En vogue trends: the good, the bad, and the ugly … opportunities and pitfalls for the ARLG

Toshi Hamasaki and Scott Evans

Why?

Chip reminded me where I would go if I refused.

A Story

- I have two kids that live with me
- Ages 80 and 81
- Names: Mom and Dad
- It is hard raising parents these days
- I give them rules to keep them out of trouble:
- Clean your room or no allowance
- No tattoos or body piercings
- In bed by 9 on school nights; 11:30 on weekends

- Clinical Trialists need rules to keep them out of trouble too

Practical Questions at Interim

- Are the interventions safe enough to justify further study?
- Has evolving medical knowledge changed the scientific validity, medical importance, ethical acceptability, or equipoise of the trial?
- Stop for efficacy or futility?
- Modify duration of follow-up due to unexpected event rates?
- Is there clinical equipoise?
- Re-calculate sample size?

Motivation

- Answering these questions has:
  - Ethical attractiveness
    - Fewer participants exposed to inefficacious/harmful therapies
  - Economical advantages
    - Smaller expected sample sizes and shorter trials
  - Public health advantages
    - Answers may get to the medical community more quickly

Major Scientific Concerns with Interim Monitoring

- Statistical
  - Error control associated with multiplicity
  - Biased estimates of treatment effect with unplanned stopping
- Operational bias

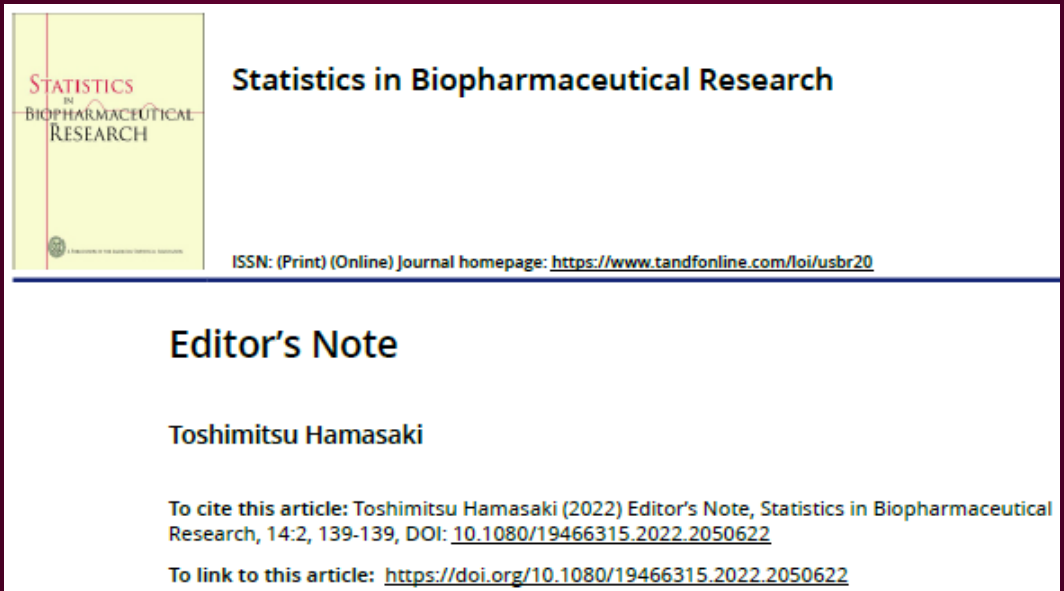
Operational Bias

- Trial results could be inferred or leaked affecting sponsor, investigator, or patient actions during the trial
  - E.g., recruitment, retention, adherence, objectivity
- Not a statistical source of bias and thus difficult for which to adjust
- May cause heterogeneity of results (before vs. after interim monitoring)
- Significant issue for adaptive designs and multi-arm studies, e.g.,
  - If sample size is adjusted based upon the observed effect size, then the quantitatively savvy may be able to back-calculate the effect
  - A treatment arm is stopped based on a comparison with the control arm. Publishing risks operational bias if the trial continues.

Addressing Concerns

- Statistical
  - Group sequential and adaptive design methods for error control
- Operational bias
  - Prevention through well-constructed DSMB and monitoring processes
    - Control of dissemination of results and adaptations

Much of what the retailors and marketers label as innovative…
is a lowering of the evidentiary standard…
camouflaged compromises in rigor and concessions of robustness.



Statistics in Biopharmaceutical Research

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/usbr20

Editor's Note

Toshimitsu Hamasaki

To cite this article: Toshimitsu Hamasaki (2022) Editor's Note, Statistics in Biopharmaceutical Research, 14:2, 139-139, DOI: 10.1080/19466315.2022.2050622

To link to this article: https://doi.org/10.1080/19466315.2022.2050622

- In clinical trials, the fewer the number of assumptions, the better. Assumptions should be reasonably justified and … verifiable through data.
- During the last two years as the Editor of SBR, I have observed that some of the models which are labeled "innovative" or "novel," require strong and nonconfirmable assumptions. These assumptions and the resulting implications are not clearly articulated.
- This makes trial results less robust and difficult to interpret. This increases the likelihood that people will draw incorrect conclusions, and the likelihood that patients will receive potentially ineffective or unsafe treatments.
- We need any innovative solutions to share certain characteristics, proven robustness with interpretable results.

JNCI: Journal of the National Cancer Institute, 2023, 115(1), 1–3
https://doi.org/10.1093/jnci/djac186
Advance Access Publication Date: 26 September 2022
Editorial

OXFORD

Editorial

Weighing evidence: robustness vs quantity

Scott R. Evans, PhD, MS,[1,2,*] Toshimitsu Hamasaki, PhD, MS[1,2]

[1]The Biostatistics Center, Milken Institute School of Public Health, George Washington University, Rockville, MD, USA
[2]Department of Biostatistics and Bioinformatics, Milken Institute School of Public Health, George Washington University, Washington, DC, USA

- Assumptions come at the sacrifice of robustness, lowering the evidentiary standard

- If one is willing to lower the standard then calculated and transparent approaches are better than ones in which error probabilities are no longer known or controlled

“Innovation”

Innovative: “featuring new methods; advanced and original”

"Innovative"?

- Real world evidence is not new … we just know it by its maiden name:

  "Observational study"

- Bayesian analyses is not new …

  Thomas Bayes, formulator of Bayes Theorem, died in 1761…

  15 years before the United States was a country

- Adaptive designs …

  Stay tuned.

- Multi-arm trials are not new…

  NIH has been supporting them for more than a half of a century

Some History

The NIH funded many of the first multi-arm trials.

These trials triggered development of:

1. the DSMB concept
2. group sequential design

The Coronary Drug Project (CDP)

- One of the first trials to use a DSMB
- Randomized, double-blind, placebo-controlled trial
- Evaluated 5 lipid-modifying agents (high-dose estrogen, low-dose estrogen, dextrothyroxine [D4T], clofibrate, and niacin) versus placebo for men 30 to 64 years of age with a documented MI within the previous 3 months
- The primary endpoint: all-cause death
- Began enrolling 8341 trial participants from 53 clinical centers in 1965

The Coronary Drug Project (CDP)

- At initiation, CDP had a steering committee of investigators to manage the trial but did not have a trial monitoring committee. Instead, investigators were informed of accumulating outcome data.

- With issuance of the Greenberg Report in 1967, concern was expressed that investigator knowledge of early trends in mortality, morbidity, or side effects may tempt investigators to select treatments that appear to be best and over-diagnose or report findings on the basis of early report summaries

- In April 1968, a decision was made that outcome data would no longer be available to CDP trial investigators and a safety monitoring committee was formed to review the data on a regular basis

The Coronary Drug Project (CDP)

- The committee eventually recommended termination of 3 of the 5 active treatment arms during the trial:
  - High-dose estrogen arm was discontinued in 1970 because of an increased incidence of cardiovascular events relative to the placebo
  - D4T arm was discontinued in 1971 owing to increased mortality relative to the placebo
  - Low-dose estrogen arm was discontinued in 1973 for futility on the basis of an evaluation indicating that it would be nearly impossible to conclude benefits with trial continuation
  - The clofibrate and niacin arms continued until the planned trial completion, but the results did not indicate a survival benefit

Cardiac Arrhythmia Suppression Trial (CAST)

- Randomized placebo-controlled trial that evaluated the effects of 3 drugs (encainide, flecainide, and moricizine), which at the time of the trial were approved for use by FDA for the treatment of cardiac arrhythmias, on the incidence of sudden cardiac death or all-cause death in patients after a myocardial infarction (MI)

- In the early 1980s, the thinking was that cardiac arrhythmia was associated with sudden or cardiovascular death and that treatment with antiarrhythmic drugs to suppress arrhythmias would thus reduce cardiac death among these patients

- CAST aimed to randomly assign 4400 patients to active treatment with one of the three drugs (encainide, flecainide, or moricizine) or placebo

Cardiac Arrhythmia Suppression Trial (CAST)

- A pre–randomized assignment run-in period was used to identify patients with a sufficient response to one of the drugs, defined by 80% arrhythmia suppression, for eligibility
- Enrollment began in 1987
- Two years later, the trial's DSMB recommended discontinuation of the encainide and flecainide arms as a result of increased mortality
- What seemed like a good treatment strategy based on the scientific knowledge at the time had been just plain wrong

Group Sequential Design (GSD)

- Fully size the trial
- Allow interim analyses before the planned end of trial
  - Efficacy or futility testing
- Retain strong control of Type I and II error using multiplicity adjustments
- A fundamental aspect of design in play for >5 decades
- Beware of necessary info on secondary information

NEJM Evidence

Published January 10, 2022
NEJM Evid 2022; 1 (1)
DOI: 10.1056/EVIDctw2100005

CLINICAL TRIALS WORKSHOP | DSMB MINI-SERIES

Independent Oversight of Clinical Trials through Data and Safety Monitoring Boards

Scott R. Evans, Ph.D.

NEJM Evidence

Published February 22, 2022
NEJM Evid 2022; 1 (3)
DOI: 10.1056/EVIDctw2100025

CLINICAL TRIALS WORKSHOP | DSMB MINI-SERIES

Data and Safety Monitoring Board Monitoring of Clinical Trials for Early Efficacy

Lori E. Dodd, Ph.D.,[1] and Michael A. Proschan, Ph.D.[1]

NEJM Evidence

Published June 28, 2022
NEJM Evid 2022; 1 (7)
DOI: 10.1056/EVIDctw2100020

CLINICAL TRIALS WORKSHOP | DSMB MINI-SERIES

Early Termination of Clinical Trials for Futility — Considerations for a Data and Safety Monitoring Board

Susan S. Ellenberg, Ph.D.,[1] and Pamela A. Shaw, Ph.D.[2]

NEJM Evidence

Published April 26, 2022
NEJM Evid 2022; 1 (5)
DOI: 10.1056/EVIDctw2100026

CLINICAL TRIALS WORKSHOP | DSMB MINI-SERIES

Stopping Trials Early Due to Harm

Thomas Cook, Ph.D.,[1] and Olive D. Buhule, Ph.D.[2]

NEJM Evidence

Published August 23, 2022
NEJM Evid 2022; 1 (9)
DOI: 10.1056/EVIDctw2100019

CLINICAL TRIALS WORKSHOP | DSMB MINI-SERIES

The Impact of Landscape Changes on Data and Safety Monitoring Board Oversight of Clinical Trials

Kaleab Z. Abebe, Ph.D.,[1] and Frank W. Rockhold, Ph.D.[2]

NEJM Evidence

Published January 24, 2023

DOI: 10.1056/EVIDctw2200220

CLINICAL TRIALS WORKSHOP | DSMB MINI-SERIES

The Data and Safety Monitoring Board: The Toughest Job in Clinical Trials

Scott R. Evans, Ph.D.,[1] Lijuan Zeng, M.H.S.,[1] and Weixiao Dai, M.S.[1]

Abstract

In this article, we discuss methods that data and safety monitoring boards (DSMBs) can use to compare the absolute and relative risks of benefits and adverse effects between trial interventions and illustrate how the DSMB can use this approach to evaluate the balance between these competing risks. Two approaches are discussed — first, the win ratio (i.e., the relative frequency by which one treatment has a more desirable result than another); and second, the desirability of outcome ranking probability (i.e., the probability of an overall more desirable result on one treatment relative to another).

Scott R. Evans, Ph.D.,
DSMB Mini-Series Editor

Jeffrey Drazen, M.D.,
Editor

Multi-arm Multi-stage (MAMS)

- Master protocol
- Platform allowing evaluation of several interventions
  - Vs. common control group
  - New treatments can be added
  - Treatment arms may stop for efficacy or futility with control of error rates using group sequential design fundamentals
    - Control of pair-wise comparison error rate (vs. trial-wise error
- Generally oriented at development rather than informing practice as generally restrict comparisons to control
- Used to evaluate antibiotic strategies in TB in the PanACEA trial

CANCER RESEARCH UK

MRC Clinical Trials Unit

NCRN National Cancer Research Network

STAMPEDE

Systemic **T**herapy in **A**dvancing or **M**etastatic **P**rostate Cancer: **E**valuation of **D**rug **E**fficacy

A multi-arm multi-stage randomised controlled trial

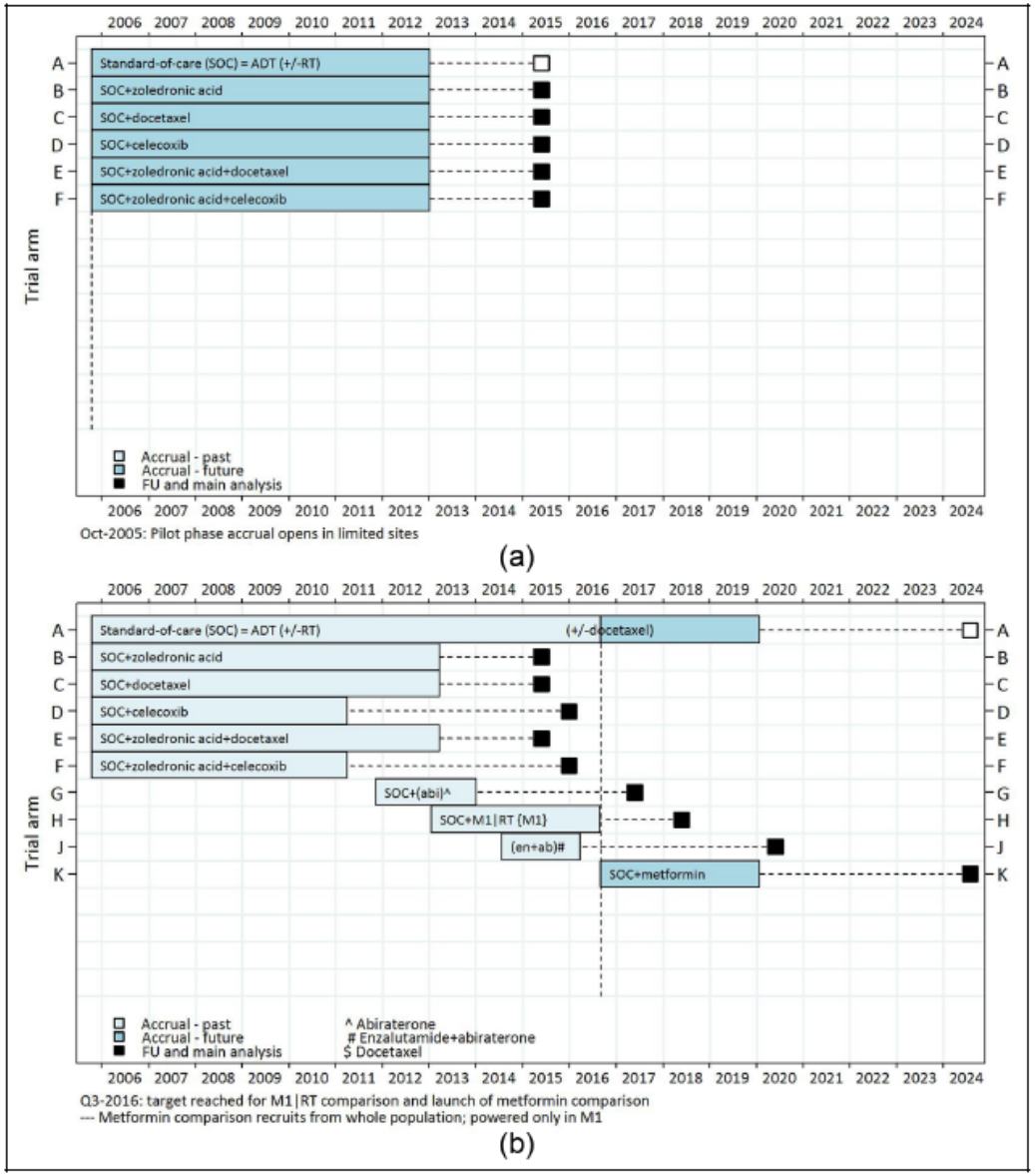


Figure 1. STAMPEDE protocol (a) at initiation (figure produced in October 2005) and (b) adapted protocol from 2005 to 2024 (figure produced in September 2016).

