Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Algorithmic fairness and transportability of machine learning models predicting cancer screening uptake in India: a nationally representative study

G. Hari Prakash
Ramaiah School of Public Health, MS Ramaiah University of Applied Sciences
Sunil Kumar D
sunilkumard@jssuni.edu.in
JSS Medical College, JSS Academy of Higher Education and Research
B D Tejaswini
Ramaiah School of Public Health, MS Ramaiah University of Applied Sciences
Vanishri Arun
B.M.S. College of Engineering

Research Article

Keywords: Machine learning, Algorithmic fairness, Transportability, External validation, Cancer screening, Health equity, Data leakage, India

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11335681/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Algorithmic fairness and transportability of machine learning models predicting cancer screening uptake in India: a nationally representative study

G. Hari Prakash[1], D. Sunil Kumar[2*], Tejaswini B D[1], Vanishri Arun[3]

1 Department of Public Health, Ramaiah School of Public Health, MS Ramaiah University of Applied Sciences, Bengaluru, Karnataka, India

2 Department of Community Medicine, JSS Medical College, JSS Academy of Higher Education and Research, Mysuru, Karnataka, India

3 Department of Computer Science and Engineering, B.M.S. College of Engineering, Bengaluru, Karnataka, India

* Corresponding author: Dr D. Sunil Kumar, Professor, Department of Community Medicine, JSS Medical College, JSS Academy of Higher Education and Research, Mysuru, Karnataka, India. Email: sunilkumard@jssuni.edu.in

Email addresses: GHP hariprakash607@gmail.com; DSK sunilkumard@jssuni.edu.in; TBD tejaswini.ph.rsph@msruas.ac.in; VA vanishriarun.cse@bmsce.ac.in

Abstract

Background: Machine learning models built from national surveys are increasingly proposed to identify people who are unlikely to be screened for cancer, so that outreach can be targeted. Whether such models capture individual behaviour or merely reproduce where programmes operate, whether they transfer to new settings, and whether they treat population groups equitably is rarely examined, particularly in low- and middle-income countries.

Methods: We used India's fifth National Family Health Survey (2019-21) to predict ever having had cervical, breast and oral cancer screening among 357353 women and oral screening among 45739 men aged 30-49 years. Logistic regression and XGBoost models were trained with

individual predictors, area predictors (state and out-of-fold district screening rate) or both. We assessed discrimination, calibration, survey-weighted performance, SHAP explanations and decision curves; transportability by leave-one-state-out validation with random-effects pooling; fairness across sex, wealth, caste, residence and religion; two bias-mitigation strategies; and the effect of resampling before the train-test split.

Results: Individual predictors alone achieved AUCs of 0.58-0.69, whereas area predictors alone achieved 0.63-0.82 and combined models 0.67-0.82, with good calibration. Area features accounted for 59-69% of SHAP importance in women. When a state was held out, individual-level models performed at chance (pooled AUC 0.51-0.54) and observed screening ranged from 0.17 to 15 times the predicted level. When the top 10% of predicted risk was flagged, sensitivity among the poorest women was 0.18-0.38 compared with 0.57-0.68 among the richest two quintiles, giving equal opportunity differences of 0.30-0.45 across wealth and 0.31-0.46 across caste or tribe groups. Group-specific thresholds reduced the wealth disparity to 0.09-0.10 without loss of AUC, and reweighing reduced it to 0.08 for cervical screening. Applying synthetic minority oversampling before the train-test split inflated the AUC from 0.64 to 0.97.

Conclusions: Survey-based screening-uptake models predict where screening programmes operate rather than who takes up screening, do not transfer across states, and allocate outreach unequally unless fairness is explicitly addressed. Such models should not be used to target individuals, and published high accuracies should be interpreted with caution where resampling precedes data splitting.

Keywords: Machine learning; Algorithmic fairness; Transportability; External validation; Cancer screening; Health equity; Data leakage; India

Background

Cancer screening programmes in low- and middle-income countries face a common problem: coverage is low and unequal. In India, which launched population-based screening for oral, breast and cervical cancers for adults aged 30 years and above in 2016 [1], fewer than 2% of women aged 30-49 years reported ever having been screened for cervical cancer and about 1% for breast or oral cancer in the fifth National Family Health Survey (NFHS-5, 2019-21) [2, 3]. Coverage varies widely between states and socioeconomic groups [3, 4].

Machine learning has been proposed as a way to identify individuals who are unlikely to be screened so that outreach can be prioritised. Recent studies have applied random forests, gradient boosting and SHapley Additive exPlanations (SHAP) to national survey data from India and sub-Saharan Africa to predict screening uptake [5-7], with some reporting near-perfect discrimination [6]. These studies raise three questions that are central to health equity but rarely examined.

First, screening uptake in a programme context depends heavily on where programmes are delivered. A model that predicts uptake from place of residence may achieve high accuracy while saying little about individual behaviour, and such a model cannot be used to find unscreened people within a given area. Second, prediction models frequently fail when applied to new settings [8-10]; internal-external validation across geographic units is the recommended way to assess this [11, 12]. Third, models trained on data that reflect existing inequities can reproduce or amplify them [13-15]. Fairness metrics such as equal opportunity [16] and mitigation techniques such as reweighing [17] and group-specific thresholds have been applied mainly in high-income clinical settings, and their use in population health prediction in low- and middle-income countries is limited [18].

A further methodological concern is class imbalance. Screening is rare, and synthetic minority oversampling (SMOTE) [19] is widely used to balance classes. Rebalancing distorts calibration

[20], and if it is applied before data are split into training and test sets, synthetic records derived from test-set observations leak into training and inflate performance [21].

We therefore aimed to (i) quantify how much of the predictive performance of screening-uptake models in India comes from individual characteristics compared with geography, (ii) assess transportability of individual-level models across states, (iii) audit fairness across sex, wealth, caste or tribe, residence and religion and test two mitigation strategies, and (iv) demonstrate the effect of resampling before data splitting on reported performance.

Methods

Data source and study population

We used NFHS-5 (2019-21), the Indian Demographic and Health Survey, conducted by the International Institute for Population Sciences with technical assistance from the DHS Program [2, 22]. NFHS-5 used a two-stage stratified cluster design covering all 36 states and union territories and 707 districts. Women aged 15-49 years were interviewed in all sampled households and men aged 15-54 years in a subsample of households. We included women (n = 357353) and men (n = 45739) aged 30-49 years, the age group targeted by the screening programme and asked the screening questions, with non-missing outcome data. The analytical framework is summarised in Fig. 1.

Outcomes

Outcomes were self-reported ever having undergone a screening test for cervical cancer, a breast examination for breast cancer and an oral cavity examination for oral cancer. Five prediction tasks were defined: cervical, breast and oral screening in women, oral screening in men, and oral screening in both sexes pooled with sex as a predictor.

Predictors and feature sets

Individual predictors were age, education, household wealth quintile, place of residence, religion, caste or tribe, health insurance, weekly media exposure, current marital status, any tobacco use, smoked tobacco, smokeless tobacco and alcohol use (Additional file 2, Table S1). Area predictors were state indicators and the district screening rate, estimated by target encoding [23] with shrinkage towards the national mean. To prevent leakage, district rates for training records were computed out-of-fold (five folds) and those for test records from the training data only. Three feature sets were compared: individual, area and individual plus area. Missing values in binary predictors (under 1%) were imputed with the training median, and unknown caste was retained as a separate category.

Model development and internal validation

Data were split into training (70%) and test (30%) sets stratified by outcome. For each task and feature set we fitted L2-penalised logistic regression and gradient boosted trees (XGBoost) [24] with prespecified hyperparameters (Additional file 2, Table S2). Classes were not rebalanced, to preserve calibration [20]. In the test set we estimated the area under the receiver operating characteristic curve (AUC) with 200 bootstrap confidence intervals, survey-weighted AUC, area under the precision-recall curve, Brier score, calibration intercept and slope [25] and the observed to expected (O/E) ratio. Predictor contributions were summarised with SHAP values [26], grouped into area and individual predictors. Within-state AUCs were calculated to assess how well models ranked individuals inside the same state. Clinical utility for outreach targeting was assessed with decision curve analysis [27].

