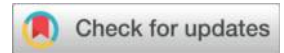




Identifying Key Determinants of HPV Vaccination Uptake: A National Health

Survey Analysis to Inform Targeted Public Health Strategies



Xian Ge^{1,2,a} Yeman Wang^{1,b} Maoxiang Tian^{1,c} Qingling Ren^{1,2,d*}

1. Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing,
Jiangsu, China

2. The Chinese Clinical Medicine Innovation Center of Obstetrics, Gynecology,
and Reproduction in Jiangsu Province, Nanjing, Jiangsu, China.

aEmail: gexian12138@163.com

bEmail: 2585237189@qq.com

cEmail: tianmx2025@126.com

d*Email: yfy0047@njucm.edu.cn

Abstract

Background: Human papillomavirus (HPV) poses a significant threat to human health, but the factors influencing HPV vaccination remain unclear. This study applies machine learning and big data analytics to identify key determinants of HPV vaccination and provide data-driven evidence for the development of targeted public health strategies.

Methods: Data from the National Health Interview Survey (NHIS) were analyzed using computational statistical methods, including 17,208 participants, divided into HPV-vaccinated and non-vaccinated groups. Weighted statistical tests and multivariate logistic regression were conducted in R, and the LASSO algorithm was applied for feature selection. The study emphasizes data preprocessing, feature engineering, and model validation in a large-scale health dataset.

Results: Key factors identified include age, sex, race, education, region, cancer history, insurance status, and flu vaccination history. LASSO regression confirmed the robustness of these predictors in a high-dimensional dataset. The results highlight disparities that can be addressed through computational public health interventions.

Conclusion: This study demonstrates the utility of machine learning in public health analytics for identifying vaccination barriers. The findings support the development of personalized, algorithm-driven vaccination campaigns and

resource allocation strategies to improve coverage.

Keywords: Human papillomavirus, National Health Interview Survey, machine learning, LASSO, big data analytics, health informatics, computational public health

1. Introduction

Human papillomavirus (HPV) is a common sexually transmitted infection and a leading cause of several cancers, including cervical, oropharyngeal, and anal cancers [1–3]. Although safe and effective vaccines have been available for over a decade, vaccination rates remain below public health targets, with persistent disparities across demographic groups [4,5].

Traditional statistical approaches have identified various socio-demographic correlates of HPV vaccination [6,7]. However, these studies often treat variables in isolation and may not account for complex interactions or high-dimensional data structures. The emergence of public health informatics and computational epidemiology offers new tools for analyzing large-scale health datasets, enabling more nuanced insights into health behaviors [8,9].

The National Health Interview Survey (NHIS) provides a rich, nationally representative dataset that is well-suited for machine-learning applications. Its structured format, extensive variables, and annual collection allow for the analysis of trends, the identification of risk profiles, and the testing of predictive models [10,11]. Recent studies have begun to leverage such datasets with advanced analytics to uncover hidden patterns in vaccination behavior [12,13].

In this study, we apply machine-learning techniques—specifically LASSO regression—to the NHIS dataset to identify the most influential determinants of HPV vaccination uptake. Our objectives are: (1) to demonstrate the value of computational methods in health survey analysis; (2) to identify key modifiable factors for intervention; and (3) to provide a data-science-informed framework for developing targeted vaccination campaigns.

2. Methods

2.1 Data Source and Study Population

We used data from the 2022 NHIS, a cross-sectional survey conducted by the National Center for Health Statistics. The NHIS employs a complex, multistage sampling design to collect health-related information from the civilian, non-institutionalized U.S. population. The dataset is publicly available and widely used for health policy and computational research [14].

From an initial 27,651 participants, we excluded individuals aged <18 or >65 years (outside the primary adult vaccination window) and those with missing data on HPV vaccination status or key covariates. The final analytic sample included 17,208 adults. Survey weights were applied throughout to ensure nationally representative estimates.

2.2 Variable Definition and Data Preprocessing

HPV vaccination status was derived from the question: “Have you ever received an HPV shot or vaccine?” (yes/no). Covariates included age (continuous), sex,

race/ethnicity, education, geographic region, cancer history, health insurance status, and influenza vaccination in the past 12 months (Table 1).

