All PICOs Great and Small

Dealing with small subpopulations in the EU HTA landscape

Patient population
Intervention
Comparator(s)
Outcomes

15th June 2026

Dave Gelb (MSD, Zurich, Switzerland), Joel Baldwin (UCB - Cogent Skills Ltd, UK), Ehsan Masoudi (Novartis, Basel, Switzerland), Milana Ivkovic (Novo Nordisk A/S, Bagsvard, Denmark), Lovisa Berggren (UCB, Monheim, Germany), Larry Leon (Merck & Co., Inc., Rahway, NJ, USA)

MSD

Agenda

1. Background and Context
2. Related guidance and literature
3. Problem statement
4. Proposed Solution
5. Discussion and conclusions

The views expressed in this presentation are those of the co-authors and do not necessarily reflect the views of their affiliated organizations

MSD

EU HTA Joint Clinical Assessment

JCA: The **mandatory** requirement of **centralised clinical assessment** for patient access of new health technologies for Member States of the EU

The assessment scope of JCA should include all relevant parameters in terms of the PICO framework:

- **P**atient population
- **I**ntervention
- **C**omparator(s)
- **O**utcomes

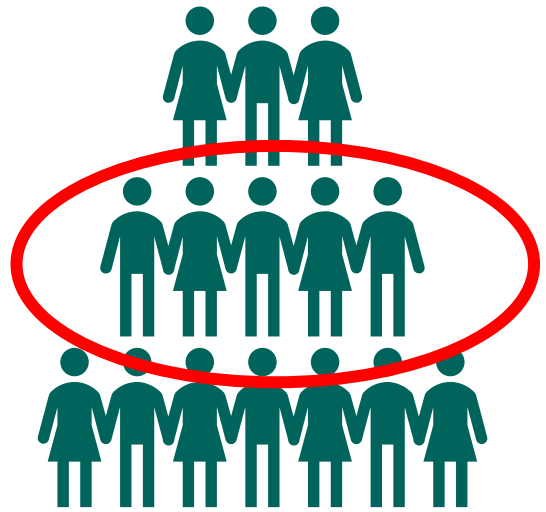
PICO selection is policy-driven, not evidence-driven.

EU Member States determine their PICO needs, which is followed by a consolidation of PICOs

EU, European Union; HTA, Health Technology Assessment; JCA, Joint Clinical Assessment; PICO, Population, Intervention, Comparator(s), Outcomes.

