Pragmatic Clinical Trial Challenges: Lessons Learned from the NIH Collaboratory Biostatistics and Design Core

Andrea J Cook, PhD
Scientific Investigator, Biostatistics Unit, KPWHRI
Affiliate Professor, Dept. of Biostatistics, University of Washington

Georgia CTSA BERD Research Forum
Morehouse School Medicine, Atlanta, GA
September 28, 2018

Faculty Disclosure

In compliance with ACCME Guidelines, I hereby declare:

- I do not have financial or other relationships with the manufacturer(s) of any commercial services(s) discussed in this educational activity.

Andrea Cook, PhD

Senior Investigator, Biostatistics Unit, KPWHRI

Acknowledgements

- NIH Collaboratory Coordinating Center
  - Elizabeth Delong, PhD, Lingling Li, PhD, Fan Li PhD, and Elizabeth Turner PhD
- NIH Collaboratory Project Biostatisticians
  - Patrick Heagerty, PhD, Bryan Comstock, MS, Susan Shortreed, PhD, Ken Kleinman, PhD, and William Vollmer, PhD
- NIH Methodologist
  - David Murray, PhD
- Funding
  - This work was supported by the NIH Health Care Systems Research Collaboratory (U54 AT007748) from the NIH Common Fund.

KAISER PERMANENTE®

Outline

- NIH Collaboratory Pragmatic Trial Setting
- Common themes across Collaboratory Studies
  - Study Design
  - Analysis/Sample Size
  - Randomization
    - Implications of Variable Cluster size on Estimation and Power
  - Outcome Ascertainment
- Next Steps and Conclusions

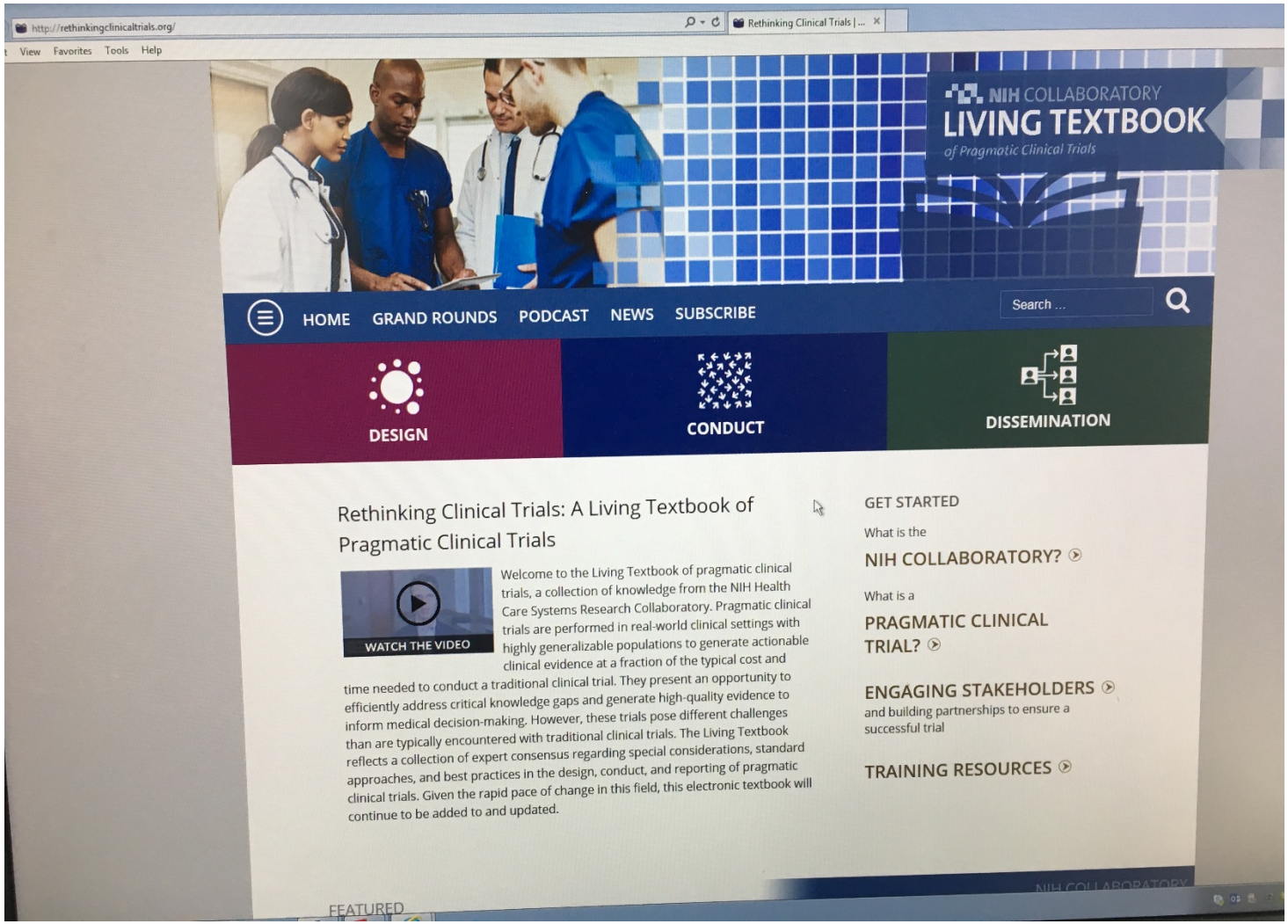
NIH Collaboratory

Health Care Systems Research Collaboratory

- Supported by The Common Fund (NIH Director's fund)
- Goal: improve the way (pragmatic) clinical trials are conducted
- Build infrastructure for collaborative research
  - "Leadership and technical expertise in all aspects of research with healthcare systems"
  - Living Textbook

NIH Collaboratory with Living Textbook

http://rethinkingclinicaltrials.org



KAISER PERMANENTE®

NIH Collaboratory

Health Care Systems Research Collaboratory

- Design and rapid execution of pragmatic trials
  - Demonstration Projects
  - UH2/UH3 funded
    - UH2 pilot phase (improve trials with collaboration amongst each other and the coordinating center through 5 core working groups)
    - UH3 where the trial is launched (most trials move to UH3 phase)

