Toward Best Practices for Interpretation of Endpoints

Moderators:
Adrian Hernandez, MD, MHS
Kevin Weinfurt, PhD
Duke University



NIH PRAGMATIC TRIALS COLLABORATORY
Rethinking Clinical Trials®

We Will…

- Identify best practices for specifying effect sizes for sample size estimation.
- Review lessons from NIH Collaboratory Trials in various settings (including under-resourced) with implications for future statistical analysis plans.
- Describe analysis pitfalls that can lead to misinterpretation of treatment effects.
- Discuss the potential for a more holistic interpretation across multiple endpoints.

NIH PRAGMATIC TRIALS COLLABORATORY
Rethinking Clinical Trials®

Closing Lineup



Andrea Cook
Kaiser Permanente Washington Health Research Institute



Corita Grudzen
Memorial Sloan Kettering Cancer Center



Kari Stephens
University of Washington



Liz Turner
Duke University



Christy Zigler
University of Pittsburgh

NIH Pragmatic Collaboratory Workshop

Andrea J Cook, PhD
Division of Biostatistics
Kaiser Permanente Washington Health Research Institute

Sept 9, 2026

BackInAction Overview

Pragmatic trial of acupuncture for chronic low back pain in older adults

Study aim

Determine whether acupuncture improves pain-related disability in adults age 65+ with chronic low back pain.

Also evaluates dose-dependence and safety in older adults.

Design

3-arm, individually randomized pragmatic trial

800 participants

Eligibility

Age ≥65 • EHR-documented uncomplicated chronic low back pain • PEG activity interference ≥3

Settings

KP Washington • KP Northern California • Sutter Health Northern California • IFH, NYC

