A new endpoint definition improved clinical relevance and statistical power in a vaccine trial.

Gaëlle Pédrono, Rodolphe Thiébaut, Ahmadou Alioum, Philippe Lesprit, Bernard Fritzell, Yves Lévy, Geneviève Chêne

A NEW ENDPOINT DEFINITION IMPROVING CLINICAL RELEVANCE AND STATISTICAL POWER IN VACCINE TRIALS

Authors. *Gaëlle Pédrono[1], Rodolphe Thiébaut[1,2], Ahmadou Alioum[1,2], Philippe Lesprit[4], Bernard Fritzell[5], Yves Lévy[4], Geneviève Chêne[1,2,3].*
1. INSERM, U897 & CIE7, Bordeaux, 33076-France
2. Bordeaux 2 University, School of Public Health (ISPED), Bordeaux, 33076-France
3. Bordeaux University Hospital (CHU Bordeaux), Bordeaux, 33076-France
4. Assistance Publique-Hôpitaux de Paris (AP-HP)Henri Mondor Hospital, Créteil, France
5. Wyeth-Research Paris, France

Corresponding Author: Prof. Geneviève Chêne, MD, PhD
Address:
INSERM U897, Bordeaux School of Public Health (ISPED), Université Victor Segalen Bordeaux 2, Case 11, 146 rue Léo Saignat, 33076 Bordeaux cedex, France
Fax +33 (0) 557 5711 72
Telephone +33 (0) 557 5712 57
E-mail Genevieve.chene@isped.u-bordeaux2.fr

ABSTRACT

Objective: Endpoints used for the evaluation of immunogenicity in vaccine trials are often the proportion of individuals with immune response or geometric means of antibody concentrations for each serotype. When a vaccine includes several types of the same species, we illustrate how an endpoint combining all responses may improve clinical relevance and statistical power.

Study design and settings: The motivating example was the ANRS-114 PNEUMOVAC trial where the effect of two vaccine strategies against *Streptococcus pneumoniae* was assessed in adults infected by the Human Immunodeficiency Virus. The power associated with several endpoints was calculated in the example and in simulations. A new endpoint based on four ordered levels is formulated and analysed by using a proportional odds model.

Results and conclusion: The analysis of this new endpoint led to an odds-ratio allowing detection of improvement and detriment. In the simulation study, this endpoint was associated with the largest statistical power by increasing the amount of information used as compared to usual endpoints. We recommend this new endpoint formulation in the formal development of a new vaccination regimen, whenever applicable.

Key words: Vaccine, endpoint, proportional odds model, HIV infection, statistical power, random effect.

Running title : New endpoint formulation for vaccine trials

Word count: abstract (179), main text (3190), 31 references, three tables, two figures

INTRODUCTION

The efficacy and safety of vaccines are assessed by several traditional parameters [1,2]: incidence of the disease(s) to be prevented, immune response (antibody response and immunocompetent cells), reactogenicity (local and systemic reactions). Phase II trials are usually built to test meaningful differences in the distribution of immunogenicity endpoints. Antibody response to the target pathogen often defines the primary immunogenicity endpoint, antibody concentrations being used as surrogate endpoints for a clinically relevant endpoint [1].

Some vaccines such as pneumococcal polysaccharide vaccines yield a response to several serotypes and are called multivalent vaccines. The immunogenicity evaluation of such vaccines must take into account the response to any serotype, all of them participating in the definition of the primary endpoint. In general, the comparison between groups involves the distribution of serotype specific antibody concentrations for each serotype using geometric means [3,4,5,6,7,8,9]. The P-value is then adjusted for Bonferroni or Family Discovery Rate (FDR)-controlling methods [1,10]. Another approach consists in comparing the percentage of responders to a minimum number (k) of serotypes defined *a priori*. First, the response to each serotype according to a given threshold is defined and then the percentage of responders to k serotypes is compared between groups [4,7,11].

Nevertheless, those endpoints are not fully satisfactory. Indeed, the former may be difficult to interpret, when results are discordant across serotypes. The latter, although clinically relevant, leads to an obvious loss of information. Furthermore, it is difficult to define the threshold for the number of responses to consider. When several serotypes are used to define immunogenicity and if they have the same relative importance, we propose to use a categorical endpoint based on the number of responses obtained and the mode of analysis using a proportional odds model [12].

Our report aims at showing the statistical and clinical relevance of this approach; firstly through the motivating example and secondly based on a simulation study.

METHODS

Study design (Figure 1)

The ANRS 114-Pneumovac trial was a phase II randomized trial comparing two vaccine strategies against pneumococcal infection among adult patients infected by the Human Immunodeficiency Virus (VIH) [13]. The rationale was that HIV-infected patients are more susceptible to infections caused by common bacterial pathogens such as *S. pneumoniae* and therefore are in need of effective preventive strategies against this infection. In this trial, 212 patients were enrolled, 106 in each arm. The tested vaccine strategy consisted in a vaccination by heptavalent pneumococcal protein conjugate vaccine: Prevenar® (PCV) at baseline (W0) and by 23-valent pneumococcal polysaccharide vaccine: Pneumo23® (PPV) at week 4 (W4). The currently recommended strategy was used in the control arm and consisted in a vaccination by PPV at W4. The 7 serotypes of PCV (4, 6B, 9V, 14, 18C, 19F, 23F) are included in PPV and are the most frequently encountered serotypes in Western countries. Antibody concentrations of each common serotype were measured at W0 and W8 for each patient. Antibody response at W8 was compared between the 2 arms (Figure 1).

Statistical methods

Direct comparison of antibody concentrations

The distribution of antibody concentrations is usually not gaussian [3,4,5]. Therefore, antibody concentrations were log-transformed and geometric means were calculated. T-test was used to compare distributions and FDR-controlling method to control for multiplicity of tests [14]. Response to each serotype was defined as two-fold increase between W0 and W8 and titre >1 μg/ml measured at 8 weeks.

Binary endpoints

Three binary endpoints were considered: (i) response to all 7 serotypes (versus non response to at least one of them), (ii) response to at least 5 serotypes among 7 (versus response to less than 4 serotypes) and (iii) response to at least 3 serotypes among 7 (versus

response to less than 2 serotypes). The comparison of responder rates according to treatment group was performed with Fisher exact test.

New endpoint formulation (figure 2)

The new formulation is based on: i) the identification of a response to each serotype, ii) the number of serotypes with response. These two variables were combined into four ordered levels: 0 response, response to 1 or 2 serotypes, response to 3 or 4 serotypes or response to 5 to 7 serotypes (Figure 2). This categorization was considered *a priori* the most relevant by clinicians and immunologists. All serotypes have the same importance and contribute in the same way in the definition of the endpoint. This endpoint is categorical but also ordered because the higher the category, the better the response. First, Chi-2 and trend Chi-2 were used for crude analysis of this endpoint.

Secondly, the Proportional Odds Model (POM) [12] was used to analyse the new endpoint formulated. It is a generalization of a logistic model, the response variable is categorical and ordered (Y=1,2,…,J) instead of binary. A definition of this model is provided in appendix.

When the only explanatory variable is the treatment group, the global odds ratio $\theta = e^{\beta}$ can be interpreted as the odds of being in the higher category above a given category when someone is vaccinated by the tested strategy compared to the control group. The comparison between the two groups can be performed by testing $\theta = 1$ using a likelihood ratio statistics or a Wald test. This is testing an absence of difference between the groups in the distribution of the patients according to each category.

The POM assumption is that the effect of any explanatory variable is the same whatever the level of grouping of the response. In other words, the odds ratio between no response at all versus at least one response should not be different from the odds-ratio between 2 responses or less versus 3 responses or more, etc.

Either for the fixed effect POM model or the mixed effect POM model, parameters could be estimated using maximum likelihood methods with standard softwares. In SAS (SAS® software v.9.1, SAS Institute, Cary, NC, USA), we used proc LOGISTIC for the former and proc NLMIXED for the latter. Code for fitting random effects POM model is provided in Appendix.

Sample size and power were calculated using Kolassa's solution in N'Query software [18].

Simulations

Simulations were performed in order to calculate the power associated with each endpoint according to different scenarii. True parameters were defined according to the distribution of antibody concentrations in the ANRS 114-Pneumovac trial. Antibody concentrations were assumed to follow a log-normal distribution. For a given patient and a given serotype, antibody concentrations were correlated between different time of measurements and for a given patient, antibody concentrations of different serotypes were also correlated. We used the following four steps approach to simulate such data:

1) $Y_0 \sim MVN(M_0, C)$ (MVN for multivariate normal)

Y_0 is the multivariate vector of the logarithm of the titre of the seven serotypes.

