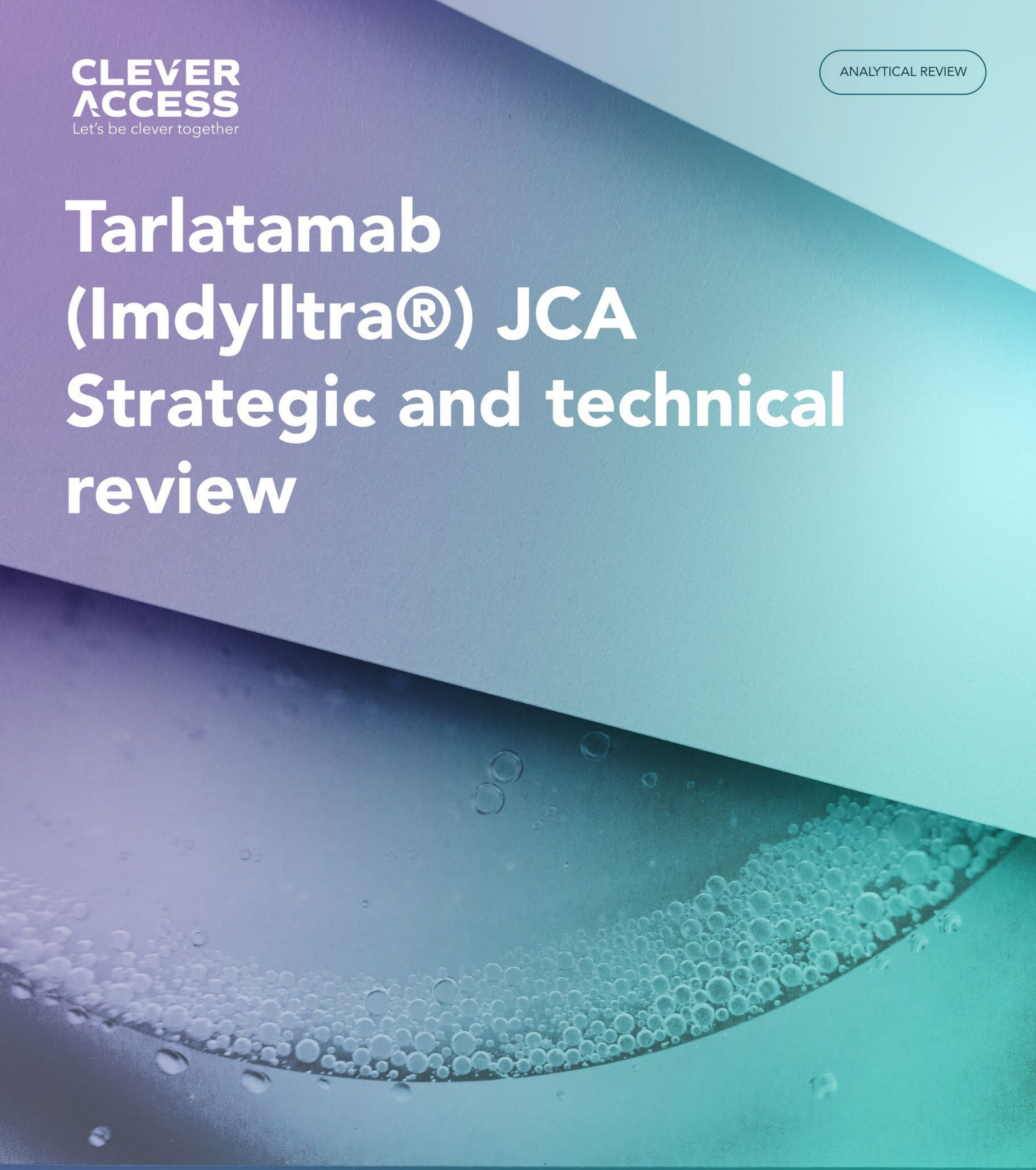
CLEVER ACCESS
Let's be clever together

ANALYTICAL REVIEW

Tarlatamab (Imdylltra®) JCA Strategic and technical review



ABOUT REPORT

Deep methodological, evidentiary and legal analysis.

AUTHORS


Mondher Toumi


Lylia Chachoua


Aleksandra Caban


Anna Kapuśniak

Published by MAS

CONTENTS

Contents

EXECUTIVE SUMMARY

Tarlatamab (Imdylltra®)

Joint Clinical Assessment

Executive summary of the deep methodological, evidentiary and legal analysis

Case	JCA-MP-2024-17 — Tarlatamab (Imdylltra®), Amgen Europe B.V.; extensive-stage small cell lung cancer after first-line platinum-based chemotherapy
Assessor / Co-assessor	IQWiG, Germany (B. Wieseler) / NCPHP, Hungary (N. Kósa)
Legal basis	Regulation (EU) 2021/2282 (HTAR); Implementing Regulation (EU) 2024/1381
Procedural status	Report v1.0; endorsed by the HTACG on 22 June 2026; EC procedural review 1 July 2026
Evidence base	One randomised, controlled, open-label phase III trial (DeLLphi-304, N=509); 4 populations, 7 PICOs
Register	Executive summary — the granular factual analysis is held in the companion technical review below

1. Headline conclusion

The assessment team (IQWiG with the Hungarian co-assessor) conducted the assessment in compliance with, and by actively enforcing, the binding methodological guidance adopted by the HTACG under Article 3(7)(d) of the HTAR. It anchored each results section to the guidance, independently verified the developer's analyses, recomputed relative effects the developer had omitted, and declined to present analyses it judged uninterpretable. The substantive limitations primarily stem from the evidence package and from the mismatch between a single pivotal trial and a sevenfold comparative scope.

What distinguishes this case from the first orphan assessments is that the dossier is based on a genuine randomised controlled trial. DeLLphi-304 is a phase III, open-label study of 509 patients comparing tarlatamab with standard second-line chemotherapy, with overall survival as the primary endpoint. For the three questions that could be answered directly from that trial, the assessment found statistically significant survival advantages. The report is therefore not an evidence-insufficiency story of the kind seen when only single-arm data exist.

The verdict is nonetheless nuanced. The report is procedurally compliant and methodologically defensible, but only partially conclusive. Of seven PICOs across four populations, three (comparisons against topotecan) are based on direct randomised evidence and show significant overall-survival effects. The remaining four rely on indirect comparisons, which the assessors themselves describe as, variously, subject to a "severely violated" similarity assumption, of "very low" certainty, or not to be "regarded as reliable information on relative effectiveness." A strong trial, re-cut into policy-defined subpopulations and supplemented by fragile indirect comparisons, produces a report far more equivocal than the underlying trial alone would suggest.

2. Compliance with the binding methodology

The assessment team's conduct is compliant and, in its rigour, exemplary. The team followed the Annex template of Implementing Regulation (EU) 2024/1381 (general information, scope, information retrieval, study characteristics and risk of bias, results per PICO), set the scope under Article 8(6) using the PICO framework, and documented the involvement of two carers and two clinical experts under Articles 8(6) and 11(4).

The team also went beyond passive review. It applied a full outcome-level risk-of-bias assessment to the randomised trial, rating overall survival as low risk of bias and progression-free survival,

objective response, duration of response, patient-reported outcomes and safety as high risk of bias, primarily because the trial was open-label. It recomputed the relative effects the scope required but the developer had not supplied, reproduced the indirect-comparison weighting to verify the figures, and declined to present single-item quality-of-life analyses it judged could not be adequately interpreted. This is the independent verification the methodology expects.

3. How the assessors handled the developer's evidence

The assessors issued a single formal request for additional information comprising nine numbered items, covering the network-meta-analysis code and effect estimates for the platinum-sensitive PICOs, the progression-free-survival estimand, quality-of-life analyses, proportional-odds testing, observation periods for patient-reported outcomes, treatment beyond progression, and discrepancies in the indirect-comparison sensitivity analyses. The developer responded in two batches, citing limited operational capacity, and a revised dossier followed.