Interventions

Standard acupuncture
15 sessions over 12 weeks

Enhanced acupuncture
6 additional maintenance sessions between 12-24 weeks

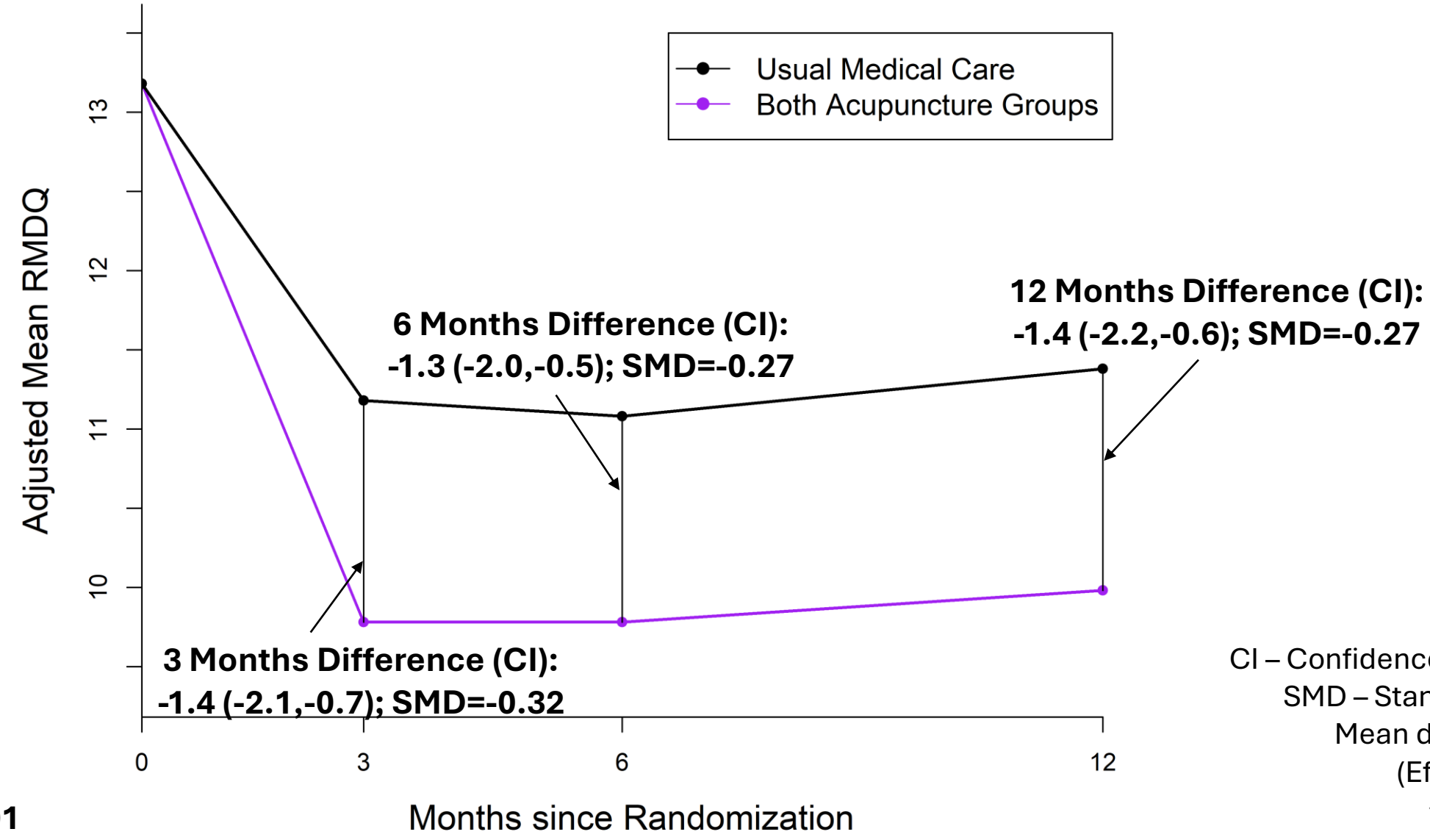
Usual medical care
Control arm

DeBar LL et al. JAMA. 2025; DeBar LL et al. Contemporary Clinical Trials. 2023; Nielsen A et al. Global Advances in Health and Medicine. 2021.

BackInAction Design/Power

- Primary Outcome: Continuous Roland Morris Disability Score (Scale from 0-24)
- Target Estimand: Adjusted Mean Difference in Change from Baseline
- Specified MCID: 2 point difference
  - Based on individuals change in score
  - Does not translate to group differences due to variability
  - Power for what you do not want to miss, but specify how to interpret the final finding of trial separately
  - Mean group differences: Standardized Mean Differences may be more understandable (0.2-0.5: small; 0.5-1.0 moderate; >01: large effects), but literature for pain outcomes show different thresholds for different outcomes (disability versus pain intensity) may be needed

Functional Disability: Acupuncture vs Usual Care



CI – Confidence Interval
SMD – Standardized Mean difference (Effect size)

All P-values < 0.001

Lessons Learned

- Continuous Outcome vs Binary Individual level MCID
  - Could have powered on binary individual-level response outcome 2-point difference or 30% improvement, but this does typically increase sample size needed for a trial
  - Interpretability vs power/sample size tradeoffs
  - Distribution of your underlying clinic population may help you choose
    - Do you expect large heterogeneity for treatment response or outliers may be better for binary outcome to detect between group differences
    - For BackInAction consistent across both, but other trials binary outcome has shown differences but not the continuous response
  - Be careful of intermingling MCID for individual vs group difference effect
- Powering for the effect you want to rule out (2 point was a 0.33 SMD) but prespecifying the interpretation of the range of effect estimates that are clinically meaningful (how are you going to interpret the findings a priori at the end of the trial)

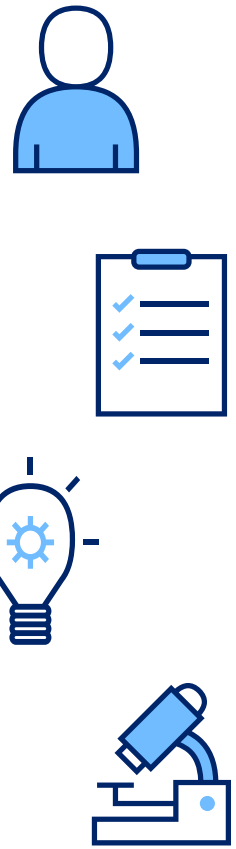
Designing Pragmatic Trials for Precise & Meaningful Treatment Effects

Corita R. Grudzen, MD, MSHS, FACEP
Division Head, Supportive and Acute Care Services
Fern Grayer Chair in Oncology Care and Patient Experience
Director, Center for Cancer Care Innovation
Memorial Sloan Kettering Cancer Center
Professor of Emergency Medicine
Weill Cornell Medical College



Memorial Sloan Kettering
Cancer Center

Best Practices for Pragmatic Trials



Enroll the right patients
Include participants who are likely to show change in the primary endpoint

Align measurement
Start with the intended impact and measure what the intervention is designed to change

