition never removed judgment from EBM — it names it as one of three required inputs, alongside best evidence and patient values (slide 3). What changes under an uncertainty-aware model is whether that judgment is disciplined or hidden: GRADE ratings and explicit confidence intervals make the degree of uncertainty auditable, not a private excuse.

An unquantified “I just know” isn’t uncertainty-awareness — it’s the overconfidence bias already linked to error, not accuracy (slide 14).

The dividing line isn’t risk vs. uncertainty — it’s whether uncertainty is stated and quantified, or hidden behind confidence. (Case in point: decks that cite a precise “5–15% diagnostic error rate” with no source attribution are doing exactly what this slide warns against.)

Source: Sackett et al., BMJ 1996, Guyatt et al., GRADE, BMJ 2008

An illustrative lens for this talk — not an established clinical finding

There are two averages, not one

Ensemble average

The outcome averaged across many parallel, independent trials or people happening at the same time. This is the expected value — the number a trial reports.

"What happens across the population."

Time average

The growth rate or outcome experienced by one entity across successive periods of its own life. This is what actually happens to a single person, one event at a time.

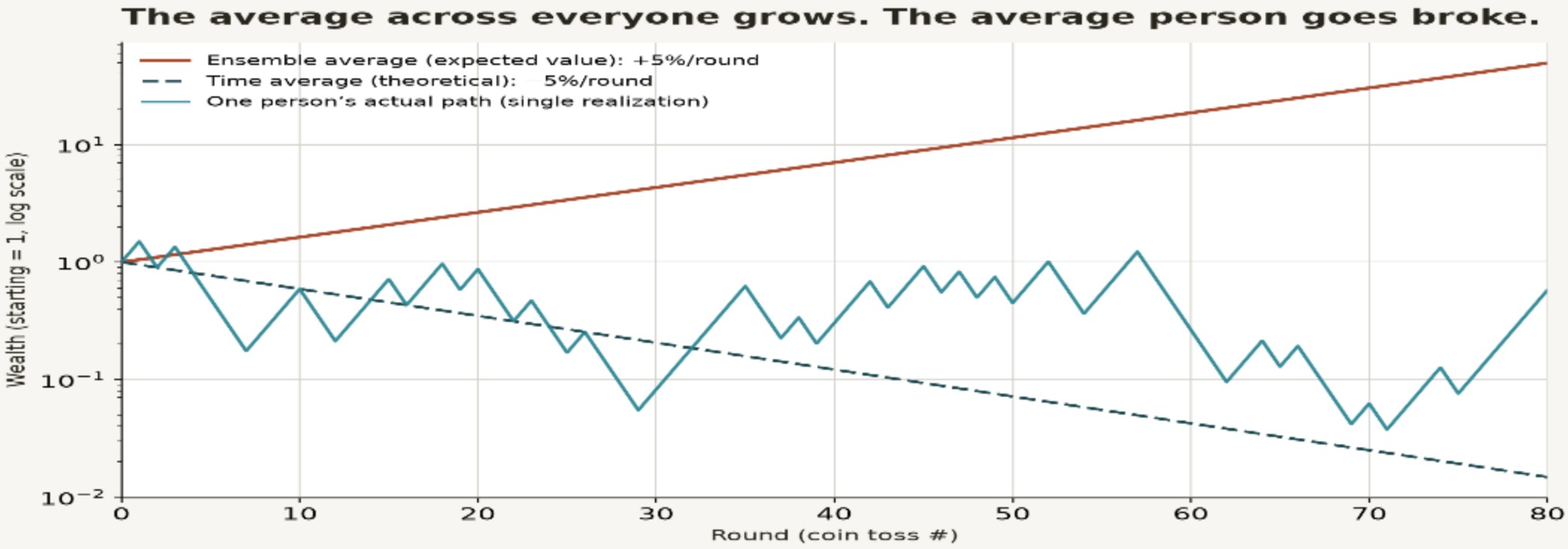
"What happens to the one patient in front of you."

"The ergodic hypothesis... underlies the assumption that the time average and the expectation value of an observable are the same." Peters shows that for compounding, growth-based processes, this assumption is usually false.

Source: Peters, The ergodicity problem in economics, Nature Physics, 2019

An illustrative lens for this talk — not an established clinical finding

The coin toss that bankrupts almost everyone who plays it



Illustrative simulation: a fair coin toss where heads gains 50% of current wealth and tails loses 40%. Expected value grows 5% per round — but almost every individual path decays toward zero because wealth compounds multiplicatively, not additively.