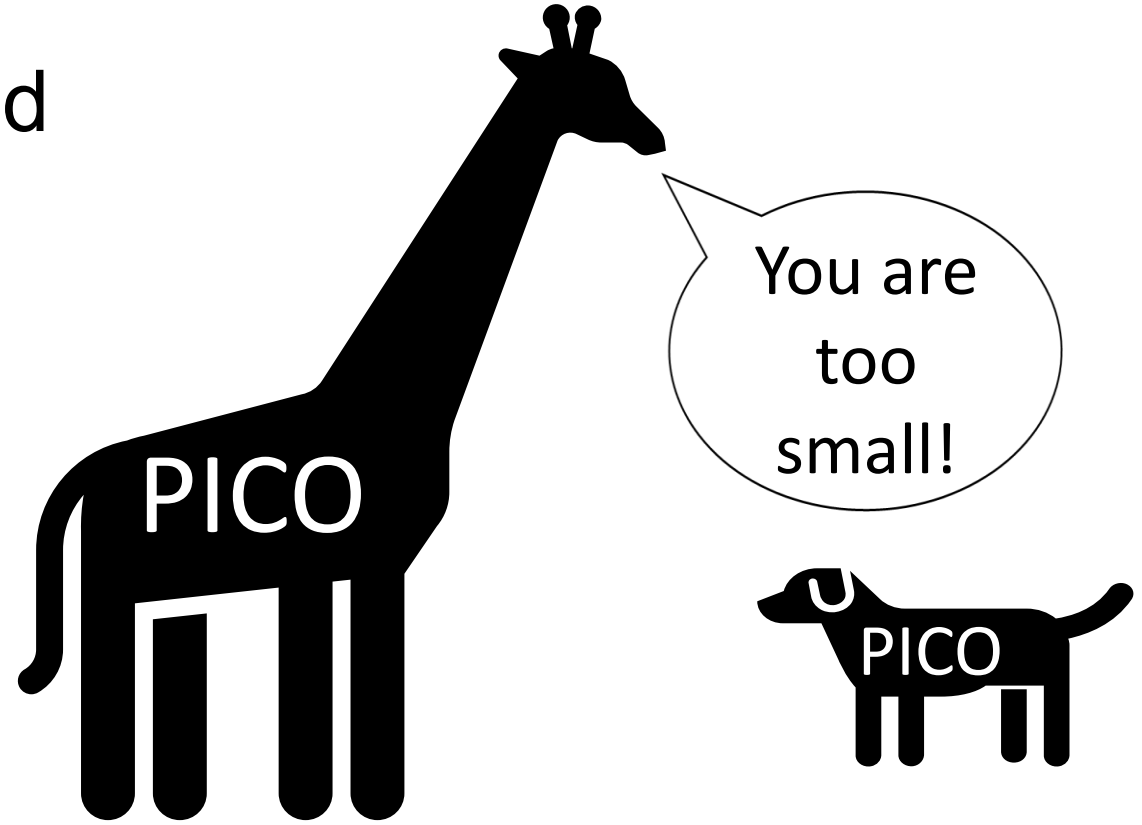
Subpopulations – the ‘P’ of the PICO

- **Subpopulations** = Target reimbursement population(s)
  - These may be defined by baseline characteristics, disease characteristics, biomarkers, treatment patterns,…
- Sponsors are required to provide full analysis & reporting package for efficacy, PROs, safety
- Conceptually subpopulations ≠ subgroups
  - *May need to provide subgroups within sub-populations*



Description of the issue

- How to address 'small' subpopulations?
- Can we “just say not feasible” and not analyse?
  - Statistical grounds to challenge?
  - What happens if we don't say no?
- A consistent approach?
  - Is this realistic within and across organisations?



MSD

The Question restated - statistical and clinical motivation

STATISTICAL OBJECTIVE

When a treatment has a perfectly uniform benefit across all patients, how often do standard subpopulation analyses produce extreme-looking results by chance alone?

WHY IT MATTERS

Stakeholders see Forest plots with apparently alarming subpopulation hazard ratios or upper confidence bounds. Are these real signals — or sampling noise from small denominators?

Answer: extreme results are routine in small subpopulations — clinical rationale provides no statistical protection.

MSD

Little literature or guidance on small subpopulations

Guidelines – only for large not small subpopulations

- Per IQWiG assessment practice, if ≥80% of the study population can be assigned to the appropriate comparator therapy-relevant subpopulation, results from the total study population may be used.
- www.igwig.de

Exploratory Literature Review (AI-Supported)

A simulation study:

- "Small samples produced larger average error, even with complete follow-up, than large samples with short follow-up ….. Correctly specifying the event distribution reduced magnitude of error in larger samples but not in smaller samples"
- "Uncertainty may not be sufficiently captured within estimated confidence intervals when extrapolating limited clinical data for use in decision models, suggesting that probabilistic analysis is not sufficient to overcome the limitations of small samples or of short follow-up in large samples."

Beca, J., Perampaladas, K., Goeree, R., & Xie, F. (2021). Impact of limited sample size and follow-up on single event survival extrapolation for health technology assessment: a simulation study. BMC Medical Research Methodology, 21, 282.

