

Electronic Persistent Pain Outcomes Collaboration

ADULT BENCHMARKS REVIEW

FINAL REPORT

*Review of ePPOC Adult Benchmark
Domains, Targets, and Methods*

March 2026

Prepared by Brigitte Phillips & Meredith Bryce
Endorsed by the ePPOC Clinical Management
Advisory Committee (CMAC)

ePPOC 
electronic persistent pain
outcomes collaboration

ACKNOWLEDGEMENTS

We acknowledge the Traditional Custodians of the lands on which the University of Wollongong is situated. We pay our respects to Aboriginal Elders past and present, who are the knowledge holders and teachers. We acknowledge their continued spiritual and cultural connection to Country. As we share knowledge, teaching, learning and research within this University we also pay respect to the knowledge embedded forever within the Aboriginal Custodianship of Country.

We also want to acknowledge the valuable contribution made by:

- The many staff from pain management services who have spent considerable time collecting, collating and correcting the data and without whose effort this report would not be possible
- Members of the ePPOC Clinical and Management Advisory Committee

DISCLAIMER

ePPOC has made every effort to ensure that the data used in this report are accurate. Data submitted to ePPOC are checked for anomalies and services asked to resubmit information where relevant.

COPYRIGHT

This work is copyright. It may be produced in whole or in part for study or training purposes subject to the inclusion of an acknowledgement of the source. It is not for commercial use or sale. Reproduction for purposes other than those above requires the written permission of ePPOC.

SUGGESTED CITATION

Phillips, B & Bryce, M. (2026). *Electronic Persistent Pain Outcomes Collaboration: Adult Benchmarks Review – Final Report*. Faculty of Science, Medicine and Health, University of Wollongong.

CONTENTS

| | |
|---|----|
| 1. EXECUTIVE SUMMARY | 3 |
| 2. INTRODUCTION | 3 |
| 3. STAKEHOLDER PERSPECTIVES ON EPPOC BENCHMARKING | 4 |
| Theme 1. Clarity: Benchmarks Must Be Unambiguous in Intent, Language and Interpretation | 4 |
| Theme 2. Fairness: Equal Expectations, Fair Comparisons..... | 5 |
| Theme 3. Achievability: Benchmark Targets Should Motivate Improvement..... | 6 |
| Theme 4. Purpose and Use: Benchmarking Serves Multiple Audiences with Different Needs | 7 |
| Theme 5. Relevance: Benchmark Domains Must Reflect What Matters to Clinicians and Consumers | 8 |
| 4. RECOMMENDATIONS & ACTIONS | 9 |
| ePPOC Benchmark Setting Criteria and Protocol | 9 |
| Benchmark Targets and Definitions of Improvement..... | 10 |
| Benchmark Domains: Retention, Refinement, and Future Direction | 12 |
| Governance, Reporting, and Appropriate Use of Benchmarking..... | 13 |
| 5. CONCLUSION | 13 |
| APPENDIX A: ePPOC BENCHMARKING CRITERIA | 15 |
| APPENDIX B: ASSESSMENT TOOLS | 16 |
| Brief Pain Inventory | 16 |
| Depression Anxiety Stress Scales..... | 17 |
| Pain Catastrophising Scale | 17 |
| Pain Self-Efficacy Questionnaire | 18 |
| APPENDIX C: ePPOC ADULT BENCHMARKS 2025 INTERIM REVIEW SUMMARY | 19 |

1. EXECUTIVE SUMMARY

Benchmarking supports continuous quality improvement by enabling services to compare outcomes against shared reference points. The electronic Persistent Pain Outcomes Collaboration (ePPOC) benchmarks were originally developed through a collaborative, consensus-based process to identify core outcome domains and set aspirational targets for adult pain services.

This review was conducted in two phases. The first phase focused on analysis of service-level and patient-level performance against the existing benchmarks. These findings were presented in the interim report released in October 2025, which identified areas of variable performance and highlighted limitations in the transparency and interpretability of some benchmark definitions and targets. In response, a second phase was undertaken to review the underlying foundations of ePPOC benchmarking, including benchmark-setting criteria, definitions of clinically significant improvement, target-setting processes, and domain relevance.

This final report synthesises feedback from ePPOC member services gathered during the Australasian Benchmarking Workshop (12 November 2025) and from governance-level representatives at the ePPOC Clinical and Management Advisory Committee (CMAC) meeting (13 November 2025). Its purpose is to ensure that ePPOC benchmark domains, assessment tools, and targets remain clinically meaningful, consumer-relevant, methodologically robust, and aligned with contemporary models of pain care.

Overall, the review confirms that ePPOC benchmarking remains a credible and valuable quality improvement mechanism. However, it also identifies the need for targeted refinement to strengthen clarity, fairness, achievability and relevance, and to ensure benchmarks continue to motivate engagement and meaningful improvement across a diverse and evolving service landscape.

Key messages

- ePPOC benchmarking remains robust and fit for purpose but requires improved clarity and refinement in benchmark definitions and target-setting processes. Benchmarks should be consistently framed as aspirational, improvement-focused indicators, and interpreted within context rather than as ranking or judgement tools.
- Consumer relevance has been formally endorsed as a core benchmark-setting criterion.
- Progress toward case-mix adjusted benchmarking is a priority to support fair and credible comparisons across heterogeneous service populations.
- Most benchmark domains remain appropriate; however, several require refinement or further consideration. These include continued monitoring of anxiety outcomes, reassessment of average pain as a benchmark measure, exploration of alternative approaches to measuring psychological wellbeing, and consideration of function- and sleep-focused domains identified as priorities.

2. INTRODUCTION

Benchmarking is a widely used method for driving continuous quality improvement across healthcare systems. At its core, benchmarking involves comparing outcomes against shared reference points to support reflection, learning and improvement, rather than judgement or ranking. When combined with complementary strategies such as audit and feedback and peer learning, benchmarking can help services understand variation in outcomes and identify opportunities for improvement.

ePPOC has established a set of adult service benchmarks to support quality improvement across participating pain services. These benchmarks were developed through a collaborative, consensus-based process involving clinicians, services, and governance representatives. They were designed to represent clinically meaningful outcomes while remaining feasible to measure across diverse service models. The benchmark targets are intentionally aspirational, guided by established criteria (Appendix A) that emphasise alignment with “good

practice” rather than average performance, with the aim of encouraging improvement rather than defining minimum standards of care. The assessment tools that underpin benchmark outcomes are summarised in Appendix B.

