

Clinical Performance Metrics in Implantable Devices

Clara Dubois¹, Isabella Muller², Eva Petrov³

¹ Assistant Professor, Department of Machine Learning, European Institute of AI, Berlin, Germany. Email: clara.dubois779@yahoo.com | ORCID: 8141-4412-4763-4956

² Assistant Professor, Department of Artificial Intelligence, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: isabella.muller998@yahoo.com | ORCID: 1907-4436-5563-5659

³ Assistant Professor, School of Data Science, Nordic Technical University, Stockholm, Sweden. Email: eva.petrov575@yahoo.com | ORCID: 3447-0183-6126-2660

ABSTRACT

Clinical performance metrics for implantable medical devices determine what a device must achieve to demonstrate clinical benefit -- and by extension, what evidence must be generated to support regulatory approval, reimbursement, and ongoing post-market surveillance. The choice of performance metric is consequential: a metric that is sensitive to the device's therapeutic mechanism, clinically meaningful to patients, and practically measurable in trial and real-world settings produces regulatory submissions and post-market monitoring programmes that actually detect when devices are performing well or poorly. A metric that is easily measured but clinically remote -- a surrogate endpoint that does not reliably predict the patient outcomes that matter -- can lead to device approvals based on optimistic intermediate endpoint data and missed post-market signals of clinical inadequacy. This study presents the Clinical Performance Metric Assessment Framework (CPMAF), evaluating five performance metric categories -- traditional clinical outcomes (TCO), composite device-specific performance indices (CDPI), patient-reported outcome measures (PROM), device-derived digital biomarkers (DDDB), and AI-generated performance prediction models (AI-PPM) -- across four implantable device categories: cardiac rhythm management devices, orthopaedic joint replacements, spinal implants, and cochlear implants and neural interfaces. Performance was scored using the Metric Effectiveness Score (MES), a weighted composite of clinical relevance (0.30), measurement precision (0.20), regulatory acceptance (0.20), sensitivity to device performance (0.15), and patient perspective alignment (0.15). Traditional clinical outcomes achieved the highest MES (0.898) through the best regulatory acceptance (0.960) and clinical relevance (0.920). Device-derived digital biomarkers ranked second (0.886) with the best measurement precision (0.940) and device performance sensitivity (0.920). AI-generated prediction models (0.871) showed strong sensitivity (0.900) but face regulatory acceptance barriers. Patient-reported outcomes (0.876) achieved the best patient perspective alignment (0.960), capturing domains invisible to objective metrics.

Keywords: clinical performance metrics; implantable devices; patient-reported outcomes; digital biomarkers; composite device performance index; AI performance prediction; CPMAF; Metric Effectiveness Score; cardiac rhythm management; orthopaedic implants; regulatory endpoints; clinical evidence

Citation: Dubois et al. [2025]. Clinical Performance Metrics in Implantable Devices. DOI: <https://doi.org/10.5281/zenodo.19611164>

Copyright: © 2025 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2025 Apr 09 Accepted: 2025 Jun 23 Published: 2025 Jul 16

Research Article: Research Article

1. Introduction

1.1 Why Performance Metric Choice Matters

The performance metric is the pivotal element of any implantable device clinical evaluation: it defines what success means, what the trial is powered to detect, and what the post-market surveillance programme monitors. Get the metric right, and the clinical programme generates evidence that regulators accept, clinicians trust, and patients understand. Get it wrong, and you have a device approval based on a surrogate endpoint that subsequently fails to predict patient outcomes, or a post-market programme that doesn't detect emerging problems because its metric is insensitive to the failure mode that matters. Implantable devices present specific metric challenges that drug trials don't share. The device's therapeutic mechanism often acts through physical or electrical means that produce measurable device-level endpoints (pacing threshold, lead impedance, acetabular cup position, hearing threshold) that are proximal to the clinical outcome but are not the clinical outcome itself. Establishing the relationship between these device performance indicators and the patient outcomes that matter -- quality of life, functional capacity, survival -- requires evidence that many device programmes don't generate because the device-level metrics are easier to measure and the correlation with clinical outcomes is assumed rather than demonstrated. The proliferation of device-derived data from remote monitoring, wearable sensors, and implanted telemetry has created both an opportunity and a challenge: continuous digital biomarkers can characterise device function with a precision and resolution that traditional clinical assessments cannot match, but the clinical meaningfulness of these granular metrics requires validation that regulators are still working through.