The NEW ENGLAND JOURNAL of MEDICINE

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Baricitinib plus Remdesivir for Hospitalized Adults with Covid-19

A.C. Kalil, T.F. Patterson, A.K. Mehta, K.M. Tomashek, C.R. Wolfe, V. Ghazaryan, V.C. Marconi, G.M. Ruiz-Palacios, L. Hsieh, S. Kline, V. Tapson, N.M. Iovine, M.K. Jain, D.A. Sweeney, H.M. El Sahly, A.R. Branche, J. Regalado Pineda, D.C. Lye, U. Sandkovsky, A.F. Luetkemeyer, S.H. Cohen, R.W. Finberg, P.E.H. Jackson, B. Taiwo, C.I. Paules, H. Arguinchona, N. Erdmann, N. Ahuja, M. Frank, M. Oh, E.-S. Kim, S.Y. Tan, R.A. Mularski, H. Nielsen, P.O. Ponce, B.S. Taylor, L.A. Larson, N.G. Rouphael, Y. Saklawi, V.D. Cantos, E.R. Ko, J.J. Engemann, A.N. Amin, M. Watanabe, J. Billings, M.-C. Elie, R.T. Davey, T.H. Burgess, J. Ferreira, M. Green, M. Makowski, A. Cardoso, S. de Bono, T. Bonnett, M. Proschan, G.A. Deye, W. Dempsey, S.U. Nayak, L.E. Dodd, and J.H. Beigel, for the ACTT-2 Study Group Members*

RANDOMI

Backgroun
treatments f
emerged in
Advisory Gro
including Lop
now been do
advice. Thes
evaluation. I
measures, a

Eligibility ar
hospitalised
assessed in
treatment an
hospital. The
suitable trea
version of th
shown in Tab
randomised
simultaneously.

Condition	Randomised compar... each vs. usual care a...		
COVID-19	High-dose corticosteroids		
	Empagliflozin		
	Sotrovimab		
	Molnupiravir		
PIMS-TS	Tocilizumab or anakin...		
Influenza	Baloxavir	(age ≥12 years)	
	Oseltamivir	✓ (any age)	×
	Low-dose corticosteroids	✓ (any age with hypoxia)[b]	×

[a] without suspected or confirmed influenza infection; [b] without suspected or confirmed SARS-CoV-2 infection. Information on completed arms is available in Section 7.

Table 1: Current comparisons

Remd... Report

J.H. Beig... Y. Chu,
A. Luetkeme... Patterson,
R. Parede... Palacios,
T. Benfiel... S. Pett,
J.D. Nea... C. Lane,

R&DBlueprint
Powering research to prevent epidemics

PLUS

...onal
...ients
...care

Operational Issues with Platform Trials

- Central decision-making
  - Who has control of the design and conduct?
- How to budget a trial that has no clear end?
  - Sample size and number of arms may be increased or decreased
  - Complex drug supply management
- Blinding with multiple arms?
  - Build in protections in the absence thereof
- “Control group” may change during the trial if one arm demonstrates superiority to original control; retain best option

Issues with Platform Trials

- Improper control of access to interim data → operational bias
- Reporting results of completed arms → operational bias

Improved Approaches: Comparative Effectiveness

- Motivated by finding the best treatment for patients
- Is it irresponsible or unethical not to evaluate all comparisons in the fact of randomized evidence that could save lives associated with a deadly disease and not report it?
- Do data sharing rules imply comparisons will be made later anyway?

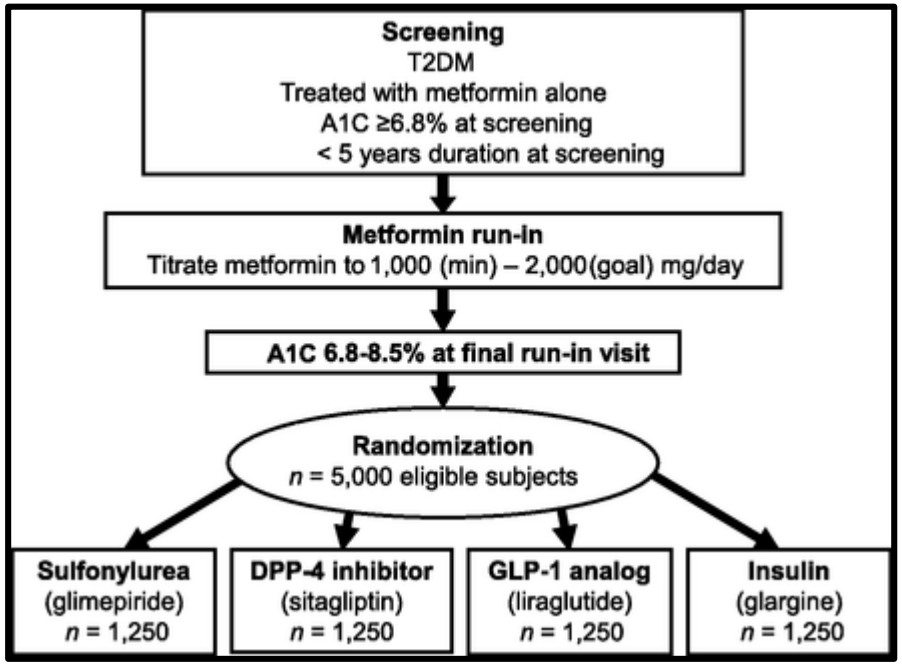
Glycemia Reduction Approaches in Diabetes (GRADE): A Comparative Effectiveness Study

- Most people with T2D eventually need 2 medications to control blood glucose levels
- Which of the many available drugs is the best choice among people already treated with metformin, the most commonly used diabetes drug, is unknown
- Pragmatic: make all comparisons to inform clinical practice

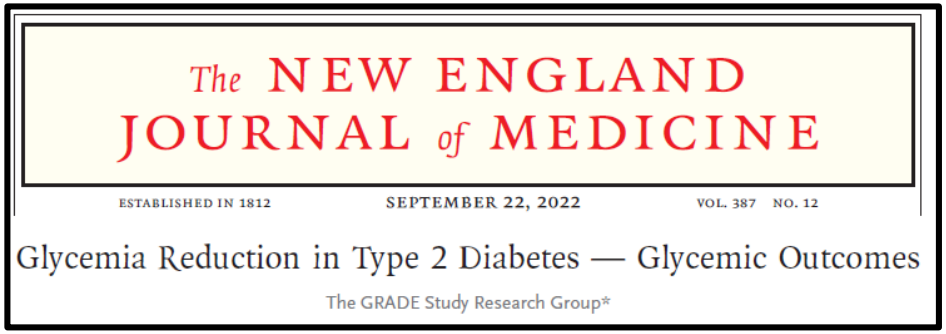
GRADE

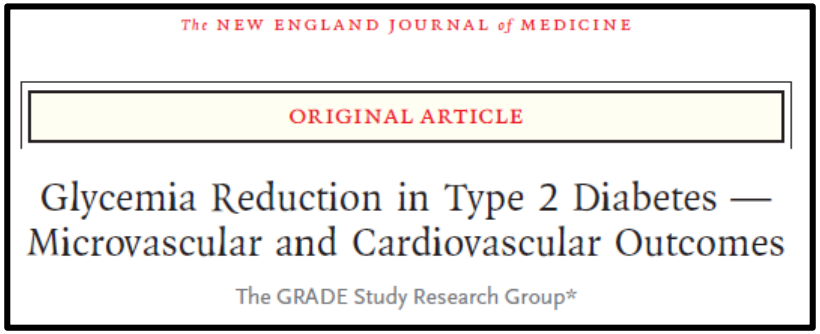
- A randomized clinical trial that enrolled more than 5000 patients from more than 40 clinical sites and followed them for up to seven years

- Compares four two-drug combinations to evaluate which is best for achieving glycemic control, having the fewest side effects, and is the most beneficial for overall health in long-term treatment for people with T2D



NIH National Institute of Diabetes and Digestive and Kidney Diseases





- All four medications, when added to metformin, decreased glycated hemoglobin levels.
- Glargine and liraglutide were modestly more effective in achieving and maintaining target glycated hemoglobin levels
- The incidences of microvascular complications and death were not materially different
- Possible differences in the incidence of cardiovascular disease

Adaptive Designs

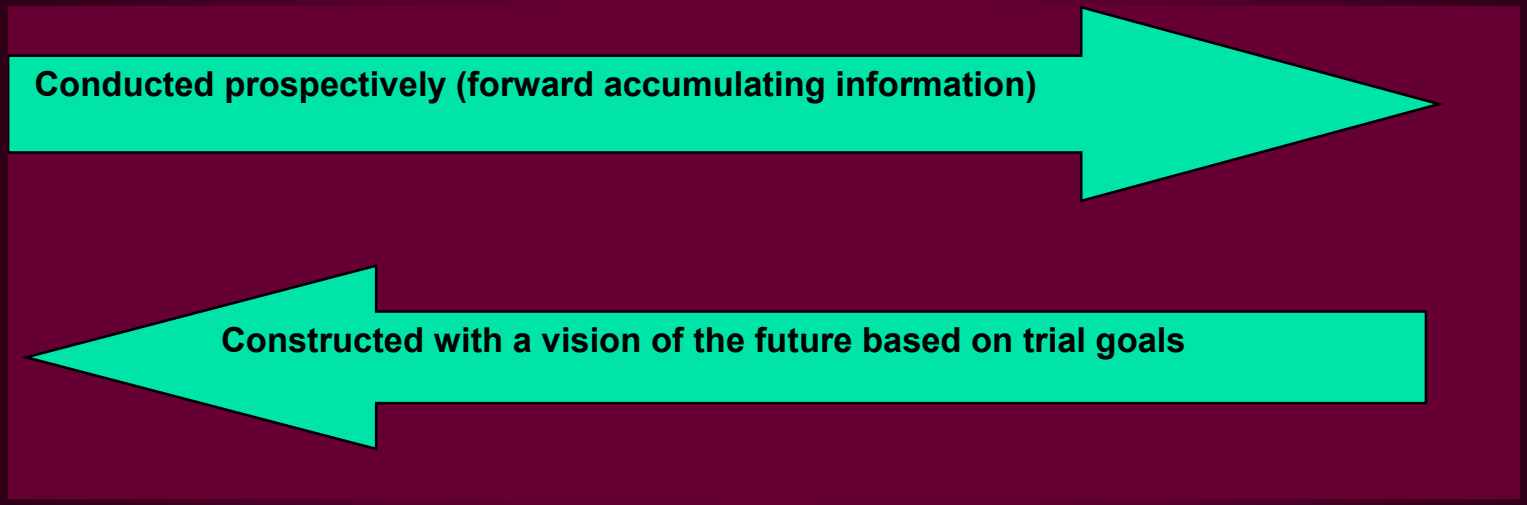
- Multiple definitions have been used
  - FDA: a clinical trial design that allows for prospectively planned modifications to one or more aspects of the design based on accumulating data from subjects in the trial
  - Narrower definitions: often restrict definition to changes made based on the observed treatment effects
  - Broad definitions: allow changes based on external data
  - Some include GSD; some do not

- A design feature with a plan for error/bias control
  - Described in protocol; requires more planning, not less
  - Not a rescue medication

- Fancy statistical methods cannot rescue poorly designed trials

- Adaptive designs have assumptions and limitations

Designing Clinical Trials



The key: vision

Integrity

- Are the data a result of the trial or the trial a result of the data?
  - Imposed circularity blurs causality and the distinction between learning and confirming, i.e., hypothesis generation vs. confirming
- Some adaptations are well grounded and understood; others less so
- Depends upon
  - Type of adaptation
  - The data utilized for decision-making
    - Has interim data (particularly endpoint data) been reviewed?
    - Blinded or unblinded
  - How adaptation is implemented
  - Who is reviewing data and making the decision to adapt

Threat to Trial Integrity

- Low
  - Adaptations prior to any data analyses
  - Adaptations based on
    - Baseline data
    - External data
    - Blinded (aggregate) data
    - Nuisance parameters (e.g. variation)
- High
  - Unplanned adaptations
  - Adaptations based on observed treatment effects
    - Quantitatively savvy people could back-calculate results

Example: Adapt Sample Size based on Observed Treatment Effect?

- Issues
  - Statistical error control
  - Conceptual
    - Trials are designed to detect *relevant* effects
    - Observed effects may not be relevant
  - Operational bias
    - If sample size is recalculated based on observed treatment effect, quantitatively savvy researchers could back-calculate the treatment effect threatening trial integrity
    - Trials can implement mechanisms such as:
      - DSMB apprised of ongoing accrual… when appropriate sample size is reached, DSMB says stop

Example of Issue with Sample-size Recalculation Based on Observed Effect Size

- Sponsor asked DSMB to conduct sample size re-calculation based on observed effect
- Trial without a mechanism to prevent back-calculation
- Easy to back-calculate results once new sample size is announced, jeopardizing trial integrity
- Despite education regarding this issue, sponsor kept original plan
- I resigned from this DSMB…
- The rest of the DSMB soon followed

When and Where?

- Longer trials where adaptation is feasible
  - Larger studies, studies with slow recruitment, or long duration of FU
  - Accumulating external medical information can influence the utility and ethics of ongoing trials of long duration
- When design characteristics (e.g., power) are very sensitive to assumptions
- High levels of uncertainty/unknowns
  - E.g., novel interventions where rates and variation is unknown

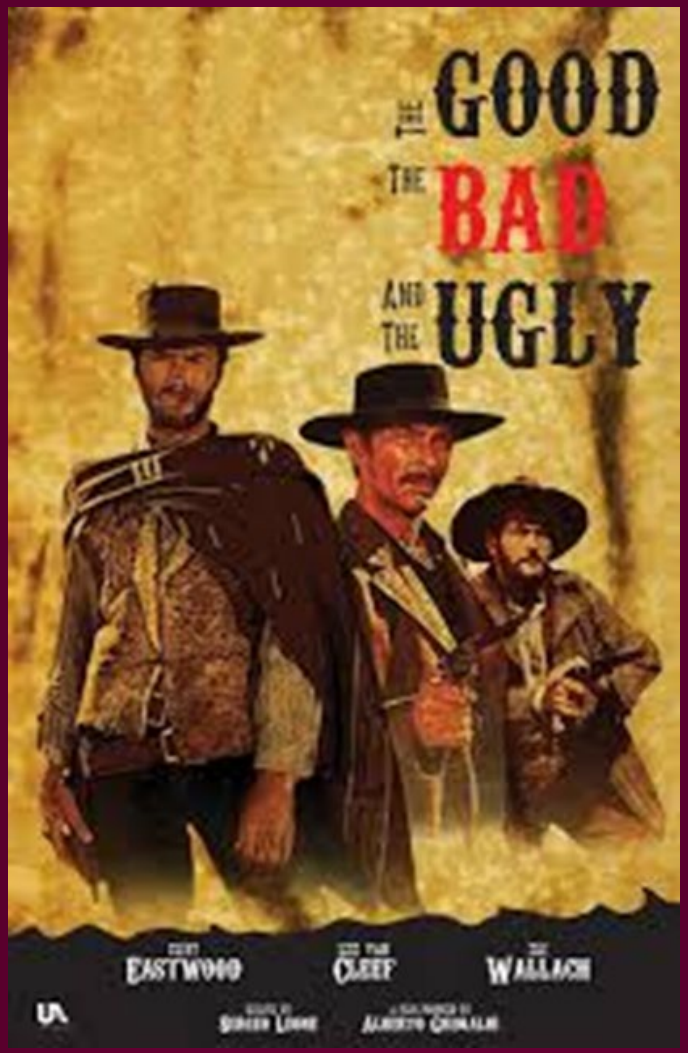
I have designed > 1500 clinical trials…

each time having to make assumptions about e.g., variation or event rates…

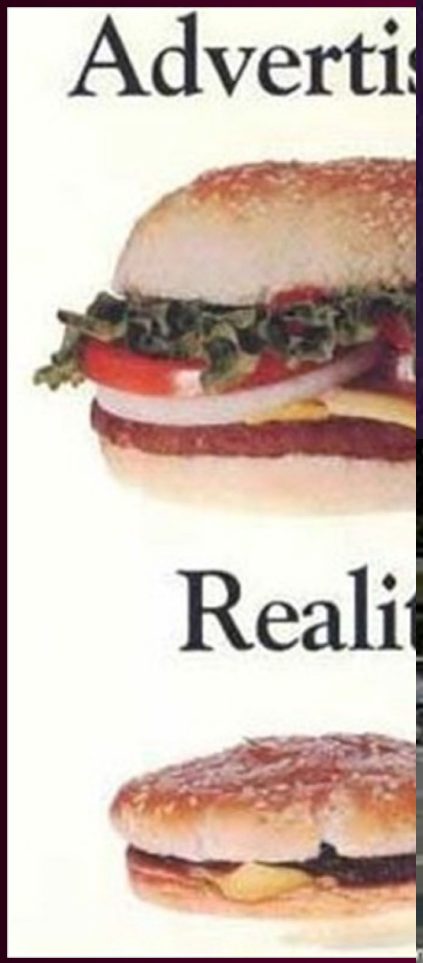
I have not been right yet.

