



Enabling Artificial Intelligence in Health: Perspectives on Health Data Challenges, Mitigation Strategies, and the Current Ecosystem in India

Vishal Deo and Kratika Pandiya

*National Institute for Research in Digital Health
Indian Council of Medical Research, New Delhi-110023, India*

Received: 18 July 2026; Revised: 8 August 2026; Accepted: 10 August 2026

Abstract

Artificial intelligence (AI) and machine learning (ML) are increasingly embedded in health systems for clinical decision support, disease surveillance and health system planning, yet their performance and fairness depend fundamentally on the quality and governance of underlying data. This paper examines health data challenges, mitigation strategies for these challenges, and explore the current health data ecosystem in India.

First, the paper synthesises key challenges in health data sharing and quality, including limited willingness to share, privacy and ownership concerns, institutional silos, irregular interoperability, and problems of accuracy, completeness, timeliness and representativeness. It also discusses the importance of statistical methods in mitigating data quality issues. It highlights how adoption of FAIR (Findable, Accessible, Interoperable, Reusable) and TRUST (Transparency, Responsibility, User focus, Sustainability, Technology) principles can underpin trustworthy repositories and data pipelines.

Second, it analyses data-related sources of bias in predictive models, distinguishing data-induced algorithmic bias from two forms of training-data bias: Type I, arising from selection and under-representation of specific subgroups, and Type II, arising from explicit use of social constructs such as race or caste.

Third, the paper maps India's health and demographic data ecosystem, including national surveys, surveillance networks, health system data, and research data repositories, and discusses opportunities and limitations for AI training and validation. Finally, it reviews national digital health frameworks, notably the Ayushman Bharat Digital Mission (ABDM), as foundational infrastructure for interoperable, consent-mediated electronic health records.

Key words: Health data infrastructure; Data repositories; Data quality; AI in health; Data sharing principles; Digital health; Data bias.

1. Introduction

Artificial intelligence (AI) and machine learning (ML) are increasingly promoted as transformative tools for clinical decision support, disease surveillance, and health system planning. Rapid growth in computing power, from teraflop to petaflop-scale hardware, and the availability of large digital datasets have made it feasible to deploy complex predictive models in routine healthcare settings (Zhu *et al.* (2023), Hassani and MacFeely (2023)).

However, the performance and fairness of these models ultimately depend on the quality, representativeness and governance of the data used to train and evaluate them. While digital health data in India are expanding, they remain fragmented, uneven in quality and often inaccessible to researchers (Salvatore *et al.* (2026), Maity *et al.* (2024), and Hassani and MacFeely (2023)). Statistics and data science provide the conceptual and methodological foundation for addressing these challenges, from designing representative data collection systems and quantifying bias, to developing robust validation frameworks and enabling responsible data sharing.

This paper provides a synthesized review of the challenges of data sharing and data quality, data induced biases in AI/ML algorithms, the role of statistical methods in mitigating these biases, the current health and demographic data ecosystem in India, and national frameworks and support system for Digital Health Data/ Electronic Health Record (EHRs).

2. Health data challenges

Health data for AI/ML in healthcare originate from multiple sources: EHRs, hospital management information system (HMIS), disease surveillance platforms, national surveys, registries and increasingly from patient-generated data streams. Despite this apparent abundance, several structural challenges constrain the effective use of these data for algorithm development and evaluation, particularly in the Indian context (Salvatore *et al.* (2026) and Maity *et al.* (2024)).

Two broad domains of concern are: (i) limitations and barriers in data sharing, and (ii) persistent problems in data quality. These domains are closely coupled, since poor quality and lack of standardisation amplify privacy, governance and trust concerns, while restricted access limits the ability of independent researchers to innovate, and also to diagnose and correct biases in deployed AI systems.

2.1. Data sharing

Lack of willingness to share data is a primary barrier, driven by concerns about privacy and confidentiality, unclear data ownership, and institutional silos (Baines *et al.* (2024) and Van Panhuis *et al.* (2014)). Many health facilities and programmes treat their data as proprietary assets, which, combined with weak data governance frameworks, leads to highly restricted access even for bona fide research purposes.

