Continuous Composite Endpoints: How Bad Is Too Bad?

James Bell – Elderbrook Solutions GmbH

On Behalf of the
Estimation Subteam
Estimands Implementation Working Group (EIWG)

PSI Conference, 10th June 2025

Continuous Composites

- **Coauthors; the Estimation Subteam of the EIWG:**
  - James Bell (Elderbrook Solutions)
  - Christian Bressen Pipper (Novo Nordisk)
  - Thomas Drury (GSK)
  - Lorenzo Guizzaro (EMA)
  - Oliver Keene (KeeneONStatistics)
  - Marian Mitroiu (Biogen)
  - Tobias Muetze (Novartis)
  - Khadija Rantell (MHRA)
  - Marcel Wolbers (Roche)
  - David Wright (AstraZeneca)

Disclaimer: The talk is based on work by the subteam. It reflects our personal opinions and may not reflect the views of our respective organisations. The example discussed is illustrative only and is not a real regulatory interaction.

Continuous Composites Introduction

- ICH E9(R1) introduced five strategies for addressing intercurrent events (ICEs)

- This talk will focus on the **composite (variable) strategy**
  - The ICE is incorporated as the patient outcome

- For binary, time-to-event and categorical outcomes, composite strategies are straightforward
  - The ICE becomes an event (typically a bad outcome)

- **For continuous outcomes the composite strategy poses fundamental difficulties**
  - How do you combine an occurrence with a number?

- This talk will look at problems with definitions and estimation of a mean
  - We will also attempt to provide some potential solutions

Setting the Scene

Continuous Composites Regulatory Conundrum

- A sponsor is looking to run a **Phase III pivotal trial for treatment of a chronic lung condition**
- Agreed endpoint: Change from Baseline in Forced Vital Capacity (FVC) at 52 weeks
  - Continuous measurement of lung capacity
- All ICEs are to be handled by treatment policy, except for death
  - 5% death expected; to be handled hypothetically
- The sponsor drafts a protocol and has an interaction with a Regulator Agency
- The Regulator Agency instead recommends a **composite strategy for death** where the defined value of (absolute) FVC is **0**.
- Seems like a reasonable request?

Continuous Composites Respiratory Example

- The Sponsor’s original (hypothetical) sample size calculation assumed the following:
  - Baseline mean of 2300 mL, with SD 500 mL
  - -180 mL change in control arm, -100 mL change in active arm
  - Fully-adjusted SD of 250 mL for the CfB
  - 5 % mortality; no other missing data

- **Sample size: 218 patients / arm, 90% power**

Continuous Composites Respiratory Example

- Just to check the Regulator Agency’s request, the Sponsor simulated the composite sample size
  - 0 mL (original scale) inserted for all dead patients; independent deaths

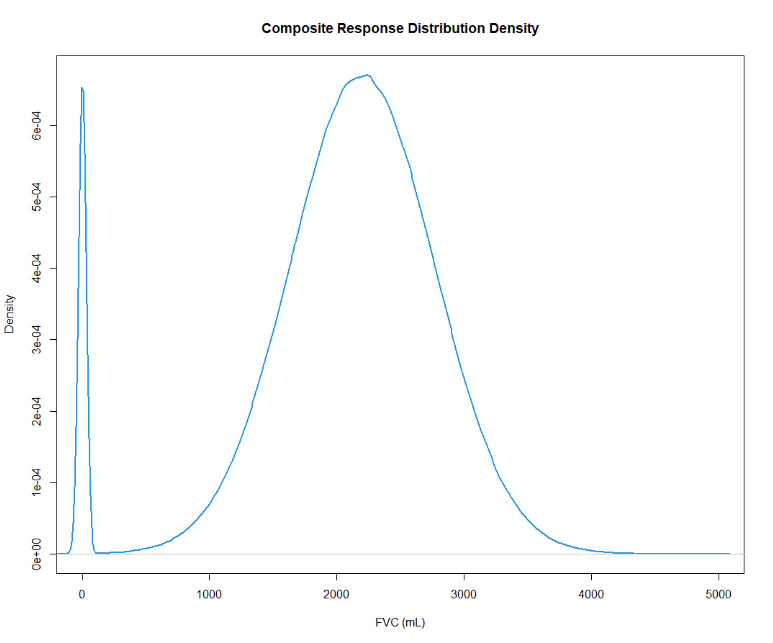
Continuous Composites Respiratory Example

- Just to check the Regulator Agency's request, the Sponsor simulated the composite sample size
  - 0 mL (original scale) inserted for all dead patients; independent deaths

- **Power: 29%**

- Variance of the estimator inflated 4.6-fold;
- New sample size, approx. 1000 patients / arm

