Comparison and selection of statistical methods in data analysis of HPV vaccine clinical trials

Yiwen Duan

Bioengineering, Tianjin University of Science and Technology, Tianjin, China

Abstract. Human papillomavirus (HPV) is a global health problem that cannot be ignored, and it is strongly associated with many types of cancer. The analysis of HPV vaccine clinical trial data is critical to accurately assess key indicators of vaccine efficacy, safety, and immune persistence. Statistical methods are indispensable tools in this field of research. In this paper, we systematically review the commonly used statistical methods in the analysis of HPV vaccine clinical trial data, including descriptive statistics, hypothesis testing, survival analysis, regression analysis, etc., and discuss the characteristics of each method, its applicability scenarios, and its application in HPV vaccine research. Through comparative analysis, this paper summarizes the basis for selecting different statistical methods, which provides a reference for researchers to choose reasonable statistical methods when facing complex data types and diverse research objectives. This, in turn, facilitates the research and development, scientific evaluation, and wide-spread dissemination of HPV vaccines, thereby advancing global public health towards a new stage of development.

Keywords: Vaccine efficacy, security, immune persistence, Human papillomavirus (HPV)

1. Introduction

Human papillomavirus (HPV) is a global public health problem that is strongly associated with many types of cancer, including cervical cancer. Studies have shown that HPV infection is the leading cause of cervical cancer, which is responsible for about 90% of all cervical cancers (Chand et al. 2025; Ferlay et al. 2021; Gore et al. 2025). Therefore, prevention of HPV infection has become a key strategy to reduce the incidence of cervical cancer. The introduction of HPV vaccines has brought hope for the prevention and control of HPV-related diseases. Since the first vaccine became available in 2006, some countries have included it in their immunization programs, reducing the rate of infection and morbidity of HPV (Liu et al. 2025; Sung et al. 2021).

The analysis of clinical trial data is of great significance to public health. Analysis of adverse reactions and antibody levels in vaccinated persons allows assessment of vaccine safety and immunogenicity; comparison of disease incidence between vaccinated and control groups al-lows determination of vaccine protective efficacy; and study of differences in effectiveness in different populations helps to develop personalized vaccination strategies(Sierra et al. 2025).Statistical methods are crucial, and commonly used descriptive statistics, hypothesis testing, survival analysis, and regression analysis each have their scope of application and preconditioning assumptions, and the correct choice of method is critical to drawing accurate conclusions. The analysis of HPV clinical trial data is key to evaluating preventive and therapeutic vaccines, and through this analysis, the immunogenicity, efficacy, and safety of preventive vaccines can be clarified. Currently, HPV prophylactic vaccines are effective in preventing infection and disease but have limitations; therapeutic vaccines have progressed but the relationship between efficacy and immune response is unclear (Wang et al. 2020).

This paper compares and selects statistical methods in HPV vaccine clinical trial data analysis. It introduces the process and data characteristics of HPV vaccine clinical trials, describes the characteristics and application scenarios of commonly used statistical methods, and discusses selecting appropriate strategies based on research objectives and data characteristics. The aim is to help researchers apply statistical methods scientifically, ensure accurate data analysis, promote the development of HPV vaccine research, provide statistical support for vaccine re-search and development, and promote the progress of global public health.

2. Overview of HPV Vaccine Clinical Trials

Clinical trials of HPV vaccines are designed to assess the effectiveness as well as the safety of the vaccine in preventing HPV infection and the development of related diseases. Trials usually involve subjects divided into a test group (vaccinated) and a control group (vaccinated with a placebo). Data collected include basic information about the subjects (e.g., age, gender, etc.), immune response indicators after vaccination, disease occurrence, and adverse reactions. Its development process includes several clinical trial phases, and these data have different characteristics and types, with different experimental purposes and statistical methods for each phase.

