NIH PRAGMATIC TRIALS COLLABORATORY
Rethinking Clinical Trials®

Endpoint Strategy in Pragmatic Trials: From Big Questions to Meaningful Outcomes

WORKSHOP SUMMARY

Table of Contents

Executive Summary **2**
Introduction **3**
Defining the Endpoint: Where Evidence Begins **3**
Aiming for Impact: What Is the Right Concept to Target? **5**
The Estimands Framework: Aligning Objectives, Endpoints, and Analyses **7**
Working With Multiple Outcomes: Multicomponent Endpoints and Beyond **9**
Moving From Measurement to Meaning **11**
Unpacking Uncertainty, Meaningfulness, and Magnitude **11**
Toward Best Practices for Interpretation of Endpoints **15**
References **18**

Prepared by: NIH Pragmatic Trials Collaboratory Coordinating Center
Version: November 9, 2026

Executive Summary

The NIH Pragmatic Trials Collaboratory held a 2-day workshop, "Endpoint Strategy in Pragmatic Trials: From Big Questions to Meaningful Outcomes," to explore the factors involved in selecting endpoints and interpreting treatment effects in embedded pragmatic clinical trials (ePCTs) to best inform decision-making. This report summarizes discussions from each workshop session.

Access the complete workshop materials, including recordings.

Day 1 Opening Presentation: Defining the Endpoint: Where Evidence Begins

The opening presentation covered foundational concepts related to endpoint selection in ePCTs. Using a case study, Dr. Weinfurt introduced considerations for each step of the process from design to interpretation.

Panel 1: Aiming for Impact: What Is the Right Concept to Target?

Outcome selection in ePCTs requires balancing the likelihood of detecting an intervention effect with the relevance of that effect to patients, clinicians, funders, payers, and health systems. The strongest design works backward from the decision the trial should inform and aligns the outcome, timing, intervention, data source, and implementation context.

Panel 2: The Estimand Framework: Aligning Objectives, Endpoints, and Analyses

The estimand framework makes the target treatment effect explicit and aligns the trial design, conduct, analysis, and interpretation. Panelists described how the framework can assist with endpoint choices and give teams a shared language for communication throughout trial design, startup, and conduct.

Panel 3: Working With Multiple Outcomes: Multicomponent Endpoints and Beyond

Multicomponent and hierarchical endpoints offer a way to preserve multidimensionality while maintaining a single prespecified primary endpoint. These approaches can be useful for conditions like pain where a single measure fails to represent the full impact.

Day 2 Opening Presentation: Moving From Measurement to Meaning

Day 2 of the workshop opened with a reframing of endpoint strategy as a question of values, tradeoffs, and interpretation rather than statistics alone. Dr. Weinfurt introduced questions for consideration when judging meaningfulness.

Panel 4: Unpacking Uncertainty, Meaningfulness, and Magnitude

Treatment effects become meaningful only after someone judges what matters, for whom, and under what conditions. Patients, clinicians, health system leaders, and regulators may use the same evidence differently because of distinct priorities, values, and contexts.

Panel 5: Toward Best Practices for Interpretation of Endpoints

While interpretation occurs after data collection and analysis, its validity depends on earlier choices. The final panel recommended best practices during trial planning that maximize the ability to estimate treatment effects precisely and for decision-makers to interpret them meaningfully.

Introduction

The NIH Pragmatic Trials Collaboratory held a virtual workshop on September 8–9, 2026, to explore the factors involved in selecting endpoints and interpreting treatment effects in embedded pragmatic clinical trials (ePCTs) to best inform decision-making.

To date, the NIH Collaboratory has launched 37 ePCTs involving >1,400 clinical sites. Through these trials, NIH Collaboratory teams and leadership have learned much about generating evidence capable of driving change in complex and evolving healthcare systems, including the many unique challenges and opportunities of pragmatic research.

In this workshop, a series of discussions addressed considerations for selecting endpoints that reliably measure concepts of interest; influential factors in the analysis and interpretation of treatment effects; applications of multicomponent and hierarchical endpoints; and examples of endpoint strategies in ePCTs.

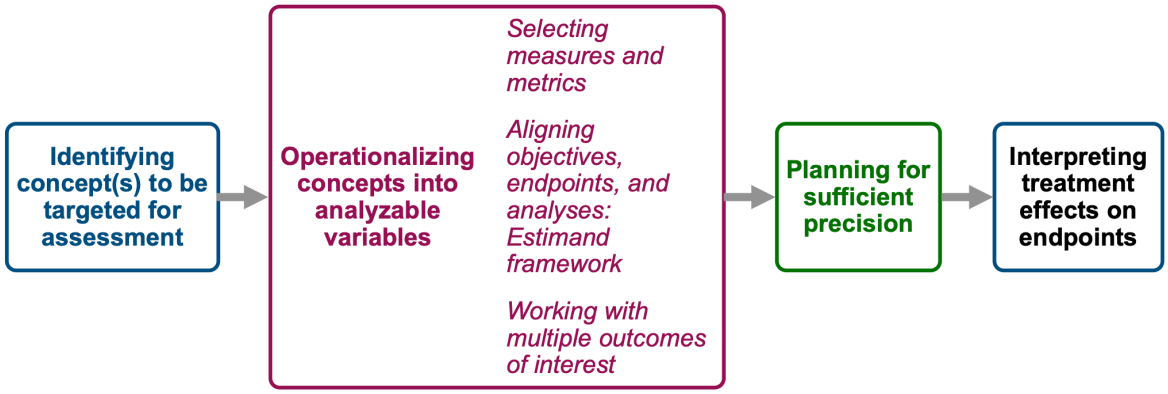
Access the complete workshop materials, including slides and recordings.

