Statistical Considerations in Surrogate Endpoint Evaluation

From evidence for surrogacy to prediction in a new trial

Annual Biopharmaceutical Applied Statistics Symposium (BASS)

Feng Li
Lead Mathematical Statistician
Division of Biometrics II, Office of Biostatistics
OTS/CDER/FDA

November 4, 2026

Disclaimer

- The views expressed in this presentation are those of the author and should not be construed as representing the official position of the U.S. Food and Drug Administration.

Roadmap

1. Regulatory and endpoint foundations
2. Cautionary cases and context of use
3. Surrogacy evaluation
4. New-trial efficacy interpretation
5. Rare disease and evidence-limited settings

Substantial evidence: statutory standard and regulatory criteria

- FD&C Act § 505(d) (21 U.S.C § 355(d))
  - "*Substantial evidence*" of effectiveness from "*adequate and well-controlled investigations*", and when appropriate, FDA may determine that *one adequate and well-controlled clinical investigation plus confirmatory evidence* is sufficient
  - "*Drug will have the effect it purports or is represented to have*"

- One of 7 criteria for adequate + well-controlled (21 CFR 314.126(b)(6))
  - "*The methods of assessment of subjects' response are well-defined and reliable. The protocol for the study and the report of results should explain the variables measured, the methods of observation, and criteria used to assess response*."

Sources: 21 U.S.C. § 355(d); FDA draft guidance, Demonstrating Substantial Evidence of Effectiveness (June 2026).

Adequate and well-controlled investigations make the treatment comparison credible

Key characteristics supporting a credible treatment comparison

1. Valid comparison with a control
2. Assignment and procedures that minimize bias
3. Well-defined, reliable response assessment
4. Appropriate, prespecified analysis of treatment effects

What a well-controlled trial helps establish

Did the intervention cause the observed treatment effect on this endpoint in this trial?

What it does not establish

Does a treatment effect on this endpoint reliably predict a meaningful effect on how patients feel, function, or survive?

Using a surrogate adds a second inference: from the treatment effect on the surrogate to the treatment effect on the clinical outcome.

Source: 21 CFR 314.126(a)-(b).

Endpoint taxonomy: what is directly observed, and what must be inferred

Endpoint type	What it measures	Typical regulatory role
Clinical endpoint	Directly measures an aspect of clinical benefit - how patients feel, function, or survive.	May support traditional approval when the effect is clinically meaningful.
Validated surrogate	Not a direct measure; clear mechanistic and clinical data providing strong evidence that its treatment effect predicts a specific clinical benefit in a defined context.	May support traditional approval within the supported context.
Reasonably likely surrogate	Not a direct measure; supported by strong mechanistic and/or epidemiologic rationale but lacks sufficient clinical validation.	May support accelerated approval when statutory criteria are met.
Intermediate clinical endpoint	A clinical endpoint measured earlier than irreversible morbidity or mortality and reasonably likely to predict later benefit.	May support accelerated approval or traditional approval if the endpoint itself reflects meaningful clinical benefit; context-dependent.

Sources: FDA-NIH BEST Resource; FD&C Act §§ 506(c) and 507(e)(9).

Why surrogate endpoints require rigorous evaluation

Why we consider them

- Greater feasibility when the target clinical endpoint is rare, slow, or difficult to observe
- Can enable shorter or smaller trials
- Potentially earlier decisions and patient access

These advantages matter only if the surrogate effect translates into meaningful patient benefit.

Why special caution is needed

- A treatment can improve the surrogate without improving the target clinical endpoint, despite an association between endpoints
- Off-target effects, confounding, subgroup heterogeneity, or insufficient surrogate effect size can lead to paradoxical findings
- More uncertainty in benefit-risk assessment due to shorter and smaller trials

Replacement is justified only when treatment effects on the surrogate predict treatment effects on the target clinical endpoint with uncertainty acceptable for the intended use.

CAST: arrhythmias were suppressed, but mortality increased

A prognostic association plus a favorable marker response did not establish patient benefit.

1 OBSERVED ASSOCIATION

After myocardial infarction, frequent ventricular premature depolarizations were associated with sudden death.

2 MARKER RESPONSE

Encainide and flecainide suppressed frequent ventricular premature depolarizations.

3 RANDOMIZED OUTCOME

The encainide and flecainide arms were stopped early because arrhythmic death or cardiac arrest and total mortality were higher than with placebo.

Surrogacy requires evidence that treatment effects on the marker predict treatment effects on the clinical outcome.

CAST Investigators. N Engl J Med. 1989;321:406–412

Two related but distinct questions

1 Surrogacy evaluation

Question

Do treatment effects on the proposed surrogate predict treatment effects on the target clinical endpoint in the proposed context of use?

Output

A bounded conclusion about whether and how the surrogate can be used in that context.

2 New-trial efficacy evaluation

Question

Given the observed surrogate effect, what target clinical effect is predicted, how uncertain is it, and is it clinically meaningful?

Output

A predictive distribution for clinical benefit that informs benefit-risk.

