Vaccine 37 (2019) 6737–6742

Contents lists available at ScienceDirect

Vaccine

journal homepage: www.elsevier.com/locate/vaccine

On the use of nonparametric tests for comparing immunological Reverse Cumulative distribution curves (RCDCs)

Robert D. Small [a], Ayca Ozol-Godfrey [b,*], Lihan Yan [c]

[a] RDS Statistical Consulting, LLC, Tampa, FL, USA
[b] Sunovion Pharmaceuticals, Inc., Marlborough, MA, USA
[c] Office of Biostatistics and Epidemiology, Center for Biologics Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD, USA

ARTICLE INFO

Article history:
Received 20 December 2018
Received in revised form 30 August 2019
Accepted 3 September 2019
Available online 16 September 2019

Keywords:
Distribution
Normality assumption
Rank-based test
Survival
Vaccine

ABSTRACT

Reverse Cumulative Distribution Curves (RCDCs) have proven to be a useful tool in summarizing immune response profiles in vaccine studies since their introduction by Reed, Meade, and Steinhoff (RMS) (1995). They are able to display virtually all of the treatment data and characterize summary statistics such as means or even their confidence intervals (CIs) that might be obscure. RMS mentioned their similarity to survival curves often used to summarize time-to-event data which are usually not normally distributed. The RCDCs, while intuitively pleasing and useful, contain important properties which allow for more powerful statistical applications. In this paper, we will suggest several widely used rank-based tests to compare the curves in the context of vaccine studies. These rank-based tests allow for comparisons between treatments, for stratified analyses, weighted analyses, and other modifications that make them the alternative of parametric analyses without the normality assumptions.

Clinical trial identification: NCT01712984 and NCT01230957.

1. Introduction

Reverse Cumulative Distribution Curves (RCDCs) have proven to be a useful tool in summarizing immune response profiles in vaccine studies since their introduction by Reed, Meade, and Steinhoff (RMS) (1995). Many researchers have used them in various ways to expound on the characteristics of the populations of immune responses measured by serum antibody levels [1,5,11,17]. They are able to display virtually all of the data from a population and characterize summary statistics such as means or even confidence intervals that might be obscure. The curves also serve as a diagnostic tool in searching for anomalies in the data that may indicate non-normality [15]. RMS mentioned their similarity to survival curves often used to summarize time-to-event data which are usually not normally distributed. The curves, while intuitively pleasing and useful, contain important properties which allow for more powerful uses of the curves, as discussed in Section 4.

Although often the discussions of RCDCs concern possible differences in the standard parametric statistical tests of point estimates [12] and the graphical displays without any statistical tests, several standard statistical tests which compare distributions can be applied to compare RCDCs since they are standard statistical quantities. In this paper, we will suggest several widely used possible test methods to compare the curves by illustrations of several real-world examples. We will also describe the use of tests, such as the Kolmogorov-Smirnov (KS) test, based on the RCDCs through the example of testing for the log-normal distribution assumption often seen in vaccine studies where log transformations are made to carry out parametric comparisons of populations of immune responses.

A nonparametric comparison is often useful since the antibody response in a population may not be normally distributed even after transformation. In fact, the discussions of RCDCs are often about that issue. We will illustrate that it is possible to stratify on covariates so that virtually any of the usual analyses that are used to compare samples of antibody response can be repeated with a nonparametric method, which is distribution free and in general can be more powerful than the counter-part of a parametric test in situations where the normality distribution assumption is deviated [9].

As described in Reed et al. [14], RCDCs are step functions based on the order statistics of the data. They begin with a value of 1.0 or 100% at an antibody titer of virtually zero and fall to a value of zero above the largest titer value in steps of 1/n. If there are ties then the step size is equal to the number of tied values times 1/n.

