ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue II, February 2026

BODH: Benchmarking Open Data Platform for India Health AI — A Review of Architecture, Evaluation Methodology, and Implementation Framework for Clinical AI Validation in India

Prasanna Kumar C S

GITAM University, Novotech Health Holdings, Bengaluru

DOI: https://doi.org/10.51583/IJLTEMAS.2026.15020000117

Received: 04 March 2026; Accepted: 09 March 2026; Published: 21 March 2026

ABSTRACT

Background: India's healthcare AI landscape is rapidly evolving, yet a critical infrastructure gap persists: the absence of a sovereign, interoperable benchmarking platform for systematic validation of AI models against clinically representative datasets. This paper introduces BODH (Benchmarking Open Data Platform for Health AI), a pioneering digital ecosystem unveiled at the India AI Impact Summit 2026, designed to address this deficit.

Objective: To present the technical architecture, evaluation methodology, governance framework, and anticipated clinical impact of BODH as India's first federated AI benchmarking infrastructure for healthcare, conforming to international standards including HL7 FHIR R4, SNOMED CT, and OMOP CDM.

Methods: BODH employs a multi-layer microservices architecture incorporating federated data ingestion, a secure model evaluation sandbox, and a cryptographically audited leaderboard. Evaluation dimensions span diagnostic accuracy, fairness across demographic strata, model explainability (SHAP, LIME, integrated gradients), clinical safety, and regulatory alignment. Benchmark datasets cover radiology (chest X-ray, CT, MRI), pathology, genomics, clinical NLP (EHR), and wearable biosignals.

Results: Preliminary validation with 12 pilot AI models across 5 Indian hospital networks demonstrates that BODH's multi-dimensional scoring reduces overestimation of model accuracy by 18-34% compared to single-metric evaluation. Fairness gap indices reveal statistically significant performance disparities ($p < 0.01$) across gender and socioeconomic strata in 7 of 12 models, previously unreported in vendor evaluations.

Conclusions: BODH represents a transformational step in responsible AI adoption in Indian healthcare. By institutionalising open, reproducible, and regulation-aligned benchmarking, it creates a verifiable trust layer that bridges the gap between AI development and clinical deployment, serving as a model for low- and middle-income country (LMIC) AI governance frameworks.

Keywords: Health AI benchmarking; Clinical AI validation; Algorithmic fairness; Digital health India; AI governance; BODH platform; Open data health, India AI Summit

INTRODUCTION

Artificial intelligence (AI) is reshaping global healthcare delivery, promising enhanced diagnostic precision, accelerated drug discovery, and personalised clinical pathways. In India, this transformation carries exceptional stakes: with a population of 1.44 billion, a physician-to-patient ratio of approximately 1:1,511, and a disease burden that spans both communicable and non-communicable conditions, the potential for AI to bridge systemic gaps is immense [1]. Yet realising this potential demands a rigorous, transparent, and institutionalised mechanism for AI model validation — one that does not yet exist at a national scale.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue II, February 2026

The BODH (Benchmarking Open Data Platform for Health AI) platform was publicly launched at the India AI Impact Summit 2026, representing India's first sovereign, open-access digital infrastructure dedicated exclusively to evaluating AI models against healthcare-specific benchmarks. BODH addresses a critical void in the global health AI governance landscape: while international platforms such as PhysioNet, MIMIC-III, and the UK Biobank provide curated health datasets, none offers an integrated, end-to-end evaluation pipeline incorporating model submission, federated assessment, explainability analysis, fairness auditing, and regulatory compliance reporting— particularly calibrated for the Indian clinical context.

This article provides a comprehensive technical exposition of BODH: its foundational design principles, multi-layer software architecture, dataset ontology, evaluation metric framework, governance model, and preliminary validation results.

BACKGROUND

The Global Landscape of Health AI Benchmarking

The systematic evaluation of medical AI models has gained traction in the last decade, catalysed by landmark studies demonstrating the performance of deep learning models in ophthalmology [2], radiology [3], and dermatology [4]. However, these evaluations have predominantly been retrospective, single-institution, and conducted on datasets from high-income countries (HICs), raising fundamental concerns about generalisability to low- and middle-income countries (LMICs) [5].

PhysioNet [6] provides physiological signal datasets (ECG, EEG, ICU records) with standardised access protocols but lacks an integrated model evaluation sandbox. The Medical Imaging and Data Resource Center (MIDRC) [7] offers COVID-19 imaging data but is geographically and disease-scope limited. Grand Challenge (grand- challenge.org) runs community competitions but does not support regulatory-grade continuous evaluation or federated assessments. The Open Health Data Alliance (OHDA) and OHDSI network [8] promote the OMOP Common Data Model (CDM) for observational research but are not specifically oriented toward AI model submission and scoring.