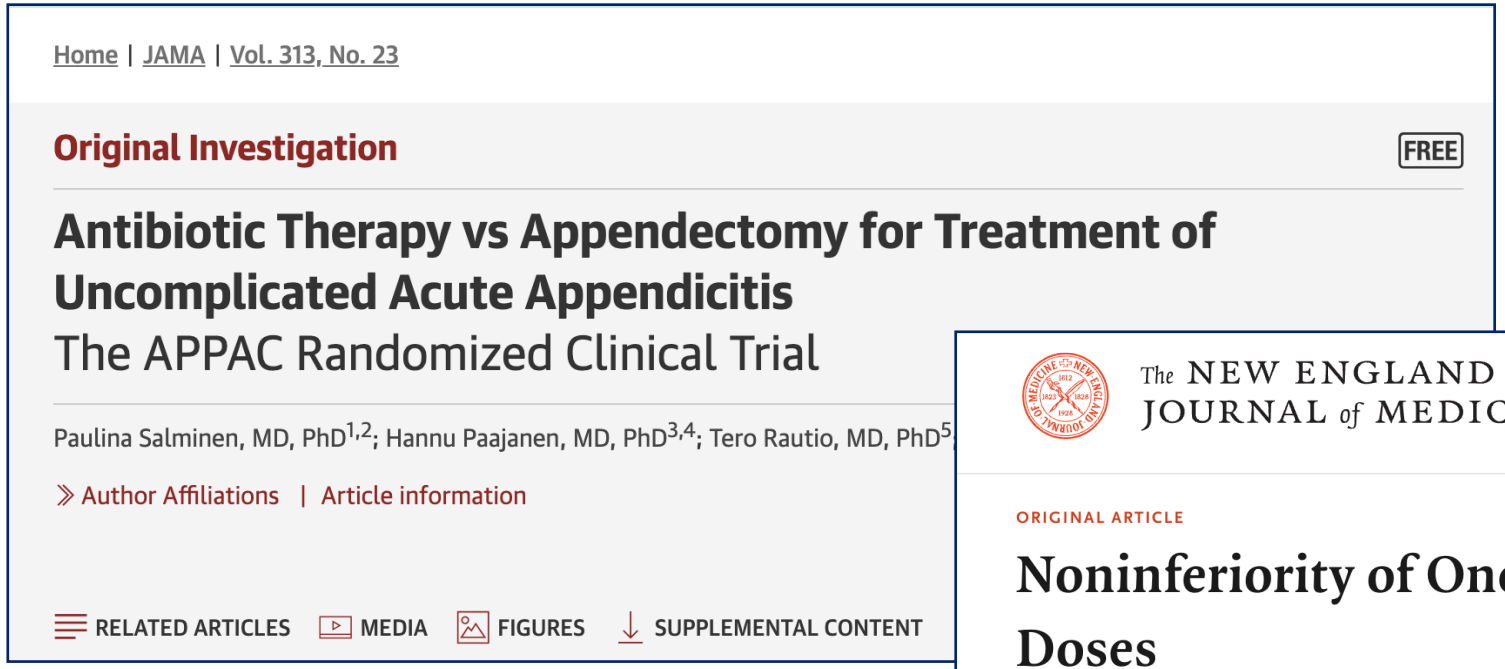
Take a holistic view
Manage the reality of a primary endpoint while integrating the broader pattern of findings

Mixed Methods
Plan nested qualitative study to explain patient/provider experience and the “why”

Source: Citation

© 2024 Memorial Sloan Kettering Cancer Center, et al. All rights reserved.

Begin with the end in mind: noninferiority or superiority



Home | JAMA | Vol. 313, No. 23

Original Investigation FREE

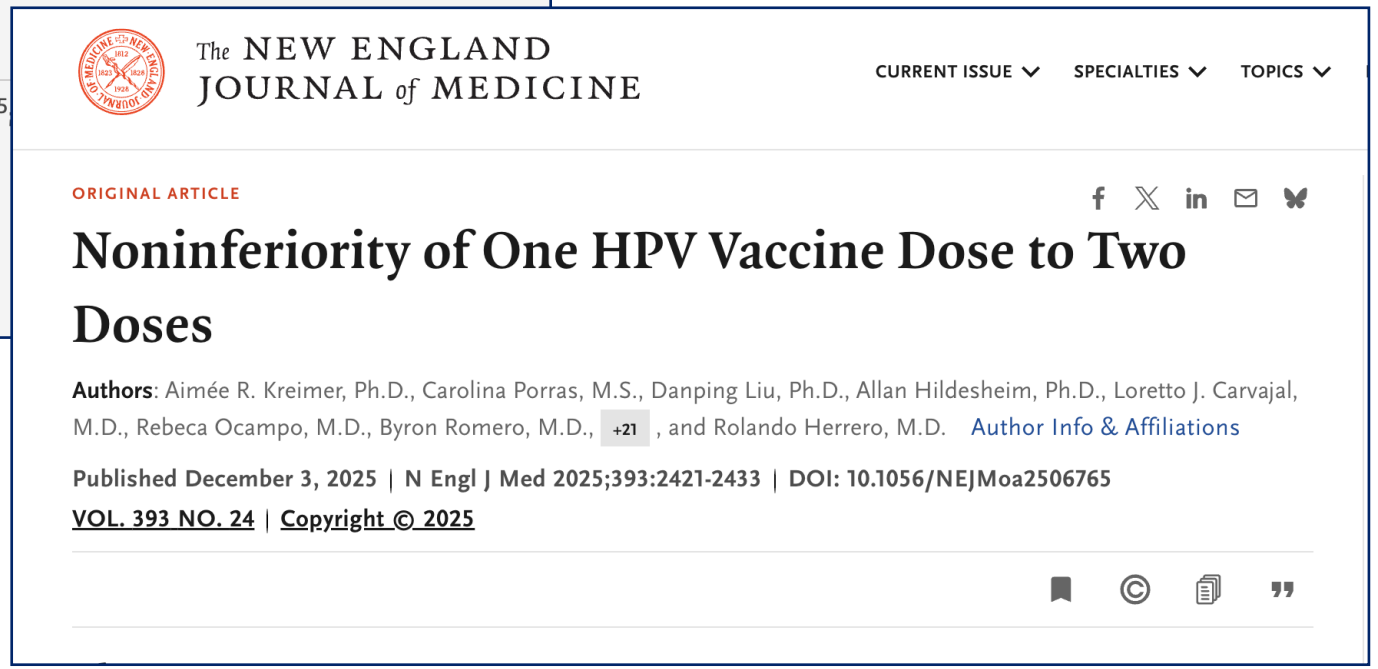
Antibiotic Therapy vs Appendectomy for Treatment of Uncomplicated Acute Appendicitis

