Benchmarking As A Tool For Assessing Cross-Country Comparability Of Survival Evidence For Health Technology Assessment (HTA)

HTA28

Scan to learn more

Philani Mpofu, PhD[1]; Elsie Horne, PhD[2]; Harlan Pittell, PhD[1]; Qianyi Zhang, MS[1]; Mohamed S Ali, SM[1]; Blythe Adamson, PhD, MPH[3]

[1]Flatiron Health, New York, NY, USA; [2]Flatiron Health UK, London, UK; [3]Formerly of Flatiron Health, New York, NY, USA

Benchmarking can help HTA judge **when cross-country survival evidence is worth borrowing**, and **how much to trust that judgement**

Introduction

- HTA bodies increasingly receive real-world overall survival (OS) data from foreign countries when local data is limited
- OS estimates from foreign countries may not reflect target/local context due to differing patients, treatments, and data systems
- The question is not only **whether** survival evidence can be borrowed, but **how credible** the evidence supporting that judgement is

Benchmarking compares OS **observed** in a target country with OS **expected** from another country after aligning measured patient characteristics

Objective

- To illustrate how structured benchmarking can indicate when OS observed in one country is relevant to another, using two case studies

Method

Figure 1. The workflow

Plan & Prespecify

Define target and source population · Understand context and feasibility · Data fit for use · Harmonize variables · Protocols registered before analysis (ISPOR RWE Registry)[1,2]

Ask five plain questions

1. Do the variables mean the same thing?
2. Do similar patients exist in both countries?
3. Are they treated similarly?
4. Do risk factors behave the same way?
5. Is follow-up comparable?

Only if the checks allow: compare OS

Assess the ability to recover OS in target country using the adjusted OS from source country after aligning patient characteristics

Data sources

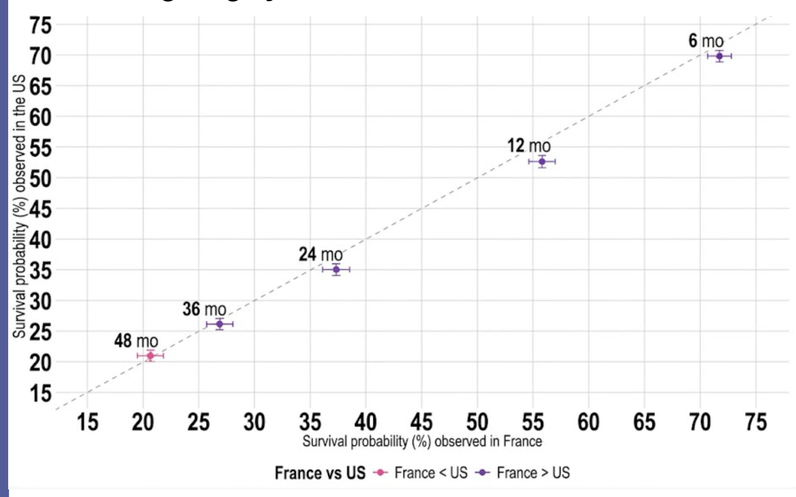
- **Metastatic NSCLC:** US Flatiron Health Research Database[3] (**source sample**; n = 41, 082) vs French UNICANCER ESME (**target sample**; n = 25, 529); diagnosed 2015–2023
- **Myeloma:** US Flatiron Health Research Database (**source sample**; n = 9,881) vs GermanOncology (**target**; n = 486)