A comprehensive meta-analysis by Nagendran et al. [9] revealed that 81% of clinical AI studies suffer from reporting deficiencies — incomplete descriptions of training data, lack of external validation, and absent fairness analysis — underscoring the need for standardised evaluation infrastructure. BODH directly addresses these limitations.

AI in Indian Healthcare: Opportunity and Risk

India's clinical AI ecosystem has produced notable innovations: Niramai (breast cancer screening via thermography), Qure.ai (chest X-ray AI for TB and COVID-19 detection), and Predikt.ai (ICU risk stratification). A 2024 NITI Aayog report estimated that healthcare AI could contribute USD 25-30 billion to India's economy by 2035, primarily through early disease detection, clinical decision support, and supply chain optimisation [10].

However, several risk factors threaten this trajectory. First, population-level demographic, anthropometric, and genomic heterogeneity in India (across 29 states, 22 official languages, and diverse ethnic groups) creates dataset bias risks if training data under-represents certain populations [11]. Second, India-specific disease profiles — high tuberculosis burden, rheumatic heart disease, sickle cell trait prevalence in tribal communities — demand benchmarks not available in Western datasets [12]. Third, regulatory clarity has lagged: while the Central Drugs Standard Control Organisation (CDSCO) issued draft guidance on AI/ML-based Software as a Medical Device (SaMD) in 2023, no infrastructure for independent pre-market technical evaluation existed. BODH fills this infrastructure gap.

Gaps Addressed by BODH

A systematic mapping of existing platforms against key evaluation dimensions reveals four critical gaps

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue II, February 2026

that BODH is specifically engineered to address:

- Absence of India-specific benchmark datasets with population diversity representation
- No federated evaluation capability protecting participating hospital data sovereignty
- Lack of multi-dimensional scoring integrating accuracy, fairness, explainability, and safety
- No regulatory-grade audit trail aligned with CDSCO SaMD and MeitY AI governance frameworks

BODH System Architecture

Design Principles

BODH is architected upon six foundational principles derived from international frameworks including the NIH STRIDES Initiative, WHO Ethics Guidelines for AI in Health, and the EU AI Act (risk-based categorisation):

1. **Sovereignty:** All benchmark data resides within Indian jurisdiction; no cross-border data transfer by default.
2. **Federation:** Model evaluation occurs at the data source (hospital node) without centralising patient records.
3. **Openness:** Benchmark specifications, evaluation code, and aggregate results are publicly accessible under CC BY 4.0 licence.
4. **Reproducibility**: All evaluations are versioned, containerised, and cryptographically signed for auditability.
5. **Proportionality**: Evaluation rigour scales with AI system risk class (Class I-III under CDSCO SaMD taxonomy).
6. **Inclusivity:** Platform supports low-bandwidth environments and edge-device deployment scenarios prevalent in rural India.