NIH Collaboratory Structure

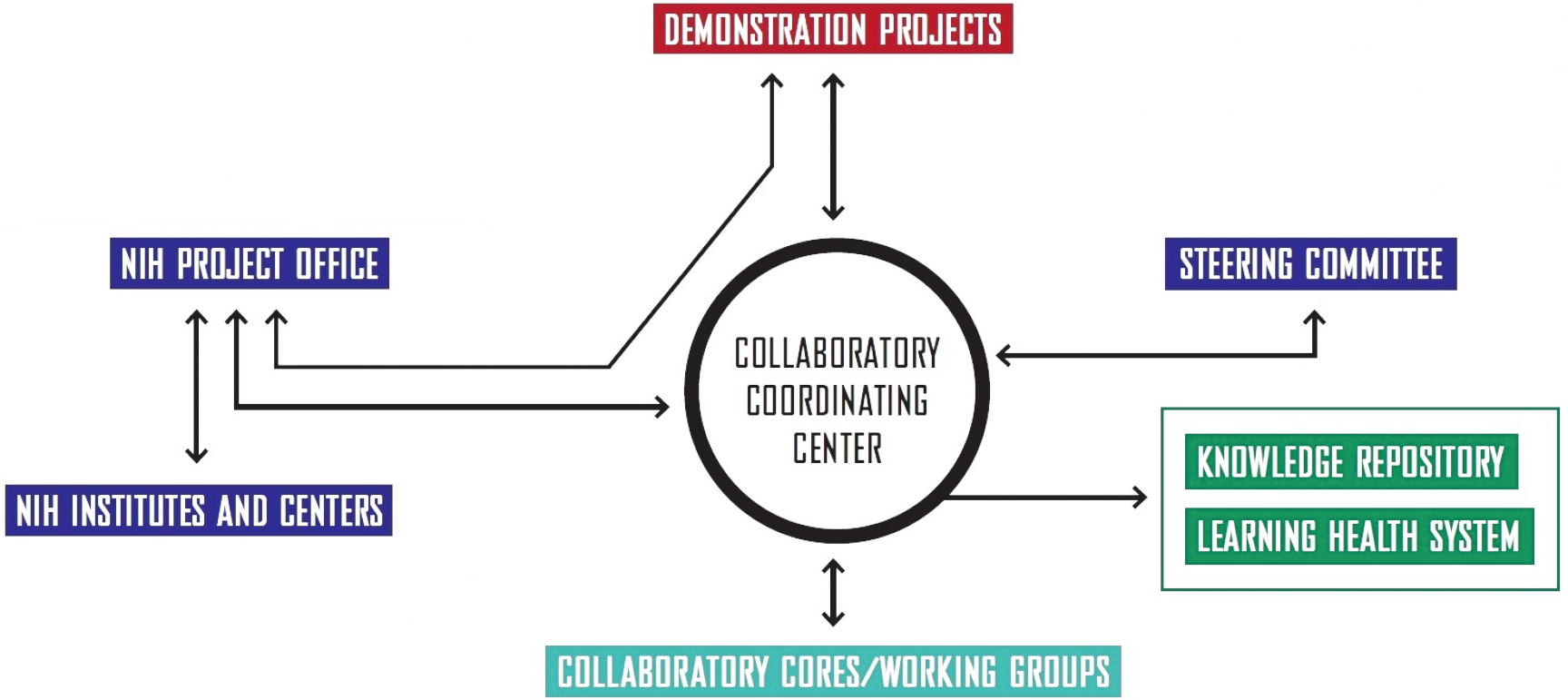


Figure downloaded from NIH Collaboratory website

KAISER PERMANENTE®

NIH Collaboratory Structure

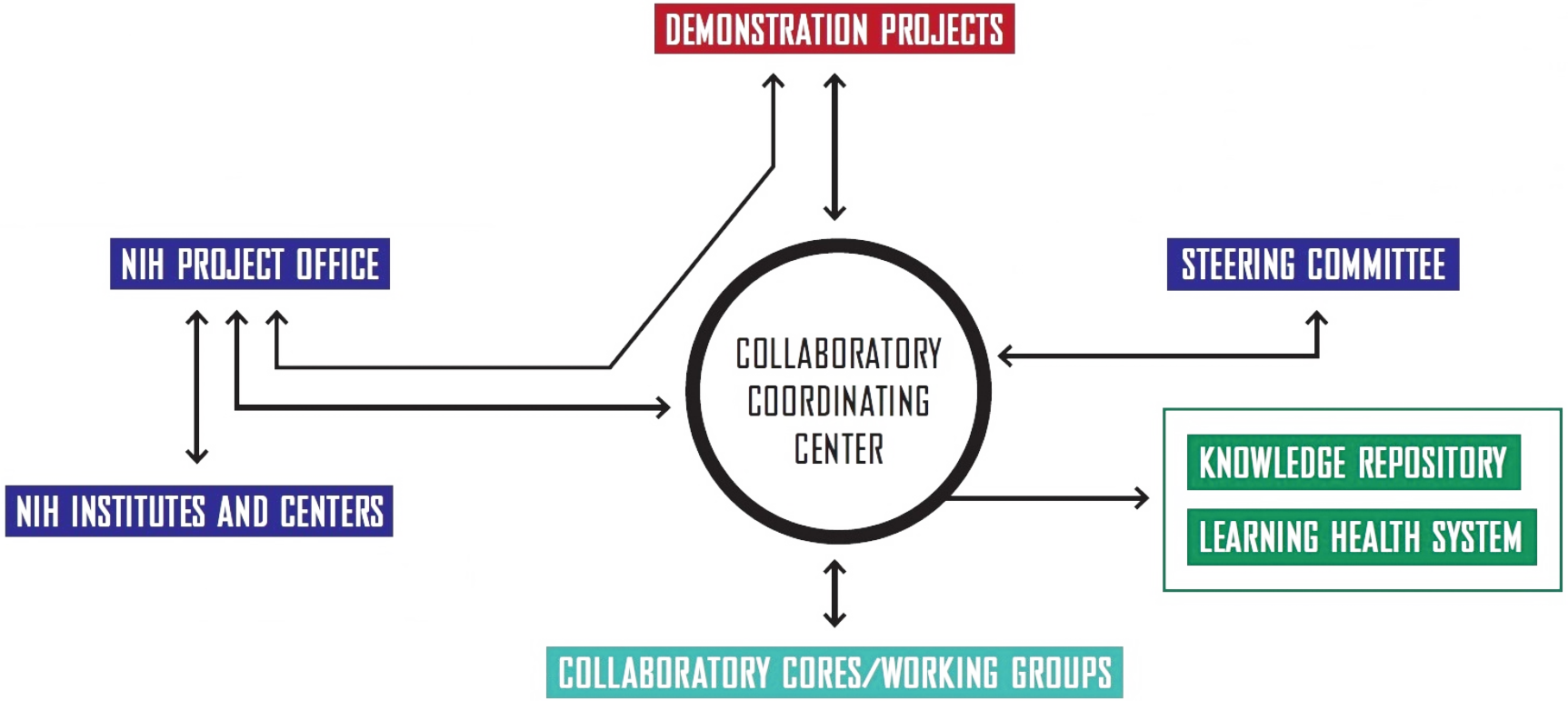


Figure downloaded from NIH Collaboratory website

Why was it created??

KAISER PERMANENTE®

Challenge #1: Clinical research is slow

- Traditional RCTs are slow and expensive—and rarely produce findings that are easily put into practice.

- In fact, it takes an average of 17 years before research findings lead to widespread changes in care.

NIH Collaboratory

Health Care Systems Research Collaboratory

Slides borrowed from NIH Collaboratory Coordinating Center Presentation

KAISER PERMANENTE®

Challenge #2: Clinical research is not relevant to practice

- Traditional RCTs study effectiveness of treatments for carefully selected populations under ideal conditions.
- Difficult to translate to real world.
- When implemented into everyday clinical practice, often see a "voltage drop"— dramatic decrease in effectiveness.

"If we want more evidence-based practice, we need more practice-based evidence."

Green, LW. *American Journal of Public Health*, 2006.

