ETZ: A Modeling Principle for Confirmability of Drug-Development Studies

Yujia Sun
Department of Mathematics, The University of Manchester;
Dept. Data Science & Big Data Technology, Zhejiang A&F University
Yang Han*
Department of Mathematics, The University of Manchester
Xingya Wang
Department of Mathematics, The University of Manchester
Szu-Yu Tang
Pfizer Worldwide Research & Development
Yushi Liu
Eli Lilly and Company
and
Jason C. Hsu*
Department of Statistics, The Ohio State University

Abstract

Transitioning from Phase 2 to Phase 3 in drug development, at a rate of ≈40%, is the most stringent among phase transitions (Hay et al. (2014)). Yet, success rate at Phase 3 leading to approval is only ≈50% (Arrowsmith (2011b)). To improve *Confirmability*, we propose a methodological shift: replacing *multiple hypothesis testing* with inference based on confidence sets, and substituting conventional *power and sample size* calculations with a Confidently Bounded Quantile (CBQ) framework.

Our confidence set inferences to answer the questions of whether to transition to a Confirmatory study as well as what to designate as the endpoint in that study. Construction of our *directed* confidence sets follows the Partitioning Principle, taking the best of each of Pivoting and Neyman Confidence Set Construction.

Rooted in Tukey's Confidently Bounded Allowance (CBA) (Tukey (1994a)), our proposed CBQ makes the transitioning decision following the Correct and Useful Inference principle in Hsu (1996). CBQ removes from "power" the probability of *rejecting for wrong reasons*, eliminating the need for informal *discounting* in power calculation that has existed in the biopharmaceutical industry.

ETZ, the modeling principle proposed in Wang et al. (2025), quantifies the impact of three variability components on *confirmability*. In repeated-measures RCTs, it separates *within-subject* and *between-subject* variability, further dividing the latter into *baseline* and *trajectory* components. This enables informed investment decisions for the sponsors on targeting variability reduction to improve *confirmability*. A Shiny-based *Confirmability App* supports all computations.

Keywords: Variability Decomposition, Randomized Controlled Trials, Decision-making Process, Multiple Comparisons.

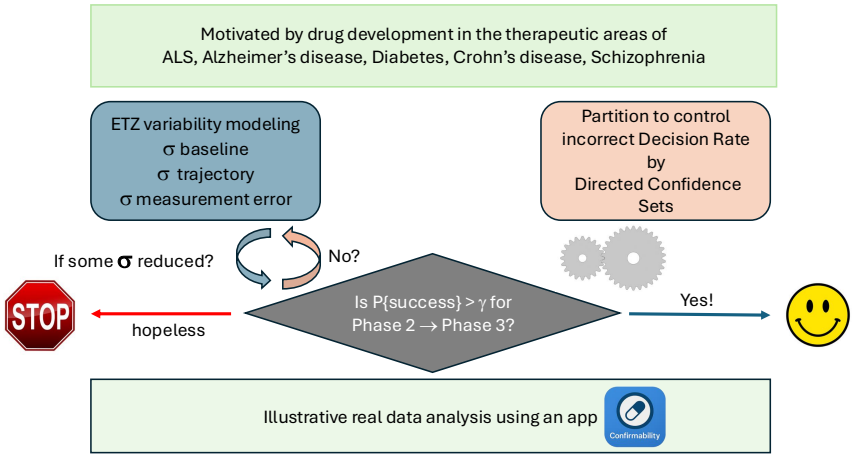

Figure 1: The structure of the article.

1 Lack of confirmability of drug development studies

Drug development proceeds in stages, from Phase 1 through Phase 3 to New Drug Application (NDA) or New Biologics License Application (BLA). Among the Phase transitions, the Phase 2 to Phase 3 transition rate at 39% as reported by Hay et al. (2014) is the most stringent, with rates of 67% for Phase 1 to 2, 68% for Phase 3 to NDA/BLA, 86% from NDA/BLA to approval. Yet, the success rate at Phase 3 leading to approval is only ≈50% as reported by Arrowsmith (2011b).

This fact seems puzzling since only Phase 2 studies with promising results graduate to Phase 3 studies, and Phase 3 studies have large sample sizes typically calculated to achieve "power" of 80%, 90%, or even higher. Within late Phase studies, we aim to increase the Phase transition success rate, so that a promising Phase 2 result will have a high probability of being confirmed in a Phase 3 study, for example.

For simplicity, we phrase the setting of this article as comparing a new treatment Rx with a control treatment C (a placebo or a standard-of-care). Our *reduce-variability principle* for enhancing the rate of success transitioning from a *feeder* study to a *confirmatory* study has the following decision path:

Misleading feeder inference? Learn, from the feeder study data itself, whether the promising result may be selective from highly variable potential results;

No Proceed to a confirmatory study.

Yes Go to whether reducing variability is worthwhile checking below.

Variability reducible? Can ETZ variability components be sufficiently reduced to achieve high confidence in confirmability?

No Terminate development.

Yes Allocate resources to reduce impactful ETZ variabilities, then proceed.

Drug development studies are typically randomized controlled trials (RCTs), so unmeasured confounders that might mislead observational studies are unlikely to be the causes of failure. For studies with repeated measures, we have a way of decomposing variability into three clinically meaningful components (abbreviated as E, $Traj$, and Z), beyond just *between-* and *within-* patient variabilities. We also have a way of assessing separately to what extent each component can cause a failure to confirm, so that resources can be judiciously allocated to reduce the most impactful component. A key innovation of our research is to show ETZ variabilities can be estimated from three variances routinely reported in RCTs.

Confirmability as distinct from reproducibility and replicability We follow the language in National Academies of Sciences, Engineering, and Medicine (2019) for *reproducibility* and *replicability*, adding the terminology of *confirmability* for clarity.

Reproducibility means that independent investigators can reproduce the reported figures, tables, point estimates, confidence intervals, etc., of a specific study from its original data, performing the same or equivalent analyses, using the same or equivalent computer codes. Reproducibility Research checks are exercised by many journals before publication for reproducibility purposes.

In the context of this article, we interpret *replicability* as the *result* of a study that can be essentially replicated by a similarly study (similar in design including sample sizes, comparing the

same or similar class of compounds, toward the same or similar indications). Replicability would apply within each Phase of an RCT. Arrowsmith (2011a) showed the overall success rate of Phase 2 studies is less than 20%, and Arrowsmith (2011b) showed approximately 50% of Phase 3 drug development trials fail, so replicability rate may differ by Phases of drug development.

Dissimilarities between a feeder study and a confirmatory study It is an oversimplification to think of a Phase 3 confirmatory study as like its Phase 2 feeder with a larger sample size.

Either a Phase 2 study with promising result, or a Phase 3 study which fails its original objective but shows tantalizing possibilities for the compound, can be a feeder to a "confirmatory" Phase 3 study. Besides sample sizes, common examples of how feeder and confirmatory studies differ are as follows.

In transitioning from a promising dose-finding Phase 2 study to a Phase 3 confirmatory study, often only the most promising dose is selected. In transitioning from a Phase 3 study which fails to show efficacy in the primary endpoint but appears to have efficacy in a key secondary endpoint, the roles of primary and secondary endpoints may be reversed in a subsequent Phase 3 confirmatory study. In both cases, if a subgroup of patients appears to obtain higher efficacy than the overall population in the feeder study, it is common for the subsequent confirmatory study to primarily target that subgroup of patients.

A study fails if, at the end of it, no clinically meaningful separation between the effects of Rx and C is observed. *High variability* in the observed endpoint separation of Rx and C, is the root cause of the low rate of confirmability. Our ability to quantify how reducing each ETZ variability component can increase the chance of successful confirmation allows the drug developer to decide whether to invest in additional resources to reduce the variability of a particularly impactful component. Our research is toward increasing the *confirmability* of study results.

1.1 Therapeutic areas with Repeated Measures studies

It takes at least two measurements to estimate a variability component, so traditional thinking is to separate between-treatment, between-patient, and within-patient variabilities. Ideally, a crossover study should be conducted, so each patient is given each treatment at least twice. However, in typical comparative efficacy clinical trials, each patient is given only one treatment.

The setting of this article is randomized controlled trials (RCTs) in which each patient is given *either* Rx or C but not both, and treatment outcomes are measured *repeatedly* at the *same* time points on each subject, at Week 0, Week 4, ..., Week 80 for example. Randomization and repeated measures allow us to separate the variability components, as we will show.

Therapeutic areas with randomized controlled trials that repeatedly measure patient outcome over time include diabetes such as type 2 diabetes mellitus (T2DM), neurodegenerative diseases such as Alzheimer's Disease (AD), psychiatric diseases such as schizophrenia, immunological diseases such as rheumatoid arthritis, ulcerative colitis, Crohn's disease, and psoriasis. Toward more confirmable drug development studies, Table 1 gives an idea of the approximate number of patients that may benefit from insights given by our research.

Therapeutic Area	Number of Patients (in millions)	Region
Crohn's disease and ulcerative colitis	3	U.S.
Rheumatoid arthritis	54	U.S.
Schizophrenia	20	worldwide
Alzheimer's disease	50	worldwide
Type 2 diabetes	400	worldwide

Table 1: Possible therapeutic areas influenced by our research.

Typically in such studies, the measure of treatment effect is *change from baseline*, the *difference* in patient outcome between the milestone visit and the first visit. Milestone visit is the visit when primary efficacy analysis is taken, while the first visit (baseline) is the only visit when patients

are not given any treatment. In most cases, the primary efficacy measure of Rx versus C is the expected *difference* of the primary outcome measure (the difference of long run change-from-baseline averages).

At a high level, one can conceptualize that there is *between* patient variability and a *within* patient variability. There are actually two components of *between* patient variability, as follows. At entry to a clinical trial, before any treatment is given, there is patient-to-patient variability (which we will later call *intercept* variability). Once treatments are given, then within each treatment arm there is variability in how patients progress over time (which we will later call *trajectory* variability). There is also *within* patient variability, from random errors in measuring treatment outcome on each patient. We will refer this variability to *measurement error* variability. Two examples illustrate how these variabilities can be big or small.

In an efficacy Type 2 diabetes mellitus (T2DM) study, the primary treatment outcome measure is glycated haemoglobin A1c (HbA1c). HbA1c measures the percent of red blood cells in the patient's blood with glucose attached to them, averaged over the last 3 months. HbA1c's change-from-baseline is a *validated* surrogate endpoint for reduction of microvascular complications associated with diabetes mellitus. See FDA-NIH Biomarker Working Group (2016). One reason for using HbA1c (as opposed to measuring microvascular complications as outcomes) is its relatively small measurement error, with a reported standard deviation (SD) of 0.17. A Trulicity® study reported patients with entry HbA1c between 6.5 and 9.5 (that being the patient entry criterion) had a standard deviation (SD) of baseline HbA1c about 1.1. Even though this baseline variability includes both measurement error variability and intercept variability (which will be separated later in the article), nevertheless at this point of the article, one can sense for T2DM measurement error variability (with a SD of 0.17) is small compared to the between-patient intercept variability.

While diabetes has been treated for more than a century, developing *disease-modifying* treatments for Alzheimer's disease (AD) remains challenging. In AD studies, patient outcomes are often their cognitive (Cog) abilities as measured on the Alzheimer's Disease Assessment Scale-Cognitive (ADAS-Cog) scale, and abilities to perform instrumental activities of daily living (iADL) as measured on

the Alzheimer Disease Cooperative Study-Instrumental Activities of Daily Living (ADCS-iADL) scale. ADAS-Cog11, for example, test 11 cognition items such as Word Recall, Naming Objects, etc. Treatment effects are commonly assessed as change-from-baseline of ADAS-Cog and ADCS-iADL. Scored by raters, a patient's ADAS-Cog measurement may be affected by whether they are having a good day or a bad day, and by rater variability.

For the EXPEDITION3 study (Honig et al. (2018)) testing solanezumab for treating Alzheimer's disease, at the first visit, mean baseline ADCS-iADL was about 45 with a standard deviation (SD) of approximately 8. Later in this article we estimate SD of measurement error of ADCS-iADL to be about 3.3, an indication that for AD measurement error variability is not small compared to the between-patient intercept variability.

1.1.1 Strategies for reducing variability

We describe four potential ways of reducing variability, with examples.

Narrow Entry Narrow the patient entry criterion to reduce patient heterogeneity

- Type 2 diabetes: patients with HbA1c $\in (7.0, 9.0)$ instead of HbA1c $\in (6.5, 9.5)$

Rater Training Use more highly trained raters to reduce measurement error variability

- Schizophrenia: rate PANSS by raters at a centralized site

Surrogate Endpoint Use a surrogate endpoint less variable than clinical outcome

- Type 2 diabetes mellitus: measure HbA1c in the blood

Subgroup Targeting Enroll only patients with sufficient drug target expression

- Alzheimer's disease: patients with amyloid beta deposition (by PET imaging)

Imaging is costly, and so is having ratings done by highly trained raters at central locations (as opposed to by local clinicians or pathologists at trial centers). Surrogacy of biomarker-based endpoints also needs to be validated initially by large-scale studies that measure both the biomarker

endpoint and clinical outcomes side-by-side. With finite resources to allocate, the questions are "How much does each variability component contribute to a potential failure to confirm efficacy? Is it worth the expense of reducing it?" The ETZ decomposition of variability components (to be developed in Section 2 of this article) provides a principled approach of assessing how much each component of variability contributes to a potential failure to confirm. It provides a rational way to design a confirmability study for success.