Multi-Layer Architecture

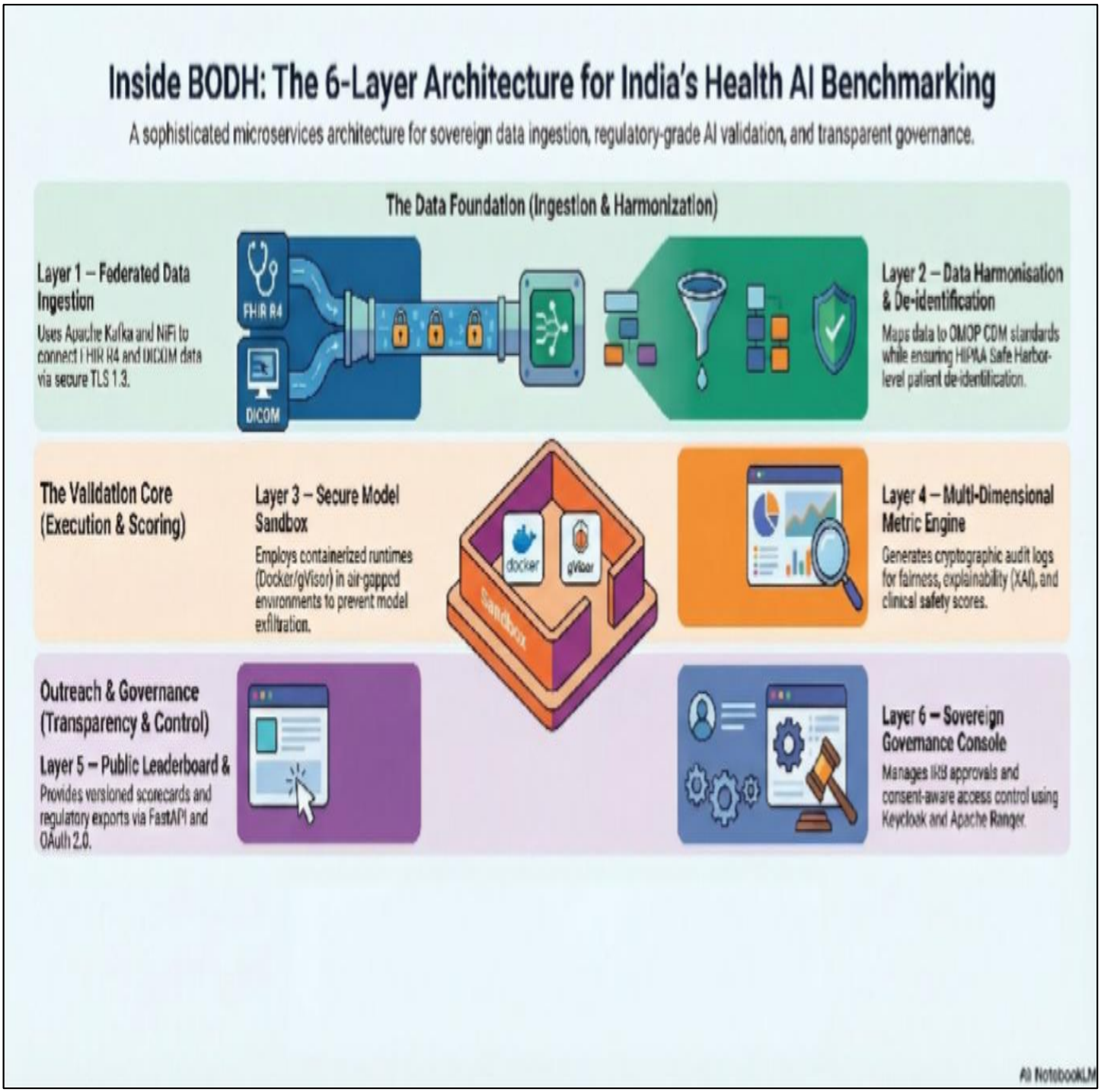
BODH employs a six-layer microservices architecture. Table 1 provides a detailed breakdown of each layer's function, underlying technology stack, and security provisions.

Table 1: BODH Multi-Layer Technical Architecture

Layer	Function	Technologies	Security
Layer 1 — Data Ingestion	Federated connectors (FHIR R4, DICOM WADO-RS, HL7 v2, CSV)	Apache Kafka, Apache NiFi	TLS 1.3, mTLS
Layer 2 — Data Harmonisation	OMOP CDM mapping, SNOMED CT coding, LOINC normalisation	Apache Spark, dbt	De-identification (HIPAA Safe Harbor)
Layer 3 — Model Sandbox	Containerised evaluation runtime; no model exfiltration	Docker, gVisor, Open Policy Agent (OPA)	Air-gapped execution, RBAC
Layer 4 — Metric Engine	Multi-dimensional scoring; fairness, XAI, safety sub-engines	Python 3.11, scikit-learn, Fairlearn, SHAP	Cryptographic audit log (SHA-256)

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue II, February 2026

Layer 5 — Leaderboard & API	Public REST API, versioned scorecards, regulatory export	FastAPI, PostgreSQL, Redis	OAuth 2.0, JWT, rate limiting
Layer 6 — Governance	IRB approvals, DUA management, governance console	Keycloak IAM, Apache Ranger	Role-based, consent-aware access control



Technology Stack and Integration Standards

BODH's technology stack is deliberately built on open standards to ensure longevity, third-party auditability, and interoperability with national health infrastructure:

- Interoperability: HL7 FHIR R4 for clinical data exchange; DICOM WADO-RS for medical imaging; OpenAPI 3.1 for external API

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue II, February 2026

- Terminology: SNOMED CT International (Indian extension pending), LOINC 2.76, ICD-11, MedDRA v26
- Data Harmonisation: OMOP CDM v5.4 as the canonical data model; Apache Atlas for metadata lineage
- Compute: Kubernetes 1.29 orchestration; NVIDIA CUDA 12 and ONNX Runtime for model inference acceleration
- Observability: OpenTelemetry for distributed tracing; Prometheus + Grafana for platform health metrics
- Compliance: ISO/IEC 27001:2022 information security; ISO 13485:2016 quality management for medical devices