Transportability

Internal-external validation [11] was performed by leave-one-state-out: individual-level models were trained on all other states and evaluated in the held-out state, which had to contain at least ten screened respondents. State-specific AUCs (logit-transformed, standard errors from the Hanley-McNeil method [28]) and log O/E ratios were pooled with DerSimonian-Laird random-effects meta-analysis [29], with 95% prediction intervals and I squared [12].

Fairness audit and bias mitigation

For the best combined models we flagged the 10% of individuals with the highest predicted probability, a plausible outreach capacity, and calculated by subgroup the sensitivity (true positive rate), false positive rate, proportion flagged, AUC and O/E ratio. Subgroups were defined by sex, wealth quintile, caste or tribe, residence and religion. Disparities were summarised as the equal opportunity difference, the range of sensitivity across groups [16], with 100 bootstrap 95% confidence intervals, and the ratio of the lowest to highest proportion flagged. Two mitigation strategies were tested for cervical screening (by wealth) and pooled oral screening (by wealth and by sex): pre-processing by reweighing [17], in which training records were weighted so that group membership and outcome were independent, and post-processing with group-specific thresholds chosen on out-of-fold training predictions to equalise sensitivity. Sex-specific models were also compared for oral screening.

Resampling before data splitting

To show how resampling can inflate reported performance, cervical screening models with individual predictors were evaluated under five pipelines: a majority-class rule; no resampling; SMOTE [19] applied to the training set only; SMOTE applied to the whole dataset before an 80:20 split; and random oversampling before the split.

Software and reporting

Analyses were performed in Python 3.10.12 with pandas 2.3.3, NumPy 2.2.6, scikit-learn 1.7.2 [30], XGBoost 3.2.0, statsmodels 0.15.0, SciPy 1.15.3 and matplotlib 3.10.9. SMOTE, target encoding, fairness metrics and meta-analysis were implemented in purpose-written code. The study is reported according to the TRIPOD+AI statement [31] (Additional file 1).

Results

Study population

The analysis included 357353 women and 45739 men (Table 1). Ever screening was reported by 5590 women for cervical cancer (weighted 1.97%), 2261 for breast cancer (0.88%) and 2467 for oral cancer (0.94%), and by 501 men for oral cancer (1.26%).

Individual versus area predictors

Models using only individual predictors discriminated modestly: XGBoost AUC 0.656 (95% CI 0.644-0.669) for cervical, 0.688 (0.669-0.706) for breast and 0.612 (0.593-0.631) for oral screening in women, and 0.579 (0.537-0.624) for oral screening in men (Table 2, Figs. 2 and 3a). Area predictors alone achieved 0.818, 0.807, 0.783 and 0.631 respectively. Adding individual predictors to area predictors increased the AUC by at most 0.01 in women (combined 0.824, 0.814 and 0.784). Logistic regression performed as well as XGBoost and better for men (combined AUC 0.744 vs 0.668). Combined models were well calibrated (slopes 0.91-0.98; O/E 1.00-1.05), and survey-weighted AUCs were similar (0.77-0.83).

Area features accounted for 69%, 59% and 62% of total SHAP importance for cervical, breast and oral screening in women, 60% in the pooled oral model and 30% in men (Fig. 3b). Within states, individual-only models ranked individuals barely better than chance (median within-state AUC 0.52-0.58), and combined models reached a median of 0.61-0.69 by exploiting differences between districts (Fig. 3c).

Transportability across states

In leave-one-state-out validation (30-32 states for women's outcomes and 17 for men), pooled AUCs of individual-level XGBoost models were 0.529 (95% CI 0.515-0.543) for cervical, 0.542 (0.522-0.561) for breast and 0.512 (0.489-0.535) for oral screening in women and 0.530 (0.500-0.560) for oral screening in men, with prediction intervals including 0.5 (Table 3, Fig. 4a). Logistic regression gave similar results (pooled AUC 0.513-0.554). Calibration did not transfer: the O/E ratio in held-out states ranged from 0.17 to 15.4, with prediction intervals as wide as 0.03 to 15 (Fig. 4b).

Fairness audit

With the top 10% of predicted risk flagged, sensitivity varied markedly across groups (Fig. 5a; Additional file 2, Table S3). Among the poorest women, sensitivity was 0.27 for cervical, 0.38 for breast and 0.18 for oral screening, compared with 0.58-0.68 in the two richest quintiles. Scheduled tribe women had sensitivities of 0.22-0.43. Equal opportunity differences were 0.30-0.45 across wealth quintiles and 0.31-0.46 across caste or tribe groups, with confidence intervals excluding zero (Table 4). The poorest quintile was flagged at 15-31% of the rate of the most-flagged quintile. In the pooled oral model, sensitivity was 0.51 for women and 0.34 for men (equal opportunity difference 0.17, 95% CI 0.08-0.26). These disparities arose despite good within-group discrimination (for example, AUC 0.82 among the poorest women for cervical screening) and O/E ratios close to 1 in most subgroups, indicating that a single threshold concentrates outreach in groups with higher baseline uptake.

Bias mitigation

Group-specific thresholds reduced the wealth equal opportunity difference from 0.357 to 0.096 for cervical screening and from 0.365 to 0.088 for oral screening, and the sex difference from

183 0.168 to 0.086, without changing the AUC and with a small increase in the proportion flagged
184 (10.0% to 11.3%) (Table 4, Fig. 5b). Reweighing reduced the wealth disparity for cervical
185 screening to 0.077 at a small cost in AUC (0.824 to 0.818), but was less effective for oral
186 screening by wealth (0.247) and increased the sex disparity (0.248). Sex-specific models reduced
187 the sex disparity only slightly (0.149).

188 ***Resampling before data splitting***

189 Without resampling, the cervical model with individual predictors had an AUC of 0.638 and an
190 accuracy of 0.984, identical to the accuracy of predicting that nobody was screened (Fig. 6a;
191 Additional file 2, Table S7). SMOTE applied to the training set alone did not improve
192 discrimination (AUC 0.606). When SMOTE was applied before the train-test split, the AUC rose
193 to 0.967 and sensitivity to 0.85, and random oversampling before the split gave an AUC of 0.787.

194 ***Decision curve analysis***

195 Across threshold probabilities of 0.5-6%, models with area predictors had higher net benefit than
196 screening everyone or no one, whereas the net benefit of individual-only models approached zero
197 above thresholds of about 2-3% (Fig. 6b-d). At a 2% threshold, the net benefit for cervical
198 screening was 7.1 per 1000 for the combined model and 2.3 per 1000 for the individual model.

199 ## Discussion

200 In this national study of 403092 adults, machine learning models of cancer screening uptake in
201 India performed well in conventional internal validation, but their performance came almost
202 entirely from geography. Individual characteristics added little beyond the state and district in
203 which a person lived, individual-level models performed at chance when applied to a new state,
204 and calibration failed across states. When used to select individuals for outreach, the models
205 systematically under-served poorer, scheduled tribe and male respondents. Group-specific

thresholds reduced these disparities substantially without loss of discrimination. Resampling before data splitting produced near-perfect but spurious performance.

These findings help to interpret recent studies. An analysis of NFHS-5 breast screening reported AUROCs of about 0.99 after SMOTE [6]; for an outcome with 0.9% prevalence and predictors similar to ours, our experiment suggests that such results are consistent with information leakage from resampling before splitting rather than genuine predictive ability [21]. A weighted random forest study of NFHS-5 cervical screening reported an AUC of 0.71 [5], close to our individual-level estimate, and a study using Demographic and Health Surveys from sub-Saharan Africa identified predictors of screening uptake with machine learning [7]. None of these studies assessed transportability or fairness. Our results agree with evidence from Indian survey data that machine learning models can perform unequally across socioeconomic and caste groups [18] and with broader evidence that prediction models often fail to generalise beyond their development setting [8-10].

