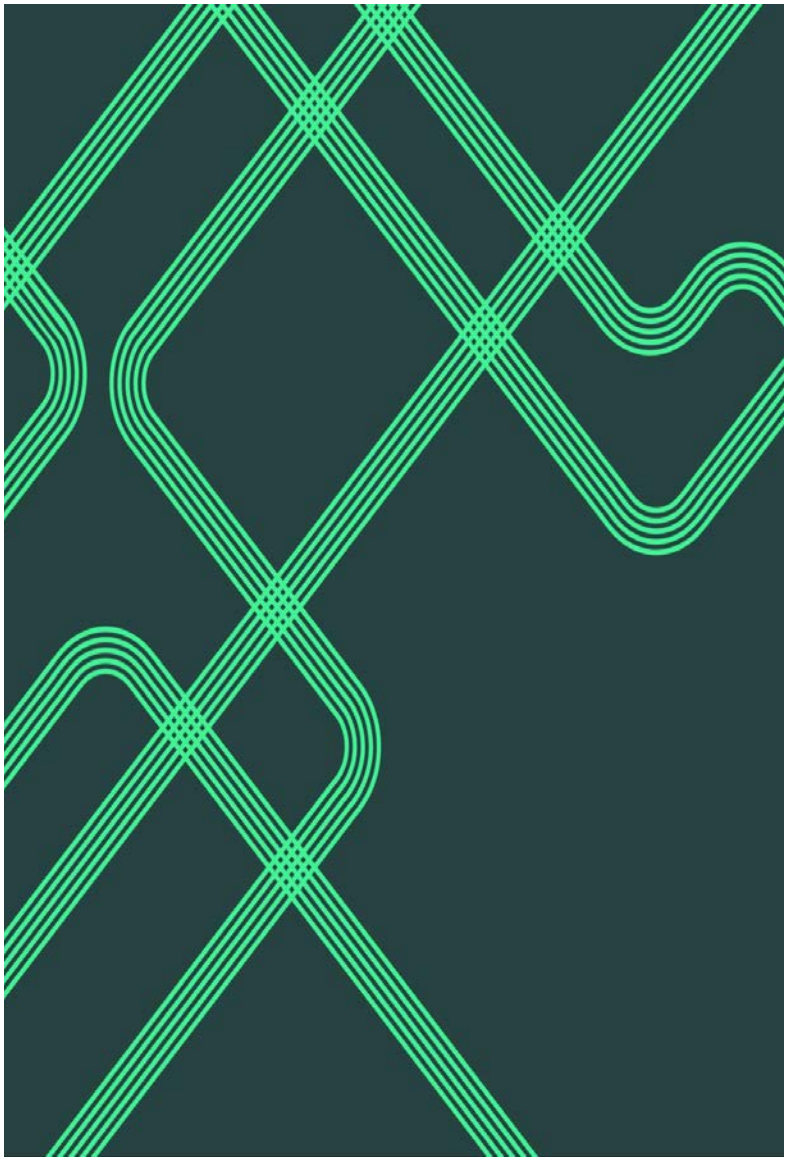
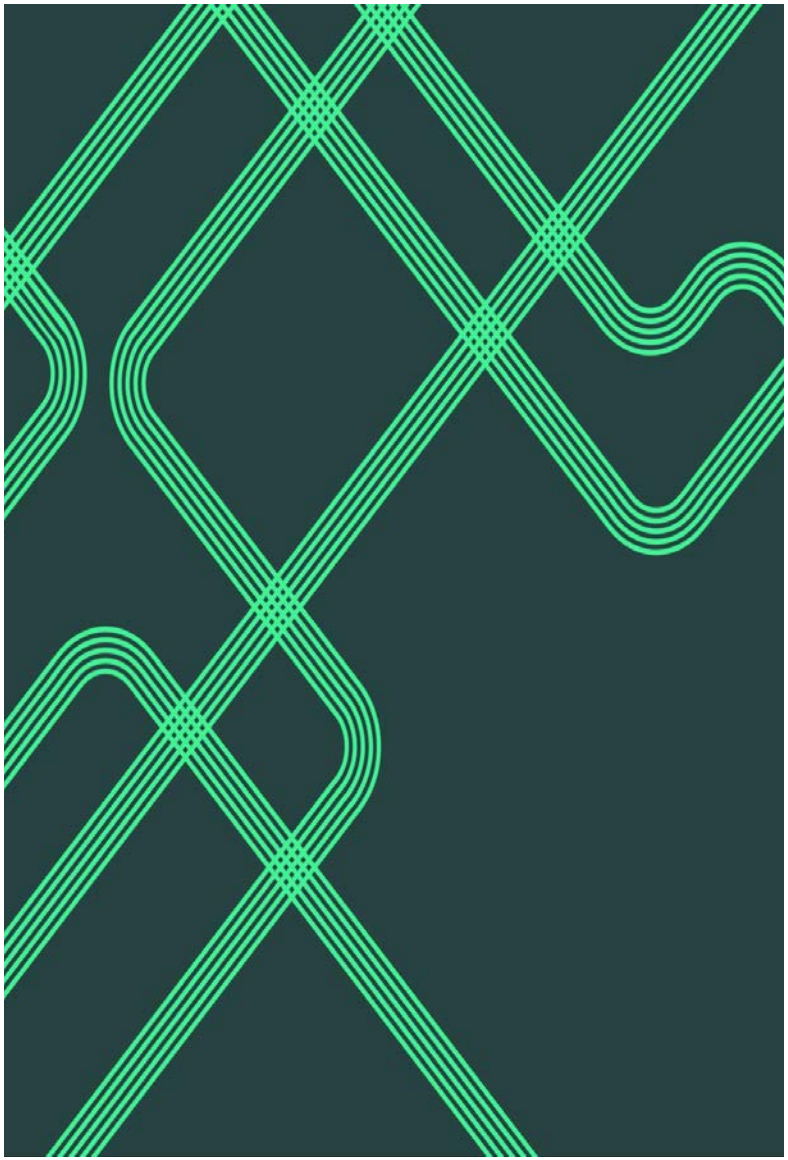