This review was undertaken in two distinct but complementary phases. The first phase, reported in the interim benchmarking report released in October 2025, focused on analysing recent performance against the existing ePPOC benchmarks at both service and patient levels. A summary of the interim review is presented in Appendix C. That analysis highlighted several important issues that warranted further consideration. In particular, it identified substantial variability in how achievable different benchmarks appear to be, with some domains consistently met by a relatively high proportion of services and others, most notably anxiety and average pain, remaining persistently difficult to achieve. These patterns raised questions about whether the benchmark targets, the definitions of clinically significant improvement, and the underlying measures are functioning as intended as quality improvement tools.

The interim review also highlighted potential limitations in the transparency and interpretability of the benchmarking framework. Differences in benchmark attainability prompted discussion about how target levels are set and reviewed over time, and whether current definitions of improvement may underestimate meaningful clinical change or introduce unintended bias based on baseline severity. Collectively, these findings suggested that while services are achieving meaningful improvements for many patients, existing benchmarks may not always reflect those improvements in a way that is fair, easily interpretable, or motivating.

The second phase of the review, presented in this final report, shifted focus from performance itself to the foundations of ePPOC benchmarking. Rather than asking whether services are meeting benchmarks, this phase asks a more fundamental question: *are we benchmarking the right things, in the right way?* Specifically, it examines whether benchmark criteria, target-setting processes, definitions of clinically significant improvement, and selected outcome domains remain appropriate, evidence-based, and aligned with contemporary clinical practice and patient priorities. The findings and recommendations presented in this report have been reviewed and endorsed by the ePPOC Clinical and Management Advisory Committee (CMAC), providing governance-level approval of the proposed refinements to the benchmarking framework.

This phase of the review draws on structured feedback from ePPOC member services gathered during the Australasian Benchmarking Workshop (12 November 2025), as well as governance-level discussion at the ePPOC CMAC meeting (13 November 2025). While clinician perspectives dominated these discussions, participants consistently emphasised the importance of incorporating consumer perspectives more explicitly into future benchmarking decisions, a need that is acknowledged throughout this report.

In summary, while ePPOC benchmarking is designed to provide a simple and accessible signal of service-level outcomes, that simplicity relies on a robust and transparent methodological foundation. This report focuses on strengthening that foundation to ensure benchmarking continues to support meaningful, fair, and motivating quality improvement across the ePPOC network.

3. STAKEHOLDER PERSPECTIVES ON EPPOC BENCHMARKING

This section synthesises stakeholder perspectives gathered during the ePPOC Australasian Benchmarking Workshop (12 November 2025) and the ePPOC CMAC meeting (13 November 2025). These sessions built on the interim benchmark review and used a combination of facilitated discussion and live polling to test assumptions, surface areas of agreement and tension, and prompt deeper reflection. The themes presented below capture the key issues raised across both forums.

Theme 1. Clarity: Benchmarks Must Be Unambiguous in Intent, Language and Interpretation

Across both the Benchmarking Workshop and CMAC discussions, participants consistently emphasised that the value of ePPOC benchmarking depends heavily on clarity—particularly around terminology, intent, and how

results are interpreted. In practice, a lack of clarity was seen to create confusion about what benchmark results are intended to represent, how they should be used, and what conclusions can reasonably be drawn from them.

Ambiguous language, most notably the use of the term “good practice,” was identified as a key source of this confusion. Although benchmarks were intended to align with the original criterion (Appendix A) that thresholds reflect “good practice” rather than average performance, participants noted that this intent is not always explicit or consistently understood. This issue is compounded by the fact that some benchmark domains are inherently more difficult to achieve than others yet are presented using the same language and framing. As a result, when benchmarks are not achieved, services may perceive this as a failure of care, rather than recognising the influence of case complexity, domain sensitivity, measurement constraints, or the intended framing of benchmarking as an opportunity for improvement. This unintended interpretation risks discouraging services, despite benchmark thresholds being intentionally aspirational and not designed to imply deficit or blame.

Participants expressed concern that this lack of clarity increases the risk of misinterpretation at multiple levels, particularly when benchmark outputs are viewed by audiences outside the participating clinic staff, such as executives or funders. Without sufficient contextual explanation, benchmarking results may be over-simplified, misused, or inappropriately applied to performance management or funding decisions.

In response, there was strong agreement that benchmarking outputs should more clearly signal what benchmarks are, and are not, intended to convey. Participants emphasised that this framing must be consistently and explicitly embedded across all benchmark documentation, reporting, and supporting materials. For example, service-level reports that present benchmark results alongside explanatory narrative, trend data, and contextual caveats were viewed as more meaningful and less prone to misinterpretation than simple binary pass/fail indicators.

Participants also recognised that benchmarking is intended to function as a high-level, accessible quality improvement mechanism that is quick to understand and practical to use. However, this surface simplicity relies on a complex methodological foundation. Transparency about how benchmarks are developed, reviewed, and governed was therefore seen as critical to maintaining trust and credibility. Processes such as this review, and the active involvement of participating services, consumers, and collaborators, were viewed as essential to demonstrating that benchmarking is both robust and responsive. The consensus view was that benchmarking should appear simple on the surface, while being underpinned by a transparent and defensible foundation.

Theme 2. Fairness: Equal Expectations, Fair Comparisons

Early polling during the Benchmarking Workshop indicated a lack of shared confidence in the fairness and motivational value of current ePPOC benchmarks. When asked whether “the current ePPOC benchmark targets are fair, meaningful, and motivating for services,” 73% of respondents indicated neutral or disagreement (n=40), suggesting that benchmarks are not uniformly experienced as equitable or credible across participating services.

Workshop and CMAC discussions indicated that perceptions of unfairness primarily stem from differences in case-mix, clinical context, and service configuration. Participants pointed to the diversification of participating ePPOC services—including increasing numbers of compensation-focused clinics and the emergence of pelvic pain clinics, as well as variation in service size and staffing, and regional–metropolitan differences—as creating unequal conditions for meeting benchmarks. Regardless of whether these factors directly influence outcomes, their perceived impact was seen to undermine confidence in the legitimacy of benchmark comparisons.

