SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Artificial Intelligence and Machine Learning in Indian Health Insurance: A Critical Analysis and the SUTRA Framework for Risk Assessment and Moral Hazard Containment

Dr. Ravi Jaiswal

Assistant Professor, National Insurance Academy, Pune, ravi.j@niapune.org.in

ABSTRACT

Purpose: Indian health insurance is expanding rapidly while remaining structurally loss-making in parts of the market, and both insurers and regulators increasingly attribute the gap to moral hazard on the demand side, supplier-induced demand on the provider side, and fraud at the interface between them. This paper critically examines what artificial intelligence (AI) and machine learning (ML) are actually delivering in the Indian health insurance market, as distinct from what is claimed for them, and develops an integrated framework through which Insurtech firms and insurers can measure risk and contain moral hazard without converting analytics into an instrument of claim denial.

Design/methodology/approach: The study adopts a pragmatist, mixed-methods design executed in two stages. Stage one is a systematic critical review of 96 sources covering the health economics of moral hazard, the AI-in-insurance literature, and Indian regulatory and industry documentation, synthesised into a gap analysis. Stage two is theory-building: the paper constructs the SUTRA framework (Sensing, Underwriting, Triage, Restraint, Assurance) and, within it, a composite Moral Hazard Exposure Index (MHEI) that decomposes exposure into demand-side, supply-side and intermediation components. A complete empirical validation protocol is specified, including construct operationalisation, sampling design, and the statistical procedures (exploratory and confirmatory factor analysis, PLS-SEM, binary logistic regression, propensity-score matching and difference-in-differences) required to test the eight hypotheses derived from the framework.

Findings: The critical analysis identifies a systematic asymmetry: AI adoption in Indian health insurance is concentrated in cost-reducing and claim-restricting applications, while applications that would reduce moral hazard at its source — prevention, care navigation, provider contracting — remain marginal. Three structural constraints explain this: fragmented and non-interoperable claims data, the absence until recently of any AI governance framework, and an incentive structure in which the returns to denying a claim are immediate and measurable whereas the returns to preventing one are deferred and diffuse. The framework responds by embedding a proportionality constraint that ties the intensity of any restraint action to the strength and auditability of the evidence supporting it.

Originality/value: The paper offers the first integrated framework that treats moral hazard containment in Indian health insurance as a measurement problem rather than a detection problem, distinguishes the three loci of hazard rather than collapsing them into "fraud", and specifies a graduated intervention ladder that is auditable against IRDAI's emerging AI governance expectations and the Digital Personal Data Protection Act, 2023.

Keywords: health insurance; artificial intelligence; machine learning; moral hazard; insurtech; India; IRDAI; risk assessment; supplier-induced demand; digital public infrastructure

Introduction

1.1 The Indian health insurance paradox

Indian health insurance presents a paradox that has become steadily harder to explain away. The segment is the fastest-growing line of non-life business in the country, with gross premium of ₹1,27,417 crore in FY 2024–25, a growth of 9.19 per cent over the previous year, and it now accounts for 41.42 per cent of all non-life premium underwritten in India (IRDAI, 2025). At the same time, the industry-wide incurred claims ratio for non-life business stood at 82.88 per cent, with public sector general insurers reporting a combined ratio of 99.84 per cent — a figure that, once management expenses of roughly 15.60 per cent of premium are added, describes a business that pays out more than it collects. Growth and structural unprofitability are occurring simultaneously, in the same product, at the same time.

Two explanations dominate industry discourse. The first is medical inflation, which IRDAI places in the region of 12 to 14 per cent annually, well above general price inflation. The second is behavioural: that the presence of insurance changes what policyholders demand and what hospitals supply, in ways that inflate claims beyond what the underlying morbidity would justify. This second explanation is the domain of moral hazard, and it is the concern of this paper.

The behavioural explanation is not merely an industry grievance. Industry estimates cited in support of IRDAI's 2025 Insurance Fraud Monitoring Framework suggest that approximately 15 per cent of health insurance claims contain some element of fraud (Ankura, 2025). The regulator's own data show that of roughly ₹1.17 lakh crore in health claims filed in FY 2023–24, only 71.29 per cent were paid; ₹15,100 crore (12.9 per cent) was

disallowed and a further ₹10,937 crore (9.34 per cent) repudiated (IRDAI, 2024). Whether one reads those disallowance figures as evidence of successful fraud control or as evidence of aggressive claim restriction depends entirely on which side of the transaction one occupies — and that ambiguity is precisely the problem this paper takes up.

1.2 Why moral hazard is a measurement problem, not a detection problem

The industry's technological response to rising claims has been to invest in detection. Machine learning models are trained to flag anomalous claims, unusual provider billing patterns, and clusters of high-cost procedures at particular hospitals. IRDAI's 2025 framework formalises this direction by requiring insurers to develop Red Flag Indicators calibrated to their own business profile and to move from post-hoc investigation towards predictive architectures (Ankura, 2025).

Detection, however, addresses the symptom at the last possible moment. A claim reaching the adjudication queue represents a hospitalisation that has already occurred, a bed that has already been occupied, and a set of clinical decisions that have already been made. Flagging it can prevent a payment; it cannot prevent the event. More seriously, detection systems are structurally biased towards a particular kind of error. Because the cost of a false negative (a fraudulent claim paid) is immediately visible in the loss ratio while the cost of a false positive (a legitimate claim denied) is diffuse, delayed, and borne largely by the policyholder, an insurer optimising narrowly on claims cost will tolerate a false-positive rate that is socially excessive. The rise in grievances — up 20 per cent to 2.57 lakh complaints in FY 2024–25, with claims accounting for 69 per cent of general insurance complaints — is consistent with this dynamic.

This paper therefore argues for a reframing. Moral hazard should be treated as a measurable exposure attaching to a policy, a provider and a distribution channel, quantified continuously and acted on proportionately, rather than as a binary property of individual claims to be detected at the point of settlement. This reframing has three consequences. It moves intervention upstream, where prevention is possible. It makes the insurer's response auditable, because a graduated response to a measured exposure can be justified in a way that a binary denial often cannot. And it separates the three distinct loci of hazard — policyholder, provider, and intermediary — which current practice collapses into the single undifferentiated category of "fraud".

1.3 The insurtech opportunity and its unrealised portion

India's digital public infrastructure has, within a short period, removed the principal technical obstacle to this reframing. The National Health Claims Exchange (NHCX), built by the National Health Authority in consultation with IRDAI on HL7 FHIR R4 standards and anchored to Ayushman Bharat Health Account identifiers, provides a single protocol through which hospitals, insurers and third-party administrators exchange structured claims data. As of April 2026 the NHCX dashboard reported 83 registered payers, 42,687 provider facilities, and more than 23.4 million claims processed (Caladrius Health, 2026); by May 2026, 160 integrators and over 12,600 hospitals had been onboarded (DreamSoft4u, 2026). For the first time, the data required to measure provider-level utilisation patterns across insurers — rather than within a single insurer's book — is becoming institutionally available.

Yet the insurtech applications that have attracted the most capital and attention in India remain concentrated at the two ends of the value chain that are easiest to digitise: distribution, where aggregators and digital brokers have transformed how policies are sold, and claims triage, where automation has compressed turnaround times. The middle of the chain — risk measurement, prevention, provider contracting and care navigation, which is where moral hazard is actually generated — remains comparatively untouched. This paper treats that gap as the central object of analysis.

1.4 Research questions

The study is organised around five questions.

1. RQ1. What is the current state of AI and ML deployment in Indian health insurance, and how does the pattern of deployment compare with the pattern of claims to which the industry attributes its losses?
2. RQ2. Which structural, informational and regulatory constraints explain the observed pattern of adoption?
3. RQ3. How can moral hazard in Indian health insurance be decomposed and operationalised as a measurable construct across its demand-side, supply-side and intermediation loci?
4. RQ4. What framework can integrate the available data infrastructure, analytical capability and regulatory constraints into a coherent system for measuring risk and containing moral hazard?
5. RQ5. What safeguards are required to ensure such a system does not degenerate into an instrument of claim restriction, and how can compliance with those safeguards be evidenced to a regulator?

1.5 Contribution

The paper makes four contributions. First, it provides a critical rather than promotional assessment of AI adoption in Indian health insurance, grounded in regulatory data and documented deployment rather than vendor claims. Second, it develops the SUTRA framework — Sensing, Underwriting, Triage, Restraint, Assurance — which organises Insurtech capability into five layers with specified inputs, outputs and

governance obligations at each. Third, it constructs the Moral Hazard Exposure Index (MHEI), a composite measure that decomposes exposure into demand-side, supply-side and intermediation components, each built from indicators that are observable in NHCX-standard claims data. Fourth, it specifies a Graduated Restraint Ladder governed by an explicit proportionality constraint, which is the paper's principal normative contribution: it establishes that the intensity of any adverse action must be bounded by the strength and auditability of the evidence supporting it.

1.6 Structure of the paper

Section 2 sets out the conceptual background on moral hazard and its Indian institutional setting. Section 3 reviews the literature across five streams and derives the research gap. Section 4 presents the critical analysis of AI adoption. Section 5 details the research methodology, including the empirical validation protocol and statistical tools. Section 6 develops the SUTRA framework, the MHEI and the associated hypotheses. Section 7 discusses theoretical, managerial and policy implications. Section 8 states limitations and directions for future research, and Section 9 concludes.

2. Conceptual Background: Moral Hazard in Health Insurance

2.1 The classical formulation

Moral hazard in insurance describes the change in behaviour that follows from the transfer of risk. Arrow (1963) first identified the phenomenon in medical care, observing that insurance reduces the marginal price of treatment at the point of consumption and therefore alters the quantity demanded. Pauly (1968) sharpened the argument by demonstrating that this response is not moral failing but rational price response: if a service that costs ₹1,00,000 is available at a co-payment of ₹5,000, the consumer faces a price of ₹5,000 and will consume up to the point where marginal benefit equals that price, not the point where it equals ₹1,00,000. Zeckhauser (1970) formalised the resulting trade-off between risk-spreading and incentive provision that any insurance contract must resolve.

The literature distinguishes two temporal forms. Ex-ante moral hazard is the reduction in preventive effort that follows from being insured — less attention to diet, exercise, screening or medication adherence, because the financial consequences of illness are now borne by the insurer. Ex-post moral hazard is the increase in the quantity of care consumed once illness has occurred — a longer stay, a private room, a more expensive implant, an additional diagnostic panel. The two have very different policy implications. Ex-ante hazard is addressed by prevention and incentive design; ex-post hazard is addressed by cost-sharing, utilisation review and provider contracting.