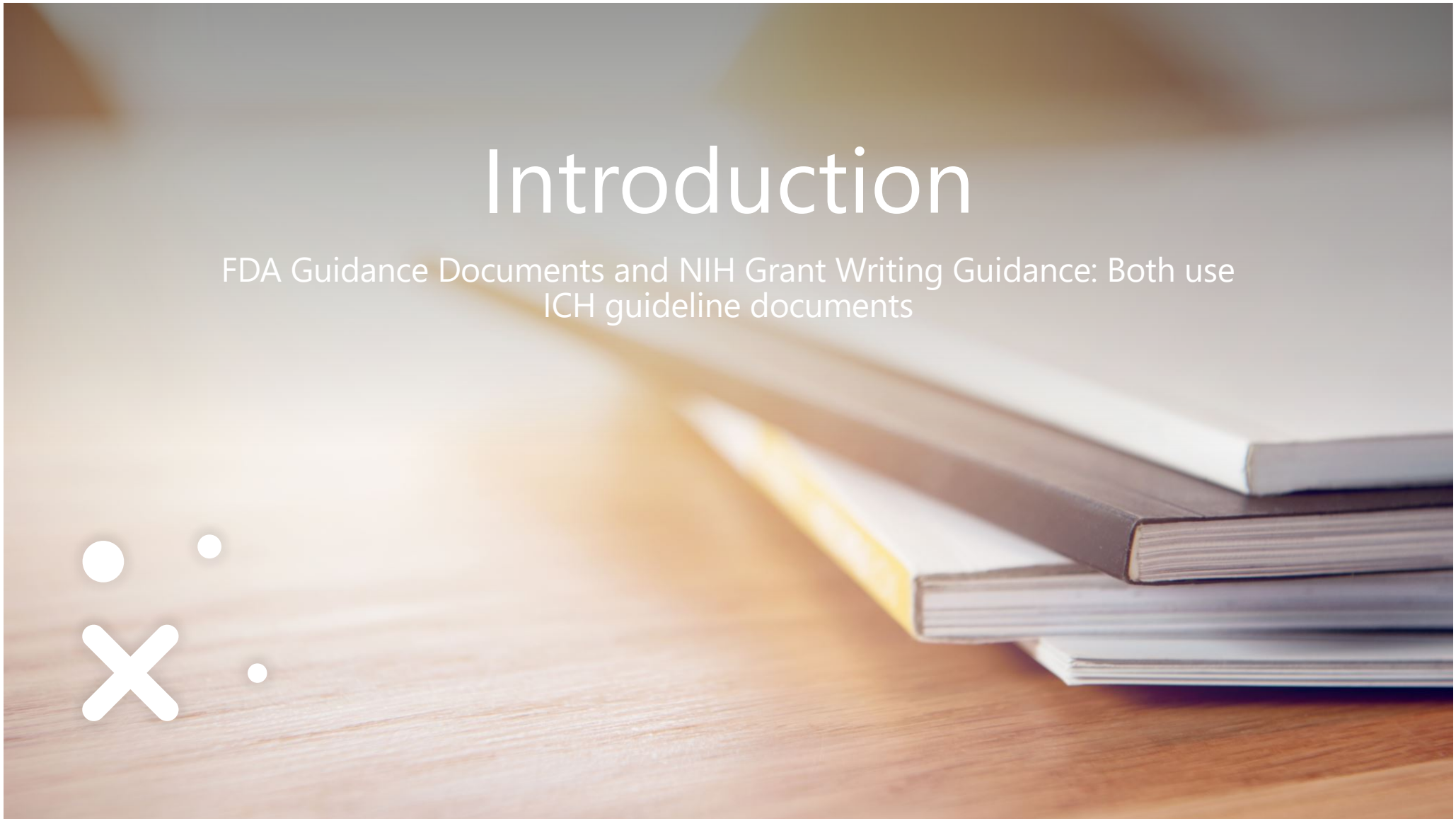
Testing Specific Hypotheses in Regression, ANOVA, and SEM: Multiplicity Issues