- The Sponsor is now in a difficult situation



Failure Values

Continuous Composites

Continuous Composites

- This is an example of a ‘**Continuous Composite Endpoint**’
  - A continuous endpoint where one or more ICEs are handled by a composite strategy
  - A fixed continuous value is assigned for all patients with the relevant ICE(s)

- We will refer to the continuous value defined for ICEs as the ‘**failure value**’

- Failure value is **part of the estimand definition**
  - Not a ‘missing value to be imputed”
  - Does not (necessarily) reflect what would have happened to the patient - not hypothetical!
  - Must be defined independently of treatment

Continuous Composites

Impact on Means

- Typically, we are interested in mean differences for continuous endpoints
- However, the **mean is highly sensitive to outliers**…
  - … such as many ‘obvious’ failure values (e.g. ‘0’) for endpoints
- For extreme failure values, continuous endpoints become **pseudo-binary**:
  - Either failure or not; sample variance becomes increasingly irrelevant
- Inappropriate choice of failure value can cause **extreme variance inflation**

Continuous Composites Inflation Formula

- **We can predict variance inflation by considering mixture distributions:**
  - Component 1: A continuous distribution with variance σ^2 and proportion $(1-w)$
  - Component 2: Point mass with variance 0 at $\Delta.\sigma$ from mean of component 1, in proportion w
  - Forms a composite distribution with variance σ_c^2

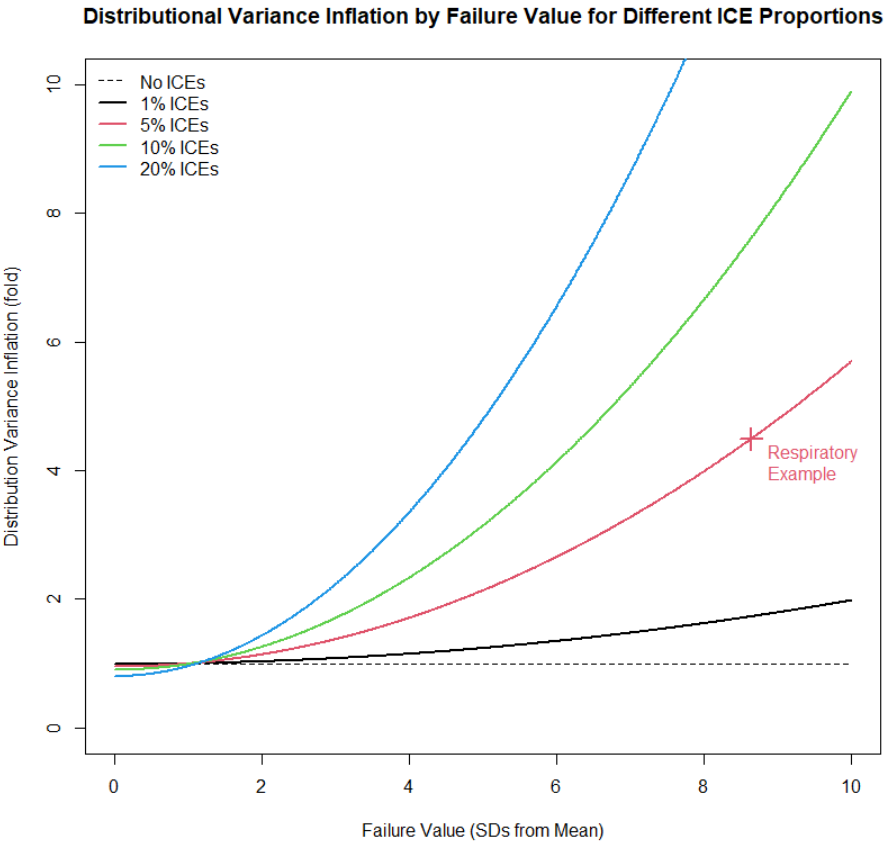
- From the mixture distribution variance formula, the variance inflation relative to component 1 is:

$$\varphi = \frac{\sigma_c^2}{\sigma^2} = (1-w)(1+w\cdot\Delta^2)$$

- Inflation of the distribution variance directly feeds into variance of estimation
  - Additional complexities such as extra data, multiple arms, correlated outcomes etc
  - Distribution variance is main driver of observed variance inflation

Continuous Composites

Inflation Formula



- Distributional variance inflation generally very high where failure value outside bulk of distribution (> 2-3 SDs)
- Less extreme failure values produce less variance inflation
- Distributional variance shrinks if failure value < 1.1 SDs approx.
- Values around 2 SDs from mean seem reasonable
  - Penalising, yet not too variance inflating