Empirical estimates of the magnitude of ex-post moral hazard rest heavily on the RAND Health Insurance Experiment (Manning et al., 1987), which found that participants facing higher cost-sharing consumed substantially less care, with a price elasticity of demand for medical care of approximately −0.2. Subsequent work using the Oregon Health Insurance Experiment (Finkelstein et al., 2012) confirmed substantial utilisation responses to coverage in a modern setting. The persistent difficulty, addressed in Section 2.3, is that observed utilisation differences between insured and uninsured populations confound moral hazard with adverse selection and with genuine unmet need.

2.2 The third locus: supplier-induced demand

The classical formulation places the behavioural change with the patient. In health care, this is at best half the story. The patient rarely chooses the quantity of care; the physician does, in a relationship characterised by profound information asymmetry. Where the physician also benefits financially from the volume prescribed, the conditions for supplier-induced demand are present (Evans, 1974; McGuire, 2000).

In the Indian context this locus is arguably dominant. The private hospital sector, which delivers the majority of insured inpatient care, operates largely on fee-for-service reimbursement in which revenue rises with the number and intensity of billable interventions. The documented forms are well catalogued: upcoding of room categories and procedure codes; unbundling of package rates into separately billed components; extension of length of stay to satisfy the 24-hour hospitalisation requirement that most indemnity policies impose; and, at the fraudulent extreme, phantom billing for admissions and procedures that did not occur.

Evidence from India's publicly funded schemes is instructive but mixed. Studies of the Rashtriya Swasthya Bima Yojana and its successor found consistent increases in utilisation following enrolment, but the evidence that these schemes reduced out-of-pocket or catastrophic expenditure is inconclusive (Sriram & Khan, 2020; Reshmi et al., 2021). Work on PM-JAY in Chhattisgarh found that enrolment neither increased hospitalisations nor decreased out-of-pocket expenditure, with the authors attributing the result partly to provider capture and double billing (Garg et al., 2020). A matching-estimator study of a single Indian state found evidence consistent with both moral hazard and supplier-induced demand in publicly funded schemes (Ghosh & Mladovsky, 2024). Conversely, an interrupted time-series analysis of PM-JAY spine surgery in Punjab found that increased utilisation was unlikely to be supplier-induced, since private-sector provider behaviour did not differ significantly from public-sector behaviour (Prinja et al., 2024). The literature, in short, establishes that the mechanism operates in India but has not settled its magnitude — a gap that motivates the measurement

approach developed here.

2.3 The identification problem

A central methodological difficulty runs through the entire literature and must be confronted by any framework that claims to measure moral hazard. Higher claims among the insured relative to the uninsured may reflect three distinct phenomena. It may reflect moral hazard, where coverage changes behaviour. It may reflect adverse selection, where individuals who anticipate needing care are disproportionately likely to purchase coverage. Or it may reflect the correction of unmet need, where coverage enables treatment that was previously forgone for financial reasons — an outcome that is the intended purpose of insurance and not a pathology at all.

Failing to separate these is not a technical nicety. An insurer that misclassifies the correction of unmet need as moral hazard will suppress exactly the utilisation that its product exists to enable, and will do so disproportionately among lower-income and first-time policyholders whose prior utilisation was suppressed by cost. This is the mechanism by which a moral hazard control system becomes a regressive denial system. Section 5.5 accordingly specifies propensity-score matching and difference-in-differences designs as mandatory rather than optional components of the validation protocol, and Section 6.6 embeds the distinction directly into the framework through the proportionality constraint.

2.4 The Indian institutional setting

Five features of the Indian market shape how moral hazard manifests and how it can be addressed.

Persistent underinsurance alongside rapid growth. Overall insurance penetration in India declined for a second consecutive year to 3.7 per cent of GDP in 2023–24, against a world average roughly double that figure (IRDAI, 2024). A large uninsured and newly insured population means that a substantial share of observed utilisation growth is the correction of unmet need rather than moral hazard — making the identification problem of Section 2.3 acute rather than academic.

High out-of-pocket expenditure. Out-of-pocket expenditure fell from 64.2 per cent of total health expenditure in 2013–14 to 39.4 per cent in 2021–22 (National Health Accounts, 2024). The decline is substantial, but the residual share remains high by international standards and means that insurance sits alongside, rather than in place of, direct payment — which weakens the price-response mechanism at the heart of the classical model.

The inpatient bias of indemnity products. Most Indian retail health products reimburse hospitalisation and exclude or narrowly cap outpatient care. This creates a structural distortion: since reimbursement requires admission, both patients and providers face an incentive to convert what could be managed on an outpatient basis into an admission. A meaningful share of what is recorded as ex-post moral hazard in Indian claims data is better understood as an artefact of product design.

Intermediated adjudication. In FY 2023–24, 72 per cent of health claims by number were settled through third-party administrators and 28 per cent through in-house mechanisms (IRDAI, 2024). The TPA sits between insurer and provider with objectives that align imperfectly with either, creating a third locus of hazard that the two-party classical model does not capture and that this paper treats explicitly.

A newly interoperable data layer. NHCX, ABHA and the Ayushman Bharat Digital Mission have begun to make claims data structured, standardised and portable across payers. This is the enabling condition for cross-insurer measurement of provider behaviour, without which supply-side moral hazard is invisible to any single insurer.

Table 1. Three loci of moral hazard in Indian health insurance and their observable signatures

Locus	Mechanism	Observable signature in claims data	Conventional control
Demand-side (policyholder)	Reduced prevention (ex-ante); higher intensity of care sought once ill (ex-post)	Claim clustering shortly after waiting-period expiry; room-category upgrades; repeated short admissions; low preventive-service uptake	Co-payment, deductible, sub-limits, waiting periods, no-claim bonus
Supply-side (provider)	Fee-for-service revenue rises with billed volume and intensity; information asymmetry over clinical necessity	Length of stay above peer median for same DRG; procedure-mix skew; unbundling of package rates; implant and consumable cost outliers	Package rates, empanelment conditions, pre-authorisation, utilisation review

Locus	Mechanism	Observable signature in claims data	Conventional control
Intermediation (TPA, agent, broker)	Compensation tied to volume settled or sold rather than to portfolio outcome	Adjudication turnaround inconsistent with claim complexity; disclosure quality varying systematically by channel; early-duration claim concentration by agent code	Commission structure, service-level agreements, channel audit

3. Literature Review

3.1 Review protocol

The review follows a structured protocol adapted from Tranfield et al. (2003). Searches were run across Scopus, Web of Science, EBSCO Business Source and Google Scholar using combinations of the terms health insurance, moral hazard, artificial intelligence, machine learning, insurtech, fraud detection, India, and supplier-induced demand, restricted to the period 2015–2026 for the technology streams and unrestricted for the foundational health-economics stream. Regulatory and industry documentation — IRDAI annual reports and circulars, National Health Authority publications, and consultancy reports — was included as grey literature because peer-reviewed coverage of the Indian institutional setting lags practice by several years. After removal of duplicates and screening for relevance, 96 sources were retained and coded into five streams.

3.2 Stream 1: AI and ML in underwriting and risk assessment

The application of machine learning to insurance underwriting is well established internationally. Gradient-boosted tree ensembles have repeatedly been shown to outperform generalised linear models on tabular insurance data (Chen & Guestrin, 2016; Noll et al., 2020), and Wüthrich and Merz (2023) provide the definitive treatment of how these methods integrate with actuarial pricing practice. The literature on the actuarial implications is more cautious: Frees et al. (2014) and Denuit et al. (2019) emphasise that predictive accuracy and pricing adequacy are distinct objectives, and that a model optimised for the former may be inadequate for the latter.

Indian-specific empirical work in this stream is thin. Most published studies apply standard classifiers to publicly available or single-insurer datasets and report accuracy metrics without addressing rate-filing constraints, IRDAI product-approval requirements, or the cross-subsidy expectations embedded in Indian retail health products. The absence of work connecting ML-based risk scores to the actual regulatory pricing process is a significant gap.

3.3 Stream 2: AI and ML in claims processing and fraud detection

Fraud detection is the most mature application area. Surveys by Abdallah et al. (2016) and Şahin et al. (2013) map the standard toolkit — supervised classification on labelled fraud cases, unsupervised anomaly detection where labels are scarce, and network or graph methods that model providers, patients and procedures jointly. Graph-based approaches have proved particularly effective at detecting collusive fraud invisible to per-claim classifiers (Van Vlasselaer et al., 2017), which is directly relevant to the provider-network structure of Indian cashless claims.

Two limitations recur across this literature and are rarely acknowledged. The first is the label problem: models are trained on claims that were investigated and confirmed as fraudulent, which is a biased sample of actual fraud, since investigation itself is triggered by existing heuristics. The model therefore learns to reproduce the investigators' priors rather than to discover fraud they missed. The second is evaluation on rebalanced data: reported accuracy figures are frequently obtained on artificially balanced datasets that bear no resemblance to a live claims queue in which the fraud base rate may be one or two per cent, and such figures do not transfer.

In the Indian setting, IRDAI's 2025 Insurance Fraud Monitoring Framework has moved this stream from optional to mandatory by requiring insurer-specific Red Flag Indicators and predictive rather than reactive architectures (Ankura, 2025). The regulatory instruction is clear; the methodological guidance on how to build such systems without amplifying the two biases above is not.

3.4 Stream 3: Insurtech, behavioural pricing and wellness-linked products

The insurtech literature addresses how digital intermediation reshapes insurance markets (Eling & Lehmann, 2018; Cappiello, 2020; Stoeckli et al., 2018). A distinct sub-stream examines behaviour-linked pricing, in which continuously observed conduct feeds into premium adjustment. The motor telematics literature is the most developed (Verbelen et al., 2018; Guillen et al., 2021) and establishes that high-frequency behavioural data carries predictive power beyond conventional rating factors. The health analogue — wearables-based wellness programmes offering premium discounts or reward points for step counts, screening completion and medication adherence — is now common in Indian retail health products.

The critical literature raises three objections that the Indian implementations have largely not answered.