Even with a validated surrogate for the intended use, a small observed effect may imply limited or uncertain clinical benefit.

Not every statistically significant surrogate effect is sufficient.

Surrogacy is context-specific

Why this matters

- Surrogacy is a property of the relationship between treatment effects in context, not a universal property of the endpoint.

What not to assume

- A surrogate supported in one setting should not be assumed to generalize automatically to another setting, even within the same disease area.

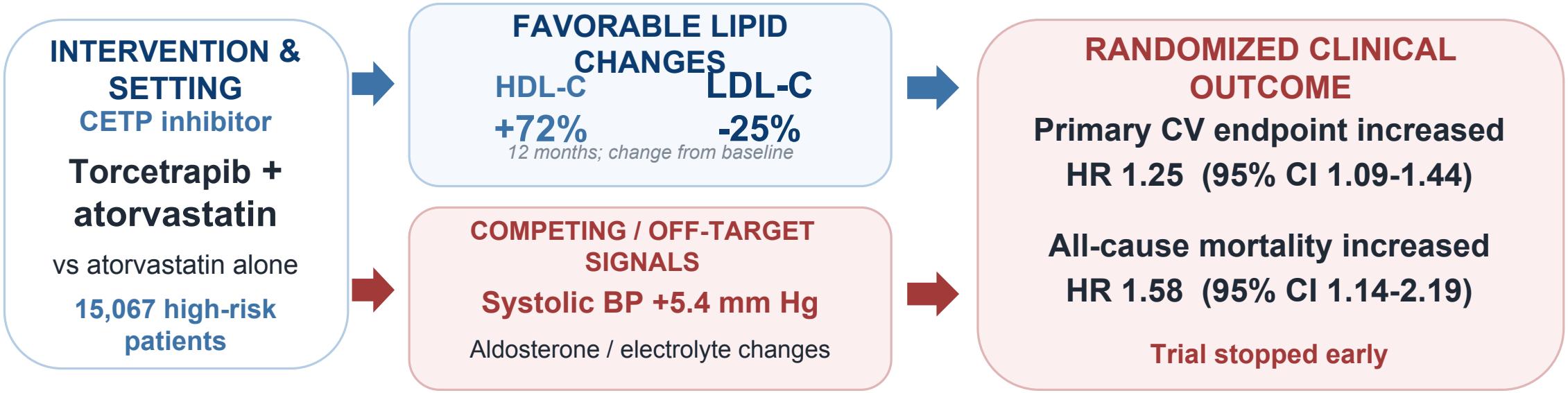
Central evaluation question

- Can treatment effects on the surrogate predict the treatment effects on the target clinical endpoint in the proposed context with a level of confidence appropriate to the intended regulatory use?

Surrogacy is a relationship among treatment effects within a defined context of use.

Why context of use matters: torcetrapib and competing pathways

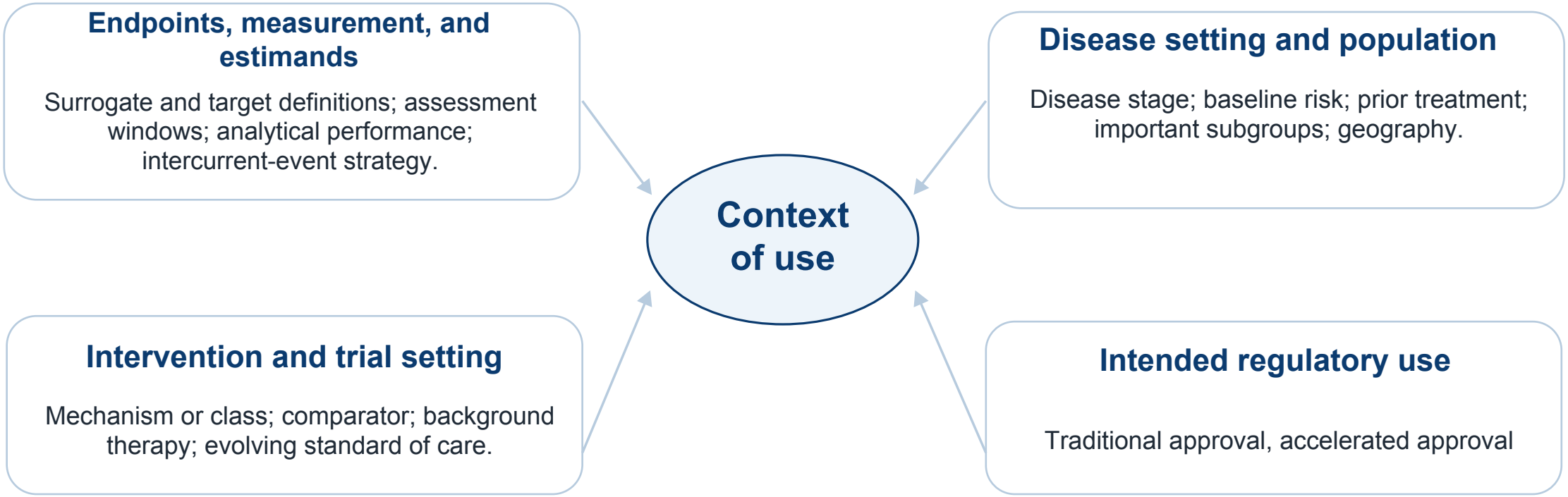
Favorable lipid changes did not ensure net cardiovascular benefit.