Interoperability challenges further inhibit data sharing. When health information systems use incompatible formats, coding standards and identifiers, aggregating data across platforms becomes technically difficult and costly (Van Panhuis *et al.* (2014)). Security concerns, limited infrastructure and absence of standardised data sharing pipelines – for

example, secure APIs, de-identification services and governance contracts – mean that even when data exist in digital form, they seldom move seamlessly between providers, programmes and researchers.

Wilkinson *et al.* (2016) introduced the Findable, Accessible, Interoperable, Reusable (FAIR) principles for data sharing. The aim of FAIR principles is to provide practical guidance for anyone seeking to make their data more reusable. The FAIR principles explicitly prioritise making data automatically discoverable and usable by machines, while still supporting reuse by individual users (Martínez-García *et al.* (2023)). Lin *et al.* (2020) underscores the importance of adopting Transparency, Responsibility, User focus, Sustainability, Technology (TRUST) principles for health data repositories. Applying FAIR/TRUST in practice implies building trustworthy digital repositories with clear organisational mandates, sustainable funding, robust technical infrastructure, external certification (such as CoreTrustSeal or ISO 16363), and transparent policies for access, consent and reuse.

2.2. Data quality

Data quality issues manifest along several dimensions, including accuracy, completeness, consistency, timeliness, representativeness, validity and source credibility (Schmidt *et al.* (2021)). In many Indian programmes, routine data are collected primarily for service reporting or administrative purposes, rather than for analytic robustness, which can lead to under-reporting, misclassification and variable adherence to definitions across states and facilities (Blanchard *et al.* (2020)).

Representativeness is particularly critical for AI applications. If datasets over-represent certain regions, facility types or socio-demographic groups, the resulting models may perform poorly, or even harmfully, for under-represented populations (Hassani and MacFeely (2023)). Timeliness and interoperability also affect the feasibility of using health data for real-time surveillance or adaptive decision support, since lagged or siloed datasets lead to outdated or incomplete signals.

Data labelling and annotation quality, and granularity are other important aspects of data quality, especially for imaging and sensor data. Without standardised annotation protocols, multi-reader verification and detailed metadata on devices, acquisition conditions and grading procedures, AI models trained on such datasets can exhibit unpredictable performance when deployed in different settings. Squires *et al.* (2023) quantifies how variability in radiologist breast-density labels affects deep-learning performance, demonstrating that inconsistent grading leads to unpredictable model behaviour and reduced reliability when deployed. Multusch *et al.* (2025) demonstrated that annotation quality varies with reader expertise and that noisy labels from less experienced annotators degrade AI performance, underscoring the need for structured multi reader verification and quality control in radiology datasets.

3. Data induced bias in AI algorithms

Bias in medical AI can be defined as systematic error or undesired dependence of an algorithm on specific attributes, often correlates of demographic or socio-economic groups, that leads to unequal performance or outcomes. Recent reviews have catalogued multiple

forms of bias in AI-based models for medical applications, highlighting implications for clinical decision-making, triage and resource allocation (Cross *et al.* (2024) and Mittermaier *et al.* (2023)).

Data-induced algorithmic bias occurs when patterns present in training data, including measurement errors, missingness, confounding and social inequities, are inadvertently learned and amplified by the model. Bias is not only about explicit encoding of protected attributes, but about any undesired dependence on data features that reflect inequitable structures in healthcare access, financing or outcomes. For example, if historical data record lower utilisation of specialist services among marginalised communities due to barriers in access, an AI system trained to predict future utilisation may learn that these groups “need” fewer referrals, thereby perpetuating under-treatment. Similarly, if cost data reflect differential spending patterns rather than clinical need, models that use cost as a proxy outcome can embed socio-economic bias in risk scores and care prioritisation.

For the sake of discussion, we categorize data-related sources of bias into two broad categories: (i) training data bias Type I, arising from selection and representativeness; and (ii) training data bias Type II, arising from the use of social constructs such as race, ethnicity or caste as explicit variables in algorithms.

3.1. Training data bias Type-I: selection bias/ imbalanced sample

Here our discussion corresponds to selection bias and imbalanced sample sizes, *i.e.*, over or under-representation of particular subpopulations in the data used to train AI models. This leads to worse performance and systematic underestimation of risk for under-represented groups.

A widely discussed example is the algorithm used by US hospitals to prioritise patients for extra care, analysed by Obermeyer *et al.* (2019). The algorithm did not include race as an explicit variable, but used healthcare cost history as a proxy for healthcare needs. Because Black patients historically incurred lower healthcare costs despite similar or greater need, the algorithm assigned them lower risk scores than white patients, resulting in fewer referrals for additional care at equivalent levels of morbidity.