Wellness programmes may select for the already-healthy rather than change behaviour, delivering discounts to those who would have exercised anyway (Jones et al., 2019). Continuous monitoring raises consent and surveillance concerns that intensify under the Digital Personal Data Protection Act, 2023. And behavioural pricing risks penalising conditions that are structurally rather than volitionally determined — an office worker with a long commute in a congested Indian metropolitan area faces different achievable step counts than a suburban professional, and a step-count-linked discount silently prices that difference.

3.5 Stream 4: Digital public infrastructure in Indian health

The literature on India Stack and digital public infrastructure has expanded rapidly, though most of it treats identity and payments rather than health. The health layer — the Ayushman Bharat Digital Mission, ABHA identifiers, the Unified Health Interface and NHCX — is documented principally in policy and industry sources rather than peer-reviewed work. NHCX is designed to replace fragmented insurer-specific portals with a single FHIR R4-based protocol for coverage verification, pre-authorisation, claim submission and settlement (NATHEALTH, 2025; India InsurTech Summit, 2025).

Adoption data indicate real but uneven progress. By April 2026 the NHCX dashboard reported 83 payers, 42,687 provider facilities and more than 23.4 million claims processed (Caladrius Health, 2026), and by May 2026, 160 integrators and over 12,600 hospitals had been onboarded (DreamSoft4u, 2026). Onboarding, however, is distinct from live claim flow, and participation remains voluntary, incentivised through the Digital Health Incentive Scheme at up to ₹500 per claim or 10 per cent of claim value, whichever is lower. Persistent frictions are documented: many smaller hospitals still run paper-based or non-ABDM-compliant systems, and a nationwide survey found that 60 per cent of health insurance claimants experienced delays of six to 48 hours between claim approval and discharge, despite IRDAI's three-hour cashless settlement requirement (LocalCircles, 2026).

What is missing from this stream is any treatment of NHCX as an analytical rather than merely transactional asset. Standardised cross-payer claims data makes provider-level utilisation benchmarking possible for the first time; no published work has considered what this implies for moral hazard measurement.

3.6 Stream 5: Algorithmic accountability, fairness and the critical perspective

A substantial critical literature documents how algorithmic systems in health and insurance reproduce and amplify existing disadvantage. Obermeyer et al. (2019) demonstrated that a widely deployed US health risk-prediction algorithm systematically understated the needs of Black patients because it used historical cost as a proxy for health need, and historical cost reflected differential access rather than differential illness. This finding is directly transferable to India, where using historical claims expenditure as a proxy for health risk would systematically understate the needs of populations whose prior utilisation was suppressed by cost or distance.

The fairness literature supplies formal criteria — demographic parity, equalised odds, calibration within groups — together with impossibility results establishing that these cannot generally be satisfied simultaneously (Hardt et al., 2016; Chouldechova, 2017; Kleinberg et al., 2017). The explainability literature offers post-hoc attribution methods (Lundberg & Lee, 2017; Ribeiro et al., 2016), while Rudin (2019) argues forcefully that for high-stakes decisions such methods are inadequate substitutes for intrinsically interpretable models.

India's regulatory position on these questions was, until recently, largely undefined. The Digital Personal Data Protection Act, 2023 established consent and purpose-limitation obligations, and MeitY's AI Governance and Guidelines Development report of January 2025 mapped gaps in existing frameworks. The decisive development came on 18 June 2026, when IRDAI constituted a seven-member working group on artificial intelligence, chaired by Sandeep K. Shukla of IIIT Hyderabad, with a three-month mandate to develop an AI governance framework covering ethical, transparent and explainable use of AI in claims management and fraud detection, together with a pre- and post-deployment AI audit framework (IRDAI, 2026; Business Standard, 2026). This followed IRDAI's revised information and cyber security guidelines of April 2026. The framework developed in this paper is designed to be auditable against precisely these emerging expectations.

3.7 Synthesis and research gap

Table 2 positions the five streams against the requirements of an integrated moral hazard containment system. The pattern is consistent: each stream addresses the problem it was constituted to address and is silent on the others. Fraud detection is sophisticated about anomaly identification and silent about prevention. Behavioural pricing addresses ex-ante hazard on the demand side and ignores the supply side entirely. The digital-infrastructure literature is transactional rather than analytical. The critical literature diagnoses harms without offering an operational alternative.

Table 2. Positioning of literature streams against requirements for an integrated moral hazard framework

Requirement	S1 Underwriting	S2 Fraud	S3 Insurtech	S4 DPI	S5 Critical
R1. Distinguishes demand-, supply- and intermediation-side hazard	No	Partial	No	No	No
R2. Separates moral hazard from adverse selection and unmet need	Partial	No	No	No	Partial
R3. Operates preventively rather than at point of settlement	Partial	No	Yes	No	No
R4. Uses cross-payer provider benchmarking	No	Partial	No	Partial	No
R5. Constrains adverse action proportionately to evidence strength	No	No	No	No	Partial
R6. Auditable against IRDAI and DPDP obligations	No	Partial	No	Partial	Partial
R7. Calibrated to Indian product design and institutional structure	No	Partial	Partial	Yes	No

Note. S1–S5 correspond to the streams reviewed in Sections 3.2–3.6. DPI = digital public infrastructure. "Partial" indicates the stream addresses the requirement incidentally or for a restricted case.

Three gaps follow. The conceptual gap is that no framework in the literature treats the three loci of moral hazard as distinct measurable constructs requiring distinct interventions. The methodological gap is that no operational measure exists that separates moral hazard from adverse selection and unmet need using data that an Indian insurer actually possesses. The normative gap is that no framework specifies what limits should constrain an insurer acting on an algorithmic risk signal. This paper addresses all three.

4. Critical Analysis of AI Adoption in Indian Health Insurance

This section addresses RQ1 and RQ2. It sets what is claimed for AI in Indian health insurance against what is demonstrably deployed, and identifies the constraints that explain the difference.

4.1 The asymmetry of deployment

Reviewing insurer disclosures, regulatory documentation and industry reporting yields a consistent picture. AI and ML deployment in Indian health insurance clusters in four areas: document processing and optical character recognition for claim intake; rules-plus-ML triage that routes claims to automated or manual adjudication; fraud and anomaly scoring; and customer-service automation through chatbots and voice systems. All four share a common property — they reduce the insurer's administrative cost per claim or reduce the amount paid out per claim.

Applications that would reduce moral hazard at its source are markedly less developed. Provider-level utilisation benchmarking across insurers, chronic-disease care navigation that substitutes managed outpatient care for admission, predictive intervention on policyholders showing early deterioration, and outcome-linked provider contracting are each technically feasible with available data and each remain marginal in practice. Table 3 sets out the asymmetry.

Table 3. Asymmetry between deployed and under-deployed AI applications in Indian health insurance

Application	Maturity in India	Primary beneficiary	Effect on moral hazard
OCR and document intake automation	High	Insurer (cost)	None — processes claims already generated

Application	Maturity in India	Primary beneficiary	Effect on moral hazard
Rules-plus-ML claim triage	High	Insurer (cost, speed)	None — accelerates the same decisions
Fraud and anomaly scoring	Moderate to high	Insurer (loss ratio)	Deters detected forms; displaces rather than prevents
Conversational service automation	High	Insurer (cost)	None
Wellness and step-count reward programmes	Moderate	Insurer and healthy policyholders	Targets ex-ante hazard; selection effects unproven
Cross-payer provider utilisation benchmarking	Low	Insurer, regulator, honest providers	Directly addresses supply-side hazard
Chronic-disease care navigation	Low	Policyholder and insurer jointly	Addresses both ex-ante and ex-post hazard at source
Predictive early intervention on deteriorating risk	Very low	Policyholder and insurer jointly	Prevents the admission rather than contesting the claim
Outcome-linked provider contracting	Very low	Insurer, policyholder	Removes the volume incentive generating supply-side hazard

Note. Maturity assessed from insurer public disclosures, IRDAI regulatory filings, and industry reporting through mid-2026. Classification is indicative and would benefit from systematic primary verification (see Section 5.3).

The pattern is not accidental, and it is not explained by technical difficulty. Provider benchmarking is computationally simpler than fraud detection. The explanation lies in the structure of returns.

4.2 Constraint one: the asymmetric returns to denial and prevention

The return to denying a claim is immediate, attributable and measurable. A denied claim of ₹3 lakh appears in this quarter's loss ratio, and the analytics team that flagged it can claim the credit. The return to preventing a claim is deferred, diffuse and counterfactual. A care-navigation programme that averts an admission produces a claim that never appears — and demonstrating that it would have appeared requires a control group, a time horizon exceeding the typical annual policy term, and an attribution methodology that most insurers do not maintain.

This asymmetry is compounded by India's annual retail health policy cycle and high portability. An insurer investing in preventive management of a diabetic policyholder bears the cost immediately and captures the benefit only if that policyholder renews. Where portability is easy and price competition intense, a meaningful share of the benefit accrues to whichever insurer holds the policy when the averted event would have occurred. Prevention thus exhibits a positive externality across insurers, and, as standard theory predicts, is undersupplied relative to the social optimum. This is a structural market failure, not a managerial failing, and it has a direct implication: prevention-oriented analytics will remain undersupplied until either policy terms lengthen or the benefit is collectivised through an industry-level or regulator-convened mechanism.

4.3 Constraint two: data fragmentation and the label problem

Until NHCX, each insurer saw only its own book. A hospital billing four insurers appeared to each as a moderately active provider; the aggregate pattern was legible to none. Cross-payer measurement of provider behaviour — the precondition for identifying supply-side moral hazard — was structurally impossible.

NHCX changes the availability of the data but not, by itself, the governance of its analytical use. Two questions remain unresolved. Who is permitted to compute cross-payer provider benchmarks, and under what authority? And what recourse does a provider have when a benchmark computed from pooled data triggers adverse consequences — heightened pre-authorisation scrutiny, revised package rates, or de-empanelment? These are competition-law and natural-justice questions as much as technical ones, and the framework in Section 6 addresses them through the assurance layer rather than treating them as external.

The label problem compounds the data problem. Indian insurers' fraud labels derive from investigations triggered by existing heuristics — typically claim size, provider reputation, or policy duration. A supervised model trained on these labels learns the investigators' priors with high fidelity and discovers little that they missed. Worse, if past investigation focused disproportionately on particular hospital tiers or geographies, the

model encodes that focus as a risk factor and reproduces it at scale with the appearance of objectivity.