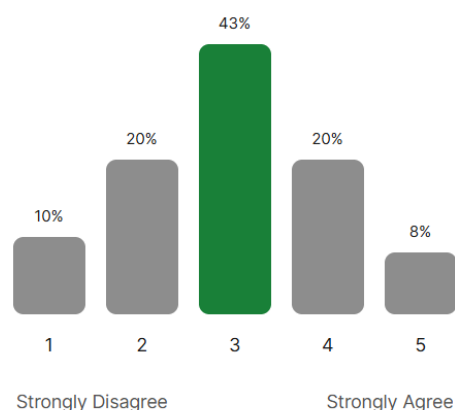


Figure 1 Benchmarking workshop participants responses to the statement ‘the current ePPOC benchmark targets are fair, meaningful, and motivating for services’ (n=40).

Importantly, these concerns were not framed as a desire for lower standards or exemption from benchmarking. Rather, participants challenged the validity of comparing services without accounting for patient-level differences that influence outcomes. While some interest was expressed in benchmarking similar services together, both the Benchmarking Workshop and CMAC cautioned against service-type stratification. CMAC emphasised that patients reasonably expect consistent standards of care regardless of where they are treated, and that stratifying benchmarks based on service type risks diluting expectations and introducing new inequities.

Discussions further distinguished between factors that warrant adjustment and those that benchmarking is intended to highlight. Differences in service resourcing, staffing, or workforce composition were viewed as legitimate contributors to benchmark performance that should remain visible. Benchmarking was seen as a mechanism to highlight potential system and service-level improvement needs, rather than obscure them. Alternatively, case-mix factors, including baseline clinical complexity, were widely accepted as appropriate for adjustment. Advances in statistical methods make this increasingly feasible, with early efforts recommended to focus on domains such as pain severity, which are strongly influenced by clinical complexity and are poorly suited to unadjusted, uniform benchmarking. There was strong consensus that case-mix adjustment represents the most appropriate long-term strategy for improving fairness while preserving common expectations.

Overall, stakeholders viewed fairness in benchmarking as producing comparisons that are credible, meaningful, and trusted by patients, clinicians, and services. Benchmarking should maintain consistent expectations of care quality and outcomes across all service types, while accounting for differences in patient mix, including baseline clinical complexity. This aligns with Social Comparison Theory (Festinger, 1954), which suggests that benchmarking supports reflection and learning only when comparisons are perceived as fair, relevant, and potentially achievable.

Theme 3. Achievability: Benchmark Targets Should Motivate Improvement

Discussions across the Benchmarking Workshop and CMAC consistently highlighted achievability as a critical determinant of whether benchmarking functions as a meaningful quality improvement tool. Participants noted that while benchmarks are intended to be aspirational, some targets are perceived as beyond reach for many services, particularly when presented through rigid thresholds and binary “meets/does not meet” framings. Substantial variation in the proportion of services meeting different benchmarks, for example, approximately 10% for depression compared with 36% for stress, reinforced perceptions that some benchmarks may be inherently more difficult to attain than others.

When benchmarks are perceived as unattainable or poorly calibrated, their motivational effect is diminished. Services reported that absolute targets can feel demoralising when sustained improvement is occurring but remains unrecognised because a fixed threshold has not been reached. This was seen to undermine engagement, reduce confidence in benchmarking, and weaken its role as a learning-oriented feedback mechanism. From a motivational perspective, participants noted that engagement is most likely when targets are perceived as understandable and realistically challenging.

This concern prompted deeper discussion about how benchmark targets are set and reviewed. Participants explored a range of approaches, including statistical thresholds (such as top quartile performance or targets based on average outcomes), consensus-defined targets, and targets anchored to each service’s historical performance. While data-driven approaches to target setting (such as averages or percentiles) were acknowledged as appealing in principle, they were also recognised as potentially unstable over time and less suited to live reporting environments. In contrast, the existing consensus-based approach was viewed as better able to balance clinical judgement and aspirational intent. Although no single method achieved full agreement, there was broad support for retaining the consensus approach as the foundation for target-setting, alongside improved transparency, consultation, and explanation of how targets are derived. Participants also expressed interest in complementing fixed targets with more explicit performance- or improvement-focused metrics that highlight trajectories over time.

Achievability concerns were further amplified when discussion turned to the definitions of clinically significant improvement. For several benchmark domains—specifically pain catastrophising, pain self-efficacy, depression, anxiety, and stress—clinically significant improvement is currently defined as requiring both a minimum point

change and a shift to a lower severity category. Participants identified the categorical severity shift requirement as problematic, noting that severity categories are based on arbitrary thresholds and that crossing (or not crossing) a category boundary does not necessarily reflect a meaningful difference in the magnitude of change. As a result, equivalent improvements can be classified differently purely because they fall on either side of a category cut-off. Polling of workshop participants indicated strong support for removing the categorical shift requirement (Figure 2).

A related but distinct concern was raised regarding the use of absolute point change as the sole indicator of improvement. Participants noted that the same absolute change can represent very different proportional improvements depending on a patient’s baseline severity. For example, a five-point reduction may constitute a modest relative improvement for a patient starting at the severe end of a scale, but a substantially larger proportional change for someone starting at a lower baseline. This raised questions about whether improvement should be defined purely in absolute terms, or whether relative (percentage-based) change from baseline may better capture clinically meaningful progress in some domains.

Overall, participants emphasised that achievability is not simply about where targets are set, but about how improvement is defined, recognised, and communicated. Benchmarks that acknowledge progress, support learning, and reflect clinically meaningful change were seen as more likely to sustain engagement and drive improvement, reinforcing the need for refinements to both target-setting processes and definitions of clinically significant improvement.

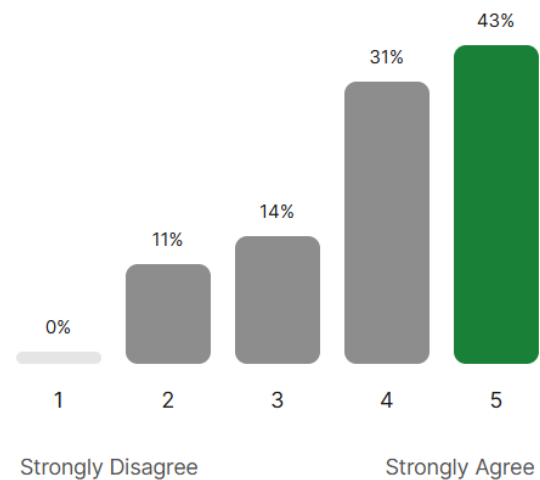


Figure 2 Benchmarking workshop participants responses to the statement ‘ePPOC should remove the required categorical shift from the definitions of clinically significant improvement’ (n=35).