Choosing Failure Values

Continuous Composites Defining Failure

- Failure values **penalise** an arm for each ICE occurrence
  - E.g. “each death should be penalised by defining the CfB FVC as -1000 mL”

- Ideally, value should be **justifiable**
  - Rooted in external data source?
  - Clinical relevance?
  - May be difficult to explain why it is e.g. -16 instead of -17 or -15 though…

- In practice, failure values risk being seen as somewhat arbitrary and subjective

- Ideally, failure values should be consistent per endpoint across trials
  - Or estimands and estimates may not be comparable
  - May need indication standards

Continuous Composites Defining Failure

- Can justify as **clinical representation of failure on the continuous scale**
  - Works best for boundary values that are clinically reasonable and populated
    - Only need to justify not choosing a higher value
    - Populated values will not be too extreme outliers
  - E.g. death defined as walking 0 metres in a 6-minute walk test

- **But usually difficult to assign a sensible value for death**, e.g.
  - Lung capacity:
    - 0 mL not realistic for a living patient!
  - Weight:
    - 0 kg weight not remotely feasible for a living patient
      - ..and weight loss is desirable in many situations
    - ‘Weight gain’ value would be arbitrary and clinically unreasonable
    - ‘0 kg change’ would not be that bad an outcome for placebo

Continuous Composites Distributional Definitions

- Can we define failure values **relatively**?
  - "ICE implies patient should perform worse than (almost) all others"

- **Distributional definition**

- Use e.g. quantiles or standardised penalties

- Existing precedent for this concept:
  - Clinical relevance of treatment effects based on e.g. Cohen's d

Continuous Composites Distributional Definitions

- We note two of the estimand principles from Mütze et al:*

- **Estimands should be defined at the population level**

  → The definition must be based on the population, not sample, distribution

  - Note; this rules out "worse observed outcome"
- **Estimands should be causal**

  → The same defined value should apply to all arms

- ICH E9(R1) also emphasizes that estimands should be **clinically relevant**

  → Control treatment should be most relevant and transferable

*Pharm Stat 2025 Jan-Feb;24(1):e2432. doi: 10.1002/pst.2432

Continuous Composites Distributional Definitions

- We therefore make two proposals:

  1. For both arms, the failure value is defined as the value that is the ***Xth centile*** of the distribution of the population taking the control arm treatment

  2. For both arms, the failure value is defined as the value that is ***Y standard deviations below the mean*** of the distribution of the population taking the control arm treatment

- They:
  - Provide some interpretability, clinical relevance
  - May be transferable between trials / endpoints / indications etc
  - Are 'well-defined' estimands
  - Can control variance inflation

Continuous Composites **Uncertainty**

- Distributional definitions code for specific, defined, values (not imputations!)
  - They exist (e.g. ‘22’) but are **unknown population parameters**

- Definitions can be estimated from ‘current’ trial or historical data

- Need to introduce a **small variance correction**
  - Reflects uncertainty in definition, i.e. parameters only
  - E.g. we estimate definition as ‘23’, but true definition is ‘22’

- Could be implemented through e.g.
  - Multiply imputing parameter values
  - Resampling (e.g. bootstrapping)
  - Possibly analytically?

Continuous Composites Alternative Summaries

- **Mean requires values (observed or otherwise) for every patient**
  - Highly sensitive to outliers
  - Other summary measures don’t need specific values & often insensitive to outliers

- E.g. **Median** and **pseudomedian** (Hodges-Lehmann) work up to ‘break points’ of ICE quantities:
  - Median – 50% ICEs
  - Pseudomedian – 29% ICEs

- **Rank-based** or **pairwise comparison procedures** also possible (e.g. Win Odds)

- Main drawbacks to moving away from the mean:
  - Alternative summary measures **may not be of interest**
    - But may be useful for statistical testing
  - Often **loss of power** vs mean with mild failure value

Discussion & Conclusions

Continuous Composites Discussion

- **Discuss with clinicians & regulators** strategies for defining failure values
  - Try to identify values that are both meaningful and statistically workable

- **ICE-handling driving large variance inflation is inappropriate**
  - Unnecessarily, massively inflating sample size requirements is unethical and unfeasible…
  - Treatments may have larger effects but less significance than under e.g. hypothetical

- **Failure values beyond '3 SDs from mean' are probably unworkable**
  - '2 SDs from mean' may represent a balance between penalising and estimable
    - Aligns with typical acceptance that 95% interval represents 'bulk' of distribution

- **Patient-level uncertainty in post-ICE values is incompatible with a composite strategy**