4.4 Constraint three: the regulatory vacuum and its recent closure

Until mid-2026, Indian insurers deployed AI in consequential decisions under no sector-specific governance framework. The applicable obligations were general: the DPDP Act, 2023 for personal data; IRDAI's information and cyber security guidelines, revised in April 2026; and the protection-of-policyholders'-interests regulations, which impose disclosure and grievance obligations but were not drafted with algorithmic decision-making in view.

The constitution of IRDAI's AI working group on 18 June 2026, chaired by Sandeep K. Shukla with members drawn from CERT-In, ReBIT, SBI Life, Star Health and ICICI Lombard, marks the closure of that vacuum. Its mandate — to map current adoption, produce a governance framework ensuring ethical, transparent and explainable AI use in claims and fraud detection, and design a pre- and post-deployment audit framework — defines the standards against which existing deployments will be assessed (IRDAI, 2026). Comparative analysis places India among the jurisdictions still formulating their approach, alongside the United Kingdom and the United States, and behind Singapore's MAS framework and the European Union's AI Act, whose binding high-risk obligations for insurance were deferred to December 2027 under the Digital Omnibus agreement.

Two implications follow for practice. First, insurers with AI already in production face retrospective assessment against standards not yet published; systems built without explainability and audit logging may require rebuilding. Second, the working group's emphasis on a pre- and post-deployment audit framework signals that the operative regulatory question will be documentation of process, not merely accuracy of outcome. A framework designed to generate that documentation as a by-product of operation, rather than to reconstruct it under examination, has a material compliance advantage. This design objective is built into the assurance layer of Section 6.

4.5 The denial-machine risk

The most serious critique of the current trajectory is that AI in Indian health insurance risks becoming an efficient denial machine. Three pieces of evidence sustain the concern. Health claim rejections and disallowances rose 19.10 per cent to ₹26,000 crore in FY 2023–24 (IRDAI, 2024). Grievances rose 20 per cent to 2.57 lakh in FY 2024–25, with claims accounting for 69 per cent of general insurance complaints. And 60 per cent of claimants reported discharge delays of six to 48 hours despite the mandated three-hour cashless turnaround (LocalCircles, 2026).

None of these figures establishes that AI caused the deterioration; they are consistent with rising claim volumes, medical inflation, and more claims of genuinely doubtful admissibility. But they establish the environment in which AI is being introduced, and the direction of the incentive gradient within it. A model trained to minimise claims cost, deployed in an organisation measured on loss ratio, in a market where the policyholder's recourse is a grievance process with months of delay, faces no counterweight to over-restriction. The absence of a counterweight is not remedied by better model accuracy. It requires a structural constraint, which is what Section 6.6 supplies.

4.6 Equity and the digital divide

A final critical consideration concerns distribution. Every mechanism discussed — wearables-linked pricing, digital claim submission, app-based care navigation, ABHA-linked record portability — presupposes smartphone access, digital literacy, reliable connectivity and, frequently, English or Hindi proficiency. These are unevenly distributed across India by income, geography, gender and age.

The consequence is that benefits accrue disproportionately to those already advantaged, while costs fall on those least able to bear them. A policyholder who cannot operate a wellness application forgoes the discount; a rural claimant treated at a non-ABDM-compliant hospital faces manual processing and longer settlement; an elderly claimant unable to navigate a digital appeal is less likely to contest an incorrect denial. Because the population underserved by these mechanisms substantially overlaps with the population whose prior utilisation was suppressed by cost, and whose current utilisation therefore represents the correction of unmet need rather than moral hazard, the two failure modes compound. A poorly specified system will identify the newly insured rural policyholder as a moral hazard risk precisely because coverage is doing what it was designed to do.

This is why Section 6 treats equity constraints as constitutive of the framework rather than as an appended compliance obligation.

5. Research Methodology

5.1 Research philosophy and design

The study adopts a pragmatist philosophical position. Pragmatism holds that the choice of method should follow from the research problem rather than from prior ontological commitment, and it accommodates the combination of interpretive and positivist elements that this problem requires (Creswell & Plano Clark, 2018). The problem has an objective component — claims patterns, utilisation rates, settlement ratios — that is

amenable to quantitative measurement, and a socially constructed component — what counts as "unnecessary" care, what an insurer owes a policyholder — that is not.

The design is exploratory sequential mixed-methods, executed in two stages of which the first is complete and reported here and the second is specified for subsequent execution.

- Stage 1 (completed, reported in Sections 3, 4 and 6). Systematic critical review of 96 sources and analysis of publicly available secondary data, synthesised into a gap analysis and used to construct the SUTRA framework and the Moral Hazard Exposure Index.
- Stage 2 (specified in Sections 5.3–5.6, not yet executed). Empirical validation combining expert interviews, a policyholder survey, and analysis of insurer claims data, to test the hypotheses of Section 6.7.

This distinction is stated explicitly and repeated in Section 8, because the contribution claimed for this paper is conceptual and methodological. No empirical results are reported, and none should be inferred.

5.2 Secondary data sources

The critical analysis of Section 4 draws on the following secondary sources, all publicly available and independently verifiable.

Table 4. Secondary data sources used in the critical analysis

Source	Publisher	Data drawn	Period
Annual Report	IRDAI	Segment premium, incurred claims ratios, settlement and repudiation ratios, grievance volumes, penetration, TPA share of settlement	FY 2019–20 to FY 2024–25
Handbook on Indian Insurance Statistics	IRDAI	Company-level time series for panel construction	FY 2015–16 onward
Insurance Fraud Monitoring Framework	IRDAI	Regulatory requirements on Red Flag Indicators and predictive fraud architecture	2025
Working group constitution on AI	IRDAI	Governance and audit framework mandate and composition	June 2026
NHCX adoption dashboard	National Health Authority	Payer, provider and claim-volume onboarding counts	To May 2026
National Health Accounts estimates	MoHFW	Out-of-pocket expenditure share, government health expenditure share	To FY 2021–22
PM-JAY scheme documentation and evaluations	NHA and peer-reviewed	Utilisation, provider behaviour, out-of-pocket outcomes	2018–2026

Note. All secondary figures cited in this paper are traceable to these sources. Where an industry estimate rather than an official statistic is used — notably the 15 per cent fraud-incidence figure — it is identified as such at the point of use.

5.3 Proposed primary data collection

5.3.1 Qualitative phase

Semi-structured interviews are proposed with 24 to 30 purposively sampled experts, distributed across six strata to ensure that no single perspective dominates: senior claims and analytics executives at general and standalone health insurers (n ≈ 8); third-party administrator operations heads (n ≈ 4); hospital revenue-cycle and insurance-desk managers (n ≈ 5); insurtech founders and product heads (n ≈ 5); actuaries and IRDAI-facing compliance professionals (n ≈ 4); and consumer-side representatives from ombudsman offices or consumer organisations (n ≈ 4). The final stratum is included deliberately: a study of moral hazard containment that interviews only those who contain it will systematically understate the costs of over-containment.

Interviews would be conducted until thematic saturation, transcribed, and analysed through Gioia methodology (Gioia et al., 2013), producing first-order concepts, second-order themes and aggregate dimensions. Intercoder reliability would be established on a 20 per cent subsample with a Cohen's κ threshold of 0.75.

5.3.2 Quantitative phase

A structured survey of retail health policyholders is proposed to test the demand-side components of the model. Using Cochran's formula for an infinite population at 95 per cent confidence and 5 per cent margin of

error, the minimum sample is 385; allowing for a design effect of 1.5 arising from clustered sampling across cities and an anticipated 70 per cent usable-response rate, a target of 825 distributed questionnaires yielding approximately 578 usable responses is proposed. Stratification would be by city tier (Tier 1 / Tier 2 / Tier 3), age band, and policy vintage, with the last of these essential to distinguishing new entrants correcting unmet need from established policyholders.

A parallel request would be made to two or three partner insurers for de-identified claims data covering a minimum of three policy years, structured to NHCX schema, sufficient to construct provider-level and policy-level indicators for the index of Section 6.5.

5.4 Constructs and instrument design

Seven latent constructs are proposed, each measured reflectively on a seven-point Likert scale using items adapted from validated instruments and pilot-tested on 40 respondents.

Table 5. Latent constructs, definitions and indicative measurement sources

Construct	Definition	Items	Adapted from
Perceived coverage generosity (PCG)	Policyholder's subjective sense of how comprehensively costs are covered	5	Adapted from RAND HIE cost-sharing salience measures
Ex-ante hazard propensity (EAH)	Self-reported reduction in preventive effort attributable to being insured	6	Newly developed; grounded in Pauly (1968) and Zeckhauser (1970)
Ex-post hazard propensity (EPH)	Stated willingness to accept higher-intensity care when insured	6	Newly developed; vignette-anchored
Provider influence salience (PIS)	Extent to which the treating provider rather than the patient determines care intensity	5	Adapted from supplier-induced demand literature (Evans, 1974)
Digital engagement capability (DEC)	Capability and willingness to use digital insurance and wellness tools	5	Adapted from UTAUT2 (Venkatesh et al., 2012)
Perceived procedural fairness (PPF)	Belief that claim decisions are made through a fair and contestable process	5	Adapted from Colquitt (2001) organisational justice scale
Institutional trust in insurer (ITI)	Confidence that the insurer will honour legitimate claims	4	Adapted from Morgan & Hunt (1994)

Note. EAH and EPH require newly developed items because no validated Indian-context instrument separates the two temporal forms of moral hazard. Vignette anchoring is used for EPH to reduce social desirability bias in self-reported over-utilisation.

5.5 Statistical tools and analytical procedures

The analysis plan proceeds in six steps, each matched to a specific inferential purpose. The tools are stated here so that the framework of Section 6 is testable rather than merely assertible.

5.5.1 Descriptive and preliminary analysis

Descriptive statistics, normality assessment through skewness and kurtosis within ±2.0, Levene's test for homogeneity of variance, and Harman's single-factor test together with a marker-variable procedure to assess common method bias. Non-response bias would be assessed by comparing early and late respondents.

5.5.2 Instrument validation

Exploratory factor analysis with principal axis factoring and promax rotation on a randomly split half of the sample, retaining factors with eigenvalues above unity and item loadings above 0.60, with Kaiser–Meyer–Olkin above 0.80 and a significant Bartlett's test. Confirmatory factor analysis on the holdout half, with model fit assessed against CFI and TLI above 0.90, RMSEA below 0.08 and SRMR below 0.08. Convergent validity requires average variance extracted above 0.50 and composite reliability above 0.70; discriminant validity is assessed through the Fornell–Larcker criterion and the heterotrait–monotrait ratio with a 0.85 threshold (Henseler et al., 2015).