A validation and implementation study for real world evaluation of AI-driven diabetic retinopathy (DR) screening in India done by Duggal *et al.* (2025) reported that AI algorithms validated on international datasets showed substantially degraded performance when deployed in Indian public screening sites. Sensitivity dropped to 60–80% and specificity to as low as 14% on Indian images, largely due to the absence of standardised, high-quality and regionally representative training data and poor documentation of device types, image quality and grading procedures.

Importantly, AI developers were often unable or unwilling to specify the provenance, characteristics or size of their training datasets, and metadata on camera type, image quality, population demographics and labelling protocols were rarely accessible (Duggal *et al.* (2025)). Such opacity prevents meaningful assessment of contextual validity and bias, and underscores the need for transparent data curation standards and regulatory requirements for documentation.

3.2. Training data bias Type-II: unintended bias due to use of social constructs as factors

These biases arise when social constructs such as race, ethnicity or caste are directly encoded in clinical algorithms, often without robust evidence on their biological or public health relevance. These constructs are not biological categories and may introduce unintended bias if used naively in risk prediction or diagnostic models.

A key example is the use of race in estimating kidney function, specifically estimated glomerular filtration rate (eGFR). Mohottige *et al.* (2023) reported limited justification for using "Black race" as an adjustment factor in eGFR equations. They showed how including race led to overestimation of kidney function in Black patients, delaying referral for nephrology care and transplantation, thereby exacerbating health inequities.

In such situations, mitigation strategies include replacing race-based adjustments with variables that more directly capture socio-economic deprivation, such as area-level indicators derived from zip codes or location information (Cross *et al.* (2024)). More broadly, the message is that algorithm designers should critically interrogate any variable that reflects historical discrimination or structural inequity, and preferentially model underlying mechanisms (for example, access barriers, environmental exposures, co-morbidities) rather than coarse social categories.

Other training-data-based biases include non-random missingness (for instance, laboratory data available only for those who can afford tests), lack of capture of important clinical variables in routine datasets, and misclassification or noisy labels in electronic records and registries.

3.3. Role of statistical methods in mitigating these biases

From a statistical perspective, handling such biases requires explicit modelling of missing data mechanisms, careful design of validation studies, and robust sensitivity analyses that assess the impact of plausible bias scenarios on model performance and fairness (Hauschildt *et al.* (2025), Cross *et al.* (2024) and Mittermaier *et al.* (2023)). Selection bias and imbalanced samples can be corrected using inverse-probability weighting, propensity-score matching, and resampling methods such as Synthetic Minority Oversampling Technique (SMOTE), Adaptive Synthetic Sampling (ADASYN), and stratified under/oversampling (Krautenbacher *et al.* (2017) and Chawla *et al.* (2002)). Under- or over-representation is addressed through data rebalancing, reweighting, and fairness-aware pre-, and post-processing pipelines, covering fairness constraints, adversarial debiasing, and threshold calibration (Chen *et al.* (2024)). Non-random missingness requires multiple imputation, inverse-probability weighting, or selection/pattern-mixture models instead of complete-case analysis, since ignoring the missingness mechanism biases estimates (Little and Rubin (2019)). Misclassification is handled through inverse-propensity reweighting of mislabeled points and regression-calibration-type corrections, improving classifier reliability when trained on biased or noisy data (Song *et al.* (2025)).

4. Current health and demographic data ecosystem in India

India has developed a rich, though fragmented, health and demographic data ecosystem comprising large-scale household surveys, disease surveillance platforms, civil registration systems and facility-based HMIS. In this article, we have reviewed key components of this ecosystem and their relevance for AI-enabled health applications, while also noting limitations in coverage, granularity and accessibility.

From an AI/ML standpoint, these data sources have the potential to provide essential ground truth and contextual information for training, validating and monitoring predictive models related to maternal and child health, infectious diseases, non-communicable diseases and health service utilisation. Yet, their differential periodicity, representativeness and openness to researchers pose challenges for integrating them into unified data strategies.

