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.

Microscopic verification of cervical cancer diagnoses in the Mashhad population-based cancer registry, 2017–2021: variation by age, place and period

Veda Vakili

vakiliv@mums.ac.ir

Mashhad University of Medical Sciences

Fatemeh Khabbaz

Mashhad University of Medical Sciences

Mojtaba Mousavi Bazaz

Mashhad University of Medical Sciences

Azadeh Soltanifar

Mashhad University of Medical Sciences

Research Article

Keywords: cervical cancer, cancer registry, microscopic verification, data quality, incidence, machine learning, Iran

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11506791/v1

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

Additional Declarations: No competing interests reported.

Abstract

Background

Microscopic verification (MV) is a core indicator of cancer registry data quality, but its variation within registries is rarely examined. We described MV and the proportion of death-certificate-only (DCO) registrations for cervical cancer in the Mashhad University of Medical Sciences (MUMS) population-based cancer registry, estimated registered incidence, and examined which registry-available information is associated with non-verification.

Methods

We analysed all 284 cervical cancer records reported in Persian years 1396–1399 (2017–2021), extracted on [extraction date]; one woman registered twice was counted once (283 women). MV was a histopathological or cytological diagnosis. Fields created by the diagnostic process were excluded to avoid leakage. We estimated MV and DCO% by year, half-year, age and residence; registered incidence per 100,000 women (age-standardized, Segi; gamma confidence intervals; Poisson trend); and adjusted odds ratios for MV. Four prediction algorithms were compared as a benchmark (repeated cross-validation, calibration, temporal validation, permutation test).

Results

MV was 67.5% (95% CI 61.8–72.7) and DCO 7.1% (4.6–10.7). MV was lowest in 1396 (55.3%), among women aged ≥ 70 years (46.7%) and in MUMS counties other than Mashhad (40.5%), and was 100% in 16 women with unknown city. Adjusted odds of MV were lower with older age (OR 0.68 per 10 years) and in other MUMS counties (0.38; 0.62, 0.24–1.56 without Taybad). Registered incidence among residents of MUMS counties was 2.20 per 100,000 women (age-standardized 2.26, 95% CI 1.96–2.60), falling by about 15% per year (rate ratio 0.85, 0.76–0.96; 0.87, 0.73–1.05 without 1399), not separable from incomplete or late registration. Cross-validated AUC was about 0.69 for all four algorithms (0.67–0.68 without year); temporal validation gave 0.58–0.71 and the permutation test $p \leq 0.005$.

Conclusions

About one third of registered cervical cancers lacked microscopic verification, with gradients by age, area and period. Registry-available variables predicted non-verification only modestly, and algorithms of very different complexity performed alike, so the choice of algorithm is immaterial. The findings support routine audit of MV and DCO by county and period rather than individual prediction. Registered incidence needs a completeness assessment.

Background

Cervical cancer remains one of the most frequent cancers among women worldwide, with the heaviest burden in low- and middle-income countries [1, 2]. Population-based cancer registries underpin surveillance, and their data quality is commonly judged by the proportion of cases with microscopic verification (MV) and the proportion registered from a death certificate only [3, 4]. A diagnosis made on clinical grounds or from a death certificate is less certain, and carries little information on morphology, grade and stage. In Iran, a quality assessment of the national cancer registry for 2010 reported MV of 68.7% overall (72% in females) [5], and national data for 2014 reported microscopic confirmation in 68.3% of all cancers and a cervical cancer age-standardized rate of 1.78 per 100,000 women [6]. A nationwide analysis of 2016 data reported MV of 84.7% for cervical cancer, with an age-standardized incidence of 1.90 per 100,000 women nationally and 2.75 per 100,000 in Khorasan Razavi Province [7].

Registries rarely examine which patient and service characteristics are associated with the absence of MV, although such information could guide audit and improve case ascertainment. In Iran, screening participation was 39.4% and HPV vaccination coverage 0.85% among women aged 18–45 years in 2021 [8], so diagnosis may depend largely on symptomatic presentation and on access to diagnostic services. Registry records can also be linked to primary health care (PHC) files, which are used in research on cancer risk factors, but the representativeness of linked cases has seldom been assessed. Machine-learning methods are increasingly applied to registry data, but their advantage over regression is often modest and is easily overstated when validation is optimistic [9, 10].

We analysed the cervical cancer records of the Mashhad University of Medical Sciences (MUMS) registry for 2017–2021 (Persian years 1396–1399). Our aims were to (i) describe MV and death-certificate-only registration and their variation by period, age and place; (ii) estimate registered incidence with population denominators for the registry catchment area, as a by-product that depends on registry completeness; (iii) compare linked and unlinked women to assess the representativeness of PHC-linked cases; and (iv) assess, as a benchmark exercise, whether information available at registration is associated with MV, comparing penalized regression with tree-based ensembles under leakage-free, temporally validated and selection-aware evaluation. This analysis is not meant for clinical use; it quantifies how much of non-verification is explained by information the registry holds.