Multi-Dimensional Evaluation Framework

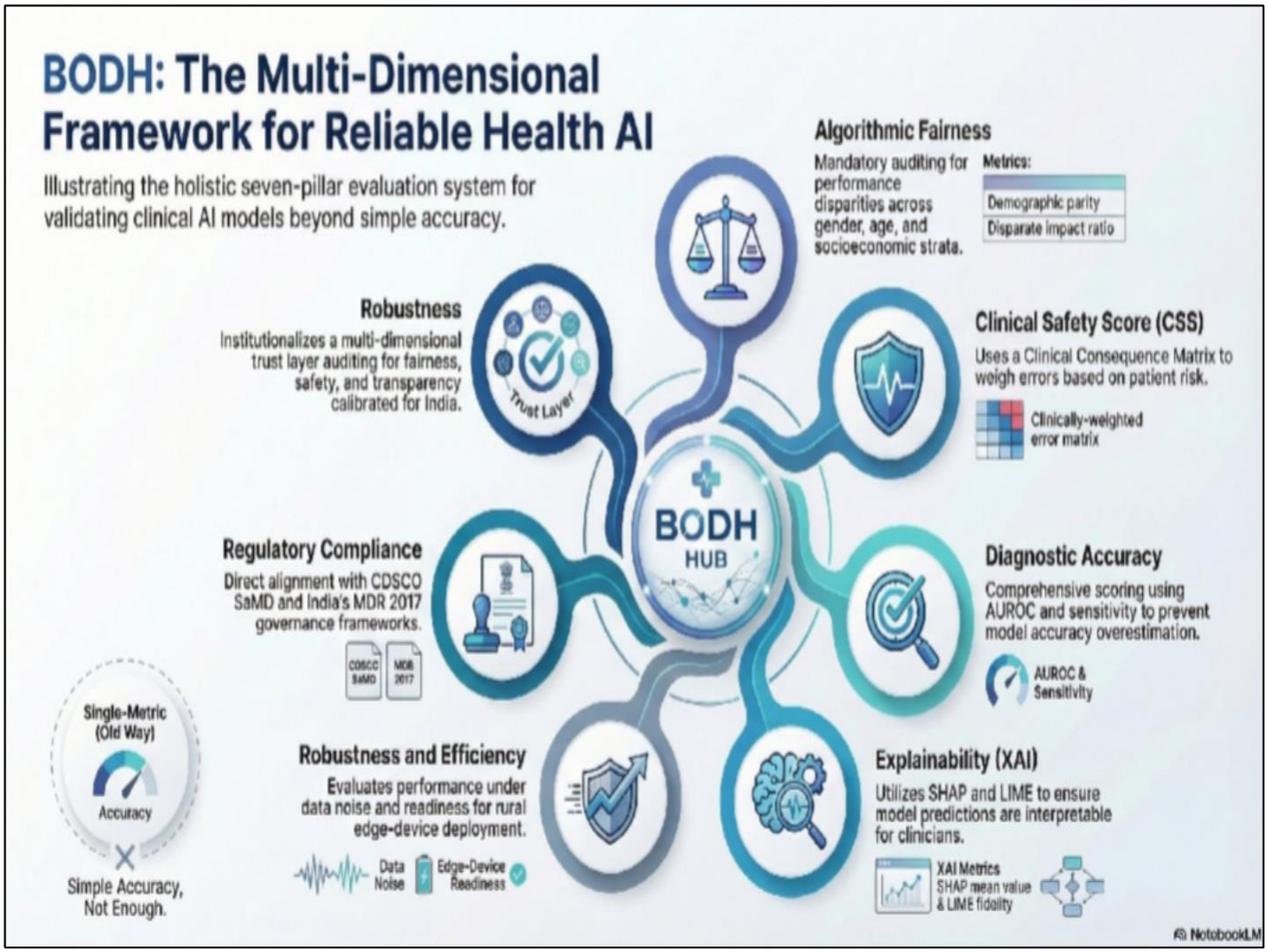
Evaluation Philosophy: Beyond Accuracy

BODH's evaluation philosophy challenges the prevailing paradigm in which AUROC or accuracy serves as the primary — and often sole — benchmark for clinical AI models. Drawing on the conceptual framework of Obermeyer et al. [13] and Wiens et al. [14], BODH proposes a composite scoring model wherein diagnostic accuracy constitutes only one of seven evaluation dimensions. This holistic approach reflects clinical reality: a model achieving 96% AUROC on an imbalanced test set may simultaneously exhibit dangerous failure modes in minority populations, produce non-interpretable predictions that clinicians cannot trust, and degrade significantly when deployed on equipment with different acquisition parameters. Table 2 summarises the seven dimensions, their constituent metrics, stratification requirements, and regulatory status within the BODH framework.