5.5.3 Structural model estimation

Partial least squares structural equation modelling (PLS-SEM) is proposed for the structural model in preference to covariance-based SEM, for three reasons: the model is predictive rather than confirmatory of an

established theory, it contains a formatively specified composite (the MHEI) alongside reflective constructs, and PLS-SEM imposes weaker distributional assumptions (Hair et al., 2019). Path significance would be assessed by bootstrapping with 5,000 subsamples; explanatory power by R^2 and effect size f^2; predictive relevance by the Stone–Geisser Q^2 through blindfolding, and out-of-sample prediction by PLS-Predict.

5.5.4 Claim-level prediction

For the claims dataset, a binary logistic regression provides an interpretable baseline in which the probability that claim i exhibits excess utilisation is modelled as

$$logit(p_i) = \beta_0 + \beta_1 DMH_i + \beta_2 SMH_i + \beta_3 IMH_i + \gamma^T X_i \quad (1)$$

where DMH, SMH and IMH are the three sub-indices defined in Section 6.5 and X is a vector of controls comprising age, sex, policy vintage, sum insured, city tier, diagnosis-related group and hospital tier. Coefficients are reported as odds ratios with 95 per cent confidence intervals. A gradient-boosted tree ensemble would be estimated in parallel to establish the accuracy ceiling, with SHAP values (Lundberg & Lee, 2017) used to compare the variables the two models rely upon. Where the ensemble materially outperforms the logistic model, that gap is itself a finding about how much predictive signal is being purchased at the cost of interpretability — a trade-off that Rudin (2019) argues should be resolved in favour of interpretability in consequential decisions, and which IRDAI's explainability mandate may in any case require.

Because the outcome is rare, model evaluation would use precision–recall AUC rather than ROC-AUC, with the operating threshold selected on a cost matrix in which the cost of a false positive is set from the estimated welfare loss of an incorrectly restricted legitimate claim rather than from the insurer's administrative cost alone.

5.5.5 Causal identification

This step is the methodological core of the design, because it addresses the identification problem of Section 2.3. Two procedures are specified.

Propensity-score matching to separate moral hazard from adverse selection. Insured and comparable uninsured or newly insured individuals would be matched on observable health status, income, education, distance to facility and prior utilisation, following the approach of Barros et al. (2008) as applied in the Indian setting by Ghosh and Mladovsky (2024). The average treatment effect on the treated estimates the utilisation difference attributable to coverage among those who selected into coverage; Rosenbaum bounds would assess sensitivity to unobserved confounding.

Difference-in-differences to evaluate any insurtech intervention. Where an insurer introduces a wellness programme, a care-navigation service or a revised co-payment structure to part of its book, the change in utilisation among the treated group relative to an untreated comparison group over the same period identifies the intervention effect, conditional on the parallel-trends assumption being supported by pre-period testing. The estimating equation is

$$Y_{it} = \alpha + \delta\,(Treat_i \times Post_t) + \lambda_i + \mu_t + \theta^T X_{it} + \varepsilon_{it} \quad (2)$$

where Y is utilisation or claim cost, λ and μ are unit and time fixed effects, and δ is the effect of interest. Standard errors would be clustered at the level of treatment assignment. Where treatment timing is staggered across groups, the estimator of Callaway and Sant'Anna (2021) would be used in preference to two-way fixed effects, which is biased under heterogeneous treatment timing.

5.5.6 Index weighting

The weights combining the three sub-indices into the MHEI would be derived by two independent methods and compared. The entropy weight method derives weights from the information content of each indicator, is fully data-driven, and avoids expert bias. The analytic hierarchy process elicits pairwise comparative judgements from the expert panel of Section 5.3.1, with a consistency ratio below 0.10 required for acceptance. Agreement between the two is evidence that the weighting is robust; material disagreement is itself a finding about the divergence between what the data emphasises and what practitioners believe matters.

5.6 Validity, reliability and ethics

Content validity would be established through expert review by five academic and five industry reviewers, with the content validity ratio computed per item. Construct validity follows from the CFA procedures of Section 5.5.2. External validity is bounded by the sampling frame and is discussed as a limitation in Section 8. Reliability is assessed through Cronbach's alpha and composite reliability, with 0.70 as the threshold.

Ethical clearance would be obtained from the institutional review board before any data collection. All primary participation would be voluntary with written informed consent. All claims data would be de-identified before transfer, with processing under the Digital Personal Data Protection Act, 2023 on a documented lawful basis and with purpose limitation observed. Because health data is sensitive personal data, no attempt would be made to re-identify individuals, and outputs would be reported only in aggregate. Provider-level results would be reported anonymously, since identifying a named hospital as a high-hazard provider on the basis of a research instrument that has not been validated for that purpose would be both methodologically and ethically

improper.

6. The SUTRA Framework and the Moral Hazard Exposure Index

6.1 Design principles

The framework is built on five principles derived from the critical analysis of Section 4. Each principle responds to a specific failure identified there.

6. Measure, do not merely detect. Moral hazard is a continuous exposure attaching to a policy, a provider and a channel, not a binary property of a claim. Continuous measurement permits early, proportionate response; binary detection permits only late, absolute response.
7. Separate the three loci. Demand-side, supply-side and intermediation hazard have different mechanisms, different signatures and different remedies. Collapsing them into "fraud" guarantees that the wrong instrument is applied to at least two of the three.
8. Prevention before restriction. Where a measured exposure admits both a preventive and a restrictive response, the preventive response is applied first. Restriction is a residual instrument, not a first resort.
9. Proportionality is binding, not advisory. The severity of any adverse action must be bounded by the strength and auditability of the evidence supporting it. This is the framework's answer to the denial-machine risk of Section 4.5.
10. Auditability by construction. Every consequential decision generates its evidentiary record at the moment it is made, not retrospectively when a regulator asks. This anticipates the pre- and post-deployment audit framework mandated to IRDAI's AI working group.

6.2 Architecture

SUTRA organises insurtech capability into five layers. The acronym denotes Sensing, Underwriting, Triage, Restraint and Assurance, and the Sanskrit sūtra — thread, or formula — is intended to convey that the layers are connected rather than independent modules. Figure 1 sets out the structure.

A — ASSURANCE LAYER Explainability records · fairness monitoring · appeal and contestation channel · IRDAI audit trail · DPDP consent ledger
R — RESTRAINT LAYER Graduated Restraint Ladder · proportionality constraint · preventive interventions · cost-sharing adjustment · provider contracting action
T — TRIAGE LAYER Claim routing · anomaly and fraud scoring · pre-authorisation intensity · utilisation review prioritisation
U — UNDERWRITING AND MEASUREMENT LAYER Moral Hazard Exposure Index (MHEI) · DMH, SMH and IMH sub-indices · risk scoring · pricing input
S — SENSING LAYER NHCX / FHIR R4 claims · ABHA-linked records · policy administration · provider master · consented wearable and wellness data

Figure 1. The SUTRA framework. Information flows upward from Sensing through Restraint; the Assurance layer constrains every layer below it and is the point of regulatory and consumer accountability.

6.3 The Sensing layer

The Sensing layer defines what the system is permitted to observe. Four sources are specified. Structured claims data exchanged through NHCX in HL7 FHIR R4 format, providing diagnosis and procedure coding, length of stay, itemised billing and provider identity in a form comparable across payers. Policy administration data, providing sum insured, vintage, prior claims, waiting-period status and distribution channel. Provider master data, providing empanelment status, accreditation, bed strength, specialty mix and package-rate agreements. Consented behavioural data from wearables and wellness applications, where and only where the policyholder has given specific, revocable consent under the DPDP Act.

Two design constraints apply at this layer. Consent must be granular and revocable, and revocation must degrade the model's inputs rather than the policyholder's entitlements — a policyholder who withdraws wearable consent must not thereby become a higher-priced or more heavily scrutinised risk, since permitting that consequence would make consent nominal. And the layer must record data provenance for every input, because the Assurance layer cannot explain a decision whose inputs it cannot trace.

6.4 The Underwriting and Measurement layer

This layer computes the Moral Hazard Exposure Index. The construction proceeds in three steps: indicator normalisation, sub-index aggregation, and composite formation.

6.4.1 Indicator normalisation

Each raw indicator is expressed on a common zero-to-one scale by min–max normalisation against its peer distribution. For an indicator x observed for unit j, with the peer set defined by city tier, hospital tier and diagnosis-related group as appropriate,

$$z_j = (x_j - x_{min}) / (x_{max} - x_{min}) \quad (3)$$

where x_{min} and x_{max} are the 5th and 95th percentiles of the peer distribution rather than the observed extremes, so that a single outlier does not compress the scale for every other unit. Values outside the range are truncated to zero or one. Indicators for which a lower raw value indicates higher hazard — preventive-service uptake, for instance — are reverse-coded as 1 − z.

The choice of peer set is consequential and is where the equity concern of Section 4.6 enters the computation. Benchmarking a Tier 3 district hospital against Tier 1 metropolitan tertiary centres would attribute to provider behaviour what is in fact a difference in case mix and capability. The peer set must therefore be defined so that units within it are genuinely comparable, and the specification of peer sets should be documented and auditable rather than left to modelling discretion.

6.4.2 The three sub-indices

Each sub-index is a weighted mean of its normalised indicators. Table 6 gives the indicator dictionary.