4.1. Survey datasets

The National Family Health Survey (NFHS) is one of the largest demographic and health surveys globally, conducted by the Ministry of Health and Family Welfare (MoHFW) with International Institute for Population Sciences (IIPS), Mumbai, as the nodal agency (IIPS (2026)). Six rounds have been completed between 1992-93 and 2023-24. Since NFHS-4, the survey provides district-level estimates on fertility, mortality, maternal and child health, family planning, nutrition, and selected non-communicable disease indicators.

Other survey datasets include the District Level Household and Facility Survey (DLHS) and Annual Health Survey (AHS), which historically provided District Level reproductive and child health and mortality indicators in selected states, as well as health-related rounds of the National Sample Survey (NSS) on morbidity, health expenditure, disability, and aging. Although DLHS and AHS have been discontinued, their functions have largely been absorbed into NFHS and HMIS while NSS continues to provide critical information on out-of-pocket expenditure, catastrophic spending and disability that can inform equity-focused AI models Dandona *et al.* (2016)).

Limitations of such survey datasets include high cost and resource intensity, recall and self-reporting biases, limited coverage of adult health and mental health, and challenges in comparability over time due to changes in questionnaires and definitions. Moreover, their periodic nature means that rapidly evolving health issues, such as emerging infectious diseases or changing risk factors, are not captured in real time, which restricts their direct use for dynamic model updating.

4.2. Disease surveillance datasets

Disease surveillance platforms provide near real-time information on epidemic-prone conditions and outbreaks. The Integrated Disease Surveillance Programme (IDSP), launched in the early 2000s and now part of the National Health Mission, aims to strengthen decentralised, laboratory-based IT-enabled surveillance for epidemic-prone diseases and to support early detection and response through trained rapid response teams (Ministry of Health and Family Welfare (2026)).

Building on IDSP, the Integrated Health Information Platform (IHIP) is a web-based extension launched in 2019 that enables near real-time reporting, analysis and geo-tagged facility data for over 50 conditions (<https://ihiplearning.in/idsp/>). In parallel, the Virus Research and Diagnostic Laboratory (VRDL) network coordinated by the Department of Health Research (DHR) and the Indian Council of Medical Research (ICMR) has established more than 160 laboratories that generate test data on viral diseases, shared in near real time with IDSP and other programmes (Department of Health Research (2026)).

From an AI perspective, these platforms offer high-frequency signals for outbreak detection, syndromic surveillance and dynamic risk prediction, but there are some limitations as well. Facility-based surveillance may not be representative of community-level disease burden; fragmentation across systems leads to duplication and inconsistency; private sector facilities, which deliver a substantial share of care in India, often do not report systematically; and data are rarely open, timely or user-friendly for external researchers.

These limitations constrain the ability to train robust forecasting models or evaluate algorithm performance across heterogeneous settings. Addressing them will require stronger mandates and incentives for reporting, integration of private sector data, improved data quality assurance, and governance frameworks that support controlled access for research and innovation.

4.3. Hospital management information system

HMIS of the MoHFW is a nationwide facility-based reporting platform covering sub-centres, primary health centres, community health centres, district hospitals and higher facilities (<https://www.hmis.mohfw.gov.in>). Launched in 2008, HMIS provides monthly aggregated indicators on service utilisation, maternal and child health, immunisation, family planning, non-communicable diseases and programme performance.

HMIS, together with NFHS, now serves as a primary source of district-level health and demographic data in India. Yet, HMIS faces challenges of limited private sector participation, variable timeliness and completeness across states, frequent staff turnover among data entry personnel, and inconsistent validation and feedback mechanisms (Gupta *et al.* (2024)).

For AI/ML applications, HMIS offers a potential backbone for learning models of service utilisation, coverage gaps and programme performance at fine geographic granularity. However, realising these potential demands investments in data quality, standardisation of coding and aggregation rules, and creation of anonymised research-ready data extracts. Linking HMIS with EHRs and registries could eventually support more granular, patient-level analytics within robust governance and consent frameworks.

4.4. National repositories of health and biomedical research datasets

As a new initiative, ICMR has established the National Health Research Data Repository (<https://data.icmr.org.in/>) with the vision to empower researchers with secure, centralized access to high-quality datasets, advancing medical research innovation while safeguarding data integrity and privacy through rigorous verification standards. High-quality datasets, along with comprehensive metadata, data snapshots, and data descriptions are available on the portal which can be seen to anyone through free portal registration. Cur-

rently, researchers are provided access to complete datasets for analysis through the federated principle, whereby analysis code is sent to the data rather than the data being shared with researchers.