The APPAC Randomized Clinical Trial

Paulina Salminen, MD, PhD[1,2]; Hannu Paajanen, MD, PhD[3,4]; Tero Rautio, MD, PhD[5]

» Author Affiliations | Article information

RELATED ARTICLES MEDIA FIGURES SUPPLEMENTAL CONTENT



The NEW ENGLAND JOURNAL of MEDICINE

CURRENT ISSUE SPECIALTIES TOPICS

ORIGINAL ARTICLE

Noninferiority of One HPV Vaccine Dose to Two Doses

Authors: Aimée R. Kreimer, Ph.D., Carolina Porras, M.S., Danping Liu, Ph.D., Allan Hildesheim, Ph.D., Loretto J. Carvajal, M.D., Rebeca Ocampo, M.D., Byron Romero, M.D., +21, and Rolando Herrero, M.D. Author Info & Affiliations

Published December 3, 2025 | N Engl J Med 2025;393:2421-2433 | DOI: 10.1056/NEJMoa2506765

VOL. 393 NO. 24 | Copyright © 2025

© 2024 Memorial Sloan Kettering Cancer Center, et al. All rights reserved.

Consider interpretation

JAMA Dermatology

RCT: Home- vs Office-Based Narrowband UV-B Phototherapy for Patients With Psoriasis

POPULATION

407 Males, 376 Females



Adults and children ≥12 y old with plaque or guttate psoriasis

Mean age, 48 y

INTERVENTION

783 Participants randomized



393 Home phototherapy
Home-based narrowband UV-B phototherapy 2-3 times per wk for 12 wk

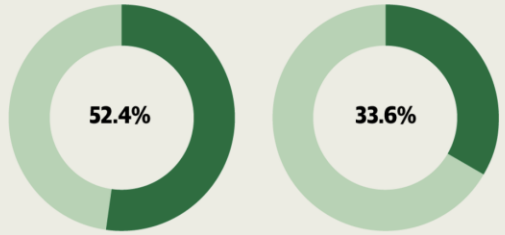


390 Office phototherapy
Routine office-based narrowband UV-B phototherapy 2-3 times per wk for 12 wk

FINDINGS

Home phototherapy was noninferior to routine office-based phototherapy for achieving PGA score of ≤1 and DLQI score of ≤5 at wk 12

DLQI score ≤5 at 12 wk



SETTINGS / LOCATIONS



42 US dermatology practices

PRIMARY OUTCOME

Proportion of participants achieving a Physician Global Assessment (PGA) score of ≤1 (clear or almost clear skin) or Dermatology Life Quality Index (DLQI) score of ≤5 (no to small impact on quality of life) at wk 12

Adjusted response differences between home vs office phototherapy

PGA score ≤1: 7.2 (95% CI, 0.8-13.5) percentage points; *P*<.001 for noninferiority
DLQI score ≤5: 18.6 (95% CI, 11.8-25.3) percentage points; *P*<.001 for noninferiority

© 2024 Memorial Sloan Kettering Cancer Center, et al. All rights reserved.

EquiP PC

Equitable Primary Care for Pain Care

Lessons Learned from Planning Endpoints

PI: Kari A. Stephens, PhD

Professor, Vice Chair of Research

UW Medicine

UNIVERSITY of WASHINGTON

FAMILY MEDICINE

World Café Activity – Signal Loss

- Team building activity during ***2-day hybrid all hands meeting***
- Goal: ***ID strategies to address potential signal loss***
- ***Split into 4 groups:*** 3 in-person, 1 online group
- Rotated between ***3 questions*** every 20 minutes
    1. Where do we anticipate potential signal loss?
    2. What methodological strategies could we use to address or mitigate signal loss?
    3. Are the currently planned data sources sufficient to answer our proposed research questions? If not, what additional data sources or data elements are needed?

Benefits: aligned team, identified ideas and strategies and next steps to plan endpoint data capture and methods

Community Advisory Board (CAB): Recruitment & Engagement Strategies

- ***Measured profiles*** across members by developing a framework and CAB profile survey to guide recruitment
- Recruited and onboarded ***members reflecting key community and stakeholder perspectives***
- ***Hybrid engagement model*** combining virtual meetings, pre-work activities, and asynchronous feedback opportunities
- ***Track CAB*** recommendations, decisions, actions taken, and report-backs through a structured feedback system
- ***Phased approach*** to allow for expansion during UH3 using profiles to aim recruitment towards key missing perspectives

Outcome Measures – Patient vs Practice

Study Goal: Expand access to high-quality, team-based chronic pain care in primary care, address barriers in access, demonstrate that adapted behavioral health integration [+ digital therapy tools] can improve patient outcomes

- **Primary outcome:** patient improvement in pain interference and intensity (PEG)
- **Secondary outcomes:** physical function, quality of life, the practice's integration of behavioral health

- Patient outcomes are more distal; Practice outcomes are more proximate
- Using a RE-AIM framework – ***need mixed methods approach***

Selecting endpoints is challenging!