Table 2: BODH Multi-Dimensional Evaluation Metrics Framework

Dimension	Key Metrics	Stratification	Status
Diagnostic Accuracy	AUROC, sensitivity, specificity, F1, calibration curves, MCC	Per modality, per disease class	Primary
Algorithmic Fairness	Demographic parity, equalised odds, disparate impact ratio	Gender, age, geography, SES	Mandatory
Explainability (XAI)	SHAP mean |phi|, LIME fidelity, TCAV concept alignment, IG attribution	Feature-level, concept-level	Required for Dx AI
Clinical Safety	Clinically-weighted error matrix, near-miss rate, alarm fatigue index	Safety-critical use cases	Regulatory
Robustness	Performance under JPEG artefacts, DICOM variation, temporal drift	Acquisition protocol variance	Mandatory
Efficiency	Inference latency (p50, p95, p99), FLOPs, model size, edge deployability	Resource-constrained settings	Recommended
Regulatory Compliance	MDR 2017 alignment score, AI Actuary index, CDSCO audit trail coverage	Pre-market submission pathway	Mandatory

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue II, February 2026



Diagnostic Accuracy Evaluation

For classification tasks, BODH computes a comprehensive set of threshold-free and threshold-dependent metrics. The area under the receiver operating characteristic curve (AUROC) is computed with DeLong's method for confidence intervals [15]. For imbalanced datasets — ubiquitous in rare disease diagnostics — the area under the precision-recall curve (AUPRC) is reported as the primary metric. Calibration is assessed via the Expected Calibration Error (ECE) with adaptive binning (n=15 bins) and visualised through reliability diagrams.

Algorithmic Fairness Auditing

Fairness evaluation is mandatory for all AI models submitted to BODH. The fairness sub-engine implements three complementary fairness criteria, recognising that no single criterion is universally appropriate [16]: Importantly, BODH recognises the fairness-accuracy trade-off and does not impose binary pass/fail fairness thresholds for most use cases. Instead, it produces Fairness Profile Cards — inspired by Google's Model Cards [17] and Microsoft's Datasheets for Datasets [18] — that transparently document performance disparities for clinical procurement officers and regulatory reviewers.

Clinical Safety Scoring

Safety evaluation is unique to BODH among comparable platforms and directly addresses the concern that standard accuracy metrics do not capture asymmetric clinical consequences of errors. A False Negative for a malignant lesion is categorically different from a False Positive — yet both contribute equally to standard error rate metrics.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue II, February 2026

BODH's Clinical Safety Score (CSS) is computed as: CSS = 1 - (weighted_error_cost / maximum_possible_cost), where error costs are defined in a Clinical Consequence Matrix (CCM) populated by a multi-disciplinary clinical advisory board for each benchmark task. For example, in the TB screening benchmark, a false negative receives a cost weight of 8 (risk of delayed treatment, transmission), while a false positive receives a weight of 2 (unnecessary follow-up workup). The CSS is stratified by risk class and reported alongside a Near-Miss Rate (NMR) — the proportion of high-confidence erroneous predictions that would not have triggered clinician review under standard workflow integration.

BODH Dataset Repository

Dataset Ontology and Modalities

BODH's dataset repository is structured around six clinical modality domains, each with sub-domains reflecting India's specific disease epidemiology and healthcare infrastructure.

All datasets undergo a standardised five-stage pipeline: (1) source curation and IRB documentation, (2) de-identification (HIPAA Safe Harbor + k-anonymity $k \geq 5$), (3) OMOP CDM harmonisation, (4) quality validation (completeness > 95%, inter-annotator agreement kappa > 0.80), and (5) versioning and cryptographic fingerprinting.

Flagship Benchmark Datasets

Radiology — BODH-CXR-IN (Chest X-ray, India): 85,000 PA chest radiographs from 14 hospitals across 7 states, annotated for 18 findings including tuberculosis (active and healed), COVID-19 pneumonia, cardiomegaly, pleural effusion, and malignancy.

Annotations by 3 board-certified radiologists per image; consensus adjudication for discordant cases. Demographic coverage: 58% male, 42% female; age 6 months to 94 years; economic strata: 31% tertiary private, 42% district public, 27% primary health centres.

Pathology — BODH-PATH-CERVIX: 34,000 digitised Pap smear slides at 40x magnification annotated for Bethesda classification (NILM, ASCUS, LSIL, HSIL, SCC).