Methods

Study design and data source

This was a registry-based descriptive study with prediction-model development and internal–temporal validation, reported according to the STROBE statement for observational studies [11] (Additional file 1) and, for the prediction component, TRIPOD + AI [12] (Additional file 2). Data came from the population-based cancer registry of MUMS, which registers cancers among residents of Mashhad and surrounding counties in Khorasan Razavi Province and also includes some women resident in other areas. The

registry file contained 284 records of cervical cancer (topography C53) reported in Persian years 1396–1399 (21 March 2017 to 20 March 2021). Two records belonged to the same woman (identified by her national identification number, used only for this check and then removed together with all personal identifiers); we retained her earlier record, giving 283 women (Fig. 1). Seventy-four of these women could be linked to PHC records and were the case group of our registry-linked case–control study [13]; registry records were linked to PHC records by national identification number. Because identifiers were removed before analysis, the linkage flag was carried over to the de-identified registry file by matching date of birth and date of report with the de-identified case–control file [confirm: number of ambiguous matches]. Data were extracted from the registry on [extraction date], [x] months after the end of 1399; notifications and death-certificate follow-back arriving after that date are not captured, so the most recent registrations, especially in late 1399, may be incomplete.

Outcome

The primary outcome was microscopic verification, defined as a diagnosis based on histopathology or cytology (the IARC definition [3]). Non-verified diagnoses were those based on clinical evidence only, clinical evaluation, or a death certificate. Histopathology alone as the verified category was used in a sensitivity analysis. The proportion of cases registered from a death certificate only (DCO%) was reported as a second quality indicator [3, 4]. A mortality-to-incidence ratio could be computed only with death-registry counts, which are not part of the registry file [M:I ratio to be inserted from the 1396–1399 death registry].

Variables and leakage control

Candidate predictors were those known at registration: age at report, year of report, residence (Mashhad, other MUMS counties, outside the MUMS area, unknown), registry region code (1 = urban, 2 = suburban, 3 = rural [confirm coding]; missing in 31 women) and an indicator of missing address (unknown city or missing region code, 33 women). Region was coded as suburban and rural indicators, so the 31 women with missing region were in the reference category with urban residents and were distinguished by the missing-address indicator; no other predictor had missing values, and median imputation within training folds served only as a safeguard. The missing-address indicator is not a patient characteristic but a marker of the notification source, because records without an address probably came from notification sources that carry no residence information (the source itself was not recorded in our file). Registered tumour grade, morphology and topography subsite were not used as predictors because they are determined by the diagnostic process and almost perfectly identify the outcome: all 92 non-verified women had unknown grade, a non-specific morphology code (M8000/3, which also occurred in 9 verified women) and topography C53.9, whereas all 12 C53.0, 11 C53.1 and 2 C53.8 records were verified. Persian years were coded as indicators. PHC variables (parity, body mass index, smoking, hypertension and others) were available only for the 74 linked women and were analysed separately.

Descriptive analyses and linkage comparison

MV was described as a proportion with Wilson 95% confidence interval by year, half-year, age group, residence and region, and DCO% by year. MV in the individual other MUMS counties is listed descriptively only, because counts are small; as a sensitivity analysis the county contrast was repeated without Taybad (1 of 7 verified). The p-values for the comparisons of linked and unlinked women and for subgroups are exploratory and were not adjusted for multiplicity. A linear trend over years was tested by logistic regression. The 74 PHC-linked women were compared with the 209 unlinked women, and separately with the 146 unlinked residents of MUMS counties (the only group in which linkage is possible), using the Mann–Whitney U, chi-square or Fisher's exact test.

Registered incidence

Registered incidence was calculated for women resident in the MUMS counties for which population denominators were available (Mashhad and 15 other counties; 216 women), using female population by five-year age group and year for 1396–1399 from the health-system (university) population reports, not from a census; the denominators therefore carry their own uncertainty. Women with unknown address (16) or residing outside the MUMS area (51) could not be assigned and were excluded from the numerator. Crude rates (exact Poisson 95% CI) and age-standardized rates (ASR, Segi world standard population [14]; 95% CI by the gamma method of Fay and Feuer [15]) per 100,000 women were computed by year and for the four years combined. The time trend was tested by Poisson regression of counts on year with an age-group adjustment and woman-years as offset, with a sensitivity analysis excluding 1399, whose registration may be incomplete. We had no independent measure of completeness, so these are rates of registered cancers and cannot be read as true incidence.