Continuous Composites

Discussion

- **Often no clinically meaningful value is workable**
  - Calls into question composite-mean approach
  - Likely need to change something (e.g. summary measure)

- **Should mortality be incorporated into continuous outcome estimation?**
  - May not be necessary, relevant
  - Often too rare to be informative
  - Mortality imbalances already separately assessed

- **Mortality often incidental to clinical question of interest**
  - e.g. vision quality, weight loss

  → Hypothetical with MAR may be appropriate estimand & analysis

Continuous Composites Summary

- **Continuous composite endpoints integrate a continuous outcome with a defined failure value**

- **Extreme failure values will cause large, predictable, variance inflation**
  - Potentially ruinous for sample size and/or power

- **Failure values must be chosen with great care and justified**
  - Aim for clinically relevant and 'observable' value

- **Distributional definitions may help where no obvious failure value**
  - Quantile or 'SDs from mean' of population control distribution

- **Alternatively consider other summary measures or ICE strategies**

EIWG estimation subteam is working on a paper on this topic

Acknowledgements

- **Boehringer-Ingelheim Pharma GmbH & Co. KG** for sponsoring James’ work with the EIWG

Thank you for your attention!

Backup Slides

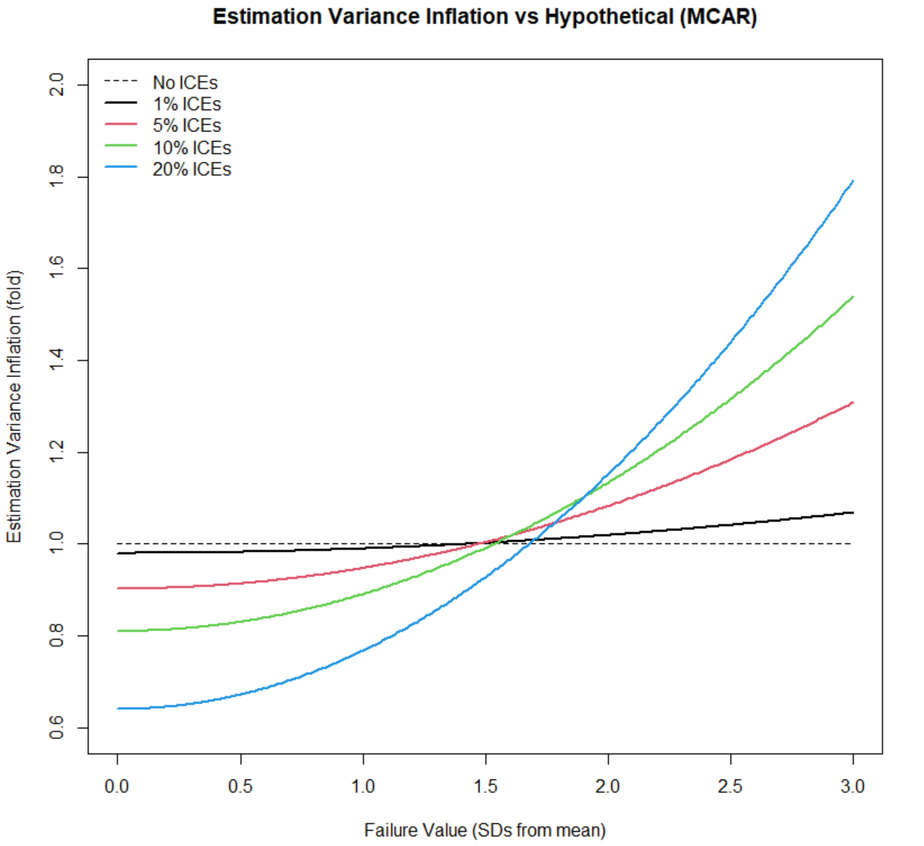
Continuous Composites Defining Failure

- Alternatively, could consider '**point of death**', or '**asymptotic**' values as death is approached
  - E.g. for FVC, define as a typical lung capacity of dying, rather than dead, patients

- Composite values for rescue could be defined by '**expected bad outcome**' had rescue not occurred
  - Gets conceptually close to MI for hypothetical 'value if rescue not administered'
    - Difference is that it would be 'fixed' across all patients, rather than varying per patient

Continuous Composites

Inflation Formula



- Sample size affected by other factors too
- Hypothetical is usual alternative strategy to composite for death
- Composite has extra ‘data’, offsetting inflation by additional factor of $(1 - w)$
  - Assumes MCAR
- Other (minor) complexities too
  - Affect exact values, not overall message
- Relative variance is low (< 1.2-fold) up to failure values of ≈ 2 SD from mean