Thomas N Templin, Ph D
Office of Health Research
Wayne State University
January 19, 2021

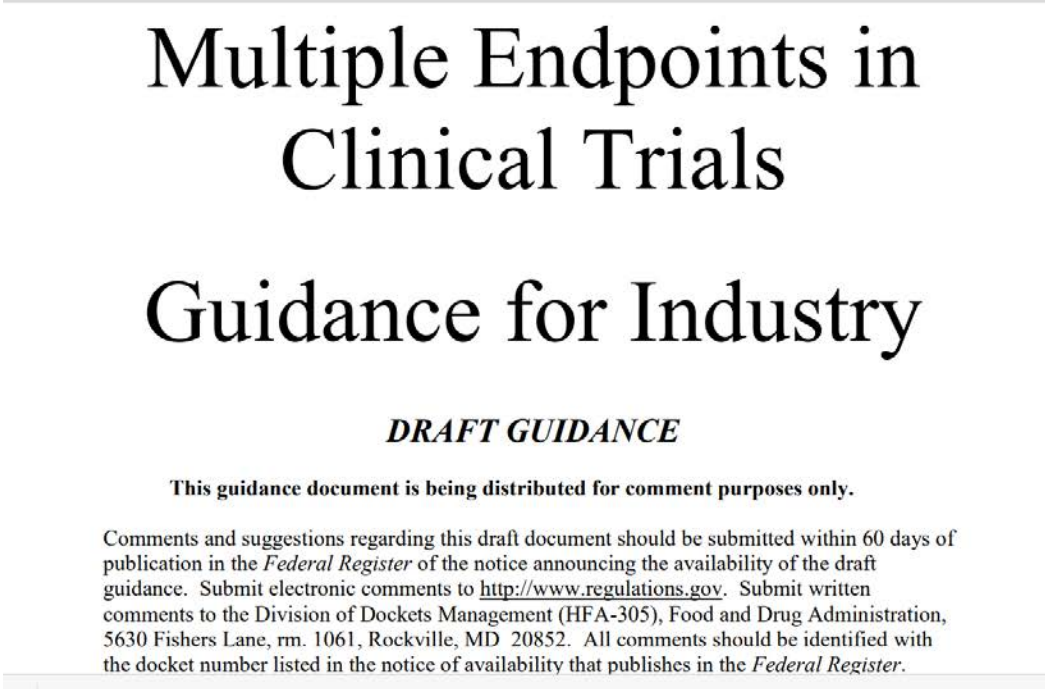


Multiple Endpoints in Confirmatory Clinical Trials: Multiplicity Issues

Thomas N Templin, Ph D
Office of Health Research
Wayne State University
January 19, 2021


Introduction
FDA Guidance Documents and NIH Grant Writing Guidance: Both use ICH guideline documents

Primary resource for today's seminar:



Multiple Endpoints in Clinical Trials

Guidance for Industry

DRAFT GUIDANCE

This guidance document is being distributed for comment purposes only.