The dominance of geography is substantively meaningful. Screening uptake in India reflects where population-based screening has been implemented, which differs greatly between states, rather than individual readiness to be screened. A model that learns district screening rates can rank people well nationally but cannot tell a programme manager which residents of a low-coverage district are least likely to be screened, because individual-level discrimination within states is close to chance. For programme planning, district coverage estimates provide the same information more simply and transparently. Targeting individuals with such models would also be inequitable: because outreach would be concentrated where uptake is already higher, the poorest and most marginalised groups would receive the least outreach, an algorithmic form of the inverse care law [32].

Fairness-aware deployment is possible. Group-specific thresholds, a simple post-processing step, nearly eliminated wealth and sex disparities in sensitivity while preserving discrimination, at the cost of a modest increase in the number flagged. Reweighing was effective for one outcome but not another and worsened the sex disparity, illustrating that mitigation strategies must be evaluated for each model and attribute. Decisions about which disparities to correct, and at what cost, require engagement with programme managers and affected communities and cannot be delegated to the algorithm [13, 15].

Our findings have practical implications. First, screening-uptake models built on survey data should report performance with and without area predictors, within-area discrimination and leave-one-region-out validation, so that readers can judge what the model actually learns. Second, fairness audits with subgroup sensitivity and flag rates should accompany any proposal to use such models for targeting. Third, class imbalance should not be corrected by resampling, and any resampling must be confined to training data within cross-validation [20, 21]. Fourth, for equitable screening policy, investment in programme coverage in under-served districts is likely to achieve more than individual-level risk targeting.

This study has strengths, including a large nationally representative sample of both sexes, multiple screening outcomes, explicit separation of individual and area information with leakage-safe target encoding, internal-external validation across 36 states with meta-analytic pooling, a fairness audit across five attributes with confidence intervals, comparison of mitigation strategies, decision curve analysis and a reproducible demonstration of resampling leakage. It also has limitations. Screening was self-reported as ever screened, without timing or setting. The data are cross-sectional and from a single survey round, so temporal validation was not possible. Models were trained without survey weights, although weighted evaluation gave similar results. The outcome is screening uptake, which itself reflects existing inequities, so fairness here concerns

the equitable allocation of outreach rather than clinical need. Hyperparameters were prespecified rather than tuned, but the equivalence of logistic regression and XGBoost suggests that tuning would not change the conclusions. Men were sampled in a subsample of households, which limited power for men.

Conclusions

Machine learning models of cancer screening uptake built from Indian national survey data predict where screening programmes operate rather than who takes up screening. They do not transfer between states, and when used to target outreach they allocate it away from the poorest and most marginalised groups unless fairness is addressed explicitly. Such models should not be used to target individuals; reporting of geographic contribution, transportability and subgroup performance, and avoidance of resampling leakage, should become standard for population health prediction models in low- and middle-income countries.

List of abbreviations

AUC: Area under the receiver operating characteristic curve

CI: Confidence interval

DHS: Demographic and Health Surveys

NFHS-5: Fifth National Family Health Survey

O/E: Observed to expected ratio

OBC: Other backward class

SC: Scheduled caste

SHAP: SHapley Additive exPlanations

SMOTE: Synthetic minority oversampling technique

ST: Scheduled tribe

TRIPOD+AI: Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis, artificial intelligence extension

XGBoost: Extreme gradient boosting

Declarations

Ethics approval and consent to participate

This study is a secondary analysis of anonymised, publicly available data. The NFHS-5 protocol was approved by the Ethics Review Board of the International Institute for Population Sciences, Mumbai, and by the ICF Institutional Review Board, and informed consent was obtained from all participants by the survey teams. Access to the data was granted by the DHS Program. The study was conducted in accordance with the Declaration of Helsinki. As a secondary analysis of anonymised, publicly available data, it was exempt from institutional ethics committee review.

Consent for publication

Not applicable.

Availability of data and materials

The NFHS-5 datasets analysed during the current study are available from the DHS Program to registered users on request (https://dhsprogram.com) [22]. The analysis code is available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable. This research received no specific grant from any funding agency.

Authors' contributions

GHP conceived the study, designed the methodology, curated the data, performed the analyses, developed the software code and figures and wrote the first draft. DSK contributed to the conceptualisation and interpretation, supervised the study and reviewed and edited the manuscript. TBD contributed to the methodology, validated the analyses and reviewed and edited the manuscript. VA contributed to the machine learning methodology, validated the models and code, and reviewed and edited the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors thank the DHS Program and the International Institute for Population Sciences for access to the NFHS-5 data.

References

1. Ministry of Health and Family Welfare, Government of India. Operational framework: management of common cancers. New Delhi: Ministry of Health and Family Welfare; 2016.

2. International Institute for Population Sciences (IIPS), ICF. National Family Health Survey (NFHS-5), 2019-21: India. Mumbai: IIPS; 2021.

3. Gopika MG, Prabhu PR, Thulaseedharan JV. Status of cancer screening in India: an alarm signal from the National Family Health Survey (NFHS-5). J Family Med Prim Care. 2022. https://doi.org/10.4103/jfmpc.jfmpc_1140_22

4. Olickal JJ, Devasia JT, Thankappan KR. Prevalence and factors associated with oral cancer screening among tobacco users in India: findings from the National Family Health Survey-5. Indian J Public Health. 2025;69(4):483-9. https://doi.org/10.4103/ijph.ijph_534_24

5. Kolasseri AE, Bhimavarapu V. Determinants of cervical cancer screening in early middle-aged women in India using weighted random forest. Indian J Gynecol Oncol. 2025;23:98. https://doi.org/10.1007/s40944-025-01027-8

6. Tanveer A, Ali RH, Majhi J, Mukherjee M. Predicting and identifying correlates of inequalities in breast cancer screening uptake using national level data from India. Front Artif Intell. 2026;8:1729796. https://doi.org/10.3389/frai.2025.1729796

7. Baykemagn ND, Aweke MN, Mesfin A, Baffa LD, Agimas MC, Abuhay HW, et al. Identifying predictors of cervical cancer screening uptake in sub-Saharan Africa using machine learning: cross-sectional study. JMIR Public Health Surveill. 2025;11:e71677. https://doi.org/10.2196/71677

8. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. Lancet Digit Health. 2020;2(9):e489-92. https://doi.org/10.1016/S2589-7500(20)30186-2

9. Yang J, Soltan AAS, Clifton DA. Machine learning generalizability across healthcare settings: insights from multi-site COVID-19 screening. NPJ Digit Med. 2022;5:69. https://doi.org/10.1038/s41746-022-00614-9

10. Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ. 2016;353:i3140. https://doi.org/10.1136/bmj.i3140

11. Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol. 2016;69:245-7. https://doi.org/10.1016/j.jclinepi.2015.04.005

12. Debray TPA, Damen JAAG, Snell KIE, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. BMJ. 2017;356:i6460. https://doi.org/10.1136/bmj.i6460

13. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med. 2018;169(12):866-72. https://doi.org/10.7326/M18-1990

14. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-53. https://doi.org/10.1126/science.aax2342

15. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical machine learning in healthcare. Annu Rev Biomed Data Sci. 2021;4:123-44. https://doi.org/10.1146/annurev-biodatasci-092820-114757

16. Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. Adv Neural Inf Process Syst. 2016;29:3315-23.