Digital Therapeutics / Mobile Apps

No inventory to select from, few have FDA authorization

- ***Expanded from chronic pain*** (CP) apps to CP adjacent apps that address to meet clinician-patient need:
  - sleep and insomnia, anxiety, depression, panic disorder and PTSD, general wellness and mindfulness
- ***Developed app rating system*** across 8 dimensions: engagement, usability, life integration, implementation fit, regulatory status, reimbursement, symptom change, and evidence
- ***Dynamic endpoints***, given apps vary and need to partner with industry to share data

Towards Best Practices for Interpretation of Endpoints

NIH Collaboratory Workshop

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

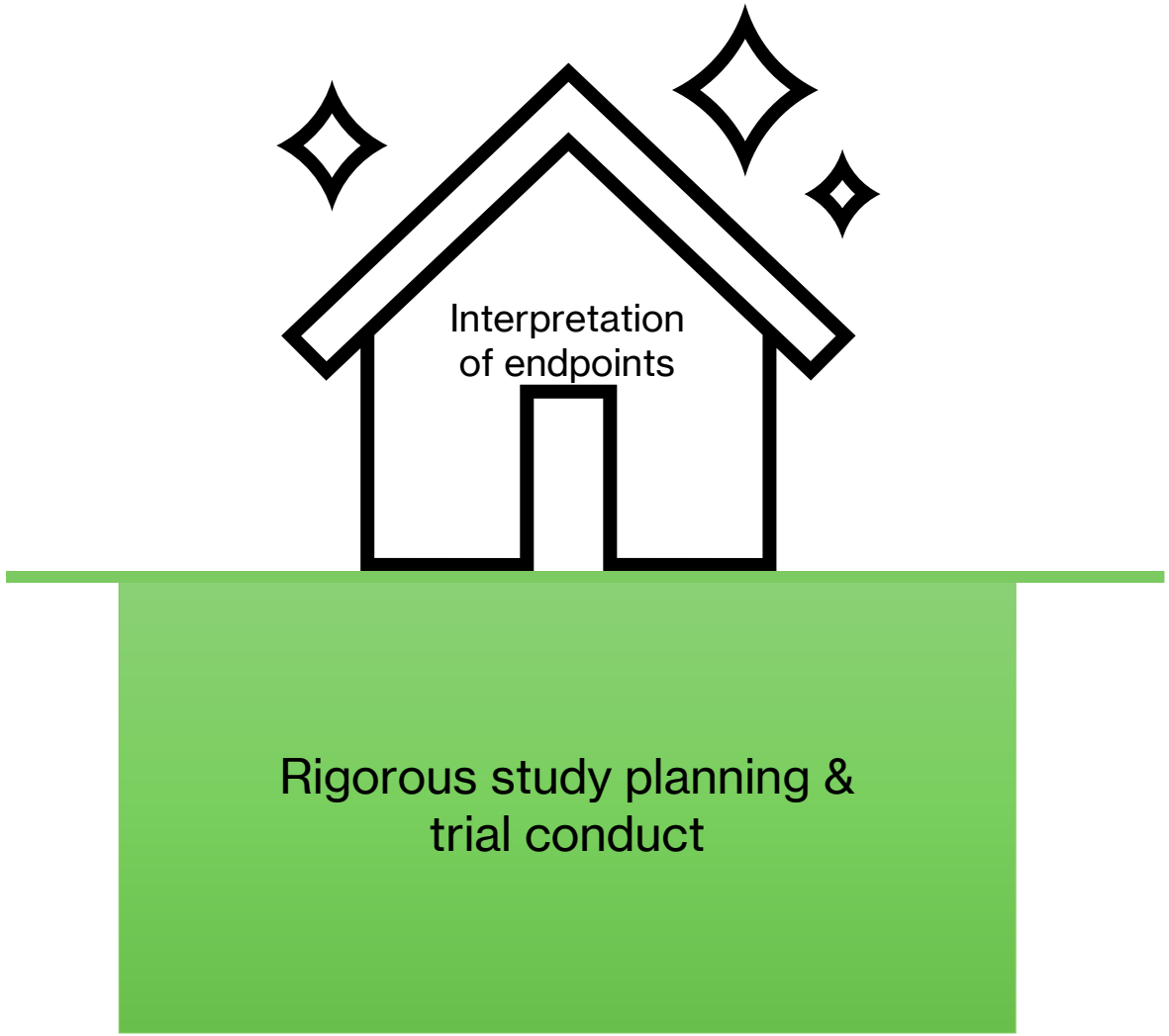
Christy K. Zigler, PhD, MSEd

Associate Professor | Physical Medicine & Rehabilitation | University of Pittsburgh School of Medicine

Director | Health Measurement Core | Division of Data Science and Rehabilitation Analytics