Two features of the exchange matter. First, at least one requested analysis, a Bucher indirect comparison for the platinum-sensitive comparison against CAV chemotherapy, was not provided, so that question rests on a network meta-analysis alone. Second, the assessors record that the results tables and graphs in the main body of the dossier were not updated to reflect the analyses actually assessed, which exist only in the dossier appendices. The consequence is an auditability gap: the document a reader consults does not depict the numbers on which the assessment turned. This is a milder version of the factual-accuracy paradox seen in earlier cases, but it is the same in kind. The most consequential material is the least conveniently visible.

4. Certainty versus uncertainty

The template requires that the degree of certainty of the relative effects be assessed. In practice, the report documents factors affecting certainty extensively — open-label design, differential attrition in patient-reported outcomes, unequal safety observation windows, residual confounding and violated similarity assumptions in the indirect comparisons — but does not translate them into an explicit, graded certainty rating. There is no high/moderate/low/very-low classification derived through a transparent algorithm; the phrase "very low" appears only as a narrative description of the two indirect-comparison networks. Uncertainty-family vocabulary recurs throughout, and the recurring formula is "the certainty of the evidence for this outcome is affected by …," followed by a list of limitations rather than a conclusion. Certainty is therefore left to be inferred narratively rather than assessed independently, as a structured framework such as GRADE would require.

5. Evidence-based medicine (EBM) compliance

The report invokes "international standards of evidence-based medicine." It applies them selectively: it imports the internal-validity-appraisal half of the discipline (risk of bias, comparability, imprecision, similarity) while setting aside the parts of EBM that retain and grade all available evidence and integrate patient values. The divergence is clear. Analyses judged flawed are excluded rather than retained at low certainty: single-item quality-of-life analyses are "not presented"; certain subgroup analyses are deemed implausible and omitted; and the PICO 7 indirect comparison, which yields the only available estimate for that population, is characterised as descriptive and "not reliable." These are predictable consequences of using a comparative, price-informing instrument under EBM's banner, not random failures of execution. The important nuance is that, because a randomised trial exists, EBM's core preference for randomised evidence is satisfied for the direct comparisons; the selective application bites only at the edges of the scope.

6. Implicit judgement and reader influence

The report respects the HTAR prohibition on value judgement: it offers no overall added-value conclusion, no benefit ranking and no place-in-therapy recommendation. Yet it is not judgement-free in the ordinary sense. It continuously decides which analyses are admissible and which estimates are reliable. The most consequential implicit judgements are the decision to treat the two "severely violated" networks and the PICO 7 comparison as effectively uninformative, and the pervasive high-risk-of-bias ratings that follow from the open-label design. Each is defensible as scientific judgement, but together they steer the national reader towards a cautious, evidence-limited frame for four of the seven questions, even though the direct randomised comparisons are positive and significant. The report stays on the scientific side of the assessment/appraisal line, but its presentation choices clearly shape downstream national appraisal.

7. Legal fragility

If the assessment were challenged before the EU courts, the strongest arguments would focus on the quality of reasoning rather than the science: an inadequate statement of reasons under Article 296 TFEU (why a given uncertainty is treated as decisive; why one indirect comparison is reported and another withheld) and inconsistent treatment of uncertainty in the absence of an explicit threshold. Weaker arguments would assert manifest error in the indirect comparisons. A defence would likely prevail: EU bodies enjoy wide scientific discretion; review is limited to manifest error, misuse of powers and procedural irregularity; and the JCA report may not be a challengeable binding act because it does not itself determine price, reimbursement or market access. Because this case is anchored in a randomised trial with significant direct effects, the orphan-methodology theme that dominated earlier cases is materially weaker here; the residual legal exposure lies in the reasoning surrounding the indirect comparisons and the certainty gap.

8. Learnings for future JCA development

- **Align scope with ambition.** A single randomised trial cannot robustly answer seven comparative questions across four populations. A feasibility screen at scoping would prevent a recurrence of the pattern in which most questions can be addressed only by fragile indirect methods.
- **Adopt a structured uncertainty-to-certainty step.** The report catalogues uncertainty in detail but never grades it; a reproducible conversion to an explicit degree of certainty is the single highest-leverage improvement.
- **Require the dossier body to present the assessed analyses**. Analyses that appear only in appendices, while the main tables show superseded figures, undermine the auditability on which the framework's legitimacy rests.
- **Trace stakeholder influence.** A carer described the reviewer role as "rather symbolic"; a brief record of how input shaped scope, comparators and outcomes would make participation auditable.
- **Use Joint Scientific Consultation.** No JSC preceded this case; comparator alignment and the selection of trial arms for the indirect comparisons could have been addressed earlier.

The consistent bottom line: an RCT-anchored assessment, conducted rigorously and defensibly, still exposes the structural friction between a demanding multi-PICO comparative methodology and the evidence a single pivotal trial can bear. It also confirms that the missing step from uncertainty to certainty is a feature of the framework, not an artefact of the single-arm setting. It raises a sharper structural question: where a scope can be addressed only by indirect comparisons that the guidance will predictably reject, the requirement is self-defeating. Developers should be

able to decline an infeasible analysis, and the framework should decide whether indirect comparisons belong in the joint layer at all, rather than in national appraisal.

STRATEGIC REVIEW

Tarlatamab (Imdylltra®)

Joint Clinical Assessment

Strategic lessons from an RCT-anchored assessment of a rare, aggressive cancer

Subject	Joint Clinical Assessment of tarlatamab (Imdylltra®), a bispecific T-cell engager for extensive-stage small cell lung cancer after platinum-based chemotherapy
Framework	Regulation (EU) 2021/2282 (HTAR); methodological guidance of the HTACG
Significance	An early JCA anchored in a randomised phase III trial — a complement to the single-arm orphan cases
Register	Strategic / governance — the granular factual analysis is presented in the companion technical review

OVERVIEW

Executive Summary

> **IN BRIEF** An early assessment anchored in a randomised phase III trial, the tarlatamab JCA is procedurally robust and, for its direct comparisons, clinically positive. Yet the seven-fold scope forces four of seven questions into fragile indirect comparisons, and the report never grades certainty. The structural tensions are largely the same as those exposed by the single-arm cases — proof that they are features of the framework, not of the evidence base.

The Joint Clinical Assessment of tarlatamab is instructive precisely because it is not an orphan single-arm case. The dossier rests on DeLLphi-304, a randomised controlled phase III trial with overall survival as its primary endpoint. Whereas earlier assessments asked whether any comparative evidence could be generated at all, this one begins with a positive, adequately powered trial. It therefore isolates a different question: what does the JCA framework do with strong randomised evidence?

The framework's architecture partly dissipates the trial's strength. A single trial is re-cut into four policy-defined populations and seven comparisons. Three of these, against topotecan, are answered directly and show statistically significant survival benefits. The other four require indirect

comparisons: one unanchored matching-adjusted comparison and three anchored networks, two of which the assessors describe as resting on a “severely violated” similarity assumption and of “very low” certainty, and one of which they say should not be regarded as reliable. The result is a report whose overall tenor is far more cautious than its own direct evidence.

Beneath that outcome lie the same structural tensions documented in the first cases: uncertainty assessed but certainty never graded; assessors performing analytical work the dossier should have contained; stakeholder input recorded but its influence untraced; and judgement that is appraisal-free yet not judgement-free. That these recur even with a randomised trial in hand is the strategic point. They are not symptoms of thin evidence. They are properties of the framework itself.

SECTION 01

Tarlatamab as a Second Kind of Stress Test

> **IN BRIEF** If the orphan single-arm cases tested the framework at the point where evidence is scarcest, tarlatamab tests it where evidence is comparatively strong. That makes it a control condition: any remaining tensions cannot be attributed to the absence of a trial.

New systems are revealed by their edge cases, but they are also revealed by their easy cases. If a framework struggles where evidence is abundant and randomised, the struggle cannot be attributed to the evidence. Tarlatamab provides close to a best case for the JCA: a rare but not vanishingly rare disease, a large multinational randomised trial, a primary endpoint of overall survival, and a statistically significant result on that endpoint in the head-to-head comparisons.