Regression and machine-learning analysis

Design decisions were guided by the PROBAST domains (participants, predictors, outcome and analysis) [16] and by published steps for developing and validating prediction models [17]. A multivariable logistic regression (age per 10 years, year, residence, region, missing address) gave adjusted odds ratios for MV (events per parameter 10.2). For prediction we compared five approaches: a prevalence-only reference; logistic regression with L2 penalty (penalty tuned by inner cross-validation); elastic-net logistic regression [18] with tuned penalty and L1 ratio; a shallow random forest [19] (300 trees, maximum depth 3, minimum leaf size 5); and shallow gradient boosting (maximum depth 2, learning rate 0.05, 100 iterations). Hyper-parameters were tuned within training folds and continuous variables were standardized. Performance was estimated by 20 times repeated stratified five-fold cross-validation (AUC and precision–recall AUC, with between-repeat SD). Uncertainty was assessed by an out-of-bag bootstrap (200 resamples with model fitting on each resample and evaluation on the women left out). Calibration was assessed by the calibration slope (95% interval from 1,000 bootstrap resamples), calibration-in-the-large and the Brier score compared with the prevalence-only value [20, 21]. Because

1396–1397 and 1398–1399 differ in MV prevalence, temporal validation trained all models on 1396–1397 (n = 147) and tested them on 1398–1399 (n = 136), a design that mimics prospective use [22]. Because year indicators carry no information for years unseen in training, temporal validation was run both without year columns (reported as the main temporal result) and with them. No formal sample-size calculation was made because all registered women were included; with 92 non-verified women and nine predictor parameters (about 10 events per parameter) the sample is small for flexible algorithms, so we used penalization and shallow trees and report intervals that include model fitting. The algorithm comparison is presented as a benchmark; we did not expect, and do not claim, that any algorithm is superior.

Statistical support for above-chance performance used a label-permutation test [23] that accounts for choosing the best of four models: in each of 200 permutations all four models were cross-validated (three repeats of five-fold cross-validation, applied identically to the observed data) and the maximum AUC was retained; p = (1 + number of null maxima at least as large as the observed maximum)/(1 + 200). Clinical utility for audit was explored by decision-curve analysis [24]: women with a predicted probability of non-verification at or above a threshold were flagged and net benefit was compared with flagging all women, using out-of-fold elastic-net predictions averaged over repeats and a bootstrap 95% interval (1,000 resamples). Sensitivity analyses used histopathology-only verification, exclusion of residents outside the MUMS area, exclusion of death-certificate cases, models with age only, with year only, and without the missing-address indicator. As a secondary analysis, the 14-variable PHC model was fitted in the 74 linked women with the same procedures; details are in Additional file 5. Analyses used Python (scikit-learn [25], statsmodels); code is in Additional file 3 and in a public repository [repository DOI to be inserted].

Results

Microscopic verification

Of the 283 women (mean age 51.4 years, SD 13.6; median 51, range 18–95), 191 (67.5%, 95% CI 61.8–72.7) had MV: 189 by histopathology and 2 by cytology. The 92 non-verified diagnoses were clinical only (n = 70), death certificate (n = 20) and clinical evaluation (n = 2). MV differed by period, age and place (Table 1, Fig. 2). It was lowest in 1396 (55.3%), peaked in 1398 (78.6%) and fell to 61.5% in 1399, without a monotonic trend (odds ratio per year 1.19, p = 0.142). MV declined with age from 83.9% below 40 years to 46.7% at 70 years or more. It was 66.5% in Mashhad but 40.5% (95% CI 26.3–56.5) in the other MUMS counties, 80.4% among residents outside the MUMS area, and 100% in all 16 women with unknown city; overall 90.9% of the 33 women with missing address information were verified. Death-certificate-only registrations numbered 20 (7.1%, 95% CI 4.6–10.7; 5, 3, 8 and 4 in 1396–1399). By half-year, MV ranged from 54.3% (1396 H2) to 86.8% (1398 H1) and was 40.0% (6 of 15 women) in the last half-year of 1399, which also had the fewest registrations. Among the 37 women in other MUMS counties, MV was 1 of 7 in Taybad, 5 of 8 in Quchan and 3 of 6 in Chenaran; the other eight counties had 3 or fewer women each, so this pooled category is heterogeneous and county-level estimates are unstable. Without Taybad, MV

in the other MUMS counties was 46.7% (14 of 30; 95% CI 30.2–63.9) and the adjusted odds ratio for these counties was 0.62 (0.24–1.56; 0.39, 0.17–0.90 in the same model with Taybad), so the county difference rests partly on one county.

Table 1
Microscopic verification of 283 registered cervical cancer cases by period, age, residence and region.