NIH Collaboratory

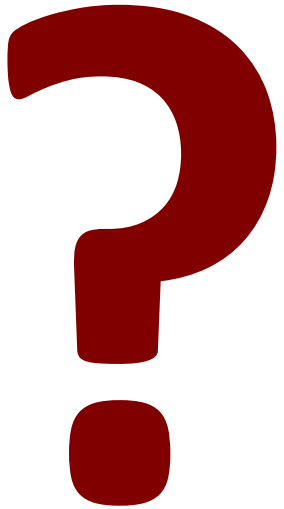
Health Care Systems Research Collaboratory

Slides borrowed from NIH Collaboratory Coordinating Center Presentation

KAISER PERMANENTE®

Challenge #3: The evidence paradox

- >18,000 RCTs published each year—plus tens of thousands of other clinical studies.
- Yet systematic reviews consistently find not enough evidence to effectively inform clinical decisions providers and patients must make.



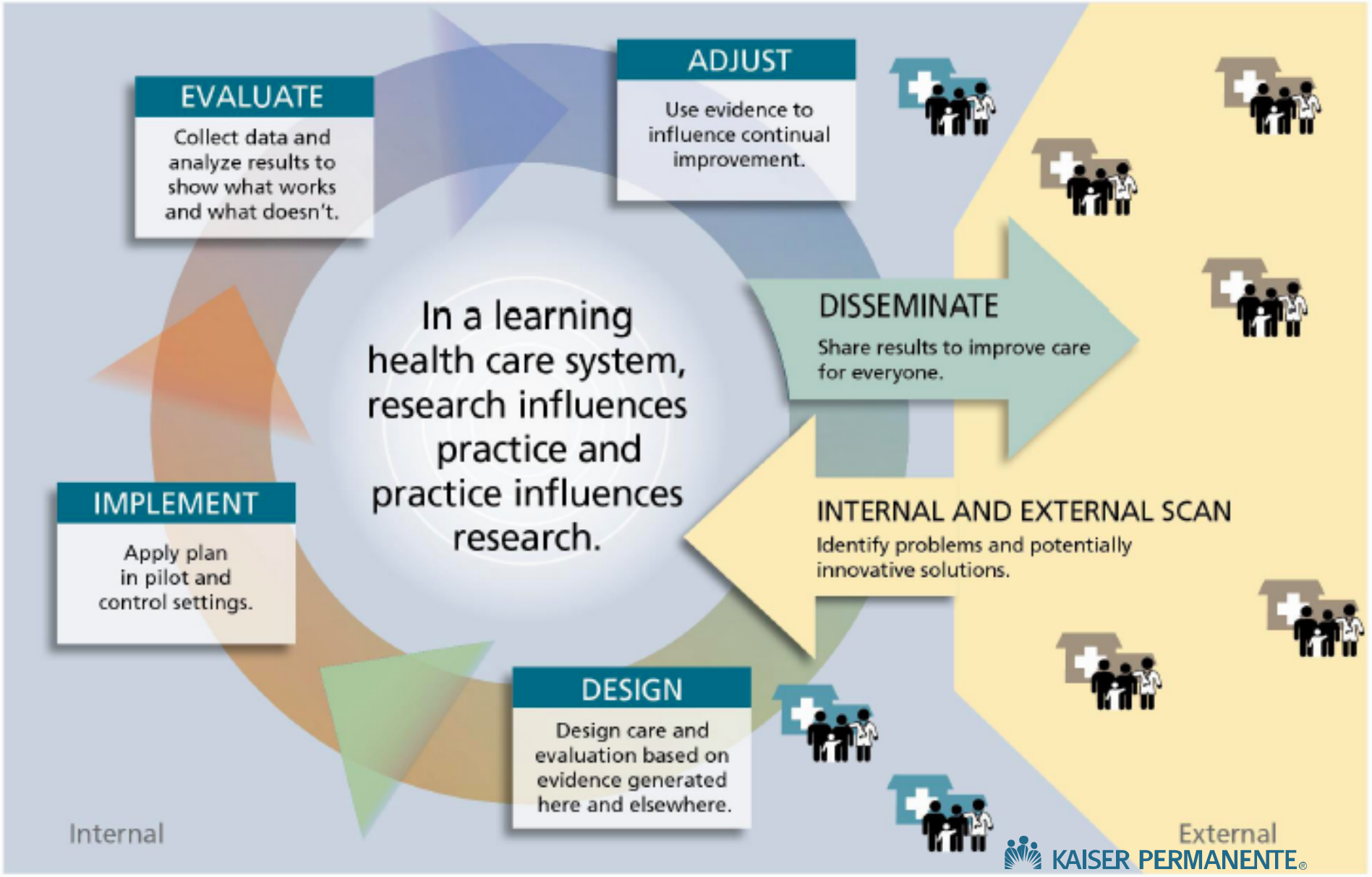
NIH Collaboratory

Health Care Systems Research Collaboratory

Slides borrowed from NIH Collaboratory Coordinating Center Presentation

KAISER PERMANENTE®

Learning Healthcare System



Pragmatic vs. Explanatory Trials

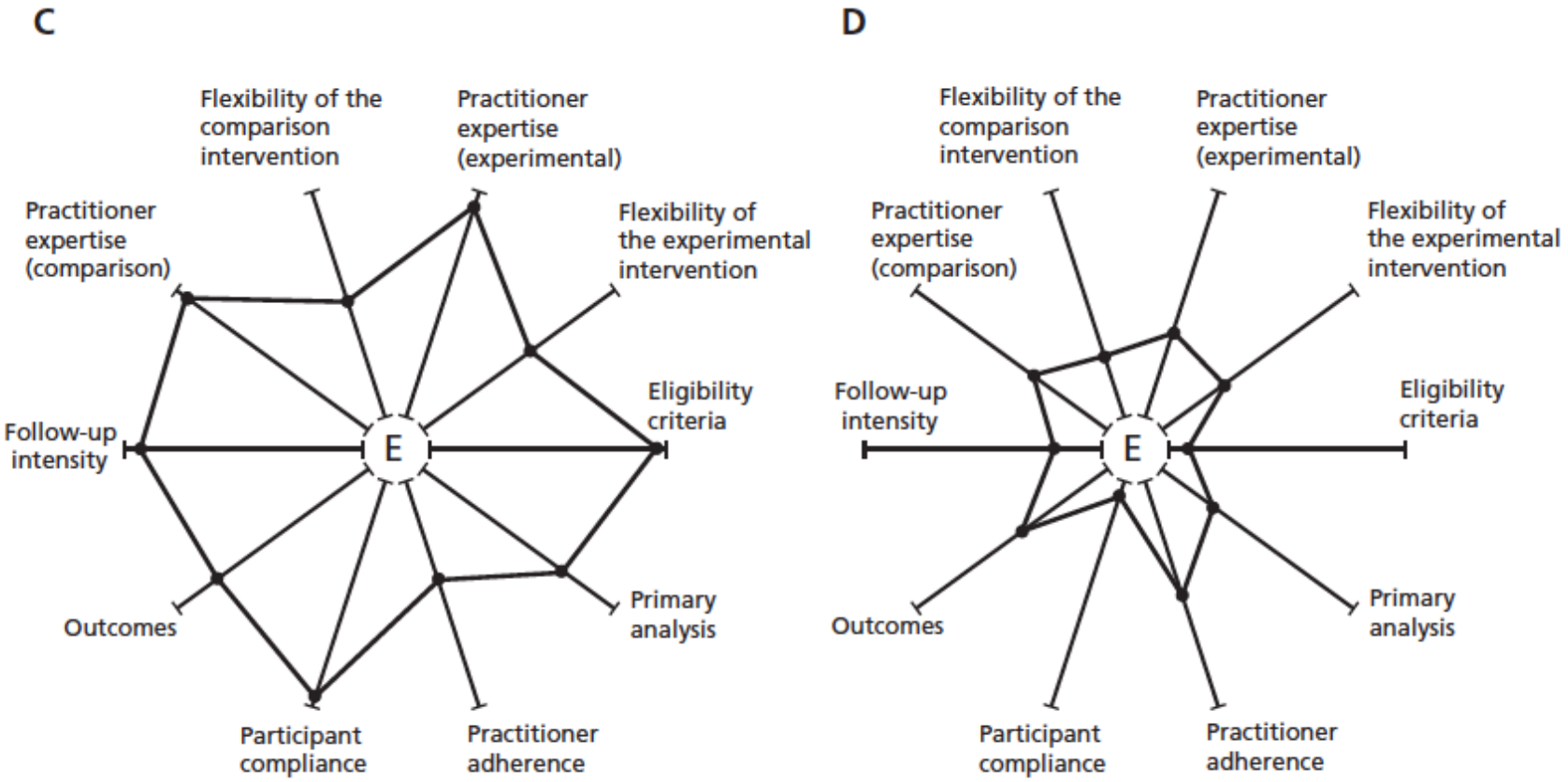
CMAJ ANALYSIS

A pragmatic–explanatory continuum indicator summary (PRECIS): a tool to help trial designers

Kevin E. Thorpe MMath, Merrick Zwarenstein MD MSc, Andrew D. Oxman MD, Shaun Treweek BSc PhD, Curt D. Furberg MD PhD, Douglas G. Altman DSc, Sean Tunis MD MSc, Eduardo Bergel PhD, Ian Harvey MB PhD, David J. Magid MD MPH, Kalipso Chalkidou MD PhD

Published at www.cmaj.ca on Apr. 16, 2009. An abridged version of this article appeared in the May 12 issue of *CMAJ*. This article was published simultaneously in the May 2009 issue of the *Journal of Clinical Epidemiology* (www.jclinepi.com).

Pragmatic vs. Explanatory Trials



KAISER PERMANENTE®

Key Features of most PCTs

Use of electronic health records (EHRs)

- EHRs allow efficient and cost-effective, recruitment, participant communication & monitoring, data collection, and follow up

Randomization at clinic or provider level

- Protocols can be tailored to local sites and can adapt to changes in a dynamic health care environment

Pragmatic Trials Concept

- Size: Large simple trials→ precise estimates, evaluate heterogeneity
- Endpoints: patient oriented usually with minimal adjudication
- Setting: integrated into real world
  - Non-academic centers
  - Leverage electronic data
  - Patients as partners

Outline

- NIH Collaboratory Pragmatic Trial Setting
- Common themes across Collaboratory Studies
  - Study Design
  - Analysis/Sample Size
  - Randomization
    - Implications of Variable Cluster size on Estimation and Power
  - Outcome Ascertainment
- Next Steps and Conclusions

STUDY DESIGN

Study Design: Cluster RCT

- Mostly Cluster RCTs
  - Randomization Unit:
    - Provider < Panel < Clinic < Region < Site
- Average Size of Cluster
  - Initial Proposals: Most large clinic level clusters
  - Goal: Smallest Unit without contamination
    - More clusters are better if possible
  - Smaller number of clusters increase sample size along with estimation issues (GEE)
  - Potential Solutions: Panel-level or physician-level

Study Design: Variable Cluster Size

- Variable Cluster Size
  - Sample Size calculations need to take this into account
    - Design effects are different
    - Depends on the analysis choice
  - Analysis Implications: What are you making inference to?
    - Cluster vs Patient vs Something in-between
    - Marginal versus conditional estimates

DeLong, E, Cook, A, and NIH Biostatistics/Design Core (2014) Unequal Cluster Sizes in Cluster-Randomized Clinical Trials, *NIH Collaboratory Knowledge Repository*.

DeLong, E, Lokhnygina, Y and NIH Biostatistics/Design Core (2014) The Intraclass Correlation Coefficient (ICC), *NIH Collaboratory Knowledge Repository*.