Defining the Endpoint: Where Evidence Begins

Speaker: Kevin Weinfurt, PhD, Duke University

Pragmatic clinical trials are studies designed primarily to inform decision-makers about the comparative balance of benefits, burdens, and risks of biomedical or behavioral health interventions at the individual or population level.[1] A benefit generally begins as a health concept, such as reducing the impact of pain on daily living. Moving from concept to evidence involves 4 steps (Figure 1).

Figure 1. Key Endpoint Considerations in ePCTs



Certain events, called *intercurrent events*, can affect whether an outcome is observed or how it is interpreted. Examples include study discontinuation, death, and nonadherence. The estimated treatment effect depends partly on how these events are handled.

The International Council for Harmonisation introduced the *estimand framework* to make the target of estimation explicit and to align the research question, endpoint, analysis, and interpretation.[2] There is an extension for cluster-randomized trials.[3]

Case Study: ACP PEACE

Using the ACP PEACE[4,5] trial as an example, Dr. Weinfurt illustrated the process of endpoint selection. The trial tested whether a video decision aid for patients, combined with goals-of-care communication training for oncology clinicians, would increase advance care planning among older patients with advanced cancer.

Identifying the Target Concept

The intervention could plausibly influence several concepts along a causal pathway. These included the occurrence of a goals-of-care conversation, delivery of care aligned with patient goals, receipt of that care, and, ultimately, achievement of the patient's goals. The research team selected the occurrence of a goals-of-care conversation as the concept to assess.

Operationalizing the Endpoint

The trial assessed advance care planning documentation in electronic health record (EHR) notes with a binary endpoint: documentation was present or absent. Natural language processing assisted human adjudication to identify qualifying conversations. The trial used an intention-to-treat approach. In estimand terminology, this corresponds to a treatment-policy strategy for relevant intercurrent events. The analysis assesses the consequence of assigning an individual to receive the intervention, regardless of adherence.



At minimum, an endpoint definition should specify the *assessment* (eg, PROMIS Pain Interference), *timing* (eg, baseline and 3 months), and *operation* (statistical or logical transformation applied; eg, change-from-baseline).

Planning for Precision, Magnitude, and Meaningfulness

Trials must be designed with enough precision to detect and characterize a meaningful treatment effect. In ACP PEACE, planning assumptions provided >90% power to detect a 10-percentage-point absolute increase in advance care planning documentation. The published interpretation described the adjusted estimated absolute effect of 7 percentage points as modest but clinically and statistically meaningful when considered across a population. This illustrates investigators' responsibility to define a meaningful effect during planning and to explain the practical significance after results are available.

Interpreting: Balancing Statistical and Clinical Information

Endpoint discussions often reveal a tension between preserving statistical information and maximizing clinical interpretability. A continuous scale retains more information and may provide greater power than a categorical measure. Clinicians, however, may prefer categories such as *normal function* vs *dysfunction*, because they are easier to interpret. A proposed resolution is to use the most granular endpoint and an appropriate statistical model, then use the model to express findings in clinically meaningful terms. For example, a model based on continuous scores can estimate the proportion of participants in each trial arm that exceed a clinically relevant threshold.

Working With Multiple Outcomes

Complex, multicomponent interventions may affect several outcomes, particularly among patients with multiple morbidities. Investigators can consider several approaches:

- **Separate endpoints:** Outcomes may be analyzed individually, with one designated as primary and others as secondary or exploratory. Trials may also use co-primary endpoints where benefit must be demonstrated on both outcomes.
- **Composite endpoints:** A subject is considered to have experienced the endpoint if any one of several prespecified component events occurs. An example is the major adverse cardiovascular event endpoint (MACE), which incorporates events such as death, myocardial infarction, revascularization, stroke, or heart failure.
- **Multicomponent endpoints:** Two or more components are combined within a subject using weights or a prespecified rule.
- **Hierarchical composite endpoints:** Components are ordered or prioritized so that outcomes can be compared according to their clinical hierarchy.

Key Takeaways

- **Identify:** Start with the health concept the intervention is meant to change, ensuring it is meaningful to decision-makers and can be plausibly impacted by the intervention during the duration of the planned trial.
- **Operationalize:** The endpoint definition should specify the assessment, timing, and operation. Also prespecify how intercurrent events will be handled.
- **Plan:** Design for enough precision to characterize a meaningful effect, not merely to reach statistical significance.
- **Interpret:** Analyze the most granular form of the endpoint, then translate the results into terms decision-makers can use.

Aiming for Impact: What Is the Right Concept to Target?

Panelists: Michael Ho, Frederick Masoudi, Miguel Vazquez, Tracy Wang, Wendy Weber; Moderator: Steven George

The session explored how outcome selection in ePCTs requires balancing the likelihood of detecting an effect with the relevance of that effect to decision-makers.

The Proximate-Distal Continuum

Dr. George described tradeoffs around the selection of outcomes along a causal chain relative to an intervention's presumed mechanism of action. Proximate outcomes are close to the intervention's immediate action, while distal outcomes occur farther downstream. For an EHR alert that identifies candidates for specialty referral, for example, the number of referrals would be a proximate outcome and patient quality of life a distal one. A review of completed NIH Collaboratory Trials suggested a pattern: proximate outcomes were more likely to show statistically significant effects, while distal outcomes were more likely to be null. Proximate outcomes may be more sensitive but could also be less compelling to

decision-makers, whereas distal outcomes may be more meaningful for practice and policy but harder to influence within a trial's duration.



The outcome most likely to change may not be the outcome interest holders care about most.