* Corresponding author.
E-mail address: ayca.ozol-godfrey@sunovion.com (A. Ozol-Godfrey).

https://doi.org/10.1016/j.vaccine.2019.09.007
0264-410X/

In mathematical notations, let the order statistics of the titer values be $t_{(1)}, t_{(2)}, \ldots\ldots t_{(n)}$. Let $I(\cdot)$ be the indicator function. Then the RCDC, defined as 1 minus the empirical distribution function, is denoted as

$$\widehat{C_n}(x) = \frac{1}{n}\sum_{i=1}^{i=n} I(t_{(i)} > x),$$

where $\widehat{C_n}(x) = \widehat{C_n}(t_{(i)})$ for all $x \in [t_{(i)}, t_{(i+1)}]$, i = 1,..., n-1, implying that $\widehat{C_n}(.)$ would be a step function.

Note that if there are ties, the relative order of the tied values is arbitrary and the corresponding step is a multiple of 1/n. Since the RCDCs are based entirely on the order statistics, it is reasonable to consider rank methods for comparing two such curves.

In this paper, we consider the Wilcoxon and log rank tests. These tests are the two common tests used on these kinds of data, in particular on survival curves. But our situations are slightly different than survival curves. First, we do not consider censoring, since it is unlikely to be present in our situation. Second, survival curves have steps at events, while we have an observation for every subject. We are not considering missing values. In our graphs, the curves do not look like step functions as the graphs are drawn as continuous curves connecting the midpoints of the steps. It is merely for convenience.

In Section 2 we give a short summary of comparisons of immunological data from vaccine studies and indicate where there are opportunities for advantages of nonparametric approaches. Section 3 shows the usefulness of the RCDCs to summarize immune response data and describes how they give information that the normal distribution based summaries may miss. Section 4 describes the use of the tests in various situations with examples from clinical trials. Summaries and conclusions and some other thoughts are given in Section 5.

2. The comparison of group antibody responses

A common measure of the effectiveness of a vaccine is the antibody response levels in serum samples [12]. Antibody response is the result of an assay of a subject's serum sample at a fixed time point before or after vaccination. It is a measure of immunological activity, usually a specific antibody, generally measured either as a concentration or a titer. The concentrations are usually measured on a single serum sample whereas the titers are the results of dilution assays. Both tend to have long tails and therefore are usually transformed with a logarithm function for analyses. Without loss of generality, the discussed statistical methods could be applied to any type of antibody response.

It is usually considered that the level of the response is a measure of risk of infection, with higher responses implying lower probability in risk of infection. Thus, to compare the likely efficacies between vaccines, the geometric means (GMs) are compared. The most common assumption is that the logarithms of the individual responses are normally distributed and therefore inferential statistics (hypothesis tests and confidence interval estimations) can be based on normal theory. For example, the confidence interval of a ratio of two GMs is constructed using usual approaches for the log-transformed data and then the limits of the intervals are back transformed to get intervals for the ratios of two GMs. Details and other extensions are described in Nauta [12].

Comparing the means of the logarithms and assuming that the samples have a common variance does present an issue sometimes. This could happen, for example, if a sample had a large number of non-responders (very low antibody response) yet had a number of very large values which would make the GM greater than or equal to another sample where the values are less sparse. The non-responders have no protection while the very large values may provide no more protection than moderate ones and there are fewer of them. Lachenbruch et al. [8] have given a method that considers a comparison using both mean and variance differences. Nauta et al. [13] have shown that population values with equal means but varying other shapes can lead to differences in the VE.

Thus, even though the usual comparisons of antibody response using logarithms and normality assumptions have done very well and have a substantial history, it may at times be useful to use techniques that are free from some of these assumptions for the reasons discussed above. In fact, they are used to summarize and enlighten. The use of the RCDCs, since their introduction in RMS, has expanded and is now common. Since there are a set of nonparametric methods that have proven valuable in a number of similar data, it only makes sense to use them to give additional support to discussions of the curves in evaluating the curves.