Study Design: Which Cluster Design?

- Cluster
  - Randomize at cluster-level
  - Most common, but not necessarily the most powerful or feasible
  - Advantages:
    - Simple design
    - Easy to implement
  - Disadvantages:
    - Need a large number of clusters
    - Not all clusters get the interventions
    - Interpretation for binary and survival outcomes:

      - Mixed models within cluster interpretation problematic

      - GEE marginal estimates interpretation, but what if you are interested in within cluster changes?

Study Design: Which Cluster Design?

- Cluster with Cross-over
  - Randomize at cluster but cross to other intervention assignment midway
  - Feasible if intervention can be turned off and on without “learning” happening
  - Alternative: baseline period without intervention and then have half of the clusters turn on

Study Design: Which Cluster Design?

	Cluster	Period 1	Period 2
Simple Cluster	1	INT	
	2	UC	
	3	UC	
	4	INT	
Cluster With Crossover	1	INT	UC
	2	UC	INT
	3	UC	INT
	4	INT	UC
Cluster With Baseline	1	UC	INT
	2	UC	UC
	3	UC	UC
	4	UC	INT

Study Design: Which Cluster Design?

- Cluster with Cross-over
  - Advantages:
    - Can make within cluster interpretation
    - Potential to gain power by using within cluster information
  - Disadvantages:
    - Contamination can yield biased estimates especially for the standard cross-over design
    - May not be feasible to switch assignments or turn off intervention
    - Not all clusters have the intervention at the end of the study

Study Design: Which Cluster Design?

- Stepped Wedge Design
  - Randomize timing of when the cluster is turned on to intervention
  - Staggered cluster with crossover design
  - Temporally spaces the intervention and therefore can control for system changes over time

Study Design: Which Cluster Design?

	Cluster	Baseline	Period 1	Period 2	Period 3	Period 4
Stepped Wedge	3	UC	INT	INT	INT	INT
	2	UC	UC	INT	INT	INT
	1	UC	UC	UC	INT	INT
	4	UC	UC	UC	UC	INT

Study Design: Which Cluster Design?

- Stepped Wedge Design
  - Advantages:
    - All clusters get the intervention
    - Controls for external temporal trends
    - Make within cluster interpretation if desired
  - Disadvantages:
    - Contamination can yield biased estimates
    - Heterogeneity of Intervention effects across clusters can be difficult to handle analytically
    - Special care of how you handle random effects in the model
    - Relatively new and available power calculation software is relatively limited

ANALYSIS/SAMPLE SIZE

Analysis: Variable Cluster Size

- Analysis Implications
  - What are you making inference to?
    - Compare intervention across clinics
      - Marginal cluster-level effect
    - Compare within-clinic intervention effect
      - Within-clinic effect
    - Compare intervention effect across patients
      - Marginal patient-level effect
    - Compare an in-between cluster and patient-level effect

DeLong, E, Cook, A, and NIH Biostatistics/Design Core (2014) Unequal Cluster Sizes in Cluster-Randomized Clinical Trials, *NIH Collaboratory Knowledge Repository*.

Cook, AJ, Delong, E, Murray, DM, Vollmer, WM, and Heagerty, PJ (2016) Statistical lessons learned for designing cluster randomized pragmatic clinical trials from the NIH Health Care Systems Collaboratory Biostatistics and Design Core *Clinical Trials* **13(5)** 504-512.