Case Study: ICD-Pieces

ICD-Pieces[6] tested whether information technology combined with practice facilitation could improve care for patients with chronic kidney disease, diabetes, and hypertension. Primary care practices were randomized to receive usual care or an intervention that supported guideline-directed care, and the primary outcome was hospitalization at 1 year. Hospitalization rates were similar in both groups, and secondary clinical outcomes also showed no meaningful differences. However, a blinded manual review of a random 10% sample showed that the intervention improved several processes of care, including problem-list updates, treatment-goal setting, patient education, initiation of renin-angiotensin system inhibition, and achievement of blood pressure goals.



A null distal outcome does not necessarily mean an intervention had no effect. Trials should collect proximate and implementation measures to help determine whether the intervention reached its target and when effects emerged.

Case Study: Nudge and Chat 4 Heart Health

The Nudge[7,8] trial used a text-message intervention to improve adherence to chronic cardiovascular medications. Adherence improved modestly during the first 3 months but did not differ at 12 months. Clinical outcomes were also unchanged. The mismatch between intervention timing and outcome assessment informed the ongoing Chat 4 Heart Health[9] trial, which delivers AI-enabled text messages focused on reducing cardiovascular risk. The team added a 7-month assessment to examine the outcome closer to intervention delivery while retaining a 12-month assessment to evaluate durability.

Funder, Patient, and Health System Perspectives

- **Funders** want outcomes that are patient-centered, practice-changing, and appropriate to the scale of the investment.
- **Patients and community members** want an outcome that is meaningful to them.
- **Health systems** prioritize outcomes that align with existing quality initiatives, value-based payment models, operational priorities, and available data.
- **Frontline staff** value workflow and process outcomes.

Key Takeaways

- Begin with the decision the trial should inform and align outcomes with the interest holders who will use the result.
- Match outcome timing to the expected time course of intervention effects. Consider collecting proximate measures even when the primary outcome is distal.

The Estimands Framework: Aligning Objectives, Endpoints, and Analyses

Panelists: Elizabeth Colantuoni, Fan Li, Richard Skolasky, Jr, Hulin Wu; Moderator: Liz Turner

An *estimand* is a precise description of the treatment effect a study aims to quantify (Figure 2). A statistical method does not uniquely define the scientific question. Yet, most medical research reports do not state the treatment effect being estimated, leaving readers to infer it and inviting misinterpretation.[10]



The *estimand* is a precise statement of the treatment effect that answers the trial's question; the *estimator* is the statistical method used to estimate it.

The estimands framework is particularly relevant to ePCTs because intervention delivery, adherence, concomitant care, provider changes, and health system events may alter the meaning or availability of an endpoint. These attributes should be treated as part of the scientific question rather than only as analytic complications.

Figure 2. The 5 Attributes of an Estimand

Attribute	What it makes explicit
1 Population	The individuals for whom the treatment effect is defined—not necessarily identical to the eligibility or analysis population.
2 Treatment conditions	The treatment versions being compared, including how discontinuation, switching, rescue therapy, and adherence are conceptualized.
3 Endpoint / outcome variable	The outcome, measurement scale, and time frame that answer the scientific question.
4 Intercurrent-event strategy	How each post-baseline event that changes the endpoint's interpretation or existence is handled.
5 Population-level summary measure	How outcomes are summarized and contrasted across treatment conditions (for example, a mean difference, risk ratio, or odds ratio).

The framework guides investigators to start with the question, then choose the analysis:

1. **Specify the estimand**. For each outcome, write down all 5 attributes.
2. **Choose an analysis**. Align the design, trial conduct, data collection, and estimator.
3. **Check assumptions**. State assumptions and assess robustness with sensitivity analyses. Before analysis, confirm the estimator targets the prespecified estimand.

Strategies for Intercurrent Events

There are several strategies for addressing intercurrent events, and selection involves a tradeoff between relevance and reliability. A highly relevant strategy may require strong assumptions, while a more reliable approach may answer a less central question. Research teams should make this tradeoff explicit.

- **Treatment policy:** evaluate treatment assignment regardless of the event, consistent with a familiar intention-to-treat perspective.
- **Composite:** incorporate the intercurrent event into the endpoint definition.
- **While alive or while on treatment:** use endpoint information observed before the event.
- **Hypothetical:** estimate the effect in a scenario where the event would not have occurred.
- **Principal stratum:** target a subgroup defined by potential values of the intercurrent event under the treatment conditions.

Extension to Cluster-Randomized Trials

For cluster-randomized trials, the estimands framework is modified in 3 areas. The study population includes both individuals and clusters, and intercurrent events may occur at either level. The population summary depends on how individuals and clusters are weighted. Finally, treatment may be delivered at either the cluster or individual level. These features require attention to the unit and level of inference.

Case Studies: APA-SM and ARBOR-Telehealth

Two NIH Collaboratory Trials reviewed their estimands (Table) and choices when defining their endpoints and analysis strategy. For APA-SM,[11] the team weighed momentary assessment vs 7-day recall, end-of-treatment change vs average effect over the intervention period, and a 2-timepoint comparison vs longitudinal modeling. The secondary estimand in ARBOR-Telehealth[12] raises a question central to ePCTs: when does a hypothetical strategy clarify the treatment effect, and when does it move too far from real-world practice?