Read against the single-arm orphan assessments, tarlatamab therefore serves as a control condition. Any tension that appears here, such as a certainty step that is never completed, stakeholder input that leaves no visible trace, or scope that outruns the evidence, is a tension the framework generates from its own design rather than one imposed on it by a difficult evidence base. The strategic value of the case is that it lets us separate the framework’s intrinsic features from the artefacts of scarcity. The question this review returns to is not whether tarlatamab works (the trial addresses that), but what the assessment of tarlatamab reveals about the instrument used to assess it.

SECTION 02

A Strong Trial Meets a Seven-Fold Scope

> **IN BRIEF** One randomised trial was asked to address seven comparative questions across four populations. Three could be answered directly; the other four had to be reconstructed using indirect methods. The scope did not collapse for want of evidence; it fragmented, and the fragments differ sharply in reliability.

The agreed scope defined four populations, distinguished by treatment-free interval and suitability for platinum re-treatment, and seven PICOs across them. The single pivotal trial compared tarlatamab with a standard-of-care chemotherapy arm combining topotecan, lurbinectedin and amrubicin. To answer the scope’s questions, the trial’s comparator arm had to be split into subpopulations, and comparisons with agents not directly tested in the trial, namely CAV chemotherapy, platinum re-treatment and carboplatin plus etoposide, had to be constructed using indirect methods and external trials.

This is a different failure mode from the scope collapse observed where no comparator evidence exists at all. Here, the scope does not collapse; it fragments. Three questions, namely the

comparisons against topotecan in the platinum-resistant, platinum-sensitive and not-suitable-for-retreatment populations, are answered directly and well. The remaining four are answered only by indirect comparison, and the assessors' own language marks their fragility. The strategic lesson is that a comparative framework can convert a single strong trial into a mosaic of unequal pieces, in which robust and unreliable findings are presented side by side and left for the reader to weigh. A scope calibrated to what the trial could actually support, rather than to the full theoretical comparator landscape, would have produced fewer, firmer answers.

SECTION 03

The Assessor Intervention Problem

> **IN BRIEF** The Regulation assumes that developers generate evidence and assessors evaluate it. Here, the assessors recomputed missing effects, reproduced the indirect-comparison weighting, and withheld analyses they judged uninterpretable. This is a rigorous assessment, but it again raises the question of where evaluation ends and generation begins.

The Regulation establishes what appears to be a clear division of labour: the developer generates and submits evidence; the assessors evaluate it; completeness review and Commission confirmation police the boundary. In practice, the assessors did analytical work themselves, computing relative effects omitted from the dossier, reproducing the matching-adjusted weighting to test the developer's figures, and declining to present analyses they considered flawed.

As in earlier cases, this is better read as the independent verification the methodology expects, rather than as a drift into evidence generation. The sharper concern is procedural. A dossier that has cleared completeness review and received Commission confirmation still required the assessors to recompute core quantities and contained a main body whose tables did not match the analyses ultimately assessed. That invites the question of whether completeness review is doing the work the architecture assumes, or whether the methodological adequacy of a submission is in practice discovered only during assessment. Where legitimate assessment ends and evidence generation begins is a line the framework does not draw, and, because the numbers that matter increasingly originate with, or are corrected by, the assessors, the provenance of the analytical record becomes harder to audit. That is a governance question, not merely a methodological one.

SECTION 04

Direct and Indirect: Where the Evidence Holds and Where it Breaks

> **IN BRIEF** The direct comparisons with topotecan carry the assessment; the indirect comparisons carry its caveats. The distinction is not incidental; it maps almost exactly onto which national decisions the report can and cannot support.

The three direct comparisons with topotecan show statistically significant overall-survival advantages, with hazard ratios below 0.6 in two of the three. These findings are ones a national body can act on with confidence, subject to the open-label caveat discussed below. The four indirect comparisons are another matter. The unanchored matching-adjusted comparison with CAV loses roughly a third of its effective sample through weighting and cannot adjust for several prognostic variables, so its bias runs in an unknown direction. Two anchored networks are described as resting on a "severely violated" similarity assumption and yielding "very low" certainty.

The comparison for the re-treatment-suitable population yields only a single progression-free-survival estimate, which the assessors say should not be regarded as reliable.

The strategic significance is that the reliability of the assessment tracks the method, and the method tracks the accident of which comparator the trial happened to include. Tarlatamab's trial isolated topotecan cleanly; it did not isolate the other comparators the scope demanded. A framework that lets the choice of trial comparator determine which national questions can be answered reliably is, in effect, outsourcing part of its own scope to decisions the developer made years earlier, for regulatory rather than assessment purposes. This is where early Joint Scientific Consultation would have had the greatest value. However, little is known about the usefulness of Joint Scientific Consultation, as nothing is transparent and no external assessment of the process and outcome is possible.

SECTION 05

The Factual-Accuracy and Auditability Paradox

> **IN BRIEF** The main dossier tables were not updated to show the analyses actually assessed; the assessed figures live in the appendices. Impropriety was not stated, but the document a reader consults does not depict the numbers on which the assessment turned.

In theory, the factual-accuracy exercise is straightforward: developers flag errors, assessors correct them, and interpretation remains with the assessors. Tarlatamab reveals a subtler version of the paradox seen in earlier cases. The assessors record that, following the request for additional information, the developer's revised analyses were provided, but the results tables and graphs in the main body of the dossier were not updated to reflect them. The analyses on which the assessment rests therefore appear only in the dossier appendices, while the main document a reader naturally consults still presents superseded figures.

The risk this creates is not that anyone acted wrongly, but that the most consequential material is the least conveniently visible. A framework whose legitimacy depends on being auditable is weakened whenever the assessed evidence and the presented evidence diverge. The paradox is the same as in disputes resolved "outside factual accuracy": the places where the real analysis lives are the hardest for a later reader, or a court, to reconstruct. The remedy is procedural and cheap: require the dossier body to depict the analyses actually assessed.

SECTION 06

Evidence-Based Medicine With an RCT in Hand

> **IN BRIEF** Because a randomised trial exists, EBM's core preference for randomised evidence is satisfied for the direct comparisons. The selective application of EBM therefore bites only at the edges, where flawed analyses are excluded rather than retained and graded.

In single-arm cases, the tension between EBM and the JCA was acute: the best available evidence was set aside because it addressed activity rather than a relative effect. Tarlatamab softens that tension in its direct comparisons, because randomised evidence is exactly what EBM prizes and what the trial supplies. The purpose mismatch does not vanish, but it recedes.

It reappears at the edges. EBM retains the best available evidence, however weak, and grades it; the JCA tends to exclude analyses once a limitation is identified. Single-item quality-of-life analyses are not presented because they cannot be adequately interpreted; certain subgroup analyses are judged implausible and omitted; the only estimate available for the re-treatment-suitable population is characterised as descriptive and unreliable. EBM would typically retain such analyses at low

certainty and bound their interpretation. The divergence is therefore narrower here than in the single-arm orphan cases, but it is the same divergence: the JCA imports EBM's apparatus for appraising validity while declining the part of EBM that retains and grades weak evidence.

SECTION 07

Uncertainty Without Certainty

> **IN BRIEF** The report catalogues the sources of uncertainty in each estimate but never states how much confidence the estimate deserves. With significant direct effects and "very low" indirect effects in the same document, the absence of a graded certainty step leaves national bodies to weigh them unaided.

The report's treatment of uncertainty is thorough. For each outcome, it lists the factors affecting certainty, including open-label design, differential attrition in patient-reported outcomes, unequal safety observation windows, residual confounding, and violated similarity assumptions. What it does not do is convert that catalogue into an explicit statement of how much confidence each estimate deserves, nor does it systematically specify the potential direction of uncertainty. There is no starting certainty by study design, no transparent movement between levels, and no final high / moderate / low / very-low category derived by algorithm or intellectual aggregation. The phrase "very low" appears, but as a narrative description of two networks, not as the output of a grading step.