Analysis: Variable Cluster Size

- What is the scientific question of interest?
  - Marginal cluster-level effect
    - "What is the average expected clinic benefit if all clinics in the health system changed to the new intervention relative to Usual Care?"
  - Within-clinic effect
    - "What is the expected benefit if a given clinic implements the new intervention relative to Usual Care?"
  - Marginal patient-level effect
    - "What is the average expected patient benefit if all the clinics in the health system changed to the new intervention relative to Usual Care?"

Analysis: Variable Cluster Size

- Simplified Example:
  - Y_{ci} is a binary outcome for patient *i* at clinic *c*
  - n_c is the number of patients at clinic *c*
  - X_c is 1 if clinic *c* was randomized to intervention or 0
  - Estimate a simple marginal clinic-level effect (difference in clinic means amongst those randomized to intervention relative to those not randomized)

$$\widehat{\Delta}^c = \frac{\sum_{c=1}^{N} \hat{\mu}_c X_c}{\sum_{c=1}^{N} X_c} - \frac{\sum_{c=1}^{N} \hat{\mu}_c (1 - X_c)}{\sum_{c=1}^{N} (1 - X_c)}$$

where $\hat{\mu}_c = \sum_{i=1}^{n_c} \frac{Y_{ci}}{n_c}$ is the mean outcome at clinic *c*

Analysis: Variable Cluster Size

- Simplified Example:
  - Y_{ci} is a binary outcome for patient *i* at clinic *c*
  - n_c is the number of patients at clinic *c*
  - X_c is 1 if clinic *c* was randomized to intervention or 0
  - Estimate a simple marginal patient-level effect (difference in patients amongst those clinics randomized to intervention relative to those not randomized)

$$\widehat{\Delta}^p = \frac{\sum_{c=1}^{N}\sum_{i=1}^{n_c} Y_{ci}X_c}{\sum_{c=1}^{N} X_c\, n_c} - \frac{\sum_{c=1}^{N}\sum_{i=1}^{n_c} Y_{ci}(1-X_c)}{\sum_{c=1}^{N}(1-X_c)\, n_c}$$

Patients are weighted equally and clustering is really just nuisance in terms of variance and not of interest

Analysis: Variable Cluster Size

- Some ways to estimate these quantities in practice
  - Marginal cluster-level effect
    - GEE with weights the inverse of the cluster size with independent correlation structure and robust variance
  - Compare within-clinic intervention effect
    - GLMM but need to get correlation structure correct but most often just a cluster random effect
  - Marginal patient-level effect
    - GEE with no weights with independent correlation structure and robust variance
  - In-between cluster and patient-level effect
    - GEE with no weights but exchangeable cluster correlation structure and robust variance
    - Exchangeable weights based on statistical information, but not necessarily the most interpretable

Sample Size: Variable Cluster Size

- Sample Size calculations need to take variable cluster size into account
  - Design effects (amount sample size is inflated due to cluster randomization relative to individual patient randomization) are different
  - Depends on the analysis of choice and the estimate of interest

- Example: Estimating marginal clinic-level mean difference
  - Design effect:

$$1 + \left(\frac{\sum_{c=1}^{N} n_c^2}{\sum_{c=1}^{N} n_c} - 1 \right) \rho > 1 + (n_a - 1)\rho \text{ where } n_a \text{ is a constant}$$

DeLong, E, Lokhnygina, Y and NIH Biostatistics/Design Core (2014) The Intraclass Correlation Coefficient (ICC), *NIH Collaboratory Knowledge Repository*.

Eldridge, S.M., Ashby, D., and Kerry, S. (2006) Sample size for cluster randomized trials: effect of coefficient of variation of size and analysis method. *Int J Epi* **35**:1292-1300.