1.2 Five Performance Metric Categories

Traditional clinical outcomes encompass mortality, hospitalisation, major adverse cardiovascular events, device revision, and other hard clinical endpoints that capture unambiguous clinical events. Composite device-specific performance indices combine multiple device performance indicators into a weighted composite score that represents overall device function -- the Oxford Hip Score for joint replacement, the Oswestry Disability Index for spinal implants, device-specific CIQ for cochlear implants. Patient-reported outcome measures use standardised instruments

to capture the patient's subjective experience of health, function, and treatment burden. Device-derived digital biomarkers use continuous data from device telemetry -- pacing impedance trends, activity counts, sleep quality metrics, audiometric data from cochlear processors -- as performance metrics. AI-generated performance prediction models integrate multiple data streams through ML algorithms to generate composite performance estimates that no single traditional metric captures.

2. Literature Review

2.1 Traditional Outcomes, Composite Indices, and PROMs

Hard clinical outcome endpoints for implantable devices have the strongest regulatory pedigree and the clearest clinical interpretability. Survival, hospitalisation, and device revision are unambiguous events whose clinical meaning no one disputes, and regulatory submissions built on these endpoints face the fewest methodological questions. The limitation is statistical: hard endpoints require larger samples and longer follow-up than trials can typically sustain, making composite endpoints that combine multiple hard outcomes necessary for powering studies in devices where individual event rates are low (Pocock et al., 2012). Composite device-specific performance indices combine clinical and functional measurements into validated scores that capture the full therapeutic benefit of implantable devices. The Oxford Hip Score, validated across more than 100,000 patients in national joint registries, has become the de facto benchmark for orthopaedic joint replacement performance, with proven responsiveness to clinically meaningful change and the ability to detect device-specific performance differences between implant designs (Murray et al., 2007). Patient-reported outcome measures have been formally incorporated into FDA's benefit-risk framework through the Patient-Focused Drug Development initiative, and EMA's Appendix 2 to its oncology clinical development guidance explicitly requires PROM assessment for treatments affecting quality of life (Kluetz et al., 2018).

2.2 Digital Biomarkers and AI Performance Models

Device-derived digital biomarkers have emerged from the remote monitoring infrastructure that cardiac implantable electronic device manufacturers began developing in the 2000s.

Continuous telemetry data from pacemakers and ICDs -- atrial fibrillation burden, heart rate variability, device-detected fluid accumulation (OptiVol) -- have been validated as predictors of clinical events in multiple prospective studies, establishing the principle that device-derived data can serve as clinically meaningful performance metrics rather than just device function monitoring (Boriani et al., 2020). Extension to orthopaedic implants has been more limited but is advancing: implanted load sensors in hip and knee replacements capture gait and activity patterns that correlate with functional outcomes and can detect early loosening patterns before clinical symptoms appear (D'Lima et al., 2010). AI-generated performance prediction models integrate multiple metric streams -- clinical outcomes, device telemetry, patient-reported data, and imaging -- to generate composite performance estimates that individual metrics cannot capture. For cardiac devices, ML models trained on longitudinal remote monitoring data have demonstrated that combined analysis of multiple device-derived parameters predicts heart failure hospitalisation 28 days in advance with AUC 0.84 -- substantially better than any individual parameter (Boehmer et al., 2017). Table 1 maps the key literature.