This matters more, not less, because the direct evidence is strong. A national body now holds a document in which a significant survival benefit from a randomised comparison and a "very low" certainty estimate from a violated network appear with the same formal status, each accompanied by a list of limitations. Without a graded certainty conclusion, the reader must supply the decisive weighting personally, and the same evidence may support different conclusions in different jurisdictions for interpretive rather than scientific reasons. If the purpose of the JCA is a common assessment of relative effectiveness, the missing link from uncertainty to certainty is not a detail; it is the gap at the centre, and it is visible even when the evidence is good.

SECTION 08

The Judgement-Free and Decontextualisation Paradoxes

> **IN BRIEF** The report issues no value judgement, yet scientific judgement runs through every exclusion and every risk-of-bias rating. By excluding context, it risks relocating contextual reasoning rather than removing it.

The report respects the HTAR's separation of assessment from appraisal: it reaches no conclusion on clinical added value, reimbursement or place in therapy. Yet it is saturated with scientific judgement, rating most outcomes as high risk of bias, excluding analyses, declaring similarity assumptions violated, and deciding which of the developer's comparisons to report. This is the first paradox: appraisal-free is not judgement-free.

The second follows from decontextualisation. The report assesses validity and uncertainty while largely excluding the practical realities raised by clinical experts, including the difficulty of delivering an inpatient step-dose regimen, the signature cytokine-release and neurological toxicities of this drug class, and the poor prognosis after platinum failure. It is difficult to believe such considerations vanish from the reasoning; the more plausible account is that they become implicit. Implicit contextualisation cannot be audited, challenged or reproduced in the way explicit contextualisation

can. The risk of decontextualisation is therefore not the removal of context but the creation of hidden context that shapes conclusions without appearing in the report.

SECTION 09

The Participation Paradox: Input Heard, Influence Unseen

> **IN BRIEF** Two carers and two clinical experts were given a genuine route into the process, and their input was annexed. But the report never shows how it changed anything, and one carer said the reviewer role felt “rather symbolic.”

The framework built participation into the procedure: two carers and two clinical experts provided written input on the scope and the revised draft reports, and their contributions were annexed. On the procedural test, was the stakeholder voice given a route in? The answer is yes.

Yet the file contains a quietly damning observation from a carer, who described the patient-reviewer role as “rather symbolic,” asked why there was no plain-language overall summary across the seven PICOs, and requested a cohort of patients with brain metastases. The carer also noted that low-grade adverse events, which the analysis treats as minor, can impose a substantial long-term burden on quality of life. These are precisely the considerations a family weighs, and precisely the ones the report either did not restructure to address or classified as beyond the comparative frame. The report records the input but never traces its influence on the scope, the comparators or the outcome set.

This is the third leg of EBM: the patient’s values, invited in, then left without a trace in the reasoning. An influence that cannot be demonstrated is, for every purpose of audit and accountability, indistinguishable from one that never occurred. The participation paradox is not separate from the certainty gap; it is the same transparency deficit, viewed from the input side.

SECTION 10

The Common Root: One Omission Beneath the Paradoxes

> **IN BRIEF** The selective EBM, the uncertainty-without-certainty gap, the surviving judgement and the untraced patient voice are not four findings. They are projections of one omission: the report explains why evidence is uncertain but never grades how far it can be trusted.

It is tempting to read the preceding tensions as independent. They are not. Three are projections of a single choice: the report explains why the evidence is uncertain but never completes the step of converting that uncertainty into an explicit, graded degree of certainty. The fourth, participation, reflects the same deficit on the input side: input recorded, influence untraced.

That one missing step explains the others. It is the EBM departure, because grading certainty is exactly the part of the discipline the JCA declines to import. It is, by definition, the uncertainty-without-certainty gap. And it is the mechanism by which judgement becomes implicit: the conversion from “here are the limitations” to “this is how far the estimate can be trusted” does not disappear when left unstated; it migrates out of the report and into the reader. A national body cannot act on a list of weaknesses alone; someone must decide how much they matter. When the report declines to make that judgement explicit, it does not avoid judgement; it delegates it invisibly. That tarlatamab, with a strong randomised trial, exhibits the same root as the single-arm cases is decisive evidence that the omission is structural.

SECTION 11

Open-Label Design and the Discounting of a Positive Trial

> **IN BRIEF** The trial's open-label design leads to high risk-of-bias ratings across almost all outcomes, except overall survival. This is defensible, but it means a positive trial is downgraded for a feature the developer could not easily have avoided in this disease and drug class.

Because tarlatamab is administered very differently from chemotherapy and produces a distinctive toxicity profile, blinding the pivotal trial was impractical. The assessors, correctly, rate overall survival as low risk of bias, as death is an objective endpoint, but rate progression-free survival, objective response, duration of response, all patient-reported outcomes, and safety as high risk of bias, largely because the open-label design permits detection bias and produces differential attrition and unequal observation windows.

Each rating is individually defensible. Their cumulative effect is a quiet discounting of a positive trial for a design feature that was, in practice, unavoidable. The pattern echoes the conservatism reflex seen elsewhere: where a limitation can be named, it is; and the accumulation of named limitations shades towards caution. The strategic question is whether the framework has a consistent, stated rule for weighing evidence whose imperfections are structural to the therapy, such as an immunotherapy that cannot be blinded or a toxicity that reveals the arm, rather than treating each such limitation as if better evidence were available and had simply not been produced. Without such a rule, the framework risks systematically under-crediting exactly the classes of therapy whose mechanism makes clean blinding impossible.

SECTION 12

Harmonisation and Its Limits

> **IN BRIEF** For the direct comparisons, the joint layer delivers something a national body could not easily reproduce alone. For the indirect comparisons, it hands down an inventory of weaknesses — relocating the work downstream rather than removing it.

The HTAR's promise is harmonisation: one rigorous clinical assessment conducted once, replacing duplicated national reviews. Tarlatamab shows the promise half-kept. For the three direct comparisons, the joint assessment delivers exactly what harmonisation should, a single, reproducible, high-quality analysis of randomised evidence that no national body would gain from redoing. For the four indirect comparisons, the joint layer produces an inventory of limitations and a set of estimates it declines to vouch for, leaving each national body to decide how much the documented weaknesses matter and, potentially, to commission its own analysis.

The seam between the harmonised assessment and national appraisal runs straight through the middle of this single case. Where the joint layer can speak confidently, direct comparisons make harmonisation efficient and real. Where it can speak mainly of uncertainty, indirect comparisons shift duplication downstream rather than eliminating it. The credibility of the European project rests on the joint layer providing national decision-makers with something they could not readily have produced alone; tarlatamab shows it can do so for part of the scope and not for the rest, and that the dividing line is the availability of direct randomised evidence.

SECTION 13

Is the Tarlatamab JCA Fit for Purpose?

> **IN BRIEF,** procedurally and methodologically, yes; and for the direct comparisons against topotecan, it is decision-useful. For the four indirect questions, it provides a rigorous account of why the evidence is uncertain, but the evidence it assesses was, in effect, compelled by a scope that its own guidance will not accept, so the report is fit for some of its readers far more than others. That is a description of the framework, not a criticism of the assessors.

Procedurally, the report follows the framework and applies the guidance consistently. Methodologically, it is coherent and transparent about limitations. For the national bodies that must use it, the picture divides along the direct/indirect line. For the topotecan comparisons, the report is close to decision-ready: a significant survival benefit from a randomised trial, appropriately caveated for its open-label design. For the other four questions, it identifies uncertainty clearly but communicates certainty poorly, and allows appraisal to begin without equipping it to conclude.

“Fit for purpose” conceals the prior question: fit for whose purpose? For the framework’s stated aim, a transparent assessment of relative effectiveness, the report largely succeeds. For a national body seeking a usable measure of how much tarlatamab helps in each population, it succeeds for three and defers on four. For the families the carer spoke for, it is largely silent on the outcomes they named. These are three audiences a single document cannot equally serve, and recognising that is itself the strategic insight: the framework’s next increment of value lies in closing the gap on the questions its own structure leaves unresolved.