Comments and suggestions regarding this draft document should be submitted within 60 days of publication in the *Federal Register* of the notice announcing the availability of the draft guidance. Submit electronic comments to http://www.regulations.gov. Submit written comments to the Division of Dockets Management (HFA-305), Food and Drug Administration, 5630 Fishers Lane, rm. 1061, Rockville, MD 20852. All comments should be identified with the docket number listed in the notice of availability that publishes in the *Federal Register*.

https://www.fda.gov/media/102657/download

DRAFT GUIDANCE

This guidance document is being distributed for comment purposes only.

Comments and suggestions regarding this draft document should be submitted within 60 days of publication in the *Federal Register* of the notice announcing the availability of the draft guidance. Submit electronic comments to http://www.regulations.gov. Submit written comments to the Division of Dockets Management (HFA-305), Food and Drug Administration, 5630 Fishers Lane, rm. 1061, Rockville, MD 20852. All comments should be identified with the docket number listed in the notice of availability that publishes in the *Federal Register*.

For questions regarding this draft document contact (CDER) Scott Goldie at 301-796-2055 or (CBER) Office of Communication, Outreach, and Development, 800-835-4709 or 240-402-8010.

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)

[January 2017]
Clinical/Medical

FDA's guidance for industry *E9 Statistical Principles for Clinical Trials* (International Council on Harmonisation E9 guidance, or "ICH E9") is a broad ranging guidance that includes discussion of multiple endpoints.

http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm. (p 2)

NIH grant writing guidance for randomized behavioral clinical trials

Templates and reference to ICH guidance

https://osp.od.nih.gov/clinical-research/clinical-trials/

Questions about the tool can be sent to SciencePolicy@od.nih.gov.

Take me to the e-Protocol Writing Tool

Final "Phase 2 and 3 Clinical Trial" Template Documents

- Word Version of Final Template
- NIH Guide Notice
- NIH Director's Statement
- *Under the Poliscope* blog
- FDA Voice blog

Final "Behavioral and Social Sciences Research Involving Humans Template" Documents

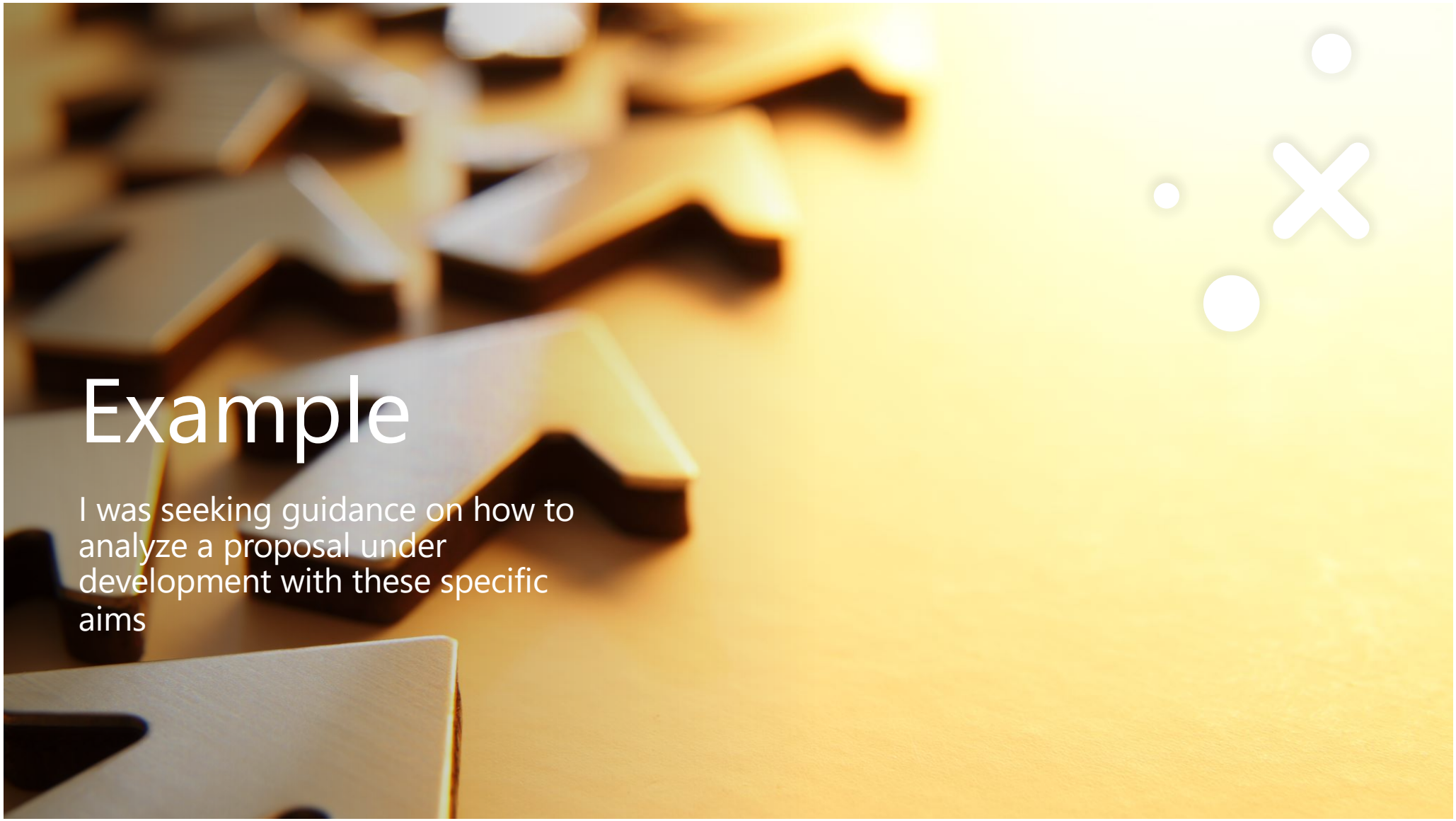
- Word Version of Final Template
- NIH Guide Notice
- *Under the Poliscope* Blog