3. Usefulness and advantages of RCDCs to compare samples of vaccine antibody response

We have already discussed the advantages of using a nonparametric method for comparing a set of samples of vaccine antibody response. The usual transformations of the antibody response do not always produce normal data and therefore an alternative is needed. The reason that the RMS suggestion to use the curves is so widely accepted is that the detailed distribution of the data is often more important than the summary statistics alone. If the curves are presented and the details are more complex than the sample mean discussed, it is reasonable to have a measure such as a statistical test to compare the curves.

There are in fact a number of tests based on ranks for comparing data like those depicted in a RCDC. The most commonly used ones are the log rank test and the Wilcoxon rank sum test. These are easy to use and understand, well studied, and powerful. They also address the issue raised earlier that the shape of the curves may be more important than the summary statistic. There is another interesting characteristic of these tests that is sometimes useful. The tests have weighted versions. For example the log rank test gives equal weight to each rank, while the Wilcoxon gives greater weight to lower values. There are other rank based tests that give varying weights to the individual ranks. These are discussed in general by Schoenfeld [16] as well as by Harrington and Fleming [4]

The weights could be useful in several situations. For example, a response above a certain threshold may be considered fully protective. Thus, there is no advantage for still a higher response. Therefore for deciding differences, that are related to VE, we may want to give greater weight to values below the threshold rather than those above. In other situations, a higher response might imply longer persistency. Other idiosyncrasies, such as number of values below the lower limit of quantitation or extremely high antibody responses, may also be considered differentially.

Some rank tests allow stratification and in that sense mimic the usual parametric comparisons with covariates. The log rank test has a straight forward method for stratified analyses. The Wilcoxon rank sum test has a stratified version called the van Elteren test [19]. It is less useful when the treatment effects vary among the strata. Some authors have attempted to address this weakness with varying success [10]. In general, the stratification will increase power and allow more direct analogies with parametric alternatives.

Use of the rank tests eliminates concern over the appropriate transformation. The antibody responses are seldom normally distributed and the logarithm is often used to normalize the data. In flow cytometry assays the arc sine is sometimes used. Even then

the data are sometimes not normal. Use of ranks eliminates all of these concerns.

Finally it is well known that if the distribution of the data is, in fact, known, then the rank tests give similar results [7]. There is very little loss in power, with the loss decreasing with growing sample size. At the same time there is no dependence on these assumptions.

4. The use of nonparametric tests to compare antibody response from vaccine trials -some examples

In Sections 4.1, 4.2, and 4.4, we will use the baseline data from a Sanofi Pasteur influenza vaccine trial ([3], NCT01712984) even though post-vaccination antibody response would be more of an interest in general. Section 4.3, uses the post-vaccination immune response data from another Sanofi Pasteur vaccine trial (see [2], NCT01230957). These trials were not designed to make the comparisons we will show here nor are they of primary interest. They do provide reasonable examples of the kinds of comparisons the rank based methods can make.

4.1. Comparison of two groups

Fig. 1 is a comparison of two curves from the pooled baseline data set and the pre-vaccination titers taken on each subject for an antigen. It compares two age groups at baseline—one 18 to 49 years old and the other 50 to 64 years old. There are 1403 and 710 subjects, respectively in the two age groups.

The graph seems to indicate a difference, and both the log rank and the Wilcoxon tests indicate significant differences (p = <0.0001 and 0.0003, respectively). Note that there are little differences for very high values above 1100 and for very low values below 10. Yet there is a strong separation in the middle. The fact that there is strong separation in the middle and none in the tails indicates some deviation from normality.

In contrast, Fig. 2 is a comparison between the same age groups but with a different antigen. Here there seems to be no separation of the curves and the log rank and Wilcoxon tests indicate no significant differences (p = 0.30 and 0.24, respectively).