17. Kamiran F, Calders T. Data preprocessing techniques for classification without discrimination. Knowl Inf Syst. 2012;33(1):1-33. https://doi.org/10.1007/s10115-011-0463-8

18. Lee JT, Hsu SH, Li VCS, Anindya K, Chen MH, Wang C, et al. Evaluating algorithmic fairness of machine learning models in predicting underweight, overweight, and adiposity across socioeconomic and caste groups in India: evidence from the longitudinal ageing study in India. PLOS Digit Health. 2025;4(11):e0000951. https://doi.org/10.1371/journal.pdig.0000951

19. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321-57. https://doi.org/10.1613/jair.953

20. van den Goorbergh R, van Smeden M, Timmerman D, Van Calster B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. J Am Med Inform Assoc. 2022;29(9):1525-34. https://doi.org/10.1093/jamia/ocac093

21. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns. 2023;4(9):100804. https://doi.org/10.1016/j.patter.2023.100804

22. The DHS Program. India: Standard DHS, 2019-21 (NFHS-5) dataset. Rockville, Maryland: ICF. https://dhsprogram.com/data/dataset/India_Standard-DHS_2020.cfm. Accessed 10 Oct 2026.

23. Micci-Barreca D. A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. ACM SIGKDD Explor Newsl. 2001;3(1):27-32. https://doi.org/10.1145/507533.507538

24. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM; 2016. p. 785-94. https://doi.org/10.1145/2939672.2939785

25. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):230. https://doi.org/10.1186/s12916-019-1466-7

26. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30:4765-74.

27. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565-74. https://doi.org/10.1177/0272989X06295361

28. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29-36. https://doi.org/10.1148/radiology.143.1.7063747

29. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986;7(3):177-88. https://doi.org/10.1016/0197-2456(86)90046-2

30. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825-30.

31. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. https://doi.org/10.1136/bmj-2023-078378

32. Hart JT. The inverse care law. Lancet. 1971;1(7696):405-12. https://doi.org/10.1016/S0140-6736(71)92410-X

Figure titles and legends

Fig. 1 Study design and analytical framework

Overview of the data, outcomes, feature sets and models, and the four evaluation components: internal validation, transportability (leave-one-state-out internal-external validation), fairness audit with bias mitigation, and demonstration of resampling before the train-test split. NFHS-5, fifth National Family Health Survey.

Fig. 2 Discrimination and calibration of XGBoost models by feature set

a Receiver operating characteristic curves in the test set for cervical screening and breast examination in women and oral examination in both sexes pooled, comparing individual, area and combined (individual plus area) predictors; AUC values are shown in the legend. b Calibration plots by deciles of predicted probability; the dashed line indicates perfect calibration.

Fig. 3 Contribution of geography to predictive performance

a Test-set AUC with 95% bootstrap confidence intervals for each outcome by feature set and model (XGB, XGBoost; LR, logistic regression). b Share of total mean absolute SHAP value

attributable to area predictors (state and district screening rate) in the combined XGBoost models. c Distribution of AUCs calculated within each state for individual-only and combined XGBoost models; each point is a state with at least ten screened respondents in the test set; the dotted line indicates chance.

Fig. 4 Transportability across states (leave-one-state-out validation)

a AUC of individual-level XGBoost models in each held-out state; point size is proportional to the number of screened respondents; diamonds show random-effects pooled estimates with 95% prediction intervals. b Observed to expected ratio in each held-out state on a log scale; diamonds show pooled estimates. The dotted lines indicate chance discrimination (a) and perfect calibration-in-the-large (b).

Fig. 5 Fairness audit and bias mitigation

a Sensitivity by subgroup when the 10% of individuals with the highest predicted probability are flagged, for combined XGBoost models. b Equal opportunity difference (range of sensitivity across groups) under a single threshold, reweighing and group-specific thresholds; numbers above bars give the overall AUC. SC, scheduled caste; ST, scheduled tribe; OBC, other backward class.

Fig. 6 Effect of resampling before data splitting and decision curve analysis

a AUC and accuracy of cervical screening models with individual predictors under five pipelines; resampling before the train-test split (shaded) leaks information from the evaluation set into training. b-d Net benefit per 1000 people across threshold probabilities for models with individual, area and combined predictors, compared with screening everyone or no one.

Tables

Table 1 Characteristics of the study population, women and men aged 30-49 years, NFHS-5

Characteristic	Category	Women, n (weighted %)	Men, n (weighted %)
Age group (years)	30-34	99084 (27.5)	12612 (27.1)
	35-39	96074 (26.7)	12330 (27.1)
	40-44	79775 (22.4)	10397 (22.6)
	45-49	82420 (23.4)	10400 (23.3)
Education	No education	130054 (35.4)	7137 (15.3)
	Primary	55241 (15.5)	6806 (15.3)
	Secondary	139755 (38.7)	24822 (53.2)
	Higher	32303 (10.4)	6974 (16.3)
Wealth quintile	Poorest	72074 (17.8)	8927 (16.4)
	Poorer	76424 (19.1)	9975 (19.3)
	Middle	74540 (20.5)	9763 (21.3)
	Richer	69800 (21.2)	9169 (22.9)
	Richest	64515 (21.3)	7905 (20.0)
Residence	Rural	264779 (66.1)	33790 (65.1)
	Urban	92574 (33.9)	11949 (34.9)
Religion	Hindu	271320 (82.3)	34714 (80.5)
	Muslim	40352 (12.1)	5150 (14.2)
	Christian	26913 (2.6)	3465 (2.7)
	Other	18768 (3.0)	2410 (2.5)
Caste or tribe	SC	66434 (21.2)	8422 (20.6)
	ST	66777 (9.1)	8961 (8.9)
	OBC	136093 (42.9)	17416 (41.6)
	None of these	69003 (21.3)	8698 (21.1)
	Not known	19046 (5.5)	2242 (7.8)
Region	North	72192 (13.9)	9097 (8.2)
	Central	76015 (22.4)	9774 (10.3)
	East	55598 (21.9)	6630 (24.8)
	Northeast	52928 (3.8)	7308 (6.0)
	West	37811 (14.8)	5333 (24.0)
	South	62809 (23.2)	7597 (26.7)
Currently married	Yes	323923 (91.0)	41701 (91.6)
Health insurance	Yes	129940 (34.4)	18326 (37.3)
Weekly media exposure	Yes	186143 (55.7)	26678 (63.4)
Any tobacco use	Yes	33302 (6.2)	24577 (50.7)
Smoked tobacco	Yes	2747 (0.5)	12541 (24.8)
Smokeless tobacco	Yes	29866 (5.4)	16795 (33.6)
Alcohol use	Yes	9705 (1.1)	15344 (29.3)
Outcomes (ever screened)	Cervical screening	5590 (1.97)	Not applicable
	Breast examination	2261 (0.88)	Not applicable
	Oral cavity examination	2467 (0.94)	501 (1.26)

Unweighted numbers and survey-weighted percentages. SC, scheduled caste; ST, scheduled tribe; OBC, other backward class

Table 2 Test-set performance by outcome, feature set and model

Feature set and model	AUC (95% CI)	Weighted AUC	PR-AUC	Brier	Calibration slope	O/E