The Simulation Setup

Real-data Data Generating Mechanism with a perfectly uniform treatment effect

1. **Source data**
   GBSG breast cancer trial: N = 686 patients, 299 events

2. **Weibull AFT model**
   Fit to the source data; build a super-population (N = 5,000)

3. **Uniform treatment effect**
   True hazard ratio = 0.70 for every patient — no heterogeneity, by construction

4. **Censoring calibration**
   uniroot-based search matches simulated censoring to the source data

5. **10,000 synthetic trials**
   Each trial: 686 patients × **56 pre-defined subpopulations** × Cox stratified by grade

Filtering on subpopulations most prone to extreme upper bounds

FILTER RULE

Pr(Upper Bound (HR) >= 2.0) >= 10% across 10,000 simulated trials

All Patients (ITT) included for comparison regardless of whether it meets the filter.

Two views, one set of subpopulations

UB(HR) panel asks: how alarming can the upper confidence bound look by chance?

HR panel asks: how misleading can the point estimate itself be?

11 subpops + ITT comparator

Random benchmarks ('random15' / 20 / 40 / 60) lead the list — pure sampling noise.

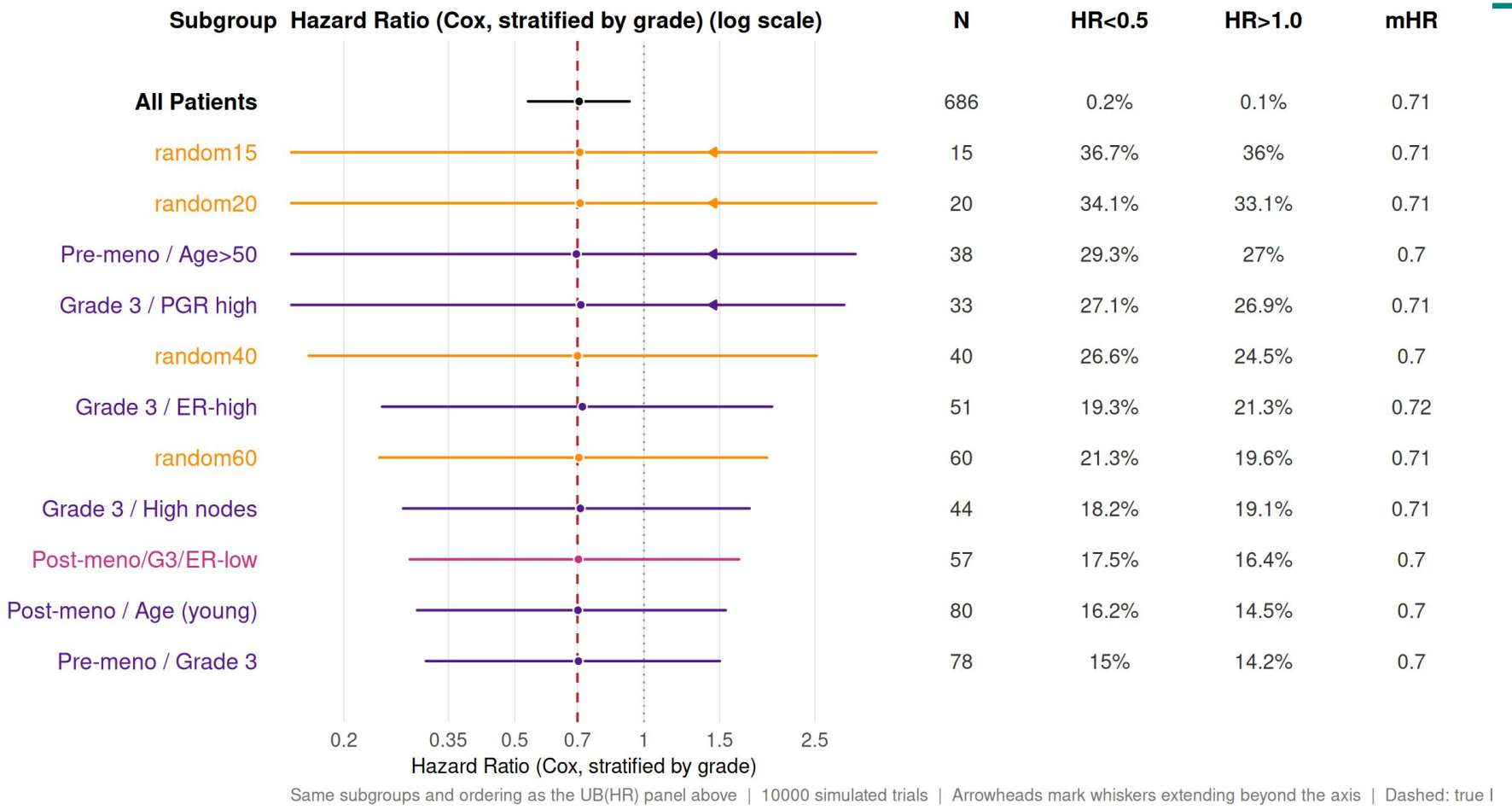
Clinical and biomarker-defined subpops of similar size are statistically indistinguishable from them.