4.2. Differences in rank tests

Fig. 3 shows curves comparing the age groups for a third antigen. The curves seem to deviate for values below about 50 and then converge for the larger values. In this case, the log rank test finds no difference (p = 0.18) while the Wilcoxon test declares a larger difference (p = 0.0002). This is because the Wilcoxon test automatically gives larger weights to lower ranks. The log rank test gives equal weights to all ranks, but it is possible to use a weighted log rank test [16,18] if justified. The usual *t*-test on the logarithms of the data gives a significant result. This is probably because the mean is usually different if any part of the curve is different. In fact, Tarone and Ware did an extensive study of the weighting of these two tests. The effective weights are functions of the risk population for the Wilcoxon tests; while for the log rank, equal weight is given. It is also known that the log rank is the most efficient comparator if the proportion hazard assumption is satisfied. If the response to one vaccine (or antigen) was proportional to the other, then the log rank would be the preferred test. If the curves were close for part of the range and then separated for another part, the weighting could make a large difference. As a diagnostic, Hess [6] discusses some graphical techniques for assessing the proportional hazard assumption.

Though there may be cases where one vaccine gives a more favorable response than another in some range than in another part of the range of data, the particular situation would need to be assessed and decisions made on the well documented differences of these two tests.

4.3. Stratified tests

When making the usual parametric comparisons of population immune measures, a covariate is often included in the model. We can do this with the log rank test as well. It can also be done with the Wilcoxon test, but that is a bit more difficult and has some issues that the log rank test does not have (see discussion of the Wilcoxon test in Section 3). The very definition of the log rank test makes it straightforward to do as much stratification as the data will bear. In Fig. 4 we have four RCDCs (see [2], NCT01230957). They represent two treatments each in two age categories. It

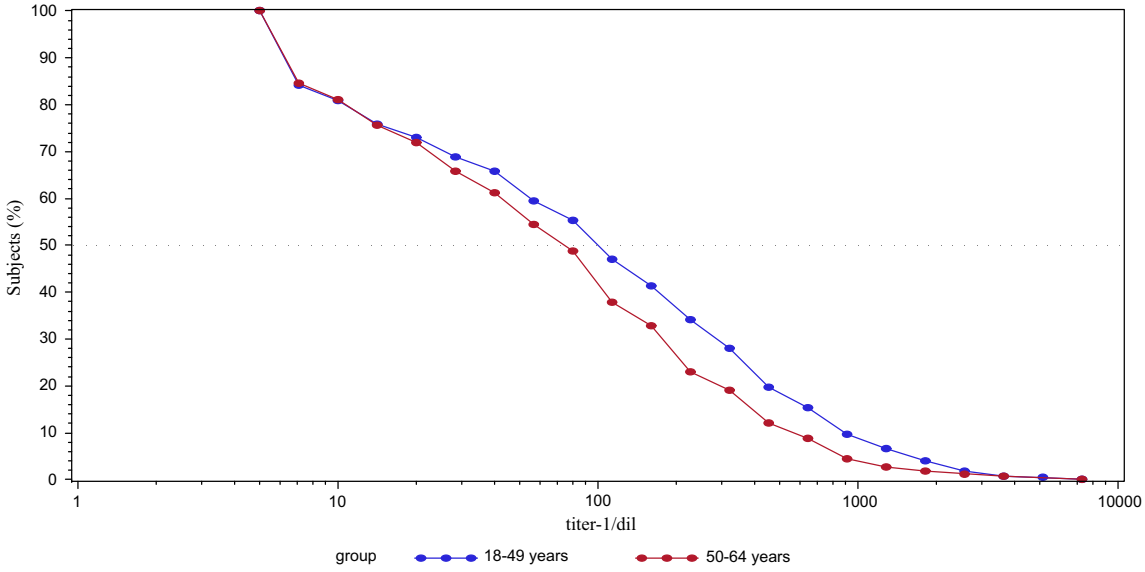

Fig. 1. RCDCs for different age groups at baseline for an influenza antigen.