Table 6. Indicator dictionary for the three Moral Hazard Exposure sub-indices

Sub-index	Indicator	Construction	Rationale
DMH — demand-side	Waiting-period proximity	Inverse of days between waiting-period expiry and first claim	Claims concentrated immediately after waiting-period expiry suggest anticipated utilisation
	Elective intensity ratio	Share of claimed procedures classified elective rather than emergent	Elective care is more price-responsive and therefore more exposed to ex-post hazard
	Room-category upgrade rate	Frequency of claiming above the policy's eligible room category	A direct and well-documented Indian manifestation of ex-post hazard
	Preventive uptake (reverse-coded)	Share of age- and sex-appropriate screenings completed	Low uptake among the insured is the observable signature of ex-ante hazard
	Short-admission frequency	Count of admissions of 24–36 hours in the observation window	Detects admissions structured to satisfy the hospitalisation trigger
SMH — supply-side	Length-of-stay deviation	Provider mean LOS minus peer median LOS for the same DRG, scaled	The clearest cross-payer signature of supplier-induced demand
	Billing intensity deviation	Provider mean billed amount minus peer median for the same DRG	Captures upcoding and intensity inflation independent of stay length
	Unbundling ratio	Share of package-rate procedures billed as separate line items	A documented Indian mechanism for exceeding agreed package rates
	Consumable and implant share	Share of claim value in consumables and implants against peer median	Concentrates margin in the least clinically transparent category

Sub-index	Indicator	Construction	Rationale
	Cross-payer concentration	Share of provider revenue from insured admissions across all payers	Only computable post-NHCX; identifies dependence on insured volume
IMH — intermediation	Adjudication time anomaly	Deviation of settlement time from complexity-predicted time	Both unusually fast and unusually slow adjudication signal process irregularity
	Channel early-claim rate	Claim rate in the first policy year by distribution channel	Isolates mis-selling and disclosure failure at the point of sale
	Disclosure completeness by channel	Share of proposals with complete medical disclosure by channel	Incomplete disclosure at sale generates disputes at claim
	Repudiation-to-grievance ratio	Grievances raised per repudiation by TPA or in-house unit	A high ratio suggests repudiations that do not withstand scrutiny

Note. LOS = length of stay; DRG = diagnosis-related group. All indicators are computable from NHCX-standard claims data combined with policy administration records. The final IMH indicator deliberately measures the insurer's own restriction quality alongside intermediary behaviour.

Formally, for sub-index S with n normalised indicators z and weights v summing to one,

$$S = \Sigma_k v_k z_k, \quad \Sigma_k v_k = 1, \quad S \in [0, 1] \qquad (4)$$

Weights within each sub-index are derived by the entropy method described in Section 5.5.6, which assigns greater weight to indicators exhibiting greater dispersion and therefore greater discriminating power, and are cross-checked against the analytic hierarchy process judgements of the expert panel.

6.4.3 The composite index

The three sub-indices combine into the composite Moral Hazard Exposure Index,

$$MHEI = w_1\, DMH + w_2\, SMH + w_3\, IMH, \quad \Sigma w = 1 \qquad (5)$$

A note on the choice of an additive rather than multiplicative form is warranted, because it is a substantive design decision rather than a convenience. A multiplicative form would imply that hazard at one locus is worthless as evidence unless accompanied by hazard at the others, which is false: a provider with severe billing anomalies is a legitimate concern irrespective of its patients' behaviour. The additive form also has the practical virtue that the contribution of each locus to the total is directly readable, which the Assurance layer requires in order to explain a decision.

The composite is nonetheless never used alone. Because the three loci demand different remedies, the framework requires that the sub-index vector (DMH, SMH, IMH) be carried alongside the scalar MHEI at every point of use. An MHEI of 0.62 arising as (0.20, 0.85, 0.10) is a provider-contracting problem; the same 0.62 arising as (0.80, 0.15, 0.25) is a product-design and policyholder-engagement problem. Acting on the scalar alone is precisely the error the framework exists to prevent.

6.5 The Triage layer

The Triage layer converts measurement into operational routing. Its function is to decide how much scrutiny a claim receives, not whether it is paid. Four routes are defined: straight-through processing for low-exposure claims within cost thresholds; standard adjudication; enhanced review with pre-authorisation scrutiny; and referral to investigation. The routing rule is a function of the MHEI vector, claim value, and the marginal value of additional review, so that scrutiny is allocated where it has the greatest expected effect rather than uniformly by claim size.

One design constraint governs this layer: routing is not a decision. A claim routed to enhanced review has not been denied, and the routing decision must not be visible to the adjudicator as a prior probability of fraud, since doing so would convert an allocation of attention into an anchor on the outcome. This separation is standard in credit adjudication and is under-observed in insurance claims practice.

6.6 The Restraint layer and the proportionality constraint

This layer is the paper's principal normative contribution. It specifies what an insurer may do in response to a measured exposure, and — more importantly — what it may not.

6.6.1 The Graduated Restraint Ladder

Responses are ordered by their severity to the affected party. The ladder is climbed from the bottom, and each rung must be exhausted before the next is available.

Table 7. The Graduated Restraint Ladder: response tiers, evidentiary requirements and locus applicability

Tier	Response	Applicable locus	Evidentiary requirement
1	Engagement: care navigation, second-opinion offer, preventive outreach, health coaching	DMH	Measured exposure above peer median. No adverse consequence, so no adjudicative standard applies.
2	Information: disclosure of package-rate comparison, network alternatives, expected out-of-pocket exposure	DMH, SMH	As above. Provides the policyholder with the information needed to constrain intensity themselves.
3	Process: enhanced pre-authorisation, utilisation review, requirement of clinical justification	SMH, DMH	Sub-index above the 75th percentile of peer distribution, sustained over at least two observation periods.
4	Contractual: package-rate renegotiation, revised co-payment at renewal, network-tier reassignment	SMH, IMH	Sub-index above the 90th percentile, sustained over three periods, with the affected party notified and given a documented right of response before the action takes effect.
5	Restrictive: claim repudiation, de-empanelment, policy non-renewal, referral for investigation	All	Specific, claim-level or provider-level documentary evidence independent of the index. The index may direct attention to a case; it may never by itself justify a Tier 5 action.

Note. The evidentiary requirement rises with tier severity. Tiers 1 and 2 carry no adverse consequence and therefore no adjudicative standard; Tier 5 requires evidence that would stand independently of the model.

6.6.2 The proportionality constraint

The ladder is enforced by a formal constraint. Let T denote the tier of a proposed action, on the scale of one to five, and let E denote an evidence-strength score on the same scale, computed from three components: the magnitude of the measured exposure relative to its peer distribution, the persistence of that exposure across observation periods, and the availability of case-specific documentary evidence independent of the model. The constraint is

$$T \leq E \quad (6)$$

stated deliberately in its simplest form, because its force lies in being non-negotiable rather than in being intricate. An insurer may always act more leniently than the evidence would permit; it may never act more severely. Every action taken must record the T proposed, the E computed, and the components from which E was derived. Where an action is taken with $T > E$, the system must block it and record the attempt — a design choice that makes over-restriction visible to internal audit and to any subsequent regulatory examination rather than invisible in aggregate statistics.

The constraint also carries an equity provision addressing Section 4.6. Where a policyholder falls within a first-time-insured cohort, or is drawn from a population whose historical utilisation was suppressed — identified by policy vintage below two years combined with Tier 3 city or rural residence — the demand-side sub-index is discounted by a factor calibrated from the propensity-score analysis of Section 5.5.5. The rationale is direct: for these cohorts the estimated proportion of excess utilisation attributable to the correction of unmet need rather than to moral hazard is materially higher, and applying an undiscounted index would systematically penalise the population that insurance exists to serve.

6.7 The Assurance layer

The Assurance layer makes the framework accountable. It performs four functions, each mapped to a specific obligation now emerging in Indian regulation.

- Explainability record. For every Tier 3 action and above, the system records the sub-index values, the indicators contributing most to each, the peer set used for normalisation, the model version, and the evidence-

strength computation. This is the artefact a pre- and post-deployment AI audit of the kind mandated to IRDAI's working group would examine.

- Fairness monitoring. Sub-index distributions and Tier 3-and-above action rates are monitored continuously across city tier, gender, age band and policy vintage. Divergence beyond a pre-registered threshold triggers review before it triggers a report, so that the finding is acted upon rather than merely disclosed.
- Contestation channel. Any party subject to a Tier 3 action or above may request the explainability record and submit a response, with a defined turnaround. Contestation outcomes feed back as labels into model retraining — which is the mechanism by which the label problem of Section 4.3 is progressively corrected, since a successfully contested action is a documented false positive of exactly the kind that conventional fraud labelling never captures.
- Consent ledger. Every behavioural data element carries its consent basis, scope and revocation status, satisfying DPDP purpose-limitation obligations and enabling the guarantee of Section 6.3 that revocation degrades model inputs without degrading entitlements.

6.8 Hypotheses

Eight hypotheses follow from the framework and constitute the empirical agenda of Stage 2.

11. H1. Perceived coverage generosity is positively associated with ex-ante hazard propensity.
12. H2. Perceived coverage generosity is positively associated with ex-post hazard propensity, and the association is stronger than for H1, reflecting the greater salience of the price effect at the point of illness.
13. H3. Provider influence salience moderates the relationship between coverage generosity and ex-post hazard, such that the relationship weakens as provider influence rises — because where the provider determines intensity, the patient's own price response is less consequential.
14. H4. Digital engagement capability is negatively associated with ex-ante hazard propensity, mediated by preventive-service uptake.
15. H5. The supply-side sub-index (SMH) explains a greater share of variance in claim cost than the demand-side sub-index (DMH), consistent with the argument of Section 2.2 that the provider locus is dominant in India.
16. H6. Policy vintage moderates the relationship between coverage and utilisation, such that the relationship is strongest in the first two policy years — the pattern predicted jointly by anticipated-utilisation selection and by the correction of unmet need, which H6 alone cannot distinguish and which the propensity-score analysis is required to separate.
17. H7. Perceived procedural fairness is positively associated with institutional trust, and institutional trust is negatively associated with ex-post hazard propensity — implying that fair claim handling reduces rather than increases moral hazard, contrary to the intuition that leniency invites abuse.
18. H8. Insurers deploying graduated restraint achieve lower claim cost growth and lower grievance incidence than insurers deploying binary restriction, tested through difference-in-differences on the panel of Section 5.2.

H7 and H8 are the hypotheses on which the framework's central claim rests. If fair process and graduated response were shown to increase rather than reduce hazard, the proportionality constraint would impose a real efficiency cost that would have to be justified on other grounds. The framework predicts the opposite, and that prediction is falsifiable.

ANTECEDENTS Perceived coverage generosity (PCG) · Digital engagement capability (DEC) · Institutional trust (ITI) · Perceived procedural fairness (PPF)
↓ moderated by: Provider influence salience (PIS), Policy vintage
HAZARD CONSTRUCTS Ex-ante hazard propensity (EAH) · Ex-post hazard propensity (EPH)
↓
COMPOSITE MEASURE MHEI = w_1 DMH + w_2 SMH + w_3 IMH
↓ bounded by proportionality constraint T ≤ E
OUTCOMES Claim cost growth · utilisation intensity · grievance incidence · renewal retention

Figure 2. Hypothesised structural model. Solid paths denote direct effects (H1, H2, H4, H5, H7); the provider influence and policy vintage constructs enter as moderators (H3, H6); H8 is tested at insurer rather than individual level.