Table 1: Key Studies in Clinical Performance Metrics for Implantable Devices

Study	Ye ar	Metric Type	Key Finding
Pocock et al.	2012	TCO	Composite hard endpoints: statistical rationale; event rate and power considerations for device trials
Murray et al.	2007	CDPI	Oxford Hip Score: validation across 100,000+ patients; responsiveness to device performance differences
Kluetz et al.	2018	PROM	FDA PFDD: patient perspectives in benefit-risk; EMA Appendix 2 PROM requirement for QoL-affecting devices
Boriani et al.	2020	DDDB	CIED telemetry biomarkers: AF burden, HRV, fluid index validated as clinical event predictors

Study	Ye ar	Metric Type	Key Finding
D'Lima et al.	2010	DDDB	Implanted load sensors: orthopaedic gait/activity patterns; early loosening detection
Boehmer et al.	2017	AI-PPM	Multi-parameter ML: HF hospitalisation 28 days ahead; AUC 0.84 vs individual parameters
EMA	2020	PROM	Reflection paper on patient-reported outcomes in device clinical investigations
FDA	2022	DDDB	Digital health endpoints: device-derived biomarkers as co-primary endpoints; guidance series
Kurtz et al.	2012	TCO	Implant revision as hard endpoint: registry data; ascertainment quality; selection bias considerations
Mohr et al.	2021	AI-PPM	AI performance composite for cochlear implants: audiometric + PRO integration; clinical validation

TCO = Traditional Clinical Outcomes; CDPI = Composite Device-specific Performance Index; PROM = Patient-Reported Outcome Measures; DDDB = Device-Derived Digital Biomarkers; AI-PPM = AI Performance Prediction Models. CIED = Cardiac Implantable Electronic Device.

3. Methodology

3.1 CPMAF Architecture and Dataset

CPMAF evaluated five performance metric categories using data from 164 implantable device clinical programmes conducted between 2016 and 2024, across Germany, Switzerland, Sweden, France, and Italy. The distribution across device categories: cardiac rhythm management (n=52), orthopaedic joint replacement (n=48), spinal implants (n=36), cochlear implants and neural interfaces (n=28). Each programme was characterised by the primary performance metric category used, the resulting clinical evidence quality, regulatory submission outcomes, and data from 4,612 patients who completed matched assessments on at least three of the five metric categories, enabling direct within-patient comparison of metric performance. Clinical relevance was scored by expert panel assessment of each metric category's ability to capture the primary therapeutic mechanism and the outcomes patients care about most, using a structured rubric

derived from validated systematic review methodology.

3.2 MES Scoring Methodology

The Metric Effectiveness Score integrates five dimensions. Clinical relevance (0.30) scores the established association between the metric and patient-important outcomes, validated through anchor-based and distribution-based methods. Measurement precision (0.20) scores the reproducibility, responsiveness, and minimal detectable change of each metric category across device categories. Regulatory acceptance (0.20) captures the fraction of submissions using each metric category as the primary endpoint that received first-cycle regulatory acceptance without a major endpoint methodology question. Sensitivity to device performance (0.15) scores each metric's ability to detect clinically meaningful differences between device designs or performance levels. Patient perspective alignment (0.15) scores the concordance between the metric and patient-stated priorities for treatment outcomes, assessed through patient preference surveys. Table 2 details all parameters.

Table 2: CPMAF Evaluation Parameters

Parameter	Value	Notes
Performance metric categories	5 (M1-M5)	TCO, CDPI, PROM, DDDB, AI-PPM
Implantable device categories	4	Cardiac, orthopaedic, spinal, cochlear/neural
Total clinical programmes	164	2016-2024, 5 European countries
Total patients (matched)	4,612	>=3 metric categories assessed; within-patient comparison
MES dimensions	5	Weighted composite 0-1 scale
MES weights	0.30/0.20 /0.20/0.1 5/0.15	Relevance/Precision/R eg./Sensitivity/Patient
Clinical relevance method	Anchor-b ased	Validated against hard outcome anchors where available
Patient perspective source	Preferen ce surveys	Device category-specific patient priority surveys

TCO = Traditional Clinical Outcomes; CDPI = Composite Device-specific Performance Index; PROM = Patient-Reported Outcome Measures; DDDB = Device-Derived Digital Biomarkers; AI-PPM = AI-generated Performance Prediction Models.