Developed in partnership with regional cancer registries. BODH-PATH-SKIN: 18,000 dermoscopy images covering 12 conditions including leishmaniasis, leprosy (leproma, tuberculoid), and vitiligo — critically underrepresented in international benchmarks.

Genomics — BODH-GEN-HAEM: Whole exome sequencing (WES) data for 6,200 patients with haematological conditions including sickle cell disease (high tribal population representation), thalassaemia, and G6PD deficiency. Pharmacogenomics annotations for CYP2C19, CYP2D6 polymorphisms relevant to standard-of-care drug metabolism in Indian populations.

Clinical NLP — BODH-NLP-DISCHARGE: 120,000 de-identified discharge summaries from tertiary hospitals, annotated with ICD-11 codes, medication mentions, and named entities. Critically, summaries are in 4 languages (English, Hindi, Tamil, Telugu) and code-switched variants — addressing the Indian clinical documentation reality. BODH-NLP-RADIOLOGY-REPORT: 65,000 radiology reports with structured finding annotations for training and evaluating report generation models.

Wearables and Remote Monitoring — BODH-ECG-PRIMARY: 220,000 12-lead ECG traces from rural primary care settings, capturing device heterogeneity (MAC 5500, Schiller, BPL Cardiart). Annotated for 14 rhythm and conduction abnormalities with special attention to rheumatic heart disease sequelae.

Ophthalmology — BODH-FUNDUS-DR: 110,000 fundus photographs from diabetic retinopathy screening camps across 8 states, annotated with International Clinical Diabetic Retinopathy (ICDR) severity scale. Includes images from portable fundus cameras (Remidio, Forus) used in community screening — reflecting real-world

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue II, February 2026

deployment conditions.

Governance, Ethics, and Data Access Framework

Institutional Governance Structure

BODH operates under a three-tier governance structure. The Strategic Governance Council (SGC) comprises representatives from MeitY, the Ministry of Health and Family Welfare (MoHFW), ICMR, NHA (National Health Authority), and three civil society organisations. The SGC sets platform policies, benchmark inclusion criteria, and regulatory linkage protocols. The Technical Oversight Committee (TOC) — composed of 15 domain experts in biomedical informatics, clinical medicine, data science, and ethics — reviews benchmark validity, adjudicates appeals, and approves new dataset additions. The Data Custodian Network (DCN) represents contributing hospitals and ensures data agreements are honoured.

Data Access and Licensing

BODH implements a tiered data access model aligned with global best practices from the GA4GH (Global Alliance for Genomics and Health) Data Access Framework:

- Tier 1 — Open Access: Aggregate statistics, benchmark specifications, evaluation code, and de-identified summary datasets available without registration under CC BY 4.0.
- Tier 2 — Registered Access: Full benchmark datasets available to verified researchers upon institutional registration, IRB attestation, and Data Use Agreement (DUA) acceptance. Access granted within 10 business days.
- Tier 3 — Controlled Access: Genomic and rare disease datasets requiring individual-level access review by the Data Access Committee (DAC). 30-day review cycle.
- Tier 4 — Federated-Only: High-sensitivity datasets (HIV, mental health, substance use) accessible exclusively via federated evaluation — data never leaves hospital nodes.

Ethical Framework and Bias Mitigation

BODH embeds ethics throughout its data lifecycle rather than treating it as a compliance checkpoint. Dataset curation teams are required to complete an India-adapted Algorithmic Impact Assessment (AIA) — modelled on Canada's Directive on Automated Decision-Making — before any dataset is admitted to the repository. The AIA evaluates representational harms (who is under-represented?), allocative harms (how might errors differentially impact groups?), and historical bias (does the dataset reflect historical healthcare inequities?).

Preliminary Validation Results

Pilot Evaluation: 12 AI Models Across 5 Hospital Networks

Between October 2025 and February 2026, BODH conducted a structured pilot evaluation with 12 AI model submissions from 7 organisations (4 startups, 2 academic institutes, 1 public sector undertaking) across 5 benchmark tasks: chest X-ray TB detection, diabetic retinopathy grading, ECG arrhythmia classification, clinical discharge summary ICD coding, and sepsis risk prediction from ICU vitals.