Cervical screening (women):						
Individual, LR	0.650 (0.637-0.663)	0.651	0.030	0.0153	1.01	1.00
Individual, XGB	0.656 (0.644-0.669)	0.651	0.032	0.0153	0.90	1.00
Area, LR	0.821 (0.811-0.829)	0.827	0.091	0.0147	0.97	1.00
Area, XGB	0.818 (0.808-0.827)	0.826	0.089	0.0147	0.96	1.00
Individual + area, LR	0.824 (0.814-0.833)	0.830	0.095	0.0147	0.97	1.00
Individual + area, XGB	0.824 (0.814-0.832)	0.830	0.101	0.0147	0.97	1.00
Breast examination (women):						
Individual, LR	0.681 (0.660-0.701)	0.674	0.014	0.0063	0.94	1.00
Individual, XGB	0.688 (0.669-0.706)	0.691	0.015	0.0063	0.88	1.00
Area, LR	0.811 (0.792-0.829)	0.833	0.060	0.0061	0.98	1.01
Area, XGB	0.807 (0.788-0.827)	0.831	0.055	0.0061	0.96	1.04
Individual + area, LR	0.817 (0.798-0.835)	0.836	0.062	0.0061	0.98	1.01
Individual + area, XGB	0.814 (0.798-0.832)	0.832	0.059	0.0061	0.94	1.05
Oral examination (women):						
Individual, LR	0.617 (0.601-0.637)	0.643	0.011	0.0069	0.99	1.01
Individual, XGB	0.612 (0.593-0.631)	0.646	0.011	0.0069	0.70	1.00
Area, LR	0.785 (0.764-0.804)	0.816	0.056	0.0067	1.01	1.01
Area, XGB	0.783 (0.762-0.804)	0.815	0.054	0.0067	0.99	1.04
Individual + area, LR	0.789 (0.770-0.809)	0.819	0.067	0.0066	1.01	1.01
Individual + area, XGB	0.784 (0.764-0.805)	0.820	0.061	0.0066	0.97	1.04
Oral examination (men):						
Individual, LR	0.611 (0.560-0.657)	0.683	0.018	0.0108	1.13	1.00
Individual, XGB	0.579 (0.537-0.624)	0.655	0.019	0.0108	0.62	0.99
Area, LR	0.733 (0.691-0.779)	0.821	0.043	0.0106	1.19	0.99
Area, XGB	0.631 (0.582-0.680)	0.727	0.026	0.0107	0.92	1.02
Individual + area, LR	0.744 (0.708-0.787)	0.832	0.044	0.0106	1.12	0.99
Individual + area, XGB	0.668 (0.614-0.714)	0.768	0.035	0.0107	0.91	1.02
Oral examination (pooled):						
Individual, LR	0.625 (0.607-0.643)	0.650	0.012	0.0073	1.02	1.00
Individual, XGB	0.627 (0.610-0.647)	0.667	0.012	0.0073	0.81	1.00
Area, LR	0.758 (0.742-0.776)	0.801	0.049	0.0071	1.00	1.00
Area, XGB	0.760 (0.744-0.776)	0.803	0.050	0.0071	0.97	1.03
Individual + area, LR	0.773 (0.757-0.790)	0.812	0.056	0.0071	1.00	1.00
Individual + area, XGB	0.775 (0.760-0.790)	0.820	0.060	0.0071	0.96	1.02

LR, logistic regression; XGB, XGBoost; AUC, area under the receiver operating characteristic curve with 200 bootstrap confidence intervals; weighted AUC, survey-weighted AUC; PR-AUC, area under the precision-recall curve (baseline equals outcome prevalence); O/E, observed to expected ratio

Table 3 Leave-one-state-out validation of individual-level models: random-effects meta-analysis

Outcome	Model	States	Pooled AUC (95% CI)	95% prediction interval	I squared (%)	Pooled O/E (range)
Cervical screening (women)	XGBoost	32	0.529 (0.515-0.543)	0.473-0.584	55.3	0.67 (0.17-7.23)
Cervical screening (women)	Logistic regression	32	0.539 (0.522-0.557)	0.461-0.616	71.2	0.66 (0.17-6.96)
Breast examination (women)	XGBoost	30	0.542 (0.522-0.561)	0.473-0.609	42.4	0.69 (0.24-10.22)
Breast examination (women)	Logistic regression	30	0.554 (0.534-0.574)	0.485-0.621	42.3	0.70 (0.24-10.68)
Oral examination (women)	XGBoost	30	0.512 (0.489-0.535)	0.413-0.610	66.6	0.78 (0.20-15.39)

Oral examination (women)	Logistic regression	30	0.522 (0.500-0.544)	0.431-0.612	62.4	0.78 (0.22-15.41)
Oral examination (men)	XGBoost	17	0.530 (0.500-0.560)	0.477-0.583	9.6	1.14 (0.58-5.04)
Oral examination (men)	Logistic regression	17	0.513 (0.475-0.552)	0.398-0.627	42.3	1.12 (0.58-5.05)

441 States with at least ten screened respondents were held out in turn. Pooled AUCs were estimated on the logit scale;

442 O/E, observed to expected ratio (pooled estimate and range across states)

443 **Table 4 Fairness audit and bias mitigation for combined XGBoost models**

Model, attribute and strategy	Equal opportunity difference (95% CI)	Flag-rate ratio (lowest/highest)	Subgroup AUC range or overall performance
Panel A. Disparities with a single threshold (top 10% flagged):			
Oral examination (pooled), sex	0.168 (0.084-0.258)	0.59	0.65-0.79
Oral examination (pooled), wealth quintile	0.365 (0.264-0.469)	0.31	0.70-0.80
Oral examination (pooled), caste or tribe	0.347 (0.241-0.536)	0.34	0.70-0.83
Oral examination (pooled), residence	0.063 (0.011-0.127)	0.68	0.74-0.78
Oral examination (pooled), religion	0.348 (0.228-0.515)	0.28	0.64-0.79
Cervical screening (women), wealth quintile	0.357 (0.290-0.432)	0.18	0.81-0.84
Cervical screening (women), caste or tribe	0.456 (0.348-0.574)	0.26	0.71-0.84
Cervical screening (women), residence	0.123 (0.076-0.178)	0.59	0.82-0.82
Cervical screening (women), religion	0.319 (0.204-0.408)	0.28	0.73-0.83
Breast examination (women), wealth quintile	0.303 (0.224-0.444)	0.15	0.73-0.83
Breast examination (women), caste or tribe	0.312 (0.227-0.504)	0.38	0.72-0.84
Breast examination (women), residence	0.160 (0.097-0.222)	0.34	0.80-0.81
Breast examination (women), religion	0.447 (0.235-0.656)	0.23	0.75-0.85
Oral examination (women), wealth quintile	0.445 (0.337-0.574)	0.22	0.65-0.81
Oral examination (women), caste or tribe	0.419 (0.285-0.535)	0.23	0.71-0.82
Oral examination (women), residence	0.046 (0.004-0.114)	0.53	0.78-0.78
Oral examination (women), religion	0.178 (0.089-0.367)	0.36	0.73-0.79
Panel B. Bias mitigation:			
Cervical screening (women), wealth quintile: Baseline (single threshold)	0.357	0.18	AUC 0.824; sensitivity 0.531; flagged 0.100
Cervical screening (women), wealth quintile: Reweighing (pre-processing)	0.077	0.83	AUC 0.818; sensitivity 0.524; flagged 0.102
Cervical screening (women), wealth quintile: Group-specific thresholds (post-processing)	0.096	0.72	AUC 0.824; sensitivity 0.544; flagged 0.113
Oral examination (pooled), sex: Baseline (single threshold)	0.168	0.59	AUC 0.775; sensitivity 0.483; flagged 0.099
Oral examination (pooled), sex: Reweighing (pre-processing)	0.248	0.85	AUC 0.769; sensitivity 0.481; flagged 0.096
Oral examination (pooled), sex: Group-specific thresholds (post-processing)	0.086	0.42	AUC 0.775; sensitivity 0.490; flagged 0.099
Oral examination (pooled), sex: Sex-specific models	0.149	0.58	AUC 0.774; sensitivity 0.494; flagged 0.099
Oral examination (pooled), wealth quintile: Baseline (single threshold)	0.365	0.31	AUC 0.775; sensitivity 0.483; flagged 0.099
Oral examination (pooled), wealth quintile: Reweighing (pre-processing)	0.247	0.81	AUC 0.765; sensitivity 0.488; flagged 0.107
Oral examination (pooled), wealth quintile: Group-specific thresholds (post-processing)	0.088	0.48	AUC 0.775; sensitivity 0.484; flagged 0.115

444 Equal opportunity difference: range of sensitivity across groups when the top 10% of predicted probability is flagged