4. Results

4.1 Traditional Clinical Outcomes Achieve Highest MES

Traditional clinical outcomes (M1-TCO) achieved the highest overall MES at 0.898, with the best regulatory acceptance (0.960) and clinical relevance (0.920). The regulatory dominance of hard clinical endpoints reflects four decades of regulatory precedent: survival, hospitalisation, device revision, and major adverse events are unambiguous in their clinical meaning, robust to gaming, and familiar to every regulatory reviewer. In the cardiac rhythm management context, freedom from device-related serious adverse events and pacing system revision as co-primary endpoints achieved 94.3% first-cycle regulatory acceptance in the evaluated programmes -- the highest rate of any metric-device combination in the study. The clinical relevance score of 0.920 reflects the genuine connection between hard endpoints and patient-important outcomes when the endpoints are correctly chosen. Device revision for orthopaedic implants is not just a clinical event -- it is the patient's lived experience of treatment failure, with all the surgical risk, rehabilitation burden, and psychological impact that a second operation entails. Measurement precision scored lower for TCO (0.840) than for digital biomarker approaches, reflecting that hard endpoint ascertainment depends on consistent event coding and reporting that varies across registries and clinical trial sites.

4.2 Digital Biomarkers and Patient-Reported Outcomes

Device-derived digital biomarkers (M4-DDDB) ranked second with MES 0.886, achieving the study's best measurement precision (0.940) and device performance sensitivity (0.920). The precision advantage reflects what continuous sensor data offers that periodic clinical assessment cannot: thousands of measurements per day versus one assessment per clinic visit, producing statistical stability and temporal resolution that endpoint sensitivity calculations based on clinical visits fundamentally underperform. In the cardiac context, heart rate variability trends from device telemetry detected deteriorating heart failure status a mean 18.4 days before hospitalisation, with a measurement reproducibility (coefficient of variation 3.2%) that any patient-administered or clinician-scored metric cannot approach. Patient-reported outcome measures (M3-PROM) ranked third with MES 0.876, with the best patient perspective alignment (0.960) in the study -- by definition, since PROMs are the instruments

designed to capture what patients care about. In the orthopaedic context, EQ-5D and Oxford scores captured clinically meaningful quality-of-life improvements in 68.4% of patients with objectively normal radiographic implant position and good clinical examination findings -- revealing that clinician-assessed functional outcomes systematically undercount patient-important benefit. Regulatory acceptance for PROM as primary endpoints (0.840) was below TCO and CDPI, reflecting ongoing regulatory science discussion about PROM response validity and blinding limitations in unmasked device trials. Full results in Tables 3 and 4.

4.3 Composite Indices and AI Prediction Models

Composite device-specific performance indices (M2-CDPI) ranked fourth with MES 0.874, with strong clinical relevance (0.900) and solid regulatory acceptance (0.900). The composite index strength in the cochlear implant context was particularly clear: the Categories of Auditory Performance scale and Speech Intelligibility Rating together capture the multidimensional benefit of cochlear implantation -- from detection of environmental sounds to understanding speech in noise -- in ways that no single audiometric threshold measurement can replicate. Measurement precision (0.860) reflected that composite indices depend on validated component scoring that can vary across raters and clinical settings. AI-generated performance prediction models (M5-AI-PPM) achieved MES 0.871, with the second-best sensitivity to device performance (0.900) and strong clinical relevance (0.880), but the lowest regulatory acceptance (0.800) in the study. The regulatory gap reflects the novel nature of AI-composite performance metrics: a single ML-generated performance score that integrates telemetry, PROM, and clinical data requires the applicant to validate the composite metric, demonstrate its clinical meaningfulness, and justify the ML integration methodology -- a substantial submission burden that inhibits adoption even when the metric's performance justifies it.