6.9 Illustrative application

A worked illustration clarifies how the framework operates in practice. The figures below are constructed for exposition and are not empirical.

Consider a 400-bed private hospital in a Tier 2 city, empanelled with six insurers. Post-NHCX benchmarking against its peer set — private hospitals of 300 to 500 beds in Tier 2 cities — yields normalised indicators of 0.81 for length-of-stay deviation, 0.74 for billing intensity, 0.88 for unbundling ratio and 0.69 for consumable share, giving an SMH of approximately 0.78 against a peer median of 0.42. The associated policyholder cohort shows a DMH of 0.31, close to the median, and the intermediation sub-index is 0.24.

The scalar MHEI, at equal weights, is 0.44 — unremarkable. The vector (0.31, 0.78, 0.24) is not. It states unambiguously that the exposure is concentrated at the provider locus, and it directs the response accordingly. Under the ladder, an SMH at the 90th percentile of its peer distribution sustained across three observation periods supports a Tier 4 contractual response: package-rate renegotiation and network-tier reassignment, with the hospital notified and given a documented right of response before the action takes effect. It does not support de-empanelment, which is Tier 5 and requires case-specific documentary evidence independent of the index.

Critically, the framework directs no adverse action towards the policyholders treated at that hospital. Under a conventional fraud-detection approach, claims from a high-anomaly provider would attract elevated scrutiny irrespective of the individual claimant's conduct — which is how patients come to bear the consequences of their hospital's billing practices. Separating the loci prevents that transfer of consequence, which is the practical value of the decomposition.

7. Discussion

7.1 Theoretical implications

The paper makes three theoretical moves. The first is the tri-locus decomposition. The classical treatment of moral hazard, from Arrow (1963) and Pauly (1968) onward, is a two-party model in which the insured responds to a reduced marginal price. Health economics subsequently added supplier-induced demand as a separate literature (Evans, 1974; McGuire, 2000), but the two have largely remained separate bodies of work with separate empirical traditions. This paper argues that in a market with intermediated adjudication — where 72 per cent of Indian health claims by number pass through third-party administrators — a third locus exists that neither literature captures, and that any measurement framework which omits it will misattribute intermediation failure to one of the other two loci.

The second is the reframing from detection to measurement. This is a claim about the appropriate mathematical object. Detection treats moral hazard as a latent binary property of a claim, to be inferred at settlement. Measurement treats it as a continuous exposure attaching to a policy, provider and channel, evolving over time. The change is consequential because it alters what counts as a successful system. A detection system succeeds by classifying accurately; a measurement system succeeds by enabling proportionate action early. The two objectives can diverge sharply, and the industry has been optimising the first while claiming the second.

The third is the proportionality constraint as a design primitive. The fairness literature has produced formal criteria — demographic parity, equalised odds, calibration — together with impossibility results establishing that these cannot generally be jointly satisfied (Kleinberg et al., 2017). Those criteria constrain the distribution of a model's errors across groups. The proportionality constraint is a different kind of object: it constrains the severity of action permitted given evidence strength, irrespective of group membership. It is therefore not subject to the impossibility results, and it addresses a harm those criteria do not reach — a model can be perfectly calibrated across every protected group and still license actions disproportionate to what it actually knows.

7.2 Managerial implications

For insurers, the analysis carries four implications. First, the returns to further investment in claim-side detection are diminishing, while the supply-side measurement made newly possible by NHCX is largely uncontested. An insurer building provider-benchmarking capability now acquires an advantage that will erode once the capability becomes standard. Second, the sub-index decomposition should be reflected in organisational structure: supply-side hazard is a provider-network and contracting problem, not a claims-investigation problem, and organisations that route it to the special investigations unit will apply the wrong instrument. Third, the discipline of recording an evidence-strength score before each adverse action is valuable independently of the model, because it forces an explicit statement of what is known at the moment of action. Fourth, insurers should anticipate that IRDAI's audit framework will ask for documentation of process rather than accuracy of outcome, and build systems that generate it continuously.

For insurtech firms, the analysis identifies where the addressable opportunity actually lies. Distribution and claims automation are crowded and increasingly commoditised. Cross-payer provider benchmarking, chronic-disease care navigation and outcome-linked contracting infrastructure are not — and Section 4.2 explains why: their returns are diffuse and cross-insurer, which makes them poorly suited to single-insurer investment and well suited to a shared-infrastructure or industry-utility business model. The externality that makes these

capabilities undersupplied by insurers is precisely what makes them viable as a third-party product.
For hospitals, the implication is more uncomfortable but should be stated plainly. Cross-payer benchmarking makes billing patterns visible in a way they have never previously been. Providers whose length-of-stay and billing intensity sit close to peer medians have little to fear and something to gain, since undifferentiated scrutiny currently falls on them equally. Providers whose patterns are substantially above peer medians should expect that this will become known, and would be better served by examining the reasons before a payer does.

7.3 Policy implications

Four recommendations follow for IRDAI and the National Health Authority.
19. Establish authority for cross-payer provider benchmarking. The data now exists but the governance does not. A regulator-convened or industry-utility mechanism, with defined access rules, provider notification obligations and a right of response, would allow supply-side measurement to proceed without each insurer constructing its own opaque provider blacklist — which is the alternative that will otherwise emerge by default.
20. Incorporate a proportionality standard into the AI governance framework. The working group's mandate covers explainability and audit. Explainability tells a policyholder why an action was taken; it does not establish whether the action was warranted given what was known. A standard requiring insurers to record and justify the severity of adverse actions against evidence strength would close that gap, and is straightforward to audit.
21. Address the inpatient bias in product design. A significant share of what Indian claims data records as ex-post moral hazard is an artefact of products that reimburse admission and exclude outpatient management. Regulatory encouragement of outpatient and day-care coverage would reduce the incentive to convert manageable conditions into admissions, addressing the behaviour at its source rather than contesting its consequences.
22. Require disaggregated reporting of restriction outcomes. Insurers currently report repudiation and disallowance in aggregate. Reporting these disaggregated by city tier, policy vintage and channel, alongside the proportion subsequently reversed on grievance or ombudsman review, would make the over-restriction failure mode visible in the same way that the under-detection failure mode already is. Asymmetric visibility produces asymmetric incentives.

7.4 Implications for policyholders

The framework has a consumer-facing consequence worth stating directly. Under a binary detection regime, a policyholder learns of an adverse assessment at the point of denial, when the treatment has already occurred and the financial exposure is immediate. Under a graduated regime, the first contacts are engagement and information — a second-opinion offer, a comparison of network alternatives, a statement of expected out-of-pocket exposure — delivered before the decision that generates the cost. The policyholder is better served not because the insurer is more generous but because the intervention arrives when it can still change the outcome. This is the practical content of the claim that prevention dominates restriction.

8. Limitations and Future Research

8.1 Limitations

The empirical validation has not been executed. This is the principal limitation and it is fundamental. The paper develops a framework and specifies in full the protocol required to test it, but reports no primary data. The hypotheses of Section 6.8 are derived, not tested; the index weights are specified as to method, not estimated; the illustrative application of Section 6.9 is constructed for exposition. The contribution claimed is conceptual and methodological, and the framework should be regarded as a proposal awaiting evidence rather than a validated instrument.
Data availability constrains what can be validated. The supply-side sub-index requires cross-payer claims data, which is available at scale only through NHCX and only for participating providers. Since NHCX onboarding remains voluntary and is concentrated among larger and better-resourced facilities, any near-term validation will be conducted on a non-random subset of providers, and results will not generalise straightforwardly to the smaller and paper-based facilities that remain outside the exchange.
Self-report is a weak instrument for moral hazard. The EAH and EPH constructs ask respondents to report behaviour that is socially undesirable and, in the ex-ante case, may not be consciously accessible. Vignette anchoring and marker-variable techniques mitigate but do not eliminate the problem. Triangulation against observed claims behaviour is essential, and where the two diverge the observed data should be preferred.
The identification problem is mitigated, not solved. Propensity-score matching addresses selection on observables only. Unobserved health status — precisely what an individual knows about their own body and an insurer does not — is the central confound and is by construction unobservable. Rosenbaum bounds quantify the sensitivity of conclusions to this confounding; they cannot remove it. Any claim that the framework separates moral hazard from adverse selection should be read as a claim about degree, not kind.
The proportionality constraint imposes an unquantified cost. Requiring evidence proportionate to action will permit some claims to be paid that a less constrained system would have restricted. The framework asserts, in H7 and H8, that this cost is offset by improved trust, retention and grievance outcomes. That assertion is an

empirical claim that has not been tested, and if it fails, the constraint must be justified on normative grounds alone.

Institutional specificity limits transferability. The framework is calibrated to Indian conditions — NHCX, ABHA, IRDAI regulation, TPA-intermediated adjudication, inpatient-biased indemnity products. Its architecture may transfer to comparable emerging markets, but its indicators and thresholds will not without recalibration.

8.2 Directions for future research

23. Execute the Stage 2 validation specified in Section 5, beginning with the expert interviews, which would refine the indicator dictionary before the survey instrument is finalised.
24. Estimate the price elasticity of demand for health care in the Indian insured population. The RAND estimate of approximately −0.2 is drawn from a very different health system and financing structure; no comparable Indian estimate exists, and the calibration of any cost-sharing instrument depends on it.
25. Investigate the externality problem of Section 4.2 directly. If prevention is undersupplied because its benefits accrue partly to competitors, the natural experiments are multi-year policy terms and industry-level prevention pools. Neither has been studied in the Indian market.
26. Conduct a randomised evaluation of graduated versus binary restraint. An insurer willing to randomise response protocol across comparable segments could test H8 directly, which observational panel data can only approximate.
27. Examine the distributional incidence of algorithmic claim restriction across city tier, gender and policy vintage using ombudsman and grievance records, which are an underused source for exactly this question.
28. Study NHCX adoption as an organisational phenomenon. The gap between onboarding and live claim flow is where the practical value of the infrastructure is being determined, and it is currently undocumented.