Figure: Power Curve
ICC is 0.03 and effect size 0.1σ

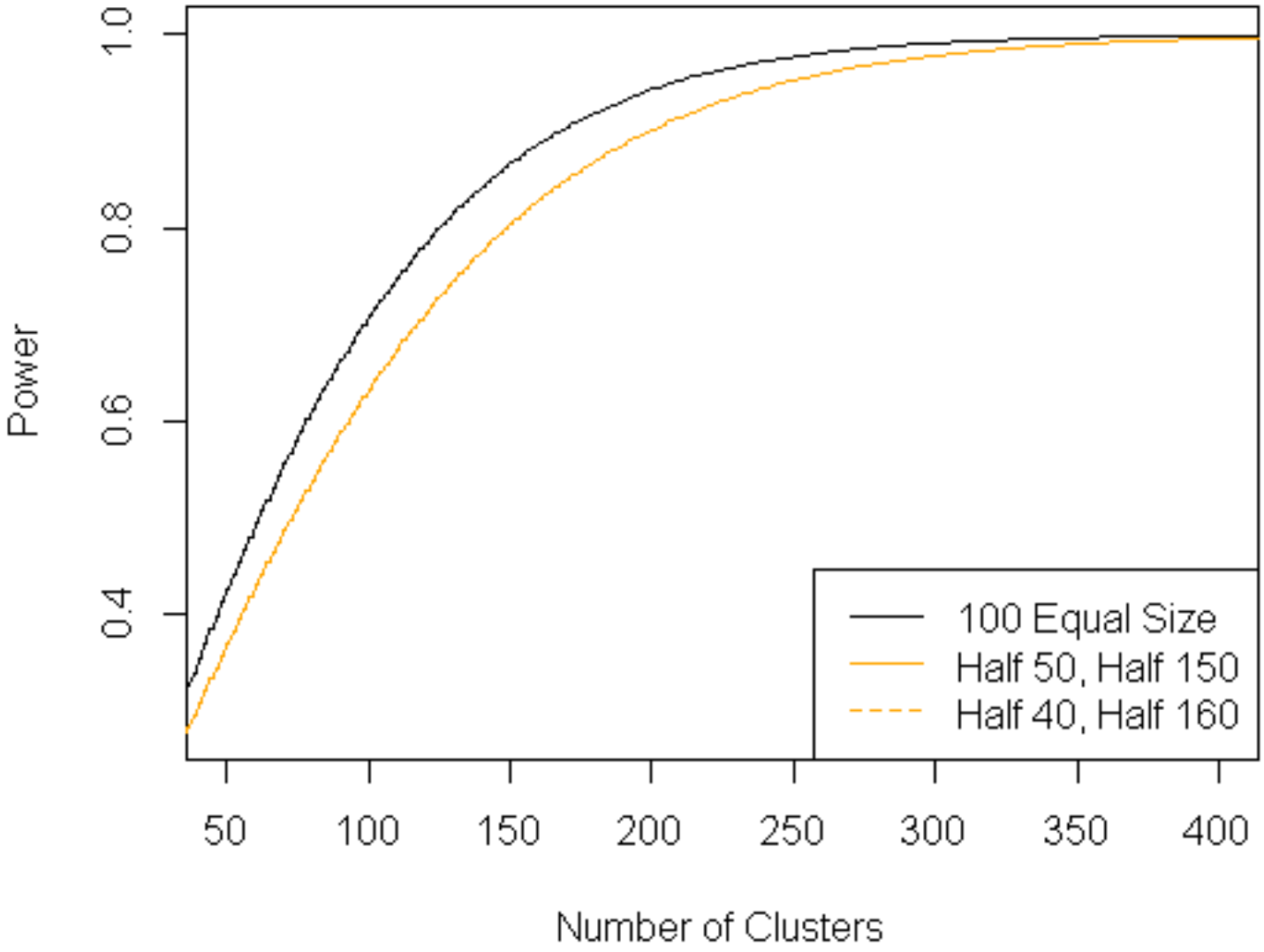


Figure: Power Curve
ICC is 0.03 and effect size 0.1σ

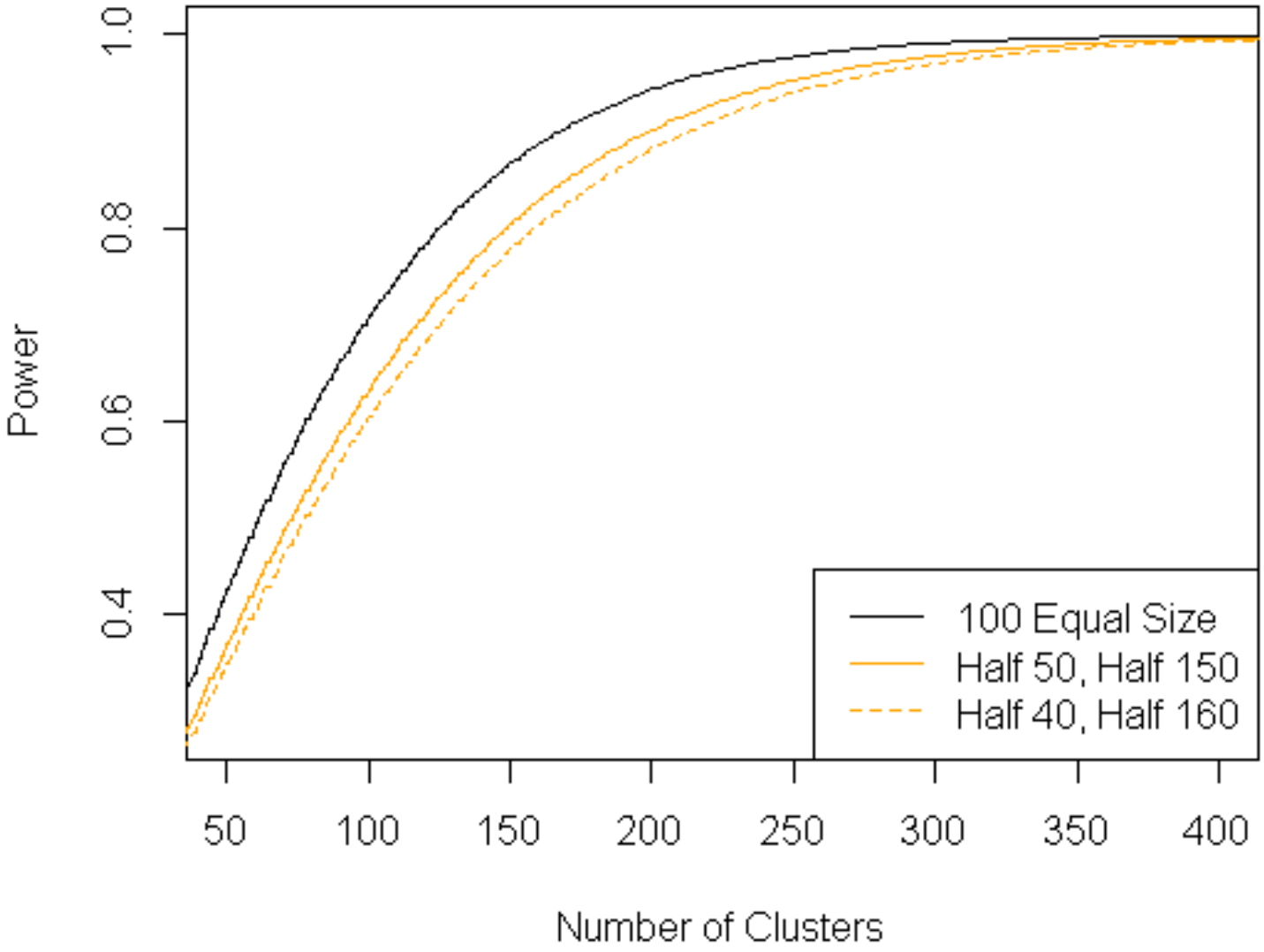


Figure: Power Curve