Another important initiative in this direction is the Indian Biological Data Centre (IBDC) established by the Department of Biotechnology (<https://ibdc.dbt.gov.in>). It is the first national repository for life science data in India. IBDC is mandated to archive all life science data generated from publicly funded research in India. As a national life-sciences data archive IBDC houses sub-repositories like the Indian Nucleotide Data Archive, Indian Proteome Databank, Indian Structural Data Archive, Indian Biomedical Image Archive, and Indian Phenome Repository.

As a more generic repository, AIKosh (<https://aikosh.indiaai.gov.in/home>) is the national AI datasets and resources platform launched under the IndiaAI Mission of the Ministry of Electronics and Information Technology (MeitY). It is designed to democratize access to high-quality, India-specific AI artefacts to researchers and innovators. It serves as a unified hub offering over 14,000 datasets, over 300 AI models, and more than 30 toolkits spanning sectors such as health, agriculture, education, and governance, along with benchmarks, use cases, and a secure sandbox environment for model training and experimentation. Organizations, including academic institutions, startups, and non-profits, can contribute non-personal, anonymized datasets and models via an Expression of Interest process, retaining control over licensing while ensuring compliance with ethical AI and data-governance standards.

5. Overview of available frameworks and support system for digital health data/ EHRs

Recognising the need for interoperable, citizen-centric digital health infrastructure, India has launched several national initiatives that define standards, registries and governance mechanisms for health data and EHRs. Two key components of this ecosystem are the Ayushman Bharat Digital Mission (ABDM) and the National Resource Centre for EHR Standards (NRCeS).

5.1. Ayushman Bharat Digital Mission (ABDM)

ABDM aims to create a national digital health ecosystem that supports universal health coverage in an efficient, accessible, inclusive, affordable, timely and safe manner. It is designed as the backbone for integrated digital health infrastructure, leveraging India's broader public digital rails, including Aadhaar, Unified Payments Interface and widespread mobile connectivity, while adhering to the guiding principles of the National Digital Health Blueprint (<https://abdm.gov.in>).

Key architectural features of the ABDM include decentralised data storage, wherein health data and applications are held at points of care; patient control over processing of their health data; and consent-driven, link-based data access, with entities able to access records only through secure links subject to permissions and consent (National Health Authority (2022)). The Ayushman Bharat Health Account (ABHA) number serves as a unique digital identifier that threads longitudinal health records across facilities, with the individual

retaining control over sharing and linkage.

ABDM also maintains two major registries: the Health Facility Registry (HFR), a comprehensive repository of public and private health facilities across systems of medicine, and the Healthcare Professionals Registry (HPR), which records registered and verified healthcare practitioners. As of July 2026, government reports indicate over 94 crores of ABHA accounts created, over 106 crores health records linked, and more than five lakh facilities and 10 lakh professionals onboarded, underscoring the rapid expansion of India's digital health infrastructure (<https://dashboard.abdm.gov.in/abdm/>).

From a statistical and AI standpoint, ABDM introduces both opportunities and responsibilities. Robust, consent-mediated linkage of patient-level records across providers can enable more accurate risk prediction, longitudinal analyses and personalised decision support. However, realising these gains requires strict adherence to privacy-preserving data management policies as per India's Digital Personal Data Protection (DPDP) Act, 2023, clear rules against re-identification, and mechanisms for independent algorithmic auditing anchored in transparent data documentation.

5.2. National Resource Centre for EHR Standards (NRCeS)

To accelerate adoption of interoperable EHR standards, the MoHFW released national EHR standards in 2013 and updated them in 2016, and established the NRCeS at Centre for Development of Advanced Computing (C-DAC), Pune.

NRCeS functions as a centre of excellence and single point of contact for assistance in developing, implementing and using EHR standards in India, providing a knowledge base and tools that facilitate adherence to these standards (<https://www.nrceS.in/>).

The standards cover data formats, coding systems (such as SNOMED CT and ICD), interoperability specifications (such as FHIR/HL7) and security requirements necessary for exchange of health information across systems. For AI developers, compliance with these standards is critical. It ensures that models trained on EHR data are portable across platforms, that variables are consistently defined and coded, and that model outputs can be integrated into clinical workflows without compromising data integrity or privacy. In combination, ABDM and NRCeS provide the scaffolding for a more coherent digital health data ecosystem in India.