2.1. Phase I clinical trial

The first stage is the starting point of vaccine development, also known as the safety assessment stage. This stage aims to assess the human body's response and tolerance to the vaccine, explore safe and effective dosages, and propose reasonable immunization procedures and protocols. The subjects' immune response and physiological indicators, such as body temperature and blood tests, are also monitored to assess the vaccine's safety in the human body and the initial signs of immunogenicity, such as whether the body can be stimulated to produce antibodies. This is usually done in a few healthy volunteers, using descriptive statistics and a dose-escalation design to determine the safe dose range.

2.2. Phase II clinical trials

Phase II clinical trials are conducted based on Phase I trials, and this trial further evaluates the vaccine's safety and determines the vaccine's immunization procedures and dose ranges in the target population. The trial focuses on observing the immune response of the vaccine, including the strength and durability of the immune response induced by different doses and vaccination procedures, such as testing the level of antibody production and cellular immune response. In addition, more safety data will be collected. Through chi-square tests, researchers can com-pare the incidence of adverse reactions in different populations and assess the differences be-tween vaccines in different populations (Xing et al. 2022). The accumulation of these data not only lays the foundation for subsequent experiments, but also helps us to gain a deeper understanding of the actual effects of vaccines in specific populations, thus providing a solid scientific basis for vaccine development and promotion.

2.3. Phase III clinical trial

Phase III clinical trials are the most critical phase of testing before a vaccine is marketed and are conducted in larger target populations, usually involving thousands or even tens of thou-sands of people. Its main objective is to comprehensively evaluate the efficacy and safety of a vaccine and determine its protective efficacy under a wide range of conditions of use. For ex-ample, a randomized controlled trial (RCT) design allows researchers to compare the difference in morbidity rates between a vaccinated and an unvaccinated population (often using a placebo as a control) to assess the vaccine's protective efficacy.

At this stage, statistical methods such as survival analysis are often used to assess the effectiveness of vaccines. Survival analysis can help researchers calculate key indicators such as time to onset of disease, morbidity, and duration of vaccine protection in vaccinated populations. In addition, statistical indicators such as risk ratios can quantify the effectiveness of vac-cines in preventing disease occurrence.

The Phase III clinical trial will also provide in-depth monitoring of the vaccine's adverse effects (Xing et al. 2022), including rare, long-term adverse reactions, to assess the risk-benefit ratio of vaccine use in large populations and to provide a sound scientific basis for vaccine registration applications and marketing.

3. Methods Of Statistical Techniques Commonlt Used In Clinical Analysis

3.1. Descriptive statistics

Antibody titer/concentration evaluation index data are mostly approximately lognormal distribution. The group-level summary generally uses geometric mean. The intergroup comparison statistics for the test group and control group geometric mean ratio, etc., through the initial organization of the data to show the role of centralized and discrete trends, such as the use of the mean and standard deviation describes the level of antibody titer after vaccination. It is suitable for analyzing quantitative data such as age, height, weight, drug dosage, laboratory results, etc., and can discover the pattern of disease occurrence and drug use.

3.2. Hypothesis testing

Hypothesis testing is based on sample data to make assumptions about the overall parameters or distribution, and through specific statistical methods to determine whether the hypothesis is valid. Commonly used hypothesis testing methods include t-test, analysis of variance (ANO-VA), chi-square test and so on. For example, the t-test is used to compare whether the means of two groups of samples are significantly different, and the chi-square test is used to test whether there is an association between two or more categorical variables (Chen et al. 2025).

3.3. Survival analysis

Survival analysis is a statistical method that examines the relationship between a dependent variable (e.g., disease occurrence, death, etc.) and multiple independent variables (e.g., age, treatment, etc.) in follow-up data. It considers not only whether the event occurred, but also when it occurred. Commonly used methods include the Kaplan - Meier, Cox proportional hazard model, etc.

3.4. Regression analysis

Regression analysis is a statistical method to study the quantitative dependence between one or more independent variables and the dependent variable. Commonly used are linear regression, logistic regression and so on. Linear regression analyzes the relationship between the dependent and independent variables when the dependent variable is a continuous variable; Logistic regression is used when the dependent variable is a categorical variable (e.g., whether or not a disease occurs, etc.)