Monitoring has value.

Adaptive Design



The “Bad”: A Lot of Overselling



Journal of Biopharmaceutical Statistics, 20: 1150–1165, 2010

ISSN: 1054-3406 print/1520-5711 online
DOI: 10.1080/10543406.2010.514457

Taylor & Francis
Taylor & Francis Group

ADAPTIVE METHODS: TELLING "THE REST OF THE STORY"

Scott S. Emerson and Thomas R. Fleming
Department of Biostatistics, University of Washington, Seattle, Washington, USA

The Food and Drug Administration (FDA) draft guidance on adaptive design randomized clinical trials provides in-depth consideration of the difficulties that unblinded adaptation of clinical trial design might introduce. We provide extended discussion of these difficulties, with focus on the problems that the adaptive designs pose in the scientific interpretation of randomized clinical trial results, for regulatory authorities as well as for patients and caregivers who wish to make evidence-based decisions regarding the choice of treatment. We consider implications in adequate and well-controlled studies of the use of unblinded measures of treatment effect to make adaptive selection/modification of treatments, adaptive selection of primary endpoints, adaptive modification of maximal sample size, adaptive modification of randomization ratios, and adaptive modification of target populations (adaptive enrichment), and then we consider the special topic of seamless phase 2–3 designs. We examine the extent to which the adaptive designs do not meet the goals of having greater efficiency, being more likely to identify truly effective treatments, being more informative, and providing greater flexibility. We fully support the FDA's continued requirement of adequate and well-controlled confirmatory studies, complete with prospective, detailed specification of the entire randomized clinical trial design in a way that allows accurate and precise estimation of treatment effectiveness.

STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:3270–3296
Published online 11 August 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2626

Issues in the use of adaptive clinical trial designs

Scott S. Emerson*,†,‡

Department of Biostatistics, University of Washington, Box 357232, Seattle, Washington 9815, U.S.A.

SUMMARY

Sequential sampling plans are often used in the monitoring of clinical trials in order to address the ethical and efficiency issues inherent in human testing of a new treatment or preventive agent for disease. Group sequential stopping rules are perhaps the most commonly used approaches, but in recent years, a number of authors have proposed adaptive methods of choosing a stopping rule. In general, such adaptive approaches come at a price of inefficiency (almost always) and clouding of the scientific question (sometimes). In this paper, I review the degree of adaptation possible within the largely prespecified group sequential stopping rules, and discuss the operating characteristics that can be characterized fully prior to collection of the data. I then discuss the greater flexibility possible when using several of the adaptive approaches receiving the greatest attention in the statistical literature and conclude with a discussion of the scientific and statistical issues raised by their use.

- "Adaptive modification of clinical trial designs poses at least as many problems as it is intended to solve."

Adaptive Designs for Clinical Trials

Insightfully Innovative

or

Irrelevantly Impractical

Stuart Pocock

London School of Hygiene and Tropical Medicine

Adaptive Designs for Clinical Trials

Insightfully Innovative

Occasionally

or

Irrelevantly Impractical

Often

Are Adaptive Designs Useful?

flexibility appeals to trial sponsors,
especially when trial is in “unexplored territory”

new methodology is fun

but they break the rules:
interim results highly confidential

statistical penalties: need to preserve type I error
practical issues: preserving trial’s integrity

Potential Problems with Adaptive Approach

sponsor stays blinded throughout?

If algorithm known, others can infer (guess) interim results

risk of wider unblinding in effect

consequences re:
- investigators
- sponsor
- investment analysts
- others

could the trial itself be compromised?

STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:3305–3312
Published online 18 July 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2641

Standard *versus* adaptive monitoring procedures: A commentary[§]

Thomas R. Fleming*,†,‡

Department of Biostatistics, University of Washington, Seattle, Washington, U.S.A.

SUMMARY

In the standard approach to designing definitive clinical trials, the primary endpoint and test statistic to be used for the primary analysis are specified before trial initiation. The false positive error rate for the null hypothesis and statistical power to detect the targeted size of treatment effect are also specified. Standard monitoring procedures, such as the group sequential guidelines, enable interim monitoring while maintaining the integrity of this approach. In contrast, adaptive monitoring procedures seek to provide flexibility to modify these pre-specified design features during the course of the trial. However, these procedures have several undesirable properties, including lesser statistical efficiency, reduced interpretability of primary outcome results, basing design changes on unreliable interim estimates of efficacy, risks to the integrity and credibility of the trial, loss of flexibility to use emerging results from external sources to alter key design features, and overemphasis of the importance of statistical significance relative to clinical significance.

Is adaptive design allowing sample size increases based on observed treatment effects more efficient?

Adaptive Designs

Musings and Myths

Robert T. O'Neill Ph.D.

Presented at the P*h*RMA 'Adaptive Designs Workshop :

Myth

That adaptive designs will be more efficient and will increase success rates of late phase trials

- Statistical arguments can illustrate how and when adaptations may be beneficial or problematic
- There are risks to using AD's that may outweigh a traditional design or approach
  - Interim decisions that are faulty
  - Complexity of planning and conduct
  - Trust and integrity - leading to validity and interpretability issues

Biometrika (2003), **90**, 2, *pp.* 367–378

Printed in Great Britain

On the inefficiency of the adaptive design for monitoring clinical trials

BY ANASTASIOS A. TSIATIS

Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695, U.S.A.

tsiatis@stat.ncsu.edu

AND CYRUS MEHTA

Cytel Software Corporation, 675 Massachusetts Avenue, Cambridge, Massachusetts 02139, U.S.A.

mehta@cytel.com

SUMMARY

Adaptive designs, which allow the sample size to be modified based on sequentially computed observed treatment differences, have been advocated recently for monitoring clinical trials. Although such methods have a great deal of appeal on the surface, we show that such methods are inefficient and that one can improve uniformly on such adaptive designs using standard group-sequential tests based on the sequentially computed likelihood ratio test statistic.

"For any adaptive design, one can always construct a standard group-sequential test that will reject the null hypothesis earlier with higher probability, and, for any parameter value not in the space of alternatives, will accept the null hypothesis earlier with higher probability."

STATISTICS IN MEDICINE
Statist. Med. 2003; **22**:971–993 (DOI: 10.1002/sim.1457)

Mid-course sample size modification in clinical trials based on the observed treatment effect

Christopher Jennison[1,*,†] and Bruce W. Turnbull[2]

[1] *Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, U.K.*
[2] *Department of Statistical Science, 227 Rhodes Hall, Cornell University, Ithaca, New York 14853-3801, U.S.A.*

SUMMARY

It is not uncommon to set the sample size in a clinical trial to attain specified power at a value for the treatment effect deemed likely by the experimenters, even though a smaller treatment effect would still be clinically important. Recent papers have addressed the situation where such a study produces only weak evidence of a positive treatment effect at an interim stage and the organizers wish to modify the design in order to increase the power to detect a smaller treatment effect than originally expected. Raising the power at a small treatment effect usually leads to considerably higher power than was first specified at the original alternative. Several authors have proposed methods which are not based on sufficient statistics of the data after the adaptive redesign of the trial. We discuss these proposals and show in an example how the same objectives can be met while maintaining the sufficiency principle, as long as the eventuality that the treatment effect may be small is considered at the design stage. The group sequential designs we suggest are quite standard in many ways but unusual in that they place emphasis on reducing the expected sample size at a parameter value under which extremely high power is to be achieved. Comparisons of power and expected sample size show that our proposed methods can out-perform L. Fisher's 'variance spending' procedure. Although the flexibility to redesign an experiment in mid-course may be appealing, the cost in terms of the number of observations needed to correct an initial design may be substantial.

Biometrics 62, 664–683
September 2006

Are Flexible Designs Sound?

Carl-Fredrik Burman* **and Christian Sonesson**

AstraZeneca R & D, SE-431 83 Mölndal, Sweden
**email:* carl-fredrik.burman@astrazeneca.com

Summary. Flexible designs allow large modifications of a design during an experiment. In particular, the sample size can be modified in response to interim data or external information. A standard flexible methodology combines such design modifications with a weighted test, which guarantees the type I error level. However, this inference violates basic inference principles. In an example with independent $N(\mu, 1)$ observations, the test rejects the null hypothesis of $\mu \leq 0$ while the average of the observations is negative. We conclude that flexible design in its most general form with the corresponding weighted test is not valid. Several possible modifications of the flexible design methodology are discussed with a focus on alternative hypothesis tests.

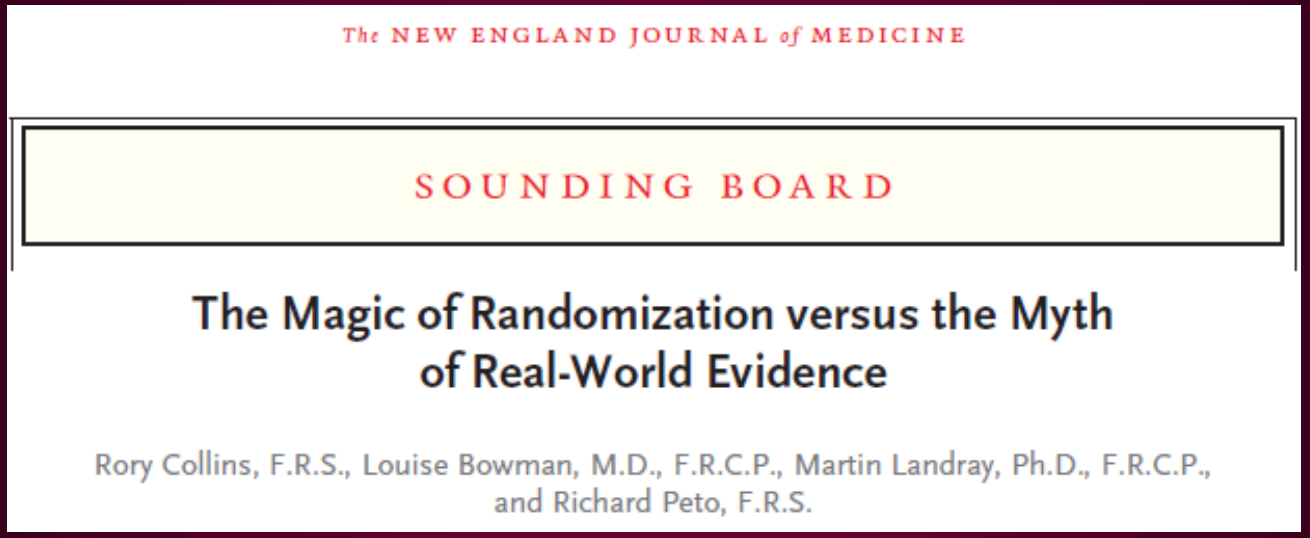
Is adaptive design allowing sample size increases based on observed treatment effects more efficient?

No.

How about response adaptive randomization?

Randomization: The Most Powerful Tool in Research

- Expectation of balance between randomized groups with respect to:
  - Known factors
  - UNKNOWN factors
    - Protects us from our own ignorance and knowledge limitations
  - Factors that cannot be measured (and thus cannot be controlled)



The NEW ENGLAND JOURNAL of MEDICINE

SOUNDING BOARD

The Magic of Randomization versus the Myth of Real-World Evidence

Rory Collins, F.R.S., Louise Bowman, M.D., F.R.C.P., Martin Landray, Ph.D., F.R.C.P., and Richard Peto, F.R.S.

Randomization

↓

Expectation of balance of factors

↓

Foundation for statistical inference.
Provides basis for error and power control.

Maintaining the Benefits of Randomization

The benefits of randomization (error and power control) can be surrendered through (sometimes ill-advised) choices regarding trial design, conduct, and analyses.

When Randomized Trials Do Not have the Benefits and Integrity of Randomization

- PP or as-treated analyses (non-ITT)
- Measuring the outcome at the end of treatment
  - Confounding by time
  - Fix outcome measurement at a fixed time from randomization
  - If of interest then include response time as an outcome
- Post randomization covariates
- Cluster randomization
  - 2 X 1000 vs. 1000 X 2
  - Number of clusters is the important parameter

When Randomized Trials Do Not have the Benefits and Integrity of Randomization

- Bayesian designs make assumptions regarding prior beliefs and the distribution of treatment effects and do not retain error control
  - Surrender the objectivity of data speaking for itself

- Response adaptive randomization
  - Confounding by time. Modeling is used to control the time effect but is an imperfect substitute for randomization
    - Imagine variant differences in COVID
    - Sites that begin and end enrollment at different times created imbalances in geographic, demographic, cultural factors

- Platform trials that use historical data

- Example: REMAP-CAP did all three but reported as “randomized”
  - Misleading to present evidence as if it has the integrity of an RCT

The Adaptive designs CONSORT Extension (ACE) statement: a checklist with explanation and elaboration guideline for reporting randomised trials that use an adaptive design

Munyaradzi Dimairo,[1] Philip Pallmann,[2] James Wason,[3,4] Susan Todd,[5] Thomas Jaki,[6] Steven A Julious,[1] Adrian P Mander,[2,3] Christopher J Weir,[7] Franz Koenig,[8] Marc K Walton,[9] Jon P Nicholl,[1] Elizabeth Coates,[1] Katie Biggs,[1] Toshimitsu Hamasaki,[10] Michael A Proschan,[11] John A Scott,[12] Yuki Ando,[13] Daniel Hind,[1] Douglas G Altman,[14] on behalf of the ACE Consensus Group

Cite this as: *BMJ* 2020;369:m115
http://dx.doi.org/10.1136/bmj.m115

Adaptive designs (ADs) allow pre-planned changes to an ongoing trial without compromising the validity of conclusions and it is essential to distinguish pre-planned from unplanned changes that may also occur. The reporting of ADs in randomised trials is inconsistent and needs improving. Incompletely reported AD randomised trials are difficult to reproduce and are hard to interpret and synthesise. This consequently hampers their ability to inform practice as well as future research and contributes to research waste. Better transparency and adequate reporting will enable the potential benefits of ADs to be realised.

Table 1. Guidelines for the Reporting of Adaptive Trials[a]

Describe	The adaptation
	Whether the adaptation was planned or unplanned
	The rationale for the adaptation
	When the adaptation was made
	The data on which adaptation is based and whether the data were unblinded
	The planned process for the adaptation including who made the decision regarding adaptation
	Deviations from the planned process
	Consistency of results before vs after the adaptation
Discuss	Potential biases induced by the adaptation
	Adequacy of firewalls to protect against operational bias
	The effects on error control and multiplicity context

[a]Adapted from Evans and Ting [14].

Response Adaptive Randomization (RAR)

- "Play-the-winner" (Zelen, *JASA* 1969) and "urn design" (LJ Wei)
- Treatment assignment is based on the observed responses of participants that are already enrolled
- Ethical advantage?
  - More patients randomized to the more effective intervention?

Response Adaptive Randomization (RAR)

- Disadvantages
  - Statistical inefficiency
  - Biased estimates of effects that are sensitive to temporal drift occurring with e.g., evolving resistance, advancements in supportive care, noteworthy for long platform trials
  - Geographic factor imbalances that occur when sites initiate after or complete before the randomization ratio modification, a concern in long-duration platform trials
  - Operational bias in open-label trials since treatment effects may be inferred from observed randomization rates

RAR designs produce allocations for which estimates of treatment effects are biased in the presence of a time trend.