SECTION 14

Practical Lessons for Health Technology Developers

> **IN BRIEF** Five operational lessons, centred on one message: in a multi-PICO scope, the comparative-evidence pathway for every population must be planned before the pivotal trial is designed, because the trial’s comparator arm silently determines which questions can later be answered directly.

- **Design the trial comparator with the full PICO scope in mind.** Tarlatamab’s trial cleanly isolated topotecan, so those questions were answerable directly; the comparators it did not isolate fell to fragile indirect methods. The comparator arm is a scope decision made years earlier.
- **Treat indirect comparisons as primary, not supporting, analyses.** Where they will carry a PICO, the identification and reporting of prognostic variables and effect modifiers must be planned prospectively; they cannot be recovered later.
- **Keep the dossier body synchronised with the assessed analyses.** Revised analyses that exist only in appendices, while the main tables show outdated figures, invite avoidable findings on transparency and auditability.
- **Anticipate that the open-label design will be read down**. For therapies that cannot be blinded, pre-specify objective endpoints and blinded assessment wherever feasible, and document why blinding was impractical.
- **Use Joint Scientific Consultation.** No JSC preceded this case; comparator alignment, trial-arm composition and outcome definitions across populations could have been addressed earlier.

A corollary deserves emphasis. Because indirect comparisons of the kind this scope requires are unlikely to satisfy the JCA guidance, a developer that submits them risks subjecting a well-conducted comparative trial to an extensive list of methodological criticisms that can depress the perceived value of the product without adding information. Where an indirect comparison cannot meet the methodological requirements, a developer may reasonably decline to generate it and instead document, on the record, why the requested analysis is infeasible under the guidance. An acknowledged, well-argued evidence gap is more defensible and more useful to national bodies than an analysis produced only to be dismissed.

SECTION 15

What Reform Would Actually Require

> **IN BRIEF** Because several tensions share a transparency root, the reform agenda is narrower than the list of problems suggests. It calls less for new methodology than for making a few implicit elements visible, above all, a reproducible step from uncertainty to certainty.

If several of the tensions share a transparency root, the reform agenda is more tractable than the catalogue of individual problems suggests. The first and highest-leverage change is a reproducible step that converts documented uncertainty into a stated degree of certainty. That single step would close the EBM gap, give national bodies the confidence signal they need, and force the decisive judgement into the open. A framework already willing to catalogue uncertainty in exhaustive detail is one short, procedural step away from grading it.

The second is a visible trail from stakeholder input to assessment decisions, a short, honest record of how carer and clinical-expert views shaped, or did not shape, the scope, comparators and outcomes. This need not manufacture influence; it need only make the influence that occurred auditable, and its absence equally so. A carer who calls the role “symbolic” has told the framework exactly what this record would address.

The third is a stated rule for weighing evidence, acknowledging that some imperfections are structural to the therapy — the un-blindable immunotherapy, the comparator the trial could not isolate, the subpopulation too small to power. The required adaptation is not a lower standard but a consistent, written rule for reasoning under unavoidable limitation, so that a positive randomised trial is not quietly discounted for features no diligence could have removed. Underlying all three is a single move: drawing, in writing, the boundaries the framework currently leaves to discretion. That is a governance task more than a scientific one.

A fourth reform follows from the first three. Where the only way to address a PICO is to present evidence that the JCA guidance itself classifies as low quality, the developer should be permitted to refrain from generating it and instead argue, on the record, that the requested analysis cannot meet the JCA's methodological requirements. Compelling the production of evidence that the guidance is designed to reject serves no evidentiary purpose and consumes resources that inform no decision.

Without such a provision, the model becomes circular and inefficient. The assessors require an analysis to be generated; the analysis cannot reach the required quality, as is foreseeable at scoping; and it is then produced only to attract a battery of criticism that adds no value and leaves Member States no better informed. The inefficiency is compounded downstream: the indirect-comparison methods of many national HTA bodies diverge materially from the JCA guidance, so national agencies, particularly those operating within a cost-effectiveness framework, must in any case re-analyse the indirect comparisons or set aside the assessors' conclusions. The joint exercise thus imposes costs at the centre without removing duplication at the national level.

IN CLOSING

Conclusion

IN BRIEF With a randomised trial in hand, the tarlatamab JCA answers three of its seven questions well and defers on four. That it still exhibits the certainty gap, the participation paradox and implicit judgement confirms that these are structural features of the framework, and that its credibility will be decided by closing a small number of boundaries rather than by adding methodology. The case also raises two pointed questions: whether the guidance for indirect treatment comparisons is fit for purpose, and whether indirect comparisons belong within the JCA at all, or should be left to national appraisal.

The tarlatamab JCA should be read as a complement to the first orphan assessments. Where those tested the framework where evidence was scarcest, this tests it where evidence is comparatively strong — and finds that the same structural tensions persist. The report is procedurally robust and, for its direct comparisons, clinically positive; yet it fragments a single strong trial into unequal pieces, downgrades a positive study for an unavoidable design feature, documents uncertainty without grading it, and records stakeholder input without tracing its influence.

None of these would, on its own, invalidate the assessment. Together, they define the reform agenda, and several are expressions of the same underlying problem: a system that records uncertainty without grading certainty, captures the patient voice without showing its impact, and leaves its most important judgements unstated. Tarlatamab's lasting contribution is to demonstrate that this is true even when the evidence is good — and therefore that the boundaries the framework must clarify are not artefacts of scarcity but properties of its own design.

TECHNICAL REVIEW

Tarlatamab (Imdylltra®)

Joint Clinical Assessment

Claim-by-claim, against the binding methodology and the source record

Case	JCA-MP-2024-17 — Tarlatamab (Imdylltra®), Amgen Europe B.V.; extensive-stage small cell lung cancer after platinum-based chemotherapy
Assessor / Co-assessor	IQWiG, Germany (B. Wieseler) / NCPHP, Hungary (N. Kósa)
Legal & methodological basis	Regulation (EU) 2021/2282 (HTAR); Implementing Regulation (EU) 2024/1381; HTACG guidance under Art. 3(7)(d)
Procedural status	Report v1.0; endorsed by the HTACG on 22 June 2026; EC procedural review 1 July 2026
Register	Technical, strong factual claims; the conceptual / governance argument is held in the companion strategic review

BOTTOM-LINE SUMMARY

Bottom-Line Summary

BOTTOM LINE Two actors, two verdicts. The assessment team's conduct is rigorous and largely exemplary; the substantive limitations lie in the evidence package and in a scope that exceeds a single trial. The report is procedurally compliant and methodologically defensible. Three of seven PICOs (versus topotecan) are based on direct randomised evidence and show significant overall-survival benefits (HR 0.491, 0.628, 0.578); the other four rely on indirect comparisons that the assessors describe as "severely violated," of "very low" certainty, or unreliable.

SECTION 01

Purpose, Scope and Method of This Review

> **KEY MESSAGES**
>
> **Two actors, judged separately.** The assessment team (IQWiG / Hungary) versus the developer's dossier — this distinction governs every finding.
>
> **One benchmark.** Every claim is anchored to the source record and measured against HTACG guidance adopted under Art. 3(7)(d) of HTAR.
>
> **Headline verdict.** Procedurally compliant and methodologically defensible; conclusive for direct comparisons, fragile for indirect ones.

This review evaluates the tarlatamab JCA against the rules that bind it, claim by claim, and against the source record: the JCA report, the developer's dossier and its appendices, and the request-and-response documentation in Appendix D. While the companion strategic review asks what the case reveals about the framework, this document asks the narrower questions — what exactly was done, under which provision, with what figures, and with what documented effect.

Central thesis. The case requires two actors to be assessed separately. The assessment team conducted the assessment in compliance with — and by actively enforcing — the binding HTACG methodology. The substantive limitations lie in the evidence package and in the mismatch between one pivotal trial and a seven-PICO scope. The defensible verdict is that the report is procedurally compliant and methodologically defensible, conclusive in outcome for the three direct comparisons and only weakly informative for the four indirect comparisons.