4. Strategies For Selecting Statistical Tools In Data Analysis Of Hpv Vaccine Clinical Trials

In HPV vaccine clinical trials, data are complex and diverse, and a rational choice of statistical methods is crucial. Different statistical methods apply to different data types and research purposes, directly affecting the accuracy and reliability of research results.

4.1. Selection of statistical tools by type of data

The type, distribution, sample size, and other data characteristics must be fully considered when choosing statistical methods. For quantitative data (e.g., antibody titer, intensity of immune response, duration of protection, etc.), if the data meet the normal distribution and the variance is homogeneous, t-test analysis can be used to compare the difference in means be-tween different groups (e.g., antibody levels between vaccine and placebo groups); ANOVA is used to control confounding factors (e.g., the effect of age on immune response). If the data did not satisfy the assumptions of parametric tests, nonparametric methods such as the Mann-Whitney test (two independent samples) or the Kruskal-Wallis test (multiple independent samples) were used. Linear regression analyzes the dose-response relationship, and survival analysis assesses the duration of vaccine protective efficacy.

For qualitative data, such as incidence of adverse reactions, HPV infection status, and efficacy ratings, the chi-square test is the central method for comparing differences in the distribution of

categorical variables between groups (e.g., proportion of adverse reactions in the vaccine group versus the control group) Logistic regression is used to analyze causality (e.g., the association of vaccination dose with risk of infection).It benefits multifactorial adjustments (e.g., covariates, such as age, sex, etc.)

4.2. Selection of statistical tools by focus of purpose

Defining the research question is the first step in selecting a statistical method. If the purpose of the study is to characterize the vaccination population and the underlying data distribution, descriptive statistics will suffice.

If one aims to study the changes in vaccine protection efficacy over time and its influencing factors, the Cox proportional hazard model can be selected. For example, the Cox proportional hazard model is analyzed in Figure 1, which plays a key role in studying factors influencing the recovery time of COVID-19 patients (Das et al. 2022). This model quantifies the time-dependent influences on vaccine protection (e.g., age, type of vaccination) and determines the significance of the risk factor recovery in urban areas through regression coefficients.

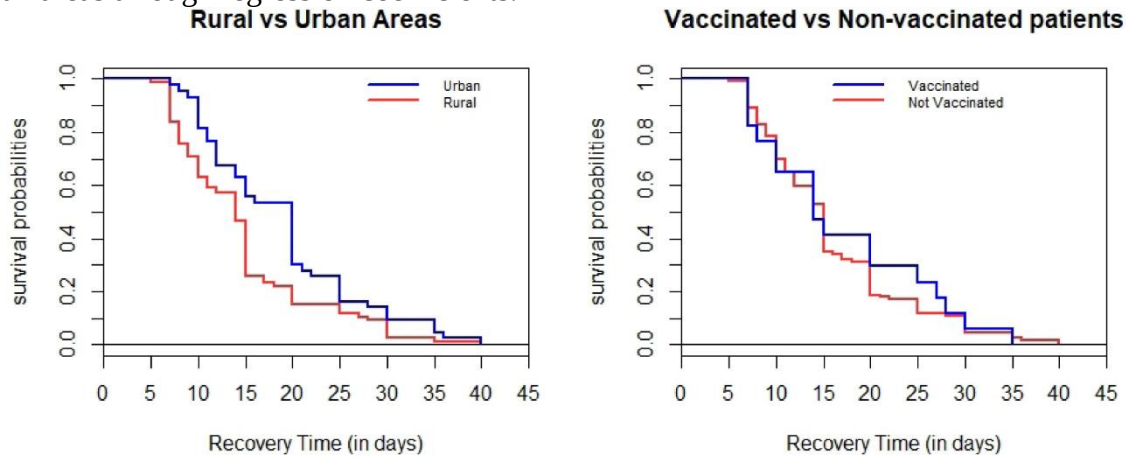

Figure 1. Survival analysis of recovery by location and vaccination status(Das et al. 2022).