M_0 (7x1) and C (7x7) were estimated from a multivariate linear mixed model using real data presented in the results section. They represent vector of means and variance-covariance matrix, respectively. R(7x7) is the correlation matrix .

2) For the reference group, the difference of antibody concentrations from baseline was generated according to a normal distribution: $Y_8 - Y_0 \sim N(p_1, \sigma_p)$ where p_1 and σ_p were based on estimations from multivariate linear mixed models fitted on data from the Pneumovac trial. For the test group : $Y_8 - Y_{0l} \sim N(p_2, \sigma_p)$ where $p_2 = p_1 + \beta_2$ with β_2 being the treatment effect

3) Test group effect was estimated using the different methods presented above

4) Steps 1 to 3 were repeated N=1000 times

Each set of simulation included 200 patients. The power is the percentage of significant tests out of 1000. Simulations were repeated for different values of β_2 (treatment effect) including the one estimated using the real data set.

RESULTS

Application with ANRS 114 Pneumovac data (table 1)

Geometric Means (GMT) were calculated for each serotype (table 1). GMT were generally higher in those vaccinated with the tested strategy. These differences were statistically significant for 2 of them (out of 7) after correction for multiple testing. The statistical power associated to each serotype varied from 35% (serotype 6B) up to 99% (serotype 18C). The power was therefore not homogeneous among serotypes and rather low (<60%) for 4 serotypes among 7. We used a chi-2 and a chi-2 for trend to test the difference of response between the two groups, p-value was respectively 0.0373 and 0.0015.

We then considered the binary endpoint: response to all serotypes *versus* non response to at least one of them. The proportion of patients with successful response was significantly different between groups according to this endpoint. The odds-Ratio (OR)=2.80 (95% Confidence Interval [CI] 1.22-6.43, p=0.0123), meaning that the tested vaccinal strategy was associated with a better immunogenicity. The statistical power of the test for the null hypothesis OR=1 was 68%.

With an endpoint defined as being responder for at least 5 serotypes versus less than 5, the OR was 2.23 (95% CI 1.28-3.89, p=0.0044) and the power was 77%. The odds to be good responder (more than 5 serotypes) was therefore better for the tested vaccine strategy.

Finally, the global OR estimated with the proportional odds model for the previously defined ordered categories was 2.12 (95% CI 1.26-3.56, p=0.0047). The power was 81%. Patients in the tested vaccine strategy group showed a higher number of responses than the control group. Morever, the proportional odds assumption was satisfied (test of non homogeneity p = 0.82). In this trial, the categorical endpoint was the most powerful as compared to other endpoints considered.

Simulations (table 2)

Estimated values for simulation parameters were:

M_0 = [-2.1484, -0.7459, -0.7377, 0.2910, -0.9306, -0.2530, -1.3869]

$$R=\begin{pmatrix} 1 & 0.6129 & 0.7192 & 0.3498 & 0.6342 & 0.6581 & 0.5685 \\ 0.6129 & 1 & 0.8517 & 0.4884 & 0.8087 & 0.8597 & 0.7902 \\ 0.7192 & 0.8517 & 1 & 0.6037 & 0.8032 & 0.7901 & 0.7821 \\ 0.3498 & 0.4884 & 0.6037 & 1 & 0.5022 & 0.4643 & 0.4849 \\ 0.6342 & 0.8087 & 0.8032 & 0.5022 & 1 & 0.7616 & 0.7841 \\ 0.6581 & 0.8597 & 0.7901 & 0.4643 & 0.7616 & 1 & 0.7519 \\ 0.5685 & 0.7902 & 0.7821 & 0.4849 & 0.7841 & 0.7519 & 1 \end{pmatrix}$$

p, = [2.0728, 1.2022, 1.3784, 1.6299, 1.4080, 0.9547, 1.2380]

σ_p = [0.3520, 0.6219, 0.5427, 0.6075, 0.6277, 0.5527, 0.8066]

The statistical power for the comparison of geometric means of antibody concentrations between the two groups increased with the treatment effect (β^t_2), as expected (table 2). Furthermore, the power was quite heterogeneous according to the serotype considered (eg from 26.0 to 42.3% when β^t_2 = D).

The statistical power was always higher with other outcomes considered (binary or ordered categories). Among binary outcomes, the binary endpoint using a threshold at 5 serotypes led to the best statistical power compared to a threshold set at 7 serotypes (13.2% vs 8.6% for β^t_2 = B; 85.0% vs 75.9% for β^t_2 = D for instance). Although the outcome based on ordered categories led to the best power, the differences with the binary endpoint using a threshold at 5 serotypes were tiny (around 3%), when the effect of the treatment (β_2) was identical according to each serotype. With a treatment effect differing from one serotype to another (very low effect for 5 serotypes and high effect for 2 serotypes, scenario F) the binary endpoint using a threshold at 5 serotypes was substantially less powerful to discriminate poor responders (2 responses for instance) than the new endpoint formulation (54.9.6% vs 63.2%, respectively). Furthermore, a null effect scenario (β^t_2 = **0**) was tested and the type I error was slightly higher for the POM model (α=0.06 as compared as α=0.05 for "$\geq$ 5 reponses" binary endpoint).

Additional simulations were performed to test the robustness of the endpoint with different correlation structures between serotypes (data not shown). Whatever the correlation structure (no correlation, slight correlation, high correlation), the power of this new endpoint formulation was better as compared to the other endpoints.

DISCUSSION

The new endpoint formulation based on ordered categories was more powerful than current endpoints used for the evaluation of a multivalent vaccine efficacy. This was even more true when the response was not homogeneous among all serotypes. With binary endpoints, the power was not satisfactory even if the test was significant in our specific study. This loss of power might be explained by a loss of information due to dichotomization. For an ordered categorical variable, a chi-2 test for trend seemed to be the most powerful test. However, the use of such a test does not yield a measurement of association, does not allow adjustment for confounders and is conditional to the existence of a linear trend of responses among categories.

The main advantage of the proposed strategy is to allow evaluating the effect of the intervention on all serotypes estimating only one parameter (the odds ratio) and performing one test. Hence, it avoids any correction for multiple testing and the main result is easy to interpret.

Indeed, it had practical relevance for clinicians. Firstly, all serotypes contributed in the same way to the construction of the endpoint. The 7 serotypes of interest were those contained in the PCV vaccine since they were the most frequent causes of pneumococcal infections. Therefore, there was no rationale to give more consideration to a serotype than another. Secondly, results can be summarized using a single parameter, i.e a global OR estimated by the POM model, and presented in a simple contingency table. Such a summary contributes to improve the understanding of results by practitioners and makes it easier to remind them and to disseminate them among those who will use the vaccines.

Another advantage is that the implementation is fairly easy and convenient using standard statistical software. It is possible to accommodate for General Linear Model and to adjust for fixed or random effects. Indeed, the model can be extended by including random effects to capture heterogeneity between centers, or between individuals with repeated measurements. In our study, we included testing for such random effects (SAS code in Appendix) and did not find significant heterogeneity among them regarding vaccine response.

A major assumption for the use a POM is the homogeneity of odds ratios among the different categories. This assumption holded in our specific example. Table 3 shows two examples illustrating extreme fictive situations when the assumption holds or not. In the first example, the vaccine strategy is efficient for all serotypes; in that case, the assumption should be satisfied. This translates into quite similar partial odds ratios. In the second example, the new vaccine strategy generates especially very good responders and the referent strategy more likely generates medium responders (example 2). The first strategy could be completely efficient (for the whole serotypes) or not at all. The second strategy could generate responses to almost 3 or 4 serotypes, leading to many responders in medium categories and few in the lowest and the strongest. According to considered cutpoints, one would differently conclude about superiority between vaccine strategies. If the cutpoint is “at least 1 response” versus none, referent strategy would be better, if the cutpoint 5-7 responses is retained, the tested strategy would be better. In this example, the criterion of definite judgment would not be relevant because partial odds-ratio would be different ($p<0.0001$), so proportional odds assumption would not be satisfied. It is especially important to check this proportionality assumption because it is not obvious to intuitively anticipate the homogeneity of partial odds-ratio. Indeed, a vaccine strategy with an effect that varies from one serotype to another may stick to partial odds homogeneity (see Table 2, scenarios A and F).