Characteristic	n	Verified, n	MV, % (95% CI)
Year (Persian)			
1396 (2017–18)	76	42	55.3 (44.1–65.9)
1397	71	51	71.8 (60.5–81.0)
1398	84	66	78.6 (68.7–86.0)
1399 (2020–21)	52	32	61.5 (48.0–73.5)
Age group, years			
< 40	56	47	83.9 (72.2–91.3)
40–49	77	51	66.2 (55.1–75.8)
50–59	77	52	67.5 (56.5–76.9)
60–69	43	27	62.8 (47.9–75.6)
≥ 70	30	14	46.7 (30.2–63.9)
Residence			
Mashhad	179	119	66.5 (59.3–73.0)
Other MUMS counties	37	15	40.5 (26.3–56.5)
Outside the MUMS area	51	41	80.4 (67.5–89.0)
Address unknown	16	16	100 (80.6–100)
Linked to PHC records	74	45	60.8 (49.4–71.1)
Not linked	209	146	69.9 (63.3–75.7)
All women	283	191	67.5 (61.8–72.7)

MV, microscopic verification (histopathology or cytology); CI, Wilson 95% confidence interval.

Linkage to primary health care files

PHC-linked women differed markedly from unlinked women (Table 2). Among residents of MUMS counties, the 70 linked women were 11 years younger than the 146 unlinked women (44.3 vs 55.3 years,

p < 0.001), more often lived outside Mashhad (34% vs 9%, p < 0.001) and in suburban or rural areas (p = 0.001), but their MV was similar (58.6% vs 63.7%, p = 0.55). Thus linkage selected younger women and peripheral residents, which limits the generalizability of PHC-based analyses, but it was not associated with the outcome.

Table 2
Characteristics of PHC-linked and unlinked women.

Characteristic	Linked (n = 74)	Not linked (n = 209)	p
Age, years, mean (SD)	44.6 (9.5)	53.8 (14.0)	< 0.001
Year of report, 1396/97/98/99, n	28/14/14/18	48/57/70/34	0.008
Microscopically verified, %	60.8	69.9	0.20
Among residents of MUMS counties	(n = 70)	(n = 146)	
Age, years, mean (SD)	44.3 (9.7)	55.3 (13.8)	< 0.001
Residence in Mashhad / other county, n	46 / 24	133 / 13	< 0.001
Region urban / suburban / rural / missing, n	28 / 29 / 12 / 1	79 / 53 / 5 / 9	0.001
Microscopically verified, %	58.6	63.7	0.55

Mann–Whitney U test (age), chi-square (categorical) and Fisher's exact test (verification, among MUMS residents).

Registered incidence

Among the 216 women resident in MUMS counties with available denominators (about 9.83 million woman-years over four years), the registered incidence was 2.20 per 100,000 women (95% CI 1.91–2.51) and the age-standardized rate (Segi) was 2.26 (1.96–2.60). Rates declined from 2.91 per 100,000 in 1396 to 1.69 in 1399 (Table 3); the age-adjusted rate ratio per year was 0.85 (95% CI 0.76–0.96, p = 0.009); excluding 1399 it was 0.87 (0.73–1.05, p = 0.15), so the decline is not statistically clear in 1396–1398. Age-specific rates were close to zero below 30 years, rose to about 7–8 per 100,000 at 45–54 years and remained at 5–8 per 100,000 at older ages. The decline and the low absolute level cannot be interpreted as true incidence changes: completeness was not measured, 67 women without an assignable MUMS residence were excluded, the 1399 half-year counts suggest late notifications, and the pandemic period may have affected both diagnosis and registration.

Table 3
Registered cervical cancer incidence among women resident in MUMS counties, per 100,000 women.

Year	Cases	Woman-years	Crude rate (95% CI)	ASR (Segi)	ASR 95% CI
1396	67	2,389,220	2.80 (2.17–3.56)	2.91	2.24–3.76
1397	52	2,432,377	2.14 (1.60–2.80)	2.19	1.62–2.92
1398	55	2,490,289	2.21 (1.66–2.87)	2.33	1.74–3.07
1399	42	2,522,962	1.66 (1.20–2.25)	1.69	1.20–2.34
1396–1399	216	9,834,848	2.20 (1.91–2.51)	2.26	1.96–2.60

ASR, age-standardized rate using the Segi world standard population, with Fay–Feuer gamma confidence intervals; denominators: female population of Mashhad and 15 other MUMS counties by five-year age group, from health-system population reports.

Multivariable analysis

In the multivariable model (Table 4), older age was associated with lower odds of MV (odds ratio 0.68 per 10 years, 95% CI 0.55–0.84), residence in a MUMS county other than Mashhad with lower odds (0.38, 0.17–0.87), and year 1398 with higher odds than 1396 (2.82, 1.31–6.04). Residence outside the MUMS area (2.15, 0.87–5.31) and a missing address (2.65, 0.72–9.76) pointed towards higher odds with wide intervals, and suburban and rural region codes towards lower odds (0.60 and 0.46).