Alt Text for the figure: Recovery curves based on survival functions were analyzed for two risk factors: residence (urban vs. rural) and vaccination status (vaccinated vs. unvaccinated). The figure is divided into three panels: the left panel compares survival probabilities across regions, the right panel compares survival probabilities by vaccination status, and the lower panel displays the distribution curve of recovery time (in days).

Similarly, in HPV vaccine studies, the Cox proportional hazard model can be used as an ex-ample to study the risk of an individual developing cervical precancerous lesions after HPV vaccination. In this case, the dependent variable is the time since the individual developed cervical precancerous lesions (counting from the start of the vaccination), which not only focuses on whether or not the lesion occurs, but is more precise as to the point in time at which it occurs. The independent variables are multiple, such as the age of the vaccinated individual, which affects the risk of lesion development due to differences in immune response at different ages; the type of vaccine received, which has different preventive effects in different vaccine valences; and the individual's lifestyle habits, such as whether or not he/she is a smoker or not, as smoking lowers immunity and increases the risk of lesion development. Through the Cox proportional hazard model researchers can comprehensively consider these independent variables, analyze the degree of their influence on the dependent variable, and determine which fac-tors have a significant effect on the time of cervical pre-cancer after HPV vaccination, which will provide powerful data support for the evaluation of the impact of HPV vaccine and the formulation of cervical cancer prevention strategies, and help to improve the level of women's health.

Regression analysis is more applicable if one wants to explore the effects of multiple factors on vaccine efficacy or safety. As shown in Figure 2, for the omicron subvariant, the curves obtained

from the regression analysis showed different trends for different comparison groups and vaccination situations(Nealon et al. 2024).For example, the efficacy of the 3-dose mRNA vaccine against mild disease declined from 62% after 4 weeks of vaccination to 48% at 20 weeks. In contrast, the protection against severe disease was initially higher but declined less. The initial efficacy of mixed immunizations was prominent and tended to decay over time. The analysis also com-pared different immunization statuses and vaccination regimens (3 doses/4 doses of mRNA), and the interaction of factors was visualized by regression analysis based on the data fitting results, providing a visual basis for assessing vaccine timeliness and optimizing vaccination strategies.