"Due to fears of bias from confounding with time trends, RAR cannot substitute for true randomization in confirmatory trials."

Professor Richard Chappell; Past-president, Society for Clinical Trials

"Most patients are not in any trial… So, if trials are going to have any large impact on practice, the sooner they yield reliable results the better.

This alone provides an important reason to keep the allocation probabilities even, for statistical sensitivity will thereby be maximized."

Sir Richard Peto, Emeritus Professor of Medical Statistics, Oxford (1985)

Editorial

CLINICAL TRIALS

Ethical concerns about adaptive randomization

Clinical Trials
2015, Vol. 12(2) 101

DOI: 10.1177/1740774515569613
ctj.sagepub.com
SAGE

Colin B Begg

Colin Begg
Editor, Clinical Trials
Past President, Society for Clinical Trials

"Adaptive randomization challenges the whole notion of equipoise and as such challenges the entire basis for randomization. In short, it is an insidious threat to the most important tool in the clinical research armamentarium."

Scores of papers by NIH and academics have not remedied the problem

Outcome-Adaptive Randomization: Is It Useful?

Edward L. Korn and Boris Freidlin

See accompanying editorial on page 606

A B S T R A C T

Outcome-adaptive randomization is one of the possible elements of an adaptive trial design in which the ratio of patients randomly assigned to the experimental treatment arm versus the control treatment arm changes from 1:1 over time to randomly assigning a higher proportion of patients to the arm that is doing better. Outcome-adaptive randomization has intuitive appeal in that, on average, a higher proportion of patients will be treated on the better treatment arm (if there is one). In both the randomized phase II and phase III settings with a short-term binary outcome, we compare outcome-adaptive randomization with designs that use 1:1 and 2:1 fixed-ratio randomizations (in the latter, twice as many patients are randomly assigned to the experimental treatment arm). The comparisons are done in terms of required sample sizes, the numbers and proportions of patients having an inferior outcome, and we restrict attention to the situation in which one treatment arm is a control treatment (rather than the less common situation of two experimental treatments without a control treatment). With no differential patient accrual rates because of the trial design, we find no benefits to outcome-adaptive randomization over 1:1 randomization, and we recommend the latter. If it is thought that the patient accrual rates will be substantially higher because of the possibility of a higher proportion of patients being randomly assigned to the experimental treatment (because the trial will be more attractive to patients and clinicians), we recommend using a fixed 2:1 randomization instead of an outcome-adaptive randomization.

VOLUME 29 · NUMBER 6 · FEBRUARY 20 2011

JOURNAL OF CLINICAL ONCOLOGY

STATISTICS IN ONCOLOGY

"RAR has several undesirable properties…including a high probability of a sample size imbalance in the wrong direction, which might be surprising to nonstatisticians, wherein many more patients are assigned to the inferior treatment arm, the opposite of the intended effect.

Compared with an equally randomized design, RAR produces less reliable final inferences, including a greatly overestimated treatment effect and smaller power to detect a treatment difference. This estimation bias becomes much larger if the prognosis of the accrued patients either improves or worsens systematically during the trial.

RAR produces inferential problems that decrease potential benefit to future patients, and may decrease benefit to patients enrolled in the trial.

For randomized trials to obtain confirmatory comparisons, designs with fixed randomization probabilities and group sequential decision rules appear to be preferable to AR, scientifically, and ethically."

Professor Peter Thall, MD Anderson Cancer Center

Clinical Infectious Diseases

INVITED ARTICLE

IDSA Infectious Diseases Society of America

hivma hiv medicine association

OXFORD

INNOVATIONS IN DESIGN, EDUCATION AND ANALYSIS (IDEA): Victor De Gruttola and Scott R. Evans, Section Editors

Resist the Temptation of Response-Adaptive Randomization

Michael Proschan[1] **and Scott Evans**[2]

[1]Mathematical Statistician, Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, Rockville, Maryland, USA, and [2]Department of Biostatistics and Bioinformatics; Director, Biostatistics Center, Milken Institute School of Public Health, George Washington University, Washington, DC, USA

Response-adaptive randomization (RAR) has recently gained popularity in clinical trials. The intent is noble: minimize the number of participants randomized to inferior treatments and increase the amount of information about better treatments. Unfortunately, RAR causes many problems, including (1) bias from temporal trends, (2) inefficiency in treatment effect estimation, (3) volatility in sample-size distributions that can cause a nontrivial proportion of trials to assign more patients to an inferior arm, (4) difficulty of validly analyzing results, and (5) the potential for selection bias and other issues inherent to being unblinded to ongoing results. The problems of RAR are most acute in the very setting for which RAR has been proposed, namely long-duration "platform" trials and infectious disease settings where temporal trends are ubiquitous. Response-adaptive randomization can eliminate the benefits that randomization, the most powerful tool in clinical trials, provides. Use of RAR is discouraged.

"The proponents of data-dependent assignment rely on an ethical argument which appears to be fallacious. Before the trial begins, clinicians are unaware of which treatment is superior, and so random allocation can be justified. After a trial which has established a treatment difference, the ethical policy will usually be to offer the superior treatment to all patients.

During the trial the evidence for superiority of one of the treatments will accumulate, and the clinician might choose to these data to help in his allocation. However, the clinician has only two choices: to randomize or to prescribe the currently more successful treatment with certainty. If assignment to the currently inferior treatment is regarded as unethical, then a reduction of the probability of this event will not redeem it."

John Whitehead, Emeritus Professor of Statistics, Lancaster University

Is adaptive design allowing sample size increases based on observed treatment effects more efficient?

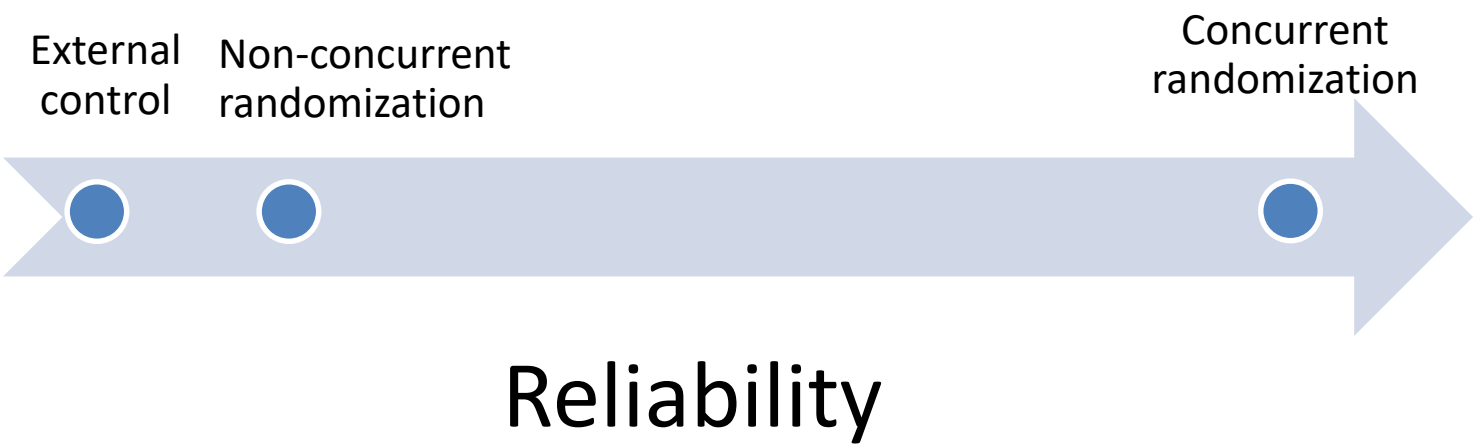
No.

How about response adaptive randomization?

No.

More data is always better?

Utilizing Concurrent Control Data



Some have used non-concurrent information to make comparisons though this removes the integrity of randomization.

Platform Trials — Beware the Noncomparable Control Group

tween trial treatment groups. Statistical modeling of trends over time and country effects can attempt to ameliorate potential bias due to a noncomparable control group, but there are two weaknesses of this approach. The first is that the more modeling conducted, the less efficient the design is in terms of required sample sizes.[3] More importantly, one never knows whether the modeling has successfully eliminated all potential bias.

An example of how nonconcurrent randomization (and the other potential aforementioned biases) can complicate interpretation of trial results is the analysis of the efficacy of interleukin-6 blockade with tocilizumab or sarilumab in patients with Covid-19 in the Randomized, Embedded, Multifactorial Adaptive Platform Trial for Community-Acquired Pneumonia (REMAP-CAP). In this international platform trial, the control group used in the analysis was not restricted to patients who had undergone concurrent randomization, and both trial agents were reported to improve survival.[5] This trial led some treatment guideline committees to recommend the use of these agents (www.cas.mhra.gov.uk/ViewandAcknowledgment/ViewAttachment.aspx?Attachment_id=103745). One cannot say with certainty that the statistical modeling was not successful in eliminating bias, especially in a complex and hard to understand platform trial such as REMAP-CAP. The added value from this trial relative to the other randomized trials with straightforward, comparable controls can be questioned.

Lori E. Dodd, Ph.D.
National Institute of Allergy and Infectious Diseases
Bethesda, MD
doddl@niaid.nih.gov

Boris Freidlin, Ph.D.
Edward L. Korn, Ph.D.
National Cancer Institute
Bethesda, MD

Disclosure forms provided by the authors are available with the full text of this letter at NEJM.org.

1. Freidlin B, Korn EL, Gray R, Martin A. Multi-arm clinical trials of new agents: some design considerations. Clin Cancer Res 2008;14:4368-71.
2. Asch DA, Sheils NE, Islam MN, et al. Variation in US hospital mortality rates for patients admitted with COVID-19 during the first 6 months of the pandemic. JAMA Intern Med 2020 December 22 (Epub ahead of print).
3. Korn EL, Freidlin B. Outcome — adaptive randomization: is it useful? J Clin Oncol 2011;29:771-6.
4. Angus DC, Derde L, Al-Beidh F, et al. Effect of hydrocortisone on mortality and organ support in patients with severe COVID-19: the REMAP-CAP COVID-19 corticosteroid domain randomized clinical trial. JAMA 2020;324:1317-29.
5. The REMAP-CAP Investigators. Interleukin-6 receptor antagonists in critically ill patients with Covid-19. N Engl J Med 2021;384:1491-502.

DOI: 10.1056/NEJMc2102446

Preserving the Integrity of Randomization

- Comparison for given drug should (typically) be against only those control patients who were eligible for and could have been randomized to drug
- Example: issues with two-stage consent process
  - Impact of knowledge of specific drug/subprotocol on consent may result in lack of comparability between drug and shared control
  - Alternative to avoid issue: single informed consent covering all active drugs prior to randomization
- Example: scenario with fixed randomization scheme resulting in changes in ratio between given drug and control over time
  (e.g., randomization ratio of sqrt(*k*):1:…:1 with *k* active drugs)
  - Analyses comparing drug to control should account for time periods with different randomization ratios, such as by stratifying by time periods defined by changes in number of active drugs

Leveraging Non-Concurrent Control Data

- Likely most reasonable in settings with different bias-variance tradeoffs such as trials in rare diseases and early-phase trials
- Rationale to leverage non-concurrent control data should include:
  - Likelihood of changes over time in prognostic factors
  - Feasibility of utilizing only concurrent control data
  - Extent of non-concurrent control data expected to be utilized
  - Ability and underlying assumptions of primary analysis methods to mitigate confounding and plan for sensitivity analyses to evaluate varying assumptions (e.g., less weighting of non-concurrent data)

Is adaptive design allowing sample size increases based on observed treatment effects more efficient?

No.

How about response adaptive randomization?

No.

More data is always better?

No.

How about Bayesian design?

Bayesian Approaches

- Start with investigator belief (prior) about the treatment effect
- Conduct a study
- Update the belief (posterior)

Bayesian Approaches

- Generally too subjective and lacking in conformity for confirmatory evidence

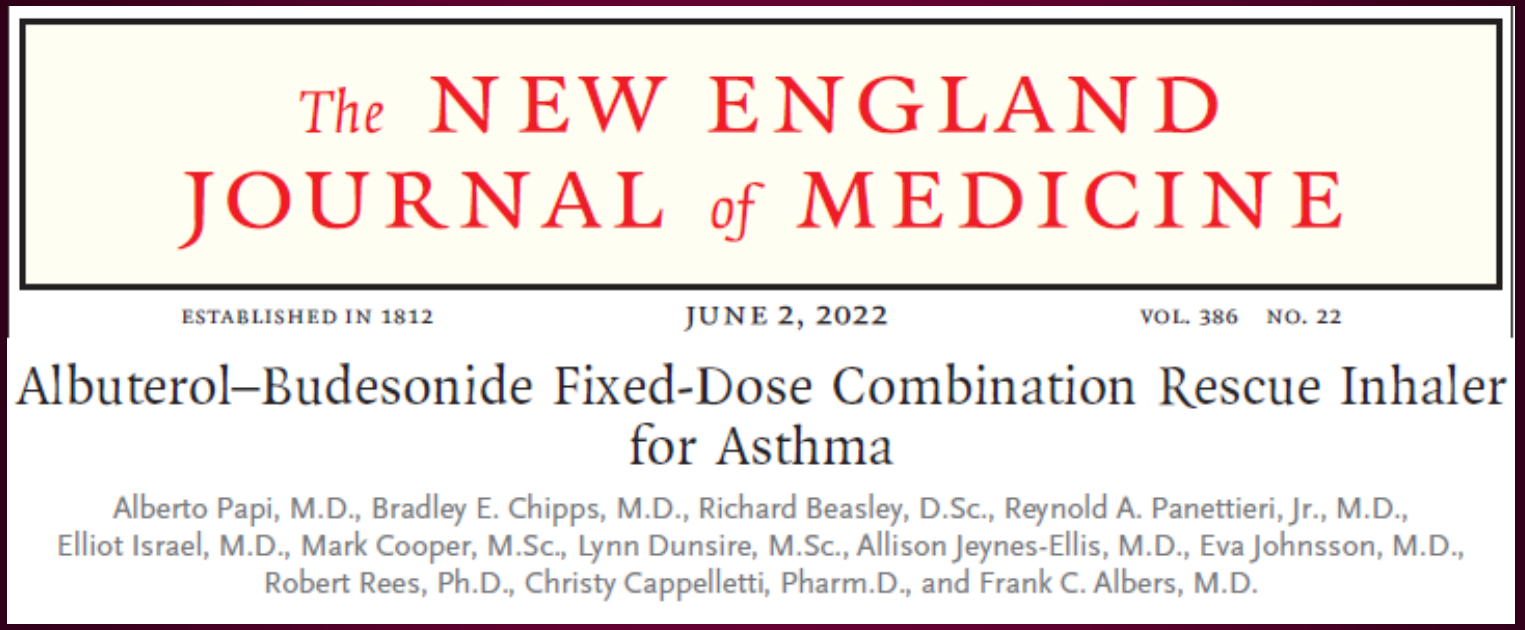
- People have different beliefs
  - Sponsor belief ≠ regulator belief ≠ public belief ≠ your/my belief ≠ …
  - Sponsor chooses one
  - May assume treatment effects are equally likely to dampen effect (noninformative)
    - Still affect precision of estimates (inflated precision from beliefs)
    - Impairs ability to evaluate the role of chance

- Danger of conflating belief with desired effect (it works!)
  - Tricky for NI (belief = no difference?)