Table. Estimand Considerations in 2 NIH Collaboratory Trials

	APA-SM	ARBOR-Telehealth
Population	Adults with chronic pain in rural areas	Adults with chronic low back pain in rural areas
Comparison	In-person acupressure training vs online training vs education control	Telehealth physical therapy vs online education modules
Primary endpoint	Change in pain impact score (7-day recall) from baseline to day 28	Change in Oswestry Disability Index from baseline to 12 weeks
Intercurrent events	New pain medications or treatments; treatment interruption or discontinuation; unrelated death	Incomplete treatment; additional pain treatments; change of therapist; handoff of intervention delivery to rural partner
Strategies	Chosen by event type: treatment policy and while on treatment; principal stratum as a possible sensitivity analysis	Primary: treatment policy (intervention as assigned) Secondary: hypothetical (effect if participants kept same therapist or completed all modules)

Ongoing Team Process

Clinical investigators, statisticians, programmers, and operational teams need a shared language and regular communication throughout trial design, startup, and conduct. Data meetings can help identify emerging events, determine whether additional information is needed, and assess whether the planned analysis still targets the intended treatment effect.

Key Takeaways

- Start with the scientific question rather than the preferred statistical method.
- State assumptions and assess them with sensitivity analyses.
- Collect information on treatment completion, concomitant care, provider continuity, and system changes when those factors may affect interpretation.
- Discuss estimand attributes throughout the trial rather than treating them as fixed.

Working With Multiple Outcomes: Multicomponent Endpoints and Beyond

Panelists: Julie Fritz, Patrick Heagerty, Eric Roseen, Pamela Shaw, Qilu Yu; Moderator: Emily O'Brien

This session examined options beyond a single primary outcome when a condition or intervention affects multiple domains.

The Challenge in Pain Trials

Pain measurement has expanded from severity scales to multidimensional instruments that capture function, mood, and other domains. Initiatives such as IMMPACT, CONSORT, and PROMIS have promoted standardized core outcome measures. However, there is a tension between the multidimensional nature of the condition and specification of a single primary outcome.

A 0 to 10 pain intensity scale remains a common choice because it is low burden, frequently available in clinical practice and EHRs, familiar, and often expected for regulatory purposes. Trials can pair it with secondary measures of function, pain interference with activity, or quality of life, or they can combine domains into a single endpoint.

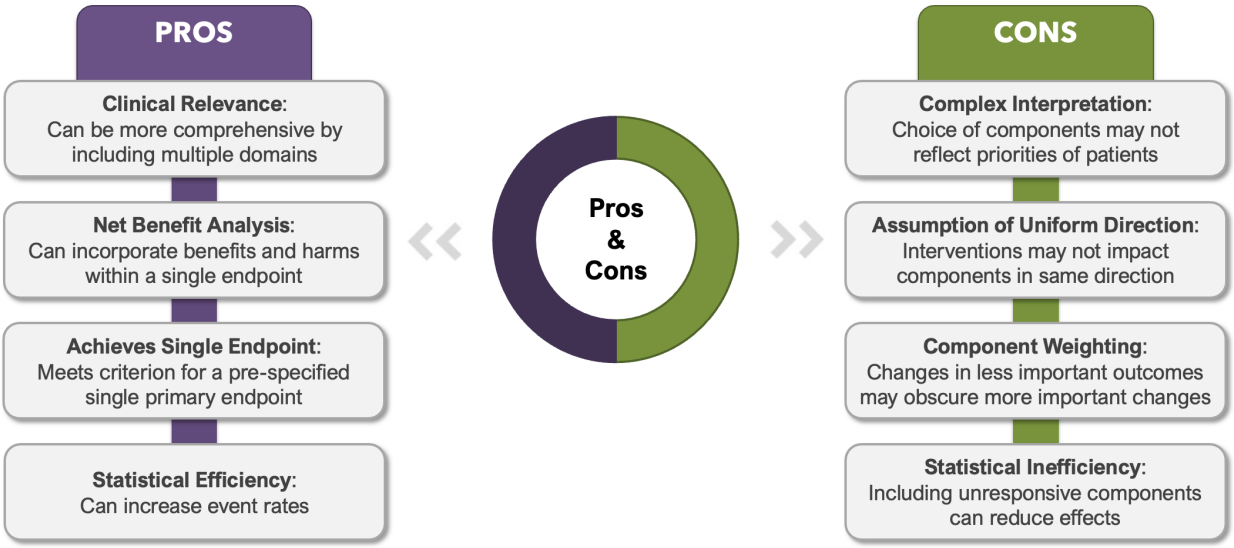
Composite Endpoints

A composite endpoint combines several measurements into one primary endpoint, with benefits and drawbacks (Figure 3).



Present individual component results even when the primary endpoint is a composite, while acknowledging that some components may be underpowered.

Figure 3. Pros and Cons of Composite Endpoints in Pain Trials



Hierarchical Endpoints: A Special Form of Composites

A hierarchical composite ranks components by clinical priority instead of weighting them equally, which may better reflect real-world values of clinicians and patients. These endpoints are well-established in cardiovascular, critical care, oncology, and infectious disease trials, but remain rare in pain trials.[13]

In the standard approach, each participant in the intervention group is compared with each participant in the control group, starting with the highest-priority outcome. The participant with the better outcome is counted as a win for their group, and a loss for the other group. If the pair ties, the comparison moves to the next outcome, continuing until one participant wins or all outcomes tie. For quantitative outcomes, investigators must pre-specify what difference counts as a "win." Results are summarized as the win ratio (wins divided by losses) or as the net treatment benefit (the proportion of wins minus the proportion of losses).[14] A related approach, the desirability of outcome ranking (DOOR), orders combinations of benefit and harm on a single ordinal scale. Creating that ranking requires explicit judgments about how to weigh benefit against harm, so interpreting those results for clinical decisions depends on whether those judgments are accepted.

Outcome order is a scientific and value choice. Researchers should consult users of the evidence (eg, patients, clinicians) to ensure their values are reflected in the choices. A reanalysis of an appendicitis trial showed that different prioritizations could be chosen which favored either appendectomy or antibiotics.[15] Investigators should prespecify a primary hierarchy and can later test how findings change with other orders.