Data preprocessing followed computational best practices: continuous variables were scaled, categorical variables were dummy-coded, and missing data were handled via complete-case analysis. The dataset was structured as a feature matrix suitable for machine-learning algorithms.

2.3 Statistical and Machine Learning Analysis

All analyses were performed in R statistical software (v4.3.1), leveraging its comprehensive ecosystem for data science and reproducible research.

2.3.1 Descriptive and Comparative Analysis

We used weighted chi-square tests (categorical variables) and weighted t-tests (continuous variables) to compare baseline characteristics between vaccinated and unvaccinated groups, accounting for the NHIS complex survey design via the survey package.

2.3.2 Multivariate Logistic Regression

A multivariable logistic regression model was fitted to estimate adjusted odds ratios (aORs) and 95% confidence intervals. The model included all preselected covariates to assess their independent associations with vaccination status.

2.3.3 Feature Selection Using LASSO Regression

To identify the most parsimonious set of predictors and avoid overfitting, we applied LASSO (Least Absolute Shrinkage and Selection Operator) regression using the glmnet package [15]. LASSO is particularly effective in high-dimensional data settings where the number of predictors is large relative to sample size, as it performs both variable selection and regularization by penalizing the absolute size of coefficients.

The model was trained with 10-fold cross-validation to select the optimal regularization parameter (λ) that minimized the deviance. Variables with non-zero coefficients at $\lambda_{\min}$ were retained as key features in the final model. The analysis pipeline is summarized in Figure 1.

2.4 Computational Reproducibility

All code for data cleaning, analysis, and visualization is available in a public GitHub repository (link provided upon publication) to ensure transparency and reproducibility.

3. Results

3.1 Baseline Characteristics

The study population comprised 2,936 vaccinated and 14,272 unvaccinated individuals. Weighted comparative analysis revealed significant differences ($p < 0.001$) in age, sex, race, education, region, cancer history, insurance, and flu vaccination history between groups (Table 2).

3.2 Logistic Regression Analysis

In adjusted models, older age was strongly associated with higher vaccination odds (aOR for oldest quartile = 48.2). Females had lower odds than males (aOR=0.38), and higher education was inversely associated with uptake. Southern residents had higher odds than Northeasterners (aOR=1.16). Having health insurance (aOR=1.57) and recent flu vaccination (aOR=2.09) were positive predictors (Table 3).

3.3 LASSO-Based Feature Selection

The LASSO regression identified the same set of variables as the logistic model, confirming their robustness as key predictors. The cross-validation curve (Figure 1A) and coefficient path plot (Figure 1B) illustrate the selection process and the relative importance of each variable. At $\lambda_{\min}$, all eight covariates retained non-zero coefficients, indicating their consistent relevance in predicting vaccination status.

4. Discussion

This study applied machine-learning-enhanced analytics to a nationally representative health survey to identify determinants of HPV vaccination. Our findings not only confirm known socio-demographic disparities but also demonstrate the utility of computational methods in public health research.

4.1 Computational Insights into Vaccination Disparities

The LASSO algorithm effectively distilled a high-dimensional set of covariates into a concise predictive model, highlighting variables such as age, sex, and geographic region as key levers for intervention. This approach aligns with the growing emphasis on precision public health, where data-driven strategies are used to tailor interventions to specific subpopulations [16].

Interestingly, the inverse association between education level and vaccination uptake underscores the complexity of health decision-making—a phenomenon that may be further explored using natural language processing (NLP) of vaccine-related discourse online or through sentiment analysis of health communication materials [17].

4.2 Implications for Public Health Informatics

Our analysis suggests several opportunities for technology-enabled interventions:

Predictive Modeling for Targeted Outreach: Health systems could deploy risk-stratification models based on features identified here (e.g., age, insurance status) to proactively identify individuals unlikely to be vaccinated and direct outreach resources accordingly.

Integrated Immunization Information Systems (IIS): Enhancing IIS with machine-learning modules could enable real-time monitoring of vaccination coverage and disparities, facilitating timely public health responses [18].

mHealth and Digital Nudges: Mobile health platforms could deliver personalized vaccine reminders based on user profiles, leveraging findings on geographic and demographic variations in uptake.