When the assumption is not satisfied, other models are available such as a continuous ratio model or a partial odds ratio model [16,17,21]. This latter model allows odds to be partially

different accross subgroups of categories provided that they are homogeneous within each subgroup. To be able to use this model the number of categories should be large enough (at least 5 categories). A recent paper [15] provides SAS code for other multinomial model, which can be useful if the proportional odds assumption is not satisfied and more sophisticated model is needed. Another paper [19] suggests solutions to analyse ordinal endpoint when the proportional odds assumption does not hold. However those alternative models estimate additional parameters which leads to a decrease of statistical power to detect an effect.

Another limitation (also shared with binary endpoints) is that the threshold for antibody responses should be defined *a priori* and according to a relevant biological and clinical rationale. In the litterature, definition of antibody response is varying according to the increase considered (e.g. 2- or 3-fold increase of antibody concentrations) [5,23], to a threshold of the titre (e.g. titre greater than 1 μg/ml) [3] or the combination of both: 2-fold increase and titre>1 μg/ml at W8 [11]. This threshold seemed to be the most relevant because it took into account not only the increase of titre between W0 and W8 but also a minimal threshold to reach response level. Indeed, a 2-fold increase from 0.01 to 0.02 μg/ml, could not be considered as a response. In the same way, a patient could present a titre > 1 μg/ml at W8 but also at W0 and be considered as a responder even if the titre was stable. That is why the endpoint definition chosen in the present study seemed to be appropriate.

In some studies [5,6] an alternative approach was used to analyze efficacy of multivalent vaccine trial. Geometric means of different serotypes were compared using mixed MANOVA (Multivariate analysis of variance). This type of model assumes independence of errors across observations. A correlation between dependant variables is modeled through a multivariate normal distribution. For the comparison of group effect, the null hypothesis is that global effect is null for all serotypes. The alternative hypothesis is that there is at least one serotype for which the group effect is significant. Thus, this model is not discriminant enough and does not really answer our initial question of whether the global immunologic efficacy is better in the test group compared to the reference group. Moreover, the fit of those distributions might need complex

assumptions such as mixture of lognormal and joint distributions in the context of antibody measurement with left-censuring issues due to undetectable values [3,20].

Composite endpoints are frequently used as primary endpoint in randomized trials because it may increase the statistical power and thus the efficiency of the trial and may facilitate interpretation [22]. However, the main rationale for using binomial composite endpoint would be the ability to include a range of potentially important outcomes in the composite like in ANRS 114 PNEUMOVAC trial. The use of categorization in vaccine area is often necessary, yet, the construction of such a categorical endpoint is not always suitable, depending on the vaccine strategy, vaccine composition, importance of every serotype. The proposed formulation might also be useful in other situations than vaccine evaluation, as far as a categorization of endpoint seems to be relevant.

CONCLUSION

Guidelines for the formulaton and analysis of primary endpoint in multivalent vaccine trials. We propose a formulation based on a combination of the responses to serotypes and the number of response achieved that is able to achieve a higher statistical power as compared to commonly used endpoints such as means of antibody concentrations and has better clinical relevance. This formulation could also be used in other contexts, whenever the outcome can be divided in few ordered categories formed by units of the same relative importance.

Akwowledgement : We thank Dr. D. Madore and C. Cimino from Wyeth Pharmaceuticals and Dr. D. Goldblatt from ICH for the processing of antibody concentrations. We thank Pr. F. Mentré and Pr. P. Morlat members of the Data and Safety Monitoring board of the study. We thank N.Sarrazin, O.Danet, J.Collin who monitored the study and maintained the database.

Source of funding. This study was performed with the financial support of the Agence Nationale de Recherches sur le Sida et les Hépatites Virales (ANRS).

Conflict of interest statement

B. Fritzell is Vice President International Scientific & Clinical Affairs Vaccines, Wyeth. G. Chêne has received grant support and lecture fees from Boeringer-Ingelheim, and Gilead. All other authors: no conflicts.

REFERENCES

1 **Hudgens MG, Gilbert PB, Self SG**. Endpoints in vaccine trials. *Stat. Methods Med Res* 2004;13:89-114.

2 **Moodie Z, Rossini AJ, Hudgens MG, Gilbert PB, Self SG, Russell ND**. Statistical evaluation of HIV vaccines in early clinical trials. *Contemp Clin Trials* 2006;27:147-60.

3 **Kroon FP, Van Dissel JT, Ravensbergen E et al.** Enhanced antibody response to pneumococcal polysaccharide vaccine after prior immunization with conjugate pneumococcal vaccine HIV-infected adults. *Vaccine* 2001;19:886-94.

4 **Rodriguez-Barradas MC, Musher DM et al.** Antibody to Capsular Polysaccharides of Streptococcus pneumoniae after Vaccination of Human immundeficiency virus-infected subjects with 23-valent penumococcal vaccine. *Clin Infect Dis.* 1992;165:553-6.

5 **Feikin DR, Elie CM et al.** Randomized trial of the quantitative and functional antibody responses to a 7-valent pneumococcal conjugate vaccine and/or 23-valent polysacharide vaccine among HIV-infected adults. *Vaccine* 2001;20:545-53.

6 **Feikin DR, Elie CM et al.** Specificity of the antibody response to the pneumococcal polysaccharide and conjugate vaccines in human immunodeficiency virus-infected adults *Clin Diagn Lab Immunol* 2004;137-41.

7 **Tasker SA, Wallace MR et al.** Reimmunization with 23-valent pneumococcal vaccine for patients infected with human immunodeficiency virus type 1 : clinical, immunologic, and virologic responses. *Clin Infect Dis.* 2002;34:813-21.

8 **Deloria-Knoll M, Steinhoff M et al.** Effect of zinc and vitamin A supplementation on antibody responses to a pneumococcal conjugate vaccine in HIV-positive injection drug users : a randomized trial. *Vaccine* 2006;24:1670-9.

9 **Jackson LA, Neuzil KM, Nahm MH et al.** Immunogenicity of varying dosages of 7-valent pneumococcal polysaccharide-protein conjugate vaccine in seniors previously vaccinated with 23-valent pneumococcal polysaccharide vaccine. *Vaccine* 2007;25:4029-37.

10 **Horne AD, Clifford J et al.** Efficacy and safety of preventive vaccines : FDA evaluation. In processings of the Annual Meeting of the American Statistical Association, Atlanta, 2001 August 5-9.

11 **Rodriguez-Barradas MC, Alexandraki I et al.** Response of human immunodeficiency virus-infected patients receiving highly active antiretroviral therapy to vaccination with 23-valent pneumococcal polysaccharide vaccine. *Clin Infect Dis.* 2003;37:438-47.

12 **McCullagh P.** Regression Model models for Ordinal data. *J Roy Statist Soc* Series B 1980;42:109-42.

13 **Levy Y, Lesprit P, Pedrono G, Molina JM et al**. Immunological efficacy of prime-boost pneumococcal vaccination in HIV-infected adults. Results of the randomized controlled trial, ANRS 114. *AIDS* 2007;21:2425-34.

14 **Hochberg, Y.** A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 1988;75:800-2.

15 **Kuss O, McLerran D.** A note on estimation of the multinomial logistic model with correlated responses in SAS. Compute *Comput Methods Programs Biomed.* 2007;87:262-9.

16 **Cole SR, Ananth CV**. Regression models for unconstrained, partially or fully constrained continuation odds ratios. *Int. J. Epidemiol.* 2001;30:1379-82.

17 **Peterson B, Harrell F**. Partial proportional odds models for ordered response variables. *Appl Stat.* 1990; 39:205–17.

18 **Kolassa JE.** A comparison of size and power calculations for the Wilcoxon statistic for ordered categorical data. *Stat Med.* 1995;14:1577-81.

19 **Bender R, Grouven U.** Using binary logistic regression models for ordinal data with non-proportional odds. *J Clin Epidemio* 1998; 51(10):809-816.

20 **Moulton LH, Halsey NA.** A mixture model detection limits for regression analyses of antibody response to vaccine. *Biometrics* 1995;51:1570-8.

21 **Cox C.** Multinomial regression models based on continuation ratios. *Stat Med.* 1998;7:435-41.

22 **Freemantle N., Calvert M.** Weighting the pros and cons for composite outcomes in clinical trial. *J Clin Epidemiol.* 2007;60:658-9.