Case Study: PACBACK

PACBACK was a large effectiveness trial of clinician-supported self-management, spinal manipulation, or their combination, for acute low back pain.[16] Dr. Roseen proposed a reanalysis using a hierarchical composite endpoint and asked workshop participants to

select the 2 most important outcomes and consider their order. High-impact chronic pain and average back-related function over 1 year received the most support, yet every outcome received votes, illustrating that people's priorities vary.



Selecting outcomes for pain research presents challenges because one patient may prioritize sleep, while another prioritizes work, pain intensity, or function.

Panelists considered whether it would be possible to allow each patient to define success according to their preference. Such an approach would be complex, and a global improvement measure may address the same goal more directly.

Key Takeaways

- Consult with users of the evidence (eg, patients, clinicians) when creating a composite endpoint so their values are reflected in the choices.
- Treat composite and hierarchical methods as optional tools, not automatic replacements for a single endpoint.
- Prespecify the primary strategy and plan sensitivity analyses using reasonable alternative thresholds or hierarchies.
- Communicate the overall result in plain language and explain what it represents.

Moving From Measurement to Meaning

Speaker: Kevin Weinfurt, PhD, Duke University

Endpoint strategy involves values, tradeoffs, and interpretation. Consider:

- **Meaningful to whom?** Health system leaders, clinicians, and patients may value different effects.
- **Meaningful how?** A change may be noticeable in daily life, or it may be worthwhile after considering cost, burden, risk, and alternatives.
- **Meaningful at what level?** A shift observed in a population may not describe what an individual patient experiences, and vice versa.
- **How do we deal with uncertainty?** An effect estimate has a confidence interval, and an average effect does not reveal the distribution of individual effects. Patients may differ in the amount of change they consider meaningful.

Unpacking Uncertainty, Meaningfulness, and Magnitude

Panelists: Devon Check, Lynn DeBar, Juan Marcos Gonzalez, Laura Lee Johnson, Natalia Morone; Moderator: Greg Simon

Dr. Simon emphasized that the primary "customers" for ePCTs are not funders, journal editors, or other researchers. Those audiences remain important, but meaningfulness should be anchored to the people making real-world decisions:

- **Patients and clinicians:** They use evidence to make decisions about care for an individual patient.
- **Health system leaders:** They decide whether programs or services should be adopted, sustained, or de-implemented.
- **Regulators:** They may make decisions about approval, labeling, or payment for products, programs, or services.



Meaningfulness depends on who will use the evidence, what decision they face, what benefits and burdens they value, and whether the measured change is reliable, implementable, and worthwhile in context.

Case Study: OPTIMUM

OPTIMUM[17] evaluated group-based mindfulness for chronic low back pain. The team chose PEG as the primary outcome because it is validated, widely used, and practical in clinical care. They supplemented it with PROMIS-29, patient global impression of change, qualitative, safety, and implementation data. A 1-point change in PEG is commonly treated as meaningful within an individual. OPTIMUM was powered for a 1-point between-group difference. The intervention group improved by 1.21 points at 6 months and 1.52 points at 12 months, while the usual care group did not reach a 1-point change. The between-group difference was statistically significant but did not reach 1 point.

More intervention participants rated themselves as minimally, much, or very much improved. Dr. Morone argued for a broader definition of meaningfulness that includes patient-perceived improvement and ease of integrating the intervention into clinical care.

Case Study: 3 Pain Trials

Dr. DeBar described 3 pain ePCTs. In PPACT[18], health system leaders and payers sought support for primary care teams managing difficult chronic pain cases. The cognitive behavioral intervention reduced PEG scores through 1 year, improved patient satisfaction, and reduced service use. Even so, sustainment was difficult because the health system was hesitant to bear implementation burden. Opioid doses decreased overall but did not differ between groups, which complicated interpretation amid concern about opioid tapering.

RESOLVE[19] adapted PPACT into virtual self-management and health coaching programs. Both improved pain outcomes compared with usual care. Coaching worked faster, but self-management caught up over time and produced a cost offset. Participants who began with self-management also found long-term maintenance easier.

BackInAction[20], prompted by funder and payer interest, found that acupuncture produced greater, clinically important improvement in back pain–related disability in older adults and was cost effective.



Compare effect magnitude with available pharmacologic and nonpharmacologic alternatives rather than evaluating it in isolation.

Patient Preferences

Value extends beyond whether an effect exists (Figure 4). Trial design involves judgments about significance, context, heterogeneity, and acceptable tradeoffs. Those decisions shape the information later used by patients, clinicians, administrators, and others.

Figure 4. From Detection to Value



Evidence must be translated for an individual. Decision-makers judge whether the evidence is strong enough, whether the study context applies to the individual, and whether the expected net clinical benefit is positive considering risks, burdens, and alternatives.

People assign value differently because their lives and priorities differ. Preference research systematically presents scenarios to estimate how much risk, cost, or burden someone is willing to accept for a benefit. This information makes tradeoffs more transparent and easier to combine with clinical data.

People also extrapolate trial evidence beyond its context, drawing on other information. Narrower trials may improve precision but do not prevent extrapolation, whereas ePCTs may be particularly useful for understanding real-world consequences.

Clinical evidence does not interpret itself. It becomes meaningful only after someone judges what matters, for whom, and under what conditions.

Health System Perspectives

Implementation determinant frameworks such as CFIR[21] organize factors that influence adoption, and evidence strength is only one of them. Treatment-effect magnitude must be weighed alongside cost, complexity, fit, organizational conditions, and the people expected to carry out the change.