Evidence supporting a surrogate in one intervention context does not ensure that it captures competing effects of another intervention.

Source: Barter et al. N Engl J Med. 2007;357:2109-2122 (ILLUMINATE).

What defines the context of use (COU)



- Transportability must be justified;
- Extrapolation beyond the supported context should be limited.

Scientific evidence spans two domains and is judged on a spectrum

Biological plausibility

- Mechanistic link to the target clinical endpoint
- Alignment with the intended mechanism of action
- Assessment of competing pathways and off-target effects
- Support from pathophysiology, nonclinical evidence, and natural history, as appropriate

Empirical evidence

- Prefer evidence linking treatment effects on both endpoints
- Trial-level treatment-effect meta-regression addresses prediction more directly when the evidence base is informative
- Subject-level association is supportive rather than primary evidence
- Assess uncertainty and transportability explicitly

Illustrative spectrum

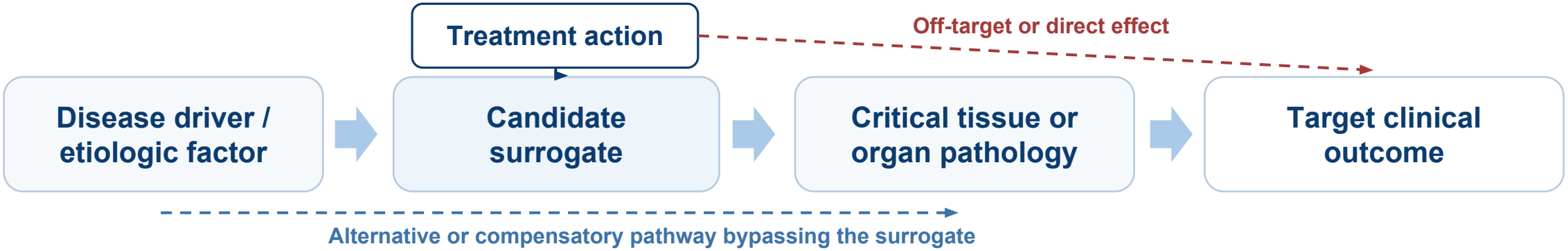
Illustrative profiles only — not rigid bins.

Evidence lies on a spectrum; judgment should be documented, not mechanically classified.

Measurement and study quality are prerequisites; mechanistic and empirical evidence should converge on treatment-effect prediction.

Mechanistic plausibility: locate the surrogate in the causal system

Ask where the surrogate sits, where treatment acts, and what pathways the surrogate does not capture.



CAUSAL POSITION

Is the surrogate an integral mediator, a downstream marker, or a bystander?

INTERVENTION ALIGNMENT

Does treatment act through the pathway represented by the surrogate?

BYPASS RISK

Could alternative pathways, tissue mismatch, feedback, or toxicity alter the outcome independently?

Mechanistic understanding provides the scientific rationale; it does not by itself establish treatment-effect prediction.

Subject-level association is not treatment-effect prediction

Subject-level question

Do subjects with better surrogate values tend to have better clinical outcomes?

What each subject contributes

- Paired surrogate and clinical-outcome measurements

What we can learn

- Prognostic value and biological coherence

Key limitation

- The subject-level relationship is prognostic; it is not a randomized treatment-effect relationship.

Trial-level prediction question

Do randomized treatment effects on the surrogate predict randomized treatment effects on the target endpoint?

What each trial contributes

- Paired treatment-effect estimates
- Their standard errors and within-trial covariance

What we want to learn

- Across relevant trials, do more favorable surrogate effects predict more favorable clinical effects?

What we ultimately need

- A predicted clinical effect for a new trial, together with its uncertainty

Subject-level association is supportive; but it cannot by itself establish or rule out treatment-effect surrogacy.

Why subject-level association alone is not enough

A surrogate can look useful within subjects and still fail to predict treatment benefit across studies.

Off-target effects

Treatment may improve the surrogate but harm the clinical outcome through another pathway that the surrogate does not capture.

Confounding

Common post-baseline causes of both endpoints can create a reassuring subject-level association that does not reflect a true treatment-effect relationship.

Subgroup heterogeneity

Differential effects on the surrogate and clinical outcome across subgroups can make pooled results nontransportable or conceal an unfavorable clinical effect.

Surrogate effect too small

The surrogate may move in the right direction, but the change may still be too small to lead to a clinically meaningful benefit.

Subject-level association can support prognosis and biological coherence; by itself, it does not establish treatment-effect surrogacy.