23 **Olander RM, Wuorimaa T et al.** Booster response to the tetanus and diphtheria toxoid carriers of 11-valent pneumococcal conjugate vaccine in adults and toddlers. *Vaccine* 2002; 20:336-41.

24 **Whitehead J.** Sample size calculations for ordered categorical data. *Stat Med.* 1993;12:2257-71.

25 **Thiébaut R, Jacqmin-Gadda H, Chêne G et al.** Bivariate linear mixed models using SAS proc MIXED. *Comput Methods Programs Biomed*. 2002;69:249-56.

26 **Moulton LH, O'Brien KL, Kohberger R et al.** Design of a Group-Randomized Streptococcus pneumoniae Vaccine Trial. *Control Clin Trials.* 2001;22:438-52.

27 **Moulton LH, Holt EA, Job JS et al.** Percentile regression analysis of correlated antibody responses. *Stat Med.* 1995;14:777-87.

28 **Subramaniam KS, Segal R, Lyles RH et al.** Qualitative change in antibody responses of human immunodeficiency virus-infected individuals to pneumococcal capsular polysaccharide vaccination associated with highly active antiretroviral therapy. *J Infect Dis.* 2003;187:758-68

29 **Biesheuvel CJ, Vergouwe Y, Steyerberg EW et al.** Polytomous logistic regression analysis could be applied more often in diagnostic research. *J Clin Epidemiol.* 2008;61:125-134

30 **Bender R, Benner A.** Calculating ordinal regression models in SAS and S-Plus. *Biom. Jl.* 2000;42:677-699

31 **Armstrong BG, Sloan M.** Ordinal regression models for epidemiologic data. *Am J Epidemiol.* 1989;129(1):191-204

Tables/Figures legends

Table 1. Results of several endpoint formulations in the ANRS 114 Pneumovac trial, measurements of association, P-value and associated statistical power.

Table 2. Probability of rejecting the null hypothesis at 0.05 level according to possible endpoints with simulated data. N=200 subjects, 1000 simulations

Table 3. Two examples of responses distributions : example 1 where proportional odds assumption holds and example 2 where the assumption does not hold

Figure 1. ANRS 114 Pneumovac trial design.

Figure 2. New endpoint definition: 1) response for each serotype: 2 fold-increase and a titer > 1mg/ml at W8 2) whole response: number of serotypes with response. Example of a patient who had response to 4 different serotypes.

A NEW ENDPOINT DEFINITION IMPROVING CLINICAL RELEVANCE AND STATISTICAL POWER IN VACCINE TRIALS

Authors. *Gaëlle Pédrono[1], Rodolphe Thiébaut[1,2], Ahmadou Alioum[1,2], Philippe Lesprit[4], Bernard Fritzell[5], Yves Lévy[4], Geneviève Chêne[1,2,3].*
1. INSERM, U897 & CIE7, Bordeaux, 33076-France
2. Bordeaux 2 University, School of Public Health (ISPED), Bordeaux, 33076-France
3. Bordeaux University Hospital (CHU Bordeaux), Bordeaux, 33076-France
4. Assistance Publique-Hôpitaux de Paris (AP-HP)Henri Mondor Hospital, Créteil, France
5. Wyeth-Research Paris, France

Corresponding Author: Prof. Geneviève Chêne, MD, PhD
Address:
INSERM U897, Bordeaux School of Public Health (ISPED), Université Victor Segalen Bordeaux 2, Case 11, 146 rue Léo Saignat, 33076 Bordeaux cedex, France
Fax +33 (0) 557 5711 72
Telephone +33 (0) 557 5712 57
E-mail Genevieve.chene@isped.u-bordeaux2.fr

ABSTRACT

Objective: Endpoints used for the evaluation of immunogenicity in vaccine trials are often the proportion of individuals with immune response or geometric means of antibody concentrations for each serotype. When a vaccine includes several types of the same species, we illustrate how an endpoint combining all responses may improve clinical relevance and statistical power.

Study design and settings: The motivating example was the ANRS-114 PNEUMOVAC trial where the effect of two vaccine strategies against *Streptococcus pneumoniae* was assessed in adults infected by the Human Immunodeficiency Virus. The power associated with several endpoints was calculated in the example and in simulations. A new endpoint based on four ordered levels is formulated and analysed by using a proportional odds model.

Results and conclusion: The analysis of this new endpoint led to an odds-ratio allowing detection of improvement and detriment. In the simulation study, this endpoint was associated with the largest statistical power by increasing the amount of information used as compared to usual endpoints. We recommend this new endpoint formulation in the formal development of a new vaccination regimen, whenever applicable.

Key words: Vaccine, endpoint, proportional odds model, HIV infection, statistical power, random effect.

Running title : New endpoint formulation for vaccine trials

Word count: abstract (179), main text (3190), 31 references, three tables, two figures

INTRODUCTION

The efficacy and safety of vaccines are assessed by several traditional parameters [1,2]: incidence of the disease(s) to be prevented, immune response (antibody response and immunocompetent cells), reactogenicity (local and systemic reactions). Phase II trials are usually built to test meaningful differences in the distribution of immunogenicity endpoints. Antibody response to the target pathogen often defines the primary immunogenicity endpoint, antibody concentrations being used as surrogate endpoints for a clinically relevant endpoint [1].

Some vaccines such as pneumococcal polysaccharide vaccines yield a response to several serotypes and are called multivalent vaccines. The immunogenicity evaluation of such vaccines must take into account the response to any serotype, all of them participating in the definition of the primary endpoint. In general, the comparison between groups involves the distribution of serotype specific antibody concentrations for each serotype using geometric means [3,4,5,6,7,8,9]. The P-value is then adjusted for Bonferroni or Family Discovery Rate (FDR)-controlling methods [1,10]. Another approach consists in comparing the percentage of responders to a minimum number (k) of serotypes defined *a priori*. First, the response to each serotype according to a given threshold is defined and then the percentage of responders to k serotypes is compared between groups [4,7,11].

Nevertheless, those endpoints are not fully satisfactory. Indeed, the former may be difficult to interpret, when results are discordant across serotypes. The latter, although clinically relevant, leads to an obvious loss of information. Furthermore, it is difficult to define the threshold for the number of responses to consider. When several serotypes are used to define immunogenicity and if they have the same relative importance, we propose to use a categorical endpoint based on the number of responses obtained and the mode of analysis using a proportional odds model [12].

Our report aims at showing the statistical and clinical relevance of this approach; firstly through the motivating example and secondly based on a simulation study.

METHODS

Study design (Figure 1)

The ANRS 114-Pneumovac trial was a phase II randomized trial comparing two vaccine strategies against pneumococcal infection among adult patients infected by the Human Immunodeficiency Virus (VIH) [13]. The rationale was that HIV-infected patients are more susceptible to infections caused by common bacterial pathogens such as *S. pneumoniae* and therefore are in need of effective preventive strategies against this infection. In this trial, 212 patients were enrolled, 106 in each arm. The tested vaccine strategy consisted in a vaccination by heptavalent pneumococcal protein conjugate vaccine: Prevenar® (PCV) at baseline (W0) and by 23-valent pneumococcal polysaccharide vaccine: Pneumo23® (PPV) at week 4 (W4). The currently recommended strategy was used in the control arm and consisted in a vaccination by PPV at W4. The 7 serotypes of PCV (4, 6B, 9V, 14, 18C, 19F, 23F) are included in PPV and are the most frequently encountered serotypes in Western countries. Antibody concentrations of each common serotype were measured at W0 and W8 for each patient. Antibody response at W8 was compared between the 2 arms (Figure 1).

Statistical methods

Direct comparison of antibody concentrations

The distribution of antibody concentrations is usually not gaussian [3,4,5]. Therefore, antibody concentrations were log-transformed and geometric means were calculated. T-test was used to compare distributions and FDR-controlling method to control for multiplicity of tests [14]. Response to each serotype was defined as two-fold increase between W0 and W8 and titre >1 μg/ml measured at 8 weeks.

Binary endpoints

Three binary endpoints were considered: (i) response to all 7 serotypes (versus non response to at least one of them), (ii) response to at least 5 serotypes among 7 (versus response to less than 4 serotypes) and (iii) response to at least 3 serotypes among 7 (versus

response to less than 2 serotypes). The comparison of responder rates according to treatment group was performed with Fisher exact test.