To concretely motivate our approach, we describe in the next section the Phase 3 EXPEDITION3 Alzheimer's disease study.

1.1.2 EXPEDITION3 studying solanezumab for treating Alzheimer's disease

As reported in Honig et al. (2018), EXPEDITION3 studied solanezumab for treating patients with mild dementia due to Alzheimer's disease. A double-blind, placebo-controlled, Phase 3 trial, EXPEDITION3 involved a total of 2129 patients randomized to receive either solanezumab at a dose of 400 mg or a placebo, every 4 weeks for 76 weeks (1057 patients were assigned to receive solanezumab and 1072 to receive placebo). Scores at the initial (Week 0) visit were considered *baseline*, and thereafter treatment outcomes were measured at weeks 12, 28, 40, 52, 64, and 80.

The primary outcome was the change (from baseline to Week 80) in the score on the 14-item cognitive subscale of the Alzheimer's Disease Assessment Scale (ADAS-cog14), with higher scores indicating greater cognition loss. A key secondary endpoint was the change (from baseline to Week 80) of Alzheimer Disease Cooperative Study-Instrumental Activities of Daily Living (ADCS-iADL), which assesses complex activities such as using public transportation, managing finances, or shopping, with lower scores indicating greater functional loss. Figure 2 shows observed profiles of ADAS-cog14 and ADCS-iADL reproduced from Figure 2 in Honig et al. (2018). EXPEDITION3 did not demonstrate efficacy in the chosen primary endpoint ADAS-Cog14, but did show promising results in the secondary endpoint ADCS-iADL. Planning of EXPEDITION3 was done according to standard power and sample size calculation techniques. What this article shows is that there are under-recognized factors that make it difficult to judge which promising results in earlier Phase or in

the same Phase studies can be confirmed, and ETZ model in Section 2 will make under-recognized factors precise, so that they can potentially be mitigated.

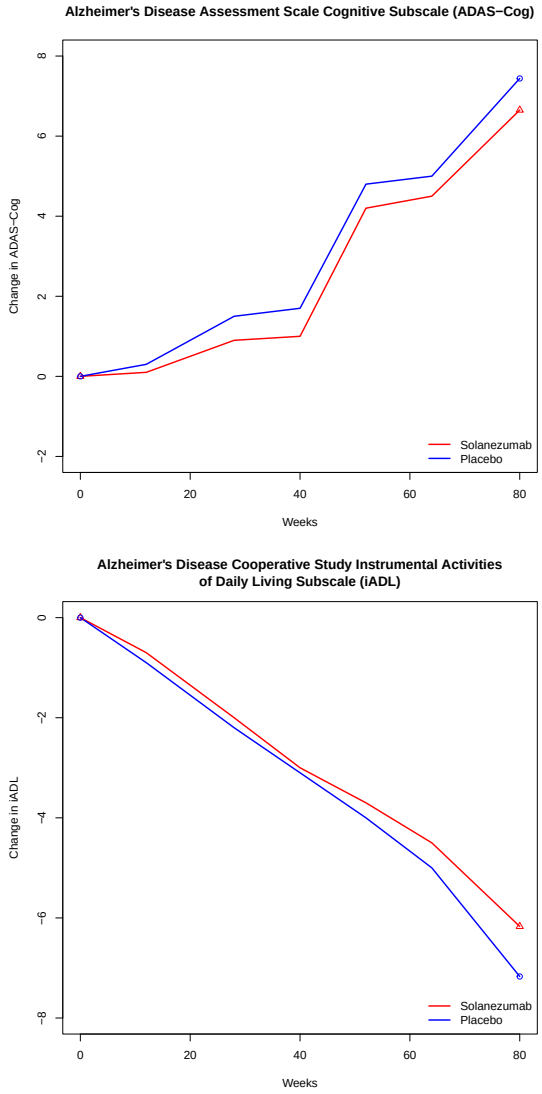

Figure 2: Observed profiles of solanezumab vs. placebo for ADAS-Cog14 and ADCS-iADL

For illustration purposes, since EXPEDITION3 indicated solanezumab may have potential in slowing the decline in ADCS-iADL (with a confidence interval for difference of change from baseline between solanezumab and the placebo being (0.17, 1.83)), we use the ADCS-iADL outcome in EXPEDITION3 to demonstrate the potential of our methodology to discern whether such observations can be confirmed in subsequent studies, or is an illusion due to variability, be it

person-to-person variability or measurement error.

Analyses in EXPEDITION3 were by mixed effects repeated measures modeling, with visits considered a categorical variable. Our methodology can, surprisingly, extract sufficient information from such routine analyses, to discern how variable future studies will be, under a variety of variability conditions.

1.2 Mixed Model Repeated Measures (MMRM)

Index each of the m visits of a repeated measures study by v, $v = 1, \ldots, m$. Denote by $Y_i^{Trt[v]}$ the outcome of a patient i given treatment Trt (either Rx or C) measured during visit v.

Denote the *time from randomization* of each subject as *Time*, which is a common factor for all the subjects in an RCT. There are two ways of modeling the repeated measurements. The first, a random coefficient model, to be discussed in Section 5.2, considers Time as a continuous variable, with the response of each subject generated randomly from a distribution of response trajectories around parametric response profiles representing average effects of treatments Rx and C over Time.

The second model is called Mixed Model Repeated Measures (MMRM)

$$Y_i^{Trt[v]} = \mu + \alpha^{[v]} + \beta^{Trt} + (\alpha\beta)^{Trt[v]} + b_i^{Trt} + w_i^{Trt[v]}, \ v = 1, \ldots, m. \tag{1}$$

The fixed effect is represented by $\mu + \alpha^{[v]} + \beta^{Trt} + (\alpha\beta)^{Trt[v]}$, the mean response under treatment Trt at time v, where $\alpha^{[v]}$ is the time effect, β^{Trt} is the treatment effect, and $(\alpha\beta)^{Trt[v]}$ is the treatment$\times$time interaction effect. Random between subject effect for the i^{th} patient assigned to treatment Trt is b_i^{Trt}. Random within subject effect is represented by $w_i^{Trt[v]}$.

In a repeated measures model, corresponding to the REPEATED statement in Proc Mixed of SAS, Time is not considered a continuous variable but rather just a label, a `CLASS` variable in SAS, or a `factor` in R. No profile of the response over time is modeled, but the `Type =` statement let the user specify a variety of structures of the $m \times m$ variance-covariance matrix of measurements on the subjects across the m visits. This specification does not separate variability from the randomness of the individuals' response trajectories and the randomness of measurement errors.

MMRM has become popular, from recommendations such as

> MMRM analysis appears to be a superior approach in controlling Type I error rates and minimizing biases, as compared to LOCF ANCOVA analysis (Siddiqui et al. (2009)).

2 The ETZ modeling principle

Index each of the m visits of a repeated measures study by v, $v = 1, \ldots, m$. Denote by $Y_i^{Trt[v]}$ the outcome of a patient given treatment Trt (either Rx or C) measured during visit v.

The ETZ modeling principle thinks of each patient as having their own random *intercept* before any treatment is given, with their random response *trajectory* around a *profile* of long-run average trajectories (be it linear, quadratic, Emax, what have you) once treatments are given. Values of profiles for Rx and C at times of visits are the fixed effects in the MMRM model (1). Similar to the MMRM model, ETZ does not assume a functional form for the profiles. Conceptually drawn as the curves in Figure 3, trajectories are not actually observed because measurements have errors. Observed, or potentially observed, are $Y_i^{Trt[v]}$, $v = 1, \ldots, m$, represented as dots and squares in Figure 3.

2.1 ETZ notations

To be more specific, conceptualize $Y^{Trt[v]}$ as having either two or three random components. At visit 1 ($v = 1$), before any treatment is given, $Y^{Trt[1]}$ has two random components, a component Z which is that patient's intercept (true baseline measurement), and a measurement error component E which is the difference between the true and the observed baseline. Thereafter, $Y^{Trt[v]}$ has three random components, a Z which is that patient's intercept (true baseline measurement), plus a random trajectory $Traj$ component (around its mean profile μ) representing that patient's response to the treatment given up to the time of visit v ($v > 1$), and a measurement error component E which is the difference between the measured outcome and that patient's true trajectory at each visit v. These components can be seen in Figure 3. Distinct from MMRM (1), the ETZ modeling

principle clinically separates measurements $Y^{Trt[1]}$ at visit 1 ($v = 1$) before any treatment is given from measurements $Y^{Trt[v]}$ at visits $v > 1$ once treatments are given. See (2) and (3) below. It is assumed that measurement error E is independent of Z and $Traj$, and i.i.d. over the visits, across all the patients.

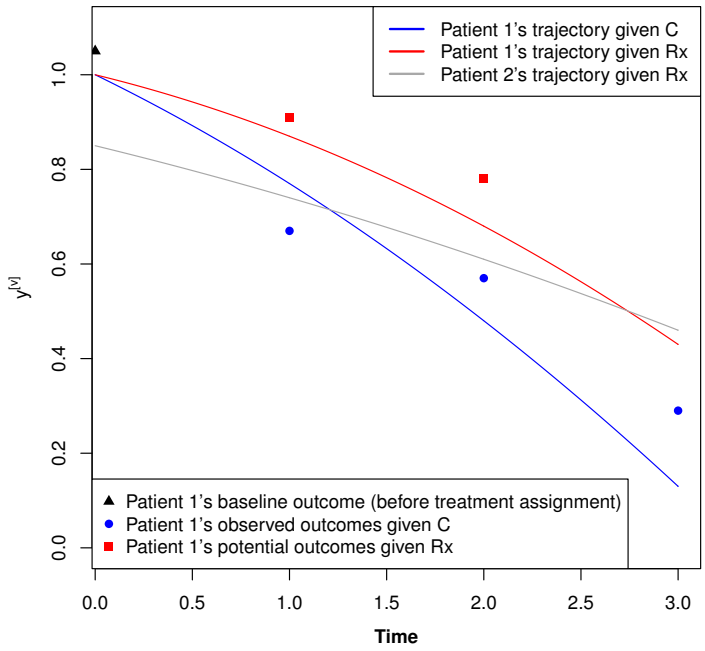

Figure 3: Components of the ETZ modeling principle for an RCT with $m = 4$ visits at Time = 0, 1, 2, and 3. The ETZ modeling principle only needs variability information for (1) baseline, (2) trajectory, and (3) milestone visits. No shape of the curve needs to be assumed, as Change-from-baseline sufficiently captures the trajectory information. The black triangle is the baseline of patient 1 before any treatment is taken. The blue curve is the trajectory of patient 1 who is assigned to C, and the blue dots are the observed Y. The red curve represents patient 1's potential trajectory had they been assigned to Rx, and the red squares are the potential Y (with the last value missing). The grey curve is the trajectory of patient 2 who is assigned to Rx. Measurement errors are differences between the outcomes and their trajectories.

Variabilities reported in clinical trials With random allocation of patients to Rx and C in a randomized controlled trial (RCT), one can estimate unbiasedly Rx versus C effect by comparing observed patient outcomes under Rx and C. Primary measure of treatment effect typically is *change* from baseline, $Y^{Trt[\text{change}]} = Y^{Trt[m]} - Y^{Trt[1]}$, so reducing $Var(Y^{Trt[\text{change}]})$ increases the chance of a successful confirmatory trial. A published result of an RCT typically includes an estimate of $Var(Y^{Trt[1]})$, and the standard error of the estimated mean of $Y^{[\text{change}]}$ from which one can infer an estimate of $Var(Y^{Trt[\text{change}]})$. Often an estimate of $Var(Y^{Trt[m]})$ is given as well.

It is not obvious how these variabilities relate to each other. For example, would narrowing the patient entry criterion, which presumably reduces $Var(Y^{Trt[1]})$, have much of an impact on $Var(Y^{Trt[\text{change}]})$? If it did, then narrowing patient entry criterion would increase the chance that a confirmatory study would be successful. It turns out that it actually does not, as we will show in Section 5.1.

Idea The $Traj$ (T) and Z parts of ETZ separate the between-subject effect b_i^{Trt} in the MMRM model (1) into two parts: a visit 1 pre-treatment part, and a treatment-dependent $Traj$ part starting with visit 2. The measurement error E part of ETZ is different from the within-subject measurement error in the model (1). Each of E, $Traj$, and Z is medically interpretable, and their variabilities can be controlled to varying extents. For example, while intercept variability $Var(Z)$ can be reduced by narrowing the patient entry criterion, trajectory variability $Var(Traj)$ can be reduced by decreasing drug target heterogeneity among patients. Variability in measurement error $Var(E)$ can be reduced by using more highly trained raters, for example.