Results

Figure 2. Same workflow, two different verdicts — Met · Partly met · Not met

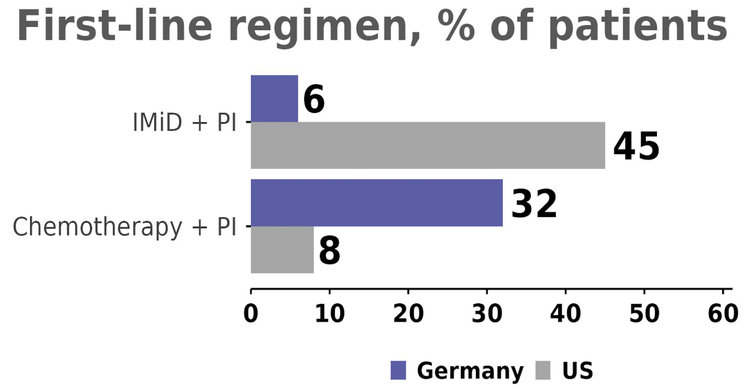
Plain question	NSCLC · US vs France	Myeloma · US vs Germany	Comment
1. Do the variables mean the same thing or could be harmonized? Measurement invariance	Staging, histology, ECOG, first-line treatment, and biomarker results could be aligned	US staged with ISS (43% missing); Germany with R-ISS (78% complete; 19% available in US)	NSCLC: large samples in both Myeloma: disease severity could not be aligned at all
2. Do similar patients exist in both countries? Positivity / covariate overlap	French cohort younger and more often male; clinical profile similar	Demographics could be aligned; disease severity could not be compared	
3. Are they treated similarly? Treatment consistency	Same order of first-line preference; French immunotherapy uptake ~2 years behind the US	Chemo + PI: 32% Germany vs 8% US. IMiD + PI: 6% vs 45%	
4. Do risk factors behave the same way? Conditional exchangeability	Same direction in both countries; size of effects differed modestly	Opposite directions for R-ISS III, deletion 17p and first-line chemotherapy	Myeloma: German n = 486 with wide confidence intervals, so reversals may partly be noise
5. Is follow-up comparable? Independent censoring	Comparable follow-up in full & exposure cohorts	Differential follow-up observed	
Verdict on comparing adjusted OS	**Informative, with caveats** Proceeded to OS comparison	**Comparison not recommended** Stopped before OS comparison	A "stop" is a result: it prevented a comparison that could have misled

Figure 3. NSCLC: survival broadly aligned after adjustment



- Good agreement over the first 48 months
- Weighting slightly **widened** the gap between point estimates

Figure 4. Myeloma: why the analysis stopped



Risk factors showed inconsistent associations with survival across countries

R-ISS stage III, deletion 17p and first-line chemotherapy were associated with **better** survival in Germany but **worse** survival in the US

Abbreviations: IMiD, immunomodulatory drug; PI, proteasome inhibitor; R-ISS, Revised International Staging System

Implications for HTA

Weigh the evidence, not just the verdict

Sample size, data completeness, prognostic consistency, and follow-up determine how much a "comparable / not comparable" call can be trusted

Know the limits

Benchmarking does not prove that survival evidence transfers between countries and does not estimate treatment effects

"Not feasible" is a useful answer

Conclusions depend on data quality, not just country differences. Assess data appropriateness when reviewing OS transferability claims

Build the evidence base

Studies across more tumour types, country pairs, and improved data will reveal where alignment is achievable versus not

Limitations

- Adjustment limited to variables captured in both countries; residual confounding cannot be excluded
- The benchmarking steps outlined cannot prove whether the transportability identifiability assumptions hold–they merely assess the plausibility that they hold from the observed data

References

1. Ling AY, et al. *Epidemiology*. 2023;34:627–36
2. FORUM protocols: osf.io/s8e36 (NSCLC); osf.io/9f5uj (myeloma)
3. Flatiron Health. Database Characterization Guide. Flatiron.com. Published March 18, 2025. Accessed October 6, 2026. https://flatiron.com/database-characterization

Acknowledgments: On behalf of the Flatiron FORUM consortium; we thank UNICANCER (ESME) and GermanOncology. We also thank Darren Johnson (Flatiron Health, Inc.) for publication support

Disclosures: This study was sponsored by Flatiron Health, Inc.—an independent member of the Roche Group. During the study period, all authors reported employment with Flatiron Health, Inc. and stock ownership in Roche

Contact information: Philani Mpofu, philani.mpofu@flatiron.com