Co-Director | Patient Center Outcomes Core | NIH Pragmatic Trials Collaboratory

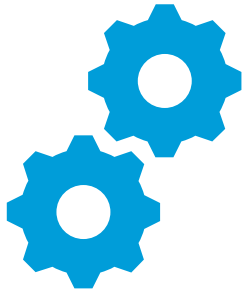
Email: cek53@pitt.edu


Interpretation
of endpoints
Rigorous study planning &
trial conduct

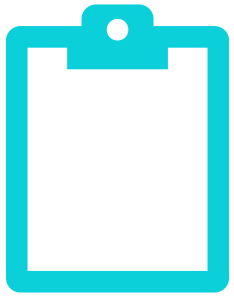
Rely on your *interpretation* ‘foundation’



Outcome should be meaningful and important



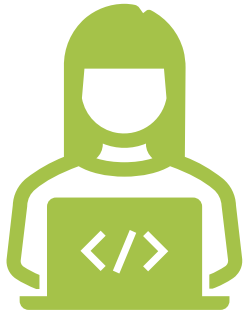
It should be related to your intervention (mechanism of action)



A valid way to measure the thing (e.g. an established clinical outcome assessment; COA)

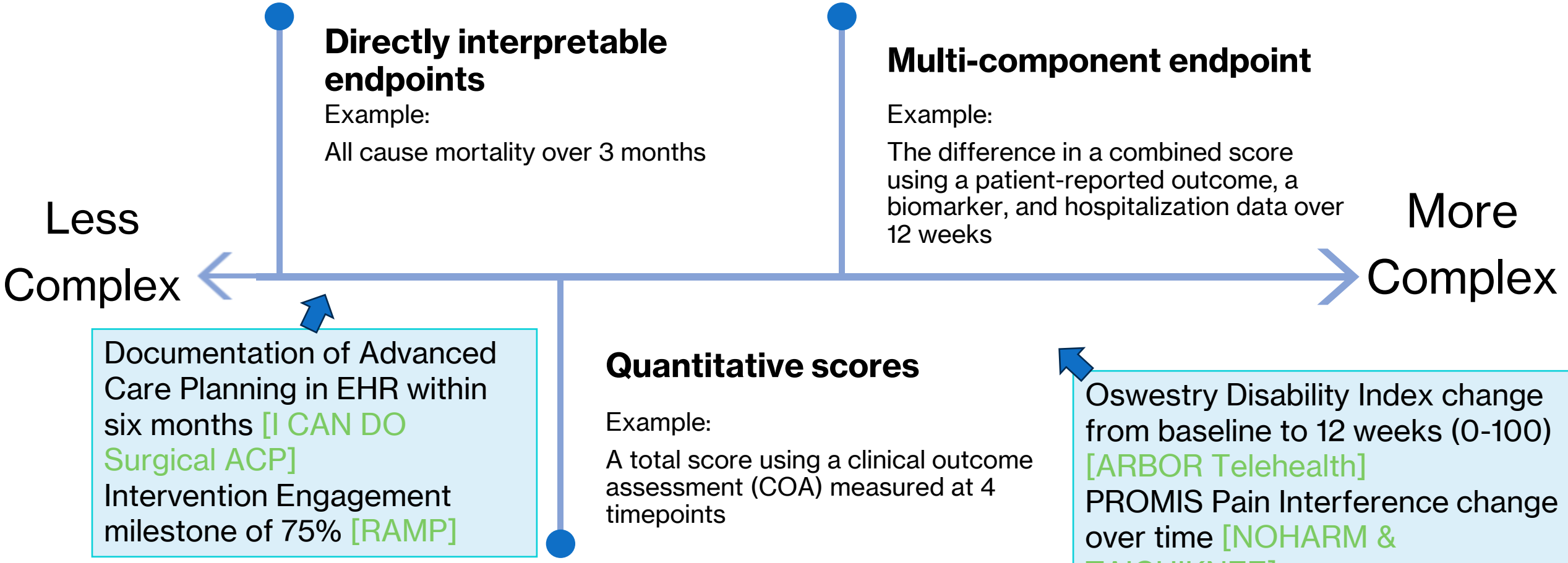


A valid data collection scheme that minimizes bias and missing data



A statistical analysis plan that matches the research questions and the way your endpoint is measured

Spectrum of Interpretability



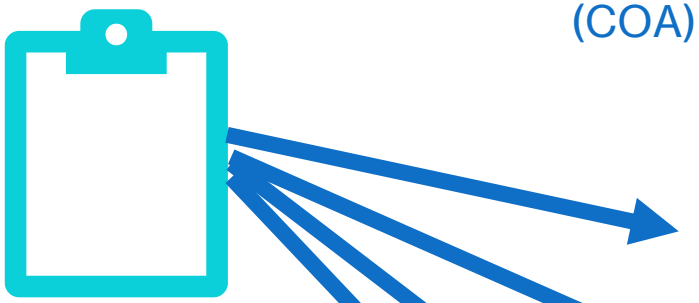
For Directly Interpretable Endpoints



A valid data collection scheme that minimizes bias and missing data

For Endpoints where interpretability is more complex:

Quantitative scores: Can be measured using clinical outcome assessments (COA) that have reference values that can be used to assist interpretation of scores within the pragmatic trial.