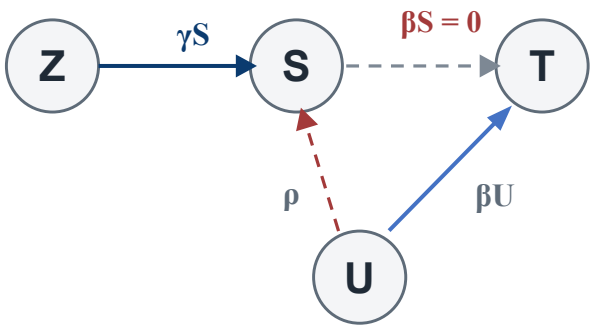
Within-trial methods are supportive with limitations

Approach	What it can add	Critical limitation
Prentice criteria / adjusted treatment effect	Checks treatment effects on both endpoints and attenuation of the treatment coefficient after adjustment for the surrogate.	The surrogate is post-randomization; a nonsignificant residual effect is not proof of full capture and depends on model form and precision.
PTE and causal mediation analyses	Proportion of treatment effect explained (PTE) describes attenuation; causal mediation decomposes effects under explicit assumptions.	Descriptive PTE can be unstable, negative, or greater than one; causal mediation requires strong identification assumptions.
Principal stratification and related causal frameworks	Explores effects within latent strata defined by potential surrogate responses and can reveal heterogeneity.	Principal strata are unobserved; identification generally requires structural assumptions or external information.

Randomization makes the treatment comparison credible; it does not randomize the post-baseline surrogate.

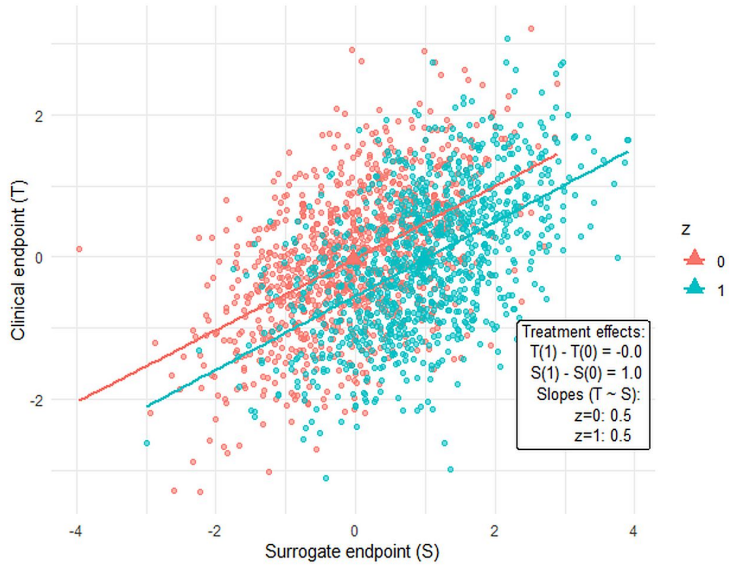
Simulation intuition: false reassurance from subject-level association

Confounding can create a positive within-trial association even when changing S has no effect on T.

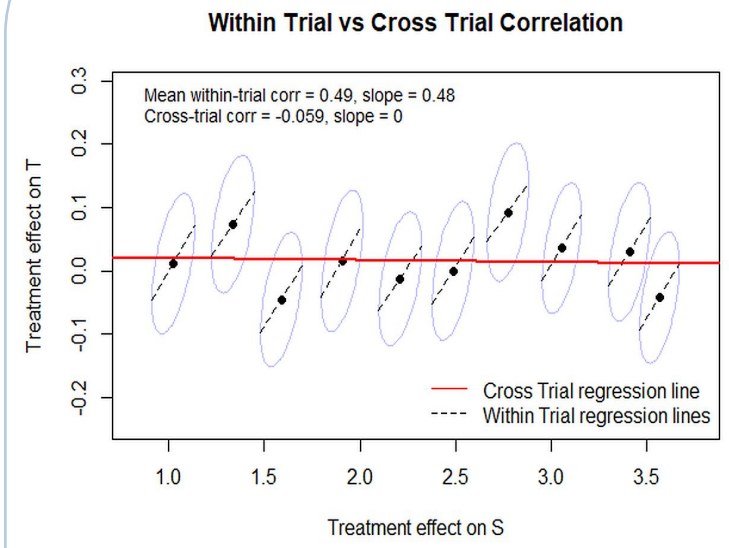


Data-generating model

Z affects S. U affects T and correlates S. The causal S→T effect is zero, so the positive subject-level slope is spurious.