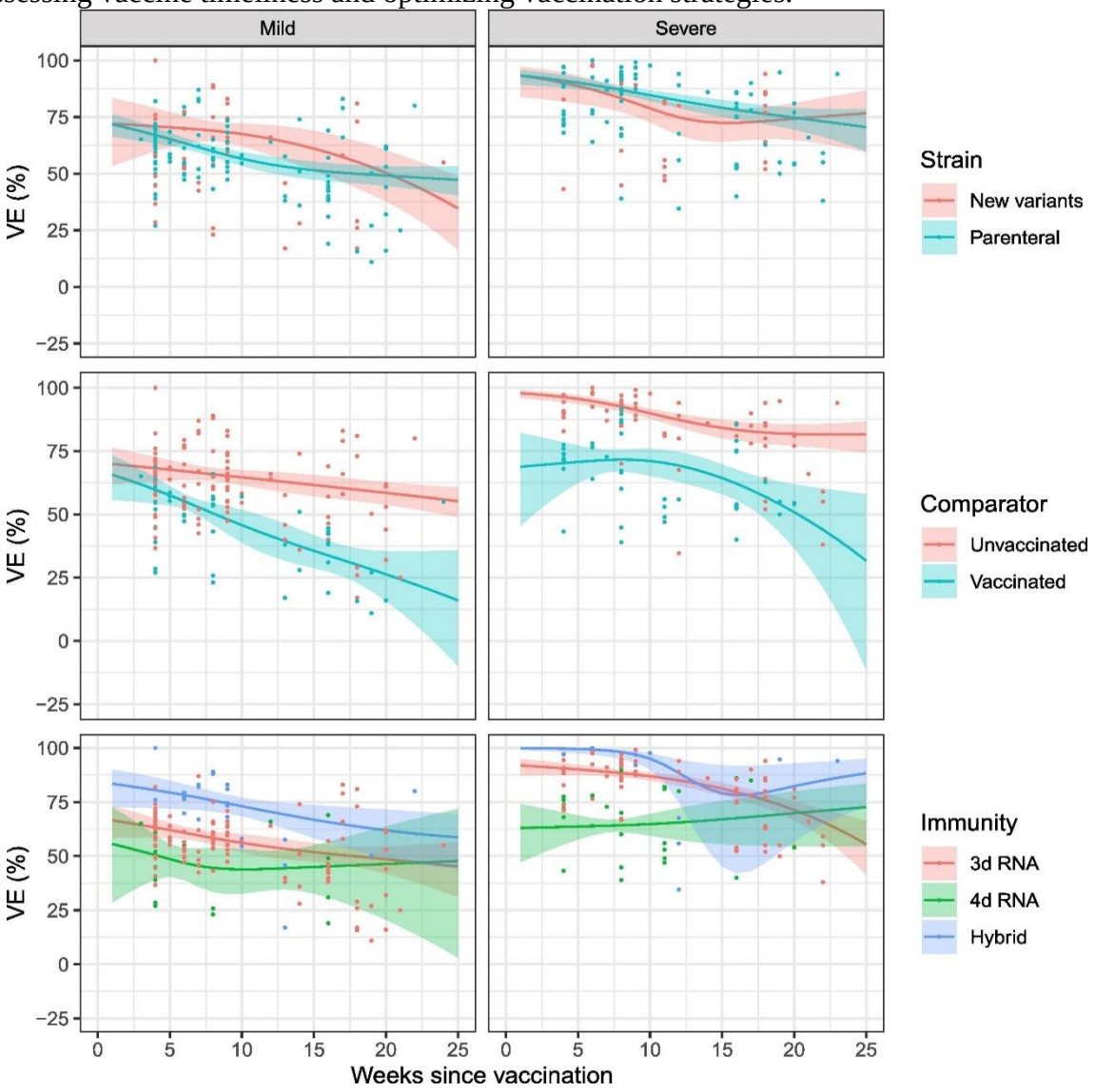

Figure 2. Vaccine efficacy changes over time (Nealon et al. 2024)

Alt Text for the figure: This figure demonstrates the influence of various factors on the change of vaccine efficacy (VE) over time (weeks after vaccination). It covers two conditions: mild and severe cases, and compares the VE data between new variants and parental strains, unvaccinated and vaccinated individuals, as well as different immunization types (3 - dose mRNA, 4 - dose mRNA, and hybrid immunity). The trends of change are presented through curves and scatter points, which is used to analyze the effectiveness of vaccines under different circumstances.

Similarly, regression analysis is a key tool for assessing HPV vaccine efficacy and dynamics, quantifying the impact of different vaccination strategies, vaccine types and temporal factors on protection. Studies have shown differences in the preventive efficacy of bivalent, quadrivalent, and nine valent HPV vaccines against HPV infection and associated lesions (Jia et al. 2025). The regression modeling can further characterize the decay of efficacy of high-priced vaccines, covering

many subtypes. It can analyze the decay trajectory of the protective efficacy of 2-dose versus 3-dose vaccination regimens, revealing the potential impact of dose differences on immune persistence. In the assessment of mixed immunization, the model portrays the evolution of protective efficacy of the composite immunization strategy by integrating the vaccination data of vaccines of different valences. During the study, the regression model calculated confidence intervals of vaccine efficacy by controlling confounding variables and comparing the infection rate and the incidence of cervical lesions in the unvaccinated and vaccinated groups to provide a basis for the vaccination strategies of different populations. In conclusion, by integrating multiple factors such as vaccination time, number of doses, and vaccine price, the regression analysis can fit a dynamic efficacy curve and visualize the evolution of the protection effect, which can provide scientific support for prolonging the protection period of the HPV vaccine and formulating enhanced vaccination strategies.