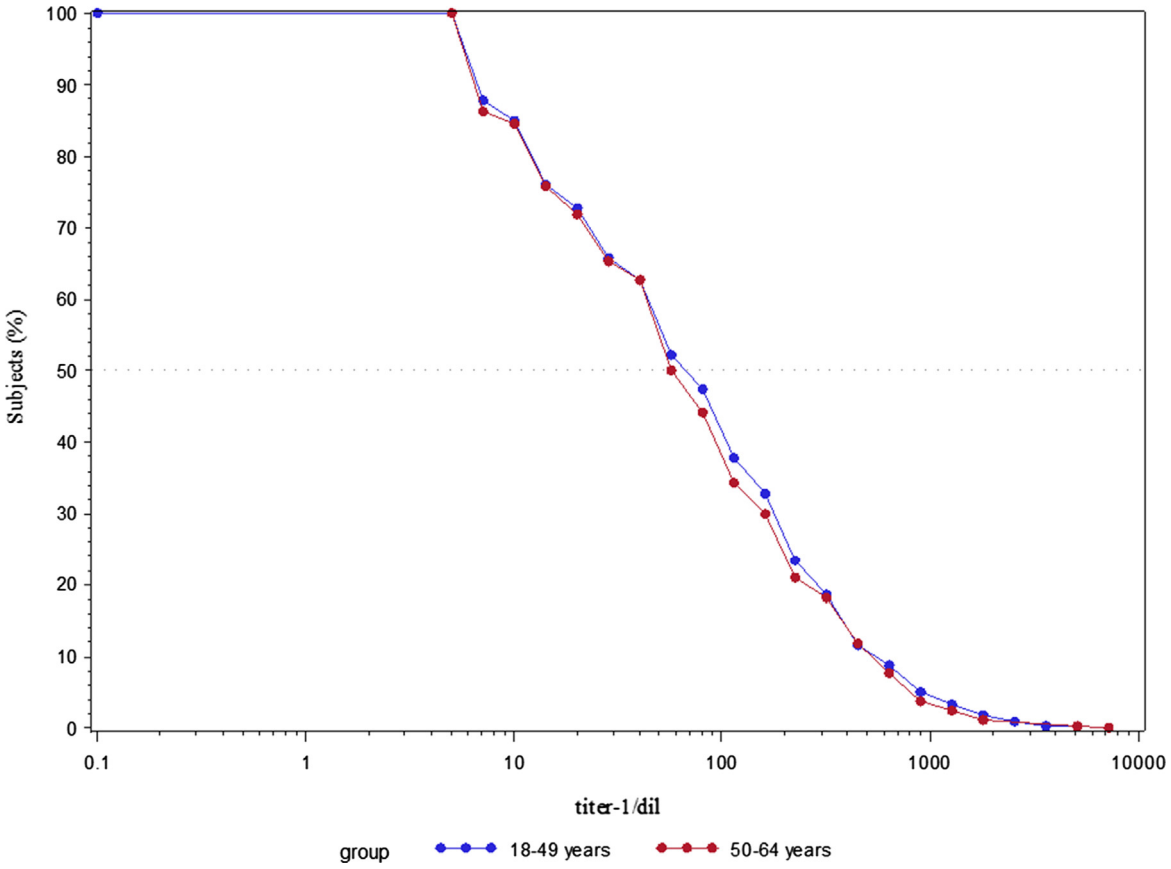

Fig. 2. A comparison of the same age groups as in Fig. 1 for a different antigen.

appears that the response is affected by age, perhaps more so in one treatment. The variation due to age should be considered in any comparison. The results for the log rank test stratified for age group is that the treatments are different with p-value = 0.011.

4.4. Tests of normality

When we transform data such as immunological response by taking logarithms, we usually assume the distribution of the resultant data is normal. We can use a rank based test to decide if, in fact the data are normal. We can compare the RCDC to the normal curve with the same mean and variance as the data with the Kolmogorov-Smirnov (KS) test. The KS test statistic is the maximum deviation between the two curves.

As an illustration, we ran the KS test against each of the previously mentioned four influenza vaccine trial baseline antigens of the combined baseline groups after taking the usual logarithm transformation. The results are given in the table below (see Table 1).