Bayesian Approaches

- No analytical solution to control error rates; relies on simulation
  - More appropriate for early-phase / exploratory evaluations
  - Confirmatory evidence still necessary

- Issues with transparency and robustness
  - Relies on black box modeling and untestable assumptions
    - Sharing computational algorithms may help



The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 JUNE 2, 2022 VOL. 386 NO. 22

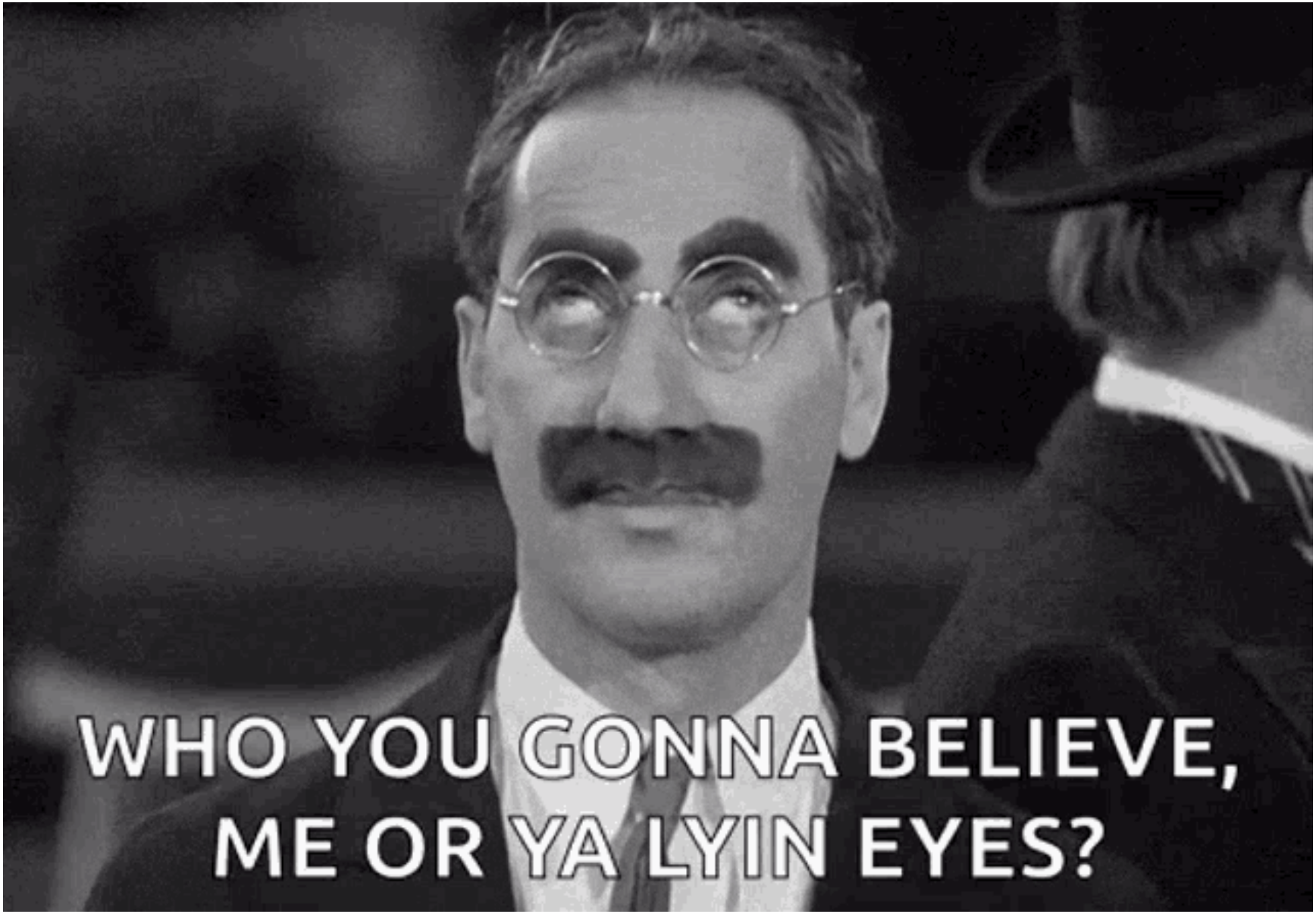
Albuterol–Budesonide Fixed-Dose Combination Rescue Inhaler for Asthma

Alberto Papi, M.D., Bradley E. Chipps, M.D., Richard Beasley, D.Sc., Reynold A. Panettieri, Jr., M.D., Elliot Israel, M.D., Mark Cooper, M.Sc., Lynn Dunsire, M.Sc., Allison Jeynes-Ellis, M.D., Eva Johnsson, M.D., Robert Rees, Ph.D., Christy Cappelletti, Pharm.D., and Frank C. Albers, M.D.

- Trial included adults and adolescents
- High dose worked well for adults, but
- Adolescents: increased severe asthma exacerbation by 44%

  HR: 1.44 95% CI = (0.54, 3.87)
- Imposing investigator belief results in an estimate of a 14% decrease:

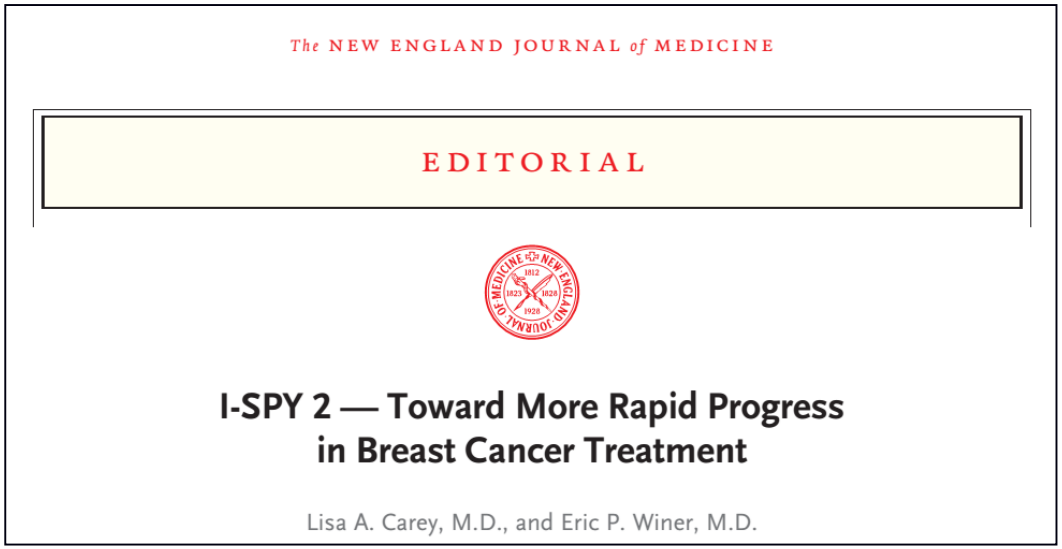
  HR: 0.86 95% CI = (0.62, 1.48)
- Rejected at FDA Advisory Committee


WHO YOU GONNA BELIEVE,
ME OR YA LYIN EYES?

A BAD Design

Marc Buyse

Founder, International Drug Development Institute (IDDI)



The NEW ENGLAND JOURNAL of MEDICINE

EDITORIAL

I-SPY 2 — Toward More Rapid Progress in Breast Cancer Treatment

Lisa A. Carey, M.D., and Eric P. Winer, M.D.

ISSUES WITH BAYESIAN ADAPTIVE DESIGNS (BAD)

- Needs
  - reliable data to form and rank biomarker groups
  - prior distribution of treatment effects
- Operating characteristics of design unknown
- Treatment comparisons potentially biased
- Analysis
  - must be model-based
  - cannot readily contribute to meta-analyses
- Adaptive randomization
  - is not ethically preferable
  - is not statistically desirable

Bayesian Approach

Brad Efron (Professor, Stanford University; ASA president)

“The moral high ground of scientific objectivity has been seized by the frequentists.”

"Scientific objectivity" is more than a catch-phrase. Strict objectivity is one of the crucial factors separating scientific thinking from wishful thinking.

Complete objectivity about one's own work is a little much to expect from a human being, even a scientist, but it is not too much to expect from one's colleagues. A prime requirement of any statistical theory intended for scientific use is that it reassures oneself and others that the data have been interpreted fairly.”

Bayesian Approach

"Although the Bayesian method has attractions - it must be recognized that its use does affect the chances of making a false conclusion and introduces biases to estimates of treatment effect. The Bayesian philosophy is not concerned with error rates and biases, which is not the same as saying that it avoids them."

"Emphasis is on getting things right at the design stage, in which case the freedom offered by adaptive designs should not be required. There is a price to be paid for that freedom, in terms of reduced statistical efficiency."

John Whitehead, Emeritus Professor of Statistics, Lancaster University

(Nature Reviews, 2004)

Bayesian Analyses and DSMBs

- Issues for DSMBs
- Principles of DSMB operations
  - Independence
  - Objective as possible
- If investigator beliefs are imposed on presented results, then:
  - Can assessment be objective?
  - Is independence violated?
- Lacks analytical error control

Is adaptive design allowing sample size increases based on observed treatment effects more efficient?

No.

How about response adaptive randomization?

No.

More data is always better?

No.

How about Bayesian design?

“It’s a mess.”
(Anonymous FDA colleague)

RCTs, Observational Studies, and Adaptive Designs

- RCTs
  - Advantages: eliminates many biases of observational studies when protected
  - Disadvantage: Time-consuming and expensive

- Observational studies
  - Advantage: faster and cheaper than an RCT
  - Disadvantage: subject to many biases

	Error Control	
	Yes	No
Inexpensive		Observational Studies
Expensive	RCTs	

	Error Control	
	Yes	No
Inexpensive		Observational Studies RWE (?)
Expensive	RCTs	

	Error Control	
	Yes	**No**
Inexpensive	**Pragmatic RCTs (?)**	**Observational Studies RWE (?)**
Expensive	**RCTs**	

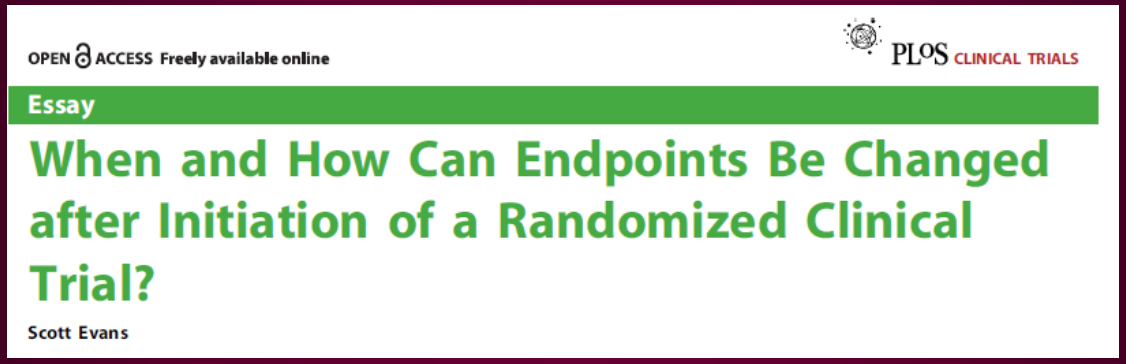
- Pragmatic RCTs … best of both worlds?
  - Right idea
  - Questions remain about expense
  - More a question about the use of ITT, clinical endpoints, etc. than the data source. Often conflated.

	Error Control	
	Yes	No
Inexpensive	Pragmatic RCTs (?) Platform RCTs (?)	Observational Studies RWE (?)
Expensive	RCTs	

Adaptive platform designs can combine the advantages of both …

	Error Control	
	Yes	No
Inexpensive	Pragmatic RCTs (?) Platform RCTs (?)	Observational Studies RWE (?)
Expensive	RCTs	Some Platform RCTs

Adaptive platform designs can combine the advantages of both …
or the disadvantages of both.



- Changes to endpoints can compromise the scientific integrity of a trial
- New information could merit endpoint changes
  - E.g., long-term trials w/ evolving medical knowledge (e.g., new biomarkers)
- OK if decision is *independent* of treatment effect (e.g., based on external data)
  - Wittes, 2002 (*SiM*): "may consider changes to the primary endpoint when the trial has airtight procedures to guarantee separation of the people making the decision from data providing insight into the treatment effect"
- Examples
  - Post-CABG (post coronary artery bypass graft)
  - 2-stage trials where stages are independent

DSMB

- Thus DSMB is not the right group to make this decision
- For similar reasons they do not adjudicate study endpoints

2-Stage Designs

- "Internal pilot": Stage 1 vs. Stage II: learn vs. confirm
  - Hypothesis generation vs. hypothesis testing

- Efficiency advantage
  - Single trial addresses objectives traditionally addressed in two trials
  - Eliminates down-time between separate trials (but less thinking time)
  - IRB advantage (vs. approval of two trials)

Phase II Evaluation of Low-Dose Oral Etoposide for the Treatment of Relapsed or Progressive AIDS-Related Kaposi's Sarcoma: An AIDS Clinical Trials Group Clinical Study

By Scott R. Evans, Susan E. Krown, Marcia A. Testa, Timothy P. Cooley, and Jamie H. Von Roenn

Journal of Clinical Oncology, Vol 20, No 15 (August 1), 2002: pp 3236-3241
DOI: 10.1200/JCO.2002.12.038

- Phase II single-arm trial of oral etoposide for AIDS KS
- Endpoint: tumor response rate (50% decrease in lesion number/size)
- 2-stages
  - Stage I: Enrolled N=14
    - If response is unacceptably low (0/14), then quit for futility
    - If true response rate is 20% then <5% chance of observing 0/14
  - Otherwise continue to Stage II (no efficacy testing)
- Minimizes expected sample size when response is low given error constraints
- Trial continued w/ final response rate = 36%

2-Stage Design for Diagnostics

- Assay results may be measured quantitatively
- Stage I
  - Identify optimal cut-off for discrimination
    - Decision based on weighing consequences of false positive and false negative errors
- Stage 2
  - Validate the utility of the diagnostic when using the identified cut-off using data from Stage II (independent from Stage 1)
- Important to separate hypothesis generation from hypothesis testing accounting for selective nature of the discrimination rule
  - Learn vs. confirm

Adaptive design is not all bad.

We have not told anyone but in fact…
the ARLG is conducting an adaptive trial!

ARLG PHAGE

- 2-stage design with adaptive features
  1. Dose selection
  2. Sample size based on precision

FDA Adaptive Design Guidance Principles

- Design, conduct, and analysis of adaptive trials intended to provide substantial evidence of effectiveness should satisfy four key principles:

1. Chance of erroneous conclusions should be controlled
2. Treatment effect estimates should be sufficiently reliable
3. Details of the design should be completely prespecified
4. Trial integrity should be appropriately maintained

FDA Guidance
Maintaining Integrity

- Strongly recommended that access to comparative interim results be limited to individuals with relevant expertise who are independent of the personnel involved in conducting or managing the trial

- This provides the greatest confidence that potential unplanned design modifications are not motivated in any way by accumulating data

- Limitation of access to comparative interim results provides the greatest assurance of quality trial conduct

FDA Guidance
Maintaining Integrity

- Establish safeguards to ensure that the persons responsible for preparing and reporting interim analysis results to the DMC or the adaptation committee are physically and logistically separated from the personnel tasked with managing and conducting the trial

- Recommended that no sponsor representatives have access to comparative interim results

- There is potential for knowledge of the adaptation decision to convey information about the interim results. Knowledge of a sample size modification algorithm and the adaptively chosen sample size, for example, can allow back-calculation of the interim estimate of the treatment effect. Steps should be taken where possible to minimize the information that can be inferred by observers.

Disseminating results

- Inadvertent dissemination of information from an ongoing trial conducted under a master protocol may pose a risk to trial integrity
- Example 1:
  - The primary endpoint is to death
  - Multiple drugs enter the platform trial at approximately the same time
  - Event-driven sample sizes for each (drug versus control) comparison
  - If target number of events is reached for one (drug versus control) comparison and the drug is superior, strongly suggests that other drugs still under evaluation are also superior as they have had event fewer deaths
- Example 2:
  - Unblinded results reported for first drug while second drug still under evaluation
  - Knowledge of comparative results for (first drug vs control) in addition to blinded pooled results for (second drug + control) may lead to partial unblinding of comparative results for (second drug versus control)

Dendreon Disclosure May Endanger Provenge Study

Matthew Herper Former Staff
I cover science and medicine, and believe this is biology's century.

Mar 25, 2009, 06:00am EDT

This article is more than 10 years old.

Has Dendreon , a money-losing drug developer, imperiled the main study of its experimental prostate cancer treatment Provenge?

On Oct. 6, Dendreon put out a press release that sent its shares up 33% in a day. Provenge appeared to cut patients' death rates by 20% compared with a placebo treatment, the company said. The release also contained statistical details that made good results seem likely when final results are released in April.