New endpoint formulation (figure 2)

The new formulation is based on: i) the identification of a response to each serotype, ii) the number of serotypes with response. These two variables were combined into four ordered levels: 0 response, response to 1 or 2 serotypes, response to 3 or 4 serotypes or response to 5 to 7 serotypes (Figure 2). This categorization was considered *a priori* the most relevant by clinicians and immunologists. All serotypes have the same importance and contribute in the same way in the definition of the endpoint. This endpoint is categorical but also ordered because the higher the category, the better the response. First, Chi-2 and trend Chi-2 were used for crude analysis of this endpoint.

Secondly, the Proportional Odds Model (POM) [12] was used to analyse the new endpoint formulated. It is a generalization of a logistic model, the response variable is categorical and ordered (Y=1,2,…,J) instead of binary. A definition of this model is provided in appendix.

When the only explanatory variable is the treatment group, the global odds ratio $\theta = e^{\beta}$ can be interpreted as the odds of being in the higher category above a given category when someone is vaccinated by the tested strategy compared to the control group. The comparison between the two groups can be performed by testing $\theta = 1$ using a likelihood ratio statistics or a Wald test. This is testing an absence of difference between the groups in the distribution of the patients according to each category.

The POM assumption is that the effect of any explanatory variable is the same whatever the level of grouping of the response. In other words, the odds ratio between no response at all versus at least one response should not be different from the odds-ratio between 2 responses or less versus 3 responses or more, etc.

Either for the fixed effect POM model or the mixed effect POM model, parameters could be estimated using maximum likelihood methods with standard softwares. In SAS (SAS® software v.9.1, SAS Institute, Cary, NC, USA), we used proc LOGISTIC for the former and proc NLMIXED for the latter. Code for fitting random effects POM model is provided in Appendix.

Sample size and power were calculated using Kolassa's solution in N'Query software [18].

Simulations

Simulations were performed in order to calculate the power associated with each endpoint according to different scenarii. True parameters were defined according to the distribution of antibody concentrations in the ANRS 114-Pneumovac trial. Antibody concentrations were assumed to follow a log-normal distribution. For a given patient and a given serotype, antibody concentrations were correlated between different time of measurements and for a given patient, antibody concentrations of different serotypes were also correlated. We used the following four steps approach to simulate such data:

1) $Y_0 \sim MVN(M_0, C)$ (MVN for multivariate normal)

Y_0 is the multivariate vector of the logarithm of the titre of the seven serotypes.

M_0 (7x1) and C (7x7) were estimated from a multivariate linear mixed model using real data presented in the results section. They represent vector of means and variance-covariance matrix, respectively. R(7x7) is the correlation matrix .

2) For the reference group, the difference of antibody concentrations from baseline was generated according to a normal distribution: $Y_8 - Y_0 \sim N(p_1, \sigma_p)$ where p_1 and σ_p were based on estimations from multivariate linear mixed models fitted on data from the Pneumovac trial. For the test group : $Y_8 - Y_{0l} \sim N(p_2, \sigma_p)$ where $p_2 = p_1 + \beta_2$ with β_2 being the treatment effect

3) Test group effect was estimated using the different methods presented above

4) Steps 1 to 3 were repeated N=1000 times

Each set of simulation included 200 patients. The power is the percentage of significant tests out of 1000. Simulations were repeated for different values of β_2 (treatment effect) including the one estimated using the real data set.

RESULTS

Application with ANRS 114 Pneumovac data (table 1)

Geometric Means (GMT) were calculated for each serotype (table 1). GMT were generally higher in those vaccinated with the tested strategy. These differences were statistically significant for 2 of them (out of 7) after correction for multiple testing. The statistical power associated to each serotype varied from 35% (serotype 6B) up to 99% (serotype 18C). The power was therefore not homogeneous among serotypes and rather low (<60%) for 4 serotypes among 7. We used a chi-2 and a chi-2 for trend to test the difference of response between the two groups, p-value was respectively 0.0373 and 0.0015.

We then considered the binary endpoint: response to all serotypes *versus* non response to at least one of them. The proportion of patients with successful response was significantly different between groups according to this endpoint. The odds-Ratio (OR)=2.80 (95% Confidence Interval [CI] 1.22-6.43, p=0.0123), meaning that the tested vaccinal strategy was associated with a better immunogenicity. The statistical power of the test for the null hypothesis OR=1 was 68%.

With an endpoint defined as being responder for at least 5 serotypes versus less than 5, the OR was 2.23 (95% CI 1.28-3.89, p=0.0044) and the power was 77%. The odds to be good responder (more than 5 serotypes) was therefore better for the tested vaccine strategy.

Finally, the global OR estimated with the proportional odds model for the previously defined ordered categories was 2.12 (95% CI 1.26-3.56, p=0.0047). The power was 81%. Patients in the tested vaccine strategy group showed a higher number of responses than the control group. Morever, the proportional odds assumption was satisfied (test of non homogeneity p = 0.82). In this trial, the categorical endpoint was the most powerful as compared to other endpoints considered.

Simulations (table 2)

Estimated values for simulation parameters were:

M_0 = [-2.1484, -0.7459, -0.7377, 0.2910, -0.9306, -0.2530, -1.3869]

$$R=\begin{pmatrix} 1 & 0.6129 & 0.7192 & 0.3498 & 0.6342 & 0.6581 & 0.5685 \\ 0.6129 & 1 & 0.8517 & 0.4884 & 0.8087 & 0.8597 & 0.7902 \\ 0.7192 & 0.8517 & 1 & 0.6037 & 0.8032 & 0.7901 & 0.7821 \\ 0.3498 & 0.4884 & 0.6037 & 1 & 0.5022 & 0.4643 & 0.4849 \\ 0.6342 & 0.8087 & 0.8032 & 0.5022 & 1 & 0.7616 & 0.7841 \\ 0.6581 & 0.8597 & 0.7901 & 0.4643 & 0.7616 & 1 & 0.7519 \\ 0.5685 & 0.7902 & 0.7821 & 0.4849 & 0.7841 & 0.7519 & 1 \end{pmatrix}$$

p, = [2.0728, 1.2022, 1.3784, 1.6299, 1.4080, 0.9547, 1.2380]

σ_p = [0.3520, 0.6219, 0.5427, 0.6075, 0.6277, 0.5527, 0.8066]

The statistical power for the comparison of geometric means of antibody concentrations between the two groups increased with the treatment effect (β^t_2), as expected (table 2). Furthermore, the power was quite heterogeneous according to the serotype considered (eg from 26.0 to 42.3% when β^t_2 = D).

The statistical power was always higher with other outcomes considered (binary or ordered categories). Among binary outcomes, the binary endpoint using a threshold at 5 serotypes led to the best statistical power compared to a threshold set at 7 serotypes (13.2% vs 8.6% for β^t_2 = B; 85.0% vs 75.9% for β^t_2 = D for instance). Although the outcome based on ordered categories led to the best power, the differences with the binary endpoint using a threshold at 5 serotypes were tiny (around 3%), when the effect of the treatment (β_2) was identical according to each serotype. With a treatment effect differing from one serotype to another (very low effect for 5 serotypes and high effect for 2 serotypes, scenario F) the binary endpoint using a threshold at 5 serotypes was substantially less powerful to discriminate poor responders (2 responses for instance) than the new endpoint formulation (54.9.6% vs 63.2%, respectively). Furthermore, a null effect scenario (β^t_2 = **0**) was tested and the type I error was slightly higher for the POM model (α=0.06 as compared as α=0.05 for "$\geq$ 5 reponses" binary endpoint).

Additional simulations were performed to test the robustness of the endpoint with different correlation structures between serotypes (data not shown). Whatever the correlation structure (no correlation, slight correlation, high correlation), the power of this new endpoint formulation was better as compared to the other endpoints.

DISCUSSION

The new endpoint formulation based on ordered categories was more powerful than current endpoints used for the evaluation of a multivalent vaccine efficacy. This was even more true when the response was not homogeneous among all serotypes. With binary endpoints, the power was not satisfactory even if the test was significant in our specific study. This loss of power might be explained by a loss of information due to dichotomization. For an ordered categorical variable, a chi-2 test for trend seemed to be the most powerful test. However, the use of such a test does not yield a measurement of association, does not allow adjustment for confounders and is conditional to the existence of a linear trend of responses among categories.

The main advantage of the proposed strategy is to allow evaluating the effect of the intervention on all serotypes estimating only one parameter (the odds ratio) and performing one test. Hence, it avoids any correction for multiple testing and the main result is easy to interpret.