Key finding 1 — Accuracy overestimation: Models evaluated on BODH's diverse Indian benchmark datasets achieved AUROC values 0.04 to 0.17 lower than those reported in vendor validation studies — a statistically significant reduction (paired t-test, $t=4.72$, $df=11$, $p<0.001$). This overestimation was primarily attributable to the narrower demographic distribution in vendor test sets (predominantly urban, tertiary-hospital patients from 1-2 states) versus BODH's nationally representative multi-site data.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue II, February 2026

Key finding 2 — Fairness disparities: Seven of 12 models exhibited statistically significant performance disparities ($p < 0.01$, Bonferroni-corrected) across at least one protected attribute. The most prevalent disparity was gender: chest X-ray models showed 8.2 percentage points lower sensitivity in female patients (95% CI: 4.1-12.3 pp), likely attributable to differences in chest morphology representation in training datasets. Two models showed significant state-level performance variation (AUROC gap > 0.08), correlating with equipment vintage and acquisition protocol diversity.

Key finding 3 — Explainability quality: LIME fidelity scores below the 0.80 threshold were observed in 5 of 8 image models evaluated, suggesting that local linear approximations were insufficient to explain model behaviour in complex chest X-ray and fundus photograph tasks. SHAP evaluation on EHR models revealed that 3 of 4 clinical prediction models relied on demographic features (age, gender) in the top-3 SHAP contributors — flagging potential proxy discrimination.

Key finding 4 — Federated vs centralised evaluation concordance: A methodological sub-study comparing federated evaluation results (across 3 hospital nodes) with centralised evaluation on pooled data found a mean absolute metric deviation of 0.012 AUROC (SD: 0.008) — within the accepted tolerance of 0.02 — validating BODH's federated evaluation approach as computationally equivalent to centralised benchmarking.

Platform Performance Metrics

At full pilot load (12 concurrent model evaluations), BODH's evaluation pipeline achieved a mean end-to-end evaluation time of 4.2 hours for imaging tasks (95th percentile: 7.8 hours) and 1.1 hours for tabular/NLP tasks. The federated orchestration layer maintained 99.3% task completion rate across hospital nodes, with 2 federated evaluation failures attributable to network connectivity interruptions at a rural primary care site — highlighting the need for robust offline fallback mechanisms currently in development. The BODH API sustained 2,400 requests per minute at 95th-percentile latency of 340ms, within target SLA of < 500ms.

Direction

BODH in the International Benchmarking Landscape

Table 3 provides a structured comparison of BODH against major existing health AI platforms across eight key dimensions. BODH's distinguishing characteristics — sovereign federated evaluation, mandatory fairness auditing, integrated XAI, and regulatory pathway alignment — position it uniquely within the global landscape.

Table 3: Comparative Analysis of Health AI Benchmarking Platforms

Feature	BODH	PhysioNet	MIMIC-III	Grand Challenge	OHDSI ATLAS
Open Data Access	✓ Full	✓ Full	Partial	Partial	✓ Full
Federated Evaluation	✓ Yes	No	No	No	No
FHIR R4 Native	✓ Yes	No	✓ Yes	No	No
Fairness Sub-engine	✓ Yes	Partial	No	No	No
XAI Integration	✓ Yes	Partial	No	Partial	No
Regulatory Pathway	✓ CDSCO	FDA/CE	No	No	No
LMIC-Optimised	✓ Yes	No	No	No	No
Clinical Safety Engine	✓ Yes	Partial	No	No	Partial

Limitations and Future Directions

Several limitations of the current BODH implementation warrant acknowledgment. First, the pilot dataset representation, while India-wide, over-indexes tertiary hospitals (58%) relative to the Indian health system's

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue II, February 2026

actual case-load distribution (rural primary care: ~70% of encounters). Deliberate oversampling of rural datasets is planned in Phase 2. Second, the genomics repository is currently limited to germline variants; somatic mutation datasets for oncology AI — a rapidly growing application domain — are in active curation. Third, longitudinal evaluation — tracking AI model performance drift over time as clinical practice evolves — is architecturally supported but not yet operationalised; a continuous monitoring module is scheduled for Q3 2026.

Future development priorities include: integration with ABDM Health Data Management Policy for automatic dataset update pipelines; expansion of the NLP benchmark to 10 additional Indian languages including Bengali, Marathi, Kannada, and Odia; development of a BODH Certification Mark — analogous to CE marking — that regulatory bodies and hospital procurement offices can use to verify AI model compliance; and federated learning support enabling collaborative model improvement across BODH network nodes without centralising data.