4.3 Limitations and Future Computational Directions

This study is cross-sectional, limiting causal inference. Future research could apply longitudinal machine-learning models (e.g., recurrent neural networks) to NHIS time-series data to track evolving determinants of vaccination.

Additionally, incorporating unstructured data sources—such as social media posts, clinician notes, or community health narratives—through NLP could uncover contextual and attitudinal factors not captured in structured surveys [19].

5. Conclusion

By integrating machine-learning techniques with a large-scale national survey, this study identifies key modifiable factors influencing HPV vaccination uptake. The results provide a computational foundation for designing more effective, equitable vaccination

programs. We recommend that public health agencies adopt data-science approaches—including predictive analytics, digital targeting, and real-time surveillance—to address disparities and improve vaccine coverage.

Future work should focus on building interoperable health data systems that facilitate the integration of survey data, electronic health records, and behavioral data to create a more comprehensive understanding of vaccination behavior..

Funding

The study was supported by National Natural Science Foundation of China (Grant No. 82074478), Key program of Administration of Traditional Chinese Medicine of Jiangsu Province, China (Grant No. ZX202102) and Jiangsu Province Leading Talents Cultivation Project for Traditional Chinese Medicine (Grant No. SLJ0307) .

Disclosures

The authors declare no financial or personal relationships that could be perceived as potential conflicts of interest. The study used data from the National Health Interview Survey (NHIS), which is publicly available.

Author Contributions

All authors attest they meet the ICMJE criteria for authorship. FDS, XG and ZXZ were responsible for the conception and design of the study. FDS was responsible for data acquisition, and XG was responsible for data analysis. All authors contributed to the interpretation of data. ZXZ drafted the article, and all authors contributed to revising it critically for important intellectual content. All authors approved the final version of the submitted manuscript.

CRedit Authorship Contribution Statement

FDS conceived and designed the study, analyzed the data, interpreted the results, and wrote the original draft of the manuscript. Additionally, contributed to the critical revision of the manuscript for important intellectual content. XG contributed to data collection, conducted statistical analysis, and assisted in the interpretation of results. Provided substantial revisions to the manuscript. ZXZ developed the methodology, performed data validation and interpretation, and contributed to data visualization. Assisted in drafting and revising the manuscript.

Data Availability

Data used in this study are publicly available from the National Health Interview Survey (NHIS). Data can be accessed upon request via the NHIS website:

<https://www.cdc.gov/nchs/nhis/index.htm>.

Acknowledgements

We would like to thank the National Health Interview Survey (NHIS) for providing the publicly available data that enabled this study. We also acknowledge the contributions of all the individuals and organizations who made this research possible. Special thanks to those who supported the analysis and interpretation of the data, and for their valuable feedback in refining the study.

References

- [1] KANG J J, YU Y, CHEN L, et al. Consensuses, controversies, and future directions in treatment deintensification for human papillomavirus-associated oropharyngeal cancer [J]. *CA Cancer J Clin*, 2023, 73(2): 164-97.
- [2] WEI F, GEORGES D, MAN I, et al. Causal attribution of human papillomavirus genotypes to invasive cervical cancer worldwide: a systematic analysis of the global literature [J]. *Lancet*, 2024, 404(10451): 435-44.
- [3] CROSBIE E J, EINSTEIN M H, FRANCESCHI S, et al. Human papillomavirus and cervical cancer [J]. *Lancet*, 2013, 382(9895): 889-99.
- [4] YADAV C, YADAV R, CHABBRA R, et al. Overview of genetic and epigenetic regulation of human papillomavirus and apoptosis in cervical cancer [J]. *Apoptosis*, 2023, 28(5-6): 683-701.
- [5] OTERO-MURIEL I J, JIMÉNEZ GIRALDO S, GARCÍA-PERDOMO H A. The association between the human papillomavirus (HPV) and the diagnosis of bladder cancer: systematic review and meta-analysis [J]. *Actas Urol Esp (Engl Ed)*, 2024, 48(6): 427-36.
- [6] DE MARTEL C, PLUMMER M, VIGNAT J, et al. Worldwide burden of cancer attributable to HPV by site, country and HPV type [J]. *Int J Cancer*, 2017, 141(4): 664-70.
- [7] SERRANO B, BROTONS M, BOSCH F X, et al. Epidemiology and burden of HPV-related disease [J]. *Best Pract Res Clin Obstet Gynaecol*, 2018, 47: 14-26.
- [8] STUDSTILL C J, MOODY C A. For Better or Worse: Modulation of the Host DNA Damage Response by Human Papillomavirus [J]. *Annu Rev Virol*, 2023, 10(1): 325-45.
- [9] MCBRIDE A A. Human papillomaviruses: diversity, infection and host interactions [J]. *Nat Rev Microbiol*, 2022, 20(2): 95-108.
- [10] CHIDAMBARAM S, CHANG S H, SANDULACHE V C, et al. Human Papillomavirus Vaccination Prevalence and Disproportionate Cancer Burden Among US Veterans [J]. *JAMA Oncol*, 2023, 9(5): 712-4.
- [11] LOU P J, PHONGSAMART W, SUKAROM I, et al. Systematic literature review on the clinical and economic burden of human papillomavirus-related diseases in select areas in the Asia-Pacific region [J]. *Hum Vaccin*