Theme 4. Purpose and Use: Benchmarking Serves Multiple Audiences with Different Needs

Discussions highlighted that ePPOC benchmarking is used by multiple audiences for different purposes, but that these uses are not always clearly distinguished. At the service level, benchmarks are primarily intended to support quality improvement, guide reflection on performance over time, and engage teams in learning. At governance and system levels, the same data are used to inform oversight, planning, and accountability. Participants also acknowledged that benchmarking data may be used externally for purposes such as marketing or broader communications, even when this is not the original intent of ePPOC benchmarking.

When the purpose and audience of benchmarking are unclear, the information provided may be poorly matched to its use. Participants noted that simple signals, such as binary pass/fail indicators, may be preferred for high-level reporting or system oversight, but can obscure important nuances and are inadequate alone for interpreting performance or guiding improvement. In these contexts, the absence of explanatory narrative, trend data, and contextual caveats limits interpretability and can lead to overly simplistic conclusions about service performance.

This mismatch was seen to create tangible risks. Participants expressed explicit concern that tying benchmarking too directly to funding or external accountability could incentivise gaming, cherry-picking of patients, or other behaviours that distort results. These risks were viewed as undermining both the credibility of benchmarking and its role as a learning-oriented quality improvement tool.

In response, stakeholders emphasised the need for transparent framing, clear caveats, and audience-appropriate reporting. Rather than prescribing a single “correct” use of benchmarking, there was agreement that the intended purpose, appropriate audience, and limitations of the data must be explicitly communicated wherever benchmark results are presented. When benchmarks are framed thoughtfully and interpreted within context, they can support improvement, learning, and accountability across system levels without distorting service behaviour or undermining trust.

Theme 5. Relevance: Benchmark Domains Must Reflect What Matters to Clinicians and Consumers

Discussions consistently emphasised that benchmarking is most effective when it focuses on outcomes that are meaningful to those delivering and receiving care. While the majority of workshop participants and CMAC representatives were clinicians, there was strong agreement that benchmark domains must also align with consumer priorities. It was further acknowledged that funders and policymakers may have additional information needs, reinforcing the importance of selecting domains that can serve multiple audiences without diluting clinical or consumer relevance. CMAC explicitly endorsed consumer relevance as a foundational principle for benchmarking and supported the addition of a new criterion—that benchmarks should be meaningful and relevant to consumers—to formally embed this expectation.

Concerns about relevance prompted discussion about whether the current benchmark set adequately reflects outcomes that matter most in real-world care. To explore perceived gaps, participants were invited to provide free-text responses identifying additional outcomes they believed should be considered for benchmarking. Responses were reviewed and consolidated prior to visualisation to address spelling variants, closely related constructs, and overlapping concepts (e.g. “sleep” and “sleep quality”). Original response frequencies were summed within each consolidated domain to preserve the relative emphasis of participant input. The resulting word cloud therefore reflects thematic priority rather than verbatim text frequency.



Figure 3 Word cloud summarising additional outcome domains proposed by participants during the ePPOC Benchmarking Workshop and CMAC discussions. Free-text responses were consolidated to address spelling variants and conceptually overlapping terms prior to visualisation. Word size reflects the relative frequency with which domains were nominated, indicating thematic priority rather than verbatim text frequency.

In contrast, average pain was consistently identified as the weakest and least trusted benchmark domain. Concerns were raised about poor interpretability for patients, lower completion rates, reliance on a single-item measure rather than a multidimensional domain score, limited clinical usefulness, and low benchmark attainability. While pain severity remains a core construct in pain care and cannot be removed entirely from benchmarking, average pain was widely viewed as provisional and not suitable as a long-term benchmark in its current form.

Similarly, the appropriateness of psychological distress domains measured using the Depression Anxiety Stress Scales (DASS), generated complex discussion. While the DASS was widely acknowledged as clinically useful—

particularly for identifying severity, informing triage, and guiding treatment pathways—concerns were raised about its suitability as a benchmarking instrument. Consumer feedback was presented, which has highlighted that the DASS can feel confronting or distressing, with its strong focus on negative emotional states, and that it may be triggering for some patients, particularly those with prior trauma or severe distress. Clinicians echoed concerns about burden and repetition, noting the length of the measure and the emotional load of repeated administration.

At the same time, CMAC discussion made clear that removing the DASS outright would be premature and potentially harmful. Many services rely on DASS subscale scores to support clinical decision-making, duty of care, and risk identification, particularly in settings where access to comprehensive psychological assessment is limited. There was also recognition that declining performance in anxiety outcomes warrants careful monitoring, and that removing the measure during a period of apparent deterioration could undermine transparency. As such, the DASS was viewed as clinically valuable but imperfect, with agreement that its continued use should be reviewed as part of broader dataset evolution (including Version 3), rather than addressed in isolation within benchmarking.

Overall, relevance was framed as central to credibility and engagement. Participants cautioned against unrestrained expansion of benchmark domains, expressing a clear preference for substitution, refinement, or consolidation rather than simple addition, due to concerns that an ever-growing benchmark set would increase burden without improving interpretability or impact. Benchmarks that reflect outcomes that matter to patients and clinicians were viewed as more likely to support meaningful reflection and improvement, whereas domains perceived as clinically weak or disconnected from lived experience risk disengagement, regardless of methodological sophistication.

4. RECOMMENDATIONS & ACTIONS

This section translates the key themes arising from the Australasian Benchmarking Workshop and CMAC discussions into concrete recommendations and actions. Together, these actions are intended to strengthen the foundations of ePPOC benchmarking by improving clarity, fairness, achievability, relevance, and the appropriate use of benchmark data. The recommendations are aligned with the structure of the workshop discussions and are organised across benchmark-setting criteria, target definition, benchmark domains, and cross-cutting governance and reporting considerations.

ePPOC Benchmark Setting Criteria and Protocol

The ePPOC benchmark-setting criteria (Appendix A) provide the foundational principles that guide which domains are benchmarked and how benchmark targets are established. Six of the original seven core criteria were endorsed by CMAC in May 2016 and have underpinned benchmark development since that time. Prior to this review, however, these criteria had not been formally revisited, despite substantial evolution in the service landscape, data maturity, and expectations of benchmarking.