Within-trial subject-level relationship



Across-trial treatment-effect relationship

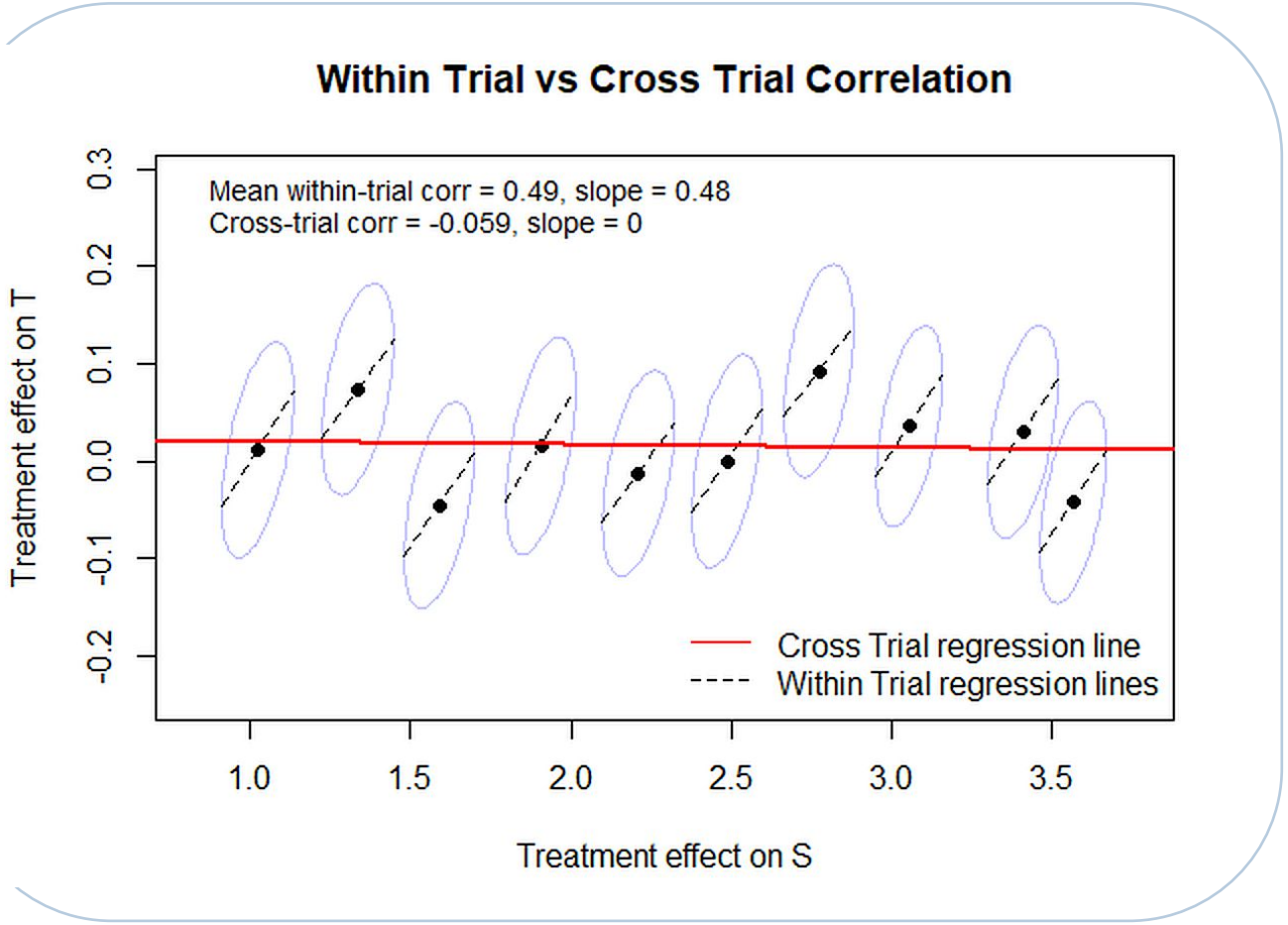
Cross-trial regression recovers the null relationship here only under the stated restricted assumptions.

Illustrative simulation; not empirical data.

Cross-trial treatment-effect evidence addresses prediction more directly

Why it matters

- Each trial contributes paired randomized treatment effects, not a subject-level correlation
- Strong subject-level association can coexist with weak treatment-effect prediction
- This is closer to the future-trial prediction question, but uncertainty and context still matter.



Same illustrative simulation as Slide 18; not empirical data.

Cross-trial meta-regression: model the true treatment effects

Propagate sampling error in estimated treatment effects on both endpoints.

1 Trial inputs

- Surrogate treatment-effect estimate and SE
- Target-endpoint treatment-effect estimate and SE
- Within-trial covariance of the two estimates
- Comparable endpoint definitions and estimands

2 Latent treatment-effect model

Model the relationship between true treatment effects; use random effects to account for between-trial heterogeneity.

3 Outputs

- Predicted target effect and prediction interval
- Posterior predictive distribution, when Bayesian
- Posterior predictive probability (target benefit exceeds a clinical threshold)
- Surrogate threshold effect (STE) or clinically meaningful surrogate range
- Validation and transportability diagnostics

The target is prediction for a new trial - not only a fitted slope, correlation, or R^2.