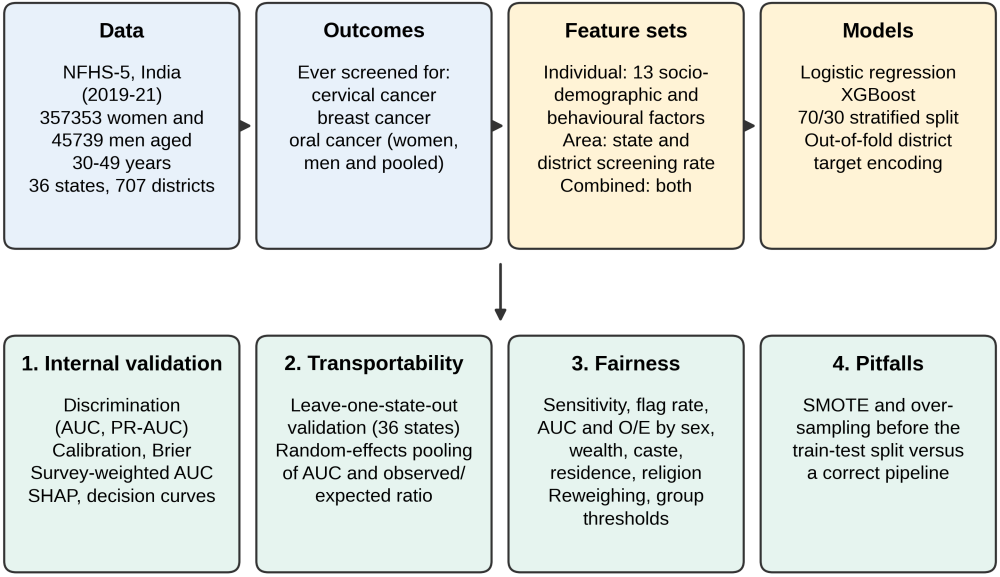
445 (100 bootstrap confidence intervals). Flag-rate ratio: lowest divided by highest proportion flagged across groups.

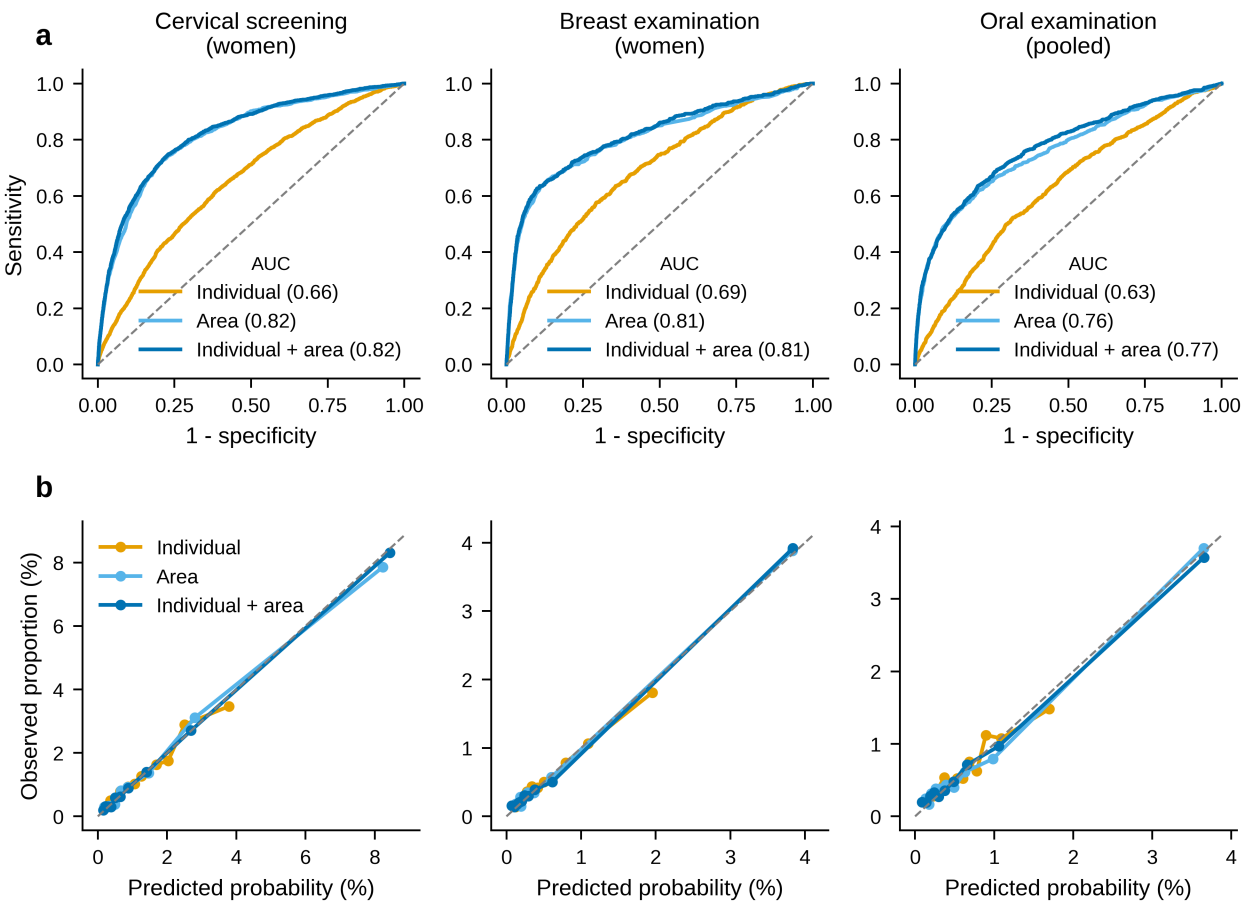
446 Group-specific thresholds were chosen on out-of-fold training predictions