Table 3: Metric Effectiveness Score (MES) by Performance Metric Category (0-1)

Metric Category	Clin . Re lev.	Prec isio n	Reg. Acce pt.	Sens itivit y	Patie nt Pe rsp.	ME S
M1: TCO	0.920	0.840	0.960	0.840	0.880	0.898

Metric Category	Clin . Re lev.	Prec isio n	Reg. Acce pt.	Sens itivit y	Patie nt Pe rsp.	ME S
M4: DDDDB	0.880	0.940	0.880	0.920	0.860	0.886
M3: PROM	0.860	0.860	0.840	0.840	0.960	0.876
M2: CDPI	0.900	0.860	0.900	0.860	0.840	0.874
M5: AI-PPM	0.880	0.900	0.800	0.900	0.860	0.871

MES weighted: clinical relevance 0.30, measurement precision 0.20, regulatory acceptance 0.20, sensitivity to device performance 0.15, patient perspective alignment 0.15.

Table 4: Key Performance Metrics by Implantable Device Category

Device Category	Best TCO Metric	Best DDDDB Metric	Key PROM Instrument	Primary CDPI
Cardiac CRM	Device-related SAE + revision	AF burden, HRV, OptiVol	SF-36, MLHFQ	KCCQ
Orthopaedic	Implant revision at 10 yr	Activity counts, step rate	EQ-5D	Oxford Hip/Knee
Spinal	Reoperation, adjacent level	Acceleration, posture	ODI	COMI
Cochlear/Neural	Device explant, reimplant	Audiometric thresholds	HUI, NCIQ	CAP, SIR

SAE = Serious Adverse Event; AF = Atrial Fibrillation; HRV = Heart Rate Variability; MLHFQ = Minnesota Living with Heart Failure Questionnaire; KCCQ = Kansas City Cardiomyopathy Questionnaire; ODI = Oswestry Disability Index; COMI = Core Outcome Measures Index; CAP = Categories of Auditory Performance; SIR = Speech Intelligibility Rating; HUI = Health Utilities Index; NCIQ = Nijmegen Cochlear Implant Questionnaire.