High variability of change-of-baseline separation between Rx and C causes a low rate of successful confirmatory studies. How much of an impact on $Var(Y^{Trt[\text{change}]})$ will reduce each of $Var(Z)$, $Var(Traj)$, $Var(E)$ have? That question will be answered in Section 5, after we show the three variances typically reported are sufficient to estimate variabilities of Z, $Traj$, and E.

Plausibility Intuitively, variability at the first and milestone visits and of Change may contain all the information we need to separate the E, $Traj$, and Z variability components. Baseline measurements at the beginning of the study (visit 1) have the random intercept Z and measurement error E components but no random trajectory component $Traj$ since patients have not been treated yet. Change, or change-from-baseline, measurement at the milestone visit minus measurement at the first visit, has the random trajectory $Traj$ and the measurement error E components (two of E in fact) but not the random intercept component since that has been subtracted. Measurements at the end of the study (visit m) have all three random components. So it seems plausible that $Var(Y^{Trt[1]})$, $Var(Y^{Trt[m]})$, and $Var(Y^{Trt[\text{change}]})$ are sufficient for us to recover the E, $Traj$, and Z variability components (assuming variability of measurement errors is constant across visits).

We assume the random intercept Z and the random trajectory $Traj$ are *independent* within each treatment arm, that is, a patient's true severity of illness before they receive either Rx or C does not inform on whether that patient's response trajectory is going to be above or below the profile μ (long-run average of the trajectories). In addition, while trajectories for Rx and C are expected to have different profiles, we assume *variabilities* of the trajectories are the same for Rx and C.

2.2 Components of the ETZ decomposition

Assuming that the variability of measurement error at visit m is the same as the variability of measurement error at visit 1, our discussion starts from the setting in which random intercept Z is independent of the random trajectory $Traj$.

With such a conceptualization, outcomes of patient r given Rx and patient s given C at visit 1

and visit m are represented as

$$\begin{aligned} Y_r^{Rx[1]} &= Z_r + E_r^{[1]} && (2)\\ Y_r^{Rx[m]} &= Z_r + Traj_r^{Rx[m]} + E_r^{[m]} && (3)\\ Y_s^{C[1]} &= Z_s + E_s^{[1]} \\ Y_s^{C[m]} &= Z_s + Traj_s^{C[m]} + E_s^{[m]}. \end{aligned}$$

Implicitly, the fixed effects in this model are

$$E(Z_r) = \alpha^{Rx} \quad = \quad E(Z_s) = \alpha^{C} \qquad (4)$$

$$E(Traj_r^{Rx[m]}) = \mu_r^{Rx[m]} \quad \text{and} \quad E(Traj_s^{C[m]}) = \mu_s^{C[m]}. \qquad (5)$$

Change, or change-from-baseline, is measured by subtracting the baseline (visit 1) measurement from the last (visit m) measurement, so it does not have a random intercept component. *Change* of patient i given treatment Trt is then

$$Y_i^{Trt[\text{change}]} = \left[Traj_i^{Trt[m]}\right] + \left[E_i^{[m]} - E_i^{[1]}\right]$$

so

$$Var(Y_i^{Trt[\text{change}]}) = Var(E_i^{[1]}) + Var(Traj_i^{Trt[m]}) + Var(E_i^{[m]}) \qquad (6)$$

which leads to

$$\begin{aligned} Var(Z_i) + Cov(Z_i, Traj_i^{Trt[m]}) &= \frac{Var(Y_i^{Trt[m]}) + Var(Y_i^{Trt[1]}) - Var(Y_i^{Trt[\text{change}]})}{2} && (7)\\ &= Cov(Y_i^{Trt[m]}, Y_i^{Trt[1]}). && (8) \end{aligned}$$

In the case where Z and $Traj$ are independent, $Cov(Z_i, Traj_i^{Trt[m]}) = 0$, we can estimate $Var(Z_i)$ from (7), and estimate $Var(Traj_i^{[m]})$ from (3). In turn, we can estimate the variance of measurement error as

$$\begin{aligned} Var(E_i^{[1]}) &= Var(Y_i^{Trt[1]}) - Var(Z_i) && (9)\\ &= Var(E_i^{[m]}). \end{aligned}$$

Note that, if Z and $Traj$ are in fact positively correlated (say), then estimates of $Var(Z_i)$ and $Var(Traj_i^{[m]})$ would be somewhat higher than they ought to be, while the estimate for $Var(E_i^{[1]}) = Var(E_i^{[m]})$ would be somewhat lower than it ought to be.

2.3 Providing input to the ETZ decomposition

The ETZ decomposition (only) requires knowledge of the three quantities: $Var(Y_i^{Trt[1]})$, $Var(Y_i^{Trt[m]})$, and $Var(Y_i^{Trt[\text{change}]})$.

If patient-level data is available, then after modeling data using Proc Mixed REPEATED (for example), $Var(Y_i^{Trt[1]})$ and $Var(Y_i^{Trt[m]})$ can be directly obtained as the $[1,1]$ and the $[m,m]$ elements of the $\mathbf{R}$ matrix. Since $Cov(Y_i^{Trt[m]}, Y_i^{Trt[1]})$ is the $[1,m]$ element of the $\mathbf{R}$ matrix, one can then use the equality between (7) and (8) to obtain $Var(Y_i^{Trt[\text{change}]})$ as $Var(Y_i^{Trt[1]}) + Var(Y_i^{Trt[m]}) - 2Cov(Y_i^{Trt[m]}, Y_i^{Trt[1]})$. Obtaining the three variance estimates needed by ETZ from the same estimation process provides consistency.

If patient-level data is not available, then we obtain estimates of these three variances as best we can, from summary statistics reported for clinical trials, whatever form they may be in. For EXPEDITION3, we use the standard deviation (SD) of ADAS-Cog and ADCS-iADL at visit 1 and at visit m (Week 80) for both Rx and C provided in Table 2 of Honig et al. (2018). Note that, in this case, both visit 1 and visit m (Week 80) estimates are computed from *raw scores*. While the estimate of $Var(Y_i^{Trt[1]})$ would be computed from raw scores, if the estimate of $Var(Y_i^{Trt[m]})$ reported for a clinical trial is computed from *last observation carried forward* (LOCF) instead, one could use that. Also, as was the case for EXPEDITION3, typically standard error (SE) of least squares mean (LSmean) of *Change* is reported instead of $Var(Y_i^{Trt[\text{change}]})$. In such a situation, we convert the reported SE(*Change*) to an estimate of $Var(Y_i^{Trt[\text{change}]})$ using the reported sample sizes for Rx and C at visit 1 and visit m, in a fashion similar to how one would using Hedge's formula in meta analysis.

3 Partition to control the incorrect decision rate

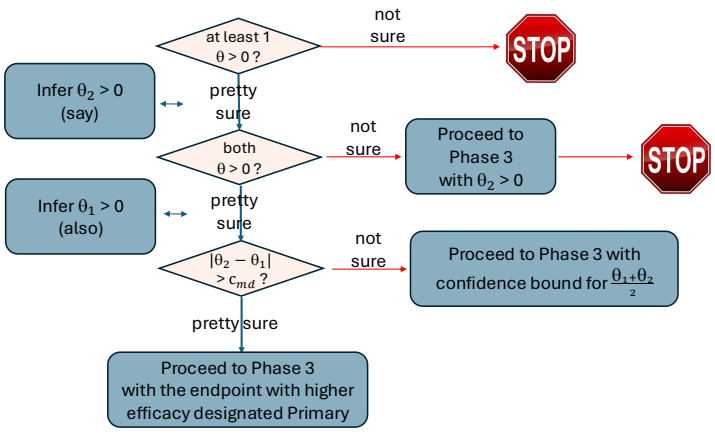

Figure 4: This flowchart illustrates the decision path for a Phase 2 study with two endpoints, guiding (1) whether to transition to Phase 3, and (2) based on inference on the efficacy of the two endpoints, whether to designate one endpoint as primary or to combine both endpoints into a single composite endpoint.

In the two-endpoint study setting, this article aims to address the following two questions:

1. **Confirmatory Transition Question:** Whether at least one of the endpoints shows positive efficacy; and

2. **Endpoint Designation Question:** If both endpoints show efficacy, which of the two should be designated as the primary endpoint, or whether a combination of the two should serve as the primary endpoint due to their comparable efficacy.

Testing equality nulls may not control Incorrect Decision rate In testing a single parameter θ, one cannot assume controlling the Type I error rate for testing $H_0^{=} : \theta = 0$ automatically controls the directional error rate, because directional errors are not counted in the Type I error definition of testing an equality null. In everyday practice, danger of rejecting for wrong reasons when the

hypotheses do not cover the entire parameter space is real. Section 7 of Liu et al. (2021) shows, for a level-5% log-rank test commonly used in survival analysis to test the null hypotheses that patients treated with Rx and C have equal survival probability within every time interval between events, the incorrect decision rate of deciding Rx increases the median survival time relative to C when in truth it does not exceed 15%.

The Partitioning Principle for hypothesis testing Partition the entire parameter space into null hypotheses where (potentially different) Type I error can occur, as well as (possibly) an alternative hypothesis where no Type I error can occur. By definition of Tukey (1953) and Tukey (1994b), Error Rate Familywise (often phrased as Familywise Error Rate, abbreviated as FWER, in current practice), testing each partitioned null hypothesis at level-α controls FWER (strongly), over the entire parameter space.

In contrast to decision-making by testing equality nulls, making decisions based on a confidence set controls the Incorrect Decision rate:

Theorem 1. *If inferences are made with no point contained in a* $100(1-\alpha)\%$ *confidence set contradicting them, then one can be confident that all such inferences, no matter how many, will be simultaneously correct with probability* $(1-\alpha)$.

Proof. By definition, a $100(1-\alpha)\%$ confidence set contains the truth with probability $(1-\alpha)$. Therefore making inferences (with statements such as parameters are greater or smaller than specific thresholds) that do not contradict a $100(1-\alpha)\%$ confidence set will not result in any incorrect inference more than α proportion of the time. □

3.1 Partition the parameter space to construct confidence sets

The so-called Neyman Construction of a confidence set (Neyman (1937)) proceeds by testing for every possible parameter value *in the entire parameter space* the null hypothesis that it is the true parameter. Since there is only one true parameter value, testing each null hypothesis at level-α controls the incorrect decision at α. Collecting all the parameter values for which their corresponding

nulls fail to be rejected results in a set C that we call a "confidence set", a (random) set which contains the true parameter value with a probability of at least $1-\alpha$. A formal proof of this straightforward fact is given on pages 421-422 of Casella and Berger (2001).

Tukey's and Dunnett's methods are confidence sets in *multi-dimensions* conveniently constructed by pivoting a multivariate-t statistic whose distribution depends neither on the unknown means of interest nor the nuisance variance-covariance parameters as follows. In the p-dimensional space, denote $(\theta_1, \ldots, \theta_p)$ by $\boldsymbol{\theta}$ and its point estimate $(\hat{\theta}_1, \ldots, \hat{\theta}_p)$ by $\hat{\boldsymbol{\theta}}$. Suppose that the event $\{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \in \boldsymbol{A}\}$ has a distribution that does not depend on any unknown parameter, with $\boldsymbol{A}$ scaled so that $P\{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \in \boldsymbol{A}\} = 1-\alpha$. Then the following 2-step process results in a level-(1-α) confidence set for $\boldsymbol{\theta}$:

Reflection Reflect $\boldsymbol{A}$ through the origin $\mathbf{0}$ to obtain $-\boldsymbol{A}$;

Translation Shift the origin $\mathbf{0}$ in $-\boldsymbol{A}$ to be at the point estimate $\hat{\boldsymbol{\theta}}$.

However, Tukey's and Dunnett's confidence sets do not target comparisons of specific interest in different parts of the parameter space. Amazingly, in the 1970s, Takeuchi (1973) developed sophisticated confidence sets targeting the determination of signs of $\theta_1, \ldots, \theta_p$. But since it was published in the Japanese language, his impressive work was largely unknown until re-discovered and reported at the 2007 MCP (Multiple Comparison Procedures) conference in Vienna, and honored at the 2009 MCP conference in Tokyo. Happily, we can now read this work in English in Takeuchi (2010).

Our Pivoting-Partitioning Confidence set construction below first uses Neyman Construction to partition the multi-dimensional parameter space into regions with different comparisons of interest in each, then uses pivoting to conveniently construct confidence subset in each region *directed* toward those comparisons, which are then coalesced into a confidence set for $\boldsymbol{\theta}$ in the entire parameter space.