What cross-trial meta-regression can and cannot establish

Can support

- Association between treatment effects over the observed effect range
- A predictive model with residual heterogeneity and uncertainty
- Prediction intervals or posterior predictive probabilities for a new trial
- Decision aids such as an STE or clinically meaningful surrogate range

Cannot by itself

- Prove causal mediation or that all clinical effects operate through the surrogate
- Eliminate cross-trial confounding, selective evidence, or treatment-class effects
- Rescue few trials, little true effect variation, or poor signal-to-noise
- Justify extrapolation beyond the supported population, class, estimand, or effect range

A precise mean line can still yield poor prediction in a new trial.

Before fitting: can the evidence support cross-trial prediction?

Core evidence requirements

1. Enough relevant, independent randomized trials
2. Adequate spread of surrogate treatment effects over clinically relevant prediction range
3. Reliable paired treatment-effect estimates, including uncertainty and covariance
4. Harmonized endpoints, estimands, follow-up, populations, and other key design features

Red flags for weak identification or unsupported extrapolation

- One trial or treatment class would dominate the relationship
- Little within treatment class variation; the pooled pattern is mainly between classes
- Clinical endpoint results are selectively reported or unavailable
- Clinically relevant predictions would extend beyond the historical surrogate-effect range

Model complexity should match the independent information in the evidence base.

A prediction model should say more than “the association looks good”

Do not stop at a summary statistic such as R^2.

What is the direction and shape of the association?

Check whether more favorable surrogate effects predict more favorable clinical effects; assess slope, nonlinearity, and the expected clinical effect when the surrogate effect is near zero.

How uncertain is prediction for a future comparable trial?

Report a frequentist prediction interval or Bayesian posterior predictive interval for clinical effect across the relevant range of surrogate effects. Wider intervals mean less precise prediction.

What surrogate effect satisfies the decision criterion?

Report the surrogate threshold effect (STE) for excluding a null clinical effect—and, where relevant, a threshold tied to a prespecified clinically meaningful benefit.

Are predictive results robust and applicable within the intended COU?

Assess sensitivity to model choices and influential trials, leave-one-trial-out prediction, and relevance of the historical evidence to the intended setting.

R^2 depends on the observed effect range, model, weighting, and treatment-class mix; no universal cutoff establishes predictive validity.

Cross-trial surrogacy evaluation: a six-step workflow

1 DEFINE

Prespecify the COU, the surrogate and clinical endpoint estimands, effect direction and scale, and the clinical decision criterion.

2 ASSEMBLE

Identify all eligible randomized trials; assess risk of bias and missing or selectively reported results.

3 HARMONIZE

Harmonize endpoint definitions, timing, effect scales, and estimands; assess population comparability; derive paired effects, standard errors, and covariance.

4 MODEL

Match complexity to information content; fit a prespecified meta-regression that accounts for uncertainty and covariance in both treatment effects; quantify residual heterogeneity.

5 STRESS-TEST

Assess influential trials or classes, leave-one-trial-out prediction, model sensitivity, and evidence that contradicts the fitted relationship.

6 CHARACTERIZE PREDICTIVE PERFORMANCE

Estimate future-trial prediction intervals across clinically relevant surrogate effects; report decision-relevant probabilities or thresholds; assess transportability.

RESIDUAL UNCERTAINTY Document critical knowledge gaps • conflicting or contrary evidence • untested assumptions • transportability limits

Prespecify the context of use, primary model, prediction target, and sensitivity analyses before examining results.

OUTPUT: Predictive uncertainty, decision-relevant thresholds, and the supported context of use.

From surrogate evidence to regulatory acceptance

Support for surrogate use lies on a spectrum; acceptable residual uncertainty depends on the decision context.

Scientific evidence strength

- Adequate support in the intended context of use
- Biological plausibility + empirical evidence both matter
- Support judged on a spectrum, not on a simple binary cut-off

Context and tolerance of residual uncertainty

- Disease seriousness, rarity and unmet need
- Feasibility and ethics of directly observing the target endpoint
- Available therapy, safety information, and overall benefit-risk

These factors shape the acceptable level of residual uncertainty

Strong evidence of reliable prediction	Reasonably likely to predict clinical benefit	Insufficient for primary efficacy use
Validated surrogate May support traditional approval	Reasonably likely surrogate May support accelerated approval	Usually not supportable as the primary efficacy basis

Context changes the tolerable residual uncertainty; it does not lower the statutory standard.

Bone Mineral Density (BMD) case study (1/2): qualified context, biology, and safeguards

FDA-QUALIFIED CONTEXT OF USE

ENDPOINT
Percentage change from baseline in total hip BMD

ASSESSMENT
24 months, measured by DXA

POPULATION
Postmenopausal women with osteoporosis at risk for fracture

TRIAL ROLE
Validated surrogate endpoint for phase 3 assessment of investigational therapies

BIOLOGICAL PLAUSIBILITY + BYPASS-RISK CONTROL

Therapy alters bone remodeling → **Favorable effect on total hip BMD** → **Contributes to greater bone strength and lower fracture risk**

Plausible pathway - but BMD does not capture every determinant of fracture.

NOT FULLY CAPTURED BY BMD

- Microarchitecture and material quality
- Bone geometry
- Falls, trauma, and off-target effects

SUBSET BIOPSY SAFEGUARD

Biopsy can assess architecture, mineralization, and remodeling for abnormalities not reflected by DXA BMD.