5. Conclusion

The choice of statistical methods in the analysis of HPV vaccine clinical trial data is directly related to the accuracy and reliability of the study results. This paper reviews the characteristics and scope of application of each commonly used statistical method, such as descriptive statistics, hypothesis testing, survival analysis, and regression analysis. Researchers should reasonably select and comprehensively apply statistical methods according to the type of data and the purpose of the research question, to ensure the scientific and accurate analysis of HPV vaccine clinical trial data. It provides a reference for the selection of appropriate statistical methods for the development of HPV vaccines. It also provides a reference to help researchers objectively measure vaccines' performance and optimize their design. Helping HPV vaccines to benefit the public more widely.

In the future, as statistical methods continue to innovate and develop, it is expected that more suitable methods for analyzing HPV vaccine clinical trial data will be born. These new methods will further enhance the precision and efficiency of analysis, deeply explore the value of data, and promote the research of HPV vaccines in preventing cervical cancer and other related diseases to new heights, thus contributing more to the global public health cause.

References

[1] Chand, T., Dubey, A.K., Misra, G. 2025. Unraveling HPV-associated cancer complexity: From molecular insights to innovative therapies. Heliyon 11:e42437.

[2] Chen, W. 2025. Problems Needing Attention in the Statistical Processing of Data in Papers [J]. The Journal of Traditional Chinese Orthopedics and Traumatology37(02): 3.

[3] Das, D., Saikia, H., Bora, D., Bhattacharjee, D., Das, J. 2022. A survival analysis approach for identifying the risk factors in time to recovery of COVID-19 patients using Cox proportional hazard model. Decision Analytics Journal 5:100137.

[4] Ferlay, J., Colombet, M., Soerjomataram, I., Parkin, D. M., Piñeros, M., Znaor, A., Bray, F. 2021. Cancer statistics for the year 2020: An overview. International Journal of Cancer.

[5] Gore, M., Kabekkodu, S.P., Chakrabarty, S. 2025. Exploring the metabolic alterations in cervical cancer induced by HPV oncoproteins: From mechanisms to therapeutic targets. BBA-Reviews on Cancer 1880:189292.

[6] Jia, X., Hu, S., Kuang, X., Qiao, Y. 2025. Long-term protective efficacy of the Escherichia coli - produced HPV-16/18 bivalent human papillomavirus vaccine in women vaccinated at 18–45 years: A 9-year follow-up study. Infectious Medicine 4:100164.

[7] Liu, Y., Chen, D., Zhao, L., Zhang et al. 2025. Stability study of recombinant 9-valent human papillomavirus vaccine based on Escherichia coli expression system. Human Vaccines & Immunotherapeutics 21(1):2455807.

[8] Nealon, J., Mefsin, Y.M., McMenamin, M.E., Ainslie, K.E.C., Cowling, B.J. 2024. Reported effectiveness of COVID-19 monovalent booster vaccines and hybrid immunity against mild and severe Omicron disease in adults: A systematic review and meta-regression analysis. Vaccine X 17:100451.

[9] Sung H, Ferlay J, Siegel RL, et al. 2021. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries［J］. CA Cancer J Clin 71(3):209-249.

[10] Sierra, M.S., Carvajal, L.J., Dull, P., Herrero, R., Schussler, J., Hildesheim, A., Schiller, et al. 2025. Human papillomavirus type 16 and 18 viral clearance and progression to precancer among women aged 18–25 years enrolled in the Costa Rica HPV prophylactic vaccine trial (CVT). Vaccine 50:126841.

[11] Wang, R., Pan, W., Jin, L., Huang, W., Li, Y., Wu, D., Gao, C., Ma, D., Liao, S. 2020. Human papillomavirus vaccine against cervical cancer: Opportunity and challenge. Cancer Letters 471:88–102.

[12] Xing, R. Y. 2022. Clinical statistical evaluation of HPV vaccine safety Master's thesis, Guilin University of Electronic Technology.