But four top statisticians now say Dendreon may have compromised the integrity of the trial by putting out the release. They say it was unorthodox for Dendreon to even know such a detailed result, much less to publicize it. The danger: The company, patients or doctors might have changed what they were doing once they knew how the study was going. If the final outcome is only marginally statistically significant, it might be tossed,

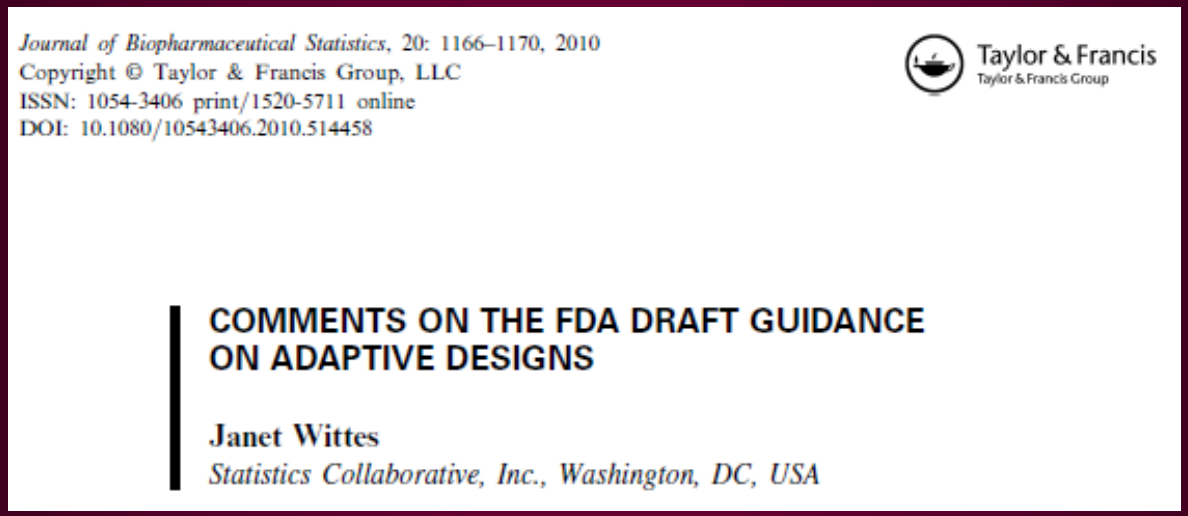
- "I have no idea what their rationale would have been," says Susan Ellenberg (University of Pennsylvania).

- Janet Wittes (Statistics Collaborative) says she was "shocked" by Dendreon's actions. "When we as a company are serving as the reporting statisticians for a trial and a company asks for this sort of information, we refuse to release it.“

The Seattle Times, August 7, 2005

Drug researchers leak secrets to Wall St

- violate medical ethics
- jeopardize vital research
- can introduce bias
- government regulators seem to know nothing about it



Journal of Biopharmaceutical Statistics, 20: 1166–1170, 2010

ISSN: 1054-3406 print/1520-5711 online
DOI: 10.1080/10543406.2010.514458

Taylor & Francis
Taylor & Francis Group

COMMENTS ON THE FDA DRAFT GUIDANCE ON ADAPTIVE DESIGNS

Janet Wittes
Statistics Collaborative, Inc., Washington, DC, USA

- A typical vignette from my own experience… the vignette stems from hype about the magic of adaptive designs.

- Companies in asking for help in designing studies come with a request for an adaptive design. Rather than starting with a discussion of the question the study is to address, they start with a question like, "Our investors want an adaptive design. Can you produce one?" The guidance will reinforce us statisticians when we answer, "Tell me the question. Then we can discuss the design." Further, the guidance will provide ammunition for the argument that caution is necessary.

- I hope that the message of the guidance — encouraging rational adaptation in early phase study but caution in Phase 3 trials — will reach investigators outside the FDA and will be embraced by reviewers inside

ICH E-20 in Development

Global harmonized regulatory guideline for adaptive trials

- Definitions related to adaptive clinical trials are inconsistent, and there are no common principles for adaptive clinical trials, especially in relationship to the risk of erroneous conclusions and maintenance of trial integrity

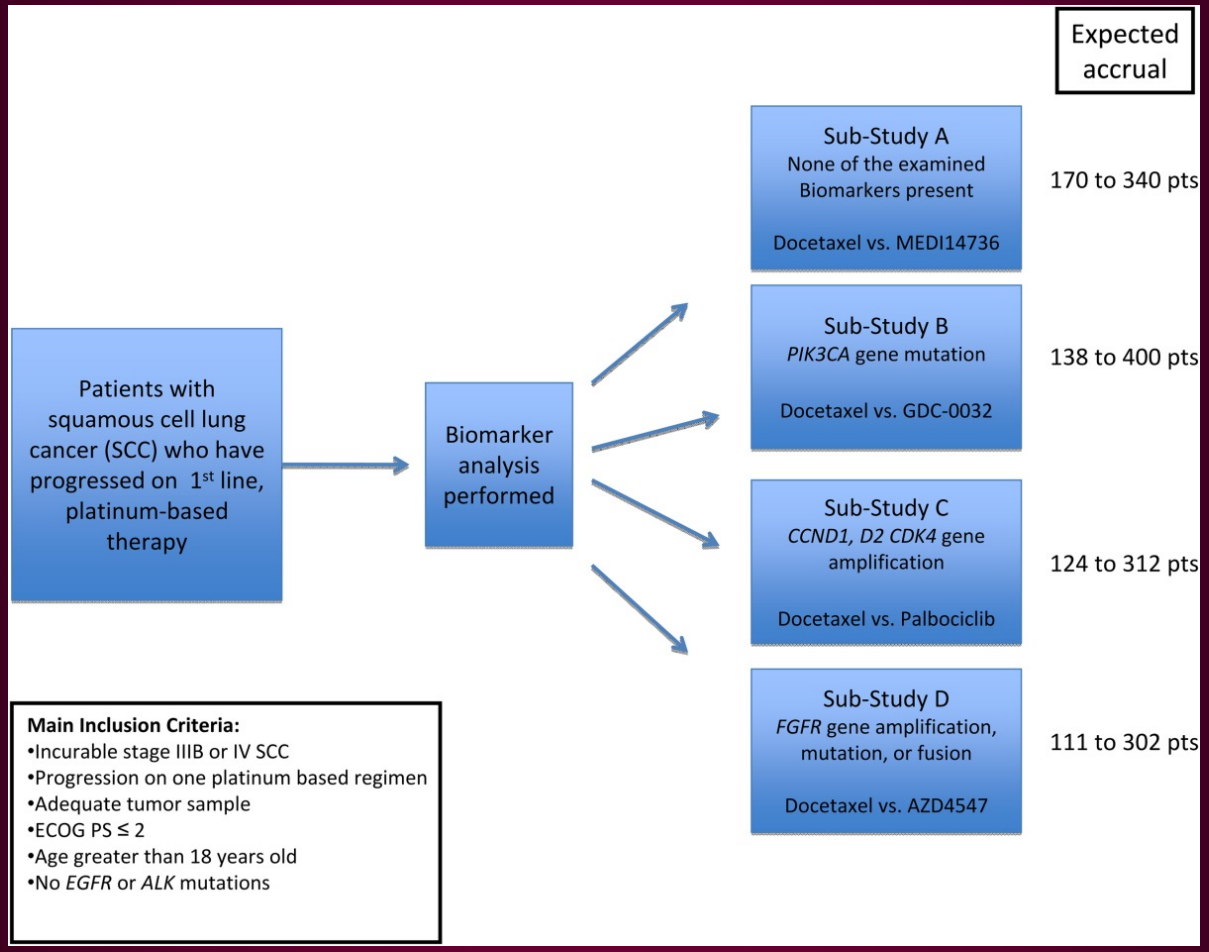
- ICH E-30 will address:
  - A common terminology
  - The potential benefits and harms and areas (e.g., study settings and design features) of meaningful applications
  - The principles for the design, conduct, analysis, and proper interpretation, including considerations of the risks of erroneous conclusions, maintenance of trial integrity, and handling of operational challenges
  - Important documentation for adaptive clinical trials

Umbrella Trials

- ≥ 2 sub-trials linked via a common screening infrastructure
- Patients screened for relevant characteristics and assigned to a targeted subtrial if eligible for that subtrial
- The screening process facilitates accrual for trials with low-prevalence diseases, acting as an enrichment process for the subtrials of interest
- Flexibility to add, drop, or modify the subtrials without affecting other subtrials
- Subtrials of the umbrella subtrials can use e.g., 2-stage designs, MAMS, SMARTs
- First used on cancer to tests the impact of different drugs on different mutations in a single cancer type
  - NCI MATCH: 20+ arms and uses Simon's optimal 2-stage design
- Low-risk; high reward

LUNG-MAP

Phase II/III Biomarker-Driven Master Protocol for 2nd Line Therapy of Squamous Cell Lung Cancer

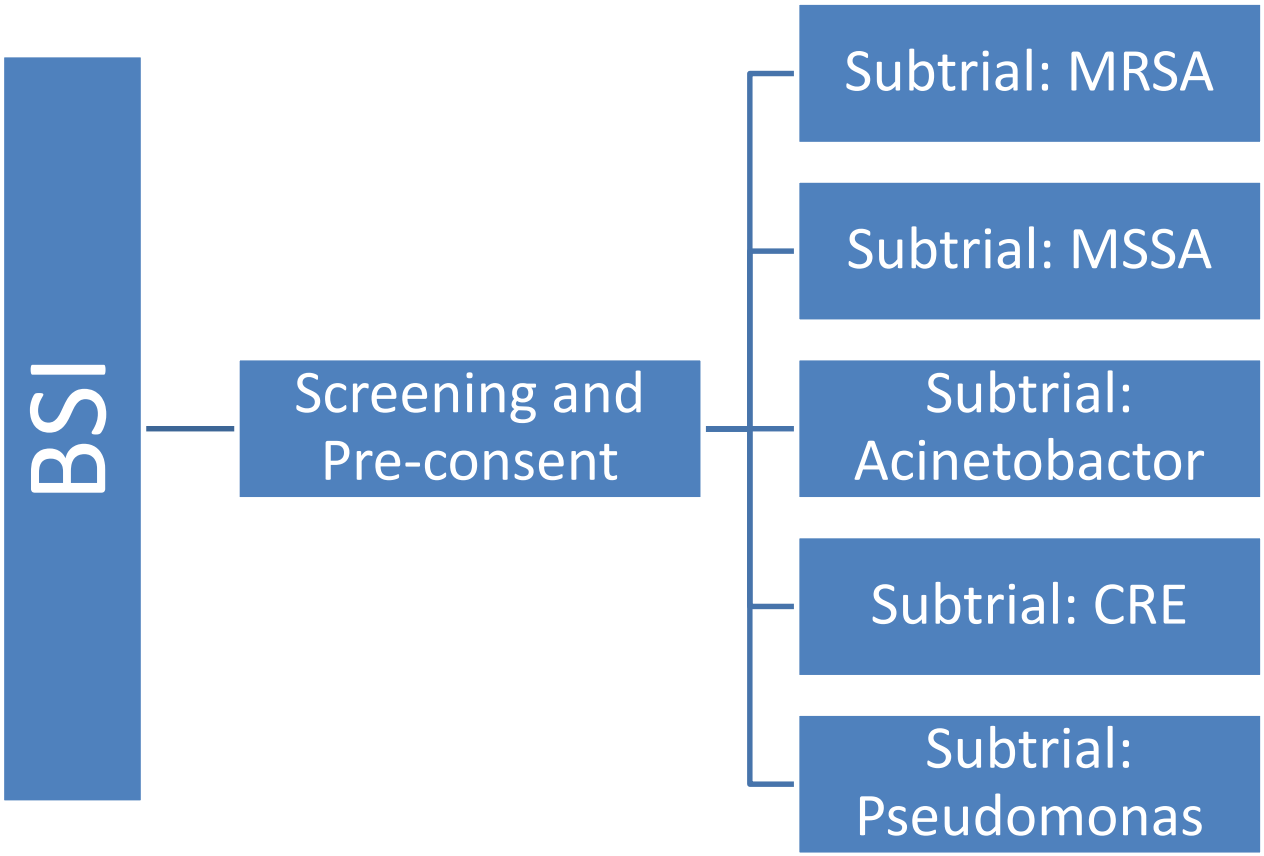


Umbrella for ARLG?

- Process
  1. Choose a site of infection
  2. Direct to sub-trial based on pathogen

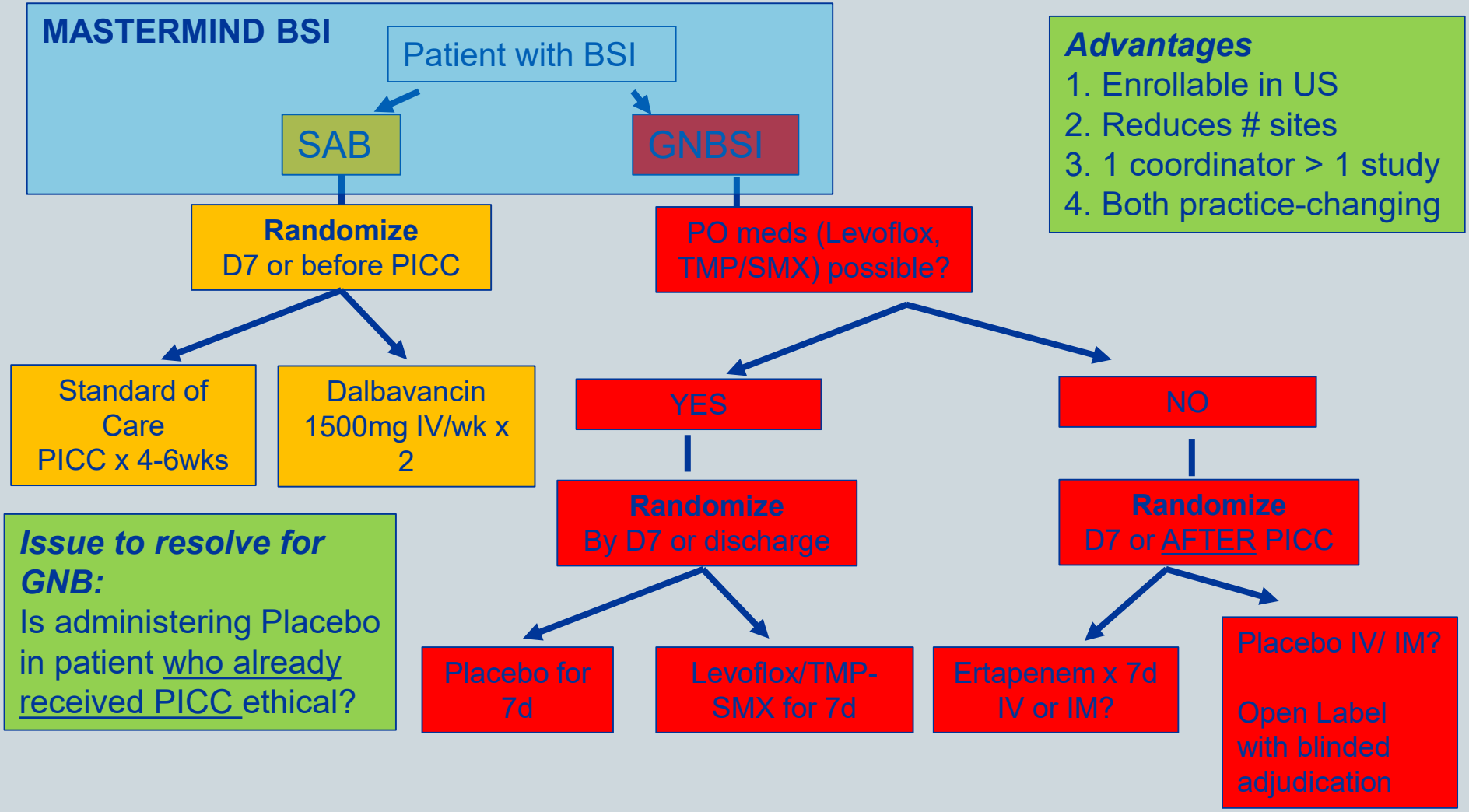
- Consider separate umbrella structure for major infection sites:
  1. cUTI
  2. cIAI
  3. HABP/VABP
  4. CABP
  5. ABSSSI
  6. BSI (ARLG to do the studies others won't; BSI not a regulatory indication)