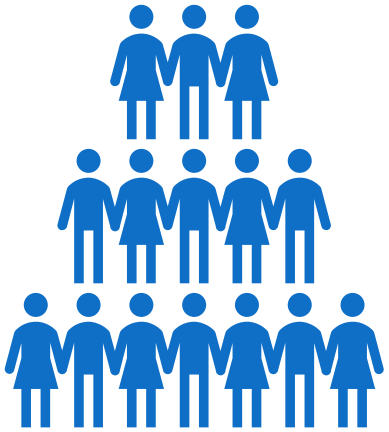
A valid way to measure the meaningful aspect of health

- Need to know possible range of scores
- Need to know what higher vs lower scores mean
- Normative data or meaningful score regions
- Meaningful score differences (MSD) / Minimal Clinically Important Differences (MCID)

Clinical versus statistical significance

Meaningful Change

- Meaningful to whom?



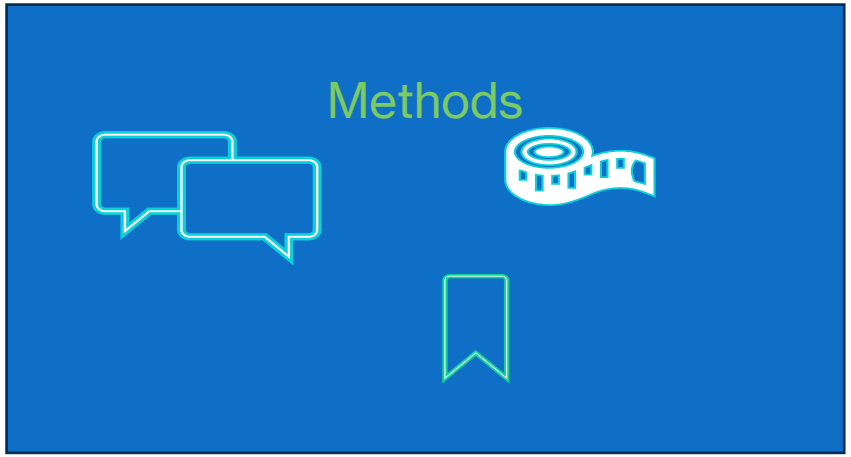


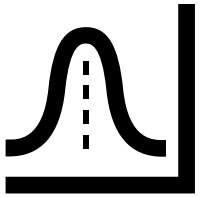
U.S. Food & Drug Administration
Patient-Focused Drug Development