HR = Hazard Ratio, UB = Upper Bound of 95 % Confidence Interval

MSD

Hazard Ratio Distribution

Same subpopulations, point-estimate view: how misleading can the headline HR be?

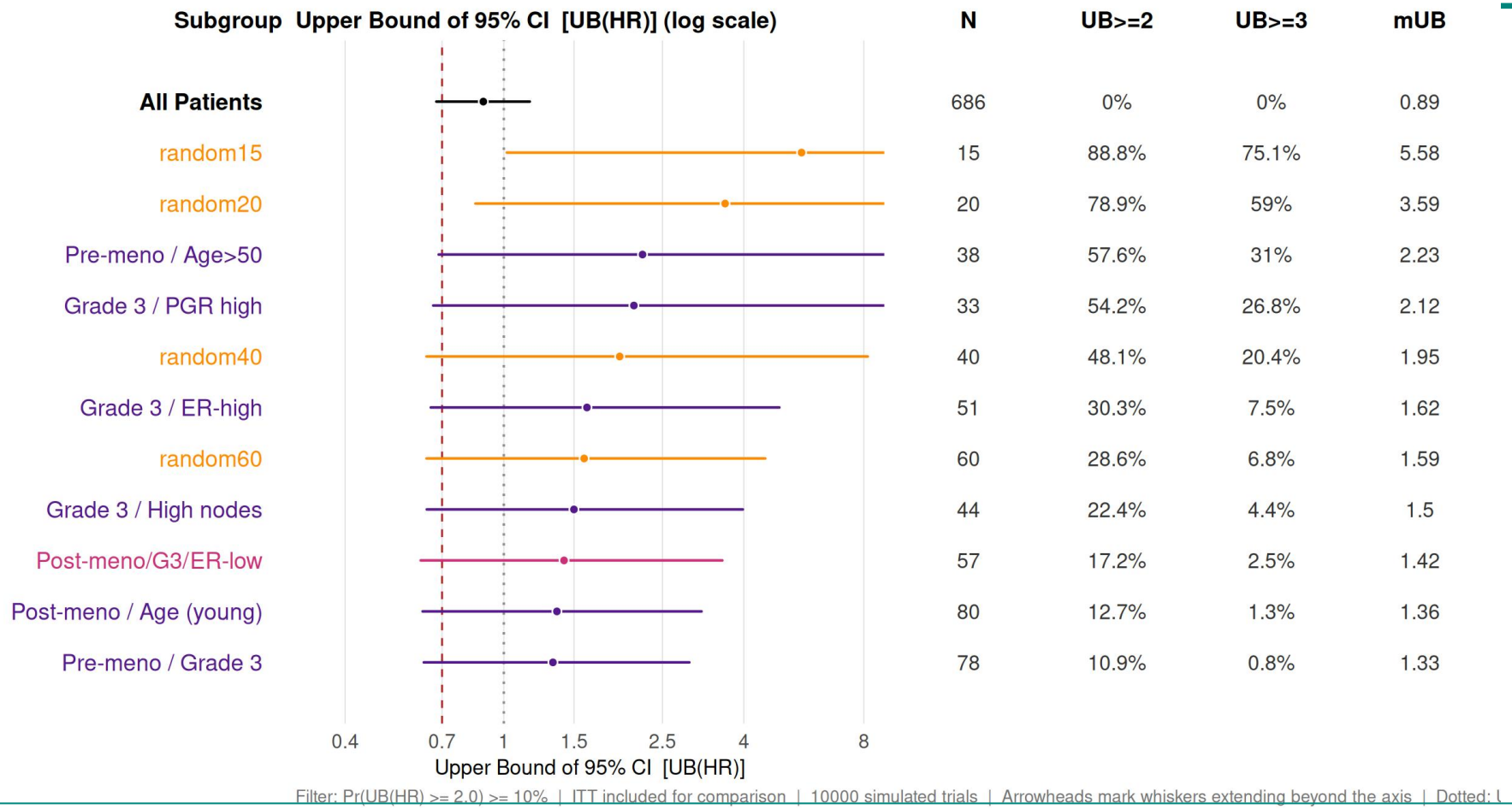


- Every row's median HR is approximately 0.70 — the Cox model is unbiased even at N = 15.
- But 27-37% of trials in subpops ≤ 40 produce HR < 0.5 (looks like a huge benefit); 25-36% produce HR > 1.0 (looks like harm).
- All this variability is sampling noise — the true HR is 0.70 by construction.

MSD

Distributions of Upper Bounds (Hazard Ratio) Displayed with 99% Estimated CIs

What a reviewer sees: the upper 95% confidence bound, across simulations



- random15 / random20 produce UB >= 2 in 75-89% of trials and UB >= 3 in 59-75%.
- Clinical / biomarker subgroups of similar size (Pre-meno/Age>50, Grade 3/PGR high, ...) show similar distributions.
- Arrowheads mark whiskers whose 99th percentile extends beyond the visible axis.

mUB = median of upper bound; UB = upper bound of 95% confidence interval

MSD

Key Takeaways

Reading the Upper Bound and Hazard Ratio panels together

1 Extremes are routine in small subpopulations

Under a uniform HR = 0.70, random15 trials produce UB(HR) > 3 about three-quarters of the time. Same true effect for everyone — the apparent heterogeneity is sampling noise.