Theorem 2 (Pivoting-Partitioning Confidence Set Construction)**.** *Partition the entire parameter space* Θ *into sub-spaces* $\Theta_1, \ldots, \Theta_M$*. If each* $A_1, \ldots, A_M$ *is scaled so that* $P\{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \in A_m\} = 1-\alpha$*,*

then a level $100(1-\alpha)\%$ *confidence region for* $\hat{\boldsymbol{\theta}}$ *is*

$$C(\hat{\theta}_1, \ldots, \hat{\theta}_p) = \bigcup_{m=1}^{M} \left(\{ -\boldsymbol{\theta} + \hat{\boldsymbol{\theta}} : -\boldsymbol{\theta} + \hat{\boldsymbol{\theta}} \in A_m \} \cap \Theta_m \right).$$

Proof. Denote by $\boldsymbol{\theta}^0$ a candidate true parameter value. By Neyman Construction, a level $100(1-\alpha)\%$ confidence set for $\boldsymbol{\theta}$ is

$$\begin{aligned} C(\hat{\boldsymbol{\theta}}) &= \{ \boldsymbol{\theta}^0 \in \Theta : H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}^0 \text{ is not ejected } \} \\ &= \bigcup_{m=1}^{M} \{ \boldsymbol{\theta}^0 \in \Theta_m : \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0 \in A_m \} \\ &= \bigcup_{m=1}^{M} \left(\{ \boldsymbol{\theta} \in -A_m + \hat{\boldsymbol{\theta}} \} \cap \Theta_m \right). \end{aligned}$$

□

As an example of how Pivoting-Partition confidence set construction has proven useful, suppose θ_1 and θ_2 are the primary and secondary endpoints respectively, and values no greater than zero indicate a lack of efficacy. Hsu and Berger (1999) partitioned the parameter space as $\Theta_1 = \{\theta_1 \leq 0\}$, $\Theta_2 = \{\theta_2 \leq 0 \text{ but } \theta_1 > 0\}$ and $\Theta_3 = \{\theta_1 > 0 \text{ and } \theta_2 > 0\}$. Since they are disjoint, no multiplicity adjustment is needed in testing $H_{0m} : \boldsymbol{\theta} \in \Theta_m$. As rejecting H_{02} indicates efficacy in the secondary endpoint only if H_{01} is rejected as well, this leads to testing H_{01} first and only if it is rejected then test H_{0s}. As another example, Stefansson et al. (1988) used Pivoting-Partitioning to construct a confidence set for the *step-down version* of Dunnett's method for comparing new treatments with a control.

Insight Partition testing of hypotheses and confidence set construction are *conditional* inferences, conditioning on *relevant subsets*, using observed data to eliminate subsets of the partitioned parameter space. Basis of the Partitioning Principle was Multiple Comparisons with the Best (MCB) as described in Chapter 4 of Hsu (1996), which Jason Hsu developed by connecting the idea of *conditional confidence* in Ranking and Selection in Section 4 of Kiefer (1977) with John W. Tukey's Multiple Comparisons framework. *Relevant subset* Neyman Construction confidence sets are unlikely to have issues such as the *marginalization paradox* that afflict Fiducial/Structural confidence sets

constructed by conditioning on *ancillary* statistics (such as *residuals*). Further, whereas Kiefer changed the *conditional confidence coefficient* (to higher or lower than the nominal confidence level) conditionally while keeping the shape of the confidence set constant, *directed* Partitioning confidence sets *shape-shift* to keep the conditional confidence level as constant as possible.

In order to answer the **Confirmatory Transition Question** and **Endpoint Destination Question** above, the two-dimension parameter space $\Theta \subseteq \mathbb{R}^2$ for θ_1 and θ_2 should be further partitioned into:

$$\Theta^{\leq\leq} : \theta_1 \leq 0 \quad \text{and} \quad \theta_2 \leq 0 \tag{10}$$

$$\Theta^{>\leq} : \theta_2 \leq 0 \quad (\text{but} \quad \theta_1 > 0) \tag{11}$$

$$\Theta^{\leq>} : \theta_1 \leq 0 \quad (\text{but} \quad \theta_2 > 0) \tag{12}$$

$$\Theta^{>>\nabla} : \theta_1 > 0 \quad \text{and} \quad \theta_2 - \theta_1 > c_{md} \tag{13}$$

$$\Theta^{>>band} : \theta_2 - \theta_1 \leq c_{md} \quad \text{and} \quad \theta_2 - \theta_1 \geq -c_{md} \tag{14}$$

$$\Theta^{>>\triangle} : \theta_2 - \theta_1 < -c_{md} \quad \text{and} \quad \theta_2 > 0 \tag{15}$$

where the c_{md} is a clinically meaningful difference predefined by clinicians.

(10), (11) and (12) correspond to the plot (b) of Figure 5 while (13), (14) and (15) can be further partitioned in Figure 6.

(10), (11) and (12) are partitioned to answer the **Confirmatory Transition Question**. A Phase 2 study can transition to a Phase 3 study when the Negative-negative quadrant (10) and at least one of the Positive-negative quadrant (11) and the Negative-positive quadrant (12) are not covered by confidence set. The contents in the brackets of (11) and (12) can be omitted because whether an endpoint shows efficacy is an one-sided problem.

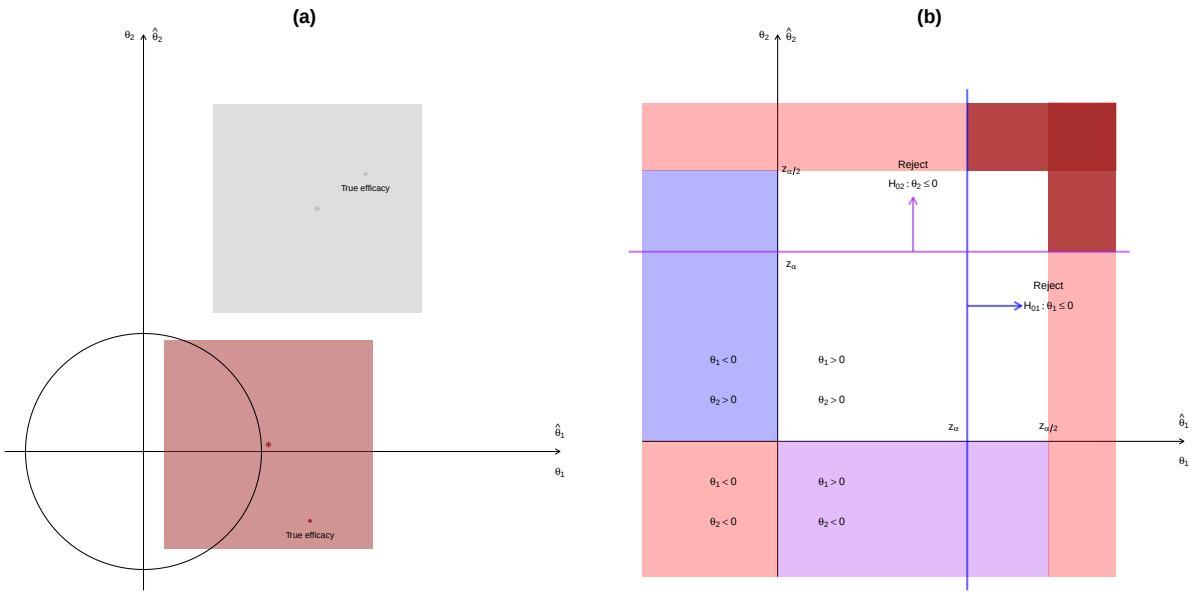

Figure 5: Correct and incorrect decision-making involving two efficacy parameters θ_1 and θ_2 (that are efficacious when positive). Plot **(a)** shows two true parameter points (red, gray) with their confidence rectangles. The F-test rejects $H_0\colon \theta_1 = \theta_2 = 0$ when $(\hat{\theta}_1, \hat{\theta}_2)$ lies outside the rejection circle. If the red point is the truth, the F-test would reject when $(\hat{\theta}_1, \hat{\theta}_2)$ is inside the gray rectangle perceiving incorrectly $\theta_2 > 0$ (leading to power calculation that includes rejecting for wrong reasons). In contrast, making decisions based on confidence sets yields confident two-sided Correct and Useful inference by Theorem 1. For example, the red confidence rectangle correctly infers $\theta_1 > 0$ only. Plot **(b)** shows typical one-sided rejection regions for the partitioned quadrants $\Theta^{\leq\leq}\colon\ \theta_1 \leq 0, \theta_2 \leq 0$; $\Theta^{>\leq}\colon\ \theta_1 > 0, \theta_2 \leq 0$; and $\Theta^{\leq>}\colon\ \theta_1 \leq 0, \theta_2 > 0$, with rejection of $\Theta^{\leq\leq}$ requiring multiplicity adjustment with bounds $b_{\rho_1}, b_{\rho_2} \in [z_\alpha, z_{\alpha/2}]$, while rejections of the single-efficacy quadrants do not.

The concept of *directed* confidence sets This concept of *directed* confidence set in Hsu and Berger (1999) is to shape the confidence set within each of partitioned region of the parameter space toward the inference desired for that region.

In Bioequivalence, the desired inference is $-\delta \leq \theta \leq \delta$ where δ is the equivalence margin. So the Neyman Construction of Hsu et al. (1994) tested for each $\theta^0 > 0$ the simple null $H_0 : \theta = \theta^0$

against the simple alternative $H_a : \theta = 0$ at level-α, while for each $\theta^0 < 0$ they also tested the simple null $H_0 : \theta = \theta^0$ against the alternative $H_a : \theta = 0$ at level-α. That each of these two sets of 1-sided tests in opposite directions is most powerful according to the Neyman-Pearson lemma makes the resulting Bioequivalence confidence interval have the smallest expected length among all level-$(1 - \alpha)$ confidence intervals if $\theta = 0$ in truth.

Instead of desiring "equivalence", to decide which endpoint should be designated Primary and the other Secondary in our Confirmability setting, the desired inference is *meaningful difference*. The directed Neyman Construction confidence set for that is given in Section 4.1 of Hayter and Hsu (1994).

Deciding whether to proceed to a confirmatory study, is a complex decision-making process involving multiple parameters.

Our view is that there are two possible strategies for choosing two endpoints:

↑↑ If a compound's efficacy is comparable in magnitude and variability in both endpoints (as in the Table 2), then combining the two endpoints into a single endpoint is reasonable both medically and as a strategy for gaining marketing approval;

⊥ if efficacy in the two endpoints are rather different, substantially higher in one than the other, then a sponsor would want to designate the stronger one to be the primary endpoint and the weaker one to be the secondary endpoint.

Endpoint	Estimated efficacy	95% Confidence Interval
ADAS-Cog14	-0.80	$(-1.73, 0.14)$
ADCS-iADL	1.00	$(0.17, 1.83)$

Table 2: Solanezumab's efficacy in ADAS-Cog14 and ADCS-iADL are comparable in magnitude and variability in EXPEDITION3

In July 2024, donanemab (Kisunla™) was approved by the FDA for treating patients with early symptomatic Alzheimer's disease. Its primary outcome measure iADRS (integrated Alzheimer's

Disease Rating Scale) combines ADAS-Cog13 and ADCS-iADL. While a higher ADCS-iADL is better, a lower ADAS-Cog is better. So iADRS is calculated by subtracting ADAS-Cog13 from ADCS-iADL, with a higher iADRS being better (less impaired). To anchor ADAS-Cog at zero, a constant 85 is added for ADAS-Cog13 (90 for ADAS-Cog14), see Wessels et al. (2015).

As one does not know whether the truth is $\uparrow\uparrow$ or $\perp$, we design the acceptance/rejection regions to achieve a "directed" confidence set.

3.2 Meaningful difference confidence sets construction

To answer the **Endpoint Designation Question**, the Positive-positive quadrant is partitioned into (13), (14), and (15), respectively, as shown in Figure 6. Confidence set decision-making for answering this question has two layers as was shown in Figure 4.

The first possibility is one endpoint shows significantly higher efficacy than the other, as indicated by all candidate values in either the top cone or the bottom cone in Figure 6 are eliminated from the confidence set, so which endpoint to designate as primary is clear. The second possibility is difference in efficacy between the two endpoints is not clinically meaningful, in which case a lower confidence bound of the combined efficacy (averaged over the endpoints) is computed.

Confidence construction for the top and bottom cones inspired by Neyman construction and section 4.1 in Hayter and Hsu (1994) is as follows.

Intuition for meaningful difference confidence construction As discussed in the insight for Theorem 1, Neyman's confidence set construction permits each partitioned region to adopt its own most informative acceptance region shape.

For each of the partitioned top and bottom cone regions, the desired inference is 1-sided, aiming to exclude parameter values lying in the opposite partitioned cone. Thus, the confidence set for the cone-shaped partitioned regions should be directed *away from* the origin, as opposed to the bioequivalence confidence set described in Section 3.1 being directed *toward* the origin.