Table 4
Adjusted odds ratios for microscopic verification (logistic regression, n = 283).

Variable	OR	95% CI	p
Age, per 10 years	0.68	0.55–0.84	< 0.001
Year 1397 (vs 1396)	1.78	0.86–3.70	0.124
Year 1398	2.82	1.31–6.04	0.008
Year 1399	1.13	0.52–2.46	0.768
Other MUMS county (vs Mashhad)	0.38	0.17–0.87	0.022
Outside the MUMS area	2.15	0.87–5.31	0.099
Address missing	2.65	0.72–9.76	0.142
Suburban region (vs urban)	0.60	0.32–1.10	0.095
Rural region	0.46	0.17–1.25	0.128

Predictive performance

All four algorithms discriminated modestly and almost identically (Table 5). Cross-validated AUC was 0.693 (SD across repeats 0.013) for penalized logistic regression, 0.697 for elastic net, 0.687 for random forest and 0.686 for gradient boosting, against 0.491 for the prevalence-only reference. Out-of-bag bootstrap intervals were 0.614–0.771 (logistic), 0.611–0.768 (elastic net) and 0.595–0.763 (random forest). The best-of-four AUC (0.697 with three repeats) exceeded all 200 values of the permutation null (null mean 0.517, 95th percentile 0.581; $p \leq 0.005$, the smallest value attainable with 200 permutations). Without year indicators, cross-validated AUC was 0.667 (logistic), 0.675 (elastic net), 0.674 (random forest) and 0.668 (gradient boosting). With training on 1396–1397 and testing on 1398–1399 and no year columns, AUC was 0.713 (logistic), 0.703 (elastic net), 0.678 (random forest) and 0.582 (gradient boosting) (0.700, 0.688, 0.683 and 0.590 with year indicators), so discrimination was maintained in later years for the regularized and bagged models but not for boosting. Age alone (AUC 0.603–0.616) or year alone (0.588–0.591) was much less informative than the combination.

Table 5
Cross-validated and temporal performance for predicting microscopic verification (n = 283).

Model	AUC, mean	SD across repeats	PR-AUC	Out-of-bag AUC (95% interval)	Temporal AUC (no year)
Prevalence-only	0.491	0.000	0.671	–	0.500
Logistic regression, L2	0.693	0.013	0.820	0.690 (0.614–0.771)	0.713
Elastic net	0.697	0.012	0.827	0.685 (0.611–0.768)	0.703
Random forest	0.687	0.014	0.830	0.681 (0.595–0.763)	0.678
Gradient boosting	0.686	0.018	0.826	–	0.582

AUC, area under the receiver operating characteristic curve; PR-AUC, area under the precision–recall curve (uninformative baseline 0.675). Out-of-bag intervals were not computed for gradient boosting. Temporal: models without year columns, trained 1396–1397 (n = 147), tested 1398–1399 (n = 136).

Calibration and potential utility

Calibration-in-the-large was close to zero (0.01 and 0.00). The calibration slope was 0.90 (95% CI 0.61–1.27) for elastic net, indicating good calibration, and 1.94 (1.29–2.73) for penalized logistic regression, indicating over-shrunk (under-confident) predictions caused by strong regularization. Brier scores of 0.205 (logistic) and 0.200 (elastic net) improved on the prevalence-only value of 0.219. In quintiles of elastic-net predicted probability, observed MV rose from 53% to 90% (predicted 42% to 89%).

In the decision-curve analysis for flagging women likely not to be verified (prevalence 32.5%), elastic-net net benefit exceeded that of flagging everyone at thresholds of 0.3 (0.120, 95% interval 0.059–0.184, versus 0.036; 144 flagged, 67 truly non-verified) and 0.4 (0.047, − 0.006 to 0.099, versus − 0.125; 85

flagged), but the benefit was small at 0.2 (0.176, 0.116–0.241, versus 0.156; 211 flagged) and null at 0.5 (− 0.004). The intervals show that the utility is limited and uncertain.

Sensitivity analyses

Results were stable (Table 6). Using histopathology alone as verified (94 non-verified women) gave AUCs of 0.692 and 0.696, excluding residents outside the MUMS area 0.683 and 0.698, excluding death-certificate cases 0.693 and 0.701, and removing the missing-address indicator 0.692 and 0.694 (logistic and elastic net).

Table 6
Sensitivity analyses: cross-validated AUC (20 × 5-fold).