+ Clinical Trial Definition

Both templates found in the electronic protocol tool meet the standards outlined in the *International Council on Harmonisation (ICH) Guidance for Industry, E6 Good Clinical Practice: Consolidated Guidance (ICH-E6)*. These are international standards of good clinical practice that apply to all clinical trials, and their goals are to ensure research integrity and protect human subjects. In addition, use of the electronic protocol tool allows researchers to interface directly with clincialtrials.gov.

More NIH tutorial resources on behavioral trials

https://obssr.od.nih.gov/training/training-supported-by-the-obssr/


Example
I was seeking guidance on how to analyze a proposal under development with these specific aims

This proposed study will use a pragmatic effectiveness design. In this design, the primary goal is to determine whether the intervention works in a real-world home hospice setting.

Our **overall objective** for this study is to increase family caregiver knowledge and performance of end-of-life symptom management of their family member with advanced stage dementia.

The following **specific aims** will guide this study:

Aim1: To compare effectiveness of family caregiver performance of symptom assessment and management to promote comfort (respiratory distress, pain, and agitation) between intervention and control groups.

Hypothesis 1: Patients in the intervention group will display comfort to a greater degree compared to the control group.

Research question 1: Does the impact of the intervention on symptom comfort vary by *APOE* and other dementia-specific genotypes?

Aim2: To compare family caregiver strain, depression, and anxiety between intervention and control groups.

Hypothesis 2: Family caregivers in the intervention group experience less strain, depression, and anxiety than those in the control group.

Aim3: To compare family use of resources and patient site of death between intervention and control groups.

Hypothesis 3a: The patients in the intervention group have fewer incidences of calls to hospice and/or calls to emergency services and/or acute hospital admissions because of symptom distress than patients in the control group.

Hypothesis 3b: More patients in the intervention group die at their preferred site of death (home) than patients in the control group.

(PIs: Campbell, Vallerand, Schutte, application in development)

Aim1: To compare effectiveness of family caregiver performance of symptom assessment and management to promote comfort (respiratory distress, pain, and agitation) between intervention and control groups.

Hypothesis 1: Patients in the intervention group will display comfort to a greater degree compared to the control group.

Aim1: To compare effectiveness of family caregiver performance of symptom assessment and management to promote comfort (respiratory distress, pain, and agitation) between intervention and control groups.

Hypothesis 1: Patients in the intervention group will display comfort to a greater degree compared to the control group.

Questions raised by H1:

- Three separate endpoints or one composite endpoint?
- Are we going to claim success if any one endpoint is significant or only if all 3 are significant?
- How should alpha be adjusted to maintain a familywise error rate at .05?

Aim2: To compare family caregiver strain, depression, and anxiety between intervention and control groups.