Surprisingly, for 1-sided meaningful difference inference, it is necessary to construct a confidence

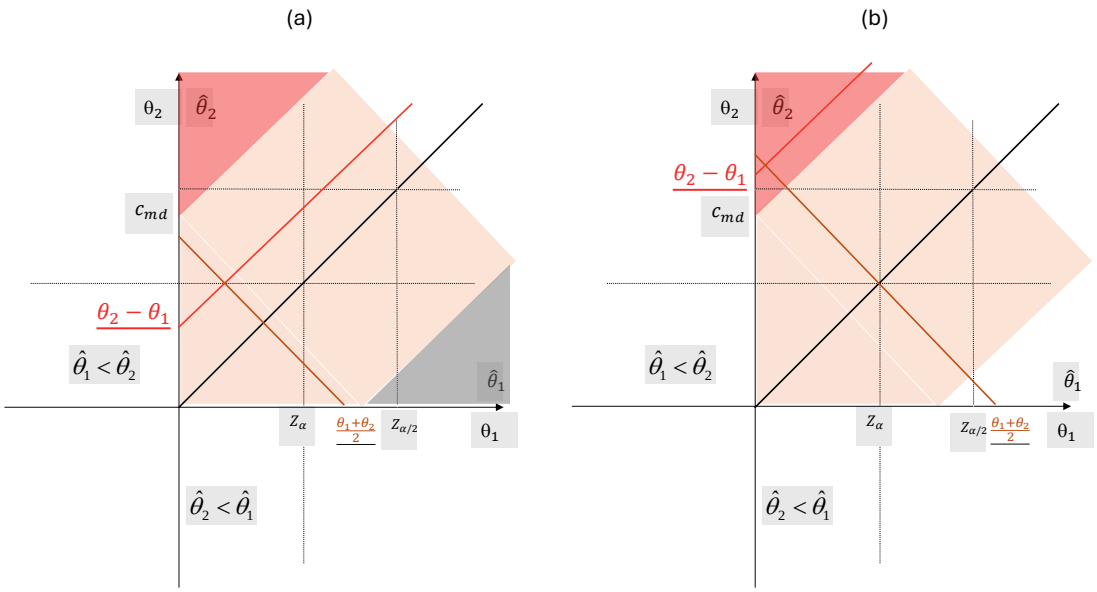

Figure 6: This figure presents two situations in the positive-positive quadrant, partitioned into three regions: the scarlet top cone $\Theta^{>>\nabla}$, the grey bottom cone $\Theta^{>>\triangle}$, and the beige middle band region $\Theta^{>>band}$. In the parameter space, if one can identify either $\theta \in \Theta^{>>\nabla}$ or $\theta \in \Theta^{>>\triangle}$, then endpoint 2 or 1 should be designated as the primary endpoint, respectively. Otherwise, one possible strategy is to combine the two endpoints into a single composite endpoint. For example, while Plot (a) suggests the two endpoints be combined into a single endpoint, Plot (b) indicates endpoint 2 should be designated as the primary endpoint.

set using a 2-sided acceptance region. That is because, as shown in Figure 7, the confidence set resulting from an 1-sided acceptance region will always contain values from both cone-shaped partitioned regions, unable to **usefully** infer that one endpoint has meaningfully higher efficacy than the other. Only with a 2-sided acceptance region can the constructed confidence set conclude that a specific endpoint should be designated as the primary one.

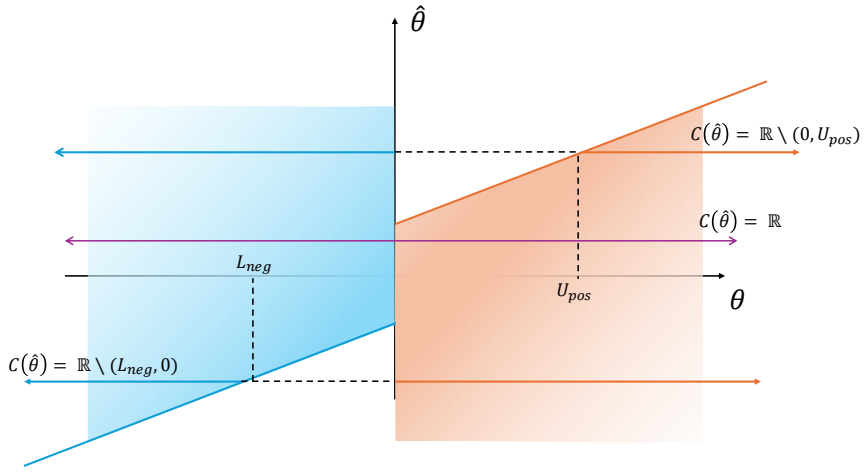

Figure 7: This plot explains why one-sided acceptance regions are ineffective. The parameter space is partitioned into $\Theta^- = \{\theta : \theta \leq 0\}$ and $\Theta^+ = \{\theta : \theta > 0\}$, and two one-sided acceptance regions in opposite directions, represented by regions of the two different colors, are considered for the partitioned Θ^- and Θ^+. Regardless of $\hat{\theta}$, the Neyman construction confidence set cannot entirely eliminate either Θ^- or Θ^+.

We show Neyman construction of a confidence set for the top cone as an example.

Example For each candidate parameter value in the top (scarlet) cone, a *lower* bound of the acceptance region testing it is needed, so that if the true parameter value is in the lower (grey) cone, the entire top cone is eliminated. For that purpose, we can use the same fixed lower bound $-\sigma_{diff} z_{\alpha/2}$ as one would in traditional 2-sided testing, without tailoring it to the specific candidate parameter value.

However, when $\hat{\theta}_{diff}$ is able to eliminate the entire lower cone, we would want the confidence bound to be as positively away from the origin as possible. For that purpose, we let the upper bound of the acceptance region testing each candidate value to increase with it, so the asymmetric acceptance region becomes

$$A(\theta_{diff}) \quad = \quad \{\hat{\theta}_{diff} : -\sigma_{diff} z_{\alpha/2} \leq \hat{\theta}_{diff} \leq \overline{e}_{\alpha,|\theta_{diff}|}\}, \tag{16}$$

satisfying

$$P(-\sigma_{diff} z_{\alpha/2} \leq \hat{\theta}_{diff} \leq \overline{e}_{\alpha,|\theta_{diff}|}) = 1 - \alpha, \tag{17}$$

where $\hat{\theta}_{diff} \sim N(\theta_{diff}, \sigma^2_{diff})$ with $\overline{e}_{\alpha,|\theta_{diff}|}$ increasing monotonically from $\sigma_{diff} z_{\alpha/2}$ to $\theta_{diff} + \sigma_{diff} z_{\alpha}$ as θ_{diff} increases from 0 to $+\infty$. The resulting directed confidence set for the top cone is a one-sided confidence interval $\theta_{diff} \in [\hat{\theta}_{diff} - d_{\alpha,|\hat{\theta}_{diff}|}, +\infty)$ when $\hat{\theta}_{diff} > \sigma_{diff} z_{\alpha/2}$ with $d_{\alpha,|\hat{\theta}_{diff}|}$ being the solution to

$$\Phi\left(\frac{d_{\alpha,|\hat{\theta}_{diff}|}}{\sigma_{diff}}\right) - \Phi\left(\frac{(-\sigma_{diff} z_{\alpha/2}) - (\hat{\theta}_{diff} - d_{\alpha,|\hat{\theta}_{diff}|})}{\sigma_{diff}}\right) = 1 - \alpha, \tag{18}$$

but is a non-informative $\theta_{diff} \in (-\infty, +\infty)$ when $|\hat{\theta}_{diff}| \leq \sigma_{diff} z_{\alpha/2}$ (corresponding to the situation that traditional 2-sided testing fails to reject $H_0 : \theta_{diff} = 0$).

A directed confidence set for the bottom (grey) cone can be constructed analogously.

Consider that the two endpoints are ADCS-iADL and ADAS-Cog14. The difference in efficacy can be denoted as $\theta_{diff} = \theta_{iADL} - \theta_{Cog}$, where θ_{iADL} is the scaled efficacy for ADCS-iADL and θ_{Cog} is the opposite of the scaled efficacy for ADAS-Cog14. Both θ_{iADL} and θ_{Cog} are scaled such that higher values indicate better efficacy.

Using the estimates from Table 2 for solanezumab as an example, and starting with the assumption of independence as a baseline, we have the estimated difference in efficacy $\hat{\theta}_{diff} = 0.2$, with an estimated standard deviation of $\hat{\sigma}_{diff} = 0.98$.

Since $\hat{\theta}_{diff} < \hat{\sigma}_{diff} z_{\alpha/2}$, the confidence interval for θ_{diff} includes both cone-shaped partitioned regions. Thus, it is reasonable to consider combining the two endpoints, as is done with donanemab.

Define the average efficacy as $\theta_{avg} = \frac{\theta_{iADL} + \theta_{Cog}}{2}$. The estimated average efficacy is $\hat{\theta}_{avg} = 0.9$, with an estimated standard deviation of $\hat{\sigma}_{avg} = 0.49$. The lower bound of the confidence interval is

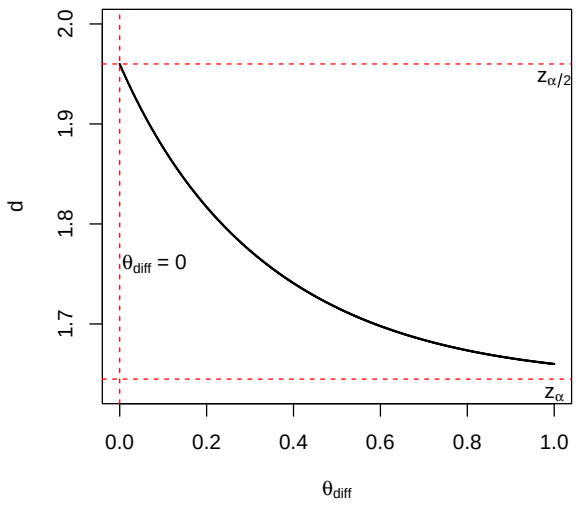

Figure 8: The allowance d decreases from z_α approximately to $z_{\alpha/2}$ as θ_{diff} increases from 0^+ to $+\infty$.

0.09, which is positive, supporting the designation of the combined efficacy as a primary endpoint.

4 Phase 2 to 3 Transition Decision-making Process

Consider designing a confirmatory study to meet a high prespecified probability of "success". Suppose the statistical analysis plan has a 5% chance of making an incorrect inference. Then for any design to claim that it has a higher than 95% chance, 97% "power" say, of leading to a "success" is unreasonable, because in that case the definition of "success" evidently includes at least a 2% chance of erroneous or false "success".

That the power is set so high is perhaps due to experience with high Phase 2 to Phase 3 transition failure rate, as observed in Hay et al. (2014) for example. Our understanding is often there is some off-the-books "discounting" of the calculated *power*, to be conservative. With the ETZ modeling principle able to take all sources of variability into account, instead of informal discounting, this article offers a direct calculation of success confidence, which is the probability of true success

in transitioning from a feeder study to a confirmatory study (in analogy to a confidence interval guaranteeing coverage probability).

4.1 Replacing $P\{$*correct and useful inference*$\}$

The concept of "power" is meant for 2-action problems such as testing a simple null hypothesis against a simple alternative. For problems with multiple parameters, such calculation of "power" is legacy practice from B.C. (before computers) when comparing k means typically used the F-test for testing $H_0 : \mu_1 = \ldots = \mu_k$ in an ANOVA setting (e.g., Scheffé (1959) Section 3.3). However, as Hsu (1989) pointed out, included in the power calculation of an omnibus test is the probability of *rejecting for wrong reasons*. Suppose, in comparing high, medium, and low doses of a psychiatric drug against a placebo, the truth is the medium dose is effective but high and low doses are not. Power calculation of the F-test (e.g., by Proc Power and Proc GLMpower in SAS) includes the probability that it rejects for the wrong reason that data indicates incorrectly high or low dose is effective but the medium dose is not, for example. With modern computing power, we believe this legacy practice should be ditched in view of the following.

The Correct and Useful inference principle Statistical inferences are truly useful only if they are correct.

To follow this principle, Hsu (1989) and Appendix C of Hsu (1996) defined the result of a study to be correct and useful if the following two events simultaneously occur:

Correct = {no new treatment worse (better) than the control will be incorrectly inferred to be better (worse) than the control}

Useful = {all treatments sufficiently better (worse) than the control will be inferred to be better (worse) than the control}

Clearly, if $\inf P\{\text{Correct}\} = 95\%$, and $P\{\text{Useful}\} < 100\%$, then $P\{\{\text{Correct}\} \cap \{\text{Useful}\}\} = P\{\text{Correct and Useful inference}\} < 95\%$ which is one aspect of the *correct* and *useful* inference concept more reasonable than that of "power".

To make the event {Correct and Useful inference} occur with sufficiently high probability requires replacing omnibus testing by simultaneous inference methods such as Tukey's and Dunnett's which are capable of inferring the magnitude and direction of individual treatment efficacy. These methods only became computationally feasible for the General Linear Model from the development of the Factor Analytic algorithm by Hsu (1992). The ProbMC function in SAS, which implements the FA algorithm, can be used to calculate the sample size needed to achieve the desired $P\{\text{Correct and Useful inference}\}$.

As illustrated in Figure 5, operationally, if the clinically meaningful threshold for a (component) parameter θ_i is δ_i, then one infers $\theta_i > \delta_i$ only if the confidence interval for θ_i (given by the confidence set) is entirely larger than δ_i (similarly for inferring $\theta_i < \delta_i$). With confidence intervals covering their true values guaranteeing *correctness*, one can design a study ensuring the event {Correct and Useful inference} occurs with desired probability by having sufficient sample size so that the confidence intervals will be sufficiently tight.