Sources and standard of proof. Every factual claim is linked to the source record. Quantitative claims are reported as they appear in the report; figures that are approximate or audit-derived are flagged. The binding benchmark is the HTACG guidance under Art. 3(7)(d) HTAR — in particular, the guidance on quantitative evidence synthesis, outcomes for the JCA, and completion of the dossier template.

SECTION 02

Procedural and Structural Compliance

> **KEY MESSAGES**
>
> **All testable procedural requirements are met.** Template, scope, stakeholder involvement and transparency each meet the requirements.
>
> **No procedural non-compliance**. The compliance question turns on the substantive methodology outlined below.

On the procedural and structural requirements, the assessment team is compliant on each testable point:

Requirement	Finding
Template	Follows the Annex template of Implementing Regulation (EU) 2024/1381 — general information, background, scope, information retrieval, study characteristics and risk of bias, and results per PICO.
Scope	Set under Article 8(6) HTAR using the PICO framework: 4 populations (by treatment-free/relapse-free interval and suitability for platinum re-treatment) and 7 PICOs.
Stakeholder involvement	Documented under Articles 8(6) and 11(4): two carers and two clinical experts provided input on the scope and the revised draft reports, as annexed in Appendix A.
Joint Scientific Consultation	None held (Table 4); consequently, no deviations from JSC propositions to report (Table 5).
Risk of bias (RCT)	Full outcome-level RoB assessment applied to DeLLphi-304 (Table 27): OS low; PFS, ORR, DoR, all PROs, and safety high, primarily due to the open-label design.

None of these constitutes a point of non-compliance. The compliance question turns entirely on the substantive methodology addressed in the sections that follow.

SECTION 03

The Evidence Base: One RCT, Four Comparison Types

> **KEY MESSAGES**
>
> **One randomised pivotal trial.** DeLLphi-304 (N=509) versus standard-of-care chemotherapy (topotecan, lurbinectedin, amrubicin), with overall survival as the primary endpoint.
>
> **Four comparison types.** Direct (vs topotecan); anchored network / Bucher (vs platinum re-treatment, CAV, carbo-etoposide); unanchored MAIC (vs CAV, platinum-resistant).
>
> **Reliability tracks method.** The direct comparisons are robust; the indirect ones carry caveats.

The assessment is based on DeLLphi-304, a phase III, randomised, controlled, open-label, multinational trial of 509 patients (254 tarlatamab versus 255 standard-of-care chemotherapy), with overall survival as the primary endpoint and a median follow-up of 11.2 months at the data cut-off used. The comparator arm combined topotecan, lurbinectedin and amrubicin; to address the scope, the trial's arms were divided into subpopulations, and comparisons with agents not isolated in the trial were made using indirect methods and external trials (GFPC 0501, Baize 2020, von Pawel 1999).

The four comparison types differ markedly in inferential strength. Direct comparisons with topotecan (PICOs 1, 4, 6) use randomised within-trial subpopulations and preserve the trial's internal validity. Anchored indirect comparisons (PICOs 3, 5, 7) rely on a common comparator (topotecan) linking DeLLphi-304 to external trials; their validity depends on the similarity of the external trials, which the assessors found inadequate. The single unanchored matching-adjusted comparison (PICO 2, versus CAV) has no common comparator and rests on the strong assumption that all relevant prognostic variables and effect modifiers have been captured — an assumption the assessors judged could not be verified. For the unanchored comparison, no formal risk-of-bias rating is performed, consistent with guidance that risk-of-bias tools are not applied to single arms used in unanchored comparisons; validity is instead appraised narratively.

3.1 Safety and the signature toxicities

Safety illustrates both the strength and the limits of the evidence. In the direct comparison of the platinum-resistant population (PICO 1), cytokine-release syndrome occurred in 59.5% of the tarlatamab arm versus 0% of the topotecan arm, with a large proportion of these events serious — a class effect of bispecific T-cell engagers that the clinical experts identified, alongside immune

effector cell-associated neurotoxicity (ICANS), as the drug's signature toxicity. These events are absent from the chemotherapy comparator by mechanism, so the safety comparison is asymmetric both qualitatively and quantitatively.

The assessors rate all safety outcomes as having a high risk of bias, primarily because the open-label design and unequal treatment duration create unequal observation windows: adverse events were recorded only for a fixed period after the last dose, while treatment duration differed between the arms. Crucially, for the indirect comparisons, no safety analysis was feasible — the report states that no safety results are available for the matching-adjusted and network comparisons. As a result, the populations addressed only by indirect methods (PICOs 2, 3, 5, 7) have, in effect, no comparative safety assessment, even though the toxicity profile is one of the most decision-relevant features of this therapy for clinicians and patients alike.

SECTION 04

The Seven PICOs and Their Results

KEY MESSAGES

Three direct comparisons, all positive on OS. PICOs 1, 4 and 6 (versus topotecan) show significant overall-survival benefits.

Four indirect comparisons, heavily caveated. PICO 2 unanchored; PICOs 3 and 5 "severely violated" similarity, "very low" certainty; PICO 7 a single PFS estimate the assessors call unreliable.

Patient-important outcomes are mixed. OS assessable across most PICOs; PFS often non-significant; duration of response not calculable; safety data unavailable for the indirect comparisons.

The results by PICO, as reported:

PICO / population	Method vs comparator	Principal reported results
PICO 1 — Pop. A (TFI <3 mo, platinum-resistant)	Direct RCT vs topotecan	OS HR 0.491 [0.332, 0.727], p<0.001; PFS HR 0.748 [0.541, 1.033], p=0.078 (ns); ORR OR 6.698 [2.426, 18.496], p<0.001; DoR — no effect measure calculable. CRS 59.5% vs 0%.
PICO 2 — Pop. A	Unanchored MAIC vs CAV	OS significant, favouring tarlatamab; PFS ns; ORR ns; no safety (ITC not feasible). ESS fell from 108 to ≈78.4 (base case); bias of unknown magnitude and direction.
PICO 3 — Pop. B (TFI ≥3 mo, platinum-sensitive)	NMA + Bucher vs platinum re-treatment	OS favours tarlatamab (95% CrI excludes 1); PFS no difference; ORR no difference; DoR not feasible. Similarity assumption "severely violated"; certainty "very low."
PICO 4 — Pop. B	Direct RCT vs topotecan	OS HR 0.628 [0.415, 0.951], p=0.028; PFS HR 0.788 [0.581, 1.069], p=0.13 (ns); further outcomes assessable; DoR not calculable.
PICO 5 — Pop. B	Bayesian NMA vs CAV	OS HR 0.65 [0.40, 1.08] (CrI includes 1); ORR OR 2.24 [0.93, 5.55] (ns); no safety; Bucher ITC not provided by developer. Similarity "severely violated"; certainty "very low."
PICO 6 — Pop. C (RFI ≤6 mo, unsuitable for re-treatment)	Direct RCT vs topotecan	OS HR 0.578 [0.426, 0.786], p<0.001; PFS HR 0.769 [0.600, 0.987], p=0.039; ORR assessable; DoR not calculable.
PICO 7 — Pop. D (RFI >6 mo, suitable for re-treatment)	Bucher ITC vs carboplatin + etoposide	Only PFS available: HR 3.62 [1.65, 7.94], p=0.001. Assessors: "subject to major uncertainties … should not be regarded as reliable information on relative effectiveness." No OS, ORR, safety, sensitivity or subgroup analyses.

Reading the pattern. The three direct comparisons convey the assessment's substantive findings: statistically significant overall-survival advantages, robust to the open-label caveat because death is an objective endpoint. The four indirect comparisons convey its caveats: an unanchored comparison with unknown bias direction, two networks the assessors judged to rest on a severely violated similarity assumption, and a single progression-free-survival estimate the assessors declined to treat as reliable. Duration of response could not be calculated in the direct comparisons because randomisation is not preserved among responders, and safety could not be assessed in the indirect comparisons at all.