ICC is 0.03 and effect size 0.1σ

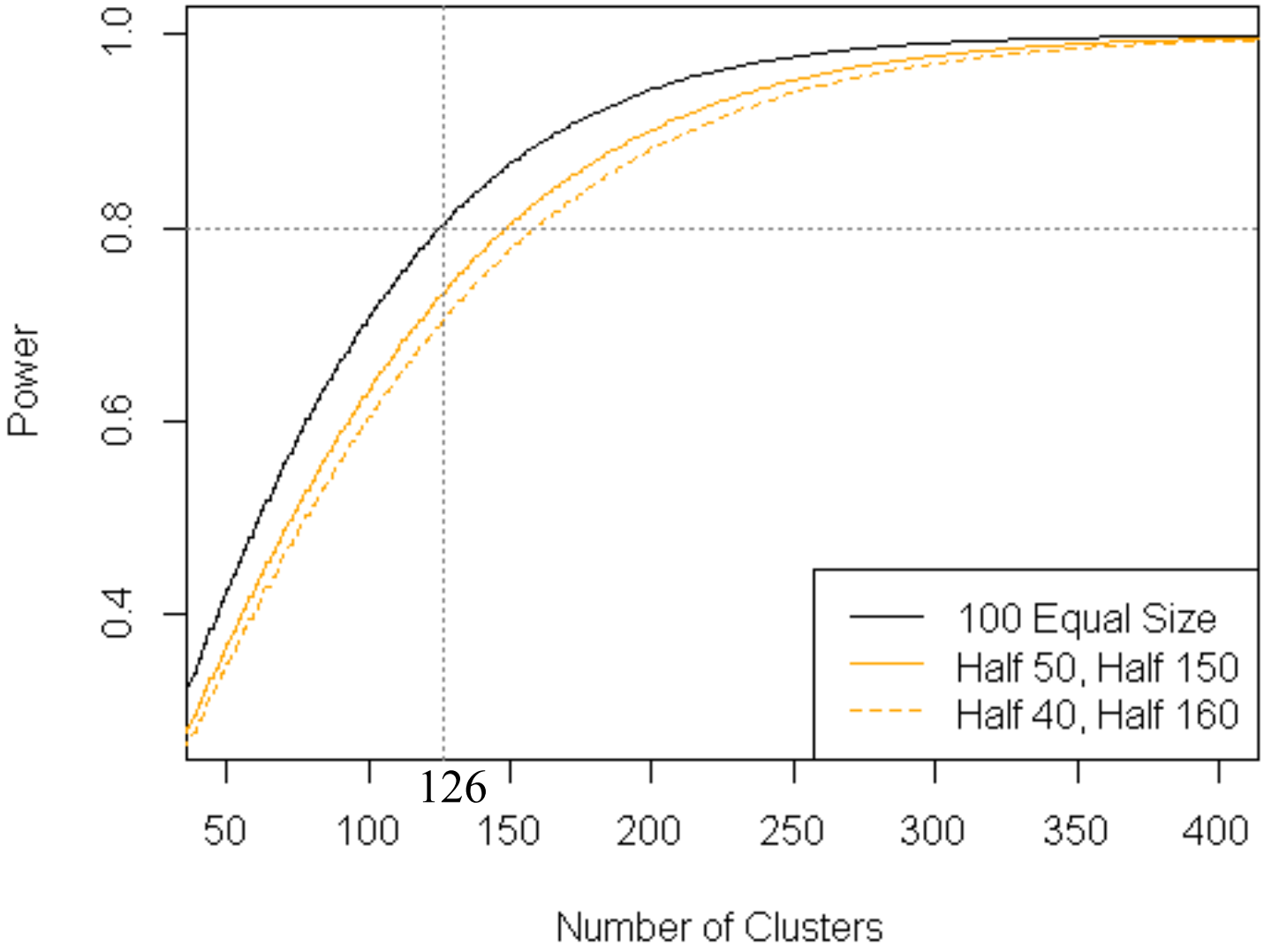


Figure: Power Curve
ICC is 0.03 and effect size 0.1σ

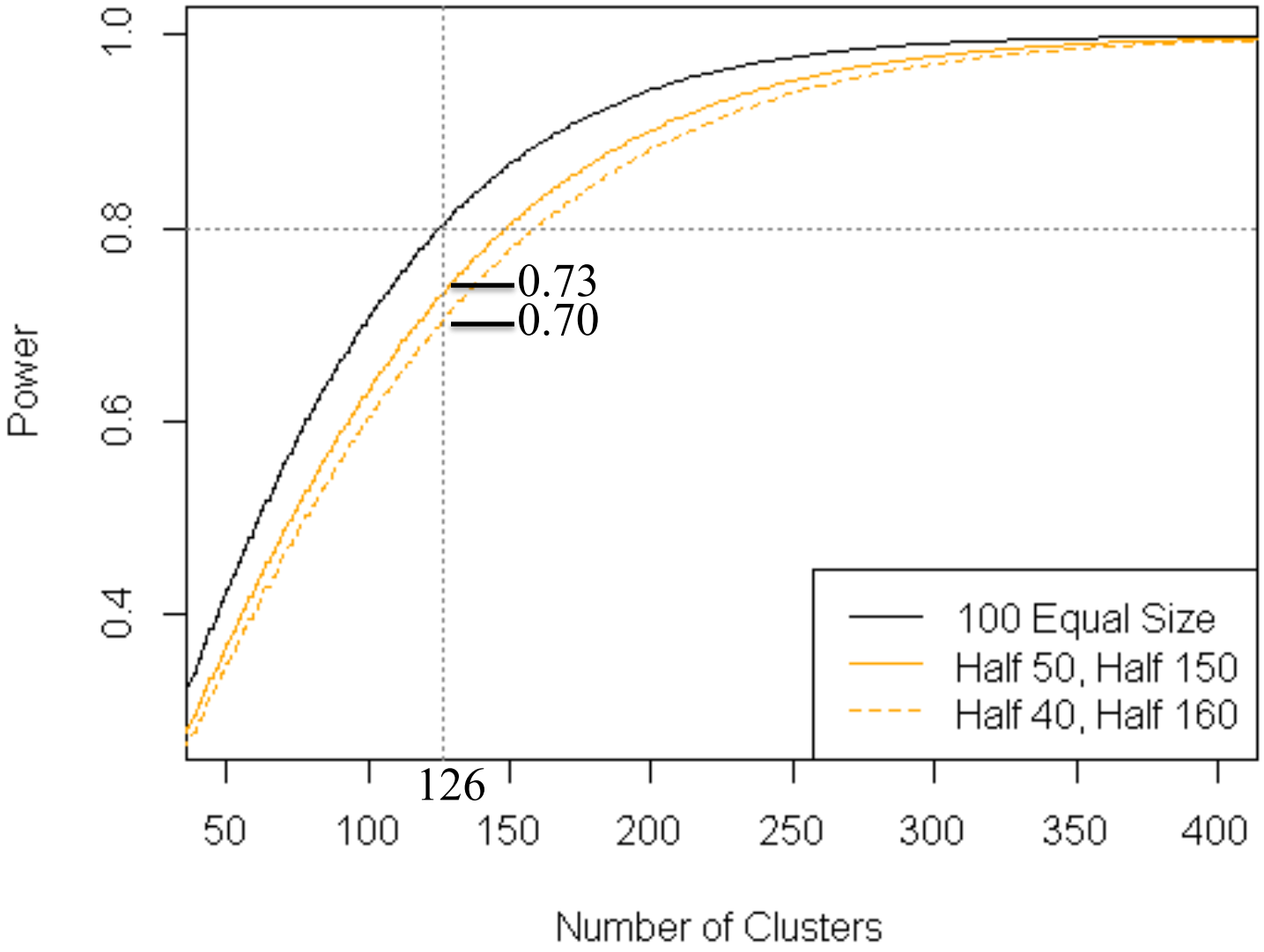
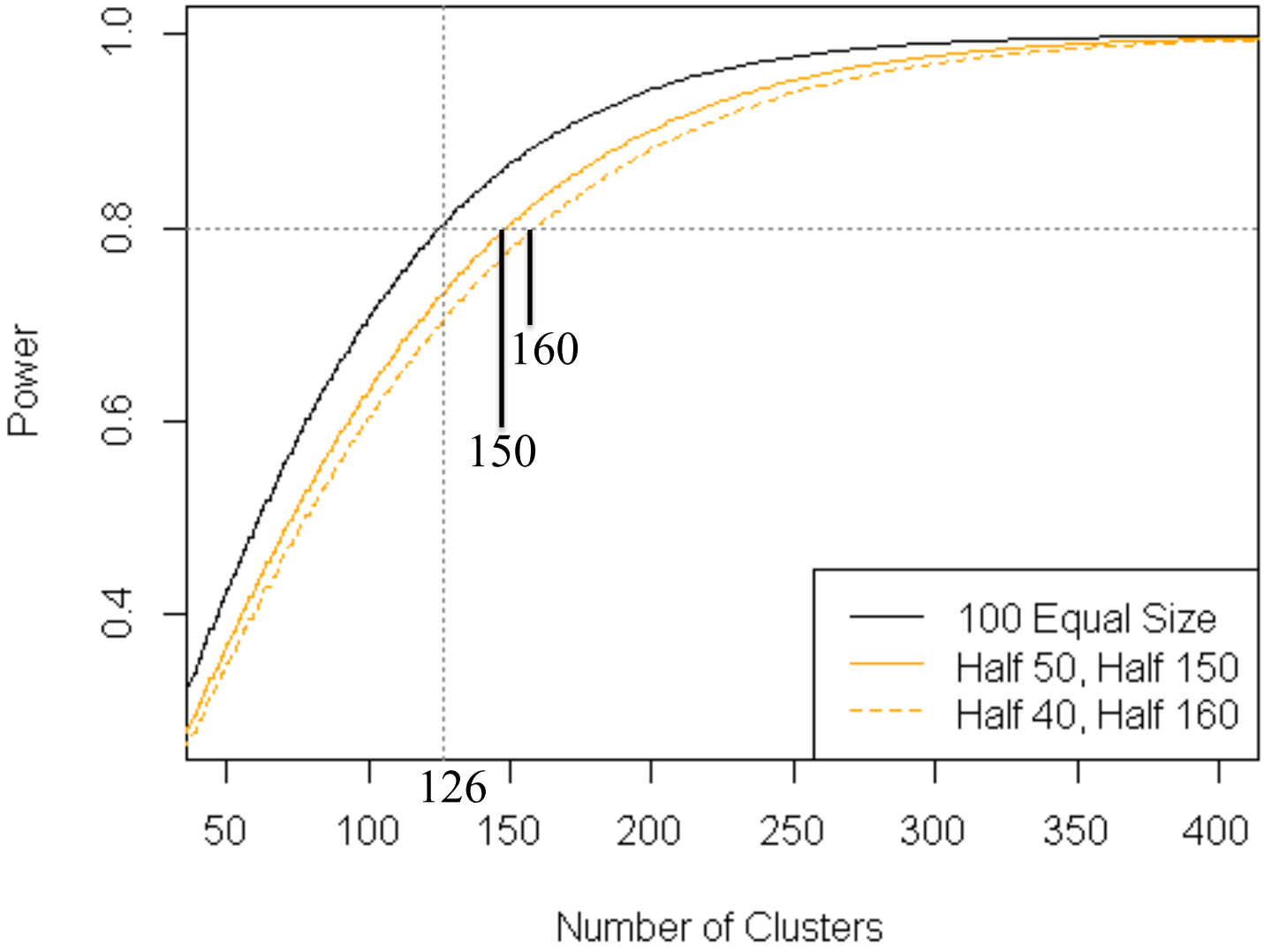