4.2 Acknowledging uncertainty in true efficacy

To decide whether it is a correct decision to proceed from a feeder study to a confirmatory study, current industry practice seems to be to take estimates from the feeder study (and other sources) as the truth, and compute the sample size needed to adequately "power" the confirmatory study, using

$$n = 2(z_{\alpha/2} + z_{\beta})^2(\sigma/\Delta)^2 \tag{19}$$

where α is the probability of type I error, β is the probability of type II error, Δ is the true separation between Rx and C at the milestone visit of the study (visit m), σ is the true standard deviation of the outcome, z_γ denotes the upper γ quantile of the standard Normal distribution.

For example, page 260 of the published protocol of EXPEDITION3 states

> "Based on data from completed Phase 3 solanezumab Studies LZAM and LZAN ... in mild AD patients with evidence of amyloid pathology ... for ADAS-Cog14, we assume a treatment difference of approximately 2 points at 18 months with a standard deviation

> of 10. For the ADCS-iADL, we assume a treatment difference of 1.5 points at 18 months with a standard deviation of 10. ... 1050 randomized patients per arm, or 2100 randomized patients total ..., will have approximately 97% power for the ADAS-Cog14 comparison and approximately 84% power for the ADCS-iADL comparison to detect a significant treatment difference at 18 months using a 2-sided significance level of 0.05."

The concept of Correct and Useful inference is a principle applicable to complex decision-making, with CBA's *coverage* being a form of *correctness* and *narrowness* being a form of *usefulness* in the simple setting of inference on a single parameter. In the Confirmability setting, we define "*true* success" as, at the end of the confirmatory study, a truly efficacious compound is correctly inferred to be efficacious. True success requires the occurrence of the two events

Transition Estimated efficacy from a Phase 2 feeder study conservatively justifies transitioning,

Confirmation The Phase 3 study successfully confirms efficacy,

the probabilities of which depend not only on true efficacy but also on variabilities.

Our Confidently Bounded Quantile (CBQ) follows the correct and useful inference principle. CBQ is a decision-making process that designs a study with adequate sample size and control of variabilities to ensure a pre-specified success confidence $P\{\text{True success}\} \geq \gamma$ (e.g., 85%).

Replacing off-the-books discounting of *power*, CBQ is a formal 2-step conditional discounting process. The first discounting is, instead of a point estimate of true efficacy, a Confident Efficacy is estimated from Phase 2 data. Then, following the Conditionality Principle which states statistical inference should be based on the experiment actually performed (Berger (1985), page 31), *conditional on this Confident Efficacy*, the second discounting calculates a *Confidently Bounded Quantile* for the Phase 3 outcome taking confirmatory study variability into account.

To be precise, visualize making transitioning decisions over infinitely many Phase 2 to potential Phase 3 studies. Define

- **event** A as the event that Phase 2 study transitions to Phase 3 study correctly, and

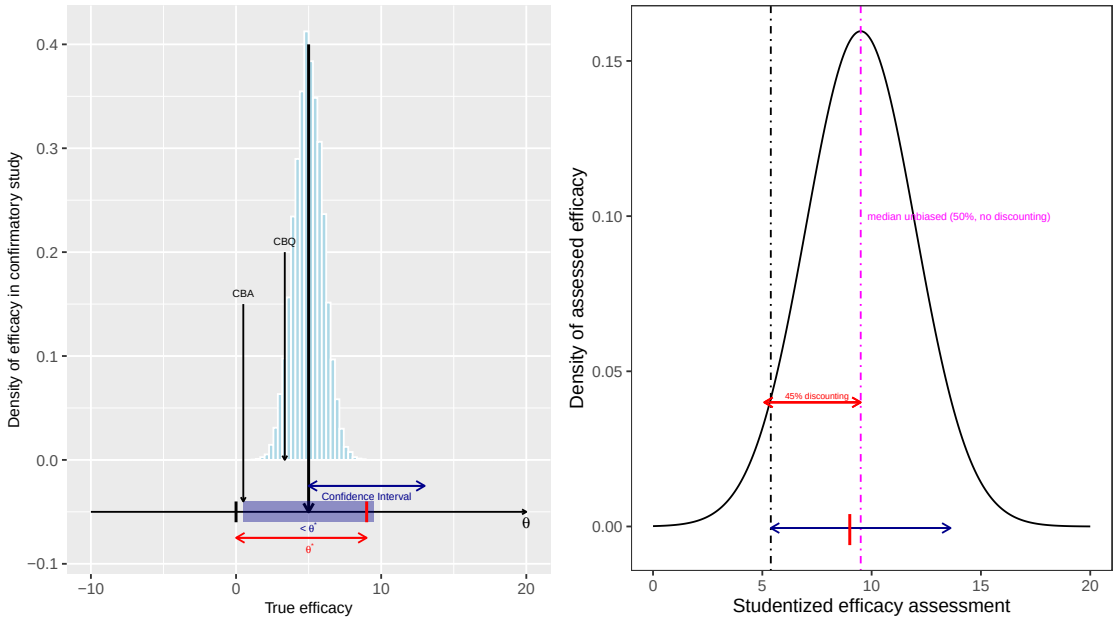

Figure 9: The plot on the left panel shows Confidently Bounded Quantile (CBQ) versus Confidently Bounded Allowance (CBA). The horizontal axis represents efficacy with the **short vertical red segment** being the true efficacy. The double-arrow horizontal bar represents what a 95% confidence interval from the feeder (Phase 2) study centered around the true value might look like. The semi-transparent dark blue area visualizes a possible Tukey's CBA with a confidence interval rather off-centered but covers the true value (so the inference will be correct) and is sufficiently narrow to not cover zero (so the inference will be useful). On the other hand, our CBQ takes a conservative limit of the **blue** double-arrow confidence interval as the true efficacy, and simulates a confirmatory study (Phase 3) 10000 times, resulting in the light blue histogram. CBQ is then the quantile of this histogram that ensures $P\{\text{True success}\} \geq \gamma$. The plot on the right panel shows Phase 2 discounting. A 95% lower confidence limit of efficacy from Phase 2 data corresponds to a Phase 2 discounting of 45%, relative to a 50/50 (median unbiased) point estimate. (A 50% confidence limit has a 50/50 chance of being higher/lower than the true value.)

- **event** B as the event that Phase 3 study shows efficacy successfully.

$P(A \cap B)$, the overall probability of making a correct decision, equals $P(A)P(B \mid A)$, where $P(B \mid A)$ is the probability that the Phase 3 study shows efficacy conditional on the confidence interval for true efficacy from the Phase 2 covering the true value. Instead of using the Bonferroni inequality as in CBA, we calculate the multivariate probability precisely using modern computing facilities, interfacing with users in the form of an interactive app.

CBQ calculates a Confident Efficacy by taking a conservative confidence limit of true efficacy, then uses it to confidently estimate a Phase 3 outcome by replicating the Phase 3 study with a fixed sample size many times. To achieve confidence, two discountings are employed for distinct purposes:

- **Phase 2 discounting** confidently assesses efficacy from the feeder study, accounting for variability between Phase 2 studies.
- **Phase 3 discounting** confidently predicts potential outcome of the Phase 3 study, accounting for variability replicating a Phase 3 study.

Each discounting is relative to a 50/50 (median unbiased) point estimate. As illustrated in Figure 9, a Phase 2 *discount* of d_{Phase2} controls $P(A)$ at $P(A) = d_{Phase2} + 50\%$. For example, if the Phase 2 discounting is $d_{Phase2} = 45\%$, then Phase 2 discounting (conservatively) calculates true efficacy as its lower $45\%+50\% = 95\%$ confidence limit from the Phase 2 data.

A Phase 3 *discount* of $d_{Phase3|Phase2}$ controls $P(B \mid A)$ at $P(B \mid A) = d_{Phase3|Phase2} + 50\%$. So, with $P\{A \cap B\} = P(A)P(B \mid A) = (d_{Phase2} + 50\%)(d_{Phase3|Phase2} + 50\%)$, given a (fixed) desired success confidence $P\{\text{True success}\} = \gamma$, setting d_{Phase2} determines $d_{Phase3|Phase2}$, and setting $d_{Phase3|Phase2}$ determines d_{Phase2}. For example, if $\gamma = 80\%$ and one sets $d_{Phase2} = 45\%$, then $d_{Phase3|Phase2} = 34.21\%$, so that $P\{A \cap B\} = (45\% + 50\%)(34.21\% + 50\%) = 80\%$.

4.3 Assessing EXPEDITION3 as a feeder example using the App

Our Confirmability App is accessible at https://xwukstatsci.shinyapps.io/confirmability/. Its *Transition Decision-Making Process Plots* tab calculates whether the confident quantile exceeds

the desired threshold. We illustrate this using summary statistics from the EXPEDITION3 study described in Section 1.1.2. For the EXPEDITION3 study, with Rx being solanezumab and C being a placebo, Figure 2 from Honig et al. (2018) shows the estimated profile for Alzheimer Disease Cooperative Study-Instrumental Activities of Daily Living (ADCS-iADL) is approximately linear, while the estimated profile for ADAS-Cog14 shows some non-linearity. The confidence interval for the difference of change-from-baseline between Rx and C at 80 weeks was $(-1.73, 0.14)$ for ADAS-Cog14 and $(0.17, 1.83)$ for ADCS-iADL. The sample sizes for both treatment arms and estimations for fixed effects in both treatment arms with standard deviation (SD) or standard error (SE) are documented in Table 3. We use ADCS-iADL to illustrate the assessment of the impact of ETZ variability components on Confirmability.

Outcome	Raw Score at Baseline		Raw Score at 80 Wk		Least-Squares Mean Change at 80 Wk	
	Placebo	Solanezumab	Placebo	Solanezumab	Placebo	Solanezumab
ADCS-iADL score	45.37±8.14	45.60±7.93	39.01±11.86	39.83±11.41	-7.17±0.32	-6.17±0.32
	Sample Size at Baseline		Sample Size at 80 Wk		Sample Size for Change at 80 Wk	
	Placebo	Solanezumab	Placebo	Solanezumab	Placebo	Solanezumab
Sample size	1063	1053	896	908	980	981

Table 3: The sample sizes for both treatment arms and estimations for fixed effects in both treatment arms with standard deviation (SD) or standard error (SE). The plus–minus (±) values for the scores at baseline and at 80 weeks are means (±SD). The estimated difference is the least-squares mean (±SE) change from baseline between the two trial groups at 80 weeks. The sample size for change is estimated by the average of sample size at baseline and 80 weeks.

Assessing clinical meaningfulness of treatment effects Table 2 in Honig et al. (2018) gave $\text{SD}(Y_i^{Trt[1]})$, $\text{SD}(Y_i^{Trt[m]})$, and $\text{SE}(Y_i^{[\text{change}]})$ for seven outcome measures, with those for ADCS-iADL reproduced in Table 4 below.

	Visit 1	Visit $m = 7$
Variance	$\text{Var}\left(Y_i^{Trt[1]}\right) = 64.580$	$\text{Var}\left(Y_i^{Trt[m]}\right) = 135.389$
Standard Deviation	8.036	11.636

	Change
Variance	$\text{Var}\left(Y_i^{[\text{change}]}\right) = 92.365$
Standard Deviation	9.611

Table 4: Reported EXPEDITION3 ADCS-iADL variances. Visit 1 and visit 7 variances are computed from raw scores. Variance for Change is converted from the reported SE for LSmeans of Change using Hedges' formula.

Applying ETZ decomposition, we obtain the estimated ETZ ADCS-iADL variability components in Table 5.

Random Intercept	Measurement Error	Random Trajectory
53.802	10.778	70.809

SD(Z)	SD(E)	SD(Traj)
7.335	3.283	8.415

Table 5: Estimated ETZ ADCS-iADL EXPEDITION3 variability components.

Having an estimate of the within patient SD of measurement error is useful for patients and clinicians to interpret the estimated effect of a specific treatment. That is, when interpreting the average decrease of ADCS-iADL over an 80-week period reported in Honig et al. (2018) being -7.17 for patients given the placebo and -6.17 for patients given solanezumab, one should keep the 3.283 SD of measurement error in mind.

Having an estimate of the between patient SD of trajectory is useful for patients and clinicians to interpret the estimated *difference* of treatment effects. That is, when interpreting the average

difference of Change in ADCS-iADL between solanezumab and placebo reported in Honig et al. (2018) being 1.0, one should keep the 8.415 SD of trajectory in mind.

4.4 Confident Efficacy calculation

Denote true efficacy $\mu_{Rx}^{[m]} + \mu_{C}^{[m]} - \mu_{Rx}^{[1]} - \mu_{C}^{[1]}$ by θ_{Phase2}, and denote its point estimate from Phase 2 data by $\bar{\theta}_{Phase2}$. With SE_{pooled} being the pooled standard error of *change*,

$$\frac{\bar{\theta}_{Phase2} - \theta_{Phase2}}{SE_{pooled}} \sim t_{n_{Rx}+n_C-2},$$

where n_{Rx} and n_C are the Rx and C sample sizes.