Cost-effectiveness and cost-utility analyses translate treatment effects into measures such as cost per case prevented or cost per quality-adjusted life year, helping systems weigh benefits relative to resources. However, positive economic results do not guarantee sustainment. PPACT illustrated the “wrong pocket” problem: the entity paying to deliver an intervention may not be the one that realizes the savings.

A discrete choice experiment found that healthcare professionals, leaders, and policymakers were often willing to pay more for interventions that addressed a recognized

clinician or patient need and that organizations were ready to implement.[22] Trial teams can measure related implementation constructs:

- **Appropriateness:** Does it fit our clinical context and needs?
- **Acceptability:** Do clinicians and staff like it?
- **Feasibility:** Can we realistically implement it?



Health systems may need to decide about sustainment before robust cost findings are available.

Regulatory Interpretation

Dr. Johnson distinguished *outputs* (ie, the measurable things produced by a trial) from *outcomes* (ie, the reasons those changes matter). A score change is an output; an outcome explains why that change affects how a patient feels, functions, survives, or uses care.

Work should not stop at a number and P value. Regulators may triangulate evidence from interviews, clinical outcome assessments, and laboratory findings. In one example, participant transcripts helped develop non-misleading label language. However, qualitative evidence must be collected rigorously and systematically.

Magnitude also depends on perspective. For alcohol treatment, a 2-level reduction in World Health Organization risk drinking level was accepted as a surrogate primary efficacy endpoint. Cutting consumption from 24 to 12 beers per day may be substantial personal progress yet still carry high risk and be insufficient from a societal perspective.

Similarly, PEG is useful clinically and may be required by some payers, but it would not usually serve the same purpose for drug evaluation, where decisions consider efficacy or effectiveness together with safety, risks, burdens, and feasibility of use. Well-designed preference studies can clarify how patients, caregivers, and providers weigh these elements and how much change is necessary before an intervention becomes worthwhile.

Precise measurement is not valuable if it resolves differences in a range that does not matter. Conversely, a small population shift can be important if it substantially changes resource use or public outcomes, even when individual patients do not notice the change. Investigators should decide in advance how endpoint magnitude will be interpreted for relevant groups and explain why the result matters, rather than expecting audiences to care about an output because it is statistically significant.

Key Takeaways

- Identify the customer and decision before defining meaningfulness.
- Use rigorous preference methods to quantify benefit-burden tradeoffs.
- Allow different measures to serve clinical, health system, and regulatory purposes.
- Plan qualitative, feasibility, and economic analyses early to inform implementation decisions.

Toward Best Practices for Interpretation of Endpoints

Panelists: Andrea Cook, Corita Grudzen, Kari Stephens, Liz Turner, Christy Zigler; Moderators: Adrian Hernandez, Kevin Weinfurt

Panelists shared lessons from NIH Collaboratory Trials and offered guidance on planning choices that support precise estimation and meaningful interpretation.

NIH Collaboratory Trial Case Studies

BackInAction

In BackInAction, the target treatment effect was the adjusted between-group mean difference in change from baseline on the Roland-Morris Disability Questionnaire. The trial was powered to detect a 2-point difference. The observed difference fell short of that threshold but was statistically significant and comparable with effects reported for other pain interventions. The team later recognized that the 2-point threshold reflected meaningful change within an individual, not between randomized groups. The two are not equivalent because individual change has intrinsic variability so a measured individual difference must be larger than what could be explained by measurement error to be reliable, while group differences reflect the shift in the population and relatively small changes may lead to important overall population-level impact.

The trial could instead have used a binary responder endpoint, such as at least a 2-point difference or 30% reduction at the individual level. Binary outcomes may be easier to explain but generally require larger samples and discard information. Continuous outcomes are more efficient but less intuitive. Outliers, worsening in a subset of participants, or heterogeneous treatment effects may make responder analyses informative, and some pain trials have found stronger binary than continuous signals when only a subset of participants improved meaningfully.

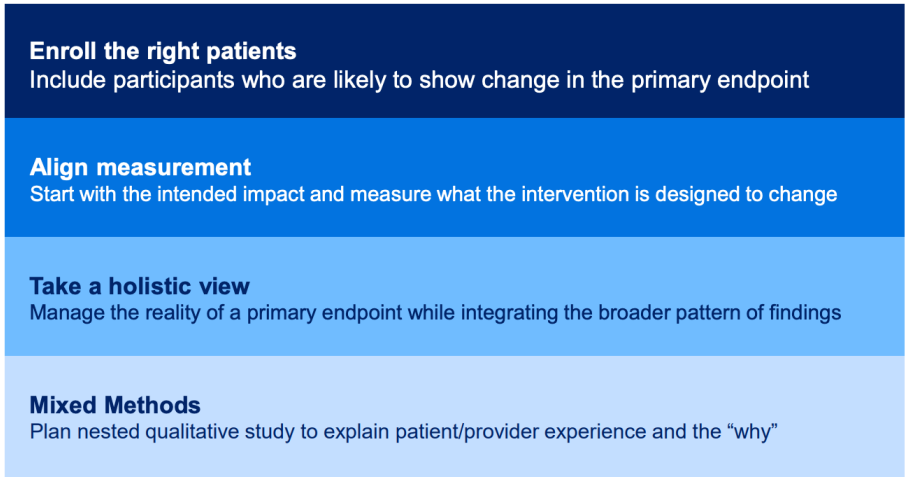
Dr. Cook recommended separating the effect used for power calculations from the range of effects that will be considered clinically meaningful, and specifying both prospectively in the statistical analysis plan.