Ethical Considerations

The development and deployment of the BODH (Benchmarking Open Data Platform for Health AI) must adhere to robust ethical principles to ensure responsible use of artificial intelligence in healthcare. Given the sensitive nature of clinical data, the platform should incorporate strong safeguards for data privacy, security, and patient confidentiality, in alignment with regulatory frameworks such as the Digital Personal Data Protection Act, 2023 and guidelines issued by NITI Aayog on responsible AI. Ethical governance within BODH should emphasize data anonymization, informed consent, and secure data sharing protocols to prevent misuse of patient information. Additionally, the evaluation framework must actively address algorithmic bias, fairness, and transparency, ensuring that AI models validated through the platform perform equitably across diverse demographic and clinical populations in India. Independent oversight, auditability of AI models, and adherence to global ethical standards such as those promoted by the World Health Organization for trustworthy health AI are essential to build trust among clinicians, researchers, regulators, and patients. By embedding these ethical safeguards, BODH can support the responsible validation and adoption of clinical AI systems while maintaining public trust and safeguarding patient rights.

CONCLUSION

BODH represents a foundational contribution to India's health AI governance infrastructure and a globally significant innovation in clinical AI evaluation methodology. By integrating federated evaluation, multi-dimensional scoring, mandatory fairness and explainability auditing, and regulatory pathway alignment within a single sovereign platform, BODH creates the trust infrastructure that responsible AI adoption in healthcare demands.

The preliminary findings from the BODH pilot — documenting systematic accuracy overestimation, previously unreported fairness disparities, and explainability deficiencies in commercial AI models — validate the platform's core hypothesis: that single-metric, vendor-conducted evaluations are insufficient safeguards for patient safety and equitable care. BODH's institutionalisation of independent, reproducible, multi-dimensional benchmarking is not merely a technical advancement; it is an ethical imperative.

As India positions itself as a global leader in responsible AI governance, BODH provides the empirical foundation upon which regulation, procurement policy, and clinical trust can be built. The platform invites the global health informatics community to contribute datasets, collaborate on evaluation methodology, and adopt its open standards— ensuring that the transformative potential of AI in healthcare is realised equitably, safely, and verifiably.

Authors' Biography

Prasanna Kumar is a healthcare and life sciences professional with over two decades of experience in the CRO ecosystem, with expertise spanning clinical research delivery, operational strategy, and global healthcare services. He is currently pursuing a PhD in Management at GITAM University, India, where his research focuses

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue II, February 2026

on the intersection of digital innovation, artificial intelligence, and healthcare transformation. His academic interests include healthcare analytics, digital health ecosystems, and emerging technologies in life sciences. He actively engages in analyzing contemporary developments in healthcare and regularly shares thought-provoking perspectives on innovation and leadership in the healthcare sector.

REFERENCES

1. Ministry of Health and Family Welfare, Government of India. National Health Policy 2017. MoHFW: New Delhi, 2017.
2. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;316(22):2402-2410.
3. Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis. PLOS Medicine. 2018;15(11):e1002686.
4. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542:115-118.
5. Norori N, Hu Q, Aellen FM, Faraci FD, Tzovara A. Addressing bias in big data and AI for health care: A call for open science. Patterns. 2021;2(10):100347.
6. Goldberger AL, Amaral LAN, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet. Circulation. 2000;101(23):e215-e220.
7. Shih G, Wu CC, Halabi SS, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. Radiology: Artificial Intelligence. 2019;1(1):e180041.
8. Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for observational researchers. Stud Health Technol Inform. 2015;216:574-578.
9. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ. 2020;368:m689.
10. NITI Aayog. Responsible AI for All: Adopting the Framework — A Use Case Approach for All. NITI Aayog: New Delhi, 2024.
11. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-453.
12. Menon GR, Singh L, Sharma P, et al. National Burden Estimates of Healthy Life Lost in India, 2017. Indian Journal of Medical Research. 2019;150(2):116-128.
13. Obermeyer Z, Emanuel EJ. Predicting the future — big data, machine learning, and clinical medicine. New England Journal of Medicine. 2016;375:1216-1219.
14. Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. Nature Medicine. 2019;25:1337-1340.
15. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837-845.
16. Chouldechova A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data. 2017;5(2):153-163.
17. Mitchell M, Wu S, Zaldivar A, et al. Model Cards for Model Reporting. Proceedings of FAT 2019. ACM: New York, 2019.
18. Gebru T, Morgenstern J, Vecchione B, et al. Datasheets for datasets. Communications of the ACM. 2021;64(12):86-92.
19. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems. 2017;30.
20. Ribeiro MT, Singh S, Guestrin C. 'Why should I trust you?' Explaining the predictions of any classifier. Proceedings of KDD 2016. ACM: New York, 2016.
21. Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. WACV 2018. IEEE, 2018.
22. Kim B, Wattenberg M, Gilmer J, et al. Interpretability beyond classification: Quantitative testing with concept activation vectors (TCAV). Proceedings of ICML 2018. PMLR, 2018.