Source: Peters, Nature Physics, 2019 (conceptual diagram redrawn for this presentation)

PRESSURE-TESTING THE ANALOGY

Second objection: isn't "antifragility" just a finance metaphor grafted onto a hip?

THE OBJECTION

Ergodicity, antifragility, and the barbell strategy were built for financial portfolios and systemic market risk — Taleb's response to 2008-style crises. Applying that vocabulary to a single patient's joint replacement risks dressing up an analogy as clinical evidence.

The risk: borrowed language mistaken for borrowed proof.

THE ANSWER

The vocabulary is Taleb's; the underlying math is not finance-specific. Peters (2019) shows the ensemble-vs-time-average gap is a general property of any non-ergodic process — the same structure Kent et al. (2016) describe for trials vs. individual patients (slide 13). That structural claim survives the metaphor.

What doesn't survive: portfolio "ruin" (lost capital) and surgical "ruin" (revision, infection, death) are different orders of event, and the 90/10 barbell has no validated clinical translation.

Borrow the structure, not the numbers. If you cite this framework, cite Taleb's own books — The Black Swan (2007) and Antifragile (2012) — not a derivative summary deck, and state plainly which part is established math and which part is analogy.

Source: Peters, Nature Physics, 2019, Kent et al., Int J Epidemiology, 2016, Taleb, The Black Swan (2007) & Antifragile (2012)

An illustrative lens for this talk — not an established clinical finding

A trial's mean is an ensemble average. Your patient lives a time average of one.

A gambler who can only play once	should not bet on the expected value of a game designed for thousands of simultaneous players.
A patient who has one surgery, one recovery, one life	should not be told their outcome equals the mean improvement score of a trial's enrolled population.
Both are the same mathematical error	treating an ensemble average as if it were a time average — assuming ergodicity that does not hold.

This is the physiology EBM training rarely teaches: population statistics describe an ensemble. Clinical judgment exists precisely to translate that ensemble into a decision for one time-average life.

CASE IN POINT

The gap between "significant" and "certain for this patient," in real orthopaedic trials

Trial / analysis	What the population-level number says	What it doesn't tell the individual patient
COMPARE ACL trial (BMJ, 2021)	Early reconstruction beat rehab-first at 2 years, p = 0.026	Authors' own caveat: "clinical importance is unclear"; 50% of the rehab-first group never needed surgery
Thromboprophylaxis after major ortho surgery	NNT to prevent one VTE vs. NNH to cause one major bleed, expressed as a benefit-to-harm ratio	NNT alone says nothing; the harm side of the ledger changes the answer for higher-bleeding-risk patients
Osteoporotic hip-fracture prevention	NNT over 3 years ranges from 48 (strontium ranelate) to 200 (denosumab)	NNT is driven mainly by the placebo group's baseline absolute risk — not a fixed property of the drug

Source: Reijman et al., BMJ 2021; Mt-Isa et al., PDS 2014; Ferrari et al., Arch Osteopor 2016

A PRACTICAL PROTOCOL

Five questions before you trust a number

1	**Is this risk or uncertainty?**	Does a stable reference class actually exist for this patient, or is this a one-off situation dressed up in a probability?
2	**Absolute, or just relative?**	Convert every relative risk reduction back to an absolute number and an NNT before it changes your decision.
3	**How fragile is "significant"?**	Check (or calculate) the Fragility Index. A p-value that a handful of patients could flip is not a fact — it's a fragile estimate.
4	**Statistically real, or clinically real?**	Compare the effect size to the MCID, not just to $p < 0.05$.
5	**Whose average is this?**	An ensemble average across a trial population is not a forecast for this patient's time-average life. Say so, out loud, in the consent conversation.

The bottom line

Evidence-based orthopaedics needs an anatomy course and a physiology course — not just a literature search.

Know the structure of the evidence. Know how it actually functions in a decision. Know the difference between risk and uncertainty, between relative and absolute, between the average across everyone and the average for the one person on your table.

REFERENCES

Foundations and critical appraisal

1. Sackett DL, Rosenberg WMC, Gray JAM, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ. 1996;312:71–72. doi:10.1136/bmj.312.7023.71.

2. Oxford Centre for Evidence-Based Medicine. OCEBM Levels of Evidence. 2011. cebm.ox.ac.uk/resources/levels-of-evidence.