As part of the interim benchmark review, participants at the Australasian Benchmarking Workshop were asked to consider and rank the importance of the existing criteria. This was followed by a formal re-voting process at the subsequent CMAC meeting, enabling governance-level representatives to review, refine, and update the criteria in light of contemporary practice and stakeholder feedback. This process resulted in several important clarifications and amendments that are detailed in Appendix A.

There was unanimous endorsement of the core clinical and methodological principles underpinning benchmarking. CMAC reaffirmed that benchmark domains must be clinically important, amenable to intervention, supported by sufficient data, and framed against aspirational targets rather than average performance. In particular, the criterion stating that benchmarks should reflect “*good practice*” was formally reworded to clarify that benchmark thresholds represent aspirational targets intended to reflect high performance, rather than minimum acceptable or average practice. This change directly addresses concerns

raised throughout the review regarding clarity of intent, achievability, and the motivational purpose of benchmarking.

Two criteria were explicitly rejected following discussion. The requirement that benchmarks only include patients with at least moderate baseline severity did not achieve consensus and was rejected to preserve flexibility across domains and populations. Similarly, the criterion requiring sufficient variation between services performance was retained as rejected, reflecting the sentiment that there may be situations where all services are performing similarly poorly where benchmarking is warranted.

One of the most substantive areas of discussion related to whether benchmarks should be applied uniformly across all services. While this criterion generated considerable debate, it was ultimately endorsed with important qualification. CMAC agreed that benchmark targets should be consistent across services, regardless of public or private status, role, or configuration, in order to preserve shared expectations of quality and outcomes. This endorsement was explicitly coupled with recognition that benchmarks must take account of patient-level variability. Accordingly, there was a clear commitment to progress towards case-mix adjusted benchmarking, rather than adjusting benchmarks based on service characteristics such as resourcing, staffing levels, or workforce composition.

Importantly, CMAC also unanimously endorsed a new criterion: benchmarks should be meaningful and relevant to consumers. This addition reflects strong alignment between clinician and consumer perspectives expressed throughout the review process and formally embeds consumer relevance as a core benchmark principle alongside clinical and methodological considerations.

Actions

- Adopt the revised benchmark-setting criteria endorsed by CMAC, including the new consumer relevance criterion, as the formal framework for future benchmark development and review.
- Clearly document and communicate the benchmark-setting criteria, including the rationale for accepted and rejected criteria, to improve transparency and shared understanding among participating services.
- Progress the development and piloting of case-mix adjustment methods for relevant domains to support fair comparisons across heterogeneous patient populations while maintaining consistent benchmark targets.
- Apply domain-specific judgement when considering baseline severity thresholds, rather than enforcing a uniform rule across all benchmarks.
- Establish a mechanism for periodic review of benchmark-setting criteria to ensure ongoing alignment with evolving evidence, data quality, and stakeholder expectations.

Benchmark Targets and Definitions of Improvement

Defining ePPOC benchmark targets requires consideration of two interrelated elements:

1. how the benchmark target level is set, and
2. how clinically significant improvement is defined for each domain.

Benchmark target level

There was extensive discussion regarding empirical approaches to target setting (such as averages or percentile-based thresholds) compared with consensus-defined targets. While empirical approaches were recognised as appealing due to their apparent objectivity, participants noted that percentile-based targets are inherently unstable and may shift over time as the dataset evolves, particularly within live reporting environments. This instability was viewed as potentially undermining interpretability and confidence in benchmark targets.

In contrast, the existing consensus-based approach was widely supported as better able to balance clinical judgement and consumer relevance. Participants emphasised that benchmarks should function as stretch

targets—challenging but potentially attainable—and should motivate improvement rather than define minimum standards of care. Although no single approach achieved unanimous agreement, there was broad support for retaining the consensus-based approach as the foundation for target setting, alongside improved transparency and clearer articulation of how targets are derived and reviewed. Participants also expressed interest in complementing fixed benchmark targets with additional performance- or improvement-focused metrics that highlight trajectories over time.

Definitions of clinically significant improvement

Strong consensus emerged that current definitions of clinically significant improvement do not consistently reflect meaningful clinical progress. For several benchmark domains—including pain catastrophising, pain self-efficacy, depression, anxiety, and stress—clinically significant improvement is currently defined as requiring both a minimum point change and a categorical severity shift. Participants identified the categorical severity shift requirement as problematic, noting that severity categories are based on arbitrary cut-points and that crossing (or not crossing) a category boundary does not necessarily reflect a meaningful difference in the magnitude of change. As a result, equivalent improvements may be classified differently solely due to their position relative to category thresholds.

A related but distinct concern was raised regarding reliance on absolute point change alone. Participants noted that the same absolute change can represent very different proportional improvements depending on baseline severity, raising questions about whether improvement should always be defined in absolute terms or whether relative (percentage-based) change may better capture meaningful progress in some domains or populations.

Polling during the Benchmarking Workshop demonstrated strong support for removing the categorical severity shift requirement. CMAC confirmed that revisiting definitions of clinically significant improvement is legitimate and aligned with the intent of benchmarking as an actionable quality improvement tool.

Actions

- Maintain a consensus-based approach as the primary method for setting ePPOC benchmark target levels.
- Clearly articulate benchmark targets as aspirational, developmental goals that support quality improvement, rather than indicators of “good” or “best” practice, and ensure consistent terminology across reports, dashboards, training materials, and communications.
- Improve transparency by documenting how each benchmark target is derived, including the rationale, evidence base, and consultation processes involved.
- Implement a regular benchmark target review process, led by the ePPOC team and informed by CMAC, with benchmarks flagged for review where more than 40% of services meet the target for two or more consecutive reporting periods.
- Remove the categorical severity shift requirement from definitions of clinically significant improvement.
- Refine definitions of clinically significant improvement to better reflect meaningful clinical change, accounting for variation across domains, baseline severity, patient populations, and established concepts such as minimal clinically important difference (MCID). Support this work with targeted consumer consultation to inform future threshold setting.
- Trial staged implementation of revised definitions to ensure changes are interpretable, actionable, and supportive of quality improvement.
- Explore complementary performance-focused metrics that emphasise improvement trajectories and continuous change, rather than reliance on binary outcomes alone.