Hypothesis 2: Family caregivers in the intervention group experience less strain, depression, and anxiety than those in the control group.

Aim2: To compare family caregiver strain, depression, and anxiety between intervention and control groups.

Hypothesis 2: Family caregivers in the intervention group experience less strain, depression, and anxiety than those in the control group.

Questions raised by H2: Similar questions here and some additional ones.

- Are we going to claim success if H2 is supported and H1 is not, or should we consider success if either hypothesis is supported?

- Should alpha be adjusted to maintain a familywise error rate across hypotheses or within only?

Multiple Endpoints

+H1: more comfort (respiratory distress, pain, and agitation)

+H2: less strain, depression, and anxiety

+H3a: fewer incidences of **calls to hospice** and/or **calls to emergency services** and/or **acute hospital admissions** because of symptom distress

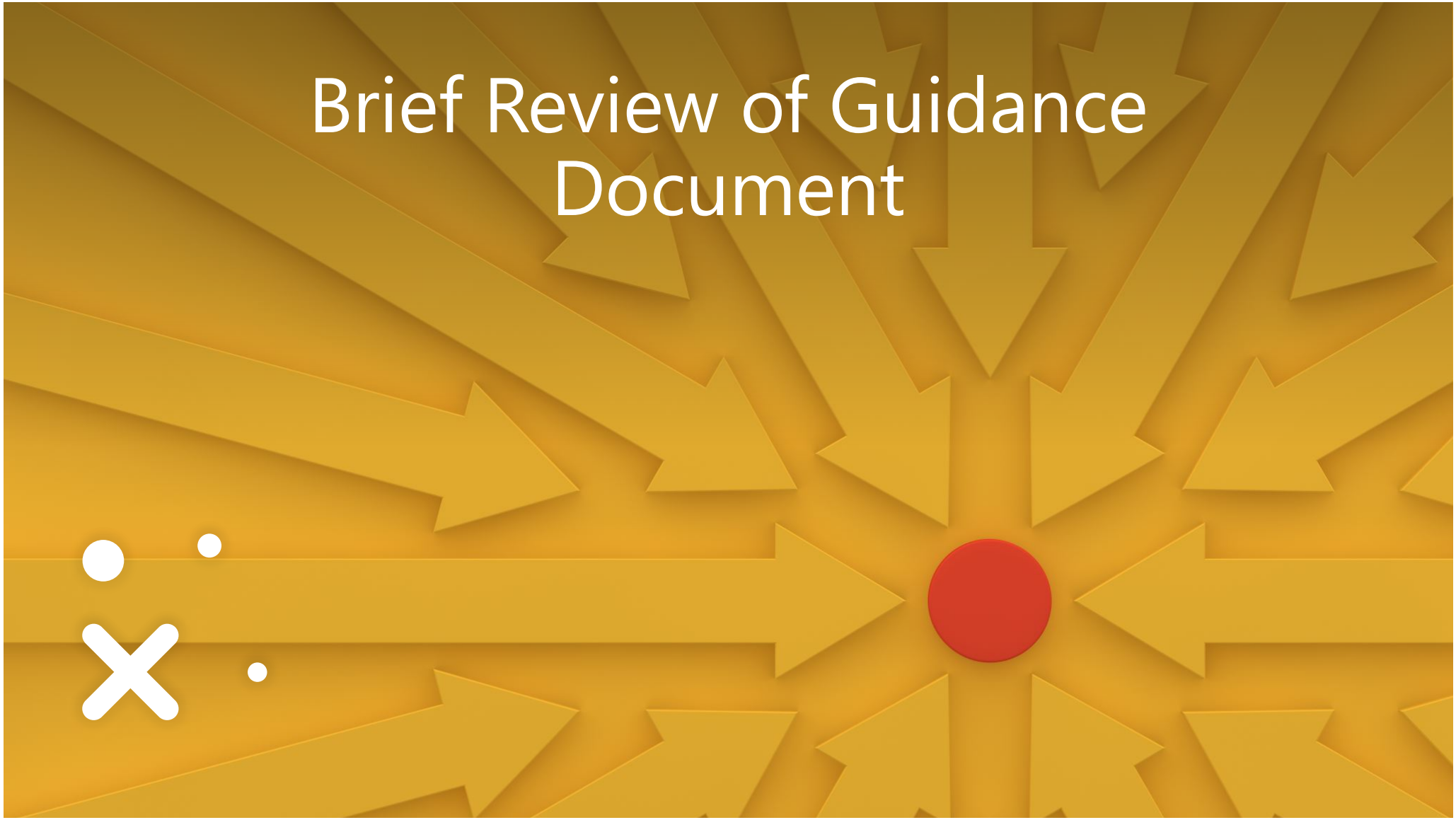
+H3b: die at their preferred site of death (home)

Endpoints not apparent from hypotheses

- Patient's and family caregiver's endpoints will be assessed weekly after the intervention until the death or withdrawal of the patient.