PRIM-ER

Drawing on PRIM-ER[23] and other trials, Dr. Grudzen shared best practices for ePCTs (Figure 5). She recommended that teams envision drafting the discussion section of their publication for every plausible result before finalizing the trial design. This exercise can reveal whether the hypothesis and design match the decision the trial is meant to inform. For example, a noninferiority design may be more appropriate than superiority when a low-resource intervention would be valuable if it performs nearly as well as standard care.

She also cautioned against eligibility criteria so broad that many participants have little room to improve on the primary endpoint. Complex interventions should stay centered on their target endpoint, so that worthwhile but peripheral activities do not displace it.

Figure 5. Best Practices for ePCTs



EquiP PC

Dr. Stephens described a World Café exercise her team used to examine signal loss from multiple perspectives.[24] The process aligned the team around the trial's aims and quickly generated mitigation strategies. EquiP PC also benefited from a community advisory board, measurement of both practice-level and patient outcomes, and mixed methods.

Building a Valid Interpretation

Dr. Zigler described valid interpretation as resting on a strong foundation: a meaningful outcome, a clear relationship to the intervention, a valid measure and data collection scheme, and an analysis plan aligned with the question and endpoint. Interpreting quantitative scores also requires context-specific evidence about what constitutes meaningful change.

Interpretation occurs after data collection and analysis, but its validity depends on earlier choices.

Because large ePCTs can produce very small P values for modest effects, interpretation requires evidence about clinical meaning, not statistical significance alone. Established measures such as PEG and PROMIS have published interpretive evidence, but those thresholds may not transfer to every population or context, and new measures may need empirical work. The FDA's Patient-Focused Drug-Development draft guidance 4 offers methods for engaging patients and communities to establish meaningful differences.[25] Multicomponent endpoints need particular interpretive support, because their construction can obscure which component drove the result.

Design and Analysis Choices

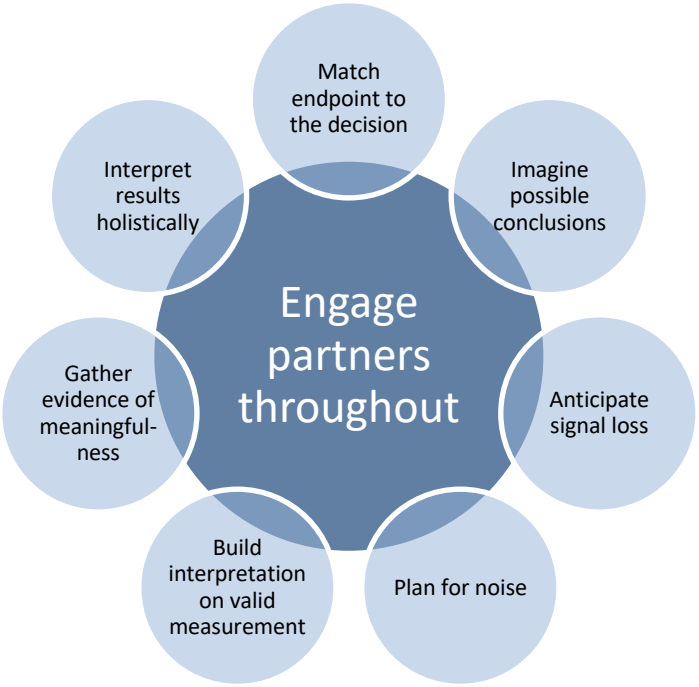
Dr. Turner offered guidance for design and analysis choices that support informed, rigorous interpretation:

- **Select the endpoint** as a team after considering who will use the evidence.
- **Select the target treatment effect** by powering for what the team does not want to miss, documenting it in the statistical analysis plan, and revisiting it before analysis in case standards have changed.
- **"Select" the expected noise** by anticipating sources of variance, accounting for them in the design and analysis, and monitoring variability during the trial.
- **Select (and adapt) the analysis plan** using appropriate covariate adjustment to improve precision, account for clustering and other design features, address missing data, and include sensitivity analyses. Revisit and finalize plans before unmasked analysis in case of methodological advances.
- **Interpret the totality of evidence** beyond the primary endpoint, including cost-effectiveness, subgroup effects, implementation, and acceptability.

Key Takeaways

The moderators constructed a mind map of the lessons from the panel, a simplified version of which is shown in Figure 6.