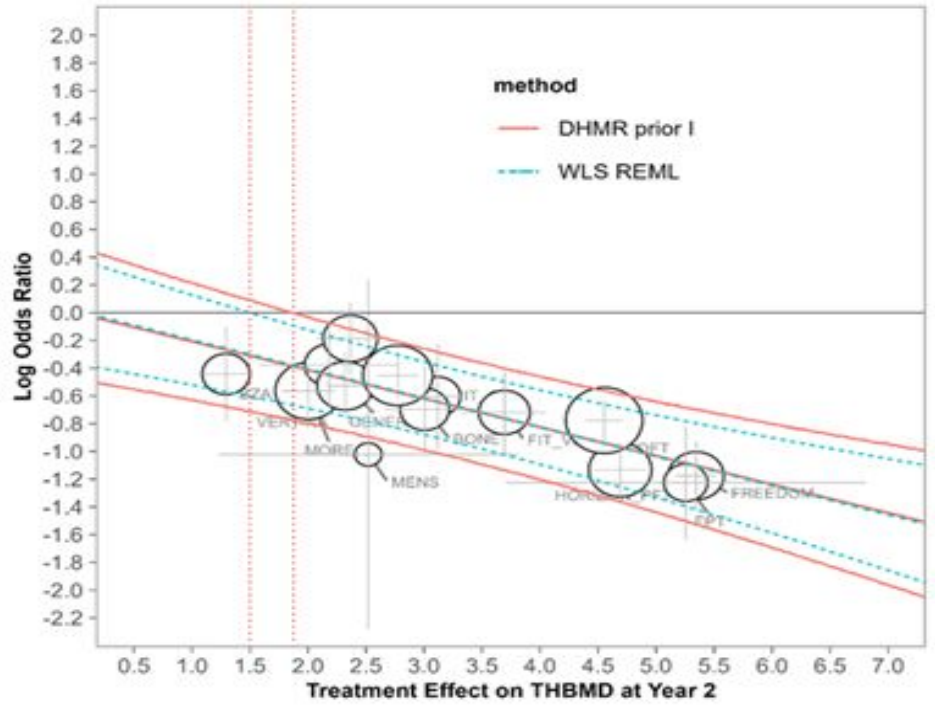
Program-specific: exact scope depends on mechanism and FDA agreement.

Additional assessment of bone quality and off-pathway harm may be needed, depending on the mechanism and program.

Sources: FDA BMD qualification announcement (Dec. 19, 2025); Eastell et al., J Bone Miner Res. 2022;37:29-35.

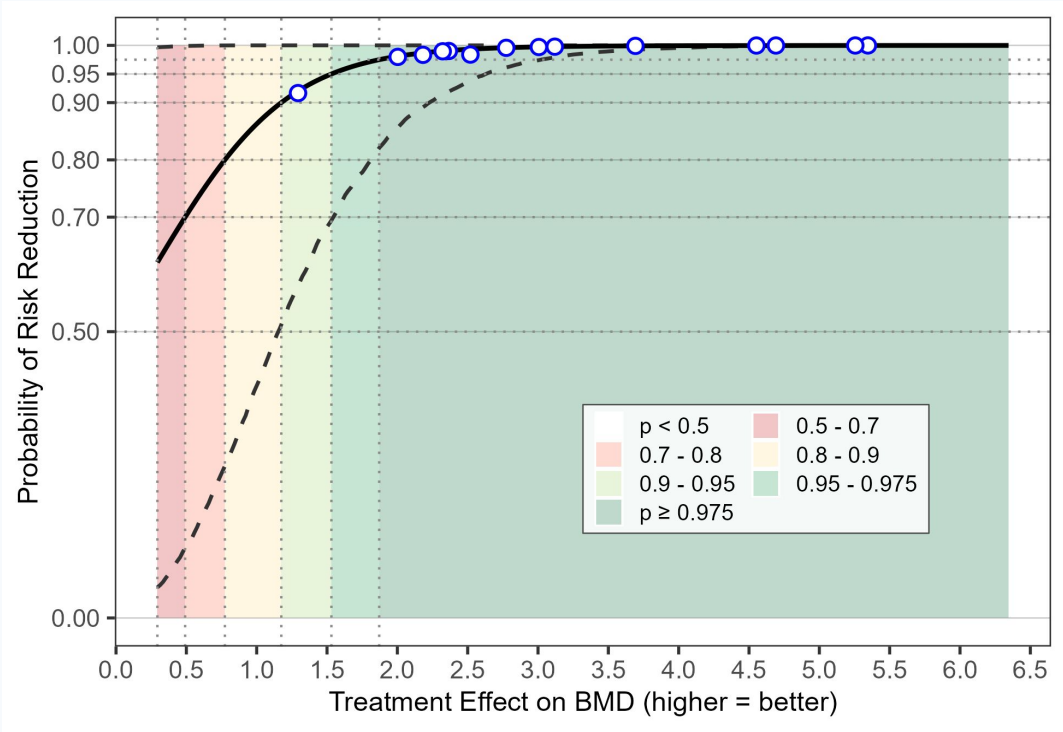
BMD case study (2/2): from total hip BMD effect to vertebral fracture-risk prediction

Illustrative cross-trial treatment-effect relationship



Each bubble represents a trial. 95% prediction interval based on two methods

Illustrative predictive probability of vertebral risk reduction



Solid curve denotes the posterior predictive probability of any risk reduction; with dotted lines showing the 95% equal-tail interval. Blue circles mark the predicted value for each study.