- Each follow-up data collection is a potential endpoint and has to be counted in computing the family-wise error rate.

Brief Review of Guidance Document



The document is 37 pages so I will only cover parts relevant to our example. That does not leave much out except for the Section IV which describes many of the statistical methods available for adjusting for multiplicity.

The first 20 pages deals with research design issues that come up in designing a clinical trial involving the use of multiple endpoints.

.

I. Introduction

Most clinical trials performed in drug development contain multiple endpoints to assess the effects of the drug and to document the ability of the drug to favorably affect one or more disease characteristics. (p. 2)

As the number of endpoints analyzed in a single trial increases, the likelihood of making false conclusions about a drug's effects with respect to one or more of those endpoints becomes a concern if there is not appropriate adjustment for multiplicity. (p. 2)

Efficacy endpoints are measures intended to reflect the effects of a drug.

- They include assessments of clinical events (e.g., mortality, stroke, pulmonary exacerbation, venous thromboembolism),
- patient symptoms (e.g., pain, dyspnea, depression),
- measures of function (e.g., ability to walk or exercise), or surrogates of these events or symptoms. (p. 3)

When there are many endpoints prespecified in a clinical trial, they are usually classified into three families: ***primary, secondary, and exploratory***.

- The set of primary endpoints consists of the outcome or outcomes … that establish the effectiveness, and/or safety features, of the drug in order to support regulatory action.
- When there is more than one primary endpoint and success on any one alone could be considered sufficient to demonstrate the drug's effectiveness, the rate of falsely concluding the drug is effective is increased due to multiple comparisons (p. 3).

Secondary endpoints may be selected to demonstrate additional effects after success on the primary endpoint.

- For instance, a drug may demonstrate effectiveness on the primary endpoint of survival, after which the data regarding an effect on a secondary endpoint, such as functional status, would be tested.
- Secondary endpoints may also provide evidence that a particular mechanism underlies a demonstrated clinical effect (e.g., a drug for osteoporosis with fractures as the primary endpoint, and improved bone density as a secondary endpoint) (p. 3-4).

All other endpoints are referred to as exploratory. (p. 4)
Endpoints intended to serve the purpose of hypothesis generation should not be included in the secondary endpoint family. These should be considered exploratory endpoints. (p. 12)

Primary Endpoint Family

Multiple primary endpoints occur in three ways

Each has different effects on Type 1 and Type 2 error

1. Multiple primary endpoints are used, giving the study multiple chances to succeed. This results in increase of familywise error rate by formula

$$\text{FWER} = 1 - (1 - \alpha)^n$$

Multiple primary endpoints are perfectly ok as long as appropriate adjustments for multiplicity are made.

2. Determination of effectiveness depends on the success of all of two or more related endpoints.

These are called ***coprimary endpoints***.

No increase in Type 1 error but power is reduced

3. Composite Endpoints

No increase in Type 1 error and or decrease in power but all components need to be of equal importance.

Descriptive analysis of individual components should also be done.

Two types of composite endpoints

- A summation of low frequency events
- When the components correspond to distinct events, composite endpoints are often assessed as the time to first occurrence of any one of the components, but in diseases where a patient might have more than one event, it also may be possible to analyze total endpoint events (p. 16)

- A sum or average of individual domain scores
- Examples of this type are the Positive and Negative Syndrome Scale (PANSS) in schizophrenia research; … the Brief Psychiatric Rating Scale; and many patient-reported outcomes (PROs) (p. 17).

It was interesting to read about all the different ways in which composite Endpoints could be defined:

The primary endpoint in clinical trials of allogeneic pancreatic islet cells for Type 1 diabetes mellitus be a composite in which patients are considered responders only if they meet two dichotomous response criteria: normal range of HbA1c and elimination of hypoglycemia (p. 17)

Secondary Endpoint Family:

- Secondary endpoints are those that may provide supportive information about a drug's effect on the primary endpoint or demonstrate additional effects on the disease or condition.