Indeed, it had practical relevance for clinicians. Firstly, all serotypes contributed in the same way to the construction of the endpoint. The 7 serotypes of interest were those contained in the PCV vaccine since they were the most frequent causes of pneumococcal infections. Therefore, there was no rationale to give more consideration to a serotype than another. Secondly, results can be summarized using a single parameter, i.e a global OR estimated by the POM model, and presented in a simple contingency table. Such a summary contributes to improve the understanding of results by practitioners and makes it easier to remind them and to disseminate them among those who will use the vaccines.

Another advantage is that the implementation is fairly easy and convenient using standard statistical software. It is possible to accommodate for General Linear Model and to adjust for fixed or random effects. Indeed, the model can be extended by including random effects to capture heterogeneity between centers, or between individuals with repeated measurements. In our study, we included testing for such random effects (SAS code in Appendix) and did not find significant heterogeneity among them regarding vaccine response.

A major assumption for the use a POM is the homogeneity of odds ratios among the different categories. This assumption holded in our specific example. Table 3 shows two examples illustrating extreme fictive situations when the assumption holds or not. In the first example, the vaccine strategy is efficient for all serotypes; in that case, the assumption should be satisfied. This translates into quite similar partial odds ratios. In the second example, the new vaccine strategy generates especially very good responders and the referent strategy more likely generates medium responders (example 2). The first strategy could be completely efficient (for the whole serotypes) or not at all. The second strategy could generate responses to almost 3 or 4 serotypes, leading to many responders in medium categories and few in the lowest and the strongest. According to considered cutpoints, one would differently conclude about superiority between vaccine strategies. If the cutpoint is “at least 1 response” versus none, referent strategy would be better, if the cutpoint 5-7 responses is retained, the tested strategy would be better. In this example, the criterion of definite judgment would not be relevant because partial odds-ratio would be different ($p<0.0001$), so proportional odds assumption would not be satisfied. It is especially important to check this proportionality assumption because it is not obvious to intuitively anticipate the homogeneity of partial odds-ratio. Indeed, a vaccine strategy with an effect that varies from one serotype to another may stick to partial odds homogeneity (see Table 2, scenarios A and F).

When the assumption is not satisfied, other models are available such as a continuous ratio model or a partial odds ratio model [16,17,21]. This latter model allows odds to be partially

different accross subgroups of categories provided that they are homogeneous within each subgroup. To be able to use this model the number of categories should be large enough (at least 5 categories). A recent paper [15] provides SAS code for other multinomial model, which can be useful if the proportional odds assumption is not satisfied and more sophisticated model is needed. Another paper [19] suggests solutions to analyse ordinal endpoint when the proportional odds assumption does not hold. However those alternative models estimate additional parameters which leads to a decrease of statistical power to detect an effect.

Another limitation (also shared with binary endpoints) is that the threshold for antibody responses should be defined *a priori* and according to a relevant biological and clinical rationale. In the litterature, definition of antibody response is varying according to the increase considered (e.g. 2- or 3-fold increase of antibody concentrations) [5,23], to a threshold of the titre (e.g. titre greater than 1 μg/ml) [3] or the combination of both: 2-fold increase and titre>1 μg/ml at W8 [11]. This threshold seemed to be the most relevant because it took into account not only the increase of titre between W0 and W8 but also a minimal threshold to reach response level. Indeed, a 2-fold increase from 0.01 to 0.02 μg/ml, could not be considered as a response. In the same way, a patient could present a titre > 1 μg/ml at W8 but also at W0 and be considered as a responder even if the titre was stable. That is why the endpoint definition chosen in the present study seemed to be appropriate.

In some studies [5,6] an alternative approach was used to analyze efficacy of multivalent vaccine trial. Geometric means of different serotypes were compared using mixed MANOVA (Multivariate analysis of variance). This type of model assumes independence of errors across observations. A correlation between dependant variables is modeled through a multivariate normal distribution. For the comparison of group effect, the null hypothesis is that global effect is null for all serotypes. The alternative hypothesis is that there is at least one serotype for which the group effect is significant. Thus, this model is not discriminant enough and does not really answer our initial question of whether the global immunologic efficacy is better in the test group compared to the reference group. Moreover, the fit of those distributions might need complex

assumptions such as mixture of lognormal and joint distributions in the context of antibody measurement with left-censuring issues due to undetectable values [3,20].

Composite endpoints are frequently used as primary endpoint in randomized trials because it may increase the statistical power and thus the efficiency of the trial and may facilitate interpretation [22]. However, the main rationale for using binomial composite endpoint would be the ability to include a range of potentially important outcomes in the composite like in ANRS 114 PNEUMOVAC trial. The use of categorization in vaccine area is often necessary, yet, the construction of such a categorical endpoint is not always suitable, depending on the vaccine strategy, vaccine composition, importance of every serotype. The proposed formulation might also be useful in other situations than vaccine evaluation, as far as a categorization of endpoint seems to be relevant.

CONCLUSION

Guidelines for the formulaton and analysis of primary endpoint in multivalent vaccine trials. We propose a formulation based on a combination of the responses to serotypes and the number of response achieved that is able to achieve a higher statistical power as compared to commonly used endpoints such as means of antibody concentrations and has better clinical relevance. This formulation could also be used in other contexts, whenever the outcome can be divided in few ordered categories formed by units of the same relative importance.

Akwowledgement : We thank Dr. D. Madore and C. Cimino from Wyeth Pharmaceuticals and Dr. D. Goldblatt from ICH for the processing of antibody concentrations. We thank Pr. F. Mentré and Pr. P. Morlat members of the Data and Safety Monitoring board of the study. We thank N.Sarrazin, O.Danet, J.Collin who monitored the study and maintained the database.

Source of funding. This study was performed with the financial support of the Agence Nationale de Recherches sur le Sida et les Hépatites Virales (ANRS).

Conflict of interest statement

B. Fritzell is Vice President International Scientific & Clinical Affairs Vaccines, Wyeth. G. Chêne has received grant support and lecture fees from Boeringer-Ingelheim, and Gilead. All other authors: no conflicts.

REFERENCES

1 **Hudgens MG, Gilbert PB, Self SG**. Endpoints in vaccine trials. *Stat. Methods Med Res* 2004;13:89-114.

2 **Moodie Z, Rossini AJ, Hudgens MG, Gilbert PB, Self SG, Russell ND**. Statistical evaluation of HIV vaccines in early clinical trials. *Contemp Clin Trials* 2006;27:147-60.

3 **Kroon FP, Van Dissel JT, Ravensbergen E et al.** Enhanced antibody response to pneumococcal polysaccharide vaccine after prior immunization with conjugate pneumococcal vaccine HIV-infected adults. *Vaccine* 2001;19:886-94.

4 **Rodriguez-Barradas MC, Musher DM et al.** Antibody to Capsular Polysaccharides of Streptococcus pneumoniae after Vaccination of Human immundeficiency virus-infected subjects with 23-valent penumococcal vaccine. *Clin Infect Dis.* 1992;165:553-6.

5 **Feikin DR, Elie CM et al.** Randomized trial of the quantitative and functional antibody responses to a 7-valent pneumococcal conjugate vaccine and/or 23-valent polysacharide vaccine among HIV-infected adults. *Vaccine* 2001;20:545-53.

6 **Feikin DR, Elie CM et al.** Specificity of the antibody response to the pneumococcal polysaccharide and conjugate vaccines in human immunodeficiency virus-infected adults *Clin Diagn Lab Immunol* 2004;137-41.

7 **Tasker SA, Wallace MR et al.** Reimmunization with 23-valent pneumococcal vaccine for patients infected with human immunodeficiency virus type 1 : clinical, immunologic, and virologic responses. *Clin Infect Dis.* 2002;34:813-21.

8 **Deloria-Knoll M, Steinhoff M et al.** Effect of zinc and vitamin A supplementation on antibody responses to a pneumococcal conjugate vaccine in HIV-positive injection drug users : a randomized trial. *Vaccine* 2006;24:1670-9.

9 **Jackson LA, Neuzil KM, Nahm MH et al.** Immunogenicity of varying dosages of 7-valent pneumococcal polysaccharide-protein conjugate vaccine in seniors previously vaccinated with 23-valent pneumococcal polysaccharide vaccine. *Vaccine* 2007;25:4029-37.

10 **Horne AD, Clifford J et al.** Efficacy and safety of preventive vaccines : FDA evaluation. In processings of the Annual Meeting of the American Statistical Association, Atlanta, 2001 August 5-9.