SECTION 05

Assessor Conduct: Verification, Recomputation and Exclusion

> KEY MESSAGES
>
> **The dossier was not accepted at face value.** The team recomputed missing effects, reproduced the MAIC weighting, and withheld analyses deemed uninterpretable.
>
> **Exclusion over caveated presentation.** Single-item quality-of-life analyses and implausible subgroup analyses were not presented.
>
> **Disputes are genuine and a matter of degree.** The direction of the critique is well founded; several points raised by the developer are reasonably contested.

The assessment team independently verified, recomputed and, where necessary, excluded — each step was documented and consistent with the methodology's expectation of independent verification.

5.1 Recomputation of missing relative effects

The scope required relative effects that the dossier did not provide in usable form; the assessors calculated them (for example, deriving relative effects where the dossier reported group-level rates), documented the approach in the dossier appendices, and conducted the statistical work themselves.

5.2 Reproduction of the indirect-comparison weighting

For the unanchored matching-adjusted comparison, the assessors reproduced the weighting to verify the developer's figures and quantify the loss of effective sample size — a verification of submitted analyses rather than the generation of new evidence. The effective sample size fell from 108 to approximately 78, and the assessors flagged that the comparison could not adjust for several prognostic variables, leaving a bias of unknown magnitude and direction.

5.3 Exclusion rather than caveated presentation

Where analyses were deemed uninterpretable, the assessors declined to present them. Single-item quality-of-life analyses "cannot be adequately interpreted" and are not presented in the JCA report, remaining only in a dossier appendix; certain subgroup analyses for continuous outcomes were judged implausible and omitted; and results for the majority of outcomes for which relative estimates were requested are recorded as not provided, owing to missing comparator data or non-comparable definitions.

5.4 The documented developer–assessor disputes (even-handed)

Several assessor positions are genuinely contested. The central technical disputes concern the choice and reliability of the indirect-comparison methods, the adequacy of the prognostic-variable set in the unanchored comparison, and the interpretation of the progression-free-survival estimand under treatment beyond progression. These are real methodological disagreements — a strict

reading against the developer's more pragmatic one — not merely the developer conceding fault. The fair conclusion: the assessors' criticisms are well founded under the binding guidance, though several are matters of degree, and several underlying problems (external comparator trials, heterogeneous populations, subpopulations too small to power) are structural to a rare, aggressive disease rather than failures of execution.

SECTION 06

Handling of Additional-Information and Factual-Accuracy Issues

> **KEY MESSAGES**
>
> **One request round, nine items.** Delivered by the developer in two batches, citing limited operational capacity; a revised dossier followed.
>
> **One requested analysis was not provided.** A Bucher ITC for the CAV comparison in the platinum-sensitive population was not provided.
>
> **An auditability gap.** The main dossier tables were not updated to reflect the assessed analyses, which are only in the appendices.

The assessors issued a single formal request for additional information comprising nine numbered items — covering network-meta-analysis code and effect estimates for the platinum-sensitive PICOs (excluding out-of-scope comparators), the progression-free-survival estimand and treatment-policy analyses, quality-of-life analyses, proportional-odds testing, patient-reported-outcome observation periods, missing-data counts, numbers treated beyond progression, discrepancies in the matching-adjusted sensitivity analyses, and tabulated interaction p-values.

The developer responded in two batches and submitted a revised dossier. The assessors record that most of the requested information was addressed, but note two residual technical points. First, the developer did not provide the requested Bucher indirect comparison for the CAV comparison in the platinum-sensitive population, so that question rests on a network meta-analysis alone. Second, the results tables and graphs in the main body of the dossier were not updated to reflect the analyses actually assessed, which are available only in the dossier appendices. A clinical expert raised the same point. The operative technical finding is that the assessed evidence and the presented evidence diverge — the same auditability concern seen when disputes are resolved outside the factual-accuracy channel, here arising from a documentation lag rather than a classification decision.

SECTION 07

Certainty Versus Uncertainty (Quantified)

> **KEY MESSAGES**
>
> **Uncertainty documented, never graded.** Limitations are catalogued per outcome, but never converted into a high / moderate / low / very-low rating derived by an algorithm.
>
> **"Very low" is narrative, not a grade.** It appears only as a description of the two indirect-comparison networks.
>
> **The decisive step is displaced.** Without a GRADE-style conclusion, conversion to certainty falls to the national reader.

The template requires an assessment of the degree of certainty regarding relative effectiveness and safety. The report documents uncertainty exhaustively but does not provide a graded certainty conclusion. The recurring formula is "the certainty of the evidence for this outcome is affected by

…," followed by a list of limitations. There is no occurrence of a graded certainty label derived through a transparent algorithm, and no use of GRADE; uncertainty-family vocabulary dominates the results narrative, and the phrase "very low" appears only as a narrative description of the two networks (PICOs 3 and 5).

The contrast with a structured framework is instructive:

Element	GRADE	Tarlatamab JCA report
Starting certainty	Explicit, by study design	None stated
Domains	RoB, inconsistency, indirectness, imprecision, publication bias (+ upgrading)	RoB, residual confounding, similarity, ESS, imprecision — narrative
Movement between levels	Transparent downgrade/upgrade	No algorithm visible
Output	High/moderate/low/very low	No comparable final category (only narrative "very low" for networks)
Traceability	Reader can follow the derivation	Reader sees why it is uncertain, not how it becomes a degree of certainty

Consequence. It is operational, not merely presentational. A national appraisal body cannot act on an inventory of limitations alone; it must decide how much confidence to place in each estimate. Because the report treats significant direct effects and "very low" indirect effects with the same formal status, the decisive weighting is shifted to the reader, and the same evidence may support different conclusions across jurisdictions for interpretive rather than scientific reasons.

SECTION 08

Evidence-Based-Medicine Compliance (Point by Point)

> KEY MESSAGES
>
> **EBM's core preference is satisfied in the direct comparisons.** Randomised evidence is exactly what the trial provides.
>
> **Selective application bites at the edges.** Flawed analyses are excluded rather than retained at low certainty and bounded.

The report invokes "international standards of evidence-based medicine." It imports the internal-validity-appraisal half while setting aside totality-of-evidence, certainty-grading and value-integration. Because a randomised trial exists, the divergence is narrower than in single-arm cases, but it remains the same divergence:

Point of tension	What the JCA report does	What EBM would do
Direct evidence	Uses randomised within-trial comparisons for PICOs 1, 4, and 6	Endorses this — randomised evidence is preferred
Flawed/uninterpretable analyses	Excludes (single-item QoL not presented; implausible subgroups omitted)	Retain at low certainty and bound the interpretation
Indirect estimates	Networks judged "severely violated"; PICO 7 deemed unreliable	Present with explicit low/very-low grade rather than a narrative dismissal
Certainty rating	Narrative factors affecting certainty; no reproducible instrument	Structured grade (e.g. GRADE summary of findings)
Patient values & expertise	Consulted and annexed; influence on PICOs / outcomes not shown	Integrate explicitly to select and weight outcomes

SECTION 09

Expert and Carer Input: Documented but Not Traced

KEY MESSAGES

Heard procedurally, not visibly integrated. Participation met Art. 8(6)/11(4), but no traceable record shows how input shaped PICOs, comparators or outcomes.

A carer described the role as "rather symbolic." The clearest possible signal of the participation-without-influence gap.

Article 8(6) and Article 11(4) incorporate patient and clinical-expert participation into the procedure. The report records that two carers and two clinical experts provided written input on the scope and the revised draft reports, and that the material is annexed to Appendix A.

Participant	Input and concerns raised
Carers (two)	Whether enough patients exist to populate all groups; distribution of brain-metastasis patients; a request for a plain-language overall summary across the seven PICOs; that quality-of-life statistics are hard to translate into daily life; that low-grade (grade 1–2) adverse events can impose a substantial long-term burden on quality of life; and that the patient-reviewer role currently feels "rather symbolic."
Clinical experts (two)	Poor outcomes after platinum failure; substantial metastatic and brain/liver burden; favourable baseline performance status in the trial, limiting generalisability; cytokine-release syndrome and neurological toxicity (ICANS) as signature toxicities of this drug class; and the practical difficulty of an inpatient step-dose regimen.