- Positive results on the secondary endpoints can be interpreted *only* if there is first a demonstration of a treatment effect on the primary endpoint family.

- The Type I error rate should be controlled for the entire trial, as strong control. *This includes controlling the Type I error rate within and between the primary and secondary endpoint families*. (p. 11-12)

Secondary Endpoint Family (cont.):

- If success on the secondary endpoints is important, the secondary endpoints should be considered when determining study design (e.g., sample size).

- It is recommended that the list of secondary endpoints be short, because the chance of demonstrating an effect on any secondary endpoint after appropriate correction for multiplicity becomes increasingly small as the number of endpoints increases. (p. 11-12)

Secondary Endpoint Family (cont.):

- Endpoints intended to serve the purpose of hypothesis generation should not be included in the secondary endpoint family. These should be considered exploratory endpoints. (p. 11-12)

IV. STATISTICAL METHODS

The Bonferroni Method
The Holm Procedure
The Hochberg Procedure
Prospective Alpha Allocation Scheme
The Fixed-Sequence Method
The Fallback Method
Gatekeeping Testing Strategies
The Truncated Holm and Hochberg Procedures for Parallel Gatekeeping.
Multi-Branched Gatekeeping Procedures
Resampling-Based, Multiple-Testing Procedures

Control of the family-wise error rate (FWER)

observed p-values	Bonferroni	Holm Step-down	Hochberg Step-up
p_{r_1}	α/G	α/G	α/G
p_{r_2}	α/G	$\alpha/(G\text{-}1)$	$\alpha/(G\text{-}1)$
:	:	:	:
p_{r_g}	α/G	$\alpha/(G\text{-}g\text{+}1)$	$\alpha/(G\text{-}g\text{+}1)$
:	:	:	:
$p_{r_{G-1}}$	α/G	$\alpha/2$	$\alpha/2$
p_{r_G}	α/G	α	α

Holm: Compare smallest P to smallest critical value. If NS, stop. If significant step down.

Hochberg: Compare largest P to largest critical value. If significant, stop. If NS step up.

LeeWhitmoreTutorial.pdf (upenn.edu)

IV. STATISTICAL METHODS (cont.)

There are also other statistical analysis methods, often called global procedures, …. These methods allow a conclusion of treatment effectiveness in the global sense,

but do not support reaching conclusions on the individual endpoints within the family.

These methods are generally not encouraged when study designs and methods that test the endpoints individually are feasible. (p. 25)

I am thinking, multivariate methods, e.g., MANOVA, Profile Analysis, etc. are not encouraged

Discussion and Conclusions

In thinking about how to proceed considering these guidelines we also should keep in mind that they were designed for regulatory approval of a new investigational drug or device application and other guidelines could apply to our application. We rely a lot of the experience and intuition, consultants, and senior investigators.

Multiple Endpoints (type)

- H1: more comfort (respiratory distress, pain, and agitation) (primary)
- H2: less strain, depression, and anxiety (primary)
- H3a: fewer incidences of **calls to hospice** and/or **calls to emergency services** and/or **acute hospital admissions** because of symptom distress (secondary)
- H3b: die at their preferred site of death (home) (secondary)

Family-wise Type 1 error with any endpoint signaling success

- H1: more comfort (respiratory distress, pain, and agitation) (3)
- H2: less strain, depression, and anxiety (3)
- Follow-up for two different time intervals (2)
- Number of endpoints 3x2 + 3x2 = 12
- Family-wise error rate = $1-.95^{12} = 1 - \ .54 = .46$
- Bonferroni adjusted alpha would be used for power analysis.
- Adjusted alpha = .05/12 = .0042
- Affect on power, drops from .91 to .67 with effect size = .40
- Affect on power, drops from .96 to .81 with effect size = .45

Family-wise Type 1 error with composite endpoints for H1-H3a. All 4 hypotheses as primary outcomes

- H1: more comfort (respiratory distress, pain, and agitation) (primary)
- H2: less strain, depression, and anxiety (primary)
- H3a: fewer incidences of **calls to hospice** and/or **calls to emergency services** and/or **acute hospital admissions** because of symptom distress (primary)
- H3b: die at their preferred site of death (home) (primary)

- Test each at alpha/4 = .0125, or with two follow-up endpoints use alpha/8.


Thank You