9. Conclusion

Indian health insurance is growing quickly and, in significant parts of the market, losing money while it does so. The industry attributes much of the gap to behaviour — policyholders who consume more because they are covered, hospitals that bill more because someone else is paying, and intermediaries whose incentives align with neither. It has responded by investing in the detection of individual bad claims.

This paper has argued that the response is aimed at the wrong object. Detection intervenes at the last possible moment, after the cost has been incurred, and it operates under an incentive structure with no counterweight to over-restriction. The evidence that this is already producing harm — rising repudiations, rising grievances, persistent discharge delays — is visible in the regulator's own data, even if its attribution to any particular cause remains contested.

The alternative developed here treats moral hazard as a continuous, decomposable exposure rather than a binary property of claims. The SUTRA framework organises the required capability into five layers; the Moral Hazard Exposure Index measures exposure separately at the demand, supply and intermediation loci; and the Graduated Restraint Ladder, bound by the proportionality constraint $T \leq E$, specifies what an insurer may do with the resulting knowledge. That constraint is the paper's central normative claim: an insurer may act only as severely as its evidence permits, and must record the reasoning at the moment of action rather than reconstruct it when questioned.

Two developments make this a timely rather than merely aspirational proposal. NHCX has made cross-payer measurement of provider behaviour institutionally possible for the first time. And IRDAI's AI working group, constituted in June 2026, will define the governance standards against which Indian insurers' algorithmic systems are assessed. The framework proposed here is designed to be auditable against those standards as a by-product of its operation.

The framework remains, however, a proposal. Its hypotheses are derived rather than tested, its index weights specified rather than estimated. Section 5 sets out in full what would be required to establish whether it works. The argument advanced here is that the question is worth answering, because the alternative currently taking shape — accurate, fast and unconstrained claim restriction — will satisfy the loss ratio and fail the policyholder, and a market that fails its policyholders at scale does not remain a market for long.

References

1. Abdallah, A., Maarof, M. A., & Zainal, A. (2016). Fraud detection system: A survey. Journal of Network and Computer Applications, 68, 90–113.
2. Ankura. (2025). Playbook to unlocking the power of IRDAI's 2025 insurance fraud monitoring framework. Ankura Consulting Group.
3. Arrow, K. J. (1963). Uncertainty and the welfare economics of medical care. American Economic Review, 53(5), 941–973.
4. Barros, P. P., Machado, M. P., & Sanz-de-Galdeano, A. (2008). Moral hazard and the demand for health services: A matching estimator approach. Journal of Health Economics, 27(4), 1006–1025.
5. Business Standard. (2026, June 18). Irdai sets up working group to guide AI adoption in insurance sector. Business Standard.

6. Caladrius Health. (2026). The ABDM stack: How NHCX fits into India's digital health architecture. Caladrius Health AI Studio.
7. Callaway, B., & Sant'Anna, P. H. C. (2021). Difference-in-differences with multiple time periods. Journal of Econometrics, 225(2), 200–230.
8. Cappiello, A. (2020). Technology and the insurance industry: Re-configuring the competitive landscape. Palgrave Macmillan.
9. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).
10. Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153–163.
11. Colquitt, J. A. (2001). On the dimensionality of organizational justice: A construct validation of a measure. Journal of Applied Psychology, 86(3), 386–400.
12. Creswell, J. W., & Plano Clark, V. L. (2018). Designing and conducting mixed methods research (3rd ed.). SAGE.
13. Cutler, D. M., & Zeckhauser, R. J. (2000). The anatomy of health insurance. In A. J. Culyer & J. P. Newhouse (Eds.), Handbook of health economics (Vol. 1, pp. 563–643). Elsevier.
14. Denuit, M., Hainaut, D., & Trufin, J. (2019). Effective statistical learning methods for actuaries I: GLMs and extensions. Springer.
15. DreamSoft4u. (2026). NHCX integration services: FHIR-based claims for insurers and TPAs. DreamSoft4u.
16. Einav, L., & Finkelstein, A. (2018). Moral hazard in health insurance: What we know and how we know it. Journal of the European Economic Association, 16(4), 957–982.
17. Eling, M., & Lehmann, M. (2018). The impact of digitalization on the insurance value chain and the insurability of risks. Geneva Papers on Risk and Insurance – Issues and Practice, 43(3), 359–396.
18. Evans, R. G. (1974). Supplier-induced demand: Some empirical evidence and implications. In M. Perlman (Ed.), The economics of health and medical care (pp. 162–173). Macmillan.
19. Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., & Baicker, K. (2012). The Oregon health insurance experiment: Evidence from the first year. Quarterly Journal of Economics, 127(3), 1057–1106.
20. Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. Journal of Marketing Research, 18(1), 39–50.
21. Frees, E. W., Derrig, R. A., & Meyers, G. (Eds.). (2014). Predictive modeling applications in actuarial science. Cambridge University Press.
22. Garg, S., Bebarta, K. K., & Tripathi, N. (2020). Performance of India's national publicly funded health insurance scheme, Pradhan Mantri Jan Arogya Yojana (PMJAY), in improving access and financial protection for hospital care: Findings from household surveys in Chhattisgarh state. BMC Public Health, 20, 949.
23. Ghosh, S., & Mladovsky, P. (2024). Publicly funded health insurance schemes and demand for health services: Evidence from an Indian state using a matching estimator approach. Health Economics, Policy and Law, 19(4), 429–445.
24. Gioia, D. A., Corley, K. G., & Hamilton, A. L. (2013). Seeking qualitative rigor in inductive research: Notes on the Gioia methodology. Organizational Research Methods, 16(1), 15–31.
25. Guillen, M., Nielsen, J. P., & Pérez-Marín, A. M. (2021). Near-miss telematics in motor insurance. Journal of Risk and Insurance, 88(3), 569–589.
26. Hair, J. F., Risher, J. J., Sarstedt, M., & Ringle, C. M. (2019). When to use and how to report the results of PLS-SEM. European Business Review, 31(1), 2–24.
27. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems (Vol. 29, pp. 3315–3323).
28. Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. Journal of the Academy of Marketing Science, 43(1), 115–135.
29. India InsurTech Summit. (2025). National Health Claims Exchange (NHCX): Health claims reform. India InsurTech Association.
30. Insurance Regulatory and Development Authority of India. (2024). Annual report 2023–24. IRDAI.
31. Insurance Regulatory and Development Authority of India. (2025). Annual report 2024–25. IRDAI.
32. Insurance Regulatory and Development Authority of India. (2026). Constitution of working group on artificial intelligence in the insurance sector [Order dated 18 June 2026]. IRDAI.
33. Jones, D., Molitor, D., & Reif, J. (2019). What do workplace wellness programs do? Evidence from the Illinois workplace wellness study. Quarterly Journal of Economics, 134(4), 1747–1791.
34. Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In Proceedings of the 8th Innovations in Theoretical Computer Science Conference (pp. 43:1–43:23).
35. LocalCircles. (2026). Health insurance claim settlement and hospital discharge survey. LocalCircles.

36. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (Vol. 30, pp. 4765–4774).
37. Manning, W. G., Newhouse, J. P., Duan, N., Keeler, E. B., & Leibowitz, A. (1987). Health insurance and the demand for medical care: Evidence from a randomized experiment. American Economic Review, 77(3), 251–277.
38. McGuire, T. G. (2000). Physician agency. In A. J. Culyer & J. P. Newhouse (Eds.), Handbook of health economics (Vol. 1, pp. 461–536). Elsevier.
39. Ministry of Electronics and Information Technology. (2025). Report on AI governance and guidelines development. Government of India.
40. Ministry of Health and Family Welfare. (2024). National health accounts estimates for India, 2020–21 and 2021–22. Government of India.
41. Morgan, R. M., & Hunt, S. D. (1994). The commitment–trust theory of relationship marketing. Journal of Marketing, 58(3), 20–38.
42. NATHEALTH. (2025). National Health Claims Exchange. Healthcare Federation of India.
43. Newhouse, J. P. (1993). Free for all? Lessons from the RAND health insurance experiment. Harvard University Press.
44. Noll, A., Salzmann, R., & Wüthrich, M. V. (2020). Case study: French motor third-party liability claims. SSRN Electronic Journal.
45. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453.
46. Pauly, M. V. (1968). The economics of moral hazard: Comment. American Economic Review, 58(3), 531–537.
47. Prinja, S., Rajsekar, K., Gauba, V. K., & Sharma, A. (2024). Impact of health benefit package policy interventions on service utilisation under government-funded health insurance in Punjab, India. Dialogues in Health, 4, 100178.
48. Reshmi, B., Unnikrishnan, B., Rajwar, E., Parsekar, S. S., Vijayamma, R., & Venkatesh, B. T. (2021). Impact of public-funded health insurances in India on health care utilisation and financial risk protection: A systematic review. BMJ Open, 11(12), e050077.
49. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144).
50. Rothschild, M., & Stiglitz, J. (1976). Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. Quarterly Journal of Economics, 90(4), 629–649.
51. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215.
52. Şahin, Y., Bulkan, S., & Duman, E. (2013). A cost-sensitive decision tree approach for fraud detection. Expert Systems with Applications, 40(15), 5916–5923.
53. Sriram, S., & Khan, M. M. (2020). Effect of health insurance program for the poor on out-of-pocket inpatient care cost in India: Evidence from a nationally representative cross-sectional survey. BMC Health Services Research, 20, 839.
54. Stoeckli, E., Dremel, C., & Uebernickel, F. (2018). Exploring characteristics and transformational capabilities of InsurTech innovations to understand insurance value creation in a digital world. Electronic Markets, 28(3), 287–305.
55. Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. British Journal of Management, 14(3), 207–222.
56. Van Vlasselaer, V., Eliassi-Rad, T., Akoglu, L., Snoeck, M., & Baesens, B. (2017). GOTCHA! Network-based fraud detection for social security fraud. Management Science, 63(9), 3090–3110.
57. Venkatesh, V., Thong, J. Y. L., & Xu, X. (2012). Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. MIS Quarterly, 36(1), 157–178.
58. Verbelen, R., Antonio, K., & Claeskens, G. (2018). Unravelling the predictive power of telematics data in car insurance pricing. Journal of the Royal Statistical Society: Series C, 67(5), 1275–1304.
59. Wüthrich, M. V., & Merz, M. (2023). Statistical foundations of actuarial learning and its applications. Springer.
60. Zeckhauser, R. (1970). Medical insurance: A case study of the tradeoff between risk spreading and appropriate incentives. Journal of Economic Theory, 2(1), 10–26.