Benchmark Domains: Retention, Refinement, and Future Direction

Discussions during the Benchmarking Workshop and CMAC meeting highlighted variable confidence in the current benchmark domains, particularly in relation to whether they reflect outcomes that matter most to patients and clinicians and support meaningful quality improvement. Domain-specific discussion focused primarily on two areas of concern: the assessment of pain severity (currently benchmarked using the average pain item) and the continued use of the Depression, Anxiety and Stress Scales (DASS) within the benchmark set.

While there was broad agreement that pain severity remains a core construct in pain services and should not be removed entirely from benchmarking, there was strong and consistent concern regarding the suitability of average pain as a benchmark domain. Participants described average pain as difficult for patients to interpret, associated with lower completion rates, methodologically weak due to its reliance on a single item, and of limited clinical usefulness at the service level. As a result, average pain was widely viewed as a provisional benchmark that is not fit for long-term use in its current form.

Discussion therefore shifted from whether pain severity should be benchmarked to how it should be assessed. CMAC agreed that a composite or summary approach to pain severity would better reflect meaningful change. Specifically, CMAC endorsed moving forward with a summary pain severity score combining average and worst pain as a pragmatic step toward improving interpretability while remaining within the existing dataset. These domains were prioritised as they provide the most stable and clinically meaningful indicators of overall pain burden and peak severity, whereas “right now” and “least” pain are more susceptible to short-term fluctuation and less informative for benchmarking purposes. From a patient perspective, participants also noted that improvements in overall and peak pain are generally prioritised over changes in momentary or minimum pain levels. The intent of the summary score is therefore not to incorporate all available pain ratings, but to prioritise domains that provide a consistent, clinically meaningful signal of improvement while minimising noise introduced by momentary variation.

Concerns were also raised regarding the continued use of the DASS as a benchmarked domain. While participants acknowledged the clinical utility of the DASS—particularly for identifying psychological distress and informing treatment pathways—there was recognition that it can be burdensome or confronting for some patients, and that its strong deficit focus may not align well with broader concepts of wellbeing or recovery. Importantly, there was no consensus to remove the DASS at this time, particularly given its role in identifying declining anxiety outcomes and supporting clinical triage. Instead, discussion positioned the appropriateness of the DASS as a future review issue, closely linked to the Version 3 dataset revision and potential incorporation of complementary wellbeing-oriented measures.

In contrast to these areas of concern, strong and consistent support emerged for outcome domains related to function and participation, including physical function, sleep, work participation, and relational or social functioning. As discussed in Theme 5 (Relevance), these domains were viewed as closely aligned with consumer priorities and real-world clinical goals. Importantly, many of these outcomes are already captured within the existing ePPOC dataset, presenting opportunities to refocus benchmarking without expanding data collection burden.

Across discussions, participants cautioned against unrestrained expansion of benchmark domains. There was a clear preference for substitution, refinement, or consolidation rather than simple addition, reflecting concern that an ever-growing benchmark set would increase burden without improving interpretability or impact.

Actions

- Retain pain severity as a core construct within ePPOC benchmarking, while strengthening the current approach by implementing the CMAC-endorsed summary pain severity score combining average and worst pain. This transition should be supported by a structured evaluation of the impact on benchmark results and target levels, alongside consideration of consumer perspectives and the role of case-mix adjustment.

- Maintain the DASS benchmarks in the short term, while explicitly flagging the appropriateness of the DASS for future review as part of the Version 3 dataset revision, including consideration of complementary wellbeing-focused measures.
- Prioritise future benchmarking of function and participation-focused domains—including physical function, sleep, work participation, and relational or social functioning—where data quality and feasibility thresholds are met.
- Avoid expansion of the total number of benchmark domains; where new domains are prioritised, pursue substitution or consolidation rather than addition.
- Strengthen the role of consumer perspectives in domain selection through CMAC and targeted consumer consultation.

Governance, Reporting, and Appropriate Use of Benchmarking

Benchmarking should support learning and improvement while maintaining trust, clarity, and fairness. Across the workshop and CMAC discussions, participants emphasised that the credibility and impact of benchmarking depend not only on what is benchmarked, but on how benchmarks are governed, communicated, and used in practice.

There was strong support for making benchmark development and review processes more visible and predictable, including clearer articulation of roles, review cycles, and opportunities for stakeholder input. Participants also highlighted the importance of reporting formats that provide sufficient context, narrative, and caveats to support meaningful interpretation by different audiences. Finally, concerns regarding gaming, cherry-picking, and promotional misuse of benchmarking data underscored the need for clearer guidance on appropriate and inappropriate uses of ePPOC benchmarks, particularly in funding and external communication contexts.

Actions

- Establish and document a regular benchmark review cycle, with clear triggers for review and defined roles for the ePPOC team, CMAC, and other stakeholders.
- Clarify decision-making pathways for methodological changes, including how clinical, consumer, and service-level input is incorporated.
- Review benchmark reporting formats to increase the use of explanatory narrative, trend data, and contextual information, including clearly articulating the intended purpose, appropriate audience, and limitations of benchmarking within all reports and supporting materials.
- Align language consistently across documentation, reports, dashboards, and education materials to reinforce aspirational, improvement-focused framing. Reinforce that benchmarking is designed to support learning, reflection, and quality improvement, rather than judgement or ranking.

5. CONCLUSION

This review confirms that ePPOC benchmarking remains a credible and valued mechanism for supporting quality improvement across adult pain services. Stakeholder feedback reaffirmed the core intent of benchmarking as aspirational, improvement-focused, and grounded in peer comparison, while also identifying clear areas where greater clarity, fairness, and relevance are required to maintain trust and engagement.

Importantly, this review marks a shift from focusing solely on *whether services are meeting benchmarks* to examining *whether the benchmarks themselves are fit for purpose*. Through structured feedback from services and governance representatives, the review has clarified the meaning of “good practice” as aspirational high performance, strengthened and updated benchmark-setting criteria, endorsed refinement of definitions of clinically significant improvement, and embedded consumer relevance as a core benchmark principle.

The findings also reinforce that effective benchmarking depends not only on methodological soundness, but on transparent governance, clear communication, and appropriate use. Benchmarks are most useful when they are interpretable, contextualised, and framed as tools for reflection and learning rather than judgement, ranking, or funding decisions.