We see that all four fail the test for the normality by a decisive amount. Thus, it makes sense to investigate other methods of analyses that do not require the normality assumption.

5. Summary and conclusions

We have recommended some rank-based tests be used to compare RCDCs of vaccine induced immunological responses collected in vaccine studies. The curves are already a common summary and diagnostic for comparing samples. Often, they are used to further clarify or even to minimize parametric comparisons of the GMs. A test of the comparisons based on the graphs only seems reasonable and useful. The rank tests described in this paper allow for comparisons of curves, for stratified analyses, for weighted analyses and other modifications that make them the counterpart of parametric analyses except for the normality assumptions. They therefore are both a clarifying and a diagnostic tool. While we present two-sided test examples, the idea can be easily extended to one-sided hypothesis testing situations where one-sided rank-based tests are of interest. For example, it's common in practice in both research and regulatory settings to determine whether an investigational vaccine group has an immune response profile that are non-inferior to that of another active control vaccine group.

Others have noted that the details of the distribution of the response can have substantial effects on the VE of the vaccine and that therefore more subtle comparisons than just GMs should be considered. The rank tests provide a comparison of the distribution, not merely the GMs and further allow weighted comparisons which evaluate certain characteristics of the distributions more than others.

Conflict of interest

The research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. This article reflects the views of the authors and should not be construed to represent FDA's views or policies.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