Most pointedly for compliance, the report states that expert input "was considered" before finalisation but does not trace which scope or outcome changes resulted from specific comments. The carer's request for a plain-language cross-PICO summary and a brain-metastasis cohort is recorded but not shown to have altered the report. Involvement is documented at the level of participation, not influence. This is the third leg of the EBM tripod — patient values and clinical expertise — present in the file but not demonstrably integrated. It is not a procedural breach; the opportunity was given and the input annexed. But it is a transparency limitation, and it compounds the mixed outcome set: outcomes the carer flagged as mattering — quality of life, long-term tolerability — are among those least firmly assessed.

SECTION 10

Implicit Judgement: The Scientific / Appraisal Frontier

KEY MESSAGES

On the right side of the line — but steering. The report issues no value judgement, yet its exclusions and risk-of-bias ratings serve as a pre-appraisal filter.

"Judgement-free" can only mean value-free. Scientific judgement is unavoidable; the frontier is crossed only by moving from "unreliable estimate" to "no clinical value".

The report respects the HTAR prohibition on value judgement: it issues no conclusions on clinical added value, benefit ranking, place in therapy or target population. Technically, however, it is not judgement-free. Concrete, appraisal-shaping decisions include the pervasive high-risk-of-bias ratings that follow from the open-label design; the characterisation of two networks as resting on a "severely violated" similarity assumption; the decision to treat the PICO 7 estimate as descriptive and unreliable; and the exclusion of single-item quality-of-life and implausible subgroup analyses. Each is defensible as a scientific-validity judgement; none crosses into appraisal. But together they function as a pre-appraisal filter that steers the national reader towards a cautious frame for four of the seven questions, notwithstanding the positive direct evidence. The frontier would be crossed

only by moving from “this estimate is unreliable” to “the medicine has limited clinical value.” The report stays on the scientific side of that line — but its choices are demonstrably consequential downstream.

SECTION 11

Legal Fragility: A Tiered Analysis

KEY MESSAGES

Attack the reasoning, not the science. The strongest grounds challenge the statement of reasons and the inconsistent treatment of uncertainty.

Annulment would likely fail. Wide scientific discretion; contestable standing; a robust randomised trial underpinning the science.

The exposure is narrower than in orphan cases. With randomised evidence in hand, the orphan-methodology thread largely dissolves.

Were the assessment challenged before the EU courts, the strength of the available grounds would vary. The technical reading favours arguments that attack the quality of reasoning over those that attack the science.

Tier	Argument	Strength
1	Inadequate statement of reasons (why a given uncertainty is decisive; why one indirect comparison is reported and another is withheld) — Art. 296 TFEU	Strong
1	Inconsistent treatment of uncertainty (no explicit threshold or framework)	Strong–moderate
2	Proportionality (expectations that a single trial should answer seven comparative questions)	Moderate
2	Failure to consider all relevant circumstances (comparator not isolable in the trial; therapy un-blindable)	Moderate
3	Manifest error regarding the indirect comparisons (similarity, ESS, PH assumption)	Weak
3	Assessors should have accepted the developer’s interpretation	Weak

Defence reading. Annulment would likely fail. EU bodies enjoy wide scientific discretion, and review is limited to manifest error, misuse of powers and procedural irregularity; there is no right to a particular methodology; procedural defects must be outcome-relevant; and the report may not be a challengeable binding act, since it does not determine price, reimbursement or market access. Because this case rests on a randomised trial with significant direct effects, the orphan-methodology theme that dominated the single-arm cases is materially weaker; the residual exposure is concentrated in the reasoning surrounding the indirect comparisons and the certainty gap.

SECTION 12

Technical Learnings and Recommendations

KEY MESSAGES

Calibrate the scope to the trial. Screen each PICO for feasibility so that a single trial is not asked to answer seven questions.

Synchronise the dossier body with the analyses assessed. Prevent the assessed/presented divergence.

Make certainty and disputes auditable. Add a structured uncertainty-to-certainty step and a methodological-disagreement channel.

Recommendation	What it does
Evidence-feasibility screen at scoping	Test each candidate PICO for clinical relevance, evidence feasibility, data availability and outcome compatibility, so the scope aligns with what the pivotal trial can support.
Comparator-arm alignment via JSC	Use Joint Scientific Consultation to align the trial comparator arm with the anticipated PICO scope, thereby reducing reliance on fragile indirect methods.
Dossier-body synchronisation rule	Require the main dossier tables and graphs to depict the analyses actually assessed, not superseded figures held solely in appendices.
Uncertainty-to-certainty framework	Convert the documented factors affecting certainty into an explicit, reproducible measure of certainty.
Structural-limitation rule	A stated rule for weighing evidence where imperfections are structural (unblindable therapy; comparator not isolable), so positive randomised trials are not read down for unavoidable features.
Methodological-disagreement channel	A published category for disputes between fact and interpretation, with assessor, co-assessor and expert responses.

SECTION 13

Bottom-Line Technical Verdict

Dimension	Verdict	Basis
Procedural/structural (assessors)	Compliant	Template, scope, stakeholder involvement, and transparency all met
Methodological conduct (assessors)	Compliant/exemplary	Independent verification, recomputation, guidance-aligned exclusions
Outcome — direct comparisons	Conclusive	3 of 7 PICOs; significant OS from randomised evidence (HR 0.491, 0.628, 0.578)
Outcome — indirect comparisons	Weak	4 of 7 PICOs; “severely violated” networks, unknown-direction bias, one unreliable PFS estimate
Certainty reporting	Gap	Uncertainty documented; no graded certainty conclusion
EBM alignment	Selective	Internal validity half applied; grading and value integration set aside
Expert/carer input	Documented, not traced	Participation per Art. 8(6)/11(4); influence on PICOs and outcomes not shown; role described as “symbolic”
Legal robustness	Defensible	Strongest challenge targets reasoning; science underpinned by a randomised trial.
Developer dossier	Mixed	Strong pivotal RCT; fragile indirect comparisons; one requested analysis not provided; main tables not synchronised

In sum. The tarlatamab JCA is procedurally compliant and methodologically defensible, and the assessment team's conduct is rigorous and largely exemplary. It is conclusive for the three direct comparisons against topotecan, which show significant overall-survival benefits, and only weakly informative for the four indirect comparisons, which rest on violated similarity assumptions and unknown-direction bias. The principal gap in the report itself is the absence of an explicit, reproducible link from documented uncertainty to a stated degree of certainty, a gap that persists, tellingly, even though the evidence base here is a randomised trial rather than a single-arm study. A second, equally technical point is structural: the four indirect comparisons were, in effect, required by a scope the guidance will predictably reject, so the framework should add a feasibility screen at scoping, a developer's right to decline an infeasible indirect comparison, and an explicit decision on whether indirect comparisons belong in the joint layer at all.

CLEVER ACCESS
Let's be clever together

ANALYTICAL REVIEW

Clever-Access HTA Perspectives

Is a series of strategic reviews and expert analyses exploring the evolving landscape of European Health Technology Assessment, evidence generation, market access and healthcare policy.

DISCUSS WITH CLEVER-ACCESS ABOUT JCA AND/OR JSC, CONTACT:

Aleksandra Caban
aleksandra.caban@clever-access.com

FOR ANY QUESTIONS ABOUT THIS REPORT, CONTACT:

Lylia Chachoua
lylia.chachoua@clever-access.com

Interested in EU HTA, Market Access and Evidence Strategy?

Join the **Market Access Society** community for expert discussions, training programmes, webinars and practical resources designed for professionals working across the healthcare ecosystem.

SCAN CODE QR TO JOIN:

MAS COMMUNITY

FOR ANY QUESTIONS ABOUT THIS REPORT, CONTACT:

Mondher Toumi
mto@inovintell.com

VISIT OUR WEBSITE:

www.marketaccesssociety.org

Published by MAS