6. Conclusion

A review of the current data ecosystem of India shows that India possesses a diverse array of health and demographic data sources – from NFHS, NSS and legacy district surveys to IDSP/IHIP, VRDL, HMIS, ABDM registries and emerging EHR implementations. Yet these assets remain fragmented across institutions and programmes, with heterogeneous standards, variable quality and limited interoperability. As a result, researchers and innovators often struggle to obtain timely, well-documented data needed to develop, validate and monitor AI/ML models that are both accurate and equitable.

Bridging this gap requires concerted efforts on multiple fronts. First, policymakers and programme managers must invest in harmonising data standards, strengthening qual-

ity assurance, and enabling secure, consent-mediated data sharing pipelines across surveys, surveillance systems and facility-based platforms. Second, health statisticians and data scientists should embed rigorous bias assessment, missing data modelling and fairness metrics into the lifecycle of AI system development, drawing on emerging literature on bias in medical AI and leveraging India-specific validation studies such as the diabetic retinopathy screening work highlighted in the article. Third, researchers themselves need to change behaviour and culture around data sharing. While ethical and legal safeguards are indispensable, reflexive non-disclosure and siloed data ownership impede collective learning and stifle innovation. Embracing FAIR and TRUST principles, contributing anonymised datasets and code to well-governed repositories, and collaborating across disciplines can help create a virtuous cycle where better data enable better models, which in turn inform better policies and clinical decisions.

Ultimately, enabling AI in health is not merely a technical endeavour. It is a statistical and governance challenge that hinges on how societies collect, curate, interpret and share health data. For India, the next decade offers an opportunity to transform its rich but fragmented data landscape into an integrated, interoperable and researcher-friendly ecosystem that supports responsible AI innovation and advances health equity.

Acknowledgements

The authors are grateful to the reviewers for their valuable comments and suggestions, and the editors for their guidance and counsel to greatly improve the paper.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

- Baines, R., Stevens, S., Austin, D., Anil, K., Bradwell, H., Cooper, L., Maramba, I., Chatterjee, A., and Leigh, S. (2024). Patient and public willingness to share personal health data for third-party or secondary uses: systematic review. *Journal of Medical Internet Research*, **26**, e50421.
- Blanchard, J., Washington, R., Becker, M., Vasanthakumar, N., Gopal, K., and Sarwal, R. (2020). Vision 2035—public health surveillance in India. A White Paper.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, **16**, 321–357.
- Chen, F., Wang, L., Hong, J., Jiang, J., and Zhou, L. (2024). Unmasking bias in artificial intelligence: a systematic review of bias detection and mitigation strategies in electronic health record-based models. *Journal of the American Medical Informatics Association*, **31**, 1172–1183.
- Cross, J. L., Choma, M. A., and Onofrey, J. A. (2024). Bias in medical AI: Implications for clinical decision-making. *PLOS Digital Health*, **3**, e0000651.