There are two types of original outcome measures. ADAS-Cog and HbA1c are examples of lower outcome being better, while ADL is an example of higher outcome is better. We take the type of higher outcome being better as example to illstrate and the other type is similar. Our Confident Efficacy, denoted by L_{Phase2}, is the t distribution lower $1 - (d_{Phase2} + 50\%)$ confidence limit for θ_{Phase2}

$$\bar{\theta}_{Phase2} - t_{d_{Phase2}+50\%, n_{Rx}+n_C-2} \times SE_{pooled}.$$

Different Phase 2 studies independently test different compounds for different diseases. By Kolmogorov's Law of Large Numbers (LLN) for independent but not necessarily identically distributed random variables (Theorem D on page 27 of Serfling (1980)), there is confidence that the proportion of Phase 2 studies with L_{Phase2} higher than its true θ_{Phase2} is $1 - (d_{Phase2} + 50\%)$.

4.5 Confidently Bounded Quantile calculation

The point estimate of the efficacy $\bar{\theta}_{Phase3|Phase2}$ from a Phase 3 study for $\theta_{Phase3|Phase2}$, should the sponsor decide to transition to it, is assessed as follows. $\bar{\theta}_{Phase3|Phase2}$ is assumed to follow the Normal distribution with mean L_{Phase2} and variance $Var(\bar{\theta}_{Phase3|Phase2})$. L_{Phase2} is taken as the mean of θ_{Phase2}, with

$$Var(\bar{\theta}_{Phase3|Phase2}) = \sigma_{pooled}^2 \sqrt{\frac{1}{n_{Phase3,Rx}} + \frac{1}{n_{Phase3,C}}},$$

where $\sigma^2_{pooled} = Var(Y_{Rx}^{[\text{change}]r}) = Var(Y_C^{[\text{change}]r}) = Var(Traj_i^{Trt[m]}) + 2 \times Var(E)$, $n_{Phase3,Rx}$ and $n_{Phase3,C}$ are sample sizes set for Rx and C in Phase 3, respectively. Then, as illustrated in Figure 9, a histogram of potential $\bar{\theta}_{Phase3|Phase2}$ is obtained by replicating the Phase 3 study many times with the sponsor's chosen sample size $n_{Phase3,Rx}$ and $n_{Phase3,C}$, generating $\bar{\theta}_{Phase3|Phase2}$ from a Normal distribution. The Confidently Bounded Quantile $L_{Phase3|Phase2}$ is then the $1-(d_{Phase3|Phase2}+50\%)$ quantile of this distribution. Conditional on L_{Phase2} being the true efficacy θ_{Phase2}, the proportion of replicated Phase 3 studies with outcome higher than $L_{Phase3|Phase2}$ is $1-(d_{Phase3|Phase2}+50\%)$. The uncertainty in estimating $\theta_{Phase3|Phase2}$ reduces when sample sizes increase and the Confidently Bounded Quantile changes as sample sizes change.

4.6 Making decision with variabilities as they are, an example

The *Transition Decision-Making Process Plots* tab of our App calculates the confident quantile with estimated separation and variability from the feeder study "as is". We use EXPEDITION3 as a feeder study to illustrate how this facilitates the decision-making process. The outcome considered is ADCS-iADL, which is better when it is higher. Before decision is made, two requirements are set below:

- requires the probability that Phase 2 study transitions to Phase 3 study correctly is at least, for example, 95%, and
- requires the probability that Phase 3 study shows efficacy conditioned on that Phase 2 has transitioned to Phase 3 study correctly is at least, for example, 80%.

In this case, the success confidence $95\% \times 80\% = 76\%$ is guaranteed. The two requirements are achieved by **Confident Efficacy** and **Confidently Bounded Quantile**, and visualized in Figure 9.

Referring to the concept of Confident Efficacy proposed in Section 4 and the estimations from Table 3, we can calculate the Confident Efficacy as

$$L_{Phase2} = \bar{\theta}_{Phase2} - t_{1-(d_{Phase2}+50\%),n_{Rx}+n_C-2} \times SE_{combined} = 1 - t_{95\%,1959} \times 0.32 = 0.26.$$

Similarly, we construct Confidently Bounded Quantile which is *conditional* quantile $L_{Phase3|Phase2}$ for Phase 3 study. It is conditional because it constructs on the distribution of efficacy in Phase 3 conditioned on Phase 2 study including discount.

Referring to Section 4 and setting the sample size for each arm (Rx and C) in Phase 3 as 1000, the Confidently Bounded Quantile (CBQ) can be constructed as

$$\begin{aligned} L_{\text{Phase 3|Phase 2}} &= L_{\text{Phase 2}} - Z_{d_{\text{Phase 3|Phase 2}}+50\%} \times \sigma_{\text{pooled}} \sqrt{\frac{1}{n_{\text{Phase 3,Rx}}} + \frac{1}{n_{\text{Phase 3,C}}}} \\ &= 0.26 - Z_{80\%} \times \sqrt{70.809 + 2 \times 10.778} \times \sqrt{\frac{1}{1000} + \frac{1}{1000}} = -0.1. \end{aligned}$$

EXPEDITION3 indeed failed in real life in 2016. Following CBQ proposed, $L_{Phase3|Phase2}$ calculated is negative, which means that the predefined success confidence cannot be guaranteed and consists with the result of EXPEDITION3. However, if sample size for each arm increases to 2000, the CBQ becomes 0.1 (positive), which may lead to the success of EXPEDITION3.

5 How decisions change if variabilities are changed

Given data from a feeder study, the *Profile Plots* tab of our Confirmability app assesses how much each of the variability reduction strategies described in Section 1.1.1 can impact on the probability of having a successful Confirmatory study. We use real data examples to show

- $Var(Z)$ hardly impact confirmability if effects are measured as *change-from-baseline*.
- Change in $Var(E)$ affects the *shape* of trajectory.
- Change in $Var(Traj)$ affects *separation* at the milestone visit.

After Section 5.1 shows variability of Z hardly impacts Var(Change), Section 5.3 shows how our Confirmability App reacts instantaneously to "what happens to a Phase 3 study if $Var(Traj)$ or $Var(E)$ changes" questions by visualizing the impact using profile plots. A sponsor can then decide whether to invest in one or more Section 1.1.1 strategies to reduce the variability of impactful component(s), to increase the chance of *confirmability*.

5.1 Baseline variability impacts Var(Change) little

There is the thinking that narrowing the patient entry criterion makes patients more homogeneous and thereby makes clinical trials more likely to succeed. This turns out not to be the case if treatment effects are measured in terms of Change, because (6) in the ETZ decomposition shows SD(Change) does not involve patients' baseline measurements Z. We demonstrate this using information available in the FDA's review of Trulicity® for treating patients with type 2 diabetes.

Trulicity® (compound name dulaglutide) is a once-weekly injectable prescription medicine to improve blood sugar (glucose) in adults with type 2 diabetes (mellitus), as measured by control of haemoglobin A1c (HbA1c). Biologic License Application 125469 for Trulicity/dulaglutide had five studies. Patient entry criteria, SD(Basline), and SD(Change) for studies GBDA and GBDC, reported on pages 6 and 7 in FDA center for drug evaluation and research (2014b) and on pages 58 and 61 of FDA center for drug evaluation and research (2014a), are reproduced in Tables 6 and 7.

HbA1c	Dulaglutide + Metformin + Pioglitazone	Placebo	Exenatide
n	279	141	276
SD(Baseline)	1.3	1.3	1.3
SD(Change)	0.98	0.97	1.02

Table 6: GBDA sub-study: Dulaglutide + Metformin + Pioglitazone vs. Placebo and Exenatide with baseline HbA1c $\in [7, 11]$

HbA1c	Dulaglutide	Metformin
n	269	268
SD(Baseline)	0.9	0.8
SD(Change)	1.00	0.98

Table 7: GBDC sub-study: Dulaglutide vs. Metformin with baseline HbA1c $\in [6.5, 9.5]$

As can be seen, narrowing the HbA1c baseline entry criterion from [7, 11] in GBDA to [6.5,

9.5] in GBDC decreases SD(baseline) of HbA1c from 1.3 to within [0.8, 0.9], but hardly affects SD(Change), which remains steady within [0.97, 1.02] for both studies. So decreasing variability of intercept Z may not enhance confirmability much. Thus we turn our attention to how variability in trajectory and measurement error impact confirmability.

Depending on outcome measures, response profiles can appear to have different shapes. For example, from Figure 2, in terms of Change, while ADCS-iADL response profiles appear close to be linear, response profiles of ADAS-Cog seem not to be easily parameterized. However, as we will show in Figure 10 of Section 5.3, variability of measurement error E can distort the shape of response profiles, by making linear response profiles appear non-linear for example. To understand how variability of the trajectory and measurement error separately impact observed separation between Rx and C at the end of the study and shape of the response profiles, we connect our ETZ modeling principle with the so-called random coefficients model.

5.2 Assessing the replicability of a Confirmatory study

Calculation under the *Transition Decision-Making Process Plots* tab does not assume any functional shape for response profiles so that, if the calculated quantile exceeds the desired threshold, one can be confident it is so without such assumptions. However, if the calculated quantile fails to exceed the desired threshold, then one can turn to the *Profile Plots* tab of our App to see whether investing in some strategies in Section 1.1.1 might make the Confirmatory study more likely to succeed.

The *Profile Plots* tab allows the user to change each of the ETZ variability components, and assess *replicability* of a Confirmatory study by observing how stable or unstable the response profiles are upon replications of the study, proceeding only if they are stable. Simulated replications of a Confirmatory study under the *Profile Plots* tab assumes the response profiles have idealized functional forms (a confirmatory study not replicable under an ideal situation will not be replicable under non-ideal situations). For our proof-of-concept App, at present calculations under the *Profile Plots* tab are for straight line response profiles, leaving development of more sophisticated profiles for future research.

Corresponding to the RANDOM statement in Proc Mixed of SAS, a random coefficients model considers Time as a continuous variable, with the response of each subject being generated randomly from a distribution of response trajectories as a function of Time (e.g., random intercept and random slope with a bivariate Normal distribution). Expectations of the response trajectories are the response profiles, reflecting the average effects of treatments Rx and C over Time.

In contrast to MMRM, but similar to the ETZ modeling principle, a random coefficients model has, in addition to randomness in the intercepts and the response trajectories, a residual term in the model which is analogous to our measurement error term. With treatments indexed by $Trt = Rx$ or C, a random coefficients model for linear response profiles would be

$$Y_i^{Trt[v]} = \alpha^{Trt} + a_i^{Trt} + \beta^{Trt} x_i^{[v]} + b_i^{Trt} x_i^{[v]} + e_i^{Trt[v]}, \tag{20}$$

where $x_i^{[v]}$ is the Time at which the v^{th} measurement on the i^{th} subject is taken. The fixed effects are represented by $\alpha^{Trt} + \beta^{Trt} x_i^{[m]}$. What distinguishes the ETZ modeling principle from this (general) random coefficient model (20) is $\alpha^{Rx} = \alpha^{C}$ in (4) because at Time = 0 ($x_i^{[1]} \equiv 0$) patients have yet to be treated.

Random subject effect for the i^{th} patient assigned to treatment Trt is $(a_i^{Trt}, b_i^{Trt})'$, with

$$\begin{bmatrix} a_i^{Trt} \\ b_i^{Trt} \end{bmatrix} \sim i.i.d.\ MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix} \right).$$

The residuals $e_i^{Trt[v]}$ are assumed to be i.i.d. $N(0, \sigma^2)$, and independent of $(a_i^{Trt}, b_i^{Trt})'$. Thus, with $t^{[m]}$ denoting time of the milestone visit (visit m),

$$Var(Y_i^{Trt[1]} | x_i^{[1]} = 0) = \sigma_a^2 + \sigma^2 \tag{21}$$

$$Var(Y_i^{Trt[m]} | x_i^{[m]} = t^{[m]}) = \sigma_a^2 + 2t^{[m]}\sigma_{ab} + {t^{[m]}}^2 \sigma_b^2 + \sigma^2 \tag{22}$$

$$Var(Y_i^{Trt[\text{change}]}) = {t^{[m]}}^2 \sigma_b^2 + 2\sigma^2. \tag{23}$$

Randomness of the intercept, response trajectories, and residuals, together determine the $m \times m$ variance-covariance matrix of outcome measures on the subjects across the m visits.

For a linear random coefficients model, our assumption of independence between Z and $Traj$ would be $\sigma_{ab} = 0$. In practice, a user can assess closeness of $\sigma_{ab} = 0$ to zero in a variety of ways to feel comfortable proceeding with this assumption. Matching with (7), (9), and (6) then, we see that in ETZ notations

$$\begin{bmatrix} a_i^{Trt} \\ b_i^{Trt} \end{bmatrix} \sim i.i.d. \quad MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} Var(Z) & 0 \\ 0 & Var(Traj)/{t^{[m]}}^2 \end{bmatrix} \right) \tag{24}$$

$$e_i^{Trt[v]} \sim i.i.d. \quad N(0, Var(E_i^{[m]})). \tag{25}$$

Note that, while intercepts are in units of Y measurements (e.g., ADAS-Cog), with Time being a continuous variable in the random coefficient setting, slopes β^{Trt} and b_i are in units of Y measurements divided by units of time (e.g., ADAS-Cog per week, if Time is measured in weeks).