The recommendations included in this report have been endorsed by the ePPOC CMAC, and implementation will proceed with initial priorities focused on:

- applying the revised benchmark-setting criteria endorsed by CMAC, including the new consumer relevance criterion;
- refining definitions of clinically significant improvement, including removal of the currently required categorical severity shift for pain catastrophising, pain self-efficacy, depression, anxiety, and stress;
- implementing the agreed refinement to pain severity benchmarking through adoption of a summary score combining average and worst pain, replacing average pain as a standalone benchmark; and
- strengthening governance, reporting, and guidance on the appropriate interpretation and use of benchmarking outputs, including establishment of a documented and regular benchmark review cycle with clear triggers for review and defined roles for the ePPOC team, CMAC, and other stakeholders.

Looking ahead, progression toward case-mix adjusted benchmarking and consumer-informed refinement represent key areas for ongoing development.

Taken together, this review positions ePPOC benchmarking as a living framework that evolves alongside services, evidence, and consumer expectations. Through regular review, open dialogue, and shared learning, benchmarking will continue to support meaningful improvement in care and outcomes for people living with persistent pain.

APPENDIX A: ePPOC BENCHMARKING CRITERIA

The benchmarking criteria outlined in this appendix describe the principles and considerations used to guide the development and review of ePPOC benchmark targets. The criteria were originally endorsed in May 2016 by the ePPOC governance committee, which approved six core principles for benchmark development. As part of the current benchmarking review, the criteria were revisited through stakeholder consultation, resulting in the retention of some principles, the removal of others, and the addition of a new criterion. The table below presents the full set of criteria alongside the outcomes of the review and stakeholder perspectives. Together, these criteria provide a structured framework to support transparent and consistent decision-making, ensuring that benchmark targets remain clinically meaningful, feasible to measure across diverse service models, and oriented toward improvement rather than defining minimum standards of care.

Table 1 ePPOC benchmark-setting criteria, detailing importance rankings obtained from the Benchmarking Workshop (November 12, 2025), and corresponding CMAC endorsement decisions (November 13, 2025).

| 2016 Criteria | Importance Ranking (n=33) | CMAC Decision |
|--|---------------------------|--|
| 1 The domain is clinically important | 6.15 | Accepted by simple majority |
| 2 Evidence exists that the domain is amenable to intervention | 5.97 | Accepted by simple majority |
| 3 Benchmark threshold reflects good practice rather than average practice | 3.97 | Accepted by simple majority, with rewording to: The benchmark threshold reflects <i>an aspirational target intended to represent high performance</i> , rather than average practice |
| 4 Sufficient data exists to benchmark effectively | 3.33 | Accepted by simple majority |
| 5 Clinical benchmarks should include patients with at least moderate baseline severity | 2.64 | 50% agreement; rejected in favour of a flexible approach |
| 6 Benchmark is the same across all services regardless of public/private tier or role | 2.15 | Accepted by simple majority, with recognition of intention to move towards case-mix adjusted benchmarking. |
| 7 Sufficient variation exists between services and patients | 1.91 | Rejected |
| 8 Benchmarks should be meaningful and relevant to consumers | N/A* | Accepted by simple majority |

*This criterion was not included in the Benchmarking Workshop importance ranking, as it was introduced as a new criterion following workshop discussions and subsequently considered and endorsed by CMAC.

APPENDIX B: ASSESSMENT TOOLS

The assessment tools used in ePPOC are:

- Brief Pain Inventory (BPI)
- Depression, Anxiety, Stress Scale (DASS)
- Pain Catastrophising Scale (PCS)
- Pain Self-Efficacy Questionnaire (PSEQ)
- ePPOC Patient Impression of Change (formerly referred to as the Global Rating of Change (GRC))
- Work Productivity and Impairment (WPAI)
- CARRA Body Chart.

Assessment tools relevant to ePPOC benchmarking are briefly described below.

BRIEF PAIN INVENTORY¹

The BPI items used in the ePPOC dataset measure the severity of pain and the degree to which the pain interferes with common activities of daily living. Pain severity questions are rated on a scale of 0 to 10, where 0 = 'No pain' and 10 = 'Pain as bad as you can imagine', with patients asked to rate their pain in four items including, average, worst and least pain over the last week, and their pain right now. Pain severity is then calculated as an average of these four items. Whereas the benchmark for average pain is based on the single average pain item only.

The IMMPACT group's recommendations for assessing clinical significance for 0-10 numeric pain scales are that a change of:

≥ 10% represents minimally important change

≥ 30% represents moderate clinically important change

≥ 50% represents substantial clinically important change.

The interference questions are rated on a scale of 0 to 10, where 0 = 'Does not interfere' and 10 = 'Completely interferes'. The interference subscale is an average of the seven interference questions. At least 4 of 7 questions must be completed for this subscale to be valid. The IMMPACT recommendation for assessment of clinically significant change on the BPI interference scale is a change of 1 point over the average of the 7 items².

Updated BPI severity bands From the 2022 mid-year report onwards, ePPOC had implemented a modified version of the previously reported severity bands: 0 = none | 1-4 = mild | 5-6 = moderate | 7-10 = severe. A score of 0 – no pain (or 0 – no interference) was previously included in the mild category, but is now reported separately³.

The above severity bands can be applied directly to the individual items reported on the integer scales (e.g. the 'average pain' item). When applied to the sub-scale scores for pain severity and interference, the mean values are rounded to the closest integer before the severity ranges are applied. The 'none' category should be interpreted as 'no or negligible pain' or 'no or negligible interference' for the pain severity and pain interference subscales respectively.

¹ Modified Brief Pain Inventory, reproduced with acknowledgement of the Pain Research Group, University of Texas, MD Anderson Cancer Centre, USA

² Dworkin, RH, et al 2008, 'Interpreting the Clinical Importance of Treatment Outcomes in Chronic Pain Clinical Trials: IMMPACT Recommendations.' *The Journal of Pain*, vol. 9, no. 2, pp 105-121.