3. Eriksen MB, Frandsen TF. The impact of PICO as a search strategy tool on literature search quality: a systematic review. J Med Libr Assoc. 2018;106:420–431. doi:10.5195/jmla.2018.345.

4. Wasserstein RL, Lazar NA. The ASA statement on p-values: context, process, and purpose. Am Stat. 2016;70:129–133. doi:10.1080/00031305.2016.1154108.

5. Guyatt GH, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008;336:924–926. doi:10.1136/bmj.39489.470347.AD.

6. Walsh M, et al. 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.

REFERENCES

Risk, uncertainty, and communication

1. 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;8:53–96. doi:10.1111/j.1539-6053.2008.00033.x.

2. Knight FH. Risk, Uncertainty and Profit. Boston: Houghton Mifflin; 1921.

3. Dahabreh IJ, Hayward R, Kent DM. Using group data to treat individuals. Int J Epidemiol. 2016;45:2184–2193. doi:10.1093/ije/dyw125.

4. Djulbegovic B. Ethics of uncertainty. Patient Educ Couns. 2021;104:2628–2634. doi:10.1016/j.pec.2021.07.025.

5. Djulbegovic B. Well informed uncertainties about the effects of treatment. BMJ. 2004;328:1018. doi:10.1136/bmj.328.7446.1018-a.

6. Al-Maghrabi M, et al. Overconfidence, time-on-task, and medical errors. Adv Med Educ Pract. 2024;15:133–140. doi:10.2147/AMEP.S442689.

7. Peters O. The ergodicity problem in economics. Nat Phys. 2019;15:1216–1221. doi:10.1038/s41567-019-0732-0.

8. Taleb NN. The Black Swan. 2007. Antifragile. 2012. Random House.

REFERENCES

Orthopaedic applications

1. Bloom DA, et al. The Minimal Clinically Important Difference: a review of clinical significance. Am J Sports Med. 2023;51:520–524. doi:10.1177/03635465211053869.
2. Khan M, et al. The fragility of statistically significant findings from randomized trials in sports surgery. Am J Sports Med. 2017;45:2164–2170. doi:10.1177/0363546516674469.
3. Maldonado DR, et al. The Fragility Index of hip arthroscopy randomized controlled trials. Arthroscopy. 2021;37:1983–1989. doi:10.1016/j.arthro.2021.01.049.
4. Go CC, et al. The Fragility Index of total hip arthroplasty randomized control trials. J Am Acad Orthop Surg. 2022;30:e741–e750. doi:10.5435/JAAOS-D-21-00489.
5. Poursalehian M, et al. Fragility indices of RCTs in top general orthopaedic journals. J Am Acad Orthop Surg. 2025;33:e340–e347. doi:10.5435/JAAOS-D-24-00691.
6. Herndon CL, et al. Fragility Index in adult reconstruction. Arthroplast Today. 2021;11:239–251. doi:10.1016/j.artd.2021.08.018.
7. Reijman M, et al. Early ACL reconstruction versus rehabilitation with elective delayed reconstruction: COMPARE trial. BMJ. 2021;372:n375. doi:10.1136/bmj.n375.
8. Mt-Isa S, et al. Balancing benefit and risk of medicines: a systematic review. Pharmacoepidemiol Drug Saf. 2014;23:667–678. doi:10.1002/pds.3636.
9. Ferrari S, et al. Unmet needs and approaches for osteoporotic patients at high risk of hip fracture. Arch Osteoporos. 2016;11:37. doi:10.1007/s11657-016-0292-1.

NOTES

How to read these claims

The trial, p-value, Fragility Index, MCID, and natural-frequency examples are educational summaries of the cited literature. Worked arithmetic examples are illustrations, not patient-specific forecasts. The extensions from Knightian uncertainty, ergodicity, and antifragility to orthopaedic decision-making are conceptual analogies—not validated clinical prediction tools or treatment recommendations.

The 2014 Fragility Index review included 399 eligible randomized trials: median FI 8; 25% had FI ≤3; and in 53%, FI was smaller than loss to follow-up. These are descriptive reporting findings, not estimates of treatment efficacy. The COMPARE and osteoporosis examples retain the populations, time horizons, and caveats of their cited sources.

Visual credit: original diagrams and slide design by John Walsh, DO; embedded figures were retained from the author-supplied source deck and are tied to the cited literature. Drawn from the OrthoScienius evidence library.