For illustration purpose, we pretend EXPEDITION3 is the feeder study to itself, to see to what extent a similar sample size (2000+ patients) feeder study might replicate/confirm its own finding in ADCS-iADL. Assuming true response profiles are linear, random coefficient model (20) is used for simulation. (Fixed) treatment effects are estimated from visit 1 (week 0) and visit 7 (week 80) point estimates reported in Honig et al. (2018), with

$$\alpha^{Rx} = 45.6, \quad \beta^{Rx} = -\frac{6.17}{80}, \qquad \alpha^{C} = 45.37, \quad \beta^{C} = -\frac{7.17}{80}.$$

Then, assuming the two treatment arms share a same covariance matrix, each simulated study uses a different random number seed to generate data with initial sample sizes same as in EXPEDITION3 and no patient dropout, with

$$\begin{bmatrix} a^{Trt} \\ b^{Trt} \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 53.802 & 0 \\ 0 & 70.809/t_m^2 \end{bmatrix} \right),$$

and i.i.d. measurement error $e^{Trt[v]} \sim N(0, 10.778)$, where $Trt = Rx$ or C.

5.3 Impact of Var($Traj$) and Var(E) on a Phase 3 study

The *Profile Plots* tab of our App shows interactively what happens to a Phase 3 study if any of the three ETZ variability components changes. Taking EXPEDITION3 as an example feeder study,

Figure 10 gives examples of changing profile plots in a single Phase 3 study if $Var(E)$ or $Var(Traj)$ changes. We observe from Figures (4L) and (4H) that reducing $Var(E)$ alone is insufficient to ensure stability of separation at the milestone. For EXPEDITION3, we observed from Figures (5L) and (5H) that reducing $Var(Traj)$ likely will have a bigger impact in enhancing Replicability of the Confirmatory study.

Recommended strategies For targeted therapies, a potential strategy to reduce $Var(Traj)$ is *Subgroup Targeting*, enrolling only patients with sufficient drug target expression. For Alzheimer's disease medicine whose mechanism of action is to reduce amyloid beta plaques, a sponsor may consider enrolling only patients with substantial amyloid beta deposition.

Using a *Surrogate Endpoint* to measure outcome is a strategy that potentially reduces both $Var(Traj)$ and $Var(E)$, but its benefit requires solid biomedical and statistical justifications. For example, in using HbA1c measurement as a surrogate to microvascular events in T2DM, there is epidemiological evidence that controlling HbA1c reduces microvascular complications associated with diabetes mellitus, and its $Var(E)$ is small (around 0.03).

6 Scope, limitation, key messages, and future research

The ETZ modeling principle is not only applicable to the major disease areas in Table 1, but also to rare disease areas. Of the some 7000 rare diseases, fewer than 10% have approved treatments. The challenge is large variability of functional outcome measures coupled with a small number of patients with rare diseases makes the identification of reliable clinical outcome assessments (COAs) for efficacy endpoints difficult. ETZ's ability to suggest judicious allocation of resources to reduce sources of variability should be useful for COA research.

Scope of application Applicable to studies with Before-and-after treatment Repeated Measurements, the ETZ modeling principle does not assume a particular shape for the response profiles.

Limitation However, the assessment of replicability of estimated response profiles does assume a

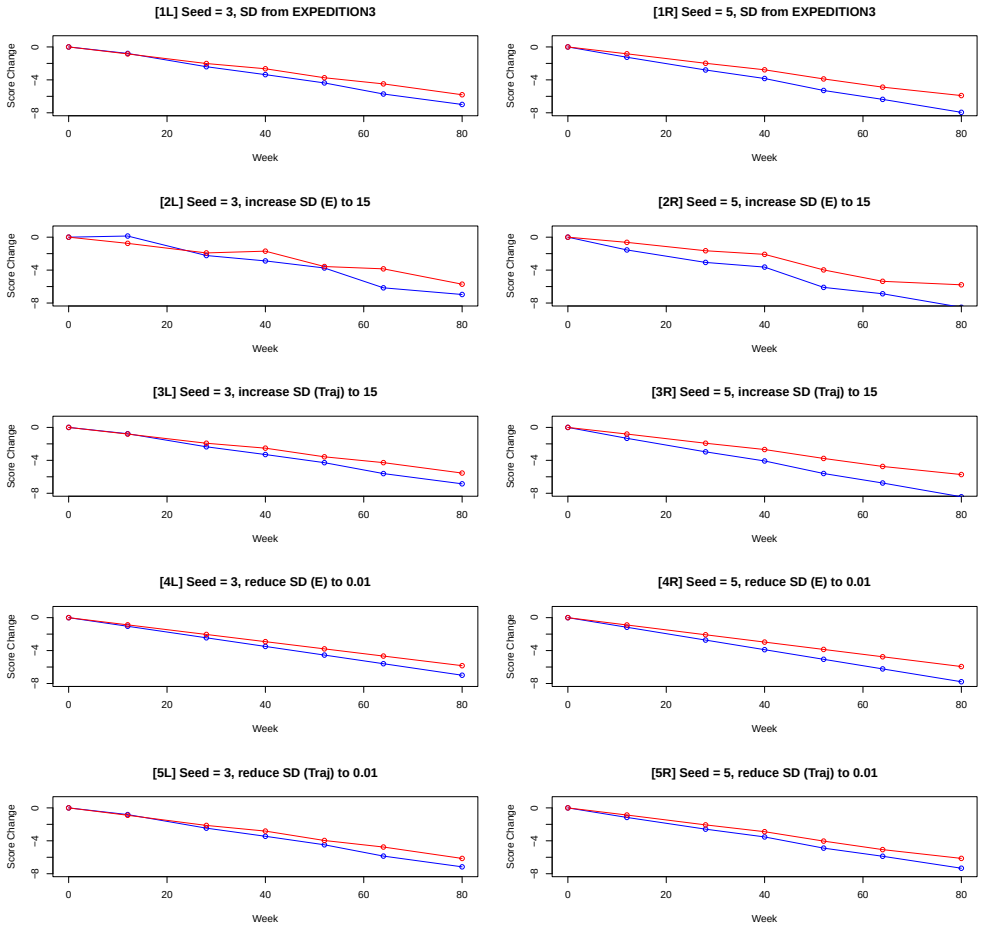

Figure 10: Changing the random number Seed (from 3 in the left column to 5 in the right column as an example) triggers the *Profile Plots* tab of our Confirmability App to simulate an independent replication of the Confirmatory study. With means and ETZ variability components as estimated for EXPEDITION 3 in Tables 3 and 5, [1L] & [1R] show separation at Week 80 lacks stability. Increasing $Var(E)$ and $Var(Traj)$ each separately while keeping the other as in Table 5, [2L] & [2R] show large $SD(E)$ distorts the shape of the profiles, while [3L] & [3R] show large $SD(Traj)$ causes separation at the milestone to be unstable. With $SD(Traj) = 8.415$ still as estimated for EXPEDITION 3, separation remains unstable even with $Var(E)$ reduced to 0.1, as [4L] & [4R] show. With $SD(E) = 3.283$ as estimated for EXPEDITION 3, reducing $SD(Traj)$ to 0.1 does make the Rx and C separation at week 80 less variable, as [5L] & [5R] show.

shape for the response profiles (linear profiles in the current implementation).

Key messages

- Efficacy measured in terms of *change* (from baseline) is largely unaffected by range of the patient entry criterion.
- Large *measurement error* variability will (even with large sample sizes) cause separation between Rx and C at the milestone visit to be unstable; and linear response profiles to appear non-linear.
- Large *trajectory* variability will cause separation between Rx and C at the milestone visit to not be replicable (even with large sample sizes).

7 Interest statement

Y.S., Y.H., X.W., S-Y.T. declare no competing interests. Y. L. is an employee and shareholder of Eli Lilly. J.C.H. is a professor emeritus of the Ohio State University and a consultant to Eli Lilly.

References

Arrowsmith, J. (2011a). Phase II failures: 2008–2010. *Nature Reviews Drug Discovery 10*, 328–329.

Arrowsmith, J. (2011b). Phase III and submission failures: 2007–2010. *Nature Reviews Drug Discovery 10*, 87.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis* (Second ed.). New York: Springer-Verlag.

Casella, G. and R. Berger (2001). *Statistical Inference*. Duxbury Press.

FDA center for drug evaluation and research (2014a). Trulicity (dulaglutide) statistical review. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2014/125469Orig1s000StatR.pdf. Accessed: 2024-06-14.

FDA center for drug evaluation and research (2014b). Trulicity (dulaglutide) summary review. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2014/125469Orig1s000SumR.pdf. Accessed: 2024-06-14.

FDA-NIH Biomarker Working Group (2016). Best (biomarkers, endpoints, and other tools) resource. https://www.ncbi.nlm.nih.gov/books/NBK326791/toc/. Accessed: 2024-06-14.

Hay, M., D. Thomas, and Craighead, J. et al (2014). Clinical development success rates for investigational drugs. *Nature Biotechnology 32*, 40–51.

Hayter, A. J. and J. C. Hsu (1994). On the relationship between stepwise decision procedures and confidence sets. *Journal of the American Statistical Association 89*, 128–136.

Honig, L. S., B. Vellas, M. Woodward, M. Boada, R. Bullock, M. Borrie, K. Hager, N. Andreasen, E. Scarpini, H. Liu-Seifert, M. Case, R. A. Dean, A. Hake, K. Sundell, V. Poole Hoffmann, C. Carlson, R. Khanna, M. Mintun, R. B. DeMattos, K. J. Selzler, and E. R. Siemers (2018). Trial of solanezumab for mild dementia due to alzheimer's disease. *New England Journal of Medicine 378*(4), 321–330.

Hsu, J. C. (1989). Sample size computation for designing multiple comparison experiments. *Computational Statistics and Data Analysis 7*, 79–91.

Hsu, J. C. (1992). The factor analytic approach to simultaneous inference in the general linear model. *Journal of Computational and Graphical Statistics 1*, 151–168.

Hsu, J. C. (1996). *Multiple Comparisons: Theory and Methods*. London: Chapman & Hall.

Hsu, J. C. and R. L. Berger (1999). Stepwise confidence intervals without multiplicity adjustment for dose response and toxicity studies. *Journal of the American Statistical Association 94*, 468–482.

Hsu, J. C., J. G. Hwang, H. Liu, and S. J. Ruberg (1994). Confidence intervals associated with bioequivalence trials. *Biometrika 81*, 103–114.

Kiefer, J. (1977). Conditional confidence statements and confidence estimators. *Journal of the American Statistical Association 72*(360a), 789–808.

Liu, Y., H. Tian, and J. C. Hsu (2021). *Handbook of Multiple Comparisons*, Chapter 13: Subgroups Analysis for Personalized and Precision Medicine Development, pp. 311–338. Chapman and Hall.

National Academies of Sciences, Engineering, and Medicine (2019). *Reproducibility and Replicability in Science*. Washington, DC: The National Academies Press.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences 236*(767), 333–380.

Scheffé, H. (1959). *Analysis of Variance*. New York: John Wiley.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: John Wiley and Sons.

Siddiqui, O., H. M. J. Hung, and R. O'Neill (2009). MMRM vs. LOCF: A comprehensive comparison based on simulation study and 25 NDA datasets. *Journal of Biopharmaceutical Statistics 19*(2), 227–246. PMID: 19212876.

Stefansson, G., W. Kim, and J. C. Hsu (1988). On confidence sets in multiple comparisons. In S. S. Gupta and J. O. Berger (Eds.), *Statistical Decision Theory and Related Topics IV*, Volume 2, pp. 89–104. New York: Springer-Verlag.

Takeuchi, K. (1973). *Studies in Some Aspects of Theoretical Foundations of Statistical Data Analysis (in Japanese)*. Tokyo: Toyo Keizai Shinposha.

Takeuchi, K. (2010). Basic ideas and concepts for multiple comparison procedures. *Biometrical Journal 52*(6), 722–734.

Tukey, J. W. (1953). The Problem of Multiple Comparisons. Dittoed manuscript of 396 pages, Department of Statistics, Princeton University.

Tukey, J. W. (1994a). The problem of multiple comparisons. In H. I. Braun (Ed.), *The Collected Works of John W. Tukey*, Volume VIII Part 1, Chapter 18: Designing for Accuracy, pp. 156–158. New York and London: Chapman & Hall.

Tukey, J. W. (1994b). The problem of multiple comparisons. In H. I. Braun (Ed.), *The Collected Works of John W. Tukey*, Volume VIII Part 1, pp. 1–300. New York and London: Chapman & Hall.

Wang, X., Y. Han, Y. Liu, S.-Y. Tang, and J. C. Hsu (2025). Counterfactual uncertainty quantification of factual estimand of efficacy from before-and-after treatment repeated measures randomized controlled trials. *Statistical Analysis and Data Mining: An ASA Data Science Journal 18*(4), e70039.

Wessels, A. M., E. R. Siemers, P. Yu, S. W. Andersen, K. C. Holdridge, J. R. Sims, K. L. Sundell, Y. Stern, D. M. Rentz, B. Dubois, R. W. Jones, J. L. Cummings, and P. S. Aisen (2015). A combined measure of cognition and function for clinical trials: The Integrated Alzheimer's Disease Rating Scale (iADRS). *Journal of Prevention of Alzheimer's Disease 2(4)*, 227–241.