2 Clinical rationale ≠ meaningful analysis

Pre-menopause & Age>50 (N = 38) behaves like random40 (N = 40). A defined subpopulation of size N has similar chance distributions of HR and UB 95% CI as a random pick of N.

3 The Cox model is working correctly

Median HR is approximately 0.70 across every row. The wide CIs accurately reflect uncertainty — the problem is interpretation, not the statistics.

4 Simulation-based calibration is a helpful tool

Without a null benchmark, no reviewer can quantify how 'extreme' UB = 3.5 in a 40-patient subpopulation actually is. The random* anchors provide that calibration.

MSD

So what?

- What happens if we say no and don’t analyse?
  - Currently unclear; wait and see what happens during the first few JCA submissions
- Why might we want to say no?
  - Mitigate risk of over- and mis-interpretation of analyses of small subpopulations
- How does this work help us?
  - A simulation-based approach can help calibrate understanding of subpopulation behavior in a given trial, allowing assessment of the most appropriate response to a given PICO (analyse vs. describe vs. push back on)

Resources

Package	github.com/larry-leon/forestsearch
Documentation	larry-leon.github.io/forestsearch/
Vignette	larry-leon.github.io/forestsearch/articles/extreme_subgroups.html
Install	pak::pak("larry-leon/forestsearch")