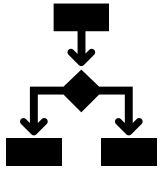
Meaningful => Patient's perspective

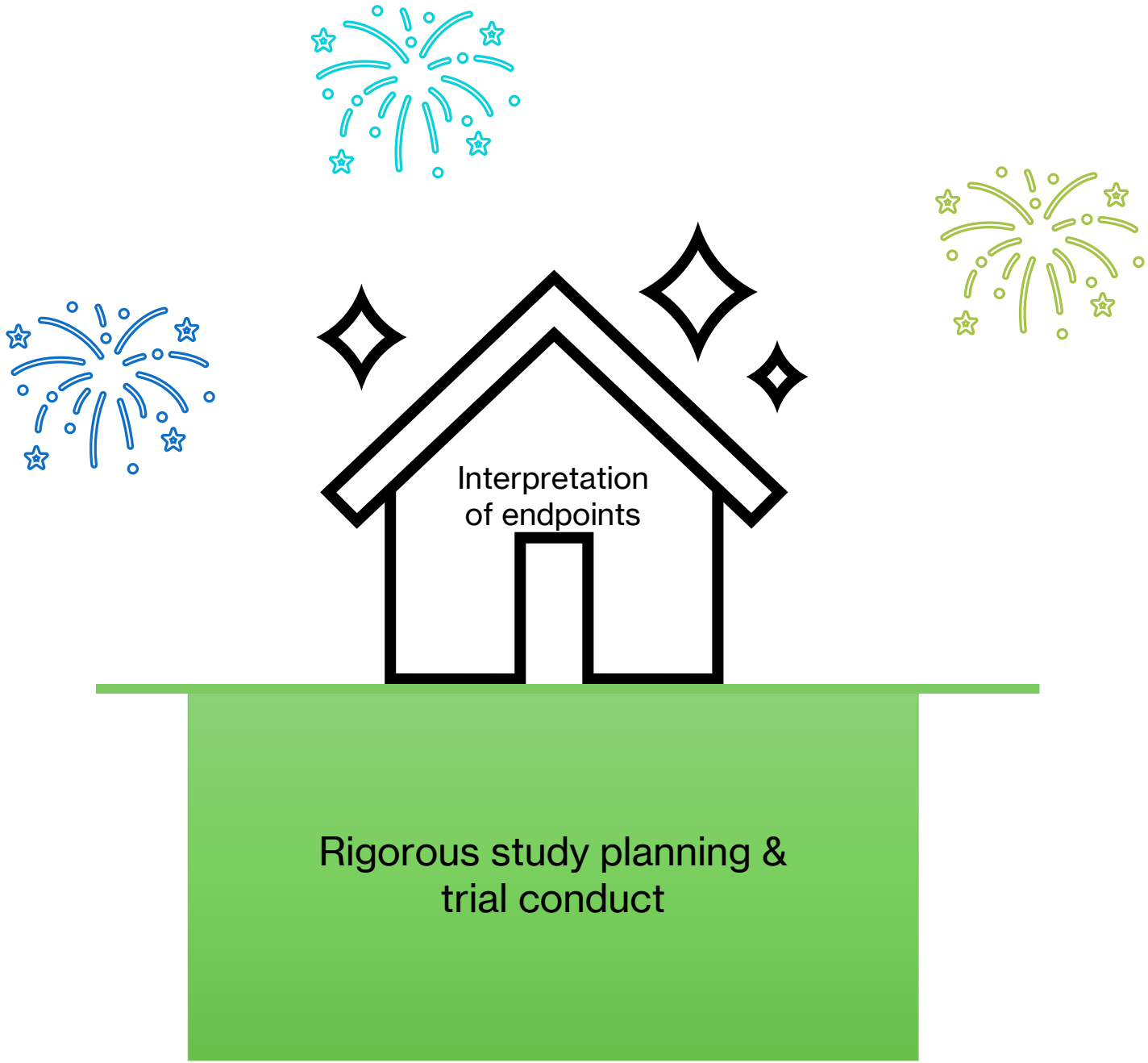


"Meaningful score difference"








Interpretation
of endpoints
Rigorous study planning &
trial conduct

Panel 5: Toward Best Practices for Interpretation of Endpoints

Reflections on Discussions of the Biostatistics Core

Liz Turner

Associate Professor of Biostatistics & Global Health, Duke University

Co-Lead, Biostatistics Core, NIH Pragmatic Trials Collaboratory

Director, Research Design and Analysis Core, Duke Global Health Institute

NIH PRAGMATIC TRIALS COLLABORATORY

Rethinking Clinical Trials®

1. Selecting endpoint of interest

- A team effort
- Invest time in decision-making process (cf. EquiP PC)
- What measure matters to whom?
  - Symptoms
  - Relative reduction from baseline (a “responder”)
  - Meeting a binary threshold of symptoms
- Proximal vs. distal measures

2. Selecting target treatment effect size

- Depends on perspective (cf. BackinAction)
- Power for what you don’t want to miss
- Before final analysis
  - Re-evaluate target magnitude & range of effects
  - Document in statistical analysis plan!

3. “Selecting” target noise

- Expected noise (as) important (as target effect size)
- Consider inclusion criteria (for sites & participants)
- Monitor “noise” (outcome variability) during the study
- (Cautiously) act on new info (e.g. FM-TIPS)

4. Selecting (and adapting) analysis plan

- Have a good plan for analysis
  - Covariate adjustment
  - Accounting for study design features (e.g. clustering)
  - Missing data
- Keep up to date on advances in statistical analysis
  - E.g., time-varying effects in SW-CRTs (Kenny et al, Stat in Med, 2022) indicated prior analyses may be flawed

5. So much more to interpreting treatment effects

- Context
- Costing & cost-effectiveness data
- Sub-group & heterogeneity of effects
- Fidelity & acceptability
- Implementation & process measures

Best Practices

- Choosing Endpoints
  - **Decision-making process**
    - Requires a team
    - Requires time
    - Transparency
  - Consider the audience(s): What measure matters to whom?
  - **Consider different participants' outcomes that might be relevant**
    - Patient-level
    - Practice-level
  - Deliberate attention to primary endpoint
    - Ensure that you enroll patients likely to show change in primary endpoint
    - Align measurement with intended effect on primary endpoint
  - Need a valid way to measure endpoint
    - Measure
    - **Data collection scheme**
- Consider signal-to-noise ratio
  - Increase signal-to-noise ratio
    - Anticipate potential sources of signal loss
    - Develop plans to mitigate signal loss
    - Inclusion criteria (sites and participants)
    - Plan on covariate adjustment
    - Consider how standardized intervention is (vs allowable variable)
  - Monitor "noise" (outcome variability)
    - Cautiously act on new info
- Targetting/ Interpreting Effects on Endpoints
  - Relevance of interpretive metrics (e.g., MCID, meaningful score differences)
    - Power for effect you DON'T WANT TO MISS
    - Specify a priori in SAP the range of effect estimates you regard as clinically meaningful
      - But re-evaluate later prior to unblinding
    - Some more complex or unfamiliar endpoints may require supportive studies to inform judgements of meaningfulness
    - Criteria of meaningfulness might depend
      - Decision-maker
      - Individual- vs Population-level
  - Holistic view
    - **Balance focus on primary endpoint and broader pattern of findings**
    - **Plan nested qualitative studies to enrich the picture (mixed-methods)**
    - Costing and cost-effectiveness
    - Sub-group and heterogeneity
    - Implementation and process
- Select and adapt analysis plan
  - **Align objectives, measures, analyses, and analytic plan (estimand framework)**
  - Covariate adjustment
  - Accounting for study design
  - Missing data
  - Monitor methodological developments and update as needed
- Imagine Discussion section under different hypothetical results (e.g., help distinguish superiority from non-inferiority)
- Work with a strong Community Advisory Board

Questions?

Please pose your questions in the Q&A

NIH PRAGMATIC TRIALS COLLABORATORY

Rethinking Clinical Trials®