11 **Rodriguez-Barradas MC, Alexandraki I et al.** Response of human immunodeficiency virus-infected patients receiving highly active antiretroviral therapy to vaccination with 23-valent pneumococcal polysaccharide vaccine. *Clin Infect Dis.* 2003;37:438-47.

12 **McCullagh P.** Regression Model models for Ordinal data. *J Roy Statist Soc* Series B 1980;42:109-42.

13 **Levy Y, Lesprit P, Pedrono G, Molina JM et al**. Immunological efficacy of prime-boost pneumococcal vaccination in HIV-infected adults. Results of the randomized controlled trial, ANRS 114. *AIDS* 2007;21:2425-34.

14 **Hochberg, Y.** A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 1988;75:800-2.

15 **Kuss O, McLerran D.** A note on estimation of the multinomial logistic model with correlated responses in SAS. Compute *Comput Methods Programs Biomed.* 2007;87:262-9.

16 **Cole SR, Ananth CV**. Regression models for unconstrained, partially or fully constrained continuation odds ratios. *Int. J. Epidemiol.* 2001;30:1379-82.

17 **Peterson B, Harrell F**. Partial proportional odds models for ordered response variables. *Appl Stat.* 1990; 39:205–17.

18 **Kolassa JE.** A comparison of size and power calculations for the Wilcoxon statistic for ordered categorical data. *Stat Med.* 1995;14:1577-81.

19 **Bender R, Grouven U.** Using binary logistic regression models for ordinal data with non-proportional odds. *J Clin Epidemio* 1998; 51(10):809-816.

20 **Moulton LH, Halsey NA.** A mixture model detection limits for regression analyses of antibody response to vaccine. *Biometrics* 1995;51:1570-8.

21 **Cox C.** Multinomial regression models based on continuation ratios. *Stat Med.* 1998;7:435-41.

22 **Freemantle N., Calvert M.** Weighting the pros and cons for composite outcomes in clinical trial. *J Clin Epidemiol.* 2007;60:658-9.

23 **Olander RM, Wuorimaa T et al.** Booster response to the tetanus and diphtheria toxoid carriers of 11-valent pneumococcal conjugate vaccine in adults and toddlers. *Vaccine* 2002; 20:336-41.

24 **Whitehead J.** Sample size calculations for ordered categorical data. *Stat Med.* 1993;12:2257-71.

25 **Thiébaut R, Jacqmin-Gadda H, Chêne G et al.** Bivariate linear mixed models using SAS proc MIXED. *Comput Methods Programs Biomed*. 2002;69:249-56.

26 **Moulton LH, O'Brien KL, Kohberger R et al.** Design of a Group-Randomized Streptococcus pneumoniae Vaccine Trial. *Control Clin Trials.* 2001;22:438-52.

27 **Moulton LH, Holt EA, Job JS et al.** Percentile regression analysis of correlated antibody responses. *Stat Med.* 1995;14:777-87.

28 **Subramaniam KS, Segal R, Lyles RH et al.** Qualitative change in antibody responses of human immunodeficiency virus-infected individuals to pneumococcal capsular polysaccharide vaccination associated with highly active antiretroviral therapy. *J Infect Dis.* 2003;187:758-68

29 **Biesheuvel CJ, Vergouwe Y, Steyerberg EW et al.** Polytomous logistic regression analysis could be applied more often in diagnostic research. *J Clin Epidemiol.* 2008;61:125-134

30 **Bender R, Benner A.** Calculating ordinal regression models in SAS and S-Plus. *Biom. Jl.* 2000;42:677-699

31 **Armstrong BG, Sloan M.** Ordinal regression models for epidemiologic data. *Am J Epidemiol.* 1989;129(1):191-204

Tables/Figures legends

Table 1. Results of several endpoint formulations in the ANRS 114 Pneumovac trial, measurements of association, P-value and associated statistical power.

Table 2. Probability of rejecting the null hypothesis at 0.05 level according to possible endpoints with simulated data. N=200 subjects, 1000 simulations

Table 3. Two examples of responses distributions : example 1 where proportional odds assumption holds and example 2 where the assumption does not hold

Figure 1. ANRS 114 Pneumovac trial design.

Figure 2. New endpoint definition: 1) response for each serotype: 2 fold-increase and a titer > 1mg/ml at W8 2) whole response: number of serotypes with response. Example of a patient who had response to 4 different serotypes.

Table(s)

Table 1: Results of several endpoint formulations in the ANRS 114 Pneumovac trial, measurements of association, P-value and associated statistical power.

Endpoint	Vaccine Strategy		Measurement of association	p-value	Statistical Power (%)
	Referent n=103	Test n=104			
Serotypes of PCV	**Geometric mean antibody titers (μg/ml)**				
4	1.0	1.5	Δ= 0.5	0.052*	55
6B	1.6	2.2	Δ= 0.6	0.112*	35
9V	1.9	2.7	Δ= 0.8	0.052*	57
14	5.9	9.2	Δ= 3.3	0.052*	51
18C	1.6	3.1	Δ= 1.5	0.002*	99
19F	2.0	2.9	Δ= 0.9	0.052*	86
23F	0.9	1.9	Δ= 1.0	0.005*	88
Chi-2				0.0373	
Trend Chi-2				0.0015	
Response to all serotypes n(%)	9 (9)	22 (21)	OR=2.80	0.0123	68
≥ 5 serotypes n(%)	41 (40)	62 (60)	OR=2.23	0.0044	77
≥ 3 serotypes n(%)	69 (66)	82 (79)	OR=1.84	0.0549	48
Responders in ordered categories n(%)					
0 serotype	10 (10)	5 (5)			81
1-2	24 (23)	17 (16)			
3-4	28 (27)	20 (19)			
5-7	41 (40)	62 (60)	OR = 2.12	0.0047	

* correction for multiple testing (FDR), significant level is 0.05. Abbreviations: Δ=difference of geometric mean, OR=Odds-Ratio

Table(s)

Table 2: Probability of rejecting the null hypothesis at 0.05 level according to possible endpoints with simulated data. N=200 subjects, 1000 simulations

Power (%)	β_2=A	β^t_2= B	β^t_2= C	β^t_2= D	β^t_2= E	β^t_2= F	β^t_2= G
Serotypes of PCV							
4	63	1	17	39	60	5	1
6B	27	2	17	33	54	5	1
9V	42	2	22	42	65	4	1
14	12	2	16	29	48	39	1
18C	89	2	20	42	64	53	1
19F	42	3	19	40	61	56	1
23F	85	2	12	26	42	3	1
Chi-2	94	7	55	82	97	40	4
Trend Chi-2	99	15	84	96	99	74	5
Response to all serotypes	91	9	53	76	91	32	3
≥ 5 serotypes	95	13	67	85	96	55	5
≥ 3 serotypes	64	7	35	52	69	32	4
Responders in ordered categories	96	14	72	89	98	63	6

A = [0.55, 0.34, 0.38, 0.20, 0.70, 0.38, 0.86] estimations from ANRS 114 PNEUMOVAC trial data
B = [0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1],
C = [0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3]
D = [0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4]
E = [0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5]
F = [0.1, 0.1, 0.1, 0.5, 0.5, 0.5, 0.1]
G = [0, 0, 0, 0, 0, 0, 0]

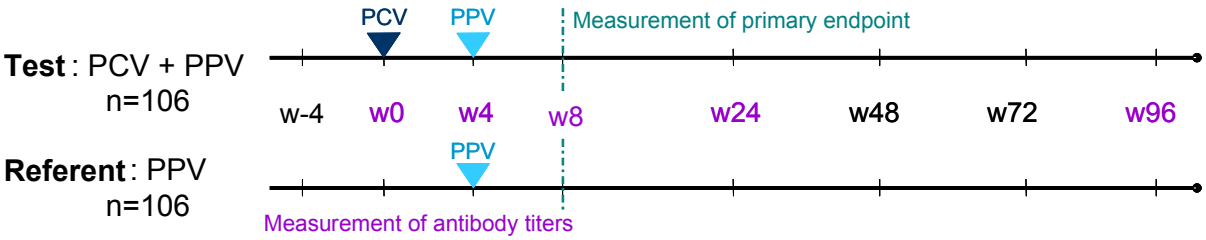
Table(s)

Table 3: Two examples of responses distributions : example 1 where proportional odds assumption holds and example 2 where the assumption does not hold