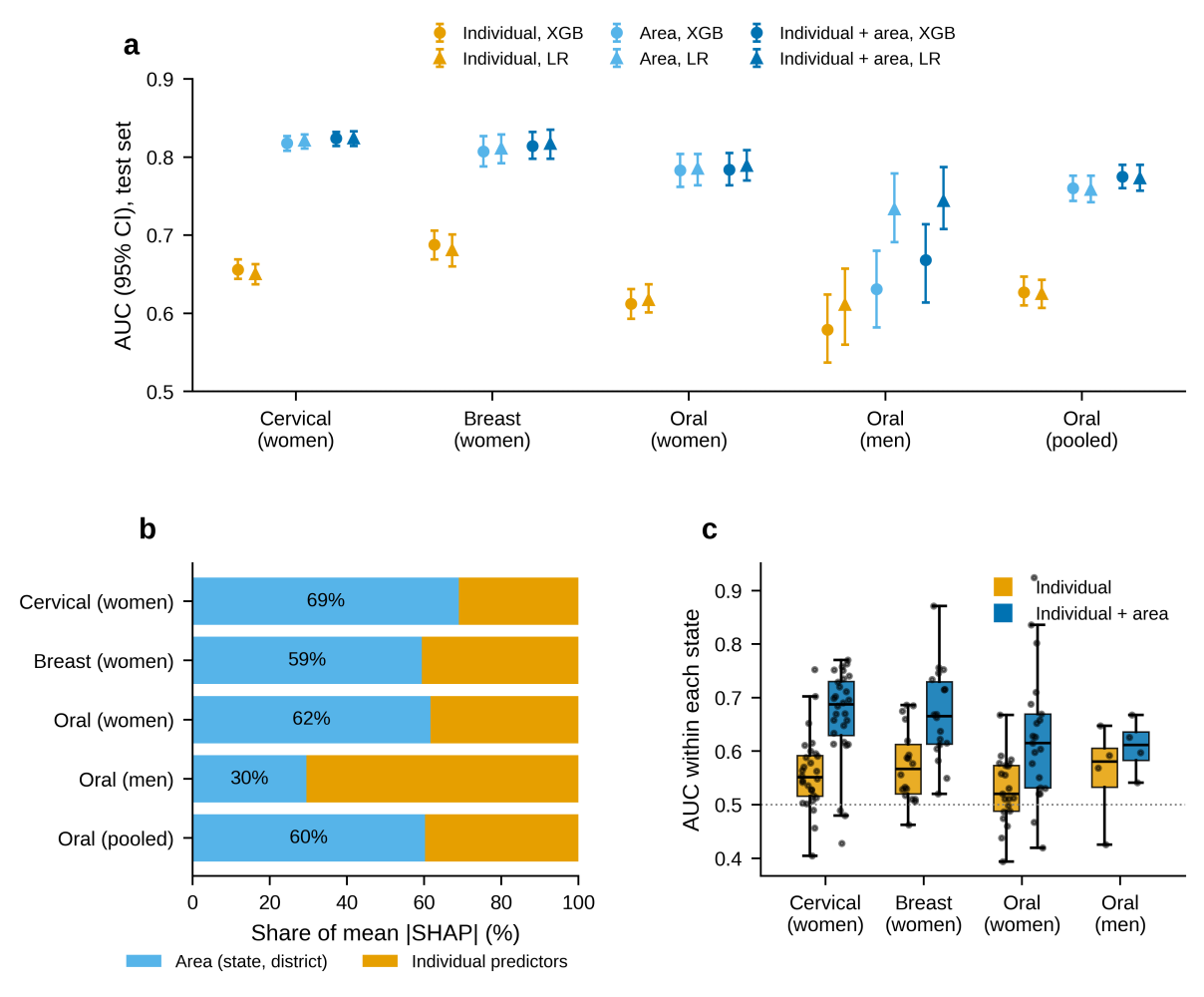
Additional files

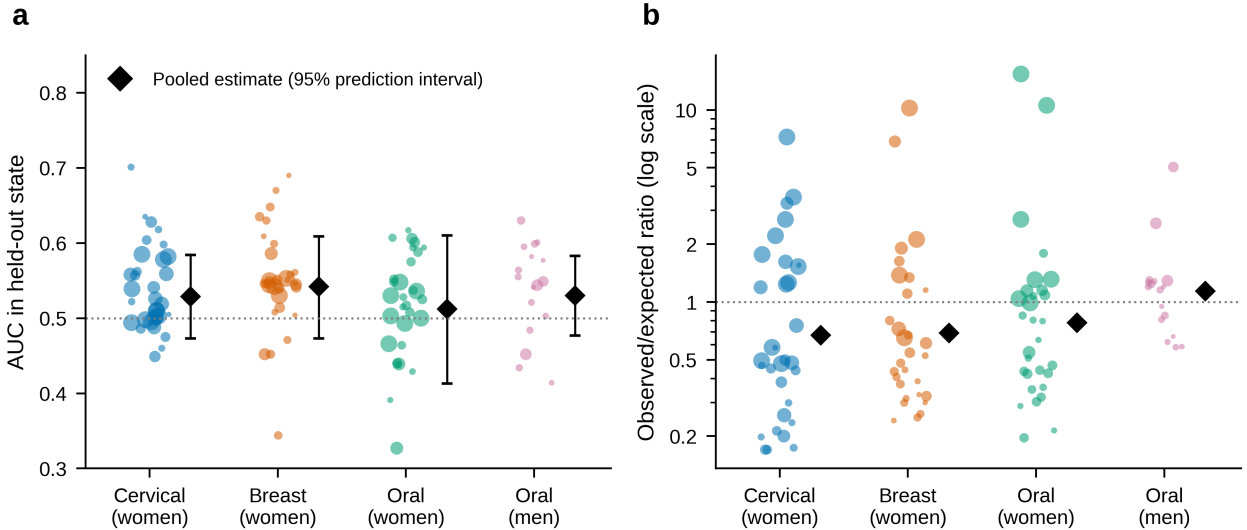
Additional file 1 (.docx): TRIPOD+AI checklist. Location of reporting items in the manuscript.

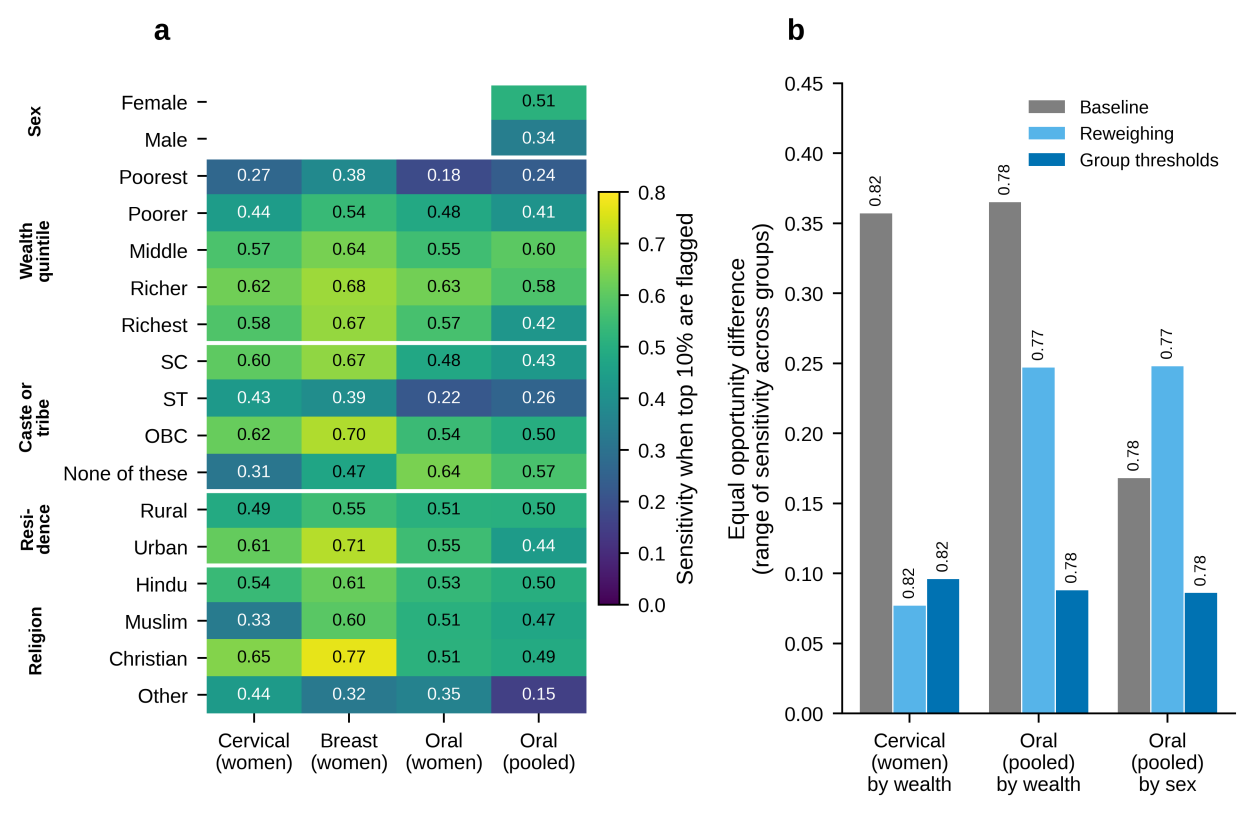
Additional file 2 (.docx): Supplementary methods and tables. Table S1, predictor definitions; Table S2, model hyperparameters; Table S3, full fairness audit by subgroup; Table S4, leave-one-state-out results by state; Table S5, within-state AUC summary; Table S6, SHAP importance by predictor group; Table S7, resampling pipelines; Table S8, full performance metrics.