Analysis	Logistic L2	Elastic net
Primary (9 predictors, n = 283)	0.693	0.697
Histopathology-only outcome (cytology as non-verified)	0.692	0.696
Excluding residents outside the MUMS area (n = 232)	0.683	0.698
Excluding death-certificate cases (n = 263)	0.693	0.701
Without missing-address indicator	0.692	0.694
Age only	0.603	0.616
Year only	0.591	0.588

PHC-linked subset (secondary analysis)

In the 74 PHC-linked women (45 verified), adding 14 PHC-recorded and registry variables gave cross-validated AUCs of 0.645 (penalized logistic regression; out-of-bag 95% interval 0.446–0.807), 0.658 (elastic net), 0.603 (random forest) and 0.556 (gradient boosting); the best-of-four AUC was borderline against a selection-aware permutation null (p = 0.050), and calibration slopes were 0.80 (0.15–1.86) and 0.58 (0.14–1.25). No PHC variable remained significant after Holm adjustment. The timing of PHC data capture relative to diagnosis could not be verified (PHC-recorded age exceeded registry age by 3–8 years in 39 of 74 women). The PHC variables therefore added no reliable information beyond the registry variables in this small subset (Additional file 5).

Discussion

In this four-year cervical cancer registry from Mashhad, one third of diagnoses (32.5%) lacked microscopic verification, with a clear gradient: lower MV in 1396, in women aged 70 years or more and, less certainly, in counties other than Mashhad (40.5%, a difference that depended partly on one county). These patterns suggest that access to and use of pathology services, rather than characteristics of the tumour, may determine registry data quality (in a setting where only 39.4% of women aged 18–45 years took part in screening in 2021 [8], diagnosis often follows symptoms and depends on access to

pathology services), and they identify specific targets for audit and case-finding. The 100% MV among women with unknown address, together with a DCO proportion of 7.1%, probably reflects notifications from pathology laboratories that carry no residence information, which also shows how the registry's sources shape its data.

Registry-available variables predicted non-verification with modest discrimination (AUC about 0.69), good calibration for elastic net, and a gain in Brier score over the prevalence. Two features strengthen our confidence in this signal: it was maintained for the regularized and bagged models (but not for boosting) when models were trained on 1396–1397 and tested on 1398–1399, and it exceeded a permutation null that accounts for choosing the best of four algorithms ($p \leq 0.005$). However, four very different algorithms performed alike, as did simple age and year terms in regression. This pattern is typical when the predictors are few and low-dimensional, and it is consistent with evidence that machine learning rarely outperforms logistic regression in such settings [9, 10]. The algorithm therefore does not matter here: the information lies in a few registry variables, discrimination is modest (AUC about 0.67–0.70 depending on whether year is included), and the regression is the more interpretable summary. The missing-address indicator marks the notification source rather than a patient trait, and the year effect reflects registry processes, which limits transport of any model to another registry.

The PHC-linked cases, used in our companion case–control study, differed substantially from the rest of the registry (11 years younger and more often from peripheral counties), although their MV was similar. Findings from PHC-based analyses therefore cannot be generalized to all registered women, and in the 74 linked women PHC variables did not add reliable predictive information. We also found that PHC age exceeded registry age in more than half of linked women, so the timing of PHC data capture needs clarification before such variables are used as pre-diagnostic information.

Registered incidence (age-standardized 2.26 per 100,000 women) is of the same order as the 2.75 per 100,000 reported for Khorasan Razavi Province in 2016 [7], and our overall MV of 67.5% is below the 84.7% reported nationally for cervical cancer in 2016 [7] and similar to the 68.3% reported for all cancers in 2014 [6]; differences in standard population, catchment area, years and denominators preclude firm comparison. Other Iranian provincial registries report MV and DCO as province-wide or cancer-type averages (East Azerbaijan 2015–2016: MV 65.1%, DCO 19.7% [26]; Fars 2015–2018 [27]), so variation within a registry by period, age and county has, to our knowledge, rarely been described. The age-adjusted rate fell by about 15% per year (rate ratio 0.85; 0.87 and not significant when 1399 is excluded), but without a measure of completeness, such as a mortality-to-incidence ratio [M:I ratio to be inserted] or a capture–recapture estimate, and with data extracted [x] months after the end of 1399 and few registrations in the last half-year, we cannot distinguish a true decline from late or incomplete registration, the pandemic or denominator changes. The rates are offered with their denominators (Additional file 4) for later comparison and do not substitute for completeness assessment.

Strengths include the use of all registered cases, explicit control of leakage from fields created by the diagnostic process, out-of-bag intervals that include model fitting, temporal validation, selection-aware

permutation inference, calibration with intervals and decision-curve analysis, and shared code. Limitations are as follows. The outcome is determined partly by registry processes, so predictors are associations with data quality and not clinical causes; the variable set is small, and no individual clinical information such as symptoms, stage or access to services was available. Region code was missing in 31 women and address in 33, and these were handled with an indicator. The study covers one registry and four years, and external validation in another registry is needed. Registered incidence rates are sensitive to completeness, to the assignment of residence and to denominators taken from health-system reports. The pooled category of other MUMS counties is heterogeneous (for example Taybad 1 of 7 verified), the region coding and the linkage procedure rest on registry conventions that we describe but could not audit, and the 40.5% estimate is based on 37 women. The decision-curve intervals are wide. Model performance was not assessed within subgroups (for example by residence or age) because of small numbers, and the models were not externally validated. Finally, the PHC-linked analysis is small and its predictor timing is uncertain.