	Categories of responses			
	0	**1-2**	**3-4**	**5-7**
Example 1				
Test : PCV+PPV (N=100)	5	15	20	60
Referent : PPV (N=100)	10	25	25	40
Partial odds-ratio	$OR_1 = 2,11$	$OR_2 = 2,15$		$OR_3 = 2,25$
p-value for proportional odds assumption		p = 0,9863		
Global odds ratio IC95%		OR = 2,21 [1,30-3,76]		
Example 2				
Test : PCV+PPV (N=100)	40	10	10	40
Referent : PPV (N=100)	10	50	30	10
Partial odds-ratio	$OR_1 = 0,17$	$OR_2 = 1,50$		$OR_3 = 6,00$
p-value for proportional odds assumption		p < 0,0001		
Global odds ratio (p-value)		OR = 1,07 [0,65 - 1,76]		

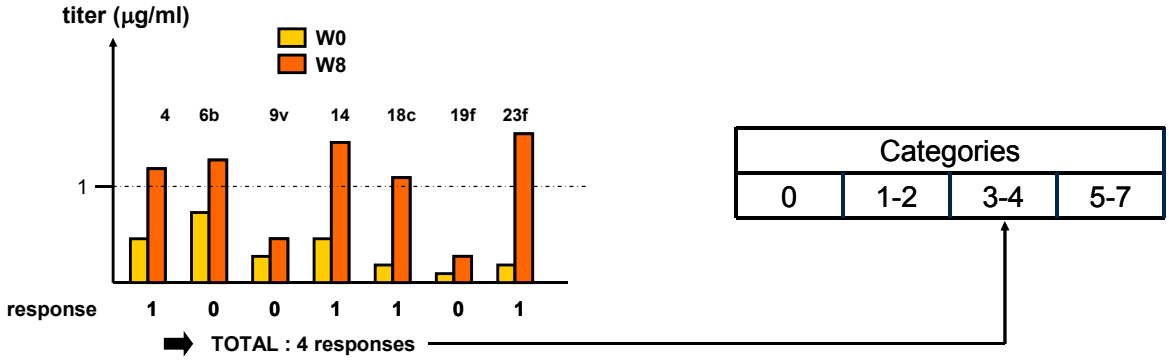
Figure(s)

Figure 1.



Figure(s)

Figure 2.



APPENDIX : Definition of Proportional Odds Model

Proportional Odds Model (POM) [**Error! Reference source not found.**] was used to analyse the new endpoint formulated. It is also called ordinal logistic model (Scott, 1997), cumulative logit model (Lee 1992, Ananth & Kleinbaum 1997), cumulative odds model (Armstrong & Sloan 1989, Greenland 1994), or McCullagh's grouped continuous model (Greenwood & Farewell, 1988 [30]. It is a generalization of a logistic model, the response variable is categorical and ordered (Y=1,2,…,J) instead of binary. The model can be written as follows:

$$\text{logit}\left[P(Y_i > j)\right] = \log\left[\frac{P(Y_i > j)}{P(Y_i \leq j)}\right] = \alpha_j + \beta' X_i \qquad \text{for i=1,..,n and j=1,…,J-1}$$

where p[(Y_i > j)] is the probability of a subject i = 1,…n to present a response in a category greater than j, αj is the log-odds of being above category j when the vector of covariate X_i is equal to 0, β the q-vector of parameters to be estimated and X_i a (q x n) matrix of covariates including the treatment group ($X_{i1} = 0$ for subject i in the control group and 1 for subject i in the tested strategy). When the only explanatory variable is the treatment group, the global odds ratio $\theta = e^{\beta}$ can be interpreted as the odds of being in the higher category above a given category when someone is vaccinated by the tested strategy compared to the control group. The comparison between the two groups can be performed by testing $\theta = 1$ using a likelihood ratio statistics or a Wald test. This is testing an absence of difference between the groups in the distribution of the patients according to each category.

With our endpoint, three logits can be calculated, one for each of the following categories: ≥1 vs 0 response, ≥3 vs <3 responses, ≥5 vs <5 responses. The POM assumption is that the parameter β does not vary according to the categories, i.e. the model assumes that effects of covariates are the same for each cutpoint. This is also called 'equal slopes assumptions' [30]. Hence, the odds is proportional to the difference between the categories of X and β is the

constant of proportionality. Because β is the same whatever j, the underlying hypothesis is that the effect of any explanatory variable is the same whatever the level of grouping of the response. In other words, the POM is invariant under collapsibility.

An extension of the model including random effects to take into account correlation between subunit (in case of a clinical trial performed in k=1,… K centres, for instance) was written as follows:

$$\text{logit}\left[P(Y_{ik} > j)\right] = \alpha_j + \beta' X_{ik} + u_k \qquad \text{for k=1,…,K, i=1,..,n}_k \text{ and j=1,…,J-1}$$

where $u_k \sim N\left(0, \sigma_u^2\right)$ is a random effect reflecting a centre-specific correction to the baseline risk [15].

APPENDIX : Definition of Proportional Odds Model

Proportional Odds Model (POM) [**Error! Reference source not found.**] was used to analyse the new endpoint formulated. It is also called ordinal logistic model (Scott, 1997), cumulative logit model (Lee 1992, Ananth & Kleinbaum 1997), cumulative odds model (Armstrong & Sloan 1989, Greenland 1994), or McCullagh's grouped continuous model (Greenwood & Farewell, 1988 [30]. It is a generalization of a logistic model, the response variable is categorical and ordered (Y=1,2,…,J) instead of binary. The model can be written as follows:

$$\text{logit}\left[P(Y_i > j)\right] = \log\left[\frac{P(Y_i > j)}{P(Y_i \leq j)}\right] = \alpha_j + \beta' X_i \qquad \text{for i=1,..,n and j=1,…,J-1}$$

where p[(Y_i > j)] is the probability of a subject i = 1,…n to present a response in a category greater than j, αj is the log-odds of being above category j when the vector of covariate X_i is equal to 0, β the q-vector of parameters to be estimated and X_i a (q x n) matrix of covariates including the treatment group ($X_{i1} = 0$ for subject i in the control group and 1 for subject i in the tested strategy). When the only explanatory variable is the treatment group, the global odds ratio $\theta = e^{\beta}$ can be interpreted as the odds of being in the higher category above a given category when someone is vaccinated by the tested strategy compared to the control group. The comparison between the two groups can be performed by testing $\theta = 1$ using a likelihood ratio statistics or a Wald test. This is testing an absence of difference between the groups in the distribution of the patients according to each category.

With our endpoint, three logits can be calculated, one for each of the following categories: ≥1 vs 0 response, ≥3 vs <3 responses, ≥5 vs <5 responses. The POM assumption is that the parameter β does not vary according to the categories, i.e. the model assumes that effects of covariates are the same for each cutpoint. This is also called 'equal slopes assumptions' [30]. Hence, the odds is proportional to the difference between the categories of X and β is the

constant of proportionality. Because β is the same whatever j, the underlying hypothesis is that the effect of any explanatory variable is the same whatever the level of grouping of the response. In other words, the POM is invariant under collapsibility.

An extension of the model including random effects to take into account correlation between subunit (in case of a clinical trial performed in k=1,… K centres, for instance) was written as follows:

$$\text{logit}\left[P(Y_{ik} > j)\right] = \alpha_j + \beta' X_{ik} + u_k \qquad \text{for k=1,…,K, i=1,..,n}_k \text{ and j=1,…,J-1}$$

where $u_k \sim N\left(0, \sigma_u^2\right)$ is a random effect reflecting a centre-specific correction to the baseline risk [15].

APPENDIX : SAS® code for Proportional Odds Model with random effect

```
proc nlmixed data=dataset corr ecorr optcheck tech=congra;
     bounds b02 b03 s > 0;
     parms b01=0.74 b02=0.5 b03=0.5 b1=0 s=0.01;
          eta1 = b01 - b1*group + u ;
          eta2 = (b01-b02) - b1*group + u;
        eta3 = (b01-b02-b03) - b1*group + u;
        /*likelihood computation*/
          p1=1-(exp(eta1)/(1+exp(eta1)));
        p2=1-(exp(eta2)/(1+exp(eta2)))-p1;
        p3=1-(exp(eta3)/(1+exp(eta3)))-p1-p2;
        p4=1-p1-p2-p3;
        L=(p1**y1)*(p2**y2)*(p3**y3)*(p4**y4);
        ll=log(L);
     random u ~ normal(0,s) subject=center;
   model cat ~ general(ll);
run;
```