Figure 6. Best Practices for Interpretation of Endpoints



References

1. Califf RM, Sugarman J. Exploring the ethical and regulatory issues in pragmatic clinical trials. *Clin Trials Lond Engl*. 2015;12(5):436-441. doi:10.1177/1740774515598334
2. European Medicines Agency. ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. Published online 2020. Accessed September 25, 2026. https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e9-r1-addendum-estimands-sensitivity-analysis-clinical-trials-guideline-statistical-principles_en.pdf
3. Kahan BC, Bretz F, Copas A, et al. CRT-Estimands Framework: consensus based extension of the ICH E9(R1) addendum for cluster randomised trials. *BMJ*. 2026;393:e089050. doi:10.1136/bmj-2025-089050
4. Lakin JR, Brannen EN, Tulsky JA, et al. Advance Care Planning: Promoting Effective and Aligned Communication in the Elderly (ACP-PEACE): the study protocol for a pragmatic stepped-wedge trial of older patients with cancer. *BMJ Open*. 2020;10(7):e040999. doi:10.1136/bmjopen-2020-040999
5. Volandes AE, Chang Y, Lakin JR, et al. An intervention to increase advance care planning among older adults with advanced cancer: a randomized clinical trial. *JAMA Netw Open*. 2025;8(5):e259150. doi:10.1001/jamanetworkopen.2025.9150
6. Vazquez MA, Oliver G, Amarasingham R, et al. Pragmatic trial of hospitalization rate in chronic kidney disease. *N Engl J Med*. 2024;390(13):1196-1206. doi:10.1056/NEJMoa2311708
7. Glasgow RE, Knoepke CE, Magid D, et al. The NUDGE trial pragmatic trial to enhance cardiovascular medication adherence: study protocol for a randomized controlled trial. *Trials*. 2021;22(1):528. doi:10.1186/s13063-021-05453-9
8. Ho PM, Glorioso TJ, Allen LA, et al. Personalized patient data and behavioral nudges to improve adherence to chronic cardiovascular medications: a randomized pragmatic trial. *JAMA*. 2025;333(1):49-59. doi:10.1001/jama.2024.21739
9. Lu AD, Chavez C, Daugherty SL, et al. Using artificially intelligent text messaging to improve health behaviors to reduce cardiovascular risk: study protocol for a pragmatic randomized control trial. *Trials*. 2026;27(1):244. doi:10.1186/s13063-026-09564-z
10. Cro S, Kahan BC, Rehal S, et al. Evaluating how clear the questions being investigated in randomised trials are: systematic review of estimands. *BMJ*. 2022;378:e070146. doi:10.1136/bmj-2022-070146
11. Kawi J, Bolin J, Wu H, et al. Auricular Point Acupressure Self-Management (APA-SM) program for chronic musculoskeletal pain among rural populations: a protocol for a pragmatic, randomized controlled trial. *BMJ Open*. 2026;16(6):e117632. doi:10.1136/bmjopen-2026-117632
12. Skolasky RL, Colantuoni E, Wegener ST, Ali KJ, McLaughlin KH. ARBOR-Telehealth study: an examination of telerehabilitation to improve function and reduce opioid use in persons with chronic low back pain in rural communities – protocol of a pragmatic, individually randomised group treatment trial. *BMJ Open*. 2025;15(6):e102773. doi:10.1136/bmjopen-2025-102773
13. Nishtar M, Mark R, Langford DJ, et al. Evaluating the balance of benefits and harms in chronic pain clinical trials: prioritizing individual participants over individual outcomes. *Reg Anesth Pain Med*. 2024;49(5):363-367. doi:10.1136/rapm-2023-104809

14. Pocock SJ, Gregson J, Collier TJ, Ferreira JP, Stone GW. The win ratio in cardiology trials: lessons learnt, new developments, and wise future use. *Eur Heart J*. 2024;45(44):4684-4699. doi:10.1093/eurheartj/ehae647

15. Salvaggio S, Monsell SE, Heagerty PJ, et al. Generalized pairwise comparisons to support shared decision-making in the CODA trial. *JAMA Netw Open*. 2025;8(3):e252484. doi:10.1001/jamanetworkopen.2025.2484

16. Bronfort G, Meier EN, Leininger B, et al. Spinal manipulation and clinician-supported self-management for preventing chronic low back pain impact: the PACBACK randomized clinical trial. *JAMA Intern Med*. 2026;186(8):952-963. doi:10.1001/jamainternmed.2026.1893

17. Morone NE, Faurot KR, Weinberg J, et al. Mindfulness-based group medical visits for persons with chronic low back pain: a randomized clinical trial. *JAMA Intern Med*. 2026;186(9):1090. doi:10.1001/jamainternmed.2026.2186

18. DeBar L, Mayhew M, Benes L, et al. A primary care–based cognitive behavioral therapy intervention for long-term opioid users with chronic pain: a randomized pragmatic trial. *Ann Intern Med*. 2022;175(1):46-55. doi:10.7326/M21-1436

19. Mayhew M, Balderson BH, Cook AJ, et al. Comparing the clinical and cost-effectiveness of remote (telehealth and online) cognitive behavioral therapy-based treatments for high-impact chronic pain relative to usual care: study protocol for the RESOLVE multisite randomized control trial. *Trials*. 2023;24(1):196. doi:10.1186/s13063-023-07165-8

20. DeBar LL, Justice M, Avins AL, et al. Acupuncture for chronic low back pain in older adults: design and protocol for the BackInAction pragmatic clinical trial. *Contemp Clin Trials*. 2023;128:107166. doi:10.1016/j.cct.2023.107166

21. Keith RE, Crosson JC, O'Malley AS, Cromp D, Taylor EF. Using the Consolidated Framework for Implementation Research (CFIR) to produce actionable findings: a rapid-cycle evaluation approach to improving implementation. *Implement Sci IS*. 2017;12(1):15. doi:10.1186/s13012-017-0550-7

22. Meredith L, Thai T, Ong JJ, Grove A, Skouteris H, Higgins AM. Factors influencing implementation decisions in health care: A discrete choice experiment of intervention adoption. *Healthcare*. 2026;14(2):100786. doi:10.1016/j.hjdsi.2026.100786

23. PRIM-ER Investigators, Chung FR, Turecamo S, Cuthel AM, Grudzen CR. Effectiveness and reach of the Primary Palliative Care for Emergency Medicine (PRIM-ER) pilot study: a qualitative analysis. *J Gen Intern Med*. 2021;36(2):296-304. doi:10.1007/s11606-020-06302-2

24. Rogers KS. *The World Café: Social Innovation in Action*. Fielding Press; 2026.

25. Center for Drug Evaluation and Research. FDA Patient-Focused Drug Development Guidance Series for Enhancing the Incorporation of the Patient's Voice in Medical Product Development and Regulatory Decision Making. October 23, 2025. Accessed September 25, 2026. https://www.fda.gov/drugs/development-approval-process-drugs/fda-patient-focused-drug-development-guidance-series-enhancing-incorporation-patients-voice-medical

The complete workshop materials, including slides and recordings, are available on the Living Textbook.