For registries, the main implication is practical: routine audit of the proportion of microscopically verified cases by county, age and period, with feedback to peripheral pathology and clinical services, is likely to be more useful than individual prediction. In MUMS counties outside Mashhad and in women aged 70 years or more, a verification target and follow-up of unverified cases would address the largest gaps. Future work should replicate the analysis with more years, add information on stage and sources of notification, and examine whether targeted audit improves MV.

Conclusions

About one third of registered cervical cancers in the MUMS registry were not microscopically verified, with lower verification in older women, possibly in counties other than Mashhad, and in the first year. Registry-available variables predicted non-verification with modest discrimination that was maintained over time for most algorithms, and algorithms of very different complexity performed alike. The results support targeted audit of peripheral counties and older women rather than individual prediction.

Abbreviations

ASR
age-standardized rate
AUC
area under the receiver operating characteristic curve
CI
confidence interval
MUMS
Mashhad University of Medical Sciences
MV
microscopic verification

OR
odds ratio
PHC
primary health care
PR-AUC
area under the precision–recall curve
STROBE
Strengthening the Reporting of Observational Studies in Epidemiology
TRIPOD + AI
Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis,extended for artificial intelligence.

Declarations

Ethics approval and consent to participate:

The study was approved by the Ethics Committee of Mashhad University of Medical Sciences (approval code: [to be inserted]). The registry file originally contained personal identifiers, which were removed after the duplicate check and before analysis; no identifiers appear in the analysis files or in this manuscript. [Confirm against the approval letter whether the committee waived individual informed consent.] The study was conducted in accordance with the Declaration of Helsinki.

Consent for publication:

Not applicable.

Competing interests:

The authors declare that they have no competing interests.

Additional files

Additional file 1: STROBE checklist (.docx). Additional file 2: TRIPOD + AI checklist (.docx). Additional file 3: Python analysis code (ml_full.py, ml_full2.py and ml_full3.py; ml_revised.py for the PHC-linked subset). Additional file 4: Female population denominators by age group, county and year, 1392–1403 (.xlsx). Additional file 5: PHC-linked subset analysis (summarized in the Results; code ml_revised.py).

Funding:

This research was supported by Mashhad University of Medical Sciences, Mashhad, Iran (Grant No. 4030952).

Author Contribution

VV conceived and designed the study, performed the formal analysis, supervised the work, and wrote and revised the manuscript. FK contributed to the conception and design, investigation and formal analysis, and wrote the original draft. MMB and AS contributed to the methodology and interpretation of results, supervised the work, and reviewed and edited the manuscript. All authors read and approved the final manuscript.

Acknowledgements:

The authors thank the staff of the MUMS cancer registry and of the PHC Family Physician Information System.

Availability of data and materials:

The de-identified analysis file is available from the corresponding author on reasonable request and subject to approval by the MUMS cancer registry. The analysis code is provided as Additional file 3 and in a public repository [DOI to be inserted] and the population denominators as Additional file 4.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71:209–49.
2. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74:229–63.
3. Bray F, Parkin DM. Evaluation of data quality in the cancer registry: principles and methods. Part I: comparability, validity and timeliness. Eur J Cancer. 2009;45:747–55.
4. Parkin DM, Bray F. Evaluation of data quality in the cancer registry: principles and methods. Part II: completeness. Eur J Cancer. 2009;45:756–64.
5. Mohammadi G, Akbari ME, Mehrabi Y, Ghanbari Motlagh A. Quality assessment of the National Cancer Registry in Iran: completeness and validity. Int J Cancer Manag. 2016;9(6):e8479.
6. Roshandel G, Partovipour E, Ghanbari Motlagh A, Hasanpour Heidari S, Salavati F, Mohammadi G, et al. Cancer incidence in Iran in 2014: results of the Iranian National Population-based Cancer

Registry. Cancer Epidemiol. 2019;61:50–8.