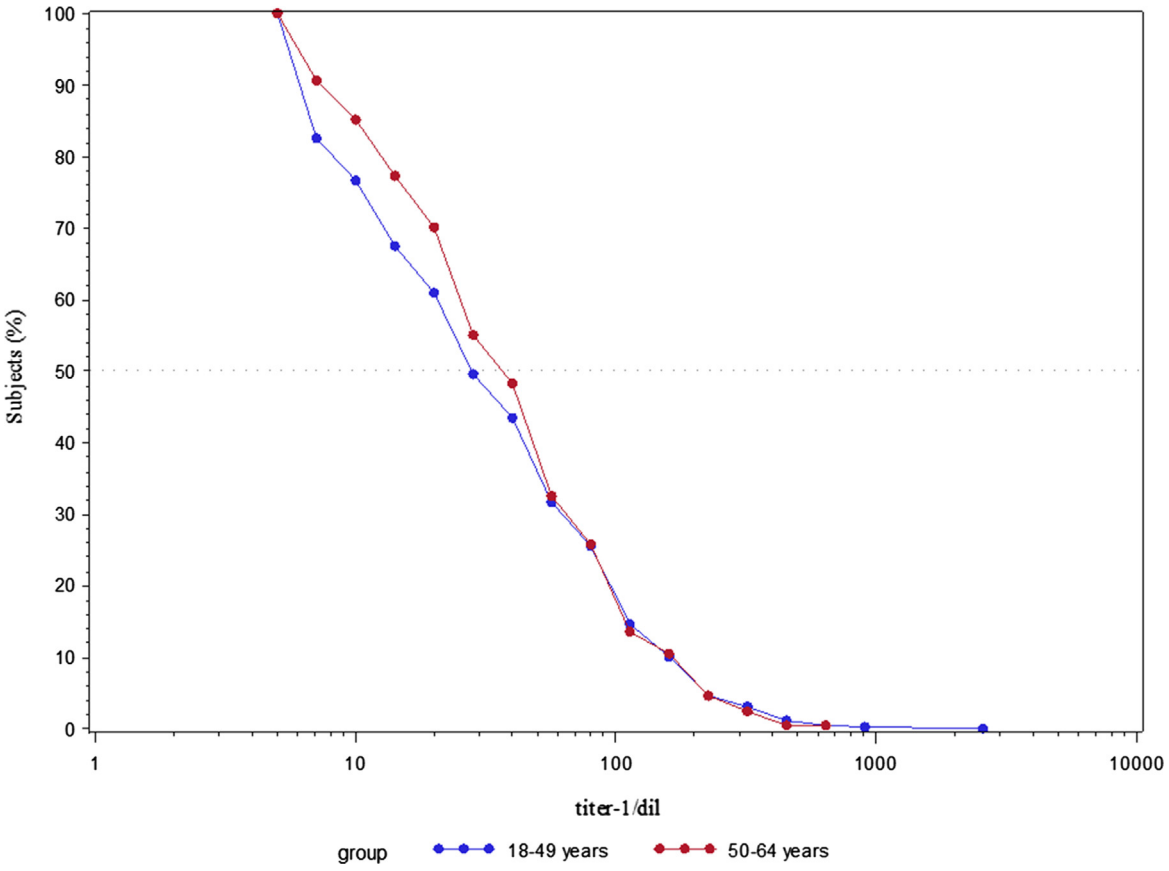

Fig. 3. RCDCs that deviate for lower values but converge for values above 50. The log rank and Wilcoxon tests give very different results.

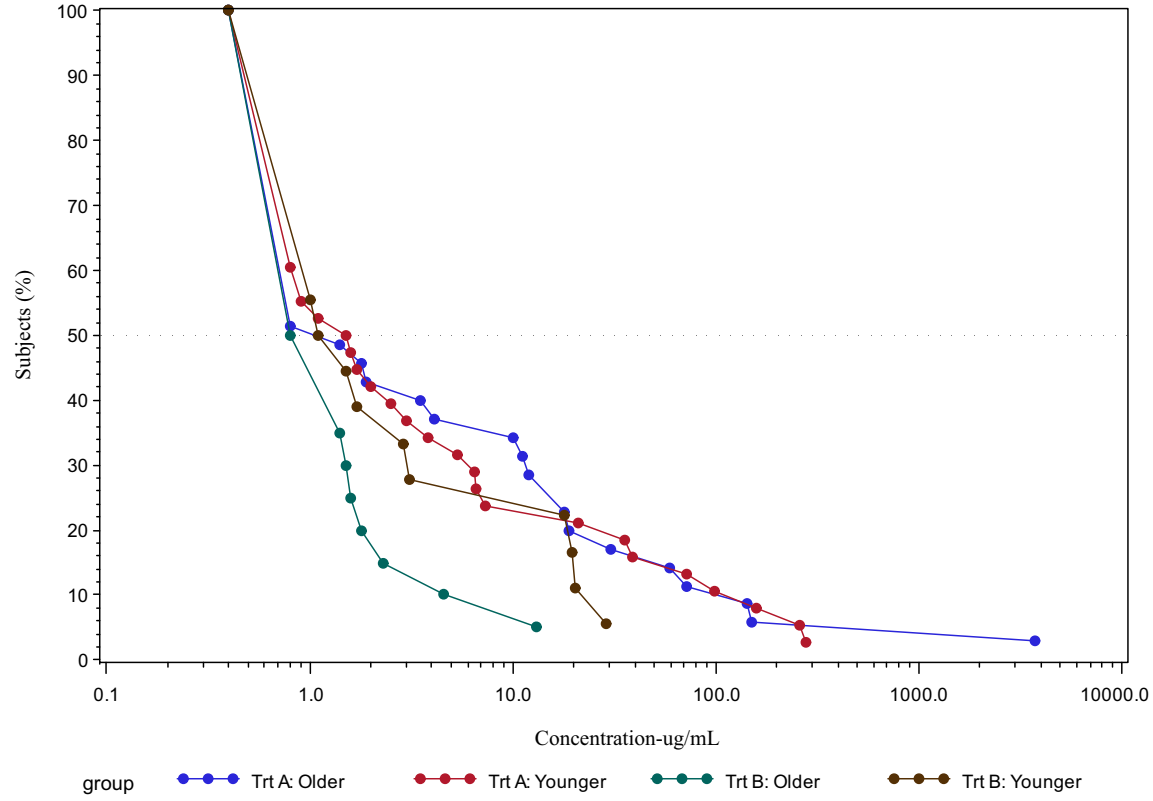

Fig. 4. RCDCS of two treatments in two age groups. One treatment (A) is a vaccine and the other is a placebo (B).