Umbrella Design



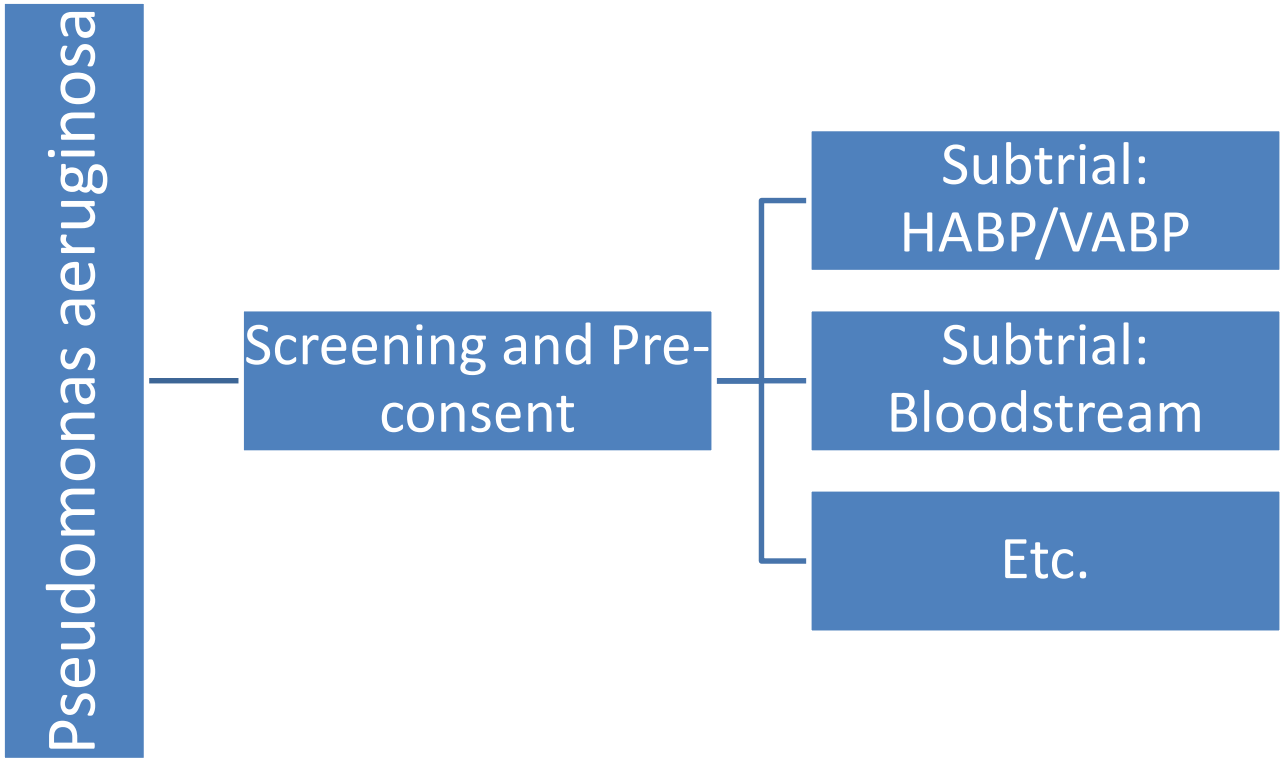
Dual Enrollment of Gram-Positive and Gram-Negative BSI Interventional Trials

S. aureus: Dalbavancin vs. SOC in Complicated SAB (DOTS)
Gram-Negative: Placebo-controlled trial of 7 v. 14d of abx for GNB-BSI



Umbrella?

What about the reverse … identify pathogen e.g., Pseudomonas aeruginosa … then direct to subtrial based on site of infection



Basket Trials: Oncology

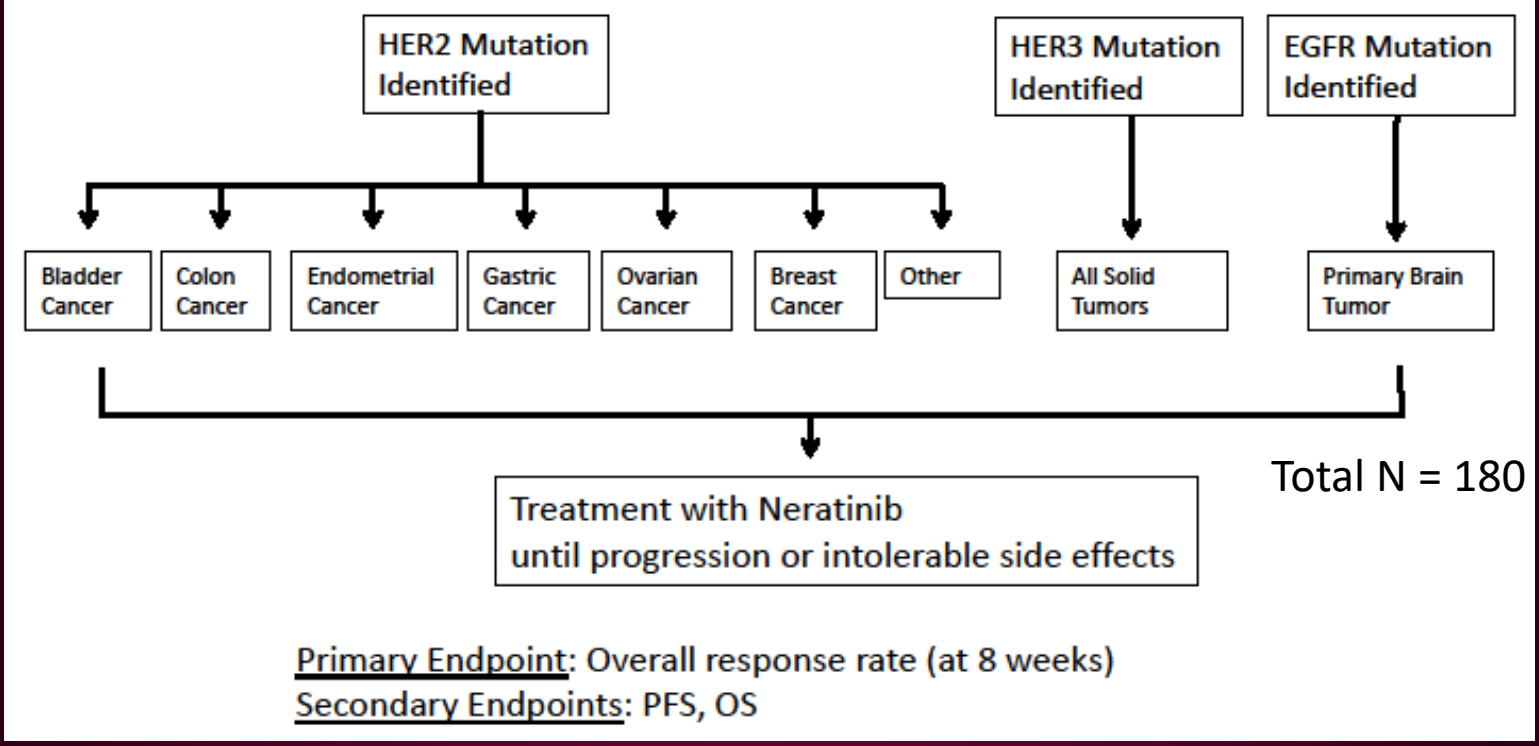
- Screens the effect of a single drug on a single mutation in a variety of cancer types (molecular classification of tumors)

- Rationale
  - Biological: focusing on mutation may be better than cancer type
  - Desperation: difficult to enroll type-specific studies

- High risk as assumptions via modeling are often made that effects in one cancer type can be used as evidence in another cancer type
  - Difficult to validate such assumptions

Basket Trial

NCT01953926: Phase 2 Study of Neratinib in Patients With Solid Tumors With [EGFR, HER2, HER3] Mutations or EGFR Gene Amplification



Basket for ARLG?

- We have selectively combined sites of infection in some studies (MDRO)
- Pooling infection sites to inform the benefits and harms of a drug?
  - Outcome in one infection site may inform for another site
  - Examples where drugs work in one infection site but not others
- Possible though risky

Basket for ARLG?

- Challenges
  - Populations differ
  - Prior therapy differs
  - Control arms differ
    - Definition of “treatment effect” (a contrast) is not well defined
  - NI tricky
    - NI in one infection site may dilute inferiority in another site
    - Limited power for identifying heterogeneity / interaction
  - Robustness issues relying on unvalidated assumptions
  - Inconsistent with regulator studies

- Benefit of the doubt or doubt of the benefit?

Clinical Infectious Diseases

INVITED ARTICLE

IDSA Infectious Diseases Society of America

hivma hiv medicine association

OXFORD

Innovations in Design, Education, and Analysis (IDEA): Scott R. Evans and Victor De Gruttola, Section Editors

Sequential, Multiple-Assignment, Randomized Trials for COMparing Personalized Antibiotic StrategieS (SMART-COMPASS)

Scott R. Evans,[1] Dean Follmann,[2] Ying Liu,[3] Thomas Holland,[4] Sarah B. Doernberg,[5] Nadine Rouphael,[6] Toshimitsu Hamasaki,[7] Yunyun Jiang,[1] Judith J. Lok,[8] Thuy Tien T. Tran,[1] Anthony D. Harris,[9] Vance G. Fowler Jr,[4] Helen Boucher,[10] Barry N. Kreiswirth,[11] Robert A. Bonomo,[12] David van Duin,[13] David L. Paterson,[14] and Henry Chambers[5]

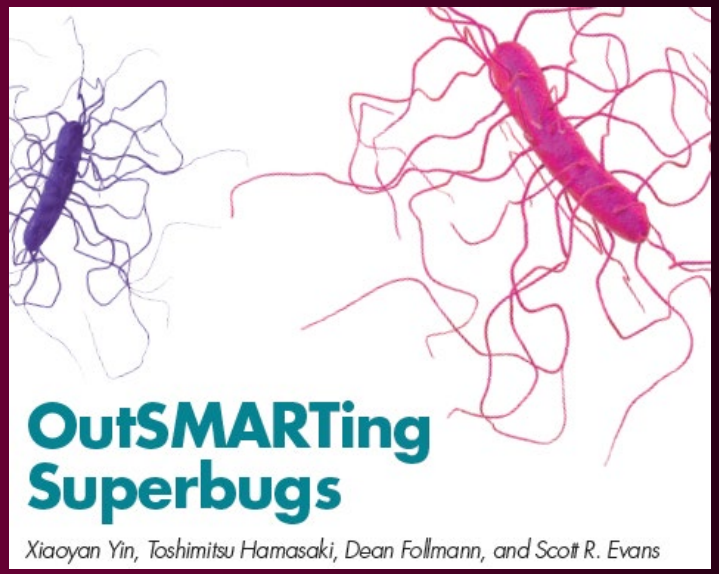
Interacting with the FDA on Complex Innovative Trial Designs for Drugs and Biological Products

Guidance for Industry

C. Sequential Multiple Assignment Randomized Trials (SMARTs)

Sequential Multiple Assignment Randomized Trials (SMARTs) are designed to inform the development of adaptive interventions. An adaptive intervention is a sequence of decision rules that specifies when and how the type and/or intensity of a treatment should be modified depending on the patient's characteristics and/or ongoing performance (e.g., response, adherence) to optimize clinically important outcomes. A SMART is comprised of multiple intervention stages, and each stage corresponds to one of the critical decisions involved in the adaptive intervention. In a SMART, patients move along multiple stages and are randomly assigned to one of several treatment options at each stage.

Elements that should be discussed and communicated with FDA in SMART proposals or similar designs include the statistical questions/hypotheses, number of stages, interventions embedded in the design, intermediate response categories, a clear illustration of the flow diagram, and methods to adjust for multiplicity if applicable.



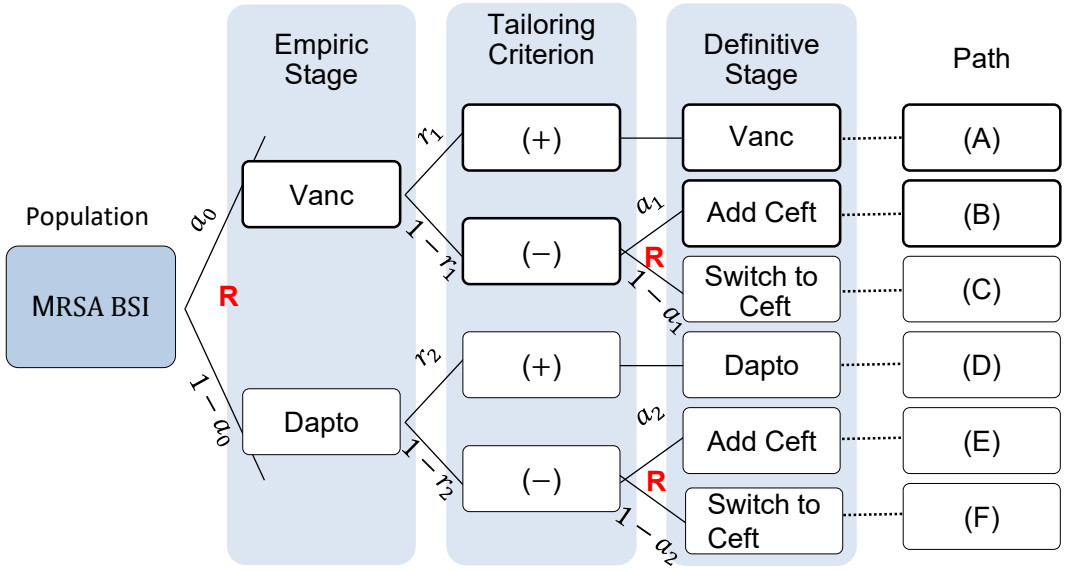
SMART-COMPASS

- **Strategy:** a decision-rule guiding patient treatment e.g., combined empiric and definitive therapy decisions based upon available data at the time such decisions are made

- **COMPASS**: Compare decision-making strategies consistent with clinical practice rather than specific treatments.

- When there are multiple empiric-therapy and multiple definitive-therapy options, a **SMART-COMPASS** can be considered.

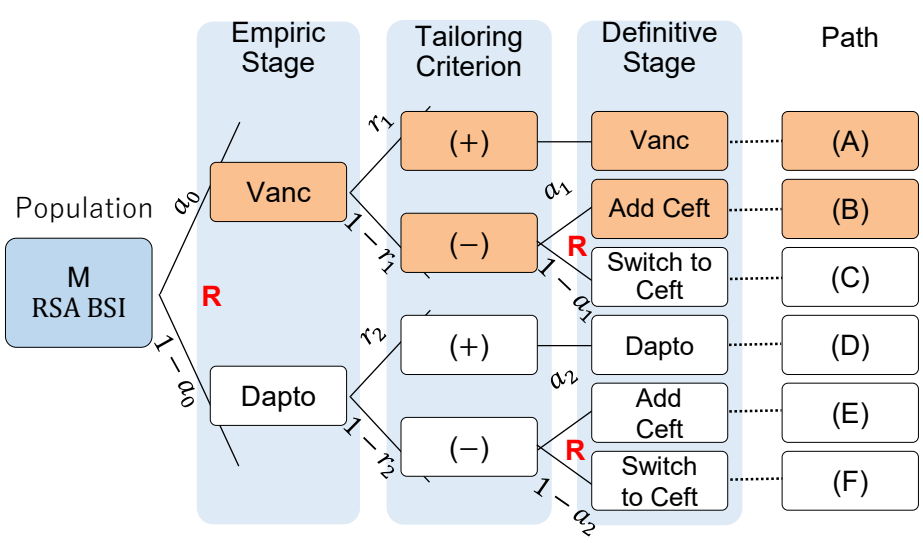
Motivating Example



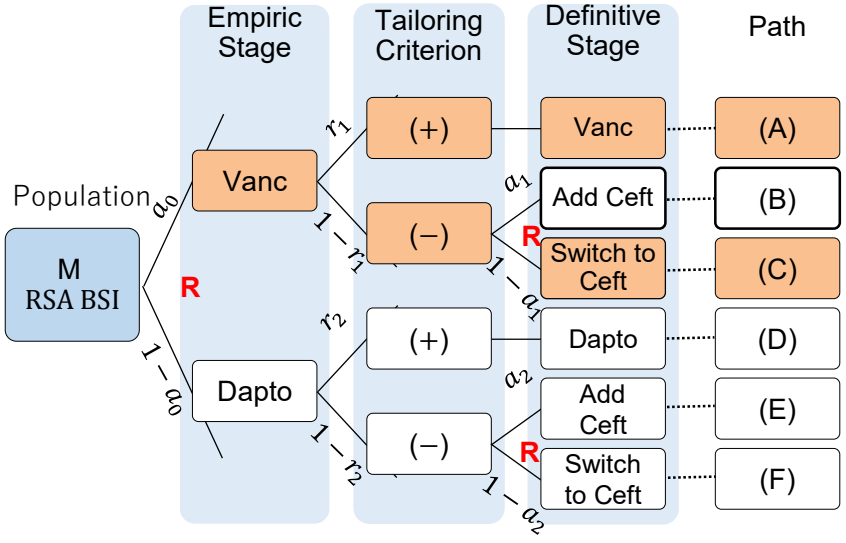
MRSA, methicillin-resistant staphylococcus aureus; BSI, bloodstream infection; Dapto, daptomycin; R, randomization; Vanco, vancomycin; Ceft, ceftaroline; (+), Culture positive; (-) Culture negative.