Figure: Power Curve
ICC is 0.03 and effect size 0.1σ



RANDOMIZATION

Randomization

- Crude randomization not preferable with smaller number of clusters or need balance for subgroup analyses
- How to balance between cluster differences?
  - Paired
    - How to choose the pairs best to control for important predictors?
    - Implications for analyses and interpretation
  - Stratification
    - Stratify analysis on a small set of predictors
    - Can ignore in analyses stage if desired
  - Other Alternatives

DeLong, E, Li, L, Cook, A, and NIH Biostatistics/Design Core (2014) Pair-Matching vs stratification in Cluster-Randomized Trials, *NIH Collaboratory Knowledge Repository*.

Randomization: Constrained Randomization

- Balances a large number of characteristics
- Concept
  1. Simulate a large number of cluster randomization assignments (A or B but not actual treatment)
  2. Remove duplicates
  3. Across these simulated randomizations assignments assess characteristic balance
  4. Restrict to those assignments with balance
  5. Randomly choose from the “constrained” pool a randomization scheme.
  6. Randomly assign treatments to A or B

Randomization: Constrained Randomization

- Is Constrained randomization better then unconstrained randomization?
- How many valid randomization schemes do you need to be able to conduct valid inference?
- Do you need to take into account randomization scheme in analysis?
  - Ignore Randomization
  - Adjust for variables in regression
  - Permutation inference

Randomization: Constrained Randomization

- Is Constrained randomization better then unconstrained randomization?
- How many valid randomization schemes do you need to be able to conduct valid inference?
- Do you need to take into account randomization scheme in analysis?
  - Ignore Randomization
  - Adjust for variables in regression
  - Permutation inference

  ⇨ Conduct a simulation study to assess these properties

Continuous Outcome Simulation Design

- Outcome Type: Normal
- Randomization Type: Simple versus Constrained
- Inference Type: Exact (Permutation) versus Model-Based (F-Test)
- Adjustment Type: Unadjusted versus Adjusted
- Clusters: Balanced designs, but varied size and number
- Correlation: Varied ICC from 0.01 to 0.05
- Potential Confounders: Varied from 1 to 4

Li, F., Lokhnygina, Y., Murray, D, Heagerty, P., and Delong, ER. (2016) An evaluation of constrained randomization for the design and analysis of group-randomized trials *Stat Med* **35(10)**: 1565-1579.

Continuous Outcome Simulation Results

- Adjusted F-test and the permutation test perform similar and slightly better for constrained versus simple randomization.

- Under Constrained Randomization:
  - Unadjusted F-test is conservative
  - Unadjusted Permutation holds type I error (unless candidate set size is not too small)
  - Unadjusted Permutation more powerful then Unadjusted F-Test

- Recommendation: Constrained randomization with enough potential schemes (>100), but still adjust for potential confounders

Binary Outcome Simulation Design

- Outcome Type: Binary
- Randomization Type: Simple versus Constrained
- Inference Type: Exact (Permutation) versus Model-Based (F-Test)
- Fitting Model Based: MLE versus linearization (restricted MLE)
- Adjustment Type: Unadjusted versus Adjusted
- Clusters: Balanced designs, but varied size and number
- Correlation: Varied ICC from 0.01 to 0.05
- Potential Confounders: Varied from 1 to 4

Li, F., Turner, E.L., Heagerty, P., Murray, D., Vollmer, W., and DeLong, ER. (2017) An evaluation of constrained randomization for the design and analysis of group-randomized trials with binary outcomes Stat Med 36:3791-3806

Binary Outcome Simulation Results

- Adjusted F-test based on maximum likelihood has inflated type I error
- Adjusted F-test based on linearization and the permutation test are valid and perform similarly and slightly better for constrained versus simple randomization in terms of power
- Under Constrained Randomization:
  - Unadjusted F-test is conservative
  - Unadjusted Permutation more powerful then Unadjusted F-Test
- Recommendation: Constrained randomization with enough potential schemes (>100), but still adjust for potential confounders; avoid using adjusted F-test based on maximum likelihood (PROC NLMIXED) due to its unsatisfactory small sample performance

OUTCOME ASCERTAINMENT

KAISER PERMANENTE®

Outcome Ascertainment

- Most trials use Electronic Healthcare Records (EHR) to obtain Outcomes
  - Data **NOT** collected for research purposes
- If someone stays enrolled in healthcare system - assume that if you don't observe the outcome it didn't happen
  - In closed system this is likely ok
  - Depends upon cost of treatment (likely to get a bill the more the treatment costs)

Outcome Ascertainment (cont)

- Do you need to validate the outcomes you do observe?
  - Depends on the Outcome (PPV, sensitivity)
  - Depends on the cost (two-stage design?)
- How do you handle Missing Outcome Data?
  - Leave healthcare system
    - Type of Missing Data: Administrative missingness (MCAR), MAR or non-ignorable?
    - Amount of Missing Data: how stable is your population being studied?
  - Depends on the condition and population being studied.

DeLong, E, Li, L, Cook, A, and NIH Biostatistics/Design Core (2014) Key Issues in Extracting Usable Data from Electronic Health Records for Pragmatic Clinical Trials, *NIH Collaboratory Knowledge Repository*

Conclusions

- Pragmatic Trials are important to be able to move research quickly into practice
- Pragmatic Trials add Complication
  - First Question: Can this study be answered using a pragmatic trial approach??
  - Study Design is essential and needs to be flexible
  - Using EHR data is valuable, but understanding the performance of all measures is important
  - Appropriate analysis taking into account design, randomization, and outcome ascertainment is key
  - Choice of which quantity to estimate should be made based on the scientific question of interest, but statistical trade-offs, including power, must also be considered.
  - Variability in cluster sizes have potentially major implications for power and analysis approach
- Lots of open design and statistical questions still to be addressed