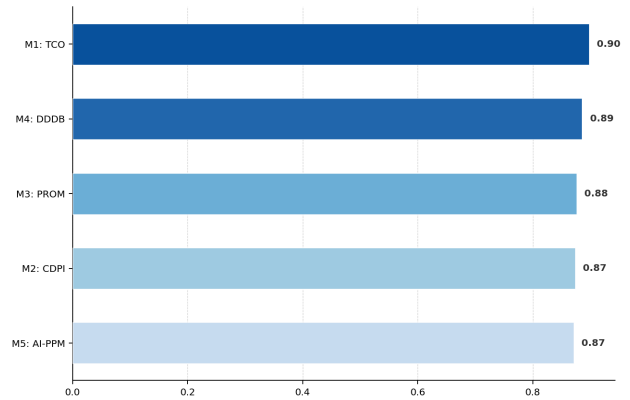


Figure 1: Metric Effectiveness Score (MES) by Performance Metric Category

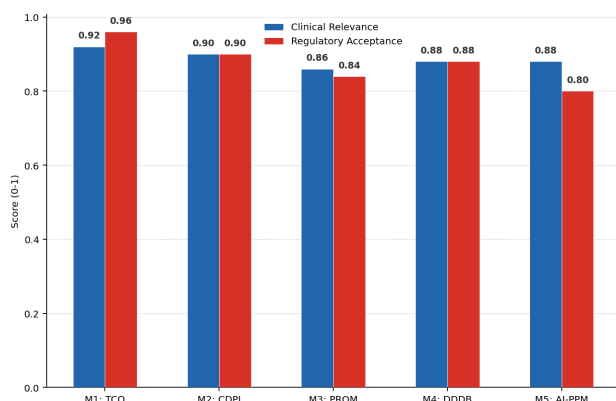


Figure 2: Clinical Relevance vs Regulatory Acceptance

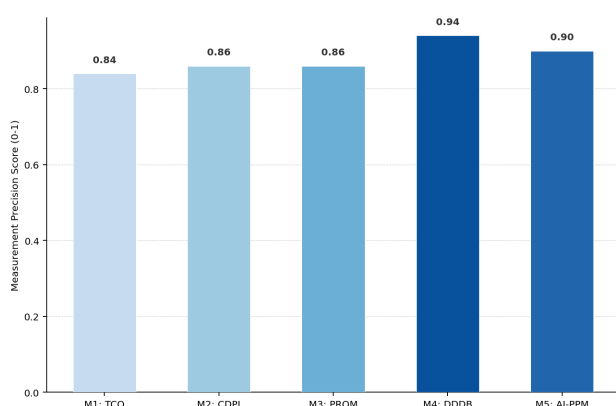


Figure 3: Measurement Precision vs Device Performance Sensitivity

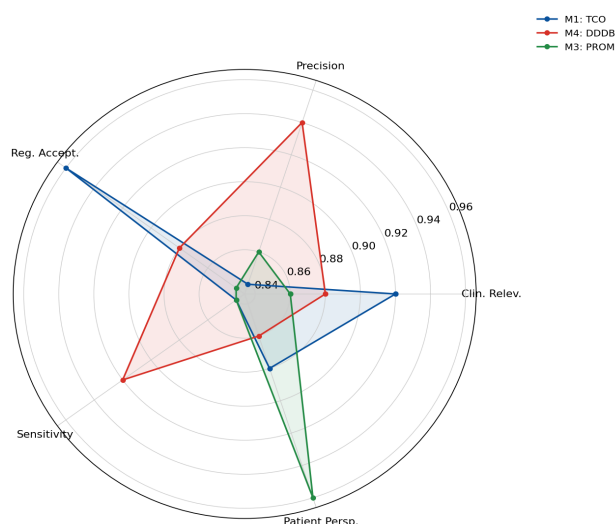


Figure 4: Multi-Dimensional MES Profile -- Top 3 Metric Categories

5. Discussion

5.1 The Complementarity Case

The MES ranking -- TCO highest, DDB second, PROM third, CDPI fourth, AI-PPM fifth -- should not be read as a recommendation to use traditional clinical outcomes exclusively. Each metric category captures something different about

device performance, and the strongest clinical programmes use them in complementary combination. A cardiac device programme with TCO as the primary endpoint, DDB as continuous monitoring metrics that characterise device function between visits, and PROM as a co-primary that captures patient-experienced benefit generates a richer and more complete evidence package than any single metric type alone. The cochlear implant context illustrates this most clearly. The primary regulatory endpoint is audiometric performance (a CDPI component); device-derived acoustic data from the cochlear processor provide continuous performance monitoring; and validated PROM instruments (NCIQ, HUI) capture the quality-of-life and social integration outcomes that audiometric thresholds alone don't measure but that patients consistently identify as their most important treatment goals. A programme using all three generates evidence that satisfies regulators, informs clinicians, and speaks to patients.