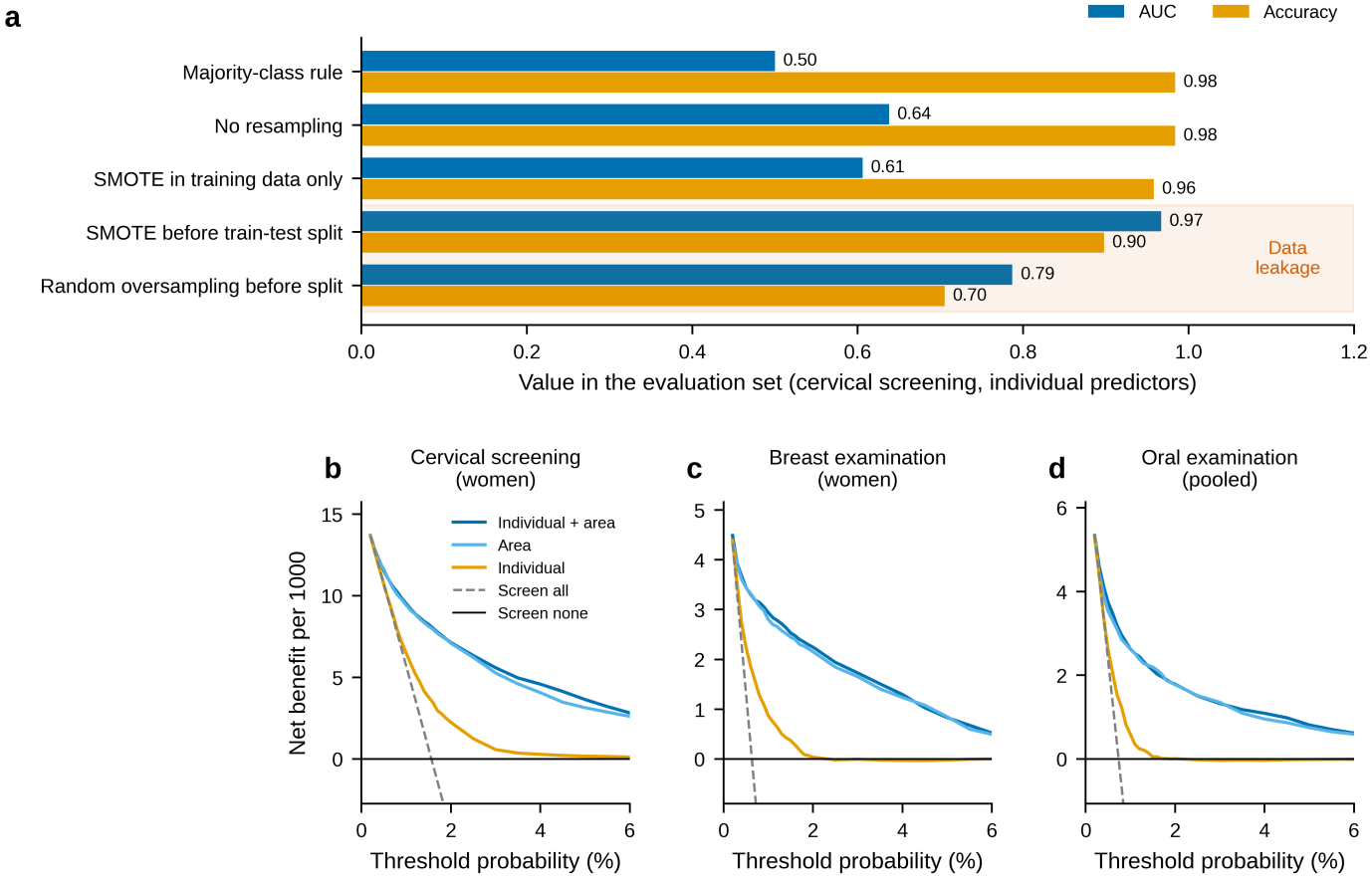

Data
NFHS-5, India
(2019-21)
357353 women and
45739 men aged
30-49 years
36 states, 707 districts
Outcomes
Ever screened for:
cervical cancer
breast cancer
oral cancer (women,
men and pooled)
Feature sets
Individual: 13 socio-
demographic and
behavioural factors
Area: state and
district screening rate
Combined: both
Models
Logistic regression
XGBoost
70/30 stratified split
Out-of-fold district
target encoding
1. Internal validation
Discrimination
(AUC, PR-AUC)
Calibration, Brier
Survey-weighted AUC
SHAP, decision curves
2. Transportability
Leave-one-state-out
validation (36 states)
Random-effects pooling
of AUC and observed/
expected ratio
3. Fairness
Sensitivity, flag rate,
AUC and O/E by sex,
wealth, caste,
residence, religion
Reweighing, group
thresholds
4. Pitfalls
SMOTE and over-
sampling before the
train-test split versus
a correct pipeline


a
Cervical screening (women)
Breast examination (women)
Oral examination (pooled)
Sensitivity
1 - specificity
AUC
Individual (0.66)
Area (0.82)
Individual + area (0.82)
Individual (0.69)
Area (0.81)
Individual + area (0.81)
Individual (0.63)
Area (0.76)
Individual + area (0.77)
b
Observed proportion (%)
Predicted probability (%)
Individual
Area
Individual + area


a
Individual, XGB
Individual, LR
Area, XGB
Area, LR
Individual + area, XGB
Individual + area, LR
AUC (95% CI), test set
0.9
0.8
0.7
0.6
0.5
Cervical (women)
Breast (women)
Oral (women)
Oral (men)
Oral (pooled)
b
Cervical (women)
Breast (women)
Oral (women)
Oral (men)
Oral (pooled)
69%
59%
62%
30%
60%
0
20
40
60
80
100
Share of mean |SHAP| (%)
Area (state, district)
Individual predictors
c
AUC within each state
0.9
0.8
0.7
0.6
0.5
0.4
Individual
Individual + area
Cervical (women)
Breast (women)
Oral (women)
Oral (men)


a
b
Pooled estimate (95% prediction interval)
AUC in held-out state
0.8
0.7
0.6
0.5
0.4
0.3
Observed/expected ratio (log scale)
10
5
2
1
0.5
0.2
Cervical (women)
Breast (women)
Oral (women)
Oral (men)


a
Sex
Female
Male
0.51
0.34
Wealth quintile
Poorest
Poorer
Middle
Richer
Richest
0.27 0.38 0.18 0.24
0.44 0.54 0.48 0.41
0.57 0.64 0.55 0.60
0.62 0.68 0.63 0.58
0.58 0.67 0.57 0.42
Caste or tribe
SC
ST
OBC
None of these
0.60 0.67 0.48 0.43
0.43 0.39 0.22 0.26
0.62 0.70 0.54 0.50
0.31 0.47 0.64 0.57
Resi-dence
Rural
Urban
0.49 0.55 0.51 0.50
0.61 0.71 0.55 0.44
Religion
Hindu
Muslim
Christian
Other
0.54 0.61 0.53 0.50
0.33 0.60 0.51 0.47
0.65 0.77 0.51 0.49
0.44 0.32 0.35 0.15
Cervical (women)
Breast (women)
Oral (women)
Oral (pooled)
Sensitivity when top 10% are flagged
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8
b
Equal opportunity difference (range of sensitivity across groups)
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45
Baseline
Reweighing
Group thresholds
0.82 0.82 0.78 0.78 0.77 0.78 0.78 0.77 0.78
Cervical (women) by wealth
Oral (pooled) by wealth
Oral (pooled) by sex


a
AUC
Accuracy
Majority-class rule
No resampling
SMOTE in training data only
SMOTE before train-test split
Random oversampling before split
0.50
0.98
0.64
0.98
0.61
0.96
0.97
0.90
0.79
0.70
Data leakage
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Value in the evaluation set (cervical screening, individual predictors)
b
Cervical screening (women)
c
Breast examination (women)
d
Oral examination (pooled)
Individual + area
Area
Individual
Screen all
Screen none
Net benefit per 1000
Threshold probability (%)
Threshold probability (%)
Threshold probability (%)