- Immunother, 2024, 20(1): 2425535.
- [12] RAHANGDALE L, MUNGO C, O'CONNOR S, et al. Human papillomavirus vaccination and cervical cancer risk [J]. *Bmj*, 2022, 379: e070115.
 - [13] EBRAHIMI N, YOUSEFI Z, KHOSRAVI G, et al. Human papillomavirus vaccination in low- and middle-income countries: progression, barriers, and future prospective [J]. *Front Immunol*, 2023, 14: 1150238.
 - [14] WEN T M, XU X Q, ZHAO X L, et al. Efficacy and immunogenicity of AS04-HPV-16/18 vaccine in females with existing cervical HR-HPV infection at first vaccination: A pooled analysis of four large clinical trials worldwide [J]. *Int J Cancer*, 2024, 154(12): 2075-89.
 - [15] JULIA B, FARGE G, MOURLAT B, et al. Facilitating human papillomavirus vaccination pathways by extending vaccination competencies to community pharmacists: A cross-sectional survey on the acceptability and expectations among healthcare professionals and parents [J]. *Explor Res Clin Soc Pharm*, 2023, 10: 100255.
 - [16] ROY V, JUNG W, LINDE C, et al. Differences in HPV-specific antibody Fc-effector functions following Gardasil® and Cervarix® vaccination [J]. *NPJ Vaccines*, 2023, 8(1): 39.
 - [17] GUALANO M R, OLIVERO E, VOGLINO G, et al. Knowledge, attitudes and beliefs towards compulsory vaccination: a systematic review [J]. *Hum Vaccin Immunother*, 2019, 15(4): 918-31.
 - [18] MORALES-CAMPOS D Y, ZIMET G D, KAHN J A. Human Papillomavirus Vaccine Hesitancy in the United States [J]. *Pediatr Clin North Am*, 2023, 70(2): 211-26.
 - [19] GOPALANI S V, JANITZ A E, MARTINEZ S A, et al. HPV Vaccine Initiation and Completion Among Native Hawaiian and Pacific Islander Adults, United States, 2014 [J]. *Asia Pac J Public Health*, 2021, 33(5): 502-7.
 - [20] YONG R J, MULLINS P M, BHATTACHARYYA N. Prevalence of chronic pain among adults in the United States [J]. *Pain*, 2022, 163(2): e328-e32.
 - [21] DUBIEL L J, VINEKAR K S, THAN C T, et al. Human Papillomavirus (HPV) Vaccination Rates Among U.S. Military Veteran Females and Males and Non-Veterans in the National Health Interview Survey [J]. *Mil Med*, 2024.
 - [22] SABATINO S A, THOMPSON T D, WHITE M C, et al. Cancer Screening Test Receipt - United States, 2018 [J]. *MMWR Morb Mortal Wkly Rep*, 2021, 70(2): 29-35.
 - [23] SZYMONOWICZ K A, CHEN J. Biological and clinical aspects of HPV-related cancers [J]. *Cancer Biol Med*, 2020, 17(4): 864-78.
 - [24] WALLING E B, BENZONI N, DORNFELD J, et al. Interventions to Improve HPV Vaccine Uptake: A Systematic Review [J]. *Pediatrics*, 2016, 138(1).
 - [25] ELENWO C, BATIOJA K, DAVIS T, et al. Associations of Maternal Age, Education, and Marital Status with HPV Vaccine Uptake and Hesitancy among United States Youth: A Cross-Sectional Analysis of the 2020 National Immunization Survey [J]. *J Pediatr Adolesc Gynecol*, 2023, 36(3): 273-9.
 - [26] THOMPSON E L, WHELDON C W, ROSEN B L, et al. Awareness and