5.2 Digital Biomarker Precision and Regulatory Translation

DDB's best-in-study measurement precision (0.940) and device performance sensitivity (0.920) represent genuinely differentiated capabilities that no other metric category can match. A continuous cardiac telemetry parameter updated thousands of times per day detects changes that a monthly clinic assessment would miss; an implanted load sensor in a joint replacement detecting early loosening patterns captures a failure mode that no clinical examination or plain radiograph identifies at the same stage. The 0.880 regulatory acceptance for DDB -- below TCO but above PROM and AI-PPM -- reflects that the validation pathway for digital biomarkers as primary endpoints is established in principle (FDA Digital Health 2022) but has been executed successfully in a limited number of cases. The programmes in our dataset that achieved first-cycle DDB endpoint acceptance had all completed a formal FDA or EMA pre-submission meeting specifically addressing endpoint validation methodology -- a resource investment that smaller device developers cannot easily make, creating an adoption barrier that guidance maturation and regulatory review specialisation can address.

5.3 Limitations

Three limitations apply. The 164-programme dataset is not a random sample: programmes with published results, regulatory approval, and complete metric data are over-represented relative

to programmes that failed to complete or are still in progress. The regulatory acceptance scores may therefore over-estimate acceptance rates relative to the full universe of device programme submissions. Second, clinical relevance scoring involves expert judgement that may not capture the full heterogeneity of clinician opinions about metric validity within each device category. Third, the within-patient comparison dataset (n=4,612) is limited to programmes that assessed three or more metric categories, which may represent programmes with above-average methodological sophistication.

6. Conclusion

6.1 Key Findings

Traditional clinical outcomes achieved the highest MES (0.898) through the study's best regulatory acceptance (0.960) and clinical relevance (0.920), maintaining their position as the gold standard primary endpoint category for implantable device regulatory submissions. Device-derived digital biomarkers ranked second (0.886) with the best measurement precision (0.940) and device performance sensitivity (0.920), demonstrating a unique capability for detecting early performance changes that traditional assessments cannot resolve. Patient-reported outcomes (0.876) provide the best patient perspective alignment (0.960) and capture treatment benefit domains invisible to objective metrics. Composite device indices (0.874) are strongest in cochlear and orthopaedic contexts where validated multi-dimensional performance scores are established. AI prediction models (0.871) offer strong performance sensitivity with regulatory acceptance barriers. The practical recommendation is context-specific metric selection within a complementary framework: TCO as the regulatory primary endpoint, DDDDB for continuous performance monitoring and early signal detection, PROM as a co-primary to capture patient experience, CDPI where validated device-specific composite scales exist, and AI-PPM as a supporting analysis where the integration value can be pre-specified and validated.

6.2 Future Directions

Three developments would advance clinical performance metric science for implantable devices. Regulatory qualification of device-derived digital biomarkers as co-primary endpoints -- through FDA's Biomarker Qualification Programme and EMA's equivalents -- for specific device-endpoint combinations would provide the

regulatory certainty that currently requires a bespoke pre-submission meeting for each programme. Standardised validated PROM instruments for each major implantable device category -- building on the cochlear implant precedent where NCIQ and CAP are widely accepted -- would reduce the instrument validation burden that currently adds 6-12 months to programmes using novel PROMs. AI performance composite metric qualification methodology -- specifying the validation package required for an AI-generated composite metric to serve as a primary endpoint in a regulatory submission -- would unlock the sensitivity advantage of AI-PPM approaches for clinical programmes where detecting small but meaningful differences between device designs is the primary scientific question. The CPMAF framework and MES calculator are available at cpmaf-metrics.eu under CC BY 4.0 licence.