New-trial efficacy evaluation: is the observed surrogate effect large enough?

Statistical significance on the surrogate does not quantify the magnitude or certainty of clinical benefit.

1 CONFIRM APPLICABILITY

Check population, treatment class or mechanism, comparator, endpoints, estimands, follow-up, and surrogate-effect range against the supported context of use.

2 ESTIMATE SURROGATE EFFECT

Estimate the new trial's surrogate treatment effect and its standard error.

3 APPLY PRESPECIFIED MODEL

Condition on the estimated surrogate effect and its standard error; propagate model uncertainty and residual between-trial heterogeneity.

4 INTERPRET WITH ALL EVIDENCE

Report the predicted clinical effect, prediction interval, and prespecified probability; integrate direct clinical and safety data.

Decision questions:

- Is the predicted clinical benefit, with its uncertainty, compelling for a favorable benefit-risk assessment?
- For accelerated approval, the model may inform confirmatory-trial planning or a prespecified predictive-probability analysis; it does not replace verification of clinical benefit.

Rare-disease settings may require a broader evidentiary package

Why a cross-trial surrogacy data set may not exist

- Few development programs or completed randomized trials
- Small and heterogeneous populations
- Slow disease progression and delayed clinical outcomes
- Limited between-trial variation or no trials with both endpoints

How the evidentiary package may change

- Broader integration of compelling biology, natural history, registries, PK/PD, observational, and nonclinical evidence
- Distinguish prognostic association from evidence about treatment-effect prediction
- Make assumptions explicit and quantify sensitivity and residual uncertainty
- Retain a feasible plan for direct or confirmatory clinical evidence where possible

Rare-disease constraints call for flexibility in evidence generation—not relaxed standards of inference.

When surrogate-only inference is not justified: partial borrowing

Illustrative Bayesian strategy: keep the target clinical endpoint as primary and borrow conditionally from surrogate evidence.

Direct clinical endpoint inference

Directly measure clinical benefit

No surrogate-based borrowing

Surrogate-informed partial borrowing

Dynamic, conditional borrowing

Clinical endpoint remains primary

Surrogate-only efficacy inference

Clinical benefit inferred from the surrogate

Only when supported in the defined COU

- Primary inference remains on the target clinical endpoint.
- Borrow more when the surrogate prediction and clinical data agree; discount when they conflict.
- May improve efficiency while preserving direct evidence of clinical benefit.
- Prespecify maximum borrowing, endpoint dependence, conflict handling, operating characteristics, and a no-borrowing sensitivity analysis.

Use when surrogate evidence is informative but insufficient for surrogate-only efficacy inference.

Key takeaways

Three principles to carry forward

1. Define the context of use precisely; do not assume transportability.

2. Prioritize randomized treatment-effect evidence; subject-level association is supportive, not sufficient.

3. In a new trial, translate the magnitude and precision of the surrogate effect into predicted clinical benefit – do not stop at the surrogate p-value.

Operational implications

1. Plan prospectively with a multidisciplinary team.
2. Prespecify the context of use, model, prediction target, and challenge analyses.
3. When surrogate-only inference is not supported, retain direct clinical evidence or consider prespecified partial borrowing.

State where prediction is supported, what uncertainty remains, and how that uncertainty affects the decision.

Selected FDA and methodological resources

FDA / regulatory resources

- FDA-NIH BEST Resource; FDA Surrogate Endpoint Table
- Biomarker Qualification Program; evidentiary framework
- Total hip BMD qualification (Dec. 19, 2025)
- Accelerated Approval; confirmatory-trial resources
- Substantial Evidence of Effectiveness - draft (June 2026)
- Bayesian Methodology - draft (Jan. 2026)
  Bioanalytical Method Validation for Biomarkers - final (Apr. 2026)

Selected methodological references

- Prentice (1989)
- Fleming and DeMets (1996)
- Daniels and Hughes (1997)
- Buyse et al. (2000)
- Burzykowski, Molenberghs, and Buyse (2005)
- Fleming and Powers (2012); VanderWeele (2013)
- Black et al. (2020)

Acknowledgments

- Office of Biostatistics Surrogate Endpoints Working Group
  - Dali Zhou, Ling-Wan Chen, Xiongce Zhao, Rebecca Hager, Therri Usher, Anup Amatya, Yu-Ting Weng, Jiwei He, Jing Zhang

- Greg Levin, Office of Biostatistics