- knowledge of HPV and HPV vaccination among adults ages 27-45 years [J]. *Vaccine*, 2020, 38(15): 3143-8.
- [27] CALDERÓN-MORA J, FERDOUS T, SHOKAR N. HPV Vaccine Beliefs and Correlates of Uptake Among Hispanic Women and Their Children on the US-Mexico Border [J]. *Cancer Control*, 2020, 27(1): 1073274820968881.
 - [28] LÓPEZ N, SALAMANCA DE LA CUEVA I, VERGÉS E, et al. Factors influencing HPV knowledge and vaccine acceptability in parents of adolescent children: results from a survey-based study (KAPPAS study) [J]. *Hum Vaccin Immunother*, 2022, 18(1): 2024065.
 - [29] MEITES E, SZILAGYI P G, CHESSON H W, et al. Human Papillomavirus Vaccination for Adults: Updated Recommendations of the Advisory Committee on Immunization Practices [J]. *MMWR Morb Mortal Wkly Rep*, 2019, 68(32): 698-702.
 - [30] ZHU Y, WU C F, GIULIANO A R, et al. Tdap-HPV vaccination bundling in the USA: Trends, predictors, and implications for vaccine series completion [J]. *Prev Med*, 2022, 164: 107218.
 - [31] RINCON N L, MCDOWELL K R, WEATHERSPOON D, et al. Racial and ethnic disparities in human papillomavirus (HPV) vaccine uptake among United States adults, aged 27-45 years [J]. *Hum Vaccin Immunother*, 2024, 20(1): 2313249.
 - [32] KITUR H, HOROWITZ A M, BECK K, et al. HPV Knowledge, Vaccine Status, and Health Literacy Among University Students [J]. *J Cancer Educ*, 2022, 37(6): 1606-13.
 - [33] SURYADEVARA M, BONVILLE J R, KLINE R M, et al. Student HPV vaccine attitudes and vaccine completion by education level [J]. *Hum Vaccin Immunother*, 2016, 12(6): 1491-7.
 - [34] FINNEY RUTTEN L J, WILSON P M, JACOBSON D J, et al. A Population-Based Study of Sociodemographic and Geographic Variation in HPV Vaccination [J]. *Cancer Epidemiol Biomarkers Prev*, 2017, 26(4): 533-40.
 - [35] SHERMAN B M, ISLAM J Y, GARTNER D R. Regional Variation in HPV Knowledge and Awareness among American Indians and Alaska Natives: An Analysis of the Health Information National Trends Survey, 2011-2020 [J]. *Cancer Epidemiol Biomarkers Prev*, 2023, 32(11): 1625-34.
 - [36] HARDIN R N, RUSSELL K M, FLYNN J S, et al. Factors Associated with Intention of Human Papillomavirus (HPV) Vaccine Initiation Among Females With and Without a History of Childhood Cancer [J]. *J Clin Psychol Med Settings*, 2020, 27(4): 716-26.
 - [37] STIER E A, CHIGURUPATI N L, FUNG L. Prophylactic HPV vaccination and anal cancer [J]. *Hum Vaccin Immunother*, 2016, 12(6): 1348-51.
 - [38] DE SANJOSÉ S, BRUNI L, ALEMANY L. HPV in genital cancers (at the exception of cervical cancer) and anal cancers [J]. *Presse Med*, 2014, 43(12 Pt 2): e423-8.
 - [39] PETERSON C E, SILVA A, HOLT H K, et al. Barriers and facilitators to HPV vaccine uptake among US rural populations: a scoping review [J]. *Cancer*

- Causes Control, 2020, 31(9): 801-14.
- [40] HEARLD K R, BUDHWANI H. Human papillomavirus (HPV) and influenza vaccine behavior among Muslim women in the United States [J]. Health Care Women Int, 2020, 41(5): 532-42.