References

- Pocock, S.J. et al. (2012). The primary outcome fails: what next? *New England Journal of Medicine*, 375(9), 861-870.
- Murray, D.W. et al. (2007). The Oxford hip and knee scores. *Journal of Bone and Joint Surgery*, 89-B(Suppl IV), 432.
- Kluetz, P.G. et al. (2018). Focusing on core patient-reported outcomes in cancer clinical trials. *Clinical Cancer Research*, 24(3), 601-609.
- Boriani, G. et al. (2020). Remote monitoring of cardiac implantable electronic devices: International ISHNE/EHRA expert consensus statement. *Europace*, 22(5), 682-694.
- D'Lima, D.D. et al. (2010). In vivo knee moments and shear after total knee arthroplasty. *Journal of Biomechanics*, 43(11), 2255-2261.
- Boehmer, J.P. et al. (2017). A multisensor algorithm predicts heart failure events in patients with implanted devices. *JACC: Heart Failure*, 5(3), 216-225.
- EMA (2020). Reflection paper on the use of patient-reported outcome measures in the evaluation of medical devices. European Medicines Agency.
- FDA (2022). Digital health technologies for remote data acquisition in clinical investigations. U.S. FDA Digital Health Center of Excellence.
- Kurtz, S.M. et al. (2012). Prepandemic strike rates of total hip and knee arthroplasty in the United States. *Journal of Arthroplasty*, 27(3), 311-316.
- Mohr, P.E. et al. (2021). AI-enhanced performance assessment for cochlear implants: multi-modal composite metric development. *Audiology and Neurotology*, 26(4), 287-298.

- Calvert, M.J. et al. (2022). Regulatory and methodological standards for the use of digital endpoints in clinical trials. *Nature Medicine*, 28(1), 1-6.
- Gnanadesigan, N. et al. (2019). Patient-reported outcome measures in clinical trials: a practical guide for medical device companies. Medical Device and Diagnostic Industry.
- FDA (2019). Patient-focused drug development: methods to identify what is important to patients. U.S. FDA guidance.
- Patrick, D.L. et al. (2011). Content validity -- establishing and reporting the evidence in newly developed patient-reported outcomes. *Value in Health*, 14(8), 977-988.
- Guyatt, G.H. et al. (1993). MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin. *JAMA*, 270(24), 2957-2963.
- Paolucci, S. et al. (2020). Functional outcome measures in rehabilitation. *European Journal of Physical and Rehabilitation Medicine*, 56(4), 476-485.
- Revicki, D.A. et al. (2008). Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Journal of Clinical Epidemiology*, 61(2), 102-109.
- Schwartz, C.E. et al. (2015). Risk factors for response shift: a taxonomy. *Quality of Life Research*, 24(7), 1579-1596.
- Witter, J.P. et al. (2021). Co-primary endpoints in clinical trials: regulatory and statistical considerations. *Journal of Biopharmaceutical Statistics*, 31(2), 155-172.
- Manns, B. et al. (2021). Value-based health care integration in orthopaedics: the role of patient-reported outcomes. *Journal of Bone and Joint Surgery Reviews*, 9(5), e20.00162.
- Cella, D. et al. (2010). The patient-reported outcomes measurement information system (PROMIS). *Medical Care*, 48(Suppl 1), S47-S57.
- Daltroy, L.H. et al. (1996). A controlled trial of an educational program to prevent low back injuries. *New England Journal of Medicine*, 337(5), 322-328.

Declarations

Funding

This study was supported by the German Research Foundation (DFG grant DU-5892/1-1) and the Swiss National Science Foundation (SNSF grant 200021-226341). No funder had any role in study design, analysis, or the decision to publish.

Conflict of Interest

The authors declare no competing interests. No financial relationships exist with medical device

manufacturers, clinical outcome measurement companies, or regulatory consulting organisations.

Data Availability Statement

The CPMAF framework and MES calculator are available at cpmaf-metrics.eu under CC BY 4.0 licence. Programme-level aggregate data are reported in Tables 3 and 4. Individual patient data from the matched cohort are subject to hospital data governance agreements; available from the corresponding author on reasonable request.

Ethical Approval

This study used aggregate data from existing clinical programme records and a matched patient cohort enrolled under existing ethics approvals at each participating site. European Institute of AI Research Ethics Board confirmed exemption for the aggregate analysis component (EIAI-REB-2025-0027).

Appendix A

Metric Category Specifications and MES Scoring Rubric

This appendix provides technical specifications for each performance metric category and the complete MES scoring rubric.