7. Anjam Majoumerd A, Hosseini S, Hosseini SMH, Roshandel G, Rahmanian V, Hazar N. Epidemiology of cervical cancer in Iran in 2016: a nationwide study of incidence and regional variation. Cancer Rep (Hoboken). 2024;7(2):e1973.
8. Haddadi M, Golestani A, Farzi Y, Rezaei N, Ghasemi E, Azadnajafabad S, et al. Iran is falling behind WHO cervical cancer elimination targets: HPV vaccination coverage and cervical cancer screening participation in 2021. PLoS ONE. 2026;21(2):e0341888.
9. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12–22.
10. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. PLoS ONE. 2019;14:e0224365.
11. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet. 2007;370:1453–7.
12. 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.
13. Khabbaz F, Vakili V, Mousavi Bazaz M, Soltanifar A, Mousa Farkhani E. Factors associated with cervical cancer among women in northeastern Iran: a registry-linked case–control study. BMC Womens Health. Accepted [citation details to be updated at publication].
14. Segi M. Cancer mortality for selected sites in 24 countries (1950–57). Sendai: Department of Public Health, Tohoku University School of Medicine; 1960.
15. Fay MP, Feuer EJ. Confidence intervals for directly standardized rates: a method based on the gamma distribution. Stat Med. 1997;16:791–801.
16. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170:51–8.
17. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J. 2014;35:1925–31.
18. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B. 2005;67:301–20.
19. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
20. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230.
21. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol. 2016;74:167–76.

22. Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med. 2000;19:453–73.
23. Ojala M, Garriga GC. Permutation tests for studying classifier performance. J Mach Learn Res. 2010;11:1833–63.
24. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Mak. 2006;26:565–74.
25. 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.
26. Somi MH, Dolatkhah R, Sepahi S, Belalzadeh M, Sharbafi J, Abdollahi L, et al. Cancer incidence in the East Azerbaijan province of Iran in 2015–2016: results of a population-based cancer registry. BMC Public Health. 2018;18:1266.
27. Rezaianzadeh A, Ghoddusi Johari M, Niazkar HR, Khosravizadegan Z, Monabati A, Shiraziyeganeh B. Cancer incidence in Southern Iran, 2015–2018: a population based study on cancer registry profile of Fars province. Health Sci Rep. 2023;6(6):e1333.

Figures

Registry records, 1396–1399	Women after removing duplicate	Residence	Basis of diagnosis
n = 284	n = 283 (1 duplicate record removed)	Mashhad 179; other MUMS counties 37; outside MUMS area 51; address unknown 16	Microscopic verification 191 (histopathology 189, cytology 2); non-verified 92 (clinical only 70, death certificate 20, clinical evaluation 2)

Figure 1

Flow of records and women.

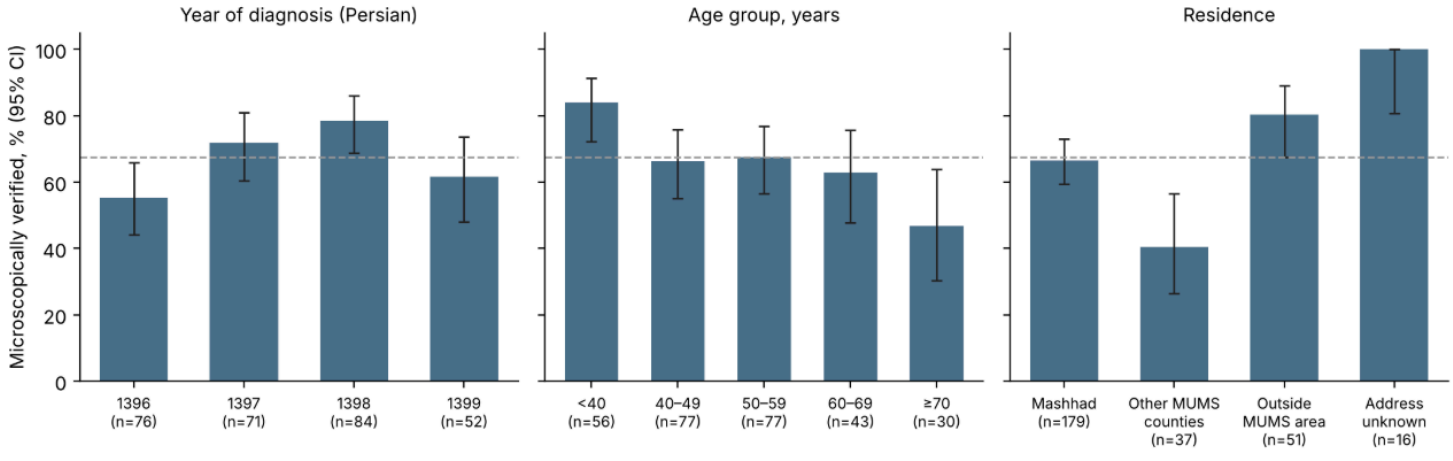


Figure 2

Microscopic verification by year, age group and residence (dashed line: overall 67.5%).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1STROBEchecklist.docx
- Additionalfile2TRIPODAIchecklist.docx