- Dandona, R., Pandey, A., and Dandona, L. (2016). A review of national health surveys in India. *Bulletin of the World Health Organization*, **94**, 286.
- Duggal, M., Chauhan, A., Gupta, V., Kankaria, A., Budhija, D., Verma, P and Miglani, V., Syal, P and Kaur, G., Kumar, L., Mutyala, N., Bezbaruah, R., Sood, N., Kernohan, A., Menon, G., and Vale, L. (2025). Real-world evaluation of AI-driven diabetic retinopathy screening in public health settings: Validation and implementation study. *JMIR Medical Informatics*, **13**, e67529.
- Gupta, A., Sharma, A., and Jha, D. K. (2024). Overcoming obstacles step by step: A comprehensive review of challenges and strategies in implementing hospital management information systems in India.
- Hassani, H. and MacFeely, S. (2023). Driving excellence in official statistics: Unleashing the potential of comprehensive digital data governance. *Big Data and Cognitive Computing*, **7**, 134.
- Hauschildt, K. E., Pan, A., Bernstein, T., Admon, A. J., Mukherjee, B., Iwashyna, T. J., and Rountree, L. (2025). Consideration of sociodemographics in machine learning-driven sepsis risk prediction. *Critical Care Medicine*, **53**, e1815–e1820.
- IIPS (2026). National family health survey user portal. retrieved July 17, 2026, from <https://www.nfhsiips.in/nfhsuser/index.php>.
- Krautenbacher, N., Theis, F. J., and Fuchs, C. (2017). Correcting classifiers for sample selection bias in two-phase case-control studies. *Computational and Mathematical Methods in Medicine*, **2017**, 7847531.
- Lin, D., Crabtree, J., Dillo, I., Downs, R. R., Edmunds, R., Giaretta, D., ..., and Westbrook, J. (2020). The trust principles for digital repositories. *Scientific Data*, **7**, 144.
- Little, R. J. and Rubin, D. B. (2019). *Statistical Analysis with Missing Data*. John Wiley & Sons.
- Maity, D., Satish, R., Jadeja, D. A., Dharmaraju, R., Chandru, V., Sundaresan, R., Singh, H., and Pal, D. (2024). MIDAS: a new platform for quality-graded health data for AI-enabled healthcare in India. *Nature Medicine*, **30**, 2704–2705.
- Martínez-García, A., Alvarez-Romero, C., Román-Villarán, E., Bernabeu-Wittel, M., and Parra-Calderón, C. L. (2023). FAIR principles to improve the impact on health research management outcomes. *Heliyon*, **9**.
- Mittermaier, M., Raza, M. M., and Kvedar, J. C. (2023). Bias in AI-based models for medical applications: Challenges and mitigation strategies. *NPJ Digital Medicine*, **6**, 113.
- Mohottige, D., Olabisi, O., and Boulware, L. E. (2023). Use of race in kidney function estimation: Lessons learned and the path toward health justice. *Annual Review of Medicine*, **74**, 385–400.
- Multusch, M. M., Hansen, L., Heinrich, M. P., Berkel, L., Saalbach, A., Schulz, H., Wegner, F., Barkhausen, J., and Sieren, M. M. (2025). Impact of radiologist experience on AI annotation quality in chest radiographs: a comparative analysis. *Diagnostics*, **15**, 777.
- Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, **366**, 447–453.
- Salvatore, M., Wang, Y., Syal, P., Wahl, B., Parikh, H., Deo, V., ..., and Mukherjee, B. (2026). The case for an integrated biobanking initiative in South Asia. *The Lancet Regional Health-Southeast Asia*, **49**.

- Schmidt, C. O., Struckmann, S., Enzenbach, C., Reineke, A., Stausberg, J., Damerow, S., Huebner, M., Schmidt, B., Sauerbrei, W., and Richter, A. (2021). Facilitating harmonized data quality assessments. a data quality framework for observational health research data collections with software implementations in R. *BMC Medical Research Methodology*, **21**, 63.
- Song, B., Zhao, S., Dang, L., Wang, H., and Xu, L. (2025). A survey on learning from data with label noise via deep neural networks. *Systems Science & Control Engineering*, **13**, 2488120.
- Squires, S., Harkness, E. F., Evans, D. G., and Astley, S. M. (2023). The effect of variable labels on deep learning models trained to predict breast density. *Biomedical Physics & Engineering Express*, **9**, 035030.
- Department of Health Research (2026). Setting up of nation-wide network of laboratories for managing epidemics and natural calamities (viral research and diagnostic labs). ministry of health and family welfare, government of India. retrieved July 18, 2026, from <https://www.dhr.gov.in/offering?page=1>.
- Ministry of Health and Family Welfare (2026). History of the integrated disease surveillance programme (idsp). retrieved July 18, 2026, from <https://ihiplearning.in/idsp/#/about/history>.
- National Health Authority (2022). A brief guide on ayushman bharat digital mission and its various building blocks. accessed july 2026 from https://abdm.gov.in/strapicms/uploads/ABDM_Handbook_19_10_2022_24c5078481.pdf.
- Van Panhuis, W. G., Paul, P., Emerson, C., Grefenstette, J., Wilder, R., Herbst, A. J., Heymann, D., and Burke, D. S. (2014). A systematic review of barriers to data sharing in public health. *BMC Public Health*, **14**, 1144.
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ..., and Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, **3**, 1–9.
- Zhu, S., Yu, T., Xu, T., Chen, H., Dustdar, S., Gigan, S., ..., and Pan, Y. (2023). Intelligent computing: the latest advances, challenges, and future. *Intelligent Computing*, **2**, 0006.