Table 1
Test for normality on each of the four antigens using the KS statistic.

Antigen	KS Statistic	P-Value
1	0.099	<0.01
2	0.095	<0.01
3	0.093	<0.01
4	0.092	<0.01

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.vaccine.2019.09.007.

References

[1] Ball LK, Falk LA, Horne AD, Finn TM. Evaluating the immune response to combination vaccines. Clin Infect Dis 2001;33:S299–305.
[2] De Bryun G, Saleh J, Workman DR, Elinoff V, Fraser NJ, et al. H-030-012 Clinical Investigator Study Team. Defining the optimal formulation and schedule of a candidate toxoid vaccine against Clostridium difficile infection: A randomized Phase 2 clinical trial. Vaccine 2016;34(19):2170–78.
[3] Gorse GJ, Falsey AR, Ozol-Godfrey A, Landolfi V, Tsang PH. Safety and immunogenicity of a quadrivalent intradermal influenza vaccine in adults. Vaccine 2015;33:1151–9.
[4] Harrington DP, Fleming TR. A class of rank test procedures for censored survival data. Biometrika 1982;69:133–43.
[5] Henckaerts I, Goldblatt D, Ashton L, Poolman J. Critical differences between pneumococcal polysaccharide enzyme-linked immunosorbent assays with and without 22f inhibition at low antibody concentrations in pediatric sera. Clin Vaccine Immunol 2006;13:356–60.
[6] Hess K. Graphical methods for assessing violations of the proportional hazard assumption in cox regression. Stat Med 1995;14:1707–23.
[7] Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. 2nd ed. New York: John Wiley and Sons; 2002.
[8] Lachenbruch PA, Rida W, Kou J. Lot consistency as an equivalence problem. J Biopharm Stat 2004;14:275–90.
[9] Lehmann EL. Parametric versus nonparametrics: two alternative methodologies. J Nonparametric Statist 2009;21(4):397–405.
[10] Mehrotra Devan V, Lu Xiaomin, Li Xiaoming. Rank-based analyses of stratified experiments: alternatives to the van Elteren test. Am Statist 2010;64:121–30.
[11] O'Brien KL, Moïsi J, Moulton LH, Madore D, Eick A, Reid R, et al. Predictors of pneumococcal conjugate vaccine immunogenicity among infants and toddlers in an American Indian PnCRM7 efficacy trial. J Infect Dis 2007;196:104–14.
[12] Nauta Jozef. Statistics in clinical vaccine trials, vol. 1. London New York: Springer; 2010.
[13] Nauta JJP, Beyer WEP, Osterhaus ADME. On the relationship between mean antibody level, seroprotection and clinical protection from influenza. Biologicals 2009;37:216–21.
[14] Reed GF, Meade BD, Steinhoff MC. The reverse cumulative distribution plot: a graphic method for exploratory analysis of antibody data. Pediatrics 1995;96:600–3.
[15] Saul A, Fay MP. Human immunity and the design of multi-component, single target vaccines. Public Library Sci one 2007;2(9):e850.
[16] Schoenfeld D. The asymptotic properties of nonparametric tests for comparing survival distributions. Biometrika 1981;68:316–9.
[17] Schuerman L, Wysocki J, Tejedor JC, Knuf M, Kim KH, Poolman J. Prediction of PNEUMOCOCCAL conjugate vaccine effectiveness against invasive pneumococcal disease using opsonophagocytic activity and antibody concentrations determined by enzyme-linked immunosorbent assay with 22F adsorption. Clin Vaccine Immunol 2011;18:2161–7.
[18] Tarone TE, Ware J. On distribution –free tests for equality of survival distributions. Biometrika 1977;64:156–60.
[19] van Elteren PH. On the combination of independent two sample tests of wilcoxon. Bull Inst Int Stat 1960;37:351–61.