Motivating Example

Strategy 1: S1

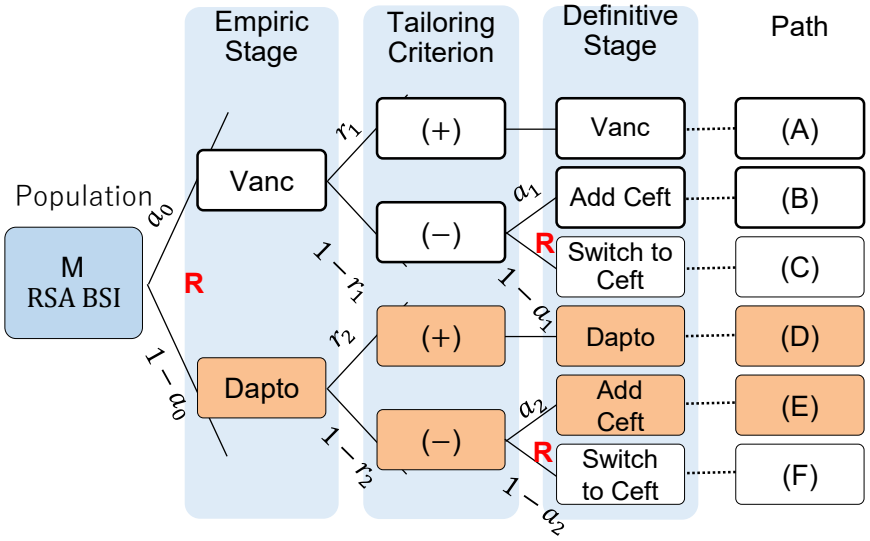


Strategy 2: S2

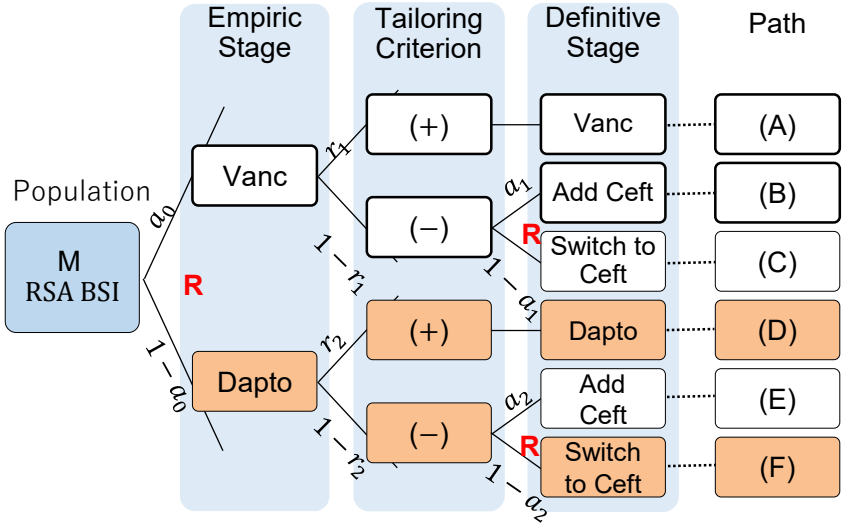


Motivating Example

Strategy 3: S3



Strategy 4: S4



Introduction: Research Questions of SMARTs

Research Questions		Example Hypotheses	Contrast of illustrative pathway
Question 1: Evaluate empiric therapies Comparisons of empiric therapies coupled with different subsequent therapies		Vancomycin is better than daptomycin.	Path (A)+(B)+(C) vs Path (D)+(E)+(F)
Question 2: Evaluate definitive therapies Comparisons of definitive therapy conditioning upon empiric therapy		Adding is better than switching to Ceftaroline conditioning upon empiric Vanc. and a positive culture.	Path (B) vs Path (C) conditioning upon E1 with a negative culture
Question 3: Strategy Comparison	Pairwise comparisons	S1 is better than S3	Paths (A)+(B) vs Paths (D)+(E)
	Identification of the best strategy	S1>S2>S3> S4	Identify best among Paths (A)+(B) vs Paths (A)+(C) vs Paths (D)+(E) vs Paths (D)+(F)

SMART-COMPASS

Advantages:

- Pragmatic approach focuses on identifying **best strategies** that mirror clinical decision-making for antibiotic treatment to improve patient outcomes.
- Personalized medicine
- Understanding the impact of therapeutic adjustments on patient outcomes is crucial in evaluating the effectiveness and cost-effectiveness of empiric use

Challenges:

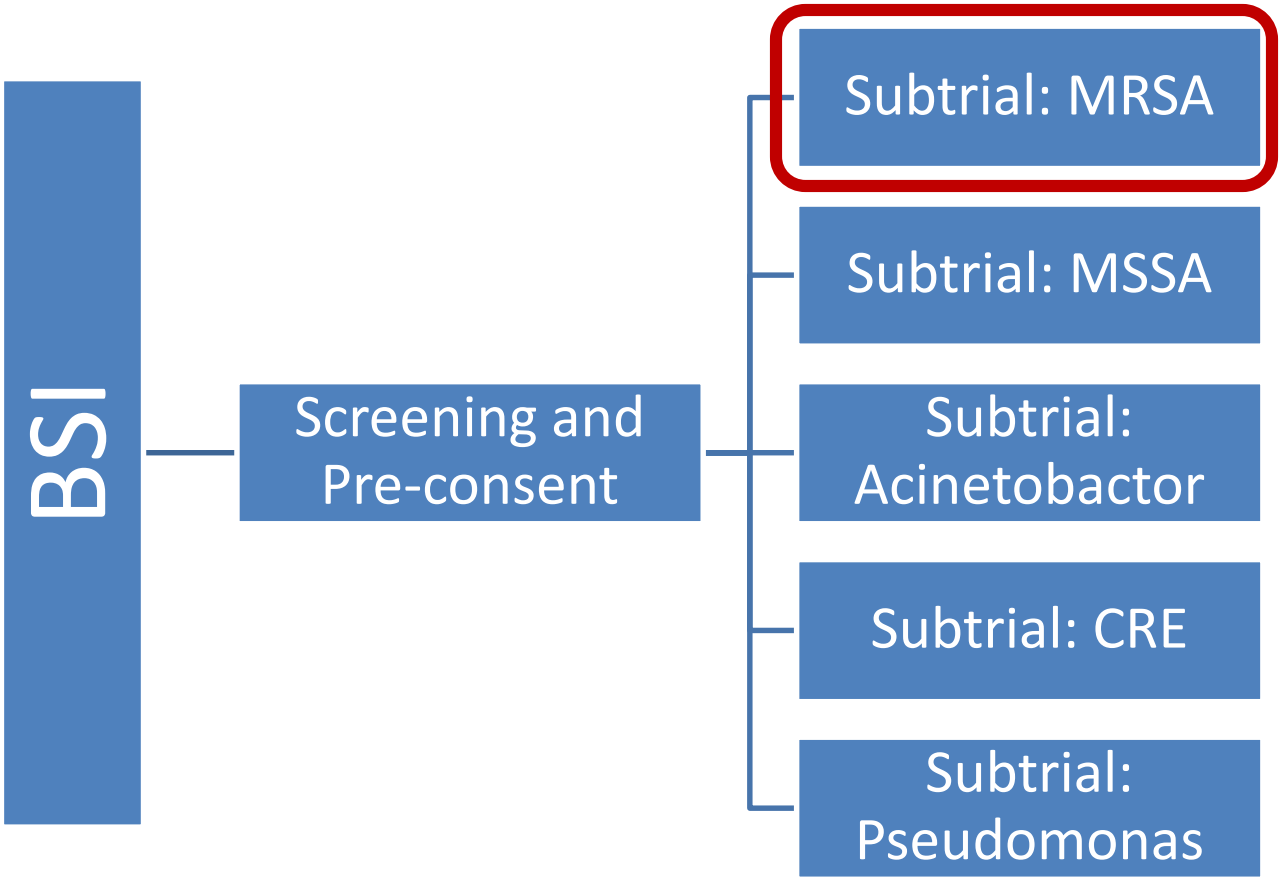
- Cultural shift of focus from drugs to strategies
- Complicated design and logistics
    - There are 2 stages of randomization and estimating the proportions of patients to be re-randomized at the definitive treatment stage is necessary for sample size calculations, which require weighting to obtain accurate estimates of strategy proportions and associated standard errors

GNB Trial evaluating: (1) mono vs. combo therapy, and (2) short vs. long term strategy

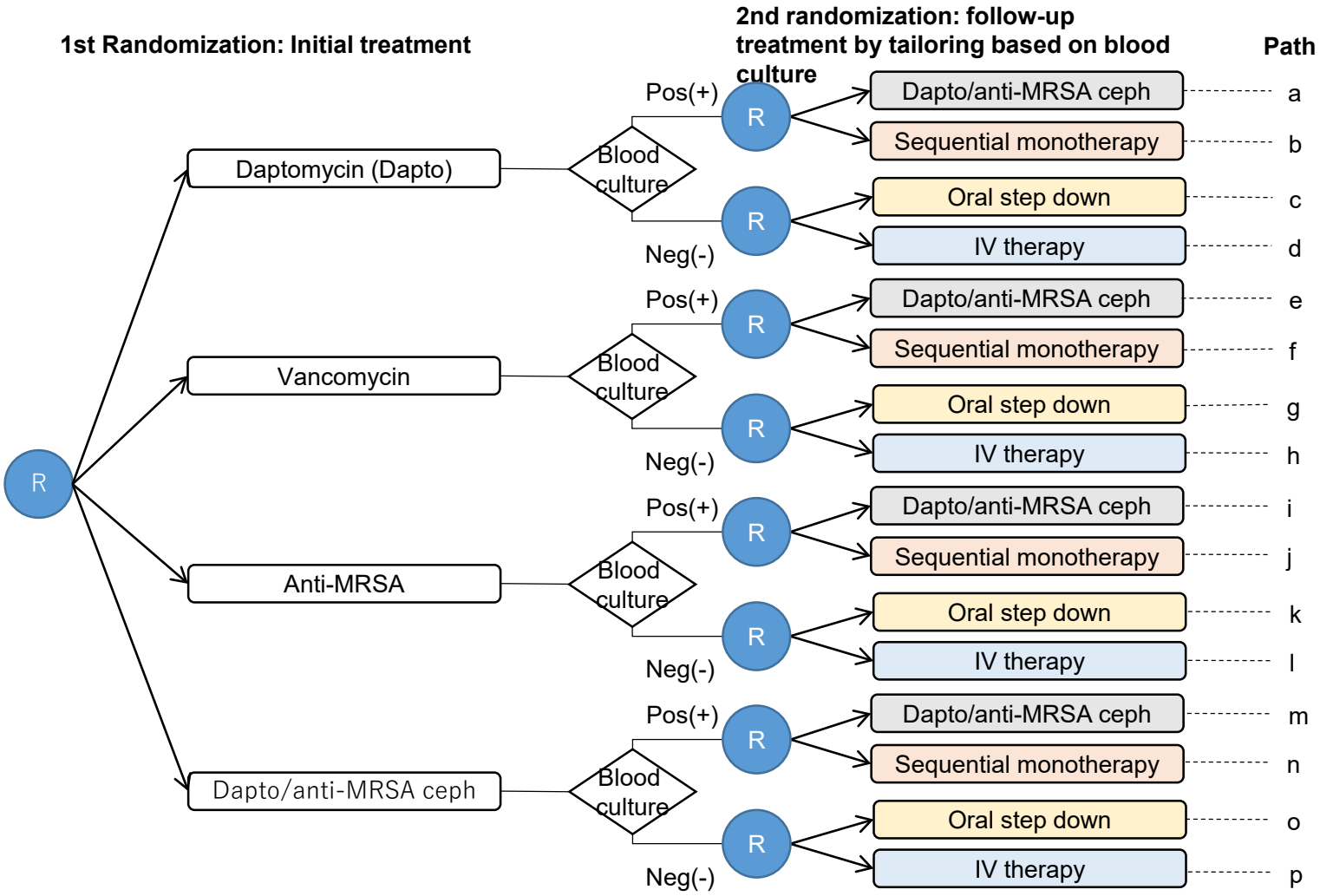
- Pick pathogen and infection site
- Randomize to mono vs. combo therapy
- Treat until the end of short term therapy
- For those that are eligible to stop therapy or transition, randomized to continue long term or transition to short
  - Decide what to do with those not eligible to step down
- Trial evaluates 4 strategies of treatment and address mono/combo question and long/step-down question

The ideas can be combined.

Umbrella Design



Design Outline



Questions: (1) initial single or dual therapy; (2) step-down options

Combat Trauma Infection Prevention and Management Wartime Preparedness

- Collaboration with Uniformed Services University and DoD
- Two multi-arm trials to investigate battlefield trauma-related infection:
  1. Prevention and early wound care products/strategies
  2. Management products/strategies
- Make all comparisons (find optimal treatment)
- People with infection from trial 1, may be randomized to trial 2

Design / Design Characteristic	Consider	Selectively Consider	Avoid	Comments
Group-sequential designs with efficacy and futility analyses	X			Solid statistical foundation.
Multi-arm / MAMS	X			Consider all possible comparisons to optimize utility / public health benefit
Umbrella	X			Screening tool. Independent sub-trials.
Basket		X		Limited robustness (assumption and model heavy). Risky and may not have a foundation for well-defined treatment effects. Inconsistent with studies submitted for regulatory approval. Generally used when options are limited.
SMARTs	X			Solid statistical foundation though complex. Pragmatic questions i.e., evaluating strategies. Difficult to understand for those focused on specific drugs.
Two-stage designs	X			

Adaptations	Consider	Selectively Consider	Avoid	Comments
Study Designs				
(1) Study objectives/hypotheses				
● Study primary objective (e.g., Superiority to noninferiority)			X	● This results in a change to the study rationale.
● Secondary or explanatory objective(s)		X		● May occasionally be appropriate.
(2) Study Design Elements				
● Endpoints (primary/secondary)				● Changing the primary endpoint during a trial is very rare.
● Study population (Enrichment)		X		● The decision-making framework must be predetermined; Data-driven, unplanned changes introduce bias into the treatment comparison, inflating the type I error.
● Select/drop/add arms	X			
● Randomization procedure (e.g., ratio, covariates)			X	● There is a risk of introducing bias into the treatment comparison, inflating the type I error.
(3) Design Parameters (for sample size)				
● Effect size, NI margin			X	● This results in a change to the study rationale.
● Sample size (Power)		X		● Fine if based on nuisance parameters e.g., variation or without unblinding. Trickier when unblinding or in the context of NI. Avoid reducing the sample size based on an observed difference in effect by group.
● Follow-up duration (event-driven)		X		
Study Conduct				
● Inclusion/exclusion criteria		X		● May occasionally be appropriate for ethical or operational reasons e.g., in long-term trials, where growing medical knowledge either from outside the trial or from interim analyses/trial monitoring may suggest the change
● Rescue/concomitant medications		X		
● Treatment regimen		X		
● Observation schedule		X		
Study Analysis				
(1) Interim analysis and early stopping				
● Stopping for efficacy, futility or both	X			● The decision-making framework must be predetermined.
● # and/or timing of interim analyses		X		● Data-driven, unplanned changes introduce bias into the treatment comparison, inflating the type I error.
(2) Final analysis				
● Analysis population		X		● The assessments must be conducted under blinded review, which occurs between trial completion (the last observation on the last patient) and the breaking of the blind, for the purpose of finalizing the planned analysis
● Analysis methods		X		
● Missing data handling		X		
● Covariates		X		

Clinical Infectious Diseases

INVITED ARTICLE

IDSA Infectious Diseases Society of America

hivma hiv medicine association

OXFORD

HEALTHCARE EPIDEMIOLOGY: Robert Weinstein, Section Editor

Adaptive Designs for Clinical Trials: Application to Healthcare Epidemiology Research

W. Charles Huskins,[1] Vance G. Fowler Jr,[2] and Scott Evans[3]

Acknowledgements

- Used slides from
  - Bob O'Neill
  - Stuart Pocock
  - Dan Rubin
  - Marc Buyse
  - Yohei Doi

Arigato gozaimasu