³ Palos, GR, et al 2006, 'Asking the community about cutpoints used to describe mild, moderate, and severe pain.' *Journal of Pain*, vol. 7, no. 1, pp 49-56

DEPRESSION ANXIETY STRESS SCALES⁴

The DASS measures the negative emotional states of depression, anxiety and stress. Due to the large number of questions in the full DASS (42 questions), the DASS21 (21 questions) is administered. The questions are rated on a scale of 0 to 3, where 0 = 'did not apply to me at all', 1 = 'applied to me to some degree, or some of the time', 2 = 'applied to me to a considerable degree, or a good part of the time', or 3 = 'applied to me very much, or most of the time'. Scores are multiplied by 2 to enable comparison with the full-scale DASS42 for which norms exist.

For each subscale (Depression, Anxiety and Stress), the 7 items are summed and then multiplied by 2. The test developers suggest that at least 6 of 7 items should be complete for each subscale to be considered valid. Table 2 shows the range of scores associated with severity categories for each subscale.

| Table 2 – DASS severity ratings | Depression | Anxiety | Stress |
|---------------------------------|------------|---------|--------|
| Normal | 0-9 | 0-7 | 0-14 |
| Mild | 10-13 | 8-9 | 15-18 |
| Moderate | 14-20 | 10-14 | 19-25 |
| Severe | 21-27 | 15-19 | 26-33 |
| Extremely Severe | 28+ | 20+ | 34+ |

Clinical significance on each of the DASS subscales requires a change of 5 or more points coupled with a move to a different severity category.

PAIN CATASTROPHISING SCALE⁵

The PCS measures a patient's thoughts and feelings related to their pain. This includes three subscales measuring the dimensions of Rumination, Magnification and Helplessness. The PCS comprises 13 questions (Rumination – 4 items, Magnification – 3 items, Helplessness – 6 items) which are rated on a scale of 0 to 4, where 0 = 'not at all', 1 = 'to a slight degree', 2 = 'to a moderate degree', 3 = 'to a great degree' and 4 = 'all the time'. For each subscale, all items must be completed to be valid. For the total to be valid, at least 12 of 13 items must be completed.

Severity bands for the PCS are:

- <20 = mild
- 20 to 30 = high
- >30 = severe.

Clinically significant change requires a change in score of 6 or more points, combined with movement to a different severity category⁶.

⁴ Lovibond, SH and Lovibond, PF 1995, *Manual for the Depression Anxiety Stress Scales*, Psychology Foundation Monograph, Sydney, Australia.

⁵ Sullivan, MJL, et al 1995, 'The Pain Catastrophizing Scale: Development and Validation', *Psychological Assessment*, vol. 7, num. 4, pp 524-532.

⁶ Sullivan, MJL, personal communication with Nicholas, MK July 2014.

PAIN SELF-EFFICACY QUESTIONNAIRE

The PSEQ measures how confident a patient is that he or she can do a range of activities despite their pain. The PSEQ Total is a sum of scores from 10 questions which are rated on a scale from 0 = 'Not confident at all' to 6 = 'Completely confident'. At least 9 of 10 items must be complete for the PSEQ Total to be valid. Increases in score represent an improvement in self-efficacy.

Severity bands for the PSEQ are:

- <20 = severe
- 20 to 30 = moderate
- 31 to 40 = mild
- >40 = minimal impairment.

Clinically significant change requires a change in score of 7 or more points, combined with movement to a different severity category⁷.

⁷ Nicholas, MK, personal communication, July 2014.

APPENDIX C: ePPOC ADULT BENCHMARKS 2025 INTERIM REVIEW SUMMARY

| Benchmark (BM) Domain | Assessment Tool | Definition of Clinically Significant Improvement | Current BM Target | Summary |
|-----------------------------|---|--|-------------------|---|
| Average Pain | Brief Pain Inventory (BPI) | ≥30% reduction from baseline for patients with moderate to severe pain | 40% | The ePPOC average pain benchmark is challenging for services to achieve. Due to the low valid completion rate of the item, and the difficulty patients have estimating ‘average’ pain, alternative measures—such as worst pain, pain frequency, or composite scores—may provide a more reliable reflection of treatment impact. Additionally, due to accumulating evidence that pain severity outcomes are influenced by baseline characteristics, benchmarking in the pain severity domain may require a more nuanced or case-mix adjusted approach. |
| Pain Interference | Brief Pain Inventory (BPI) | ≥30% reduction from baseline for patients with moderate to severe interference | 70% | Given its stable patient-level outcomes, high prevalence at episode start and strong valid completion rate, the pain interference measure remains robust, and no changes are proposed to its use as a core ePPOC benchmark. |
| Pain Catastrophising | Pain Catastrophising Scale (PCS) | ≥6-point reduction and shift to lower severity category | 70% | There has been a recent decline in pain catastrophising outcomes, which is evident at both the service- and patient-level. ePPOC will continue to monitor this trend. This review also recommends removing the required change in severity category from the current definition of clinically significant improvement for this benchmark. Simplifying the definition of improvement to a reduction of 6 or more points on the PCS scale, regardless of start and end category, may result in more consistent assessment. |
| Pain Self-Efficacy | Pain Self-Efficacy Questionnaire (PSEQ) | ≥7-point increase and shift to a less severe category | 60% | Service- and patient-level outcomes show no consistent trends across the reporting periods assessed. No changes to the domain or benchmark target are proposed, however the definition of clinically significant improvement warrants revision. Specifically, the required change in severity category may limit recognition of meaningful improvement. |
| Depression | DASS-21 Depression subscale | ≥5-point reduction and shift to lower severity category | 70% | Psychological distress is assessed in ePPOC using the DASS-21, which includes subscales for depression, anxiety, and stress. Clinically |

| | | | | |
|----------------|--------------------------------|---|------------|---|
| Anxiety | DASS-21 Anxiety subscale | ≥5-point reduction and shift to lower severity category | 50% | <p>significant improvement is defined as a ≥5-point reduction in subscale score, accompanied by a shift to a less severe category. As with pain self-efficacy and catastrophising, this categorical requirement may warrant revision.</p> <p>Among the sub-domains, anxiety and stress warrant close, ongoing monitoring and more detailed statistical assessment. Anxiety results show a consistent downward trend at both service and patient levels, while stress results have declined more recently. In contrast, while depression outcomes show no clear trend, the proportion of services meeting this benchmark is low. Feedback from services will be sought at the 2025 Benchmarking Workshop to inform whether the depression benchmark target should be reduced to a more achievable level.</p> |
| Stress | DASS-21 Stress subscale | ≥